UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Human Resources Analytics

Diploma Thesis

ELPIDA MYRTO KOUTSONI

Supervisor: Hariklia Tsalapata

Volos 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Human Resources Analytics

Diploma Thesis

ELPIDA MYRTO KOUTSONI

Supervisor: Hariklia Tsalapata

Volos 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση Δεδομένων στη Διαχείριση Ανθρωπίνου Δυναμικού

Διπλωματική Εργασία

ΕΛΠΙΔΑ ΜΥΡΤΩ ΚΟΥΤΣΩΝΗ

Επιβλέπων/πουσα: Χαρίκλεια Τσαλαπάτα

Βόλος 2022

Approved by the Examination Committee:

Supervisor **Hariklia Tsalapata**
Laboratory Teaching Staff, Department of Electrical and Computer Engineering, University of Thessaly

Member **Georgios Stamoulis**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Dimitrios Rafailidis**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

Acknowledgements

I want to express my gratitude to Professor Hariklia Tsalapata and the entire committee for overseeing my thesis.

Also I would like to deeply thank my family and the people that were close to me and supported me through this rewarding journey all of these years.

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 declarant

ELPIDA MYRTO KOUTSONI

Abstract

Human Resources Management (HRM) is important for every company, since it goes beyond finances and plays a major role in the market. Today, artificial intelligence (AI) offers a unique opportunity to improve HRM by employing AI to retrieve valuable insights from large data sets. This involves a thorough examination of HRM Analytics and an in-depth study of data from a well-known business entity.

As we progress in this project, our main focus is to evaluate advanced machine learning and deep learning techniques. These complex and sophisticated algorithms are extensively evaluated for their ability to forecast employee turnover accurately. Our study brings us to a clear conclusion: the deep learning methods, such as the stacked CNN-LSTM model, seem to be the most efficient solution for such a problem, as they combine high accuracy with minimal overhead. Nevertheless, some ensemble solutions, such as XGBoost, seem to have equally remarkable predictive performance as well.

These findings outline the advantages of each method, highlighting the best ones for resolving HR Analytics challenges. Their exceptional performance in this use case shows their ability to grasp and forecast employee behavior, aligning perfectly with the complexities of HRM. These state-of-the-art algorithms emerge as valuable allies for organizations navigating the intricacies of human capital management, enabling informed decisions that benefit the entire business community

Περίληψη

Η Διαχείριση Ανθρώπινου Δυναμικού είναι σημαντική για κάθε εταιρεία, καθώς υπερβαίνει τα οικονομικά και διαδραματίζει σημαντικό ρόλο στην αγορά. Σήμερα, η τεχνητή νοημοσύνη (ΤΝ) προσφέρει μια μοναδική ευκαιρία για τη βελτίωση της διαχείρισης ανθρώπινου δυναμικού, χρησιμοποιώντας την ΤΝ για την ανάκτηση πολύτιμων πληροφοριών από μεγάλα σύνολα δεδομένων. Αυτό περιλαμβάνει μια ενδελεχή εξέταση χρησιμοποιώντας HRM Analytics και μια εις βάθος μελέτη δεδομένων από μια γνωστή επιχειρηματική οντότητα.

Καθώς προχωράμε σε αυτή την εργασία, η κύρια εστίασή μας είναι να αξιολογήσουμε προηγμένες τεχνικές μηχανικής μάθησης και βαθιάς μάθησης. Αυτοί οι σύνθετοι και εξελιγμένοι αλγόριθμοι αξιολογούνται εκτενώς ως προς την ικανότητά τους να προβλέπουν με ακρίβεια τον κύκλο εργασιών των εργαζομένων. Η μελέτη μας μας οδηγεί σε ένα σαφές συμπέρασμα: οι μέθοδοι βαθιάς μάθησης, όπως το μοντέλο CNN-LSTM, φαίνεται να είναι η πιο αποτελεσματική λύση για ένα τέτοιο πρόβλημα, καθώς συνδυάζουν υψηλή ακρίβεια με ελάχιστο κόστος. Παρ' όλα αυτά, ορισμένες λύσεις ensemble, όπως το XGBoost μοντέλο, φαίνεται να έχουν επίσης εξίσου αξιοσημείωτες προγνωστικές επιδόσεις.

