UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

CAUSAL TECHNIQUES FOR BANKRUPTCY PREDICTION IN GREEK SMES

Diploma Thesis

Leonidas Zimianitis

Evangelia-Rafaela Frastali

Supervisor: Elias Houstis

September 2023

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

CAUSAL TECHNIQUES FOR BANKRUPTCY PREDICTION IN GREEK SMES

Diploma Thesis

Leonidas Zimianitis

Evangelia-Rafaela Frastali

Supervisor: Elias Houstis

September 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΕΧΝΙΚΕΣ ΑΙΤΙΩΔΟΥΣ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΠΤΩΧΕΥΣΗΣ ΣΕ ΕΛΛΗΝΙΚΕΣ ΜΜΕ

Διπλωματική Εργασία

Λεωνίδας Ζημιανίτης

Ευαγγελία-Ραφαέλα Φράσταλη

Επιβλέπων: Ηλίας Χούστης

Σεπτέμβριος 2023

Approved by the Examination Committee:

Supervisor **Elias Houstis**
Emeritus Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Emmanouil Vavalis**
Professor, Electrical and Computer Engineering, University of Thessaly

Member **Michael Gr. Vasilakopoulos**
Professor & Chairman of Department, Electrical and Computer Engineering, University of Thessaly

Acknowledgements

As a team, we would like to express our most profound gratitude to our Professor, Elias Houstis, who has reignited our passion for science through his academic teachings and life lessons. He has always helped us find purpose by assigning duties to us. We hope to honour him with our deeds, as he has honoured us by being our teacher.

We would also like to extend our heartfelt thanks to Professor Evangelos Rasvanis for his constant support and guidance, which have been invaluable in writing this thesis. He always provided constructive criticism on our work and has helped us enhance our research skills to evolve and become better scientists.

Next, we are deeply grateful to Professor Emmanouil Vavalis and Professor Michail Vasilakopoulos for graciously agreeing to be our supervisors and participating in evaluating this thesis. Moreover, we wish to recognise our families' endless love and support, without which none of this would have been possible.

Last, we want to express profound gratitude to each other. Though marked by differences, our partnership has proven to be very productive, and we achieve far more together than we ever could have done.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarants

Leonidas Zimianitis and Evangelia-Rafaela Frastali

Diploma Thesis

CAUSAL TECHNIQUES FOR BANKRUPTCY PREDICTION IN GREEK SMES

Leonidas Zimianitis

Evangelia-Rafaela Frastali

Abstract

Understanding the factors leading to bankruptcy is essential in the dynamic landscape of global economies. While many research efforts prioritize prediction, a comprehensive dive into causality inference is critical. With this perspective, our study aimed to uncover causal links among dataset variables, focusing on their impact on bankruptcy outcomes. For this study, we considered the bankruptcy Greek SMEs dataset [1] in which we discovered biases inconsistent with the economic backdrop of Greece. A standout feature of our methodology is its emphasis on causality to clarify and correct dataset discrepancies. This technique, coined by the term ”Construction Bias,” refers to the intentional or unintentional variable choices made during dataset creation. We applied our methodology to the bankruptcy Greek SMEs dataset in which the bankruptcy firms were underrepresented. After refinement, our strategy outperformed the traditional random sampling approach’s predictive capability, boasting a noteworthy 95.71% precision in bankruptcy predictions. This level of accuracy highlights the value of pinpointing companies in potential financial distress using a causality-centric analytical perspective.

Keywords:

Greek SMEs, Bankruptcy, Machine Learning, Causal inference, Bayesian Networks, Structure learning

Διπλωματική Εργασία

ΤΕΧΝΙΚΕΣ ΑΙΤΙΩΔΟΥΣ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑΣ ΓΙΑ ΤΗΝ ΠΡΟΒΛΕΨΗ ΠΤΩΧΕΥΣΗΣ ΣΕ ΕΛΛΗΝΙΚΕΣ ΜΜΕ

Λεωνίδας Ζημιανίτης

Ευαγγελία-Ραφαέλα Φράσταλη

Περίληψη

Εντός του δυναμικού τοπίου των παγκοσμίων οικονομιών, η βέλτιστη κατανόηση των παραγόντων οι οποίοι οδηγούν σε πτώχευση είναι απαραίτητη. Ενώ η πλειονότητα των υπαρκτών ερευνών επικεντρώνεται στην πρόβλεψη πτώχευσης, η εμβαθυμένη εξερεύνηση της αιτιατότητας παραμένει υπό ανάπτυξη. Σκοπός της έρευνας μας είναι η ανακάλυψη των σχέσεων αιτιατότητας μεταξύ μεταβλητών, σε ένα ελληνικό οικονομικό σύνολο δεδομένων, εστιάζοντας στην επιρροή την οποία οι ίδιες ασκούν στην πτώχευση. Μια πρώιμη ανάλυση, μας αποκάλυψε πως οι τάσεις του συνόλου δεδομένων δεν αντικατοπτρίζουν την ελληνική πραγματικότητα. Την κομβική πρωτοπορία της έρευνάς μας, αποτελεί η χρήση μεθόδων αιτιατής συμπερασματολογίας για την ποσοτικοποίηση της υπαρκτής μεροληψίας εντός του συνόλου δεδομένων. Κατά συνέπεια, γεννήθηκε ο όρος ”Κατασκευαστική Μεροληψία”, συμβολικός για τη συστηματική συμμετοχή ή αποχή εκάστοτε περιπτώσεων κατά τη δημιουργία του συνόλου δεδομένων. Βάσει αυτών, έπειτα μιας εξευγένισης του συνόλου δεδομένων, η τεχνική μας ξεπέρασε σε απόδοση την πολλών-ετών εγκαθιδρυμένη τεχνική απόσπασης μεροληψίας με χρήση τυχαίας δειγματοληψίας, επιτυγχάνοντας μια εντυπωσιακή ακρίβεια της τάξεως του 95.71% για την πρόβλεψη πτωχευμένων εταιρειών. Η υψηλή αυτή απόδοση αποτελεί μάρτυρα της ανάγκης για χρήση τεχνικών αναγνώρισης επικυνδινότητας πτώχευσης με οδηγό τεχνικές αιτιατότητας.

Λέξεις-κλειδιά:

Ελληνικές ΜμΕ, Πτώχευση, Μηχανική Μάθηση, Αιτιώδη Συμπερασματολογία, Μπεϋζιανά Δίκτυα, Δομική Μάθηση

Table of contents

List of figures

List of tables

Abbreviations

AANN	Auto-associative neural network
AHC	Agglomerative Hierarchical Clustering
AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ARC	Assessment of Rare Causes
ATC	Average Treatment Effect on the Controls
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
BART	Bayesian Additive Regression Trees
BNs	Bayesian Networks
BP	Bankruptcy prediction
BPNN	Back Propagation Neural Network
CATE	Conditional Average Treatment Effect
CBR	Case-Based Reasoning
CDA	Canonical Discriminant Analysis
CF	Causal Forest
CPD	Conditional Probability Distributions
CRL	Causal Representation Learning
DA	Data Augmentation
DA	Discriminant analysis
DAG	Directed Acyclic Graph
DEA	Data Envelopment Analysis
DT	Decision Trees
EDA	Exploratory Data Analysis

ETE	Effective Transfer Entropy
EWS	Early warning System
FC	Feature Construction
GA	Genetic Algorithms
GBM	Gradient-Boosted Machines
GP	genetic programming
IEWS	Integrated Early Warning System
K-NN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage Selection Operator
LDA	Linear Discriminant Analysis
LICC	Linear Intertemporal Cross-Correlation
LR	Logistic Regression
MAP	Maximum a Posteriori
MATE	Mixed average treatment effect
MCMC	Markov Chain Monte Carlo
MDA	Multiple Discriminant Analysis
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLP	Multi-Layer Perceptron
N-FC	Neuro-Fuzzy Classifier
NB	Naïve Bayes
NICC	Nonlinear Intertemporal Cross-Correlation
NNT	Number Needed to Treat
PATE	Population average treatment effect
PF	Particle Filtering
QDF	Quadratic Discriminant Function
RC	Rough Classifier
RCE	randomized controlled experiment
RPA	Recursive Partitioning Algorithm
S.e.	Standard Error
SATE	Sample average treatment effect
SBCNs	Suppes-Bayes Causal Networks

SCMs	Structural causal models
SEMs	Structural Equation Models
SFO	Structured Finance Operating
SMEs	Small-Medium Enterprises
SOM	Self-Organizing Maps
SUTVA	Stable Unit Treatment Value Assumption
SV	Stochastic Volatility
SVM	Support Vector Machines
TAN	Tree-augmented Naive Bayes
TiMINo	Time Series Models with Independent Noise
VCR	Valued Closeness Relation
iid	Independent and identically distributed random variables

Chapter 1

Introduction

In the dynamic world of global economies, the stability and growth of firms are paramount. Their prosperity or decline can ripple across industries, stakeholders, and economies. Among the myriad challenges businesses face, bankruptcy is a formidable adversary. As such, comprehending the factors that precipitate bankruptcy is of utmost importance.

Historically, the quest to predict and understand bankruptcy has been marked by the use of diverse methodologies, ranging from traditional statistical methods to contemporary machine learning algorithms. However, while many endeavours focus on prediction, a deeper understanding of causality, i.e., the factors that genuinely lead to bankruptcy, remains a frontier yet to be fully explored.While predicting its occurrence is invaluable, illuminating the underlying causes can offer transformative insights.

This introduction provides a glimpse into the landscape of financial stability, bankruptcy prediction, and the unexpected turns that research can take. As we navigate the challenges and revelations of our study, we invite the reader to engage with the evolving narrative, appreciating the intricacies and insights of our findings.

1.1 Thesis Scope

Initially, our objective was to employ causal models to shed light on the intricacies of bankruptcy, especially concerning Small and Medium-sized Enterprises (SMEs) in Greece, a nation with a distinctive economic landscape. However, as often occurs in research, we encountered unforeseen challenges. The quality and quantity of available data were not as anticipated, leading to a pivotal shift in our approach. Instead of being a limitation, this chal-

lenge enriched our journey, prompting us to innovate and adapt.

The core objective of our analysis was to establish causal relationships amidst the myriad variables present in the dataset, focusing on their influence on the outcome variable referred to as 'label'. Given the data's inherent biases, our initial analyses divulged unexpected relationships that were not congruent with our prior knowledge and the economic context of Greece. Consequently, our research journey evolved into deciphering these biases' presence, impact and mitigation.

Note on Citations

Citations within this thesis, especially those in the experiments records, do not indicate that we have directly duplicated the information from the cited materials. Instead, the referenced literature has been used as a foundation upon which we have formulated and presented our conclusions

1.1.1 Key Findings

In the following, we outline the results of our thesis,

1. **Inherent Biases**

   Our initial analyses exposed a pronounced bias in the dataset, with specific fiscal years exclusively representing bankrupt firms. Such data idiosyncrasies can severely hinder the establishment of genuine causal relationships and compromise the reliability of conclusions drawn from the data.

2. **Construction Bias**

   Delving deeper into the biases, we conceptualized the term "Construction Bias" to describe the systematic inclusion or exclusion of instances based on the outcome variable during the dataset construction phase. This bias is insidious, as it can camouflage as a genuine relationship between variables.

3. **Mitigating Construction Bias**

   By carefully selecting instances from years that did not display overt biases and ensuring balanced representation for both outcome labels, we crafted a dataset that significantly reduced the construction bias.

4. **Model Efficacy Post-Debiasing**

   The models trained on the de-biased dataset exhibited enhanced performance, especially in predicting bankruptcies. The precision for predicting 'label 1' (associated with bankrupt firms) reached an impressive 95.71%, emphasizing the model's capability. This precision is paramount since it is more crucial to accurately predict which firms will go bankrupt than to determine which ones will not.

1.2 Contribution of Thesis

This thesis makes several significant contributions. More specifically,

- it applies an extensive Exploratory Data Analysis (EDA) to the bankruptcy dataset for the Greek Small to Medium Enterprises [1], considered the largest and most extensively studied dataset for the Greek market.
- it uses causal inference, which does not rely on mere correlation metrics but delves into the fundamental relationships between variables. In doing so, it complements the Exploratory Data Analysis.
- it pioneers its application of causal techniques to Greek data and uses causal inference to quantify bias.
- it leverages the proposed causal debiasing method, that surpasses the well-established random sampling technique in eradicating bias

1.3 Thesis organisation

The thesis consists of eight chapters. The first chapter serves as the introduction. Chapter 2 delves into Causal Inference, while Chapter 3 introduces Bayesian Inference Techniques in the context of Machine Learning (ML) and establishes the theoretical background. Chapter 4 provides an extensive literature review on bankruptcy models and presents the models discussed. Chapter 5 describes the literature research that informed the implementation of Causal and Bayesian techniques in bankruptcy prediction. Chapter 6 offers detailed information about the dataset used in our study and the Exploratory Data Analysis (EDA) we conducted. In Chapter 7, causal machine learning models are developed and applied to our

custom datasets, with the accuracy of each method being interpreted. Finally, Chapter 8 concludes the study, discussing the challenges faced, and suggesting future directions.

Chapter 2

Causal Inference: From Correlation to Causation

This chapter explains causal inference, emphasizing its importance beyond mere observational data. We start by discussing why causal inference is crucial, followed by its mathematical basis. We'll explore how graphs can represent associations and review various techniques used in causal research. Later sections will address distinguishing actual causal effects, handling confounding data, and the emerging area of causal representation learning. The chapter concludes with foundational concepts in causal theory.

2.1 An Overview of Causal Inference

Causal inference is a discipline that considers assumptions, study designs, and estimation strategies allowing researchers to draw causal conclusions based on data [9]. It is a field founded in statistics that first gained popularity in the 1970s by researchers studying the causal effects of treatments in the medical sciences [10]. Later it is combined with machine learning creating Causal Machine Learning, a field of machine learning methods that strive to find the causal relationships between the variables and the data. But how did we go from machine learning to Causal Machine Learning? In the classical machine learning (ML) problems, we assume that the data are independent and identically distributed, thus in reality that is not always true. The result of considering the data independent is that even when using the most sophisticated and state-of-the-art methods the causality of the data is overlooked [11]. As a result, the outcomes of classic ML can make various predictions with high accu-

racy, but they are entirely unable to explain the outcome, letting the task explanation to the analyst, which might be inaccurate or subjective. Thus, causal ML solves this as it offers explainability and humans easily interpret its models [12].

Knowing the causal relations between the data is crucial when considering data analysis. Identifying the mechanisms in which the variables take the values they have is a key aspect of causal inference as it is not always possible to know the actual value of a variable so it is crucial to have a way to estimate it. If experimental data is not accessible, filling in the variable values usually requires taking samples from one probability distribution and inferring a variable's value in a population with a different probability distribution [13]. Moreover, even if the process of experimental interventions is possible, there are things to consider. That is very inefficient, as a large amount of data is needed to understand causal relationships between variables to perform those experiments, or sometimes is probably unethical to suggest specific interventions. For example, we could measure the effect that smoking would have on the blood pressure of a sample, but it is highly unethical to make people start smoking even for scientific reasons.

2.2 Motivation: Reasons to Employ Causal Inference

Causal inference is a fundamental aspect of scientific research. While associational (or correlational) claims show that variables are related, causal claims further suggest that one variable directly influences or causes another. That causal reasoning has many scientific applications, like estimating the effects of medical treatments, training robots to perform actions that will have the best outcome, understanding the causes of a socioeconomic incident, or even for decision-macking [3].

Moreover, correlation does not imply causation. While correlation indicates a linear statistical dependency between two variables, it does not establish a causal relationship where one variable directly influences the outcome of another. In figure 2.1, we observe a surprising correlation between two seemingly unrelated topics: the divorce rate in Maine and the consumption of margarine. These two variables are highly correlated at 99.26%. However, asserting that margarine consumption directly influences the divorce rate would be irrational. Although these two variables appear to be correlated, drawing a causal link between them would be misleading. We must be cautious, as humans can sometimes be prone to believing

such spurious correlations.

Contrary to what one might think, traditional statistics and machine learning haven't solved the 'causality' problem. Measuring causation isn't as straightforward as looking at correlation and predictive performance in data. In traditional statistics and machine learning, we cannot determine causation solely by examining metrics like correlation and predictive accuracy. This limitation underscores the importance of not conflating correlation with causation. Due to these concerns, researchers are delving deeper into novel causal inference methods.

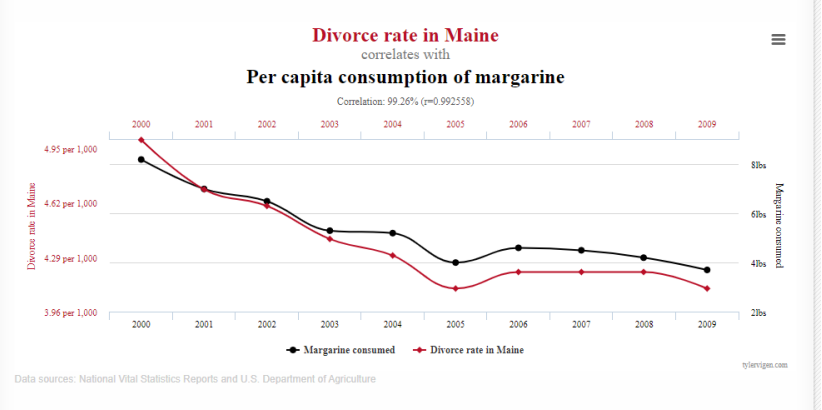


Figure 2.1: Spurious Corellations [2]

2.3 Setting the Groundwork: Mathematical Foundation of Causal Inference

Humans inherently understand causal inference as a survival mechanism. For instance, if members of a tribe ate a specific seed and died, while others who didn't consume it survived, the remaining tribe members would likely avoid eating that seed. But how can we explain this concept to someone without intuition about causality?

A causal effect can be defined as follows: When comparing the outcome of taking action A to the outcome of not taking action A, if the outcomes differ, we say that action A has a causal effect on the outcome. In statistical terms, action A can be referred to as an intervention, exposure, or treatment. If outcome Y is a dichotomous variable, it can have two values (Y=0 or Y=1). Treatment A has a causal effect on the outcome Y of an individual if $Y^{A=1} \neq Y^{A=0}$. The variables $Y^{A=1}$ and $Y^{A=0}$ are called counterfactual outcomes. If $Y^{A=1} = Y^{A=0}$ and thus

A has no causal effect the equality $Y^A = Y$ is called consistency. Individual causal effects are described by contrasting the values of counterfactual outcomes. However, for each individual, only one outcome is observed — the one corresponding to the treatment value experienced by that individual. All other counterfactual outcomes remain unseen. Because of this missing data, individual effects can't be identified; in other words, they can't be represented as a function of observed data [14].

2.4 Craphs and the Flow of Association

We explain in detail in the following section that DAGs which are a kind of graph are the structure in which we can depict causation and association between variables. So, this paragraph gives the basics of how graphs are created and their theory. Graphical building blocks are the basic components of graphs. These minimal building blocks are :

- Chain
- Fork
- Immorality
- Two unconnected nodes
- Two connected nodes

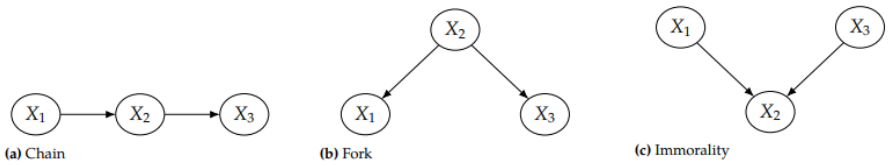


Figure 2.2: Chain Fork Immorality [3]

Figure 2.3: Two unconnected nodes [3]

Figure 2.4: Two connected nodes [3]

The flow of association is whether two nodes are associated or not in a graph. Associated we consider the statistically dependent nodes. Statistically independent nodes are not linked

in graphs. Thus we say that two nodes are not associated simply because they have no edge. In contrast, if there is an edge between two nodes, then the two nodes are associated. Regarding nodes that are not directly linked but are statistically dependent, they are conditionally dependent. In forks and chains like the ones in 2.2 typically X1 and X3 are dependent. In a chain, X1 affects X2 and then X2 affects X3. In a fork X1 and X3 are associated because changes in X2 can influence both X1 and X3. By analyzing the graph's structure, we determine which variables are conditionally independent of each other and which are not, which is crucial for causal inference. Though those associations are usually true, there are pathological cases where X1 and X2 are not associated [3]. Finally, association flow represents symmetrical correlation, meaning that if X1 affects X2 and X2 affects X1. The flow of causation is not symmetric because causation only flows in directed paths. Understanding the difference between the two is crucial in research and data analysis to avoid drawing incorrect conclusions.

2.5 Key Techniques: A Review of Popular Causal Inference Methods

Causal inference is usually separated into two broad types of models used to estimate interventions' effects [13].

- causal Bayesian networks
- Structural Causal Models

Causal Bayesian networks specify a density for a variable as a function of the values of its causes. On the other hand, Structural Equation Models (SEMs) specify the value of a variable as a function of the values of its causes (typically including some unmeasured noise terms) [13].

2.5.1 Causal Bayesian Networks

The causal relationships are graphically represented using a structure called a Causal Directed Acyclic Graph (Causal DAG). Graphs are structures consisting of a collection of nodes and edges that connect these nodes [15]. Some graphs are directed, meaning the edges

point from one node to another. A Directed Acyclic Graph (DAG) is a directed graph with no directed cycles. This means that there is no path of edges in the graph that starts and ends at the same node, node A, for example.

Graphs are especially useful for modeling probability distributions over random variables. They visually demonstrate how a joint distribution over a set of random variables can be factorized. This factorization relies on the chain rule of probability, which permits the decomposition of a joint distribution into a product of conditional distributions. This decomposition occurs due to a fundamental property of graphical models known as the Markov Condition. The Markov Condition results in the factorization of the joint distribution of all nodes in a graph into a product of conditional distributions, where each node's distribution is conditioned on its parent nodes. Mathematically, the Markov Condition is represented as follows:

Theorem 2.1. *Given a graph G of nodes X with joint distribution p(x), the Markov Condition n states that the parents pa_i of every node X_i make X_i independent of its non-descendants. This condition implies the factorization of the joint distribution*

$$p(x) = \prod_i p(x_i|pa_i)$$

This equation states that the joint distribution p(x) can be represented as a product of the conditional distributions of each node x_i given its parents pa_i. In simpler terms, the Markov Condition and the resulting factorization allow us to simplify the representation of a joint distribution using a graph. Instead of considering the relationships between all pairs of variables, we only need to consider the relationships between a node and its parents. This greatly reduces the complexity of the model, especially when dealing with a large number of variables. Finally, this is the main advantage of Bayesian Networks, they can decompose a large joint distribution p(x) into a product of several small conditional ones according to the assumed DAG relations. Consequently, Bayesian Networks leverage the structure of the DAG to provide a compact and interpretable representation of complex joint distributions. This allows efficient reasoning and learning in probabilistic domains.

2.5.2 Structural Causal Models

Previously we discussed how causal BNs enable the transition from associational distributions found in regular BNs to interventional ones. However, causal BNs fall short when

it comes to constructing counterfactual distributions. This is where Structural causal models (SCMs) come in handy. SCMs provide a formalism for causation that allows counterfactual analysis. They are also called structural equation models or functional causal models [15].

In SCMs, causal relationships are represented using deterministic functional equations. Stochasticity (or randomness) in SCMs is introduced based on the hypothesis that certain variables in the equations remain unobserved. In essence, SCMs offer a more comprehensive framework that includes counterfactual analysis. This is achieved by expressing causal relationships through deterministic equations and introducing randomness based on unobserved variables. Structural causal models use structural equations to represent causality. The equation $A = B$ is symmetric, meaning it can also be expressed as $B = A$ which implies not only that a change in A affects B, but also that a change in B affects A. This symmetry does not indicate any causal direction. To discuss causation, an asymmetric relationship is essential. This asymmetry allows one to claim that ”A is a cause of B”. In this context, it means that changing A will result in changes in B. However, altering B will not lead to changes in A. This Structural Equation for Causation will be:

$$B := f(A)$$

where f is some function that maps A to B. Although we have this rule to describe causal relationships, it is not entirely correct; this formation is deterministic meaning that only A can change B. To have a probabilistic rule that better explains the mapping when some unknown factors may also affect B we have:

$$B := f(A, U)$$

here U is some unobserved random variable. U represent the randomness or noisy data (a “noise” or “background conditions” variable) as input, it can depict any stochastic mapping, so structural equations generalize the probabilistic factors [3].

2.6 Average causal effect

The definition of an individual causal effect requires three pieces of information:

- An outcome of interest
- The actions (A = 1 and A = 0) that are being compared.

- The specific individual for whom the counterfactual outcomes ($Y^{A=0}$ and $Y^{A=1}$) are compared

However, identifying the individual causal effects is typically not feasible. That is because we do not know what will happen for both actions as the individual can only take action 0 or 1. While individual causal effects provide insights into a specific individual's outcomes under different actions, average causal effects offer a broader perspective by considering the average outcomes across a population under the same actions [14]. This shift is necessary due to the challenges of pinpointing individual causal effects.

The average treatment effect is widely used by scientists. The simplest example would be if we have an individual who is sick. If, hypothetically, we have two treatments to give them, the doctors are interested in the outcome of the individual but they cannot provide both treatments to see the individual treatment effect. Therefore, they study the literature to see what percentage of a population has a positive outcome given treatment A = 1, and what percentage of a different population has a positive outcome given treatment A = 0. If treatment A has a positive outcome on a larger number of people, the doctors would suggest that the individual take treatment A = 0 based on the average causal effects.

Therefore, the average causal effect of treatment A on an outcome Y exists if the probability that the outcome occurs under the treatment, Pr[$Y^{A=1} = 1$], is not equal to the probability of the outcome occurring without the treatment, Pr[$Y^{A=0} = 1$], in the target population

$$Pr[Y^{A=1} = 1] \neq Pr[Y^{A=0} = 1]$$

The benefit of that approach is that the average treatment effect (ATE) can be specified from data, while causal effects cannot. Subsequently, sometimes we even see in the literature that 'average causal effects' are referred to as 'causal effects'.

2.6.1 Ignorability and Exchangeability

When we previously presented the ATE, we assumed ignorability, meaning we essentially disregarded the reasons or mechanisms behind individuals' choices of treatment. Instead, the authors simplified the assumption that individuals were randomly assigned treatments. In simpler terms, ignorability means that the individual and his characteristics do not influence the treatment selection. Previously, we also assumed exchangeability when discussing the average causal treatment effect. Exchangeability means that we would see the same results

if we were to exchange the treatment groups, meaning that the treatment would have the same effect in every group, regardless of the group's characteristics. However, we can state exchangeability only when the treatment groups being compared are essentially identical in all respects, except for the treatment they received. This means that any difference in outcomes observed between the groups can be attributed to the treatment itself and not to some other underlying difference between the groups. However, in real-world scenarios, assuming ignorability can often be unrealistic. This is because most observational data is likely to have confounding variables. These confounders can influence both the assignment of treatment and the outcome, leading to potential biases in estimating causal effects. The only way to ensure ignorability is to perform Randomized Experiments [3]. In a randomized experiment, the assignment of treatment is determined by a random mechanism, such as a coin toss. This guarantees that treatment is not influenced by any other factors, including potential confounders.

2.6.2 Conditional Exchangeability or Unconfoundedness

In observational data, assuming that treatment groups are exchangeable is typically unrealistic. This means that the groups may differ in aspects other than the treatment they received. However, it might be possible to exchange subgroups by controlling for certain relevant variables (conditioning) [3]. In simpler terms, although the treatment and the outcomes are associated with confounders by controlling the treatment groups, we make them comparable, and thus there is no confounding. Does that mean that the controlled data have Exchangeability? No, but we have conditional exchangeability in the data, a crucial assumption for causal inference. Conditional Exchangeability is also known as unconfoudness.

Conditional Exchangeability is when we condition the data to diminish any noncausal association between the treatment and the outcome. Even when controlling the data unconfounded is not secured because there may exist some unobserved confounders that are not related to the data. The only way unconfoundedness is secured is when the data come from randomized experiments.

2.6.3 Positivity

The Positivity assumption (also known as the Overlap or Common Support) is a fundamental requirement in the average causal treatment effect. The positivity assumption ensures

that every subgroup (defined by covariates) has a mix of treated and untreated individuals, allowing for meaningful causal effect estimation across the entire dataset.

Given a set of covariates X and a binary treatment variable T (where $T = 1$ indicates treatment and $T = 0$ indicates control), the positivity assumption can be mathematically represented as:

$$0 < P(T = 1|X = x) < 1$$

for all values of x in the support of X.

This equation ensures that for every combination of covariates x, there is a non-zero probability of receiving both the treatment and the control. In other words, for every subgroup defined by x, some individuals receive the treatment and others don't. Without positivity, we can't compare outcomes between treated and untreated individuals within subgroups, making causal effect estimation impossible for those subgroups.

2.6.4 No Interference and Consistency

No interference, often referred to as the Stable Unit Treatment Value Assumption (SUTVA) in causal inference literature, is the assumption that the treatment assignment of one individual does not affect the outcome of another individual and that the outcome of the individual is a function of his treatment [3].

The final assumption is consistency. Consistency is the assumption that the outcome we observe Y is the potential outcome under the observed treatment T.

If the treatment is T, then the observed outcome Y is the potential outcome under treatment T. Formally

$$T = t \longrightarrow Y = Y(t) \longrightarrow Y = Y(T)$$

The intuition behind consistency is to ensure that what we observe is an outcome of the given treatment and not some random effect.

2.7 Deconfounding Data

As we briefly mentioned, confoundedness is probably the main problem when dealing with causal inference as it makes learning the true causal relationships through variables complex. It immediately comes to mind that trying to deconfound the data would tackle this

problem and make causal analysis easier. In observational studies where data are not collected from random experiments, data augmentation ensures that the relations between the variables are genuine and not due to confounding factors. Data Augmentation (DA) is usually applied in high dimensional datasets; it is a set of interventions that are happening in the dataset without causing any information loss though they deconfound the dataset. The negative of DA is that it adds much computational cost to the dataset and it is essential to apply augmentation judiciously to ensure that the augmented data remain relevant and meaningful.

2.8 Causal Representation Learning

Representation Learning is extracting representation from the data X, but the representations have lower dimensionality. Causal Representation Learning (CRL) is an emerging area in machine learning that focuses on learning representations that capture the underlying causal structure of the data. The idea is to identify and represent the high-level causal variables that generate the observed data X. These representations Z, correspond to instances of these typically latent causal variables [15]. The steps to perform CRL:

1. Causal Feature Learning: a mapping $X = g(X)$
2. Causal Graph Discovery: a causal graph G_Z among the causal variables Z
3. Causal Mechanism Learning: Understand the underlying mechanisms that drive these relationships

2.9 The Causal Theory basics

2.9.1 The Causal Null Hypothesis

The causal null hypothesis is the condition that there is no effect of one variable (treatment or exposure) on another (the outcome). Saying that the causal null hypothesis holds is equivalent to saying that variable A does not have a causal relationship with the outcome. There are another 3 ways of expressing no causal relationship:

- the causal risk difference equal to zero

$$Pr[Y^{A=1} = 1] - Pr[Y^{A=0} = 1] = 0$$

- the causal risk ratio equal to one

$$\frac{Pr[Y^{A=1}=1]}{Pr[Y^{A=0}=1]}=1$$

- the odds ratio equal to one

$$\frac{Pr[Y^{A=1}=1]/Pr[Y^{A=1}=0}{Pr[Y^{A=0}=1]/Pr[Y^{A=0}=0]}=1$$

The causal risk difference is a measure of the average individual causal effect but the causal risk ratio is a measure of causal effect on the population but is not the average of any individual [14].

2.9.2 Causal effect measures

We refer to causal risk difference, risk ratio, odds ratio, and other ratios that gauge the causal effect as Causal effect measures. They provide different perspectives on the effect of exposure or treatment on the risk of an outcome. The two most commonly used are the Causal Risk Ratio and the Causal Risk Difference. The first is used to show the strength of an association and how often treatment decreases the outcome risk compared to no treatment. The second is used to compute the absolute number of cases of the outcome attributed to treatment. The Odds Ratio expresses the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Another helpful ratio is the Number Needed to Treat (NNT) which gives the treatment's direct impact and we will explain it more in the next paragraph.

2.9.3 Number Needed to Treat

The Number Needed to Treat (NNT) is a statistical measure used normally in clinical research to convey the effectiveness of medical intervention. By definition, NNT is the number of individuals who need to receive treatment A=1 for the number of cases that have an outcome Y=1 to be reduced by one. An NNT of 2 means that one out of every two patients treated benefits. An NNT of 10 means one out of every ten patients treated benefits, and so on. The motivation behind NNT is that it gives a solid metric of how successful a treatment is, and thus is a commonly used metric for causality.

Chapter 3

A Comprehensive Study of Bayesian Inference within Machine Learning

This chapter delves into Bayesian Inference within the context of Machine Learning. We begin by discussing the progression from basic probability concepts to the more nuanced Bayes Rule. The chapter then highlights the compelling relationship between Bayesian Inference and Machine Learning, followed by a deep dive into its mathematical foundations, including topics like posterior probability and the structure of Bayesian causal inference. We also present an overview of prominent Bayesian techniques utilized in Machine Learning, from regression to structural time-series models. Concluding the chapter, we provide a list of libraries specifically tailored for Bayesian Causal Inference.

3.1 Transition from Simple Probabilities to the Bayes Rule

Bayes' theorem is a very important statistical tool with many applications in machine learning. Bayesian Machine Learning (ML) is an umbrella term that describes statistical ML models based on Bayes' theorem [16]. Classical machine learning offers deterministic predictive models that map the inputs with the outputs but do not provide explanations of their underlying processes. Thus, with Bayesian inference and Bayesian ML, we can expand our understanding of the variables and the model.

In this chapter, we will give the Bayesian approach regarding Causal inference. The chapter includes a review of the structure of Bayesian inference of causal effects and provides the necessary mathematical and statistical background needed to understand the aforementioned.

Although we try to cover the basics, we focus on the aspects of Bayesian inference that we consider more applicable to bankruptcy prediction. Finally, when considering the advantages and disadvantages of the Bayesian approach to causal inference, we approached the matter from an objective standpoint.

3.2 Motivation: The Synergy of Bayesian Inference and Machine Learning

As previously mentioned, the primary motivation for Bayesian Causal Inference stems from the limitations of classical machine learning in providing interpretable results. Bayesian Causal Inference allows us to determine cause-and-effect relationships between variables within a system. It is crucial to distinguish between variables that influence or directly affect the outcome and those that are simply associated with it. Understanding which variables are causally related to the outcome provides valuable information on which variables to manipulate (causal) and which to monitor (associative). Manipulating variables that have a causal relationship with the outcome aids in decision-making. Therefore, Bayesian inference not only seeks to predict the occurrence of events but also seeks to uncover the underlying mechanisms driving these events [17].

Bayesian inference allows for integrating prior knowledge or beliefs about a parameter or hypothesis. This prior information is combined with new data to produce a posterior distribution. This is particularly useful when data is scarce, when there is limited data available, or even when domain expertise provides valuable information that should not be ignored. If we consider a cause or a treatment as manageable, the identification and the ability to change it and affect the outcome can easily be adopted in many fields. In addition, the fact that Bayesian models can easily be updated as new data become available makes them suitable for online learning scenarios.

Another significant feature of the Bayesian approach is the flexibility that it offers. Bayesian inference provides a framework for integrating various prior distributions. As a result, Bayesian methods can be applied to general statistical analysis including automatic uncertainty quantification, complex hierarchical models, nonlinear models, and other structures that might be challenging for traditional frequentist approaches. They can adapt as more data becomes available, refining the posterior distribution. They can coherently incorporate prior knowl-

edge, and offer a rich collection of advanced models for complex data [18]. These reasons make the Bayesian approach suitable for real-world big data.

Finally, they are also very consistent. The Bayesian framework offers a consistent approach to uncertainty and decision-making. Every decision or inference is made based on a probability distribution. It treats unknown parameters as random variables, which is a more natural way to represent uncertainty.

3.3 Mathematical Foundations of Bayesian Inference

3.3.1 The posterior probability

Bayesian Inference is a technique based on Bayes' theorem that calculates the posterior probability. The posterior probability y is a conditional probability conditioned on randomly observed data. The prior probability and the likelihood of new data's occurrence define this probability's distribution [19]. Bayes' Theorem calculates the posterior probability, which is the probability that event A occurs, given that event B has occurred: [16]

$$P(A|B) = P(A)P(B|A)P(B)$$

where,

$P(A)$ =the prior probability of A occurring

$P(A|B)$ = the conditional probability of A given that B occurs

$P(B|A)$ =the conditional probability of B given that A occurs

$P(B)$ =the probability of B occurring

In Bayesian inference, the training data are treated as fixed, we aim to maximize the posterior distribution. This process is known as Maximum a Posteriori (MAP) estimation, and it is a method to obtain the most favourable values for the parameters based on prior beliefs.

3.3.2 The general structure of Bayesian causal inference

The structure of Bayesian causal inference was first introduced in [20]. There exist four quantities for every unit i. The quantities are $Y_i(0), Y_i(1), Z_i, X_i$ where $Z_i(= z)$ is the binary

variable that indicates the observed treatment status of unit i, X_i is a vector of covariates observed before treatment, and Y_i is the outcome observed after treatment. All the other quantities are observed, except $Y_i(1 - Z_i)$ which is missing. In the Bayesian inference structure, all four variables are considered random, and a model is built for them. By utilising the Bayesian model, we can make inferences on causal estimands. Causal estimands are functions of the model parameters, covariates, and potential outcomes. These inferences are drawn from the posterior predictive distributions of the parameters and the unobserved potential outcomes. We assume that the joint distribution of these random variables of all units is managed by a parameter $\theta = (\theta_x, \theta_y, \theta_z)$, conditional on the variable θ all the other variables are Independent and identically distributed random variables (iid) [18]. Then we can factorize the joint density

$$Pr\{Y_i(0), Y_i(1), Z_i, X_i|\theta\}$$

for each unit i as

$$Pr\{Z_i|Y_i(0), Y_i(1), X_i; \theta_z\} \cdot Pr\{Y_i(0), Y_i(1)|X_i; \theta_y\} \cdot Pr(Xi; \theta_x)$$

The three probability terms represent the model for the assignment mechanism, potential outcomes, and covariates, respectively. Under the assumption of ignorability, the assignment mechanism further reduces to the propensity score model $Pr(Z_i|X_i; \theta_z)$ [18] and then the joint density becomes

$$Pr(Z_i|X_i; \theta_z) \cdot Pr\{Y_i(0), Y_i(1)|X_i; \theta_y\} \cdot Pr(Xi; \theta_x)$$

In reality, we rarely model the multi-dimensional covariates X_i, it is more common to condition on the observed values of the covariates. This is the reason why most Bayesian causal inferences focus on the Mixed Average Treatment Effect (MATE). The MATE is an approximation of the Population Average Treatment Effect (PATE). The difference between MATE and SATE is subtle. The first one equals the average of the Conditional Average Treatment Effect (CATE), while the second one equals the average of the ITEs over a finite sample. Simply, MATE is like saying ”Given the varying treatment effects across different groups (like age groups), what is the average effect when we consider these differences?”, while SATE, on the other hand, is like saying ”If we were to average the treatment effect across all the individuals in our sample, what would it be?”

3.4 Key Bayesian Techniques in Machine Learning

3.4.1 Bayesian Regression

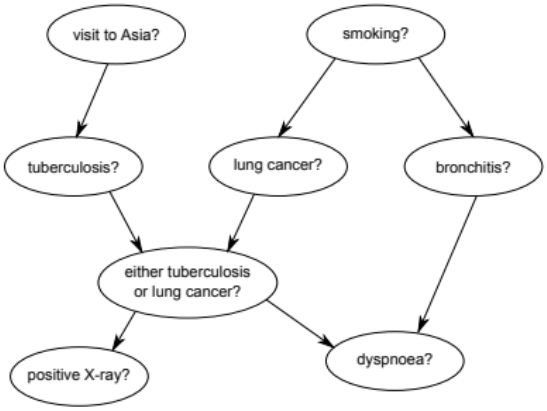
Linear Regression is a highly popular technique in machine learning. It operates under the assumption that the data follow a normal (Gaussian) distribution and that the variables exhibit a linear relationship with the outcome. But if the data validate this assumption, a better choice is to use Bayesian Regression. Bayesian Regression takes advantage of prior knowledge and Bayes' theorem to estimate the parameters of a linear regression model. Due to it being a probabilistic method, it can provide better estimates for the model parameters than OLS linear regression. Also, it is very useful if we have fewer data or if the data are poorly distributed. The output in the Bayesian regression is estimated from a probability distribution, whereas in linear regression, the outcome is estimated from a single value of each attribute [21]. In Bayesian Regression, the MAP estimation can also be used for model selection and outlier detection.

3.4.2 Bayesian Additive Regression Trees (BART)

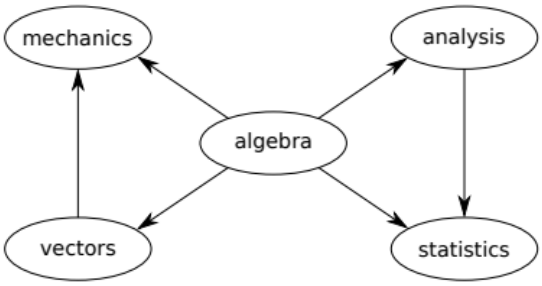
BART, which stands for Bayesian Additive Regression Trees, is a Bayesian approach to the estimation of nonparametric functions using regression trees [19]. Unlike parametric methods, nonparametric methods such as BART do not make explicit assumptions about the functional form of the data likelihood [22].]. Although Bayesian nonparametric methods can be computationally demanding, recent advancements in computer power have sparked renewed interest in these approaches. In traditional parametric Bayesian statistics, we assume a specific functional form for the likelihood and prior. Nonparametric Bayesian methods, on the other hand, allow for infinite-dimensional parameter spaces, which can adapt to the complexity of the data. A regression tree is a decision tree based on the recursive binary partitioning of the predictor space. It approximates some unknown function and thus is used for regression. The sum of trees is a multivariate additive model [22]. The predictions generated by BART are the result of successive iterations of the back-fitting algorithm.

3.4.3 Bayesian Networks

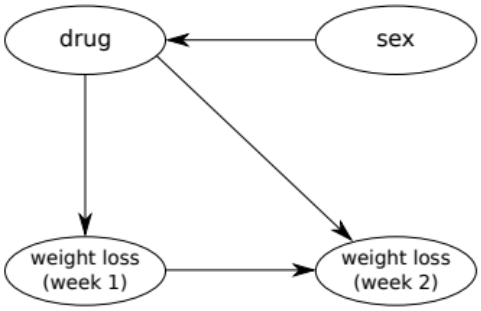
Bayesian networks(BN) are a probabilistic graphical model that represents probabilistic relationships between random variables [15] and does inference with those variables. There are three different types of BN as seen in figure 3.1.



(α′) Discrete Causal BNs: Each node has two possible states representing the responses'yes' and'no'.The direction of causality is from top to bottom. [23]



(β′) Gaussian BNs: the MARKS networks from Mardia, Kent & Bibby [24], which describes the relationships between the marks on 5 math-related topics.



(γ′) Hybrid BNs: the RATS' WEIGHT networks [25], which describes the weight loss in a drug trial performed on rats. Continuous nodes cannot be parents of discrete nodes.

Figure 3.1: Types of Bayesian Network

The type of network depends on the type of data; for continuous data, we have Gaussian Bayesian Networks, for discrete data, we have Multinomial Bayesian Networks and if the

data contain discrete and continuous variables, we have Hybrid Bayesian Networks [26]. They are very useful when studying the causal relationship between variables because of their graphical structure.

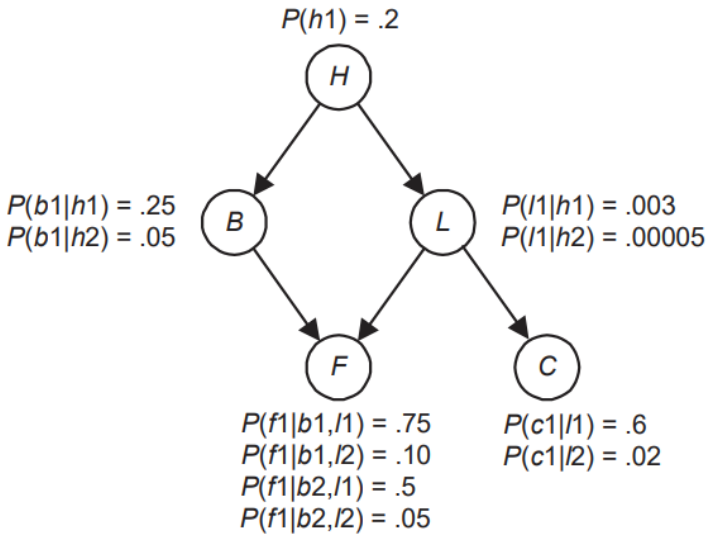


Figure 3.2: A Bayesian Network [4]

In figure 3.2 we see a Bayesian Network which represents the probabilistic relationships between the variables (H, B, L, F, C); this kind of network is also called Causal Network. The edges represent direct influence; for instance, H directly affects both Band L. Now, if we knew that H is the variable history of smoking, B is bronchitis and L is lung cancer. Thus, in everyday lingo, if we follow the path $(H \longrightarrow L \longrightarrow C)$ we can see that smoking history directly affects the presence of lung cancer and that lung cancer directly affects the result of a chest radiograph (variable C) [4]. Each node represents a feature in the data set. In Bayesian networks, the probabilities of each feature are the conditional probabilities of the values of each feature given every combination of values of their parent features; this does not apply to root nodes. In root nodes, the probabilities are prior probabilities. This network structure allows for the execution of probabilistic inferences among features. As an illustration, using this network, one can deduce the likelihood of an individual having bronchitis or lung cancer if they are known to smoke, exhibit fatigue (variable F), and possess an abnormal chest X-ray.

We employ those networks because, as we have already mentioned, DAGs are independent maps of probability [27], thus the networks explain the conditional independence relationships between variables, which helps us factorize the distribution.

The creation of the BN occurs through the creation of the causal DAG. We create a graph

where, if there is a direct edge from A to B, that means that A is a direct cause of B. That means that if we manipulate A, and A causes B then B is also affected. So if there is a manipulation in A, said relationship causes a change in the distribution of Y [4]. Therefore if we do not know if A has a causal relationship with B, we can practically examine it.

We manipulate A, by performing a randomised controlled experiment (RCE) which means that we check a population in some specific context and we manipulate A to see how B is altered. That is how causal relationships are retrieved.

3.4.4 Bayesian Structural Time-Series Models

Bayesian Structural Time-Series Models are a type of model that integrates time-series models with state-space models and Bayesian statistics [28]. They are used for time series analysis such as forecasting, decomposition and feature selection and also for causal inference. These models use two equations, the observed equation and the state equation. The observed one states that the data y_t are equal to the product of the output vector Z_t with a latent factor vector a_t plus a noise term ϵ_t.

$$y_t = Z_t^T \cdot \alpha_t + \epsilon_t$$

The state equation shows the evolution of the latent factor vector a_t through time, a_{t+1} which is the latent factor vector the time stamp $t+1$ is equal to the product of the transition matrix T_t multiplied with the latent vector in the previous time stamp a_t plus the product of the control matrix R_t with the system error term η_t

$$a_{t+1} = T_t \cdot {}_t + R_t \cdot \eta_t$$

The model's complexity is what makes it able to deeply understand the data. The model's modularity allows it to construct its components in a way that captures significant characteristics from the data. This demonstrates the patterns and seasonality of the time-series data.

3.5 Libraries to perform Bayesian Causal Inference

Although we analyzed the basis of Causality and Bayesian inference, we only saw the tip of the iceberg, thus writing code that implements the aforementioned is not only very difficult but also counterproductive. There exist already-made packages for various programming

languages that performed causality analysis, some of which we used in the implementation of this thesis. In this section, we will briefly describe the most popular packages for inference while mentioning both their advantages and disadvantages.

3.5.1 Bnlearn

It is a package for Bayesian Networks suitable for discrete, continuous and hybrid data, and it is also available to be used and incorporated in both R and Python [29]. It has many pipelines that perform causal learning, such as structure learning, parameter learning, and inference. Furthermore, it provides Bayesian network classifiers, conditional independence tests, network scores, and advanced network plotting capabilities. Therefore, a typical example to reveal causal inference from data with bnlearn involves several steps: first, constructing the graphical structure from the data (learning the DAG); next, using independence tests to prune spurious edges; then, estimating conditional probabilities based on the DAG; and finally, making inferences [17].

3.5.2 Causalpy

Causalpy is a Python package for Bayesian causal inference for quasi-experiments [30]. Quasi-experimental methods have been used to make causal claims when randomised experiments between treatment units do not exist [31]. The Causalpy package provides four Quasi-experimental methods, Synthetic control, Interrupted time series, Difference in differences, and Regression discontinuity. A typical example of using CausalPy is as follows: if we have a dataset with multiple units, but only one of them has received treatment, we would create a synthetic control as a weighted combination of the untreated units and then proceed with making inferences.

3.5.3 CausalNex

CausalNex is a Python library that uses causal Bayesian networks to make inferences and discover structural relationships for discrete data [32]. The library allows you to build causal models from data and plot them, and also to prune spurious correlations, which are seen as weaker edges. The visualisation provides an easier and beginner-friendly way to understand causality. Also, it can analyse interventions by using do-calculus. The do-calculus, introduced

by Judea Pearl, is a set of rules for reasoning about interventions in causal models. It is a formal system that allows us to make causal inferences from observational data.

3.5.4 DoWhy

DoWhy is a Python library that focuses on testing causal assumptions to determine their validity and the extent of their validity [33]. To perform those assumption tests, the library requires distinguishing the treatment and the outcome variable from the data. Additionally, it is helpful for the library to provide the DAG of the estimated causal relationships. We can model the causal graph, identify causal effects, refine outcomes based on robustness tests, check for confounders, and find the root causes of outliers or distributional changes [34]. There are certain limitations. First, the treatment variable must be binary. Second all categorical variables should be encoded to numerical values.

3.5.5 CausalImpact

CausalImapact is a library available for Python and R, it estimates the causal effects of an intervention using Bayesian structural models for time-series data [35]. When randomised experiments do not exist we can test the effect of an intervention in time series data by comparing the difference between the expected and observed values, this treatment analysis is performed by the CausalImpact. The expected values that are tested to see how the counterfactual are estimated through linear regression. Thus this model requires the outcome value to be modelled by linear regression where there is a pre-intervention period that it is not affected by the intervention. Finally, the other model assumption is for the post-intervention period to be specified.

3.5.6 Pgmpy

Pgmpy a Python library is a more advanced version of the bnlearn library demanding a deeper understanding of Bayesian inference. It distinguishes itself from the aforementioned libraries by allowing users to construct custom pipelines from causal blocks, making it considerably more extensible [36]. The main application of the library is to create probabilistic graphical models [17]. In Pgmpy, the user is responsible for building the model pipeline which involves data transformation to discrete format, collecting the results and plotting the

model.

Chapter 4

Related Work I: Statistical and Intelligent Bankruptcy Techniques

This chapter explores in-depth traditional bankruptcy prediction. Our primary reference was the comprehensive review paper, *Bankruptcy prediction in banks and firms via statistical and intelligent techniques–A review, by PR Kumar and V Ravi*, published in the European Journal of Operational Research in 2007 by Elsevier [6].

This piece of literature systematically breaks down traditional methodologies into eight key categories: (i) statistical techniques, (ii) neural networks, (iii) case-based reasoning, (iv) decision trees, (v) operational research, (vi) evolutionary approaches, (vii) rough set-based techniques—which includes subsets like fuzzy logic, support vector machine, and isotonic separation, and (viii) soft computing, a holistic amalgamation of the previously discussed techniques. Our approach to analysis employed this pivotal paper for insights into techniques and methodologies that were solidified up until 2005. This period represents the era where the foundational principles of statistical bankruptcy were carved out.

A distinctive decision in our research approach was the bifurcation of studies into pre-2005 and post-2005 segments. This strategic division was primarily influenced by the marked increase in the volume of studies post-2005, with a significant surge, particularly after the 2008 global financial crisis (see Figure 4.1. Consequently, for post-2005 insights, we are channelling our focus through the lens of *"An overview of bankruptcy prediction models for corporate firms: a systematic literature review"* by *Shi, Yin Li, and Xiaoni*. The post-2008 surge in research, as outlined in the aforementioned paper, underscores that the aftermath of the 2008 financial crisis marked a considerable uptick in papers on this topic, emphasizing

its criticality for businesses. Furthermore, the field appears to be characterised by limited co-authorship, with influential researchers often working in isolation over the past decades.

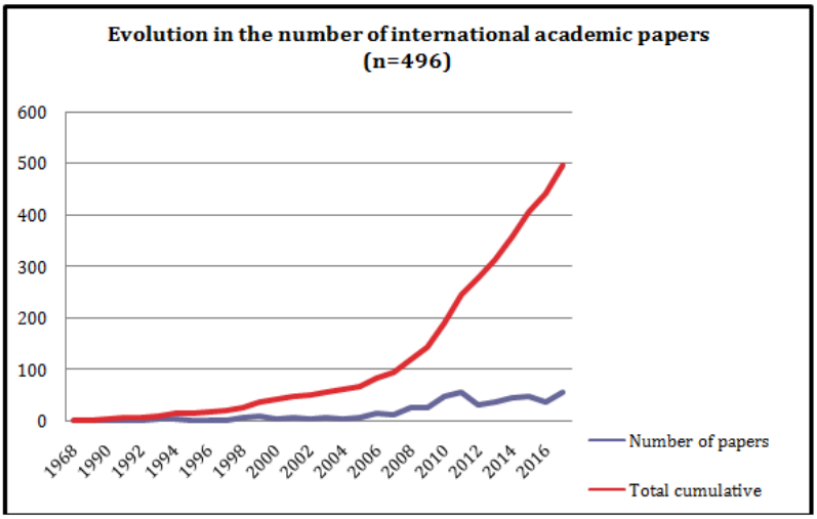


Figure 4.1: The volume of international academic publications from 1968 to 2017 [5]

Recognizing that our audience comprises individuals familiar with the more widely acknowledged techniques, we have chosen not to dwell at length on these established methodologies. Our aim here is to provide a literature overview, not an exhaustive deep-dive into the specialised intelligent techniques that have emerged. However, to achieve a balance between comprehensiveness and clarity, we have incorporated a figure (Figure **??** that succinctly captures the merits and demerits of each technique. This graphical representation simplifies the comprehension of intricate bankruptcy prediction models, enabling readers to absorb the fundamental attributes of each approach without getting entangled in intricate specifics.

4.1 Foundational Studies on Bankruptcy Prediction Up to 2005

In curating the studies for this section, we anchored our selection on two fundamental principles: a strong emphasis on the pioneers who first developed the predictive models and a preference for researchers who utilized these models to their utmost potential. This methodological severity ensures our thesis is rooted in both the origin and the optimal application of bankruptcy prediction techniques, offering a refined and impactful overview of the subject.

	Technology	Basic idea	Advantages	Disadvantages
1	FL	Models imprecision and ambiguity in the data using fuzzy sets and incorporates the human experiential knowledge into the model	Good at deriving human comprehensible fuzzy 'if–then' rules; It has low computational requirements	Arbitrary choice of Membership function skews the results, although triangular shape is the most often used one. Secondly, the plethora of choices for membership function shapes, connectives for fuzzy sets and defuzzification operators are the disadvantages
2	NN	Learn from examples using several constructs and algorithms just like a human being learns new things	Good at function approximation, forecasting, classification, clustering and optimization tasks depending on the neural network architecture	The determination of various parameters associated with training algorithms is not straightforward. Many neural network architectures need a lot of training data and training cycles (iterations)
3	GA	Mimics Darwinian principles of evolution to solve highly nonlinear, non-convex global optimization problems	Good at finding global optimum of a highly nonlinear, non-convex function without getting trapped in local minima	Does take long time to converge; May not yield global optimal solution always unless it is augmented by a suitable direct search method
4	CBR	Learns from examples using the euclidean distance and k-nearest neighbor method	Good for small data sets and when the data appears as cases; similar to the human like decision-making	Cannot be applied to large data sets; poor in generalization
5	Rough sets	They use lower and upper approximation of a concept to model uncertainty in the data	They yield 'if–then' rules involving ordinal values to perform classification tasks	It can be (a) sometimes impractical to apply as it may lead to an empty set (b) sensitive to changes in data and (c) inaccurate
6	SVM	It uses statistical learning theory to perform classification and regression tasks	It yields global optimal solution as the problem gets converted to a quadratic programming problem; It can work well with few samples	Selection of kernel and its parameters is a tricky issue. It is abysmally slow in test phase. It has high algorithmic complexity and requires extensive memory
7	Decision trees	They use recursive partitioning technique and measures like entropy to induce decision trees on a data set	Many of them can solve only classification problems while CART solves both classification and regression problems. They yield human comprehensible binary 'if–then' rules	Over fitting can be a problem. Like neural networks, they too require a lot of data samples in order to get reliable predictions
8	DEA	It uses linear programming to rank various alternatives/business units according to some input and out variables	It has found numerous applications and gives exact solution	It yields only relative scoring of the business units and not absolute ratings
9	SC	Hybridizes fuzzy logic, neural networks, genetic algorithms, etc. in several forms to derive the advantages of all of them	It amplifies the advantages of the intelligent techniques while simultaneously nullifying their disadvantages	Apparently, it has no disadvantages. However, it does require good amount of data, which is not exactly a disadvantage nowadays

Figure 4.2: Advantages and Disadvantages of Intelligent Tech [6]

4.1.1 Statistical Techniques

Statistical techniques have been at the forefront of financial distress prediction for decades. These techniques provide a systematic and quantitative approach to data analysis, allowing researchers and practitioners to make informed decisions based on empirical evidence. In the context of predicting financial distress, bankruptcy, or firm failure, several statistical methods have been employed, as highlighted in the related works

In the realm of financial distress and bankruptcy prediction, numerous studies have been conducted to develop and evaluate various models. Altman et al. [37] introduced the Zeta analysis for bankruptcy classification, utilizing data from 111 firms, each characterized by seven variables. Their findings revealed a classification accuracy ranging from 96% when predicting one period before bankruptcy to 70% for five periods prior. Notably, the ZETA model surpassed the performance of other alternative methods.