Τα ευρήματα αυτά σκιαγραφούν τα πλεονεκτήματα κάθε μεθόδου, αναδεικνύοντας τις καλύτερες για την επίλυση των προκλήσεων της HR Analytics. Οι εξαιρετικές επιδόσεις τους σε αυτή την περίπτωση χρήσης δείχνουν την ικανότητά τους να αντιλαμβάνονται και να προβλέπουν τη συμπεριφορά των εργαζομένων, ευθυγραμμιζόμενες απόλυτα με την πολυπλοκότητα της διαχείρισης ανθρώπινου δυναμικού. Αυτοί οι αλγόριθμοι τελευταίας τεχνολογίας αναδεικνύονται σε πολύτιμους συμμάχους για τους οργανισμούς που περιηγούνται στις περιπλοκές της διαχείρισης του ανθρώπινου κεφαλαίου, επιτρέποντας τη λήψη τεκμηριωμένων αποφάσεων που ωφελούν ολόκληρη την επιχειρηματική κοινότητα

Table of contents

List of figures

List of tables

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
GBM	Gradient Boosting Machine
LSTM	Long Short Term Memory
ML	Machine Learning
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic

Chapter 1

Introduction

The HR department oversees the area of human resources management (HRM), which is a fundamental component of every business. It has an impact on levels of market positioning and reputation that go well beyond simple financial transactions and earnings. The rise of artificial intelligence (AI) in the modern world has brought with it an unheard-of opportunity to inject intelligence and foresight into HRM operations. Utilizing AI, businesses may gain priceless insights from massive amounts of data, improving decision-making procedures.

The HR manager's position is crucial, comprising a range of duties including monitoring the recruiting process, managing employee benefits, keeping track of personnel data, and helping with various elements of recruitment. This position acts as a crucial conduit between a company's management and its staff, bridging the gap between operational execution and strategic vision.

At the same time, the field of artificial intelligence (AI), in particular Machine Learning (ML), provides a dynamic framework for utilizing data-driven insights to resolve complex challenges. Applications for machine learning are numerous and revolutionary, ranging from driverless vehicles and fraud detection to real-time mapping systems and virtual personal assistants. These technological developments not only shed light on consumer behavior trends, but also provide information about organizational patterns.

Human Resources Analytics (HR Analytics), a paradigm that promises to boost business income while optimizing Capital Expenditures (CapEX) and Operational Expenditures (OpEX), is the result of the convergence of HR and ML. This combination is in line with the current requirement to examine personnel results, growth trajectories, performance measures, and turnover rates holistically. Through the careful application of ML approaches, attrition,

a difficult and resource-intensive process, can be foreseen and managed more effectively.

The thesis starts making important contributions in this context. It includes a thorough analysis of pertinent research, followed by practical assessments of several ML and Deep Learning (DL) techniques using actual data. The focus is on a crucial ”target” characteristic that determines whether an employee stays with the organization or leaves. Python is used to implement ML models, utilizing the keras-tensorflow API for model development and pandas and numpy for thorough statistical analysis.

1.1 Subject of the thesis

In this thesis, we present our approach in evaluating machine learning and deep learning methods for predicting the possibility of an employee to stay or leave the company. The goal was to find the most efficient models in terms of high accuracy and minimal overhead. Our method was derived from the comparison of four machine learning classification algorithms, namely Random Forest, CatBoost, XGBoost, LightGBM and five deep learning classification models, namely Long short-term memory (LSTM), Gated recurrent unit (GRU), Feed-forward Neural Network (FNN), Convolutional Neural Network (CNN) and a stacked CNN-LSTM model. The methods predict the indend of the employee to stay or leave the company. We evaluate the results using the K-fold cross-validation technique together with ROC curve and AUC score metrics.