Ohlson [38] used a logit model to predict firm failure. The data for this study were sourced from Moody's Manual, Compustat data tapes, and 10-K financial statements. The classification accuracy of Ohlson's model was impressive, standing at 96.12% for one year, 95.55% for two years, and 92.84% when considering either one or two years.

Dietrich and Kaplan [39] proposed a three-variable linear model for loan risk classification. The researchers compared their model to the Altman model and the Wilcox bankruptcy prediction model[40], and they found that their model performed better than both.

Zmijewski [41] undertook a study to examine potential biases in financial distress models. Using data from the American and New York Stock Exchanges, Zmijewski addressed the issues of ”oversampling” and ”complete data” sample selection biases.

Kolari et al. [42] developed an Early Warning System (EWS) grounded in logit analysis and Trait recognition. Their findings indicated that Trait recognition surpassed the logit model in terms of error rates.

Finally, Canbas et al. [43] presented an Integrated Early Warning System (IEWS) that combined various methods. Using data from Turkish banks, they found that the IEWS exhibited a better predictive capability than other models in the study.

4.1.2 Neural Networks

The application of statistical techniques, particularly neural networks, in predicting financial distress and bankruptcy has been extensively researched.

Back-Propagation Neural Network

The Back Propagation Neural Network (BPNN) has emerged as a dominant model in this domain, often outperforming traditional statistical methods. Here’s a synthesis of the related works:

Tam [44] utilized the BPNN for bankruptcy prediction on Texas banks and found it superior to other methods such as Discriminant Analysis, factor-logistic, and K-Nearest Neighbors (K-NN). In a follow-up, Tam and Kiang [45] confirmed that BPNN consistently outperformed other techniques.

Salchenberger et al. [46] applied BPNN to predict the failure of savings and loan associations (S&Ls) and noted its superior performance over logistic regression. Similarly, Sharda and Wilson [47] demonstrated BPNN’s superiority over Multiple Discriminant Analysis (MDA) using Altman’s five variables.

Altman et al. [48] compared Linear Discriminant Analysis (LDA) and BPNN, noting a marginal advantage for LDA in classifying distress with financial ratios. However, Wilson

and Sharda [49] and Tsukuda and Baba [50] both documented the superior performance of BPNN over DA.

Piramuthu et al. [51] introduced Feature Construction (FC) to BPNN for bankruptcy prediction and found the enhanced BPNN to outperform the plain BPNN. In another notable study, Zhang et al. [52] introduced a new three-layered NN model trained with Generalized Reducing Gradient (GRG2), which surpassed logistic regression.

Atiya [53] underlined the importance of including novel indicators with financial ratios to boost the performance of NN in bankruptcy prediction. Swicegood and Clark [54] established BPNN as the leading model in identifying underperforming banks when compared with DA and human judgment.

Finally, Lee et al. [55] reinforced BPNN's dominance in bankruptcy prediction by showing its superiority over other techniques like Self-OrganiSing Maps (SOM), DA, and logistic regressions in a comprehensive study using data from Korean firms.

Self-Organizing Maps (SOM)

Incorporating Self-Organizing Maps (SOM) into financial distress prediction models has become a growing trend, frequently taking the form of hybrid approaches in conjunction with other statistical techniques. The following section offers a comprehensive overview of the relevant studies in this area.

Lee et al. [56] investigated three hybrid BPNN models: (i) MDA-assisted BPNN, (ii) ID3-assisted BPNN, and (iii) SOM-assisted BPNN. Using data from the Korea Stock Exchange, they concluded that these hybrid neural network models surpassed traditional methods like MDA and ID3.

Serrano-Cinca [57] compared the performance of SOM with LDA and BPNN in financial diagnosis, introducing two hybrid neural systems: (i) LDA integrated with SOM and (ii) BPNN combined with SOM. They established that their hybrid system surpassed the Z-score analysis and additionally provided insightful visual graphics elucidating bankruptcy risk.

Kiviluoto [58] used SOM variants for firm bankruptcy prediction, comparing these classifiers with methods like LDA and LVQ. He found the RBF-SOM to be slightly more effective than other classifiers in his study.

Finally, Kaski et al. [59] integrated the Fisher information matrix-based metric into SOM. Their results showed that the SOM-F, which utilizes the Fisher metric, offered a superior

representation of bankruptcy probability compared to the SOM-E.

Other types of Neural Nets

Beyond the commonly used Back Propagation Neural Network (BPNN) and Self-Organizing Maps (SOM), several other neural network topologies have been employed in the realm of bankruptcy prediction. These topologies, each with its unique architecture and learning mechanism, have shown varying degrees of success in classifying financial health and predicting bankruptcy. H. Next, we present a review of the relevant studies.

Lacher et al. [60] introduced the Cascade-correlation neural network (Cascor) to assess a firm's financial health. Using data based on Altman's five financial ratios from the Standard and Poor's COMPUSTAT database, they found that the Cascor model consistently surpassed the Altman Z-score model in classification rates.

Yang et al. [61] adopted the Propagation neural network (PNN) and the Fisher discriminant analysis (FDA) for bankruptcy prediction. When comparing PNN*, PNN, and FDA against Discriminant Analysis (DA) and BPNN, they discovered that the PNN* and BPNN yielded better classification rates with non-deflated data, while the FDA excelled with deflated data.

Last but not least, Baek and Cho [62] explored the Auto-associative neural network (AANN) for bankruptcy prediction in Korean firms. The AANN outperformed the 2-class BPNN, achieving classification rates of 80.45% for solvent firms and 50.6% for defaulted firms.

4.1.3 Case-Based Reasoning

Case-Based Reasoning (CBR) and its hybrid models have been explored as alternative methodologies for bankruptcy prediction. These models leverage past cases to make decisions about new, similar cases. Here's a synthesis of the related works:

Bryant [63] introduced the Case-based reasoning (CBR) system for bankruptcy prediction and pitted it against Ohlson's [38] logit model. Using a variety of financial variables, he found that the logit model surpassed the CBR system in accuracy.

Jo et al. [64] employed three models, including Multiple Discriminant Analysis (MDA), CBR, and Back Propagation Neural Network (BPNN) for bankruptcy prediction in Korean

firms. The results were telling: BPNN was superior, achieving an 83.79% hit ratio, outperforming both DA and CBR.

Park and Han [65] presented an innovative combination of K-Nearest Neighbors (K-NN) with the Analytic Hierarchy Process (AHP) for bankruptcy prediction. This hybrid approach, AHP-K-NN-CBR, attained the best accuracy rate of 83.0%, clearly surpassing the other models they tested.

Lastly, Yip [66] employed a unique combination of CBR with K-NN for Australian firms' business failure prediction. The model's standout performance was evident with CBR combined with weighted K-NN, which achieved a 90.9% accuracy rate, making it superior to the other models.

4.1.4 Decision Trees

Decision trees, particularly the Recursive Partitioning Algorithm (RPA), have been employed as a tool for bankruptcy prediction. These models segment the data into subsets based on certain criteria, allowing for a hierarchical decision-making process. Here's a synthesis of the related works:

Marais et al. [67] delved into the use of the Recursive Partitioning Algorithm (RPA) for bankruptcy prediction, comparing its performance with polytomous probit. Their study indicated the superiority of the polytomous probit model over recursive partitioning when all variables were evaluated.

Frydman et al. [68] expanded the exploration of RPA in the context of bankruptcy prediction. In their comparison with Discriminant Analysis (DA), one of their RPA variants (RPA1) consistently outperformed the DA models across different misclassification cost considerations. This showed RPA1's noteworthy prominence in their study.

4.1.5 Operational Research

Operational research, with its emphasis on mathematical modelling and analytical methods, has been applied to the field of bankruptcy prediction. These methods aim to optimize decision-making processes by leveraging mathematical techniques. Below, there is a compilation of the relevant research:

Banks and Prakash [69] implemented a linear programming heuristic applied to a quadratic transformation of data for predicting firm bankruptcy. Their method outshined the quadratic

discriminant function (QDF) and the approach by Johnson and Wichern, demonstrating its effectiveness.

Lam and Moy [70] presented a hybrid technique fusing various discriminant analysis (DA) methods. Their findings showed that the combined approach was more accurate than individual DA methods, pointing to the hybrid technique's importance in the field.

Cielen et al. [71] executed a comparative analysis involving three models for bankruptcy prediction. Among these, the data envelopment analysis (DEA) model emerged as the top performer, surpassing both the C5.0 and MSD models in classification accuracy.

4.1.6 Rough-Set Theory

Rough set theory offers a unique mathematical approach to handle vagueness and uncertainty in data. Its application in bankruptcy prediction has provided valuable insights into discerning patterns and relationships within data without the need for preliminary or additional information. Here's a synthesis of the related works:

Greco et al. [72] put forth a rough set technique for bankruptcy prediction, specifically focusing on attributes with ordered domains. Furthermore, their dominance-based rough set method [37] was found to capture the knowledge in the data better than the classical rough set analysis, marking it as a superior approach.

Dimitras et al. [73] utilized rough set theory with a valued closeness relation (VCR) for predicting business insolvencies. In a comparison with established methods like discriminant analysis (DA) and logistic regression, their rough set method emerged as more adept at pinpointing crucial attributes for assessing bankruptcy risk.

McKee [74] fashioned a bankruptcy prediction model based on rough set theory, which boasted an outstanding accuracy rate of 88%. This model surpassed the performance of its recursive partitioning counterpart, emphasizing the robustness of the rough set approach.

4.1.7 Fuzzy Logic and Advanced Techniques

The application of fuzzy logic and advanced machine learning techniques in bankruptcy prediction offers a nuanced approach to handling uncertainties and complexities in financial data. These methods provide a more flexible and adaptive framework compared to traditional statistical methods. Here's a synthesis of the related works:

Michael et al. [75] pioneered the use of a fuzzy rule generator method in bankruptcy prediction. Their fuzzy rule-based classifier demonstrated top-tier results, surpassing traditional methods such as LDA, QDA, logit, and probit analysis, signifying its excellence in the field.

Min and Lee [76] employed the Support Vector Machines (SVM) technique for bankruptcy prediction, marking a shift from conventional methods. Their SVM model exhibited superiority over established methods like MDA, logit, and BPNN, highlighting the effectiveness of SVM in this application.

Ryu and Yue [77] introduced the innovative isotonic separation technique for bankruptcy prediction. When matched against numerous methods, isotonic separation emerged as the most effective for short-term bankruptcy prediction. Additionally, the rough set method was identified as a top performer, shedding light on its continued relevance in the domain.

4.1.8 Soft-Computing Techniques

Soft computing techniques, which emphasize the use of approximate solutions to computationally hard tasks, have been applied to bankruptcy prediction. These techniques often combine traditional statistical methods with advanced machine learning algorithms to enhance predictive accuracy. Next, there is an overview of the relevant studies in this area.

Back et al. [78] pioneered a hybrid architecture for bankruptcy prediction that amalgamated a wide array of models including BPNN, DA, logistic regression, and several others. Through a blend of simple voting and compensation aggregation methods, the BPNN model stood out, outshining both classical and contemporary models, underscoring its potency.

Gorzalczany and Piasta [79] championed the integration of a neuro-fuzzy classifier (N-FC) with a rough classifier (RC) in a unique decision support system. In the realm of bankruptcy prediction, the N-FC demonstrated unmatched prowess, eclipsing methods such as C4.5 and CN2 in classification accuracy.

Ahn et al. [80] unveiled hybrid models, combining rough sets with BPNN, for forecasting Korean firms' failures. Demonstrably superior, these hybrid constructs, termed Hybrid I and II, surpassed the results of stand-alone BPNN and DA.

McKee and Lensberg [81] presented an innovative two-tiered hybrid method that married the rough set model with genetic programming (GP) in the context of bankruptcy prediction. This amalgamated approach exhibited enhanced performance over the standalone rough set model.

4.2 Approaches to Bankruptcy Prediction Post-2005

Since 2005, there has been a noticeable rise in studies concerning business failure. While the topic has been studied extensively, the literature reveals an inherent ambiguity and fragmentation in its definition, suggesting that scholars and practitioners haven't yet arrived at a universally accepted definition of business failure [5].

4.2.1 Post-2005 Perspectives on Business Failure and Criteria

Balcaen and Ooghe's 2006 [82] research adds depth to this discussion, emphasizing the fragmented nature of historical studies on business failure. Depending on the selected criteria, research either veers towards a legal definition, encompassing aspects like bankruptcy, or pivots to a more financial angle, concentrating on financial distress.

Altman and Hotchikiss, in 2006 [83], dive into specific terms frequently associated with unsuccessful business endeavours. Four primary descriptors emerge:

- **Failure**: Defined by an economic criterion, it's marked by a continual underperformance in returns on invested capital.
- **Insolvency**: This state occurs when a firm's liabilities overtake its assets, signalling a potential liquidity challenge.
- **Default**: It pertains to a firm's inability to meet certain obligations, especially in terms of repaying loans or complying with court decrees.
- **Bankruptcy**: The term is multi-level and includes two perspectives The first considers the net worth position of an enterprise, while the second denotes a firm's formal declaration in a federal court. This declaration could be aimed at liquidating assets or embarking on a recovery strategy.

Furthermore, the idea of "early warning" in bankruptcy prediction has gained significant attention post-2005. Initially rooted in military contexts, its scope has expanded to areas such as macroeconomics and business administration. This shift underlines the increasing emphasis on proactive measures to detect bankruptcy risks.

In essence, the period following 2005 has enriched the discourse on business failure with multifaceted perspectives. While there's a clarion call for a unified framework, the diverse viewpoints enrich the academic and practical understanding of the subject [5].

4.2.2 Bankruptcy Prediction with AI and Advanced Statistical Techniques

The literature on bankruptcy prediction has witnessed a significant evolution since Altman's seminal 1968 model. As the technological landscape has transformed, especially post-2005, the methods employed for bankruptcy prediction have become more varied and sophisticated.

Notably, advancements in machine learning and artificial intelligence have paved the way for a plethora of innovative predictive models. Techniques like the aforementioned rough set theory, which was designed to handle apparent indiscernibilities within data sets, have reported accuracies ranging between 76% and 88%. Studies by authors such as Xiao et al. in 2012 [84] and Wang & Wu in 2017 [85] testify to the increasing influence of this methodology.

When it comes to case-based reasoning works by Li & Sun in both 2009 [86] and 2011 [87] reinforced the growing relevance of this method.

Furthermore, SVM has consistently demonstrated superior performance compared to artificial neural networks, as it was later corroborated by Kim, in 2011 [88].

Insights from Previous Greek SMEs Dataset Iterations

In the context of Greek Small-Medium Enterprises (SMEs), Papadouli [89] made significant strides by utilizing a logit model to forecast firm failure, as evidenced in her research. The data for this exploration were meticulously sourced from Moody's Manual, Compustat data tapes, and 10-K financial statements, ensuring a comprehensive and robust dataset for analysis. Papadouli's model accomplishments underscores the model's effectiveness in predicting firm failure with high precision. This contribution by Papadouli builds on the extensive history of bankruptcy prediction, which has evolved from the early use of univariate statistical models of financial ratios to the modern application of sophisticated supervised machine learning models. Even in the face of challenges such as obtaining substantial labeled datasets, especially in the Greek market, the incorporation of semi-supervised and transfer learning techniques has proven invaluable. Particularly, transfer learning techniques have demonstrated superior performance, yielding high accuracy scores in various bankruptcy prediction periods, thereby enhancing the reliability and comprehensiveness of bankruptcy risk assessment in diverse market contexts.

In a subsequent paper, Papadouli [90] assembled a dataset of 170 bankrupt and 1424 non-bankrupt Greek SMEs to explore and compare various bankruptcy forecasting algorithms. Traditional models like Altman's Z-score, Springate, and Taffler's often misclassified healthy firms as bankrupt, with Taffler's model showing marginally better performance. The research revealed that semi-supervised classifiers slightly outperformed supervised ones on the imbalanced Greek SMEs dataset. Some effective alternatives for bankruptcy forecasting included supervised classifiers DT, RF, XGB, and AML with synthetic oversampling, and the sklearn implementation of the AutoML classifier. The study also successfully employed commercially available unlabeled datasets and two transfer learning algorithms for bankruptcy prediction, showing satisfactory performance and underscoring the potential of transfer learning for this application. Despite lower accuracy in other studies, this research achieved superior results, offering its dataset and software freely upon request.

In conclusion, the post-2005 period has been characterized by a transformative shift in the methods employed for bankruptcy prediction. As the impact of Artificial Intelligence and machine learning continues to grow, the field is positioned to harness these technologies to develop more precise and enlightening models in the forthcoming years [5].

4.2.3 Evolution in Bankruptcy Prediction Approaches

The burgeoning literature in the realm of bankruptcy prediction has demonstrated a heightened interest and exploration in the area. Yet, an underlying challenge remains the ambiguity and fragmentation in defining business failure. While the myriad definitions shed light on the multifaceted nature of business failure, they also underscore the need for a more consolidated and comprehensive understanding.

The increased research post-2005 indicates an acknowledgement of this challenge and showcases efforts towards employing novel and technologically advanced methodologies. These are not merely attempts to replace traditional techniques but to augment them, integrating modern computational capabilities with foundational financial and economic principles.

As we transition further into this evolving landscape, there's a promising emergence of causal approaches. These methodologies aim to enhance our grasp on bankruptcy prediction by delving deeper into root causes and mechanisms leading to business failure. The upcoming chapter will explore these causal approaches, providing insights into their potential to reshape and enrich the field of bankruptcy prediction.

Chapter 5

Related Work II: Causal and Bayesian Methods in Bankruptcy Prediction

In this chapter, we focus on two primary objectives. First, we detail the causal techniques used for bankruptcy prediction, showcasing their relevance and application. Second, we undertake a systematic literature review, providing a thorough examination of existing research in this domain. Through this dual approach, we aim to offer a well-rounded perspective on bankruptcy prediction methodologies, grounded in both current practices and historical context.

5.1 Historical Overview

As we have seen, Bankruptcy prediction (BP) has been a field of interest for many years, as it has been studied since the 1930s [91]. At first, the problem was tackled with many statistical methods. Still, the rise of machine learning and improved new automated inference techniques have caused a spike in the research for sophisticated Machine Learning (ML) bankruptcy prediction after the 2000s. Nowadays, the vast amount of bankruptcy prediction technologies has created an eco-chamber of research where not only the prediction is important, but also the assessment of danger and making changes with strategic management to shift the outcome. The most recent shift in BP research was introduced due to the development of causal ML because it is important to know what causes a firm's financial distress. We can see this trend in 5.1 where we see the rise of papers with the term causal AND bankruptcy prediction in the title, keywords or abstract that are submitted to Scopus. It is easily seen that

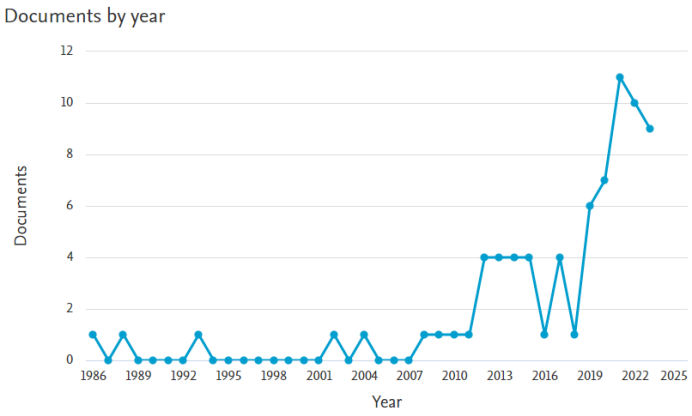


Figure 5.1: Documents with ”causal” and ”bankruptcy’ prediction in the title, keywords or abstract that are submitted to Scopus[7]

upward trend in the causal research.

5.2 From Traditional ML to Causal Bayesian ML for Bankruptcy

The evolution of bankruptcy prediction models has been a journey from simplicity to complexity, adapting to the ever-changing dynamics of the business world. The classic Machine learning methods for bankruptcy see the problem as an imbalanced, higher-dimension classification problem trying to optimise the accuracy of the model based on the data. There are two main categories of studies[92]:

- **parametric methods:** multiple discriminant analysis (MDA), linear discriminant analysis (LDA), canonical discriminant analysis (CDA), logistic regression (LR) and Naïve Bayes (NB)
- **non-parametric methods:** artificial neural networks (ANN), support vector machine (SVM), decision trees (DT), k-nearest neighbor (KNN), hazard models, fuzzy models, genetic algorithms (GA) and hybrid models, where multiple models are combined

All those models are very successful in tackling the classification problem but they lack interpretability which limits their contribution to the general economic consequences[93]. Thus the lack of explainability makes it difficult to trust the ML black box methods in a sensitive area like finance which can have tremendous socioeconomical results. As a result, there is a need for more interpretable ML with transparency on how the outcomes are estimated, to

be more easily trusted by people who are not familiar with machine learning. Instead of creating explainable models, some scientists tried to make post-hoc methods that can explain the results of black box models. A widely used post-hoc model is the Shapley values, it is used to show which variables have influenced the predictive results of the ML models. Another problem of the simple models is that while effective in estimating the posterior probability of a firm's failure based on its financial characteristics, did not consider the expected time to failure. This limitation could lead to decisions that might be too late to prevent close failures. As a response to this gap, survival analysis emerged as a potential solution. Although survival analysis has been popular in fields such as medical and technical sciences, its application in predicting financial failures remains limited[94]. However, the paper[94] by Yuri Zelenkov aims to bridge this gap by evaluating the applicability of survival analysis to bankruptcy prediction. Survival analysis offers a dynamic perspective, considering the time factor in extracting valuable information about risk dynamics and estimating the impacts of features. This transition from traditional ML models to survival analysis represents a paradigm shift, with the aim of a more holistic and time-sensitive approach to bankruptcy prediction.

Other studies tried to make a more holistic approach by incorporating external factors into the data. This study[95] bridges the gap between macroeconomic indicators, particularly the EPU, and bankruptcy prediction. Highlights the value of incorporating the EPU indicator, especially its Twitter-extracted version, into bankruptcy prediction models, thereby enhancing their accuracy. This also aligns with the emerging literature that emphasises the role of social media in predicting firm-level bankruptcy or financial distress. By integrating novel data sources like Twitter and leveraging advanced machine learning techniques, it offers a more holistic and accurate approach to predicting bankruptcy, which is paramount in the ever-evolving economic landscape.

5.3 Advantages of Causality in Bankruptcy Prediction and Financial Data Analysis

The emerging field of "causal machine learning" in Bankruptcy Prediction aims to leverage the strengths of ML to provide more precise, less biased, and reliable estimators of causal effects. In traditional econometrics, the focus has been on explanation and causality, often at the expense of predictive power. The integration of causal inference with machine learning

in bankruptcy prediction and financial data analysis offers a more comprehensive, accurate, and adaptable approach to understanding and predicting financial crises and their impacts. In this section, we will explore the benefits of the causal approach.

The first benefit is increased accuracy. Integration of machine learning with causal inference can lead to more precise, less biased, and more reliable estimators of causal effects. This means that predictions are based not only on correlations but also on understanding the underlying causal relationships of the financial data.

Furthermore, the adoption of causal bankruptcy prediction provides a more comprehensive and intricate understanding of the subject matter. Techniques, such as Causal forests, allow for a better understanding of the factors that contribute to different outcomes in different contexts[96]. This provides information on how the impact of a phenomenon varies between individuals, including potential thresholds and interactions.

Another benefit of causal Bankruptcy Prediction is that it addresses Complex Challenges. Traditional methods often struggle with issues such as overfitting or failing to control for key confounders. Causal techniques, provide a more robust framework for such complex challenges, balancing the objectives of identifying significant differences in treatment effects while also estimating causal effects accurately.

Causal models can distinguish Risk from Vulnerability, which is a very important feature [96]. This fine understanding can lead firms to make better policy recommendations and interventions.

Another benefit of causal models is that they address heterogeneity. Traditional models often treated companies as homogeneous entities. However, surveys showed that there's unobserved heterogeneity when analyzing company bankruptcy processes, which challenges the assumptions of traditional models[97]. New approaches recognize and address the heterogeneity among companies, ensuring that sector-related characteristics, capital structure, and size are considered.

They also offer flexibility in Non-Parametric Approaches. Techniques like the causal forest (CF) offer a flexible model that can handle high levels of interactions and dimensions. This allows for a more nuanced understanding of the factors influencing bankruptcy[97].

Furthermore, we can get insights into Sectoral Effects with causality. Certain sectors might be more prone to financial crises, which can exacerbate the bankruptcy process. Understanding these sectoral vulnerabilities with causal inference can guide targeted interventions

and policies.

Finally, with causal methods, we can incorporate External Factors by acknowledging the importance of both internal (company-level) and external (macroeconomic) factors in predicting bankruptcy. This comprehensive view can lead to a more holistic understanding of the factors driving bankruptcy.

5.4 Bayesian Networks Learning for Bankruptcy Prediction

A machine learning technique that has been used for causality is Bayesian Network, they are used in learning the causal relationships from the data. The use of naive Bayesian Bayesian network (BN) models in bankruptcy prediction was first proposed in this study [98]. Historically, bankruptcy prediction methodologies have transitioned from fundamental analyses in the 1960s to sophisticated techniques such as BN models, which stand out for their interpretability, adaptability, and lack of complete information dependency. Another study[99] indicates that Bayesian models can effectively predict financial distress, with information deriving from companies' financial statements being valuable. Thus, Bayesian Networks, with their probabilistic graphical models, have gained popularity for their ability to represent complex probabilistic relationships. They offer several advantages, including explicit probability outputs and a graphical model framework. However, they also have limitations, such as reliance on prior beliefs. The naïve Bayesian network, a simple structure with a common parent node, has been widely used for classification due to its simplicity and strong independence assumption. The Bayesian network model is also employed to address model uncertainty problems in analyzing firm bankruptcy and predictability. Especially when the cost ratio of Type I errors to Type II errors is high, the predictive power of the Bayesian model is stressed, suggesting its potential attractiveness in the current economic environment where significant firms face financial distress [100].

Some studies have integrated the Bayesian network model with other ML models to predict firm bankruptcy. This study's [101] model integrates the Least Absolute Shrinkage Selection Operator (LASSO) to select relevant financial ratios and subsequently establish the Bayesian Network (BN) topology and estimate its parameters. A significant advantage of the LASSO-BN model is its transparency, which provides a clear interpretation of its internal

workings by elucidating how conditional default probabilities are derived from the selected variables. This clarity in interpretation addresses the growing demand for interpretable machine learning models, especially in contexts where understanding the decision-making process is crucial. The Bayesian network model, which is powerful in its predictive capabilities, also provides a clear depiction of its internal functionality, allowing for a deeper understanding of the relationships and dependencies between variables. This dual focus on performance and transparency makes the ensemble BN models particularly relevant for investors, portfolio managers, and regulators, offering a comprehensive tool for assessing firm financial health.

In recent research on bankruptcy forecasting, approaches using a Bayesian framework for financial risk management were proposed [102]. Traditional models, such as Linear Discriminant Analysis and Artificial Neural Networks, could not incorporate prior expert knowledge, a gap addressed by the Bayesian model. Despite criticisms of subjectivity in the Bayesian method, it uniquely allows for the explicit inclusion and evolution of prior judgment. Conclusively, the studies paved the way for further exploration of Bayesian techniques in credit risk management.

5.5 Literature Review: Causal Inference in Bankruptcy Research

In the pursuit of understanding the evolution of causal inference methods within the Banking, Finance, and Insurance sectors, a comprehensive study titled "*Causal Inference for Banking, Finance, and Insurance – A Survey"* by *Satyam Kumar, Yelleti Vivek, Vadlamani Ravi, and Indranil Bose* kumar2023causal, reviewed the distribution of papers from 1992 to 2023. This seminal work will serve as our primary source of literature records in this section. The survey, underpinned by two primary themes, navigates through domain-specific applications in BFSI—ranging from corporate finance and financial governance to specific utilities like churn modelling and credit scoring—and also demystifies the causal inference methodologies employed, including Bayesian Network, Granger Causality, and counterfactuals.

Lastly, to provide a coherent structure, papers with overlapping themes were classified based on their predominant focus, determined by word count. Interestingly, an uptrend in BFSI publications over the past five years has been observed, as illustrated in Figure 5.2.

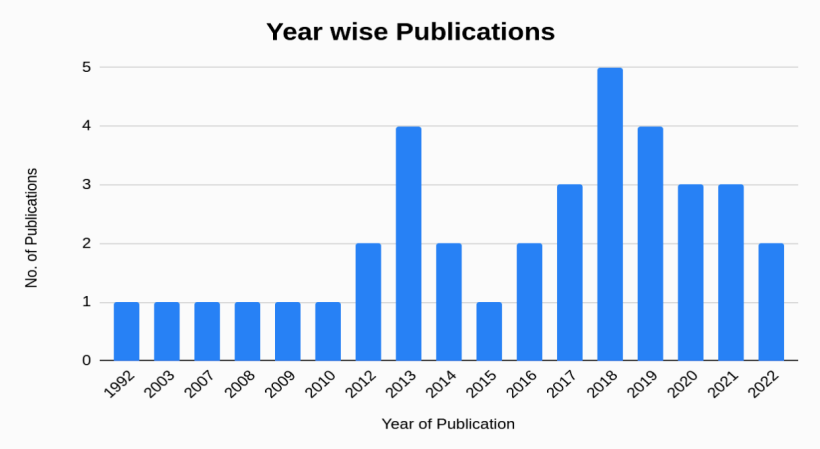


Figure 5.2: The volume of international academic publications from 1992 to 2023 [8]

5.5.1 Survey segmented by domains

Field of Banking

The banking sector, a linchpin of global financial systems, has always been a subject of keen academic scrutiny. It's intriguing to observe the multifaceted studies aimed at decoding its intricacies.

Michail [103] examined the impact of negative interest rates on inflation and bank lending in Denmark, Sweden, and Switzerland. The findings indicated that negative interest rates did not deter banks from lending. Notably, it was observed that bank funding costs and return-on-Equity served as influential limiting factors in this context.

Kolodiziev et al. [104] introduced an innovative method using causal analysis to assess the stability of the banking system in Ukraine. This methodology employed four distinct groups of indicators. The study's findings illuminated that causal analysis effectively revealed the most critical components and relationships between the indicators. This insight further emphasized the deep interconnectedness inherent in the stability of the financial system.

Stolbov & Shchepeleva [105] embarked on a detailed investigation into the causal relationships interlinking systemic risk, economic policy uncertainty, firm bankruptcies, and global volatility as represented by the VIX index. Their research unearthed that the connection between bankruptcies and factors like systemic risk, policy uncertainty, or the VIX index is intricately contingent on the scale of banks' deleveraging actions concerning the private non-financial sector.

Field of Finance

The ever-evolving realm of finance continually challenges researchers to disentangle its complex dynamics and forecast the unpredictable pathways of market trends.

Gong et al. [106] delved into a meticulous analysis of network topology, leveraging centrality measures to decipher causal relationships prevailing among financial institutions. Their study spotlighted the dynamic shifts in systemic risk within China's bustling financial markets. A pivotal revelation from their research was that their formulated framework possessed the capability to serve as a robust early warning tool during surges in systemic risk.

Davis [107] set out to investigate a selection of twelve models from literature that showcased inelastic demand. These models stood in stark contrast to their classical counterparts. Davis's exploration further delved into two specific predictor types pertinent to price alterations. The research's conclusions pointed towards the inherent difficulty in trading against price fluctuations in tangible market environments. Furthermore, a staggering count of 62 anomalies were earmarked exclusively for longstanding investments when utilizing rank correlation.

Ravivanpong et al. [108], navigating the intricate terrain of financial governance employed a multifaceted approach encompassing causal graphs, intricate visualization techniques, and anomaly detection. The primary objective was to unearth the root causes behind risk profile shifts in various investment portfolios. Their methodological arsenal included advanced techniques like Agglomerative Hierarchical Clustering (AHC) and Effective Transfer Entropy (ETE). The study's outcomes unveiled that conventional methodologies can occasionally bypass notable correlations existing between different portfolios, especially during the onset of minor crises. Moreover, when juxtaposing the efficacy of AHC analysis combined with VaR against causal graphs, the latter emerged superior both in terms of pragmatic application and technical requisites.

Rigana et al., [109] introduced an innovative measure for contagion between various currencies within the Forex market, rooted in causal inference and drawing heavily from causal graph model theory. Their findings provided a deeper understanding of the functioning of contagion paths within this complex market.

Tsapeli et al., [110] delved into the intriguing realm of social media's influence on financial markets. Their research specifically focused on gauging the causal impact of sentiment, as expressed on platforms like Twitter, on stock returns for industry giants including Apple,

Microsoft, Amazon, and Yahoo. The findings confirmed the significant sway of Twitter sentiment on stock prices, although the methodology presented inherent biases, particularly given its reliance on observational data.

Castro [111] embarked on a mission to intertwine estimation methods traditionally used in finance with contemporary causal event approaches. The spotlight was on stock holding period returns, specifically around particular, momentous events. Castro's insights underlined the standard definition of the event window that circumscribes significant financial happenings.

Kleinberg [112] brought forth the Assessment of Rare Causes (ARC) methodology, specifically crafted to decipher the causation behind infrequent events. Kleinberg's conclusions emphasized the utility of ARC in bolstering decision-making processes by elucidating the ripple effects of these rare events.

Kleinberg et al. [113] advanced an algorithmic framework meticulously designed to infer causal relationships embedded within time series data. The researchers harnessed a diverse array of price data for their study, concluding that their method presented a formidable tool for testing intricate hypotheses in time series data landscapes.

Moraffah et al., [114] tackled two pivotal causal inference tasks for time series data: treatment effect estimation and causal discovery. Their findings astutely differentiated between time-invariant and time-varying treatment effects, offering a nuanced perspective for causal analyses in the context of time series.

Peters et al., [115] presented an in-depth examination of Time Series Models with Independent Noise (TiMINo), setting it in juxtaposition with more conventional methodologies. Their research showcased that TiMINo boasts compatibility with an array of data types, with a special emphasis on the identifiability derived from constrained models.

Chikahara & Fujino, [116] harnessed the power of supervised learning classifiers to drive time series causal inference. Their methodology leaned heavily on feature representation coupled with a keen understanding of historical values. The researchers showcased their approach through its application on both synthetic and real datasets, accentuating its proficiency with both bivariate and multivariate time series.

Geiger et al., [117] proposed a pair of estimation techniques tailored for non-Gaussian, independent noise and went on to outline conditions essential for identifying causal features. Validating their methodologies, the team illustrated their effectiveness using carefully simu-

lated data sets.

Rudd et al., [118] unveiled a churn prediction system specifically tailored for businesses that do not operate on a subscription model. The core of the system's churn prediction score is powered by a multi-layer perceptron (MLP). Delving deeper into the churn's roots, causal analysis techniques were employed, drawing from structural equation models (SEMs) and Counterfactual based models. One of the standout elements of this research was the introduction of a novel feature engineering procedure, meticulously crafted around the recency, frequency, and monetary dimensions of customer engagement. This innovative approach proved pivotal, substantially elevating the overall performance of the predictive model.

Fahner [119] pioneered a method poised to tackle the longstanding challenges of selection bias and the constraints of limited historical testing that are intrinsic to credit score decision-making paradigms. This method exhibits versatility, adeptly managing multiple ordinal or categorical treatment impacts. Delving into the mechanics of this approach, it initially extracts granular information from support regions. Subsequently, it assembles a comprehensive global model. A testament to its robustness, when the model was put to the test in the realms of risk-based pricing and credit line augmentation challenges, it showcased its proficiency in discerning intricate causative relationships.

Field of Accounting

In the intricate realm of accounting, there exists an ever-growing landscape that constantly interacts with the broader economic and business environment.

Guelman & Guillén, [120] explored the effects of rate adjustments on a policyholder's choice to end a policy (lapse). To account for the intertwined dynamics, they studied the covariates of a policyholder's lapse, highlighting correlations between price elasticity and other factors. For each rate change level, a set of lapse probability models was established using gradient-boosted machines (GBM) for both training and variable selection. The team employed propensity scores and matching algorithms to correlate policyholders under varying rate change conditions, drawing counterfactual outcomes from these pairs. The study's conclusion emphasised its utility for company managers: by selecting the most favourable rate change for each policyholder, the aim is to augment company profits.

5.5.2 Survey Segmented by Methods

Bayesian Causal Network

Jacquer & Polson, [121] surveyed Bayesian econometric methods in finance, focussing on Markov Chain Monte Carlo (MCMC) and particle filtering (PF) algorithms. They highlighted MCMC's aptitude for handling complex models in the context of stochastic volatility (SV) and used PF for discrete time comparisons. Applications include optimal portfolio creation, returns predictability, and asset and option pricing.

Sanford and Moosa, [122] developed a Bayesian network for operational risk modelling, considering data heterogeneity and scarcity. Based on structured finance operating (SFO) units of an Australian bank, its three-stage methodology encompasses structural development, probability estimation from domain experiments, and model validation. They incorporated expertise from uninvolved domain experts for clear causal relations.

Gao et al., [123] introduced a stress-testing framework, merging Suppes-Bayes Causal Networks (SBCNs) with classification algorithms. Unlike traditional Bayesian networks, SBCNs employ probability causation and utilise maximum likelihood estimation (MLE) to eradicate spurious causes. Stress tests validated SBCNs' efficiency in computation and data usage.

Granger Causality

Stavroglou et al., [124] delved into financial assets, utilizing methods such as linear and nonlinear intertemporal cross-correlation (LICC and NICC). The research underscored the value of causal inferences in averting financial crises akin to 2007-2009. Data like stock indices (USA, Japan, China, India), government bonds, and oil prices revealed that both LICC and NICC shared 50% common causal links. Notably, rising oil prices influenced the China stock market downturn.

Tiffin, [96] empirically analyzed the aftermath of the financial crisis on growth using the causal random forest algorithms. This method dissected risks, pinpointed potential thresholds, inspected non-linearities, and spotlighted the crucial role of exchange rates in dictating a nation's progress.

Eichler, [125] tackled challenges in spurious causality, focusing on the application of Granger and Sims causality in empirical domains. The study introduced a distinctive iden-

tification method using latent variables for causal time-series structures.. It underscored the importance of assessing the causal effects of rare events due to their significant impact, which is relevant across finance, bioinformatics, and computational sciences.

Counterfactuals

Lundberg & Frost, [126] delved into counterfactuals in the volatile marketing arena. Through empirical testing within trading contexts based on Norm Theory, findings suggest the utility of counterfactuals in dynamic decisions. The importance of post-decision evaluations of past marketing strategies was underscored.

Svetlova, [127] posited counterfactual analysis as crucial not only in human psychology but also in societal dimensions. Contextually, the study assessed counterfactuals in portfolio management, revealing their potential pitfalls in financial markets. Key influencers in portfolio management, spanning fundamental to macro-economic factors, were highlighted.

Brodersen et al. [28] presented a novel causal impact estimation using diffusion-regression state-space models. These models outshine traditional methods by offering insights into temporal impact progression, integrating empirical priors via a Bayesian approach, and flexibly accommodating varied variations.

Gan et al. [128] crafted a model-agnostic framework for generating relevant counterfactuals in model risk management. This was automated using cloud-native algorithms, and its efficacy was gauged using the Freddie Mac dataset.

Wang et al. [129] proposed an innovative sparsity algorithm treating counterfactual explanation as an optimization problem. Tailored for high-dimensional inputs in corporate credit ratings, the algorithm resolves the challenges of non-injective functions in high-dimensional inputs by minimizing feature modifications, thus facilitating counterfactual explanations for them.

Explainability Approach

Yang et al., [130] explored the intricacies of M&As in the backdrop of the burgeoning field of XAI. By refining a transformer variant enhanced with adversarial training for M&A predictions, and identifying key words post-prediction, the study furnished counterfactual explanations. The breakthrough was its superior prediction accuracy, overtaking other methods and human expertise, along with providing more believable counterfactual rationales.

Grath et al., [131] addressed the enigma of black-box classifiers in predicting credit applications. Introducing a weight vector to underscore crucial features for counterfactual explanations, the study proposed two methods for its creation: ANOVA F-values and a Nearest Neighbors technique.

Dastile et al. [132] introduced a bespoke genetic algorithm to generate succinct counterfactual explanations for predictions made by opaque models. Primarily leveraging public credit scoring datasets, the outcome was a tool that could demystify the logic behind approved loan applications. The innovation's potential is further highlighted by the possibility of refining its fitness function via genetic programming.

Bueff et al., [133] undertook the challenge of elucidating machine learning models used in credit scoring. Through the application of counterfactual explanations, the study shed light on the intricate relationship between input variables like income or debts and credit risk. The findings emphasized the capability of counterfactual scenarios to decode the nuances in credit scores, advancing model transparency.

Chapter 6

Dataset for Bankruptcy Analysis of Greek SMEs: Data Analysis

In this chapter, we meticulously examine the nuances of a custom Greek SMEs dataset, providing a detailed overview and unravelling its multifaceted patterns. As we traverse this analytical journey, we emphasize the importance of variable selection. This critical step determines the effectiveness of our subsequent models, by diving into the very makeup of our dataset, shedding light on its source and composition and detailing the financial ratios encompassed within it. Our exploration is significantly enhanced by a suite of analytical tools, as this chapter delineates.

As we proceed, we discuss methods and techniques employed for data manipulation, such as the criteria for initial variable elimination and strategies for addressing missing values. However, The heart of our exploration lies in the comprehensive approach to Exploratory Data Analysis (EDA). We embark on an overview of EDA, followed by a dedicated section on utilizing box plots as a potent statistical tool. Moreover, we conduct an in-depth analysis to comprehend the correlations among the variables, highlighting both overarching associations and those directly related to the target variable. Hypothesis testing is integral to our analysis to confirm the validity of the correlations observed. We also touch upon some of the EDA methods that were considered but remained unused, providing insights into our analytical decision-making process. This intricate process of data exploration and variable selection provides insight into the dataset's depth and boundaries and highlights the complexities associated with bankruptcy prediction.

6.1 Data Description

This section provides a detailed description of the dataset concerning the financial data for bankrupt and active Greek small-medium enterprises (SMEs).

6.1.1 Dataset Source and Composition

The dataset has been derived from the study by *V. Papadouli, E. Houstis, and E. Vavalis* from the *Electrical and Computer Engineering Department, University of Thessaly, Volos, Greece* [1]. In this study, enterprises are recognized as small-medium sized based on two criteria:

1. They have less than 250 employees.
2. Their revenues are less than C50M or assets are less than C43M.

The dataset contains financial reports from Greek firms spanning from 2002 to 2015. The firms' financial data were collected from various sources: the commercial company iMENTOR (which is based on the ICAP database), archives from the Greek financial journal Naftemporiki, and the Datastream platform that provides financial data for numerous countries and markets.

The dataset categorizes enterprises into two distinct groups:

- Bankrupt (B)
- Non-bankrupt (NB)

A firm is classified as financially distressed if it has been declared bankrupt through a court decision; otherwise, it is labelled as non-bankrupt. This classification is carried out using the variable 'label'. The accuracy of each firm's bankruptcy status has been validated through cross-checking with the Greek ELSTAT organization. The dataset comprises records from 170 companies that have filed for bankruptcy and 1,424 firms that have remained non-bankrupt. This distribution underscores a significant class imbalance within the dataset

It is noteworthy that the dataset includes three consecutive years of data for each firm. The year a company declares bankruptcy is labelled as the benchmark year 't'. Consequently, years (t-1), (t-2), and (t-3) represent the 1st, 2nd, and 3rd years before bankruptcy, respectively. The dataset operates under the assumption that the last published balance sheet of a firm in financial distress corresponds to one year before its bankruptcy (t-1).

Later on, the dataset expanded unofficially to encompass 4,782 rows. Of these, 4,272 were labeled as ’0’ and 510 as ’1’. This revised dataset was then provided to us for our research.

6.1.2 Financial Ratios in the Dataset

The dataset integrates financial ratios that gauge firms’ soundness, stability, and performance. These ratios encompass those used in Altman’s Z-score and Taffler’s model, accompanied by additional financial indicators suggested by Carton and Hofer. The listed financial ratios are structured as shown in 6.1.

Henceforth, we will categorize the variables into two distinct groups: the non-ratio variables and the ratio variables. It’s imperative to note that the ratio variables are operationally derived from the non-ratio variables, as previously delineated.

Finally, in the subsequent Data Manipulation section (6.2, we opted to retain all financial ratio variables. These ratios, derived from established financial tenets, provide a comprehensive overview of a company’s fiscal health. They facilitate benchmarking, underpin informed decision-making processes, illuminate emerging financial trends, and enhance transparency [134]. Notably, these inherently scaled ratios neutralize the disparities arising from variations in firm sizes, ensuring a consistent and nuanced analysis across diverse entities.

6.2 Data Manipulation

In the realm of scientific research and data analytics, data manipulation stands as a pivotal process, defined as the systematic adjustment, organization, and transformation of raw data to render it suitable for in-depth analysis. This preparatory step is indispensable for several reasons. First, raw datasets, inherently replete with inconsistencies, errors, and absentees, demand rigorous manipulation to achieve a level of quality and precision vital for subsequent analysis [135], [136]. Second, the effectiveness of data-driven inquiries relies on the well-organized structure of the dataset in question, ensuring that analytical procedures are not only time-efficient but also computationally efficient. Furthermore, adept manipulation enhances the potential of the dataset, unveiling nuanced insights and trends otherwise obscured in its raw state. In essence, data manipulation not only fortifies the integrity of the dataset but also amplifies its analytical potential, thereby serving as the linchpin in the data analysis lifecycle.

6.2.1 Initial Variable Elimination

We proceed by elucidating the rationale behind the selection of specific variables for our analysis, emphasizing the importance of excluding certain variables from our dataset. The choice of variables is paramount in data analysis, as it directly influences the reliability and clarity of the results. While datasets often come replete with a plethora of variables, it's essential to discern which ones truly contribute to the analysis and which might hinder it. Our decisions pivot around three primary criteria: *Redundancy*, *Irrelevance*, *Potential Leakage*.

Judging by *Redundancy*, certain variables such as 'opening_assets', 'current_assets', 'fixed_assets', 'current_liabilities', 'opening_current_liabilities', 'long-term_liabilities', 'opening_equity', 'opening_net-income', and 'opening_capital_employed' might reflect information that their broader category counterparts already capture [137]. For instance, a firm's total assets constitute the sum of its current and fixed assets. Moreover, the term 'liabilities' encompasses all of a company's obligations and debts, including both its short-term (or current) and long-term liabilities.

As for the "opening" variables, they denote the totals of their respective measures at the start of a financial period [137]. For example, the 'opening_current_liabilities' for one period would equal the 'current_liabilities' at the end of the preceding period. As a case in point, the opening liabilities for the 2023 fiscal year would match the liabilities at the close of the 2022 fiscal year. The same principle applies to net income and capital employed.

Such redundancies can induce multicollinearity, a scenario in which two or more predictors in a model are closely correlated. This interrelation makes it arduous to isolate the unique impact of each predictor, potentially compromising the model's stability and clarity.

Furthermore, regarding *Irrelevance*, some variables do not contribute meaningfully to the analysis at hand. For instance, variables like 'vatnumber', 'company', and 'year' function as identifiers or metadata. They don't offer substantive insights into the bankruptcy prediction and can clutter the analysis, diluting the impact of more pertinent variables [135], [136].

Moreover, for *Potential Leakage*, including variables that directly or indirectly reveal the outcome can lead to overly optimistic model performance. For example, 'declared bankruptcy' is a direct indicator of bankruptcy. When such a variable is used as a predictor, it can artificially inflate the model's accuracy, leading to what is termed "data leakage". This occurs because the model gets access to information during training that it shouldn't ideally have, making the predictions trivially accurate but entirely ungeneralizable [135], [136].

To conclude, after our initial variable elimination, what we are left with are, 'fyear', 'scaling', 'inventory', 'receivables', 'assets', 'retained_earnings', 'equity', 'liabilities', 'sales', 'gross_profit', 'EBIT', 'BT', 'net-income', 'EBITDA', 'capital_employed', 'GDP'.

6.2.2 Addressing Missing Values

In data analysis, missing values refer to the absence of data in a dataset. Instead of a recognizable value or piece of data, there might be a blank space, a placeholder like "NaN" (Not a Number), or some other indicator signifying that no data is present. Such omissions can arise for various reasons: data might not be collected, it could be lost, or perhaps it was never applicable in the first place. Whatever the cause, missing values pose challenges in data analysis. They can distort statistical analyses, reduce the power of a study, and lead to potential biases. Addressing these gaps requires understanding why data might be missing and assessing the implications of different strategies to deal with these absences [135], [136].

Greek SMEs Dataset Missing Values

In the meticulous analysis of the company financial dataset, which spans 4,782 rows, two columns stood out for their missing data: 'retained_earnings' and 'inventory'. The 'inventory' column had only 12 missing values out of 4,782, representing about 0.25%. In contrast, the 'retained_earnings' column displayed a significant 525 missing entries, which is roughly 10.98% of the data, prompting an in-depth assessment of potential imputation strategies.

To discern the nature of this missingness, a thorough exploration was undertaken. It was observed that whenever a 'retained_earnings' value was missing, there was a 95.43% chance that the corresponding 'declared bankruptcy' value was also missing. This was especially prevalent among companies that had not declared bankruptcy, as indicated by their label.

However, after closely examining the data, we found a complete correlation between the 'label' column and the missing values in 'declared bankruptcy'. This is because every firm that did not go bankrupt does not have a year of bankruptcy declaration, leading to the NaN value. Looking back at the variable definitions provided earlier, this connection makes logical sense. Thus, the 'retained_earnings' missing values could either be MAR or MNAR, irrespective of the 'declared bankruptcy' column's missingness pattern.

That said, imputing missing values, especially in a column as significant as 'retained_earnings', comes with challenges. Artificial values might not truly reflect a company's financial situ-

ation. This becomes even more critical when trying to determine cause-and-effect relationships. Introducing such synthetic data could skew results, leading to inaccurate conclusions. For causal inference, having genuine and accurate data is crucial; introducing made-up data could introduce biases or hide real trends, affecting the study's credibility.

Considering these challenges, and to ensure the dataset's reliability for in-depth causal inference, the 'retained_earnings' column was removed. Without it, the dataset still boasts 52 columns, all of which provide valuable insights and are rich in information.

Lastly, upon initial inspection, the 'inventory' column, with its minimal percentage of missing values, was deemed a potential candidate for row elimination. However, a more detailed analysis revealed that all instances (100%) of NaN values within 'inventory' were associated with the minority class (label=1). Given this class's scarcity of datapoints, row removal was deemed inadvisable. Additionally, given the column's characteristics, it was assessed that its imputation would not significantly disrupt our causal inference. Consequently, an iterative imputation approach was employed, leveraging a Bayesian Ridge model.

6.3 Data Elucidation: Exploratory Data Analysis

The concept of data elucidation can be understood as the process of clarifying, explaining, and making data comprehensible. It's more than just analyzing data; it's about transforming raw, often confusing datasets into coherent, actionable insights. Through techniques ranging from data visualization to advanced analytics, data elucidation seeks to bridge the gap between mere data collection and meaningful understanding.

6.3.1 EDA: An Overview

Exploratory Data Analysis (EDA) plays a pivotal role in data elucidation, as a preliminary step in data analysis by summarizing the main characteristics of the dataset, often with visual methods. For our dataset, EDA will provide insights into the distribution, variance, and potential outliers among the financial variables of Greek SMEs. This will aid in understanding the general trends and patterns, which can be crucial for subsequent analyses [135], [136], [138].

Some fundamental steps in EDA include:

- **Box Plots:** These can be used to identify the spread and potential outliers in each fi-

nancial ratio. By comparing box plots for bankrupt and non-bankrupt firms, we can identify which ratios might be significant in determining bankruptcy

- **Correlation Analysis:** A correlation matrix or heatmap can be generated to understand the relationships between different financial ratios. High correlation between variables might indicate redundancy.

6.3.2 EDA: Box Plots as a Statistical Tool

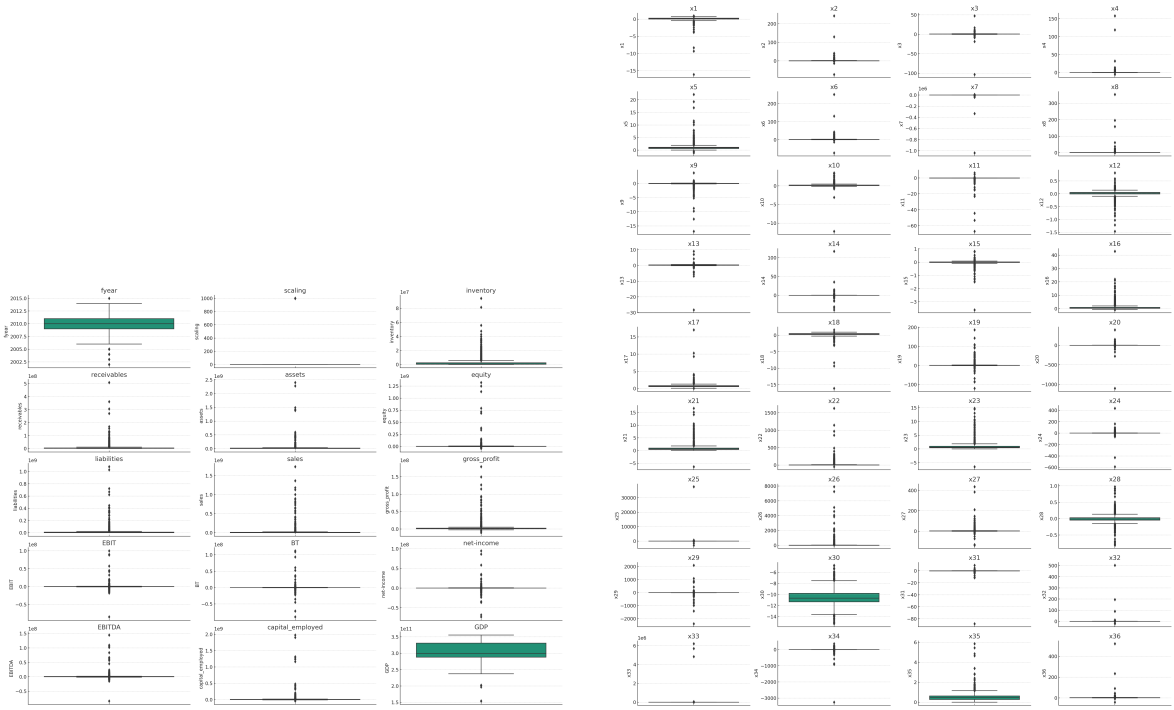
Visualization tools like box plots provide a broader perspective on data distribution. Commonly known as whisker plots, boxplots succinctly highlight the dataset's median, spread, and potential skewness, providing a concise and comprehensive view. Their compact representation ensures that the visualization results are always more collected, manageable, and less prone to misinterpretation. Leveraging these insights from box plots will be a foundational step when we delve deeper into the subsequent correlation analysis.

In our analytical approach, we've focused on the overall distribution of variables rather than plotting against the 'label'. This decision is rooted in our aim to obtain a holistic and unobstructed view of each variable's distribution, thereby ensuring clarity and simplicity in our initial exploratory phase. By examining the entire dataset's distribution, we can more efficiently detect outliers and understand our data's the central tendencies and spread. Furthermore, given that our subsequent analyses involve detailed correlation studies, where intricate patterns and relationships between variables will be explored in-depth, beginning with an overall distribution ensures we don't preemptively compartmentalize our data. This approach provides a broad benchmark, setting the stage for the nuanced, label-specific investigations that follow.

Generating Boxplots for Greek SMEs Dataset Variables

We start by generating a multiplot of boxplots for our selection of non-ratio variables against the target variable 'label. The drawn plots are showcased in Figure 6.1.

Upon inspecting the composite boxplot visualization, it becomes evident that the substantial concentration of outliers obscures any clear patterns within the data. Consequently, we've implemented a systematic outlier removal process. This step essentially helps in clearing up the boxplots and results in the visualization in Figure 6.2. It's crucial to note that the outlier



(α′) Non-ratio Boxplots (β′) Ratio Boxplots

Figure 6.1: Boxplots of Greek SMEs Variables

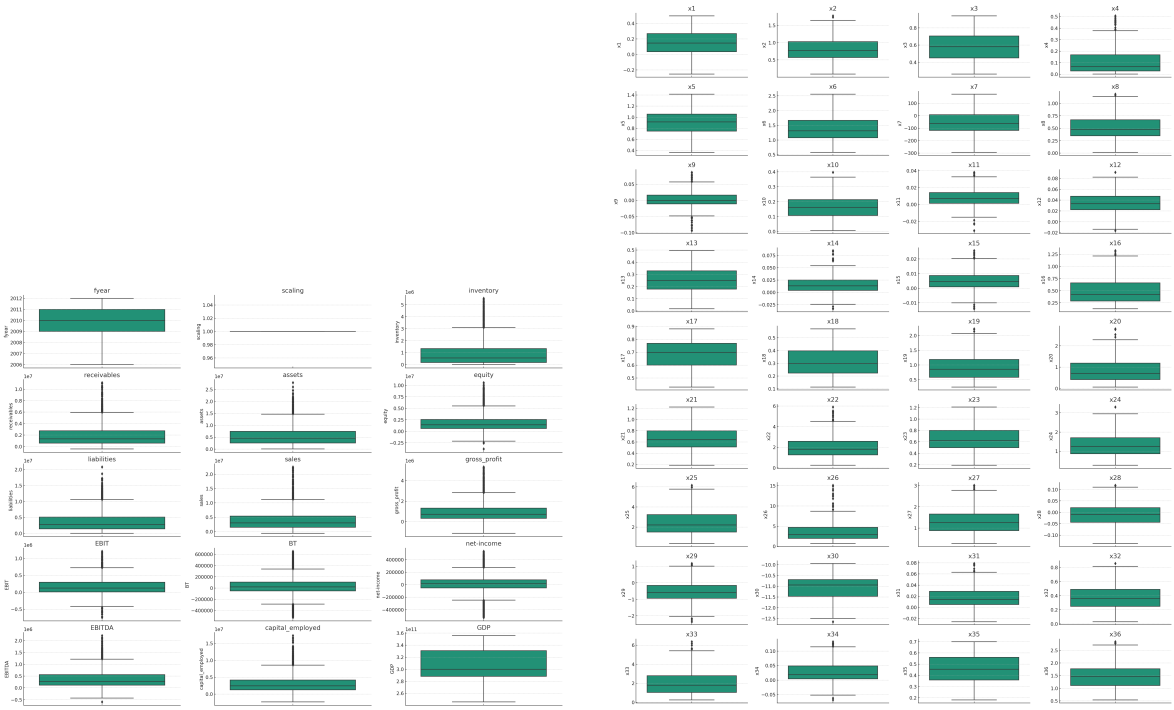
removal is primarily for demonstration purposes. Given that instances of class 1 are predominantly labelled as outliers, removing them would compromise the integrity and purpose of our analysis.

The extended presence of such outliers in the dataset also suggests a likely departure from normal distribution. In practical terms, this indicates that traditional statistical methods, which often assume normality, might not be entirely appropriate for this dataset, and alternative approaches or transformations might be required to derive meaningful insights.

Additionally, the empty boxplot for the ”scaling” variable hints at a lack of variability, suggesting all its values might be identical or near-identical across each ’label’ value. Without variability, a boxplot lacks its typical ”box” and ”whisker” structure. Due to this uniformity, the ”scaling” variable will be subsequently removed from our analyses.

Upon close inspection, it becomes evident that these outlier-free plots, while more aesthetically pleasing, do not necessarily offer substantially new insights. The core distributions, central tendencies, and data spreads remain fundamentally unchanged. Essentially, the removal of outliers served to confirm our initial observations rather than reveal novel patterns.

To further validate our findings and check our assumptions about the data’s distribution,



(α′) Non-ratio Boxplots (Post Outlier Removal) (β′) Ratio Boxplots (Post Removal)

Figure 6.2: Boxplots of Greek SMEs Variables (Post Outlier Removal)

we will be performing a normality test in the section that follows.

Normality Testing of Greek SMEs Dataset

In statistical analysis, the normality assumption is foundational for many parametric tests and methods. It's essential to determine if our data conforms to a normal distribution, as deviations can impact the validity of inferential statistics derived from the data.

To assess the normality of our dataset, we employed the Shapiro-Wilk test [139], a widely used method for testing the normality of a data sample. The test evaluates the hypothesis that a sample is drawn from a normal distribution. A low p-value (typically $p < 0.05$) indicates that the data does not follow a normal distribution. This test was applied to each variable in our dataset, both before and after a systematic outlier removal process. The detailed results of this test can be found in Table 6.2.

The results of the Shapiro-Wilk test on the pre-outlier removal data showed that none of the variables follow a normal distribution. Specifically, the p-values for all the variables were essentially zero, confirming non-normality. This observation is consistent with our earlier visual assessments from the box plots, where the presence of numerous outliers and the shapes

of the distributions hinted at non-normality.

The non-normal nature of our dataset suggests that care must be taken when applying traditional statistical methods, many of which assume normally distributed data. It underscores the importance of thorough exploratory data analysis before more advanced analyses and highlights the need for alternative methods or transformations that can accommodate non-normal data.

6.3.3 EDA: Correlation Among Variables

Following our in-depth exploration of distributions in the *Exploratory Data Analysis (EDA): Box Plots as a Statistical Tools* section, we focus on understanding the interrelationships between the variables. Correlation analysis is a fundamental step in exploratory data analysis (EDA) that helps in identifying potential associations, dependencies, or patterns among multiple variables.

Given the non-normality observed in our data, we have chosen to use Kendall's Tau for our correlation analysis. This method offers superior statistical attributes, especially when dealing with outliers [140]. A detailed explanation of Kendall's Tau will be provided in a subsequent section (see 6.3.4).

In this chapter, we will employ the correlation matrix through the heatmap technique to gauge the strength and direction of said relationships. By doing so, we can discern which variables move together, aiding in feature selection, model building, and hypothesis generation for further analysis. Additionally, this analysis will serve as a foundation for subsequent investigations into whether correlation implies causation.

Additionally, as previously established, we classify our variables into two main categories: *Non-ratio variables* and *Ratio variables*. The distinction is crucial for several reasons,

First, ratio variables are operationally derived from non-ratio variables. Studying the correlation among them separately eliminates redundancy and provides clearer insights.

Second, the division between non-ratio and ratio variables is crucial to causal inference. The ratio variable, being a composite of two or more variables, complicates the tracing back of the causal effect since we cannot readily pinpoint which component of the set of variables that constitutes the ratio is responsible. Due to this inherent complexity with ratio variables, we have an inclination to believe that ratio variables might perform better in terms of classification accuracy, a suspicion we aim to validate, offering a condensed, scaled, and interpretable

set of metrics that encapsulate the financial health and operations of firms. This is irrespective of certain characteristics, such as differences in size, which might otherwise complicate the modeling process. Consequently,

Correlation Matrix Implementation: An Overview

Correlation matrices provide a comprehensive snapshot of the linear relationships between multiple variables in a dataset. When applied to financial variables, these matrices offer insights into the interdependencies among various financial metrics, revealing patterns and associations that might not be immediately obvious. By visualizing these correlations, often through heatmaps, we can quickly discern the strength and direction of relationships between metrics, guiding deeper analysis and strategic decision-making. In the context of our dataset, understanding these correlations among financial variables is pivotal for grasping the financial dynamics and intricacies of the firms under consideration.

Correlation Matrix Implementation: Correlation Among Non-Ratio Variables

By implementing a correlation matrix through heatmap on the non-ratio variables, we achieve the following visualization (Figure 6.3,

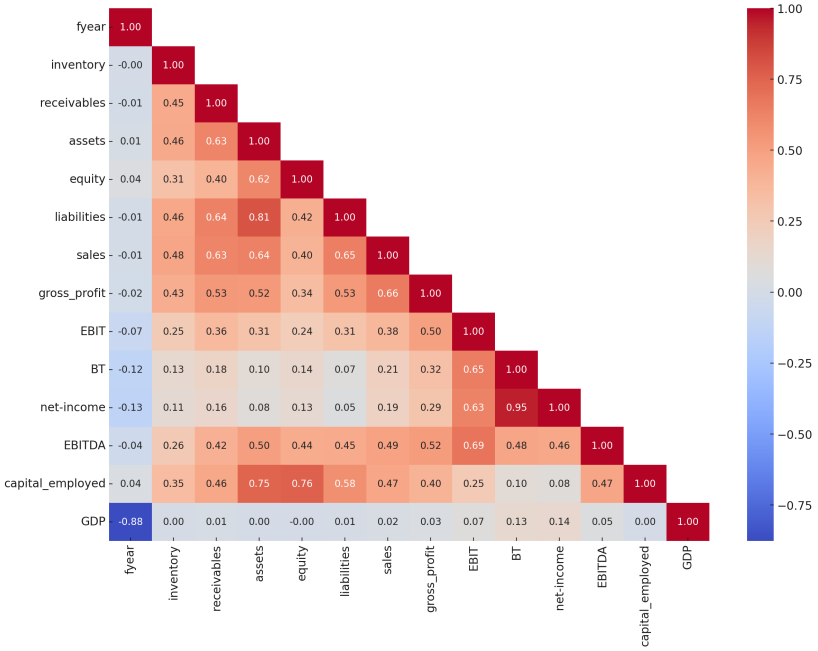


Figure 6.3: Correlation Matrix through Heatmap: Non-Ratio Variables

The insights from this procedure revealed strong correlations (greater than 0.7) which are

noted below,

- **'assets' and 'liabilities' (≈ 0.806):**

  Interpretation: This strong positive correlation suggests that as a company's assets increase, its liabilities typically increase as well. This might be because companies finance the acquisition of new assets through debts, leading to increased liabilities.

- **'assets' and 'capital_employed' (≈ 0.745):**

  Interpretation: Companies with higher assets generally have more capital employed. This indicates that larger companies, in terms of assets, tend to invest more in their business operations.

- **'BT' and 'net-income' (≈ 0.950):**

  Interpretation: Before-tax income (BT) and net income are closely linked since net income is derived after subtracting taxes from BT. A strong correlation is expected because they both are measures of a company's profitability.

- **'capital_employed' and 'equity' (≈ 0.760):**

  Interpretation: Companies with more capital employed likely have higher equity. This suggests that as companies invest more in their operations (using both debt and equity), their shareholders' equity also tends to increase.

- **'GDP' and 'fyear' (≈ -0.877):**

  Interpretation: The negative correlation might suggest that the GDP has been decreasing over the years (fyear). This could be indicative of economic trends where the overall economic growth rate is slowing down over the period covered by the dataset.

Except for the strong correlations, there exist moderate correlations (Figure 6.3 that reveal key insights into a company's financial strategy and operations. Linkages between 'receivables', 'assets', and 'sales' hint at a credit-driven sales approach, indicating the balance between growth and credit risk. The tight interplay between 'assets', 'equity', and 'liabilities' confirms the foundational accounting structure (Assets = Liabilities + Equity). Lastly, the aligned movement among profitability measures like 'EBIT' and 'BT' with 'capital employed' underscores operational efficiency.

Moreover, adding to all the aforementioned, what is most intriguing about the correlation analysis, is that a negative correlation occurs between 'fyear' and 'GDP'. Although the years covered in the dataset saw Greece experiencing both positive and negative GDP growth, there exists an intriguing negative correlation between the fiscal year (fyear) and GDP. This counterintuitive relationship goes through a deeper exploration in the section that follows.

The Negative Correlation Between fyear and GDP

The observed negative correlation between the fiscal year (fyear) and GDP in the dataset can be attributed to the dynamics of the economic trends and the distribution of data points across the years. A closer examination of the data (Figure 6.4 reveals a rise in GDP from 2002 to around 2008, followed by a more prolonged decline from 2008 to 2016.

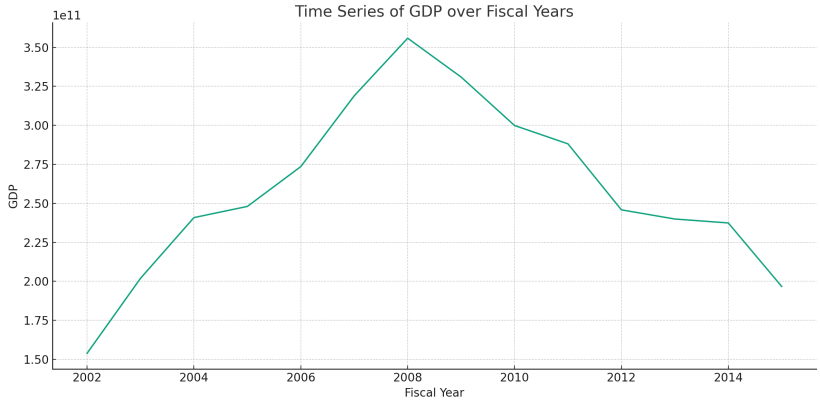


Figure 6.4: Time Series Plot of GDP Over Fiscal Years

Crucially, the dataset contains a significantly larger number of entries post-2008 (4,640 entries) compared to before 2008 (142 entries), which is presented in depth in Figure 6.5.

Given this imbalance, the more extended period of GDP decline, represented by a larger volume of data, exerts a stronger influence on the correlation calculation, thus leading to a pronounced negative correlation. This underscores the importance of understanding the context and distribution of data when interpreting statistical measures, as the correlation is not just a reflection of the economic trend but also of the data's composition over time.

Correlation Matrix Implementation: Correlation Among Ratio Variables

Following our non-ratio variables correlation analysis, we now focus on the ratio variables. By employing a correlation matrix and visualizing it through a heatmap, we obtain the subsequent representation (Figure 6.6), and make some critical observations below.

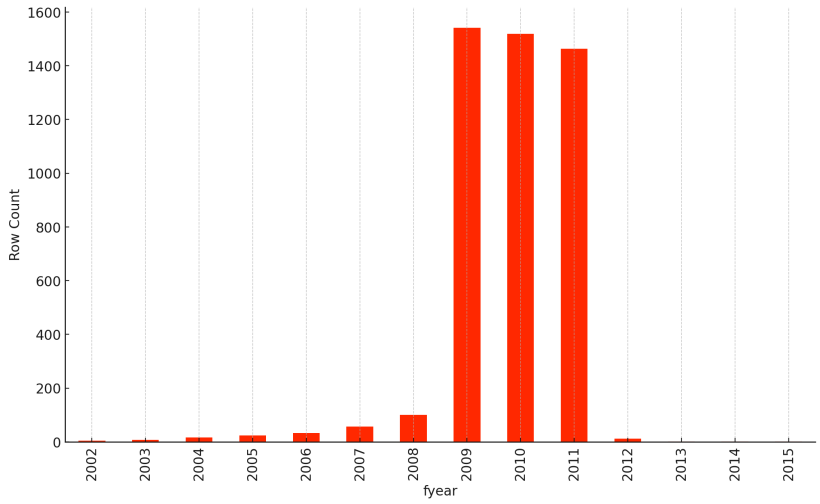


Figure 6.5: Distribution of data rows by year, highlighting the dominance of 2009-2011

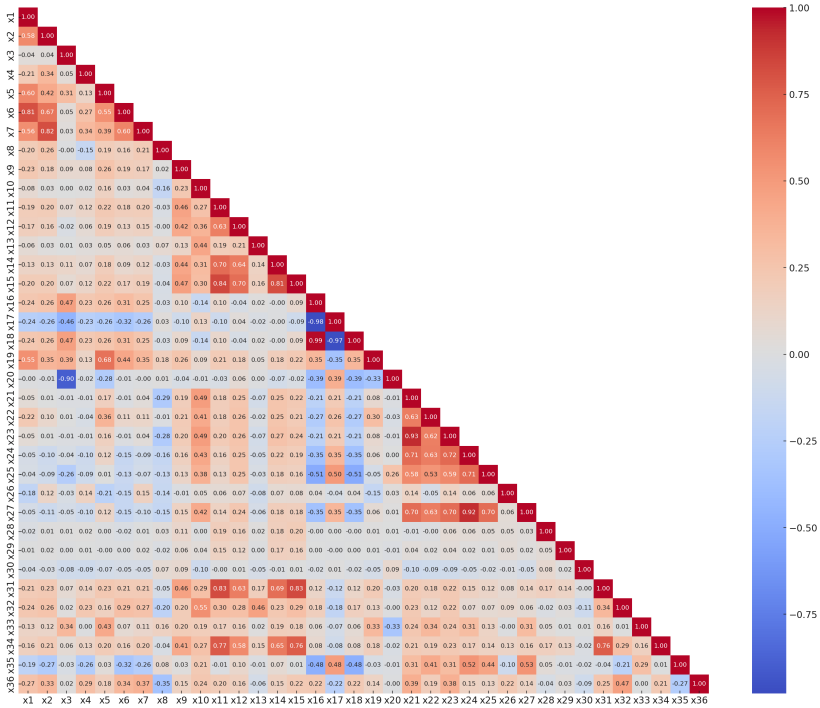


Figure 6.6: Correlation Matrix through Heatmap: Ratio Variables

- **'x1' and 'x6' (≈ 0.808):**

  Interpretation: A higher working capital to total assets ratio ('x1') is associated with a higher current liabilities to current assets ratio ('x6'). This could mean that companies with more working capital relative to their total assets might also have a larger proportion of their assets financed through short-term liabilities.

- **'x2' and 'x7' (≈ 0.822):**

  Interpretation: Companies with a higher ratio of (current assets - inventory) to short-term liabilities ('x2') tend to have longer no-credit intervals in days ('x7'). This might indicate that such companies have stronger liquidity positions and can afford to take longer to pay off their short-term debts.

- **'x11' and 'x14' (≈ 0.703):**

  Interpretation: There's a strong correlation between net profit to sales ('x11') and BT (before-tax income) to capital employed ('x14'). Companies with a higher proportion of their sales translating into net profit likely have efficient operations, leading to higher returns on the capital they've employed.

- **'x11' and 'x15' (≈ 0.843):**

  Interpretation: Net profit to sales ('x11') and net profit to total assets ('x15') being correlated suggests that companies converting a higher proportion of their sales into net profit also tend to utilize their assets efficiently to generate that profit.

- **'x11' and 'x31' (≈ 0.825):**

  Interpretation: Companies that have a higher net profit to sales ratio ('x11') also tend to have a higher profit before tax relative to their current liabilities ('x31'). This can mean that these companies are more capable of covering their short-term obligations using their before-tax profit.

- **'x11' and 'x34' (≈ 0.772):**

  Interpretation: A high correlation between net profit to sales ('x11') and net profit to inventory ('x34') suggests that companies with efficient sales operations (leading to higher net profits from sales) also manage their inventory well, converting it into profit effectively.

- **'x14' and 'x15' (≈ 0.815):**

  Interpretation: Companies that have higher before-tax income relative to their capital employed ('x14') also tend to have a higher net profit in relation to their total assets ('x15'). This indicates the effective utilization of both capital and assets to generate profits.

- **'x15' and 'x31' (≈ 0.833):**

  Interpretation: A strong correlation between net profit to total assets ('x15') and profit before tax relative to current liabilities ('x31') suggests that companies which use their assets effectively to generate net profit also maintain a comfortable profit buffer relative to their short-term obligations.

- **'x15' and 'x34' (≈ 0.761):**

  Interpretation: Companies that have a higher net profit to total assets ratio ('x15') also tend to have a higher net profit relative to their inventory ('x34'). This indicates efficient asset management coupled with effective inventory turnover.

- **'x16' and 'x17' (≈ -0.979):**

  Interpretation: The negative correlation between the book value of equity to total liabilities ('x16') and total liabilities to total assets ('x17') means that companies with a higher proportion of equity relative to their liabilities tend to have a lower proportion of liabilities relative to their total assets. This suggests a more equity-financed capital structure.

- **'x16' and 'x18' (≈ 0.991):**

  Interpretation: A very strong positive correlation between the book value of equity to total liabilities ('x16') and equity to total assets ('x18') indicates that companies with a higher proportion of their capital structure financed through equity also have a significant portion of their assets backed by equity.

- **'x17' and 'x18' (≈ -0.969):**

  Interpretation: The strong negative correlation suggests that companies with a higher proportion of liabilities relative to their total assets ('x17') tend to have a lower equity

proportion relative to their total assets ('x18'). This is expected since a company's capital structure is primarily composed of debt and equity.

- **'x20' and 'x3' (≈ -0.901):**

  Interpretation: Companies with higher long-term liabilities relative to their equity ('x20') tend to have a lower equity proportion relative to their capital employed ('x3'). This suggests that these companies might be leveraging long-term debt as a significant part of their financing strategy.

- **'x21' and 'x23' (≈ 0.930):**

  Interpretation: A strong positive correlation between sales to total assets ('x21') and sales to average total assets ('x23') indicates consistent sales performance relative to the company's asset base over the years.

- **'x21' and 'x24' (≈ 0.711):**

  Interpretation: Companies with higher sales relative to their total assets ('x21') also tend to have a higher capital employed turnover ('x24'). This suggests effective utilization of the capital employed to generate sales.

- **'x21' and 'x27' (≈ 0.705):**

  Interpretation: Companies that have a higher sales to total assets ratio ('x21') also tend to have a higher sales to capital employed ratio ('x27'). This suggests that these companies not only utilize their assets effectively to generate sales but also employ their capital (both equity and debt) efficiently in generating those sales.

- **'x23' and 'x24' (≈ 0.721):**

  Interpretation: A positive correlation between sales to average total assets ('x23') and capital employed turnover ('x24') indicates that companies that maintain consistent sales performance relative to their average assets over time also efficiently turn over their capital employed.

- **'x23' and 'x27' (≈ 0.701):**

  Interpretation: Companies with higher sales relative to their average total assets ('x23') also tend to have higher sales in relation to their capital employed ('x27'). This points to consistent and efficient sales performance relative to both assets and capital.

- **'x24' and 'x25' (≈ 0.709):**

  Interpretation: A positive correlation between capital employed turnover ('x24') and stockholders' equity turnover ('x25') suggests that companies which turn over their capital employed efficiently also tend to turn over their equity effectively. This means they effectively generate sales relative to both their total capital and their equity.

- **'x24' and 'x27' (≈ 0.920):**

  Interpretation: Companies with a higher capital employed turnover ('x24') also tend to have a higher sales to capital employed ratio ('x27'). This underscores the effectiveness of these companies in utilizing their capital to generate sales.

- **'x25' and 'x27' (≈ 0.700):**

  Interpretation: A strong positive correlation between stockholders' equity turnover ('x25') and sales to capital employed ('x27') suggests that companies which efficiently generate sales relative to their equity also tend to do so relative to their total capital.

- **'x31' and 'x34' (≈ 0.761):**

  Interpretation: Companies that have a higher profit before tax relative to their current liabilities ('x31') also tend to have a higher net profit relative to their inventory ('x34'). This indicates not only a strong liquidity position (being able to cover short-term obligations with before-tax profits) but also effective inventory management that leads to high profitability.

In conclusion, numerous financial ratios, derived from a consistent set of non-ratio variables, inherently display robust correlations among each other. This is a direct consequence of sharing common numerators or denominators in their formulations. Such interdependencies are commonplace in financial analysis, leading analysts to often select a subset of ratios to eliminate redundancy and prevent multicollinearity in financial modelling. This careful selection ensures a more robust and clearer interpretation of a company's financial health and performance.

6.3.4 EDA: Correlation Among Variables and Target Variable

This subsection aims to explore the relationships between the dataset's variables and the target variable, 'label', which indicates the financial status of the firms (bankrupt or

non-bankrupt). Knowing that our predictors our continuous variables, and our target is a binary one, we analyze Point-Biserial, Spearman, and Kendall correlations [140], [141], [142], [143], [144].

To delve deeper into our analysis, we not only employed traditional correlation methods but also integrated advanced techniques such as Mutual Information [145], [146] and Feature Importance from Tree-Based Models [147], [148], [149]. Our motivation for incorporating these sophisticated methods stems from a subtle concern: conventional correlation metrics might not fully capture each variable's predictive capacity, especially given the dataset's potential nuances in quality. By broadening our approach, we aimed to more accurately discern the strength and significance of the relationships within our data.

Brief Explanation of the Correlation Metrics

Our metrics were specifically chosen to provide a comprehensive understanding of both linear and monotonic relationships, catering to the various characteristics of our dataset.

- **Point-Biserial Correlation**, is a metric that evaluates the linear relationship between a continuous variable and a binary variable. Like Pearson, its values span from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 signifies a perfect negative linear relationship, and 0 indicates no linear association. It's especially beneficial for datasets with one continuous and one binary variable [1], allowing for insights into potential linear associations between them. However, this metric rests on the assumption of a linear relationship and can be sensitive to outliers.

- **Spearman's Rank Correlation**, gauges the strength and direction of monotonic relationships between two variables. Its values, like Pearson, range between -1 and 1. A score of 1 implies a perfect positive monotonic relationship, -1 indicates a perfect

[1] Even if we recognize our data as deviating from the expected normality, applying the biserial correlation allows us to estimate the magnitude of this deviation. In essence, we're comparing how a metric, which would traditionally be valid for a normally distributed dataset, contrasts when applied to our actual non-conforming data. This provides valuable insights into potential distortions and informs the degree of caution needed in drawing conclusions. Nevertheless, one should always remember that the underlying assumptions of biserial correlation, such as its susceptibility to outliers and the expectation of a linear relationship, remain in play. Thus, while it can be an illuminating tool, its interpretations in such contexts demand a heightened sense of discernment.

negative monotonic relationship, and 0 signifies no monotonic correlation. Spearman's Rank is particularly useful when the relationship between variables is not strictly linear but is consistent in its direction. Unlike Pearson, it doesn't assume linearity or that the data follows a specific distribution.

- **Kendall's Tau**, is another rank-based correlation coefficient that assesses the strength of monotonicity between two variables. Its values typically range between -1 and 1. While computationally more intensive than Spearman, Kendall's Tau is less sensitive to ties in the data. It provides insights into the consistency and direction of relationships, especially valuable when the dataset has a significant amount of tied ranks.

- **Mutual Information**, quantifies the amount of information shared between two variables. Higher values suggest a stronger association, whereas a value of 0 implies no shared information. Mutual Information is versatile, applicable to variables of any type (continuous or categorical), and is especially powerful for uncovering complex, non-linear relationships. It doesn't make assumptions about the functional form of the relationship between variables.

- **Feature Importance from Tree-Based Models**, is not a direct correlation measure but offers insights into the predictive power of variables in the context of a tree-based model. Higher importance scores indicate that a variable is more influential in predicting the target outcome. This approach is particularly valuable for capturing non-linear relationships and interactions between predictors. The importance scores are derived from the structure of the model and how often a predictor is used to split the data.

Correlation Among Non-Ratio Variables and Target Variable

Carrying out the methods mentioned above resulted in the correlation numbers showcased in Table 6.3.

Correlation Among Ratio Variables and Target Variable

Similarly, running the correlation models among ratio variables and target variable resulted in Table 6.4.

Results Interpretation

Based on our analysis, it is clear that most variables display weak correlations with bankruptcy risk. A handful exhibit moderate correlations, yet this underscores the importance of exercising caution when considering them as standalone predictors of bankruptcy. For instance, the 'fyear' variable in the set of financial metrics exhibits a Kendall's Tau correlation of approximately -0.29. Similarly, in the second group, 'x35' has a Kendall's Tau correlation of just above 0.29. These values, while statistically significant, do not offer strong predictive power. Our findings suggest that while some financial metrics can provide insights into bankruptcy risk, their predictive capacities are limited, confirming our suspicions that using them in more sophisticated models might enhance their predictive capabilities.

Moreover, the discrepancies observed between the Point-Biserial and Spearman correlations further underline the importance of understanding the nature and assumptions of different correlation methods. In general, the Point-Biserial combination with the Spearman Rank gives additional information about their variables and their underlying characteristics as, when the Pearson and Spearman values are not much different, our data tends not to have extreme values (outliers) [140]. Many variables exhibit significant discrepancies between their Point-Biserial values and corresponding Spearman Rank values. This confirms our previous observations regarding the variables' distributions being less than optimal.

Equally vital is the clarity brought about by advanced methods such as Mutual Information and Feature Importance derived from Random Forests, which further illuminate the relationships between variables and the target variable. This becomes particularly evident when juxtaposed with the Kendall Rank. While the Kendall Rank is often lauded as a top-tier metric for non-linear data and is regarded as superior among traditional correlation metrics, our findings suggest otherwise. Although the Kendall Rank seems to underestimate many relationships that the Spearman Rank considered more significant, it also tends to overestimate correlations for certain variables when compared to the more advanced methods. The discrepancies observed are far from trivial; in some instances, we noted differences that were an order of magnitude greater, potentially leading to undue confidence in numerous variables.

Furthermore, knowing that the direction of the correlation provides insights into the relationship between the variables and bankruptcy risk (a negative correlation indicates that as the value of the variable increases, the risk of bankruptcy diminishes, conversely, a positive correlation means that a rise in the variable's value is associated with an increased risk of

bankruptcy) a case is to be made for GDP, where it is generally counterintuitive to observe that GDP has a positive correlation with the 'label'.

Typically, one might expect that as GDP (a key economic indicator that measures the overall economic performance and size of a country's economy) increases, the overall health of firms within that country would improve, leading to fewer bankruptcies. However, in our case, it seems that the larger the GDP, the higher the chance of a firm going bankrupt. In the subsequent section, we explore this bizarre relationship between GDP and 'label'.

The peculiarity of the correlation between 'GDP' and 'label'

Upon plotting the distribution of GDP values categorized by labels (Figure 6.7), a notable pattern emerged: firms that went bankrupt had a slightly higher mean and median (306.32×10 9 308.82×10^9 and 318.94×10^9) GDP compared to those that remained solvent (306.32×10^9 and 299.90×10^9), which aligns with the positive correlation observed between GDP and the 'label' suggests that, on average, the GDP was marginally higher during periods with more bankruptcies in our dataset.

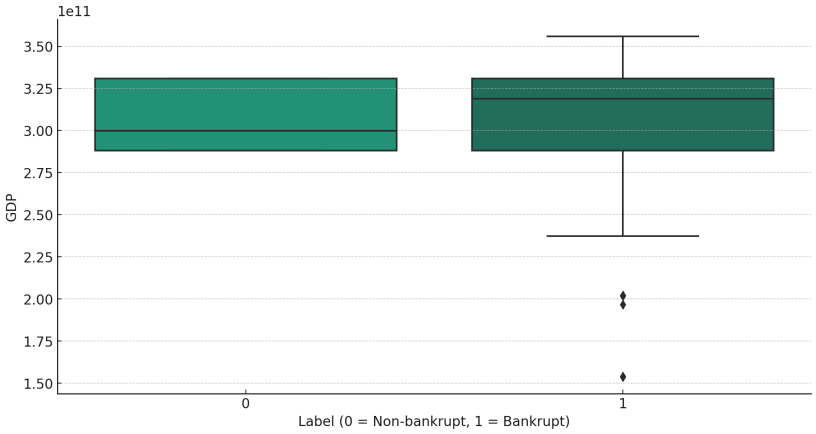


Figure 6.7: Distribution of GDP Values for Each Label

This strange finding prompted a more detailed year-by-year analysis. Intriguingly, the years that recorded the highest GDP values,

- 2007 with a GDP of approximately 318.94×10^9
- 2008 with a GDP of approximately 355.87×10^9
- 2009 with a GDP of approximately 330.95×10^9

- 2010 with a GDP of approximately 299.90×10^9
- 2011 with a GDP of approximately 288.12×10^9

, also aligned with the highest instances of bankruptcies,

- 2007 with 57 bankruptcies
- 2008 with 101 bankruptcies
- 2009 with 117 bankruptcies
- 2010 with 95 bankruptcies
- 2011 with 40 bankruptcies

There is a case to be made for the Greek economic landscape post-2008, given the constant decline in GDP that has forced many firms to close. However, when we examine the distribution of data through the lens of Figure 6.5, we find that the coexistence of GDP's highest values with the most bankruptcies is likely a matter of coincidence rather than causation. This is because it is natural for higher volumes of data (fiscal year values for 2008, 2009, 2010, 2011) to lead to higher instances of bankruptcies. As a result, one should exercise caution when attempting to draw real-world conclusions, as the data one possesses may not necessarily reflect the complexities of the real world.

6.3.5 EDA: Hypothesis Testing Among Variables and Target Variable

Examining the correlation for the variables one by one showed weak correlation between the non-ratio variables and the target variable, as well as the ratio variables and the target variable. Thus, it would be logical to consider testing what the correlation would be when the variables of each group are combined, and then see the correlation for each group with the target variable.

Hypothesis testing is a fundamental approach in statistics used to determine if a result is statistically significant and not just a mere coincidence. In the context of our correlation analysis, hypothesis testing aids us in discerning if the observed correlations between our variables and the target variable are statistically significant or might have occurred by chance.

For each correlation metric and variable under examination, we formulated two hypotheses:

- **Null Hypothesis (H_0):** There's no correlation between the independent variable and the dependent variable ('label').

- **Alternative Hypothesis (H_a):** There's a statistically significant correlation between the independent variable and the dependent variable ('label').

When we look at the p-values from the regression summaries and correlation coefficients, we are essentially testing these hypotheses. A low p-value (typically $p < 0.05$) would lead us to reject the null hypothesis in favour of the alternative, indicating a significant relationship. Conversely, a high p-value would mean that we fail to reject the null hypothesis, suggesting no significant relationship.

As a result, we run two models, one for each variable group, that explicitly states the explained variance for said groups.

The method that is used to test the model significance is the Logistic Regression [150], [151], [152], a method that is considered among the best for binary classification problems.

Correlation Among Non-Ratio Variables and Target Variable

The results from the logit model for non-ratio variables against the 'label' variable are written in Table 6.5.

The pseudo R^2[2] value is 0.3350, indicating that the model performs approximately 34% better than the null model in predicting the label. Moreover, the Log Likelihood Ratio p-value (LLR p-value) is $2.332e-223$, which indicates that the model is, statistically significant at a 95% confidence level.

Furthermore, from the p-values of the coefficients, we can determine the significance of each variable. A common threshold for significance is $p < 0.05$. Based on this,

Significant Variables (p < 0.05):

- fyear: p-value < 0.001

- receivables: p-value < 0.001

- sales: p-value $= 0.041$

[2]McFadden's pseudo R^2 is a measure of the improved fit of the model compared to a model with no predictors. A value closer to 0 indicates that your model is closer in performance to the null model, while values closer to 1 (though it rarely gets anywhere near 1) indicate increasing improvement.

- gross_profit: p-value $= 0.010$
- capital_employed: p-value < 0.001
- GDP: p-value < 0.001

While the aforementioned factors showed noticeable impacts in our analysis, all other variables were determined to be statistically insignificant and did not demonstrate a substantial effect on the outcome.

Additionally, in the quest to unravel the intricate nexus between correlation and causation among the financial variables in focus, the emergence of statistically insignificant predictors introduces a compelling dimension where it is crucial to recognize that statistical insignificance does not inherently signify the lack of a substantive real-world relationship. Rather, it indicates that within the parameters of the available data and the selected model, there isn't sufficient evidence to assert the presence of a discernible relationship. Such a scenario might arise from constrained data sets or extraneous noise masking the genuine effect.

To illustrate, the statistical insignificance of certain variables could spotlight scenarios where a mere correlation does not denote causation. For example, two variables might exhibit concurrent movements (correlation), yet it is plausible that neither directly impacts the other. Alternatively, external variables could exert influence on both, leading to an omitted variable bias. This seeming insignificance should be heeded as a cautionary indication, spurring a more exhaustive exploration into the character of the relationships observed.

This is why, given our scientific endeavour to unravel the relationship between correlation and causality in our dataset, we opt to keep the insignificant variables, as the focus should be on a holistic understanding of the data and its relationships rather than a narrow emphasis on statistical significance.

Correlation Among Ratio Variables and Target Variable

Similarly, we test the ratio variables within the context of a logit model and produce the results printed in Table 6.6.

First, the pseudo R^2 value is 0.4329, which means that the model does approximately 43.3% better than the null model when it comes to predicting the label, which is an upturn from the non-ratio model. Also, the Log Likelihood Ratio p-value is closer to zero[3] than

[3] It's essential to distinguish between statistical significance and the actual strength or magnitude of an effect

before (4.85×10^{-272}), indicating that the model is even more statistically significant.

We once again write the statistical significance state, now using the second variables, keeping in mind all that we previously discussed

Significant Variables (p < 0.05):

- x1: 0.0360
- x3: 0.0003
- x5: 1.99×10^{-8}
- x7: 0.0450
- x9: 1.69×10^{-5}
- x11: 0.0493
- x12: 0.0057
- x15: 0.0005
- x16: 0.0048
- x17: 0.0099
- x19: 9.77×10^{-6}
- x21: 0.0009
- x22: 0.0129
- x23: 0.0051
- x24: 0.0245
- x27: 1.04×10^{-4}

(like correlation). A correlation can be statistically significant but very weak. If you have a large sample size, even tiny, insignificant correlations can be found to be statistically significant. For example, imagine a correlation coefficient r of 0.02 that's statistically significant. This value is close to 0, indicating a very weak linear relationship. However, because of the large sample size, the association is found to be statistically significant. However, in practical terms, this correlation might not be meaningful.

- x28: 5.37×10^{-5}
- x30: 0.0044
- x32: 0.0041
- x35: 6.01×10^{-82}
- x36: 0.0007

Results Interpretation

The primary distinction between the two models lies in their variable composition. While the first model is built on raw financial metrics, the second utilizes ratios derived from those metrics. As we mentioned, these ratio variables not only offer a more condensed and interpretable representation of a company's financial health, but they also provide a level of standardization, crucial when analyzing firms of varying sizes. Such standardization ensures that the model's predictions aren't unduly influenced by the sheer scale of a company's financials but rather by its relative financial health. This operational transformation into standardized ratioed variables appears beneficial, as the second model's pseudo R^2 value surpasses that of the first, suggesting superior predictive capability. These ratios, condense nuanced financial narratives into singular metrics. This enables the model to capture intricate relationships more effectively than using raw metrics alone, explaining why the second model outperforms the first in its predictive prowess.

6.3.6 EDA: Unused Method

Principal Component Analysis (PCA)

PCA, or Principal Component Analysis, is a statistical technique used to simplify the complexity in high-dimensional data while retaining its essential patterns [153]. By transforming the original data into a set of orthogonal components that capture the most variance, PCA aids in visualization, reduces dimensionality, and can help in noise filtering. While it's a valuable tool for data analysis, it operates under linear assumptions and might prioritize features with large variances. The resulting components, though valuable for analysis, may not always have intuitive interpretations.

To understand the causal relationship of various predictors against the ’label’ variable, the dataset was primarily subjected to an initial Boxplot analysis and a subsequent in-depth correlation analysis. PCA (Principal Component Analysis) was not employed and for a good reason, that reason being its lack of interpretability. While PCA is a powerful tool for dimensionality reduction, it achieves this by creating a new set of orthogonal variables (principal components) that are linear combinations of the original predictors. This transformation, although effective in capturing variance, radically alters the interpretability of the original variables. Since our primary objective was to understand the direct influence of these specific predictors on the ’label’, using PCA would have obscured and complicated these direct relationships, making our insights less intuitive and more challenging to relate to the original business context.

Table 6.1: Financial Ratios

Variable	Ratio
	Liquidity Ratios
x1	WORKING CAPITAL/TOTAL ASSETS
x2	(CURRENT ASSETS - INVENTORY)/SHORT-TERM LIABILITIES
x3	EQUITY/CAPITAL-EMPLOYED
x4	(CURRENT ASSETS - INVENTORIES - RECEIVABLES)/SHORT-TERM LIABILITIES
x5	CURRENT ASSETS/TOTAL LIABILITIES
x6	CURRENT LIABILITIES/ CURRENT ASSETS
x7	NO-CREDIT INTERVAL IN DAYS
x8	RECEIVABLES/SALES
	Profitability Ratios
x9	RETAINED EARNINGS/TOTAL ASSETS
x10	GROSS-PROFIT/TOTAL ASSETS
x11	NET-PROFIT/SALES
x12	EBIT/TOTAL ASSETS
x13	GROSS-PROFIT/SALES
x14	BT/ CAPITAL EMPLOYED
x15	NET-PROFIT/TOTAL ASSETS
	Leverage Ratios
x16	BOOK VALUE OF EQUITY/TOTAL LIABILITIES
x17	TOTAL LIABILITIES/TOTAL ASSETS
x18	EQUITY/TOTAL ASSETS
x19	EQUITY/FIXED ASSETS
x20	LONG-TERM LIABILITIES/EQUITY
	Activity Ratios
x21	SALES/TOTAL ASSETS
x22	SALES/FIXED ASSETS
x23	SALES/ AVERAGE TOTAL ASSETS
	Efficiency Ratios
x24	CAPITAL-EMPLOYED-TURNOVER
x25	STOCKHOLDERS-EQUITY-TURNOVER
x26	SALES/INVENTORY
x27	SALES/CAPITAL EMPLOYED
	Growth Ratios
x28	GROWTH RATE OF TOTAL ASSETS
x29	GROWTH RATE OF NET-INCOME
	Size Ratios
x30	LOGARITHM OF TOTAL ASSETS
	Other Financial Ratios
x31	PROFIT BT/ CURRENT LIABILITIES
x32	GROSS-PROFIT/CURRENT LIABILITIES
x33	(CURRENT ASSETS - INVENTORIES)/LONG-TERM LIABILITIES
x34	NET-PROFIT/INVENTORY
x35	CURRENT LIABILITIES/TOTAL ASSETS
x36	SALES/SHORT-TERM LIABILITIES

Table 6.2: P-values from Shapiro-Wilk Test

Variable	**P-Value**	**Variable**	**P-Value**
fyear	0.000	x9	0.000
inventory	0.000	x10	0.000
receivables	0.000	x11	0.000
assets	0.000	x12	0.000
equity	0.000	x13	0.000
liabilities	0.000	x14	0.000
sales	0.000	x15	0.000
gross_profit	0.000	x16	0.000
EBIT	0.000	x17	0.000
BT	0.000	x18	0.000
net-income	0.000	x19	0.000
EBITDA	0.000	x20	0.000
capital_employed	0.000	x21	0.000
GDP	0.000	x22	0.000
x1	0.000	x23	0.000
x2	0.000	x24	0.000
x3	0.000	x25	0.000
x4	0.000	x26	0.000
x5	0.000	x27	0.000
x6	0.000	x28	0.000
x7	0.000	x29	0.000
x8	0.000	x30	7.913×10^{-19}

Table 6.3: Correlation Analysis among Non-Ratio Variables and Target Variable

Variable	Point-Biserial	Spearman	Kendall's Tau	Mutual Information	Feature Importance
fyear	-0.451043	-0.307676	-0.285593	0.138962	0.297629
inventory	-0.068637	-0.106719	-0.087149	0.024649	0.072758
receivables	-0.034923	-0.080731	-0.065924	0.027058	0.043560
assets	-0.046393	-0.172319	-0.140713	0.031140	0.045162
equity	-0.045974	-0.291770	-0.238254	0.057522	0.076513
liabilities	-0.041592	-0.093965	-0.076730	0.023723	0.032276
sales	-0.043489	-0.138686	-0.113249	0.025320	0.041947
gross_profit	-0.061030	-0.128610	-0.105021	0.017789	0.032653
EBIT	-0.028578	-0.109845	-0.089698	0.009302	0.031780
BT	-0.022015	-0.121239	-0.099002	0.021888	0.027733
net-income	-0.020544	-0.109497	-0.089414	0.010756	0.026394
EBITDA	-0.046050	-0.211478	-0.172689	0.026711	0.046469
capital_employed	-0.048158	-0.290893	-0.237538	0.074518	0.121035
GDP	0.036084	0.078571	0.072932	0.143479	0.104091

Table 6.4: Correlation Analysis among Ratio Variables and Target Variable

Metric	Point-Biserial	Spearman	Kendall	Mutual Information	Feature Importance
x1	-0.194939	-0.154050	-0.125794	0.026723	0.018050
x2	-0.038635	-0.136256	-0.111264	0.017259	0.010911
x3	0.010065	0.074306	0.060679	0.062034	0.050054
x4	-0.026360	-0.148601	-0.121345	0.017102	0.015529
x5	-0.039620	-0.044105	-0.036015	0.000000	0.020194
x6	-0.050647	-0.187182	-0.152849	0.021009	0.017796
x7	-0.057290	-0.129719	-0.105926	0.012162	0.011245
x8	0.053245	0.064071	0.052319	0.021805	0.013566
x9	-0.202394	-0.141485	-0.116177	0.028060	0.020038
x10	0.040904	0.025370	0.020717	0.017887	0.008317
x11	-0.111648	-0.122931	-0.100383	0.013405	0.007981
x12	-0.157971	-0.076174	-0.062202	0.026571	0.011903
x13	-0.045590	0.006650	0.005430	0.016507	0.008939
x14	-0.002797	-0.031891	-0.026042	0.029336	0.015228
x15	-0.203002	-0.139780	-0.114142	0.022707	0.013369
x16	-0.109088	-0.277472	-0.226579	0.050838	0.033444
x17	0.242674	0.278868	0.227719	0.046214	0.029050
x18	-0.238440	-0.276446	-0.225741	0.049428	0.021160
x19	0.048557	-0.044057	-0.035976	0.031017	0.022866
x20	-0.009666	-0.127831	-0.104436	0.085333	0.077650
x21	0.056376	0.047918	0.039129	0.006996	0.008777
x22	0.092799	0.144899	0.118322	0.031672	0.012072
x23	0.051078	0.043795	0.035762	0.005462	0.008967
x24	0.014407	0.129928	0.106097	0.046645	0.014570
x25	-0.002896	0.120050	0.098031	0.037008	0.011158
x26	-0.015600	-0.102530	-0.083731	0.032131	0.038391
x27	0.123811	0.121824	0.099479	0.046121	0.020214
x28	-0.004549	0.010595	0.008652	0.025141	0.016239
x29	-0.004744	-0.013104	-0.010700	0.008431	0.006571
x30	-0.136551	-0.129018	-0.105354	0.024154	0.015911
x31	-0.015735	-0.141755	-0.115755	0.006678	0.008241
x32	-0.015281	-0.149534	-0.122107	0.019349	0.009284
x33	-0.008735	-0.111269	-0.090905	0.081480	0.096498
x34	0.003806	-0.095087	-0.077652	0.042193	0.008140
x35	0.454503	0.349188	0.285141	0.120446	0.282151
x36	-0.038072	-0.213824	-0.174605	0.026416	0.015528

Description	Value
Dep. Variable	label
No. Observations	4782
Model	Logit
Df Residuals	4767
Method	MLE
Df Model	14
Pseudo R-squ.	0.3350
Time	19:43:28
Log-Likelihood	-1079.4
Converged	True
LL-Null	-1623.3
Covariance Type	nonrobust
LLR p-value	2.332e-223

Table 6.5: Logit Regression Results for Non-Ratio Variables

Dep. Variable	label
No. Observations	4782
Model	Logit
Df Residuals	4745
Method	MLE
Df Model	36
Pseudo R-squ.	0.4329
Time	19:43:52
Log-Likelihood	-920.59
Converged	True
LL-Null	-1623.3
Covariance Type	nonrobust
LLR p-value	4.850e-272

Table 6.6: Logit Regression Results for Ratio Variables

Chapter 7

Implementation: Causal and Bayesian Approaches in Bankruptcy Analysis

In this chapter, we delve into applying causal inference, enhanced by Bayesian techniques, to examine the intricate landscape of bankruptcy within Greek SMEs. We strive to decipher the complex causal relationships pivotal to bankruptcy events using our Greek SMEs dataset. We begin by contrasting a custom model with the 'bnlearn' package, aiming to elucidate the interplay between correlation and causation in our dataset. Proceeding further, through parameter learning, we draw inferences on the nuances of the most pronounced causal relationships and their interrelations. Subsequently, we devise an innovative strategy to address the confounding factors present in our dataset. After mitigating these confounders, we revisit the causal analysis from the ground up, culminating in a detailed presentation of our results.

7.1 Bayesian Networks: Bridging Correlation and Causation

To examine the interplay between correlation and causation within our dataset, we turned to Bayesian Networks for insights. Initially, we constructed a Bayesian network based on a correlation-centric methodology. In parallel, we utilised 'bnlearn' to craft a network to identify potential causal relationships. It's important to note, however, that while 'bnlearn' provides a stronger indication of causal relationships than mere correlation, it does not guarantee true causal relationships. If 'bnlearn' does not identify a relationship as causal despite

a correlation being present, it could possibly suggest that for that specific relationship, correlation does not imply causation. By comparing these two networks, one rooted in correlation and the other offering insights into potential causation, we aimed to elucidate the intricate correlation-causation dynamic present in our dataset.

Moreover, for the subsequent causal analysis, we deliberately chose to analyse the dataset in its entirety, rather than segmenting the variables based on non-ratio and ratio distinctions, even when the latter are operationally derived, as detailed in our preceding chapters. This decision was based on the findings from Schisterman et al.'s research on *Schisterman et al*.'s research on *"Collinearity and causal diagrams"*, [154]. Their investigation underscored that while high correlation between variables can lead to inflated standard errors, the causal relationship between these variables is paramount in ensuring valid conclusions. Hence, the primary focus should be on the interplay of causal relationships when deciding on including covariates in regression formulations. It is worth noting that the robustness of any study ultimately depends on the validity of the assumptions underpinning its conclusions. For instance, an inaccurately constructed Directed Acyclic Graph (DAG) could result in misleading conclusions, irrespective of the challenges posed by collinearity.

7.1.1 Building a Custom Bayesian Network

Inspired by the methodology presented in *"Using Bayesian Networks for Bankruptcy Prediction: Empirical Evidence from Iranian Companies"* by *Arezoo Aghaie and Ali Saeedi* [99], our analysis aims to identify the predictors influencing a firm's bankruptcy status. Recognizing the non-normal nature of the data, we once again use Kendall's rank, for our computations.

Our steps involve,

1. Using the Kendall's Tau correlations among all potential predictors and the firm's bankruptcy status.

2. Identifying first-order variables, which were those directly correlated with the bankruptcy status.

3. For each first-order variable, using Kendall's rank again, determine the second-order variables. These are predictors significantly correlated with first-order variables, but not directly with the bankruptcy status.

Initially, using Kendall's Tau correlations, we calculate the first-order variables that have a significant correlation with the firm's bankruptcy status (using an absolute correlation threshold of 0.2[1]) are, **fyear**, **equity**, **capital_employed**, **x16**, **x17**, **x18**, **x35**.

Furthermore, to identify second-order variables for each first-order variable, we use Kendall's Tau correlations between the first-order variable and all other predictors. Once again, any predictor with an absolute correlation exceeding a threshold of 0.2, and not directly correlated with bankruptcy status, was considered a second-order variable. These variables indirectly influence bankruptcy through their significant association with first-order variables.

Nevertheless, we decided to raise the lower threshold to 0.4, focusing on retaining the stronger and more significant relationships. This not only simplifies our model or representation, but also helps filter out noise, ensuring that the relationships we consider are more likely to be impactful and meaningful.

Moreover, we set an upper correlation threshold of 0.9, which should not be surpassed, we're eliminating relationships that are so closely tied that they generally are mathematical artifacts rather than meaningful insights. This is especially relevant for ratio variables. If two ratio variables share common elements in their numerators or denominators, they can exhibit a high correlation not due to a true underlying causal relationship but merely due to their shared mathematical structure. Removing such variables prevents misleading interpretations and focuses the analysis on more unique and informative relationships.

Our pre and post removal results are,

- **fyear:**
  - Pre-Threshold (0.2): GDP
  - Post-Threshold (0.4-0.9): GDP
- **x35:**
  - Pre-Threshold (0.2): equity, capital_employed, x2, x4, x6, x7, x10, x16, x17, x18, x21, x22, x23, x24, x25, x27, x32, x33, x36
  - Post-Threshold (0.4-0.9): x16, x17, x18, x22, x24, x25, x27
- **equity:**

[1] A correlation threshold of 0.2 is generally not considered substantial. However, given the magnitude of the correlations in this dataset, we consider it adequate.

  - Pre-Threshold (0.2): inventory, receivables, assets, liabilities, sales, gross_profit, EBIT, EBITDA, capital_employed, x16, x17, x18, x24, x25, x27, x30, x35
  - Post-Threshold (0.4-0.9): assets, liabilities, sales, EBITDA, capital_employed, x30

- **capital_employed:**
  - Pre-Threshold (0.2): inventory, receivables, assets, equity, liabilities, sales, gross_profit, EBIT, EBITDA, x22, x24, x27, x30, x35
  - Post-Threshold (0.4-0.9): receivables, assets, equity, liabilities, sales, gross_profit, EBITDA, x30

- **x17:**
  - Pre-Threshold (0.2): equity, liabilities, x1, x2, x3, x4, x5, x6, x7, x16, x18, x19, x20, x21, x22, x23, x24, x25, x27, x35, x36
  - Post-Threshold (0.4-0.9): x3, x25, x35

- **x16:**
  - Pre-Threshold (0.2): equity, liabilities, x1, x2, x3, x4, x5, x6, x7, x17, x18, x19, x20, x21, x22, x23, x24, x25, x27, x35, x36
  - Post-Threshold (0.4-0.9): x3, x25, x35

- **x18:**
  - Pre-Threshold (0.2): equity, liabilities, x1, x2, x3, x4, x5, x6, x7, x16, x17, x19, x20, x21, x22, x23, x24, x25, x27, x35, x36
  - Post-Threshold (0.4-0.9): x3, x25, x35

Finally, using 'networkx', a popular library for plotting Directed Acyclic Graphs (DAGs), we plot the relationships. The resulting DAG is showcased in Figure 7.1 below.

7.2 Structure learning

We undertook structure learning, a process of constructing a DAG, employing all algorithms available in bnlearn to assess the outcomes.

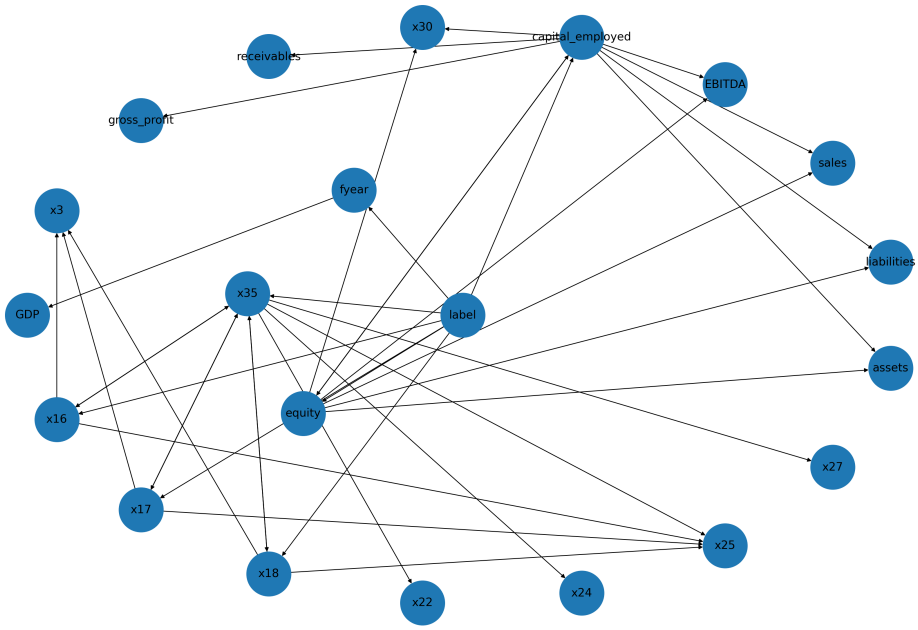


Figure 7.1: Custom Correlation-Based DAG

7.2.1 bnlearn Deployment

The Bayesian Neural Network model is implemented using the bnlearn Python package, which we thoroughly explained in Chapter 3. The package implements five key functions,

- **bn.structure_learning.fit()** Structure learning
- **bn.independence_test(model, df, test='chi_square', prune=True)** Compute edge strength with the test statistic
- **bn.parameter_learning.fit()** Parameter learning
- **bn.inference.fit()** Inference
- **bn.plot()** Plot graph

Our initial exploration was with the ExhaustiveSearch approach. This algorithm evaluates every conceivable DAG and identifies the highest-scoring one [155]. We refrained from applying this method to our data set, since it was designed for compact networks with a maximum of 15 nodes.

Our subsequent test was with the *HillClimbSearch* algorithm. This employs a greedy local search, commencing from a default disconnected DAG, termed 'start'. The process

iteratively modifies individual edges to optimise the score until a local maximum is identified [155]. Intriguingly, we saw the identical DAG meaning that even in the unpruned DAG all the edges were significant. Our pruning process involves computing edge strengths via the chi-square independence test and eliminating non-significant edges. The refined tree features three nodes: 'GDP', 'label', and 'fyear'.

A prominent observation was the apparent lack of a robust causal link between any endogenous variables and the 'label', which represents bankruptcy, in the dataset. The causal relationships are as follows: the 'label' variable is caused by 'fyear', which in turn is caused by 'GDP' as seen in Figure 7.2, which is exactly like our custom dataset.

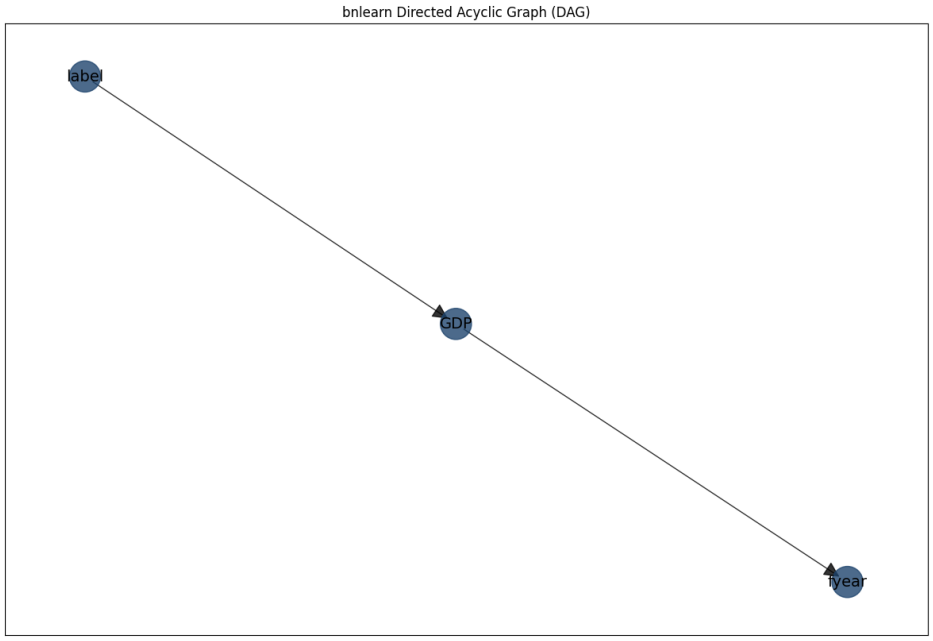


Figure 7.2: DAG with pruned edges created with Hillclimbsearch algorithm

We also employed the *Chow-Liu* method as our third approach. The Chow-Liu Algorithm is a tree search-based approach that finds the maximum-likelihood tree structure where each node has at most one parent. The complexity can be limited by restricting to tree structures which makes this approach very fast to determine the DAG using large datasets (aka with many variables) but requires setting a root node[155]. As with our previous models, we executed a pruning process. The discerned causal chain suggests that the 'label' variable is influenced by 'fyear', which, in turn, is affected by 'GDP', as depicted in Figure 7.3.

The results presented conflict since 'fyear' and 'GDP' have switched places. The implications and meaning behind this switch will be delved into further in 7.2.3.

The final algorithm we employed was the *Naive Bayes*. Naive Bayes represents a specific

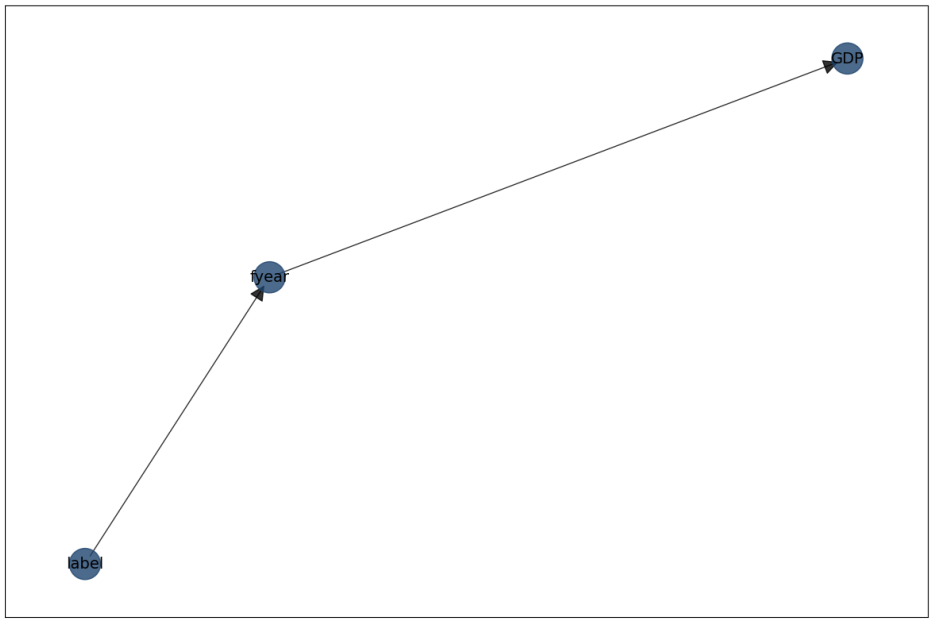


Figure 7.3: DAG with pruned edges created with Chow-liu algorithm

instance of a Bayesian model, characterised by edges only from feature variables to the dependent variable[155]. Post-pruning, the resulting DAG is more intricate than those derived from the previous two models. Notably, the 'label' variable, which is of prime interest, has connections to five nodes: 'fyear', 'GDP', 'x9', 'x20' and 'x33', as depicted in figure7.4. However, the associations with variables 'x9', 'x20' and 'x33' are less significant, implying a weaker causal relationship with the 'label'.

In the exploration of endogenous causally significant variables, three prominent ratios emerge, x9 (RETAINED EARNINGS / TOTAL ASSETS), x20 (LONG-TERM LIABILITIES / EQUITY), x33 ((CURRENT ASSETS - INVENTORIES) / LONG-TERM LIABILITIES). Intriguingly, each of these variables embodies components of both liabilities and equity, drawing a parallel to the foundational accounting equation wherein total assets are the cumulative sum of liabilities and equity (Total Assets = Equity + Liabilities). This observation accentuates the pivotal role that 'liabilities' and 'equity' have a causally significant relationship with the outcome, 'label'.

Nevertheless, the influence exerted by these endogenous ratios is not homogenously distributed across all models, especially when contrasted against potent exogenous determinants such as 'GDP' and 'fyear'. A critical examination of discrepancies in DAGs derived from alternative methodologies instead of the extant method suggests a potential shortcoming of the NaiveBayes technique. This method tends to overemphasise node interrelationships due to

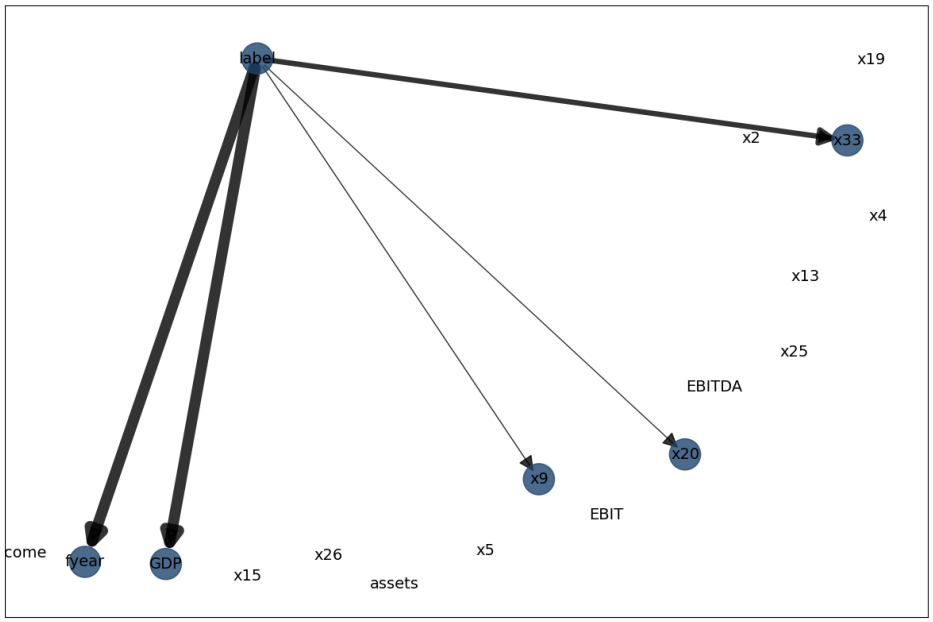


Figure 7.4: DAG with pruned edges created with Naive Bayes algorithm

its inherent inability to account for dependencies within the dataset, predicated on the presumption of predictor independence.

In our comprehensive exploration of Bayesian methodologies, as facilitated by the bnlearn Python package, we sought to unravel the intricate web of causal relationships underlying our dataset. We attempted to clarify the essence of causation from mere correlation through the systematic implementation of algorithms, including hillclimbersearch to the Chow-Liu method and finally the Naive Bayes. Of particular note were the causal relationships among 'GDP', 'fyear', and the 'label' variable, which epitomises bankruptcy. Although, each algorithm brought forth a unique perspective, inherent ambiguities make it more difficult to carry out a cohesive interpretation.

Our findings establish a compelling foundation; they underscore the causal relationships that external factors maintain with the bankruptcy status of a company. The results suggest that during years of economic downturn, the broader economic climate—reflected by the GDP factor—can directly lead certain companies to experience bankruptcy.

7.2.2 Inference on the Correlation-Causation Relationship

The observation that only two exogenous variables emerge as candidates for causal relationships prompts us to contemplate the broader implications for our dataset. Specifically, it

raises the possibility that, for the majority of variables, correlation might not necessarily imply causation. Yet, it is crucial to tread with caution. While bnlearn offers valuable insights, its determinations are not absolute. There exist inherent conditions and assumptions underpinning its outputs. As such, what we truly garner from this exercise is not definitive proof, but rather an enriched layer of probabilistic information concerning the relationships within our dataset.

7.2.3 'fyear' as a Confounding Factor

In our implementation, the variable 'fyear' is evidenced from the examples to influence a firm's bankruptcy. We can not understand if it directly affects 'label' or through the variable 'GDP but the causal relation is clear. However, if viewed purely theoretically, the year in which a firm files for bankruptcy shouldn't affect its insolvency in any conceivable manner; such a notion seems irrational. Consequently, one might suspect that 'fyear' may be a confounding variable, creating a backdoor path indicative of a false causal relationship, as it is plotted in the figure 7.5.

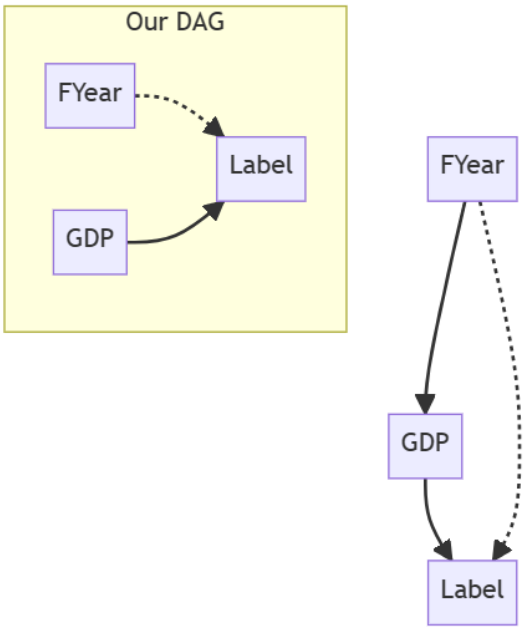


Figure 7.5: Our DAG versus the true association between 'fyear' and 'label'

GDP	fyear
153830000000	2002
201920000000	2003
240850000000	2004
248010000000	2005
273560000000	2006
318940000000	2007
355870000000	2008
330950000000	2009
299900000000	2010
288120000000	2011
245810000000	2012
239930000000	2013
237410000000	2014
196690000000	2015

Table 7.1: GDP and Fiscal Year

7.2.4 Effectiveness of Causation over Correlation

Delving deeper into the relationship between fiscal year and Gross Domestic Product, every value of 'fyear' corresponds precisely to a specific value of GDP, indicating a strong relationship between these variables as seen in Table 7.1.

In practical terms, they can be used interchangeably. However, without recognising these causal relationships, relying solely on correlation might not fully capture the depth of the link between 'fyear' and 'GDP'. This becomes particularly clear when juxtaposing this understanding with the visual representation of 'GDP' plotted against 'fyear' (Figure 6.4). If such a tight relationship were less intuitive in a scenario involving different variables, merely relying on correlation could lead us to overlook it. In summary, in this specific case, correlation alone cannot convey the true interchangeability of the two variables.

fyear	label(0)	label(1)
2002	0.007484	0.040311
2015	0.007484	0.036351
2003	0.007484	0.042291
2014	0.007484	0.036351
2013	0.007484	0.036351
2004	0.007484	0.051202
2012	0.007484	0.047242
2005	0.007484	0.059123
2006	0.007484	0.068034
2011	0.305892	0.074965
2010	0.305892	0.129420
2007	0.007484	0.091796
2009	0.305892	0.151202
2008	0.007484	0.135361

Table 7.2: CPDs of ’GDP’ and ’label’ with ’fyear’ as Index

7.3 Parameter Learning

Parameter learning involves estimating the values of Conditional Probability Distributions (CPDs). Conditional Probability Distributions (CPD) quantitatively describe the statistical relationship between each node and its parents and can be computed using Parameter Learning [29].

In our study, we opted for the Hillclimbsearch algorithm for parameter learning, primarily because of its unmatched precision in identifying inter-variable relationships. What’s remarkable is its ability to discern the desired outcome even when we have not explicitly defined it. This not only underscores the algorithm’s efficacy in pinpointing both the outcome and the inherent relationships, but also reinforces the authenticity of the data. The results corroborate our initial intuitions, further testifying to the algorithm’s robustness and reliability.

By executing the parameter learning function of bnlearn, we calculate all the Conditional Probabilities of the relationships presented in Figure 7.2 and print the results of ’GDP’ in Table 7.2.

In this table, we have interchanged the 'GDP' values with their corresponding 'fyear' values. However, this does not represent the CPD of the 'fyear' and 'label'. The interchange was done solely to enhance the table's readability and comprehension. From the table, we observe that the only years in which a company had a higher probability of not filing for bankruptcy are 2009, 2010, and 2011. This observation does not match up with the general economic state of Greece, suggesting that in all other years, a company was significantly more likely to go bankrupt. Also, it is puzzling to note that during the years 2009 to 2011, Greece was in the midst of a financial crisis. Those discrepancies prompt us to further investigate why our data do not align with the broader economic climate in Greece. After plotting the histogram for the 'fyear' variable, we recognized a bias in the labels concerning the 'fyear' as seen in figure 7.6.

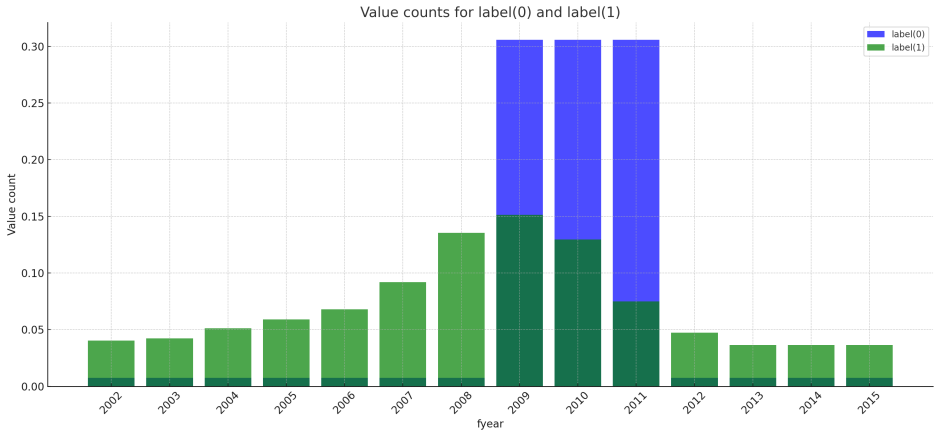


Figure 7.6: Histogram of the 'fyear' values

From 2002 to 2006 and 2012 to 2015, we only have variables labelled 1. Interestingly, our data shows a paradox during the 'fyear' 2009. While there is a pronounced surge in instances labelled as 1, indicating a specific trend or outcome, there is also a sudden burst of instances labelled as 0 which in previous years was zero. This coexistence of both labels in such a critical year creates inconsistencies in the data categorisation process.

Those as mentioned above are indicative of existing bias when the dataset was created, making our dataset challenging for establishing causal relationships. Consequently, our data do not accurately reflect the real economy of Greece. It does not serve as a representative sample or have significance for analyzing the entire time period. Yet, there is an inherent uniqueness to our data. We can derive insights specific to our dataset and draw localized conclusions, which hold value as a distinct case study.

7.4 Establishing the Treatment Effect

From a practical perspective, our dataset presents significant challenges for conducting a causal analysis of Greek SMEs. The data collection process was inherently biased, compromising the dataset's validity. As a result, drawing any logical conclusions from this data becomes an untenable task.

Given these constraints, a crucial question emerges: How can we effectively conduct an Average Treatment Effect analysis (ATE) in this context? If the data had been more comprehensive and robust, it would have served as a solid foundation for examining the financial consequences of the Greek crisis in 2012. Such an investigation would have shed light on the extent and nature of its effects on Greek SMEs during that turbulent time. Perhaps we could have analysed the impact of a significant event during the stable years of 2009-2011, but no such pivotal event occurred in that period. While examining various firms and establishing multiple treatment studies might seem like a solution, the available data for each firm are limited, especially when compared to the dataset as a whole.

Embracing our dataset's inherent flaws and characteristics is the key to devising a successful analytical approach. The first step in this direction is acknowledging and accepting the notable label imbalance spanning. This acknowledgement propels us to craft a hypothetical case study, imagining an event responsible for such a significant imbalance. Intriguingly, this event could be the inherent bias under which the dataset was constructed. We call this bias, **"Construction Bias"**.

In our scenario, such an event is evident due to its pronounced impact on bankruptcy outcomes, as demonstrated by the conditional probabilities in our parameter learning procedure. However, consider a different scenario where the effects on the outcomes are more subtle, almost imperceptible. In such cases, the application of ATE (Average Treatment Effect) becomes invaluable. It holds the potential to unearth the existence of such impactful events, providing insights that would otherwise remain concealed.

7.5 Uncovering the Effect of Construction Bias Using causal-inference

In our analytical journey to identify potential biases inherent in the dataset's construction, we turned our attention to the 'fyear' variable. As previously highlighted, this variable showcases a pronounced relationship with the label.

7.5.1 'GDP' as a Confounding Factor

We've previously identified a pronounced confounding relationship between the exogenous variables 'GDP', 'fyear', and 'label'. Earlier sections discussed the contextual logic behind 'fyear' potentially acting as a confounder. However, delving into the issue of dataset Construction Bias reveals another dimension. The decision to exclusively include bankrupt firms for specific years has inadvertently bestowed significant predictive power upon 'fyear'. This skewed distribution of label values across fiscal years is not a random occurrence, but a direct consequence of choices made during dataset creation. This prominence of 'fyear' as a predictive variable is further corroborated by specific Bayesian models we've examined.

Implementation

To further understand the nature of this bias, we introduced a new variable named 'bias'. In this variable, we labelled the years 2002-2008 and 2012-2015—where the 'label' consistently displays a value of 1—as biased, assigning them a value of 1. In contrast, the years 2009-2011, which are not indicative of such , which did not follow this trend, were marked with a value of 0, indicating their potential lack of bias.

Our analysis, we employed the 'CausalModel()' function from the 'causalinference' library. We designated 'label' as the outcome variable for this model, while 'bias' was used to represent the treatment. All other variables, except for 'fyear', were incorporated as covariates. The outcomes derived from this modelling exercise will be detailed in the following bulleted structure.

- **ATE (Average Treatment Effect)**

  Estimate (Est.): The estimated ATE is 0.921. This suggests that, on average, the treatment increases the outcome by 0.921 units compared to not receiving the treatment

(while controlling for the covariates).

Standard Error (S.e.): The standard error of the ATE is 0.083, quantifying the uncertainty associated with the estimate.

z-value: The z-value for the ATE is 11.091. This high value indicates a statistically significant difference.

P-value (P>|z|): The p-value for the ATE is 0.000, below the conventional significance level of 0.05, signifying that the treatment effect is statistically significant.

Confidence Interval ([95% Conf. int.]): The 95% confidence interval for the ATE is [0.758, 1.084]. As this interval is entirely above 0, it further affirms the statistical significance of the treatment effect.

- **ATC (Average Treatment Effect on the Controls)**

  Estimate (Est.): The estimated ATC is 0.944. This indicates that, on average, control units would experience an increase in the outcome of 0.944 units if they received the treatment.

  Standard Error (S.e.): The standard error of the ATC is 0.087.

  z-value: The z-value for the ATC is 10.838.

  P-value (P>|z|): The p-value for the ATC is 0.000, further confirming its statistical significance.

  Confidence Interval ([95% Conf. int.]): The 95% confidence interval for the ATC is [0.774, 1.115]

- **ATT (Average Treatment Effect on the Treated)**

  Estimate (Est.): The estimated ATT is 0.516. This denotes that those who received the treatment experienced an average increase in the outcome of 0.516 units compared to their outcome would have been had they not received the treatment.

  Standard Error (S.e.): The standard error of the ATT is 0.037.

  z-value: The z-value for the ATT is 14.011.

  P-value (P>|z|): The p-value of the ATT is 0.000, which solidifies its statistical significance.

Confidence Interval ([95% Conf. int.]): The 95% confidence interval for the ATT is [0.443, 0.588].

7.5.2 Filtering Out the Construction Bias

From the previous section, we determined the presence of Construction Bias by quantifying it using the Average Treatment Effect. This naturally leads to the question: What would the causal dynamics look like without Construction Bias?

To simulate the absence of Construction Bias, we removed all dataset instances that have 'fyear' values with a 100% dominance of 'label'=1 values. We then conducted the DAG analysis again using the Hillclimbsearch algorithm. This revealed 'GDP' as the sole strong causal relationship, as depicted in Figure 7.7 below.

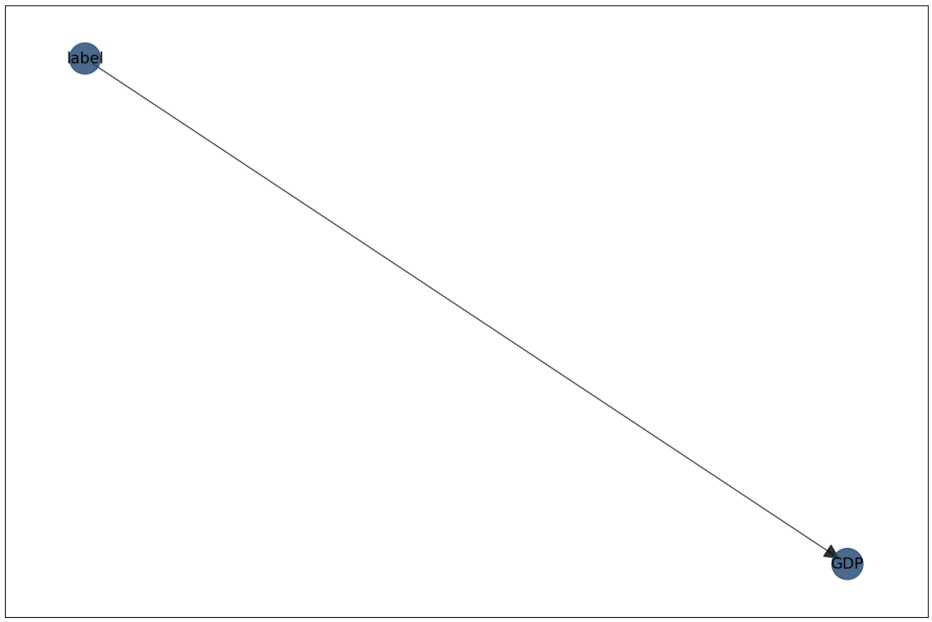


Figure 7.7: 'GDP' as the sole significant relationship

With our data removal, we eliminate the subtlety surrounding which variable, 'GDP' or 'fyear', acts as the confounder. More critically, we lose insight into why this happens, which is, of course, attributable to Construction Bias.

We validate further by running a parameter learning on the model, which resulted in the CPDs in Table (7.3).

Based on the CPDs produced, we observe a paradox: once again, larger GDP values are associated with a higher probability of bankruptcy.

label	label(0)	label(1)
2011	0.33333333333333337	0.27482269503546103
2010	0.33333333333333337	0.3479609929078014
2009	0.33333333333333337	0.3772163120567376

Table 7.3: CPDs of ’GDP’ and ’label’ with ’fyear’ as Index, post Construction Bias removal

To address this concern, we refer to Figure 7.6. Here, we note an imbalance between ’label’ values of 0 and 1. Specifically, the count of ’label’=1 has decreased over the years, which correlates with the declining probabilities mentioned earlier.

We balanced the dataset for each ’fyear’ value as a corrective measure, ensuring equal representation for both ’label’=0 and ’label’=1. Subsequently, we reran the DAG creation algorithms using ’bnlearn’.

Surprisingly, the DAG generation algorithm yielded empty plots (Figure 7.8). This indicates that no relationships were deemed sufficiently robust to be included in the diagram.

```
[bnlearn] >Computing best DAG using [hc]
[bnlearn] >Set scoring type at [bic]
[bnlearn] >Compute structure scores for model comparison (higher is better).
[bnlearn]> Nothing to plot because no edges are present between nodes.
```

Figure 7.8: DAG Results for Balanced Dataset

While this outcome initially appeared counterintuitive, a subsequent Kendall Rank Correlation on the balanced dataset still suggested potential statistically significant relationships. The powerful variable relationships in the dataset, where all relationships are relatively insignificant, are indicators of bias.

7.5.3 Assessing the Efficacy of Addressing Construction Bias

In this section, we present the efficacy of our proposed debiasing technique. We utilize three distinct datasets and employ the XGBoost [156] algorithm to evaluate their performance in predicting bankruptcy. Detailed descriptions of each dataset are provided in the ensuing subsections. Notably, all datasets were downsampled to contain 504 entries, the maximum row count permissible by our debiasing approach.

Downsampled Dataset

This dataset was derived by downsampling the primary dataset to 504 entries. During this process, it was essential to preserve the original distribution of the labels, i.e., label=1 and label=0. We first determined the proportions of these labels in the primary dataset. Subsequently, we sampled instances for each label based on these proportions to form the downsampled dataset.

Random Dataset

The creation of this dataset aimed to achieve an equal distribution of both labels, eliminating bias. Instances were randomly selected from the primary dataset. The number of samples chosen for each label was dictated by the lesser count of the two labels in the primary dataset. Consequently, this strategy assured equal representation for both labels. The final dataset size was precisely double the count of the less prevalent label from the primary dataset.

Debiased Dataset

This dataset was meticulously curated by selecting rows from the primary dataset that had fyear values of 2009, 2010, and 2011. For each of these years, an equivalent number of instances for both label=1 and label=0 were included. This strategy ensured a balanced representation of both labels for each year in the dataset. Consequently, any inherent bias in the primary dataset favouring a particular label for a specific year was mitigated.

7.5.4 Analysis of Results

Our primary objective is to evaluate the capability of predicting label 1 (bankrupt firms) relative to label 0. The performance metrics are presented in the table below.

From the table, it is evident that the ”Construction Bias-Free Dataset” offers the highest precision for label 1 at 95.71% and a commendable recall of 90.54%. This signifies its superior accuracy in predicting bankrupt firms.

To ascertain the bias favouring label 0 over label 1, we compare the recall metrics for both labels. The ”Original Dataset” has a recall disparity of 21.31% between the two labels, suggesting a pronounced bias towards label 0. Contrastingly, the ”Construction Bias-Free Dataset” narrows this difference to a mere 5.61%, indicating a substantial reduction in bias.

Metric	Dataset		
	Original	Without Bias	Without Construction Bias
Precision for label=0	97.06%	93.59%	91.46%
Recall for label=0	97.78%	93.59%	96.15%
F1-score for label=0	97.42%	93.59%	93.75%
Precision for label=1	81.25%	93.24%	95.71%
Recall for label=1	76.47%	93.24%	90.54%
F1-score for label=1	78.79%	93.24%	93.06%

Upon assessing the metrics across the datasets, we find that the model trained on the ”Bias-Free Dataset” emerges as the most effective in predicting label 1, while simultaneously minimizing bias. For precision of label 1, the ”Original Dataset” stands at 81.25%, the ”Random Sampling Dataset” at 93.24%, and the ”Bias-Free Dataset” leads at 95.71%. When predicting a firm as bankrupt (label 1), the model utilizing the ”Construction Bias-Free Dataset” is often correct.

For recall of label 1, the values are 76.47% for the ”Original Dataset”, 93.24% for the ”Random Sampling Dataset”, and 90.54% for the ”Bias-Free Dataset”. Though the ”Random Sampling Dataset” model slightly outperforms in recall for label 1, the difference is negligible. Given the paramount importance of precision (especially for accurately predicting label 1), the ”Construction Bias-Free Dataset” model remains exemplary due to its unparalleled precision.

Regarding bias assessment, the ”Original Dataset” exhibits a recall divergence of 21.31% between the two labels, highlighting a marked bias towards label 0. Conversely, the ”Bias-Free Dataset” trims this difference to just 5.61%, indicating a noteworthy bias reduction.

In summation, the ”Random Sampling Dataset” model showcases commendable performance; however, the model based on the ”Construction Bias-Free Dataset” strikes an optimal balance between precision for label 1 and bias mitigation, making it the most favourable model for predicting label 1 while accounting for bias.

Chapter 8

Conclusion

In financial stability, bankruptcy prediction models serve as invaluable tools that offer firms early warnings before encountering financial distress. This thesis delves into implementing causal machine learning algorithms and models rooted in Bayesian inference. We aimed to discern the factors prompting firms in Greece to file for bankruptcy.

This concluding chapter aims to encapsulate the salient findings, challenges, and innovative methodologies introduced throughout our research. We delve into the unique nature of our dataset, the pioneering approach we crafted in the face of scant literature, and the technical hurdles encountered in our computational endeavours. Furthermore, we touch upon this study's future directions, emphasizing the potential enhancements and broader implications for financial causality.

8.1 Fundamental Discoveries

Although our research initially focused on employing causal models to understand bankruptcy intricacies in Greek SMEs, it underwent significant evolution due to unforeseen data-related challenges. Despite these hurdles, our innovative and adaptive approach led to crucial insights. Our initial analyses unveiled pronounced and unexpected biases in the dataset.

Specifically, we introduced a novel term, "Construction Bias," for quantifying the bias present in the dataset through causal inference. This bias, identified during the dataset construction phase, obscured genuine relationships between variables and undermined the reliability of conclusions.

Diligently selecting instances from years free of overt biases and ensuring balanced rep-

resentation for both outcome labels mitigated this construction bias. This careful approach to data selection and analysis resulted in the training of models on a de-biased dataset, leading to a substantial enhancement in performance.

The models demonstrated a remarkable precision of 95.71% in predicting bankruptcies, highlighting the essential role of addressing and mitigating biases for achieving reliable and robust analytical outcomes. This research provides a deeper understanding of bankruptcy in Greek SMEs and emphasizes the importance of data handling and bias mitigation in empirical research.

8.2 Limitations

In our endeavour to uncover the causality behind Greek SMEs bankruptcy, we faced some critical issues, which can be categorized into two main categories.

8.2.1 Dataset Challenges

As showcased by both the Exploratory Data Analysis (EDA) and the Causal Analysis conducted, our model does not seem to be the optimal tool for decoding the bankruptcy factors specific to the Greek market. The dataset at our disposal has a limited number of instances from a consistent time frame, with a substantial amount of data being chronologically dispersed. A larger labelled dataset, and a more balanced one, would have enabled us to draw more expansive and consistent conclusions.

A direct result of those issues mentioned above was that the causal analysis in our research did not have a financial character, as there was significant bias in the dataset. This bias manifested in variable relationships that were much stronger than those among other variables. These relationships overshadowed the true significance of other variables, preventing probabilistic models from capturing the truly important ones and thus clouding our analysis. However, this established that one should be wary of very strong relationships in a dataset that might otherwise contain variables of lesser significance.

8.2.2 Causal Bankruptcy Prediction Challenges & Literature Gaps

Our data presented a one-of-a-kind challenge, primarily due to its nature and the extensive bias it contained. What made our quest even more daunting was the scarce literature and

implementations on Causal-Bayesian inference tailored for compromised financial data.

We extensively perused literature on Feature Importance and Selection Bias. Still, the nuanced discussions on Causal-Bayesian Structures truly illuminated our path. Drawing inspiration from these sources, we crafted a bespoke approach to navigate the unique challenges of our dataset, ensuring a robust causal analysis that was rare in its application and innovative in its methodology.

Issues with Utilised Libraries

During this thesis, we encountered significant technical challenges in implementing our experiments. While numerous libraries are available for causal and Bayesian inference, many are not widely adopted, resulting in limited online resources. For most of these libraries, we found scant information beyond the documentation provided by the authors. This lack of resources hindered our progress, as we had to experiment with multiple packages, only to discover that some did not function as anticipated.

Specifically, we dismissed the use of 'DoWhy' because it necessitates the provision of the DAG, which was incompatible with our approach, given our initial lack of knowledge regarding the causal mechanics of the data. 'CausalNex' was ruled out as it does not support continuous features, or in our scenario, features that take many values. 'CausalImpact' was not suitable since it is tailored exclusively for time-series data, and the 'BartPy' was excluded due to the absence of a Python version that supports classification.

The package we eventually settled on, 'Bnlearn', presented its own set of challenges. While it offers multiple algorithms for each step, the documentation lacks comparative insights or guidance on the appropriate contexts for each method. This necessitated extensive hands-on experimentation with our data to determine optimal performance and usage. CausalInference, though more user-friendly, also had limited online implementations. This meant we had to undergo a period of trial and error to fully grasp its functionalities.

8.3 Future Directions

In the realm of causal inference for bankruptcy prediction, particularly within the financial sector, there exists potential for further enhancements to improve both functionality and interpretability. While our study marks a significant stride towards robust forecasting perfor-

mance with model clarity, we acknowledge avenues for future refinement.

A primary enhancement to consider is the incorporation of regularization in the leaves of the BART trees implemented in Python. This would facilitate the application of the method to classification, not just regression. Currently, this feature is exclusive to the package for the R programming language.

For subsequent research, deploying our model on a more extensive dataset would be advantageous. This would provide insights into causal relationships within a different economic context and offer a comparative study of the economic dynamics between distinct eras or nations.

There is also an intriguing prospect of curating a dataset comprising multiple instances of the same companies across varied timestamps. Such a data set would pave the way for causal time series analysis, allowing tests for interventions and their associated Average Treatment Effects (ATE).

Lastly, an area we couldn't delve into during our research was the integration of Google Trends data, both for individual companies and the broader economy. Investigating the potential causal relationships between these trends and bankruptcy could yield enlightening insights.

Furthermore, it's worth noting that the realm of causality has been predominantly explored within the context of epidemiology. Its application in the financial and economic sectors remains in its infancy. This presents a vast landscape for research, as complete integration and understanding of causality in these sectors have not yet been achieved.

8.4 Concluding Remarks

This thesis aimed to investigate the causal inference of a bankruptcy prediction dataset of Greek SMEs using probabilistic techniques. While still in development, it is indeed one of the most promising domains for future research and is expected to flourish in the coming years. Even though our study found it impossible to clarify which variables affect the label the most financially, as we have seen in the literature, our innovative approach helped us better understand the dataset. In this context, a term called "Construction Bias" was devised to assist in the causal analysis by quantifying the inherent bias that arises during dataset construction. We were able to debias the dataset and introduce fairness, leading to improved

classification accuracy. This significantly enhanced the performance in the bankruptcy prediction problem. A significant achievement was the approximately 3% increase in precision for label 1, surpassing even the long-established random sampling debiasing method. While this is the main standpoint of our thesis, we also focused on elucidating the entire process of building a causal model. As it's not a common machine learning technique and is primarily used in epidemiology, we discussed our challenges.

Bibliography

[1] Emmanuel Vavalis Vasiliki Papadouli, Elias Houstis. Greek small medium sized enterprises bankruptcy dataset. pages 1–4, 2022.

[2] Spurious correlations. https://tylervigen.com/spurious-correlations.

[3] B Neal. Introduction to causal inference: From a machine learning perspective. course lect. *Notes*, 2020.

[4] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, 2004.

[5] Yin Shi and Xiaoni Li. An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, 15(2):114–127, 2019.

[6] P Ravi Kumar and Vadlamani Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques–a review. *European journal of operational research*, 180(1):1–28, 2007.

[7] Analyze search results. https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=be01d7e2eac0a695f6c144ab8df6de97&sot=a&sdt=a&sl=36&s=TITLE-ABS-KEY

[8] Satyam Kumar, Yelleti Vivek, Vadlamani Ravi, and Indranil Bose. Causal inference for banking finance and insurance a survey. *arXiv preprint arXiv:2307.16427*, 2023.

[9] Jennifer Hill and Elizabeth A Stuart. Causal inference: overview. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, pages 255–260, 2015.

[10] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[11] Bernhard Schölkopf. *Causality for Machine Learning*, page 765–804. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022.

[12] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[13] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

[14] Hernán MA and Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[15] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[16] Bayesian machine learning. https://www.datarobot.com/blog/bayesian-machine-learning/.

[17] The power of bayesian causal inference: A comparative of librarys to reveal hidden causality in your dataset. https://towardsdatascience.com/the-power-of-bayesian-causal-inference-a-comparative-analysis-of-libraries-to-reveal-hidden-d91e8306e25e.

[18] Fan Li, Peng Ding, and Fabrizia Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153, 2023.

[19] A beginner's guide to bayesian additive regression trees. https://analyticsindiamag.com/a-beginners-guide-to-bayesian-additive-regression-trees/.

[20] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[21] Implementation of bayesian regression. https://www.geeksforgeeks.org/implementation-of-bayesian-regression/.

[22] Bret Michael Zeldow. *Bayesian Nonparametric Methods for Causal Inference and Prediction*. University of Pennsylvania, 2017.

[23] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

[24] J. T. Kent K. V. Mardia and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[25] David Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2012.

[26] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. CRC press, 2021.

[27] understanding bayesian networks with examples in r. https://www.bnlearn.com/about/teaching/slides-bnshort.pdf.

[28] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. Inferring causal impact using bayesian structural time-series models. 2015.

[29] bnlearn - an r package for bayesian network learning and inference. https://www.bnlearn.com/.

[30] Causalpy - causal inference for quasi-experiments. https://causalpy.readthedocs.io/en/latest/#installation.

[31] Causalpy - causal inference for quasi-experiments. https://www.pymc-labs.io/blog-posts/causalpy-a-new-package-for-bayesian-causal-inference-for-quasi-experiments/.

[32] Introduction. https://causalnex.readthedocs.io/en/latest/01_introduction/01_introduction.html.

[33] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[34] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

[35] Causalimpact. http://google.github.io/CausalImpact/CausalImpact.html.

[36] pgmpy. https://pgmpy.org/.

[37] I Altman Edward, G Haldeman Robert, and Paul Narayanan. Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 10:29–54, 1977.

[38] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131, 1980.

[39] J Richard Dietrich and Robert S Kaplan. Empirical analysis of the commercial loan classification decision. *Accounting Review*, pages 18–38, 1982.

[40] Jarrod W Wilcox. A prediction of business failure using accounting data. *Journal of accounting research*, pages 163–179, 1973.

[41] Mark E Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, pages 59–82, 1984.

[42] James Kolari, Dennis Glennon, Hwan Shin, and Michele Caputo. Predicting large us commercial bank failures. *Journal of Economics and Business*, 54(4):361–387, 2002.

[43] Serpil Canbas, Altan Cabuk, and Suleyman Bilgin Kilic. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The turkish case. *European journal of operational research*, 166(2):528–546, 2005.

[44] Kar Yan Tam. Neural network models and the prediction of bank bankruptcy. *Omega*, 19(5):429–445, 1991.

[45] Kar Yan Tam and Melody Kiang. Predicting bank failures: A neural network approach. *Applied Artificial Intelligence an International Journal*, 4(4):265–282, 1990.

[46] Linda M Salchenberger, E Mine Cinar, and Nicholas A Lash. Neural networks: A new tool for predicting thrift failures. *Decision Sciences*, 23(4):899–916, 1992.

[47] Ramesh Sharda and Rick L Wilson. Performance comparison issues in neural network experiments for classification problems. In *[1993] Proceedings of the Twenty-sixth Hawaii International Conference on System Sciences*, volume 4, pages 649–657. IEEE, 1993.

[48] Edward I Altman, Giancarlo Marco, and Franco Varetto. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience). *Journal of banking & finance*, 18(3):505–529, 1994.

[49] Rick L Wilson and Ramesh Sharda. Bankruptcy prediction using neural networks. *Decision support systems*, 11(5):545–557, 1994.

[50] Junsei Tsukuda and Shin-ichi Baba. Predicting japanese corporate bankruptcy in terms of financial data using neural network. *Computers & Industrial Engineering*, 27(1-4):445–448, 1994.

[51] Selwyn Piramuthu, Harish Ragavan, and Michael J Shaw. Using feature construction to improve the performance of neural networks. *Management Science*, 44(3):416–430, 1998.

[52] Guoqiang Zhang, Michael Y Hu, B Eddy Patuwo, and Daniel C Indro. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research*, 116(1):16–32, 1999.

[53] Amir F Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, 12(4):929–935, 2001.

[54] Philip Swicegood and Jeffrey A Clark. Off-site monitoring systems for predicting bank underperformance: a comparison of neural networks, discriminant analysis, and professional human judgment. *Intelligent Systems in Accounting, Finance & Management*, 10(3):169–186, 2001.

[55] Kidong Lee, David Booth, and Pervaiz Alam. A comparison of supervised and unsupervised neural networks in predicting bankruptcy of korean firms. *Expert Systems with Applications*, 29(1):1–16, 2005.

[56] Kun Chang Lee, Ingoo Han, and Youngsig Kwon. Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18(1):63–72, 1996.

[57] Carlos Serrano-Cinca. Self organizing neural networks for financial diagnosis. *Decision support systems*, 17(3):227–238, 1996.

[58] Kimmo Kiviluoto. Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21(1-3):191–201, 1998.

[59] Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12(4):936–947, 2001.

[60] R Christopher Lacher, Pamela K Coats, Shanker C Sharma, and L Franklin Fant. A neural network for classifying the financial health of a firm. *European Journal of Operational Research*, 85(1):53–65, 1995.

[61] ZR Yang, Marjorie B Platt, and Harlan D Platt. Probabilistic neural networks in bankruptcy prediction. *Journal of business research*, 44(2):67–74, 1999.

[62] Jinwoo Baek and Sungzoon Cho. Bankruptcy prediction for credit risk using an auto-associative neural network in korean firms. In *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.*, pages 25–29. IEEE, 2003.

[63] Stephanie Mattox Bryant. *A case-based reasoning approach to bankruptcy prediction modeling*. Louisiana State University and Agricultural & Mechanical College, 1996.

[64] Hongkyu Jo, Ingoo Han, and Hoonyoung Lee. Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert Systems with Applications*, 13(2):97–108, 1997.

[65] Cheol-Soo Park and Ingoo Han. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*, 23(3):255–264, 2002.

[66] Angela YN Yip. Predicting business failure with a case-based reasoning approach. In *Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference, KES 2004, Wellington, New Zealand, September 20-25, 2004, Proceedings, Part III 8*, pages 665–671. Springer, 2004.

[67] M Laurentius Marais, James M Patell, and Mark A Wolfson. The experimental design of classification models: An application of recursive partitioning and bootstrapping to

commercial bank loan classifications. *Journal of accounting Research*, pages 87–114, 1984.

[68] Halina Frydman, Edward I Altman, and Duen-Li Kao. Introducing recursive partitioning for financial classification: the case of financial distress. *The journal of finance*, 40(1):269–291, 1985.

[69] William J Banks and Prakash L Abad. On the performance of linear programming heuristics applied on a quadratic transformation in the classification problem. *European Journal of Operational Research*, 72(1):23–28, 1994.

[70] Kim Fung Lam and Jane W Moy. Combining discriminant methods in solving classification problems in two-group discriminant analysis. *European Journal of Operational Research*, 138(2):294–301, 2002.

[71] Anja Cielen, Ludo Peeters, and Koen Vanhoof. Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, 154(2):526–532, 2004.

[72] Salvatore Greco, Benedetto Matarazzo, and Roman Slowinski. A new rough set approach to multicriteria and multiattribute classification. In *Rough Sets and Current Trends in Computing: First International Conference, RSCTC'98 Warsaw, Poland, June 22–26, 1998 Proceedings 1*, pages 60–67. Springer, 1998.

[73] Augustinos I Dimitras, Roman Slowinski, Robert Susmaga, and Constantin Zopounidis. Business failure prediction using rough sets. *European Journal of operational research*, 114(2):263–280, 1999.

[74] Thomas E Mckee. Developing a bankruptcy prediction model via rough sets theory. *Intelligent Systems in Accounting, Finance & Management*, 9(3):159–173, 2000.

[75] Spanos Michael, Dounias Georgios, Matsatsinis Nikolaos, and Zopounidis Constantin. A fuzzy knowledge-based decision aiding method for the assessment of financial risks: the case of corporate bankruptcy prediction. In *European Symposium on Intelligent Techniques, Crete, Greece*. Citeseer, 1999.

[76] Jae H Min and Young-Chan Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 28(4):603–614, 2005.

[77] Young U Ryu and Wei T Yue. Firm bankruptcy prediction: experimental comparison of isotonic separation and other classification approaches. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35(5):727–737, 2005.

[78] Barbro Back, Teija Laitinen, and Kaisa Sere. Neural networks and genetic algorithms for bankruptcy predictions. *Expert systems with applications*, 11(4):407–413, 1996.

[79] Marian B Gorzalczany and Zdzislaw Piasta. Neuro-fuzzy approach versus rough-set inspired methodology for intelligent decision support. *Information Sciences*, 120(1-4):45–68, 1999.

[80] Byeong Seok Ahn, SS Cho, and CY Kim. The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert systems with applications*, 18(2):65–74, 2000.

[81] Thomas E McKee and Terje Lensberg. Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European journal of operational research*, 138(2):436–451, 2002.

[82] Sofie Balcaen and Hubert Ooghe. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, 38(1):63–93, 2006.

[83] Edward I Altman and Edith Hotchkiss. *Corporate financial distress and bankruptcy*, volume 1998. New York: John Wiley & Sons, 1993.

[84] Zhi Xiao, Xianglei Yang, Ying Pang, and Xin Dang. The prediction for listed companies' financial distress by using multiple prediction methods with rough set and dempster–shafer evidence theory. *Knowledge-Based Systems*, 26:196–206, 2012.

[85] Lu Wang and Chong Wu. Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. *Knowledge-Based Systems*, 121:99–110, 2017.

[86] Hui Li and Jie Sun. Forecasting business failure in china using case-based reasoning with hybrid case respresentation. *Journal of Forecasting*, 29(5):486–501, 2010.

[87] Hui Li and Jie Sun. Predicting business failure using an rsf-based case-based reasoning ensemble forecasting method. *Journal of Forecasting*, 32(2):180–192, 2013.

[88] Soo Y Kim. Prediction of hotel bankruptcy using support vector machine, artificial neural network, logistic regression, and multivariate discriminant analysis. *The Service Industries Journal*, 31(3):441–468, 2011.

[89] Vasiliki Papadouli. Performance evaluation of statistical and machine learning Models for bankruptcy prediction of greek smes. Διπλωματική εργασία, Πανεπιστήμιο Θεσσαλίας, 2022.

[90] E. Houstis V. Papadouli and M. Vavalis. "bankruptcy prediction of greek smes using imbalance data". *Journal of Advances in Information Technology*, 2023.

[91] Elias Houstis Vassiliki Papadouli and Manolis Vavalis. Bankruptcy prediction of greek smes using imbalance data. 2018.

[92] Claudiu Clement et al. Machine learning in bankruptcy prediction–a review. *Journal of Public Administration, Finance and Law*, (17):178–196, 2020.

[93] Wei Li, Wolfgang Karl Härdle, and Stefan Lessmann. A data-driven case-based reasoning in bankruptcy prediction. *arXiv preprint arXiv:2211.00921*, 2022.

[94] Yuri Zelenkov. Bankruptcy prediction using survival analysis technique. In *2020 IEEE 22nd Conference on Business Informatics (CBI)*, volume 2, pages 141–149. IEEE, 2020.

[95] Elena Fedorova, Svetlana Ledyaeva, Pavel Drogovoz, and Alexandr Nevredinov. Economic policy uncertainty and bankruptcy filings. *International Review of Financial Analysis*, 82:102174, 2022.

[96] Mr Andrew J Tiffin. *Machine learning and causality: The impact of financial crises on growth*. International Monetary Fund, 2019.

[97] Wanderson Rocha Bittencourt and Pedro HM Albuquerque. Evaluating company bankruptcies using causal forests. *Revista Contabilidade & Finanças*, 31:542–559, 2020.

[98] Lili Sun and Prakash P Shenoy. Using bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, 180(2):738–753, 2007.

[99] Arezoo Aghaie and Ali Saeedi. Using bayesian networks for bankruptcy prediction: Empirical evidence from iranian companies. In *2009 International Conference on Information Management and Engineering*, pages 450–455. IEEE, 2009.

[100] Jeffrey Traczynski. Firm default prediction: A bayesian model-averaging approach. *Journal of Financial and Quantitative Analysis*, 52(3):1211–1245, 2017.

[101] Yi Cao, Xiaoquan Liu, Jia Zhai, and Shan Hua. A two-stage bayesian network model for corporate bankruptcy prediction. *International Journal of Finance & Economics*, 27(1):455–472, 2022.

[102] Amir Mukeri, Habibullah Shaikh, and Dr DP Gaikwad. Financial data analysis using expert bayesian framework for bankruptcy prediction. *arXiv preprint arXiv:2010.13892*, 2020.

[103] Nektarios A Michail. What if they had not gone negative? a counterfactual assessment of the impact from negative interest rates. *Oxford Bulletin of Economics and Statistics*, 81(1):1–19, 2019.

[104] OM Kolodiziev, IM Chmutova, and Vitaliy Lesik. Use of causal analysis to improve the monitoring of the banking system stability. 2018.

[105] Mikhail Stolbov and Maria Shchepeleva. Systemic risk, economic policy uncertainty and firm bankruptcies: Evidence from multivariate causal inference. *Research in International Business and Finance*, 52:101172, 2020.

[106] Xiao-Li Gong, Xi-Hua Liu, Xiong Xiong, and Wei Zhang. Financial systemic risk measurement based on causal network connectedness analysis. *International Review of Economics & Finance*, 64:290–307, 2019.

[107] Carter Davis. The elasticity of quantitative investment. *arXiv preprint arXiv:2303.14533*, 2023.

[108] Ployplearn Ravivanpong, Till Riedel, and Pascal Stock. Towards extracting causal graph structures from tradedata and smart financial portfolio risk management. In *EDBT/ICDT Workshops*, 2022.

[109] Katerina Rigana, Ernst-Jan Camiel Wit, and Samantha Cook. Using network-based causal inference to detect the sources of contagion in the currency market. *arXiv preprint arXiv:2112.13127*, 2021.

[110] Fani Tsapeli, Mirco Musolesi, and Peter Tino. Non-parametric causality detection: An application to social media and financial data. *Physica A: Statistical Mechanics and Its Applications*, 483:139–155, 2017.

[111] Carlos Castro-Iragorri. Does the market model provide a good counterfactual for event studies in finance? *Financial Markets and Portfolio Management*, 33:71–91, 2019.

[112] Samantha Kleinberg. Causal inference with rare events in large-scale time-series data. In *IJCAI*, pages 1444–1450, 2013.

[113] Samantha Kleinberg, Petter N Kolm, and Bud Mishra. Investigating causal relationships in stock returns with temporal logic based methods. *arXiv preprint arXiv:1006.1791*, 2010.

[114] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63:3041–3085, 2021.

[115] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

[116] Yoichi Chikahara and Akinori Fujino. Causal inference in time series via supervised learning. In *IJCAI*, pages 2042–2048, 2018.

[117] Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, pages 1917–1925. PMLR, 2015.

[118] David Hason Rudd, Huan Huo, and Guandong Xu. Causal analysis of customer churn using deep learning. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 319–324. IEEE, 2021.

[119] Gerald Fahner. Estimating causal effects of credit decisions. *International Journal of Forecasting*, 28(1):248–260, 2012.

[120] Leo Guelman and Montserrat Guillén. A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, 41(2):387–396, 2014.

[121] Eric Jacquier and Nicholas Polson. Bayesian methods in finance. 2011.

[122] Andrew D Sanford and Imad A Moosa. A bayesian network structure for operational risk modelling in structured finance operations. *Journal of the Operational Research Society*, 63:431–444, 2012.

[123] Gelin Gao, Bud Mishra, and Daniele Ramazzotti. Efficient simulation of financial stress testing scenarios with suppes-bayes causal networks. *Procedia Computer Science*, 108:272–284, 2017.

[124] Stavros Stavroglou, Athanasios Pantelous, Kimmo Soramaki, and Konstantin Zuev. Causality networks of financial assets. *The Journal of Network Theory in Finance*, 3(2):17–67, 2017.

[125] Michael Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1997):20110613, 2013.

[126] C Gustav Lundberg and Dean Elliott Frost. Counterfactuals in financial decision making. *Acta Psychologica*, 79(3):227–244, 1992.

[127] Ekatarina Svetlova. ” do i see what the market does not see?”: Counterfactual thinking in financial markets. *Historical Social Research/Historische Sozialforschung*, pages 147–157, 2009.

[128] Jingwei Gan, Shinan Zhang, Chi Zhang, and Andy Li. Automated counterfactual generation in financial model risk management. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4064–4068. IEEE, 2021.

[129] Dan Wang, Zhi Chen, Ionuț Florescu, and Bingyang Wen. A sparsity algorithm for finding optimal counterfactual explanations: Application to corporate credit rating. *Research in International Business and Finance*, 64:101869, 2023.

[130] Linyi Yang, Eoin M Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. Generating plausible counterfactual explanations for deep transformers in financial text classification. *arXiv preprint arXiv:2010.12512*, 2020.

[131] Rory McGrath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lécué. Interpretable credit application predictions with counterfactual explanations. corr (abs/1811.05245)(2018), 1811.

[132] Xolani Dastile, Turgay Celik, and Hans Vandierendonck. Model-agnostic counterfactual explanations in credit scoring. *IEEE Access*, 10:69543–69554, 2022.

[133] Andreas C Bueff, Mateusz Cytryński, Raffaella Calabrese, Matthew Jones, John Roberts, Jonathon Moore, and Iain Brown. Machine learning interpretability for a stress scenario generation in credit scoring based on counterfactuals. *Expert Systems with Applications*, 202:117271, 2022.

[134] Evangelos Sfakianakis. Bankruptcy prediction model for listed companies in greece. *Investment Management and Financial Innovations*, 18:166–180, 05 2021.

[135] Jason Brownlee. *Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end*. Machine Learning Mastery, 2016.

[136] Jason Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[137] Wikipedia, the free encyclopedia, financial market. https://en.wikipedia.org/wiki/Financial_market, 2023.

[138] Jason Brownlee. How to choose a feature selection method for machine learning. *Machine Learning Mastery*, 10, 2019.

[139] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[140] Pearson, spearman, and kendall-tau correlations: What are the differences? https://python.plainenglish.io/pearson-spearman-and-kendall-tau-correlations-what-are-the-differences-d0a3963b4c94, 2023.

[141] Correlation (pearson, kendall, spearman). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/.

[142] Pearson correlation coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[143] Spearman's rank correlation coefficient. https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[144] Kendall rank correlation coefficient. https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient.

[145] Mutual information. https://en.wikipedia.org/wiki/Mutual_information.

[146] Select features for machine learning model with mutual information. https://towardsdatascience.com/select-features-for-machine-learning-model-with-mutual-information-534fe387d5c8.

[147] Decision tree. https://en.wikipedia.org/wiki/Decision_tree.

[148] Random forest. https://en.wikipedia.org/wiki/Decision_tree.

[149] Explaining feature importance by example of a random forest. https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e.

[150] Jason Brownlee. A tour of machine learning algorithms.

[151] Jason Brownlee. Logistic regression for machine learning. *Machine Learning Mastery*, 1, 2016.

[152] Jason Brownlee. Difference between classification and regression in machine learning. *Machine Learning Mastery*, 25:985–1, 2017.

[153] Takio Kurita. Principal component analysis (pca). *Computer Vision: A Reference Guide*, pages 1–4, 2019.

[154] Enrique F Schisterman, Neil J Perkins, Sunni L Mumford, Katherine A Ahrens, and Emily M Mitchell. Collinearity and causal diagrams–a lesson on the importance of model specification. *Epidemiology (Cambridge, Mass.)*, 28(1):47, 2017.

[155] Causation. https://erdogant.github.io/bnlearn/pages/html/Structure

[156] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.