1.2 Organization of the thesis

The thesis is organized in the following manner. Chapter 2 gives a background on the problem statement and the current related works trying to tackle it. Chapter 3 shows our methodology, including data collection, management and the respective ML and DL algorithms. Chapter 4 presents the evaluation of the results and Chapter 5 provides a conclusion of the thesis and some thoughts about future work.

Chapter 2

Background

2.1 Problem Statement

Our objective is to anticipate whether a client intends to remain with our service or discontinue their relationship. In other words, whether the client is satisfied and wants to continue or dissatisfied and wants to leave the company.

2.1.1 The Significance

Understanding and predicting client retention is crucial for every business's sustainability and growth. It allows them to proactively address issues, tailor their services, and ultimately enhance customer satisfaction and loyalty. Thus, a healthy environmest is created where people are comfortable working being able to work more effectively, resulting in a win-win situation both for the company and the employees.

2.1.2 The Approach

To achieve this, we employ HR Analytics, harnessing the power of supervised learning on a carefully curated dataset. Supervised learning emerges as a robust and tried-and-true approach, renowned for its effectiveness and reliability in making accurate predictions. This method not only provides valuable insights into client behavior but also ensures a secure and dependable decision-making process. By leveraging the capabilities of HR Analytics and supervised learning, we are poised to make informed choices that will benefit both our clients and our organization.

2.2 Related Work

Following a thorough review of the existing literature, we highlight four noteworthy contributions.

2.2.1 Introducing HR Analytics with Machine Learning

In this book written by Christopher M.Rosett and Austin Hagerty is vividly depicted, with many real-life instances, how ML has ”invaded” in the HR domain and enhanced the process of managing employees and reaching business goals. In the first chapters the authors introduce basic concepts about HR strategies and Analytics. To adduce an example, *HR Analytics Ikigai* is a Japanese philosophy that aims to answer the question ”What are the requirements for good HR Analytics?”. According to the authors a special blend of knowledge from these four key domains is needed to make decisions based on behavioral data:

- Computing.
- Human Behaviour.
- Statistics and Research Methods.
- Business astuteness.

These four specialties, each having a significant value, are combined in order to produce equilibrium in HR Analytics. Fig. 2.1 illustrates an HR Analytics Ikigai, depicted in a four-way Venn diagram.

Another noteworthy point of the book is the analysis of two fundamental reasoning concepts the ”Deductive” and the ”Inductive”[4].

Initially, the *Deductive reasoning* is a kind of logic which through four key actions aims to develop a theory. The four key actions are the following:

- Make observations.
- Make hypotheses about the observations.
- Scrutinize the hypotheses through predictions and evaluations.
- Repeat numerous times in many different but associated experiments.

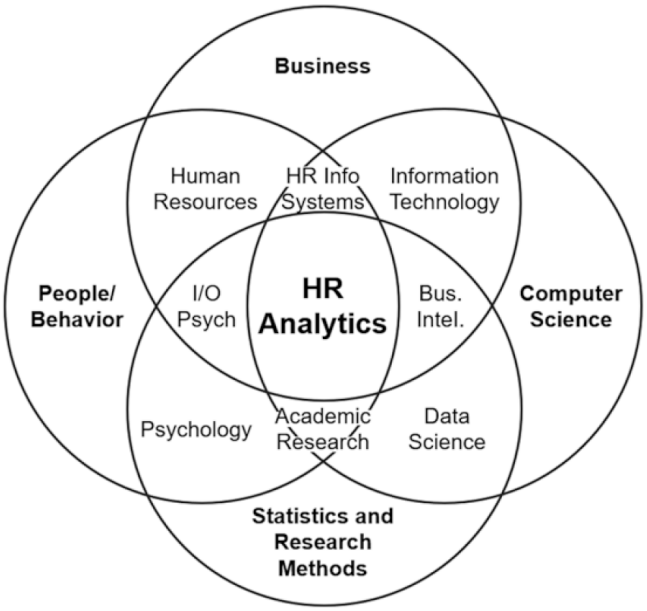

Figure 2.1: HR Analytics Ikigai

Scientists also know that the theories' contents are valid because deductive reasoning is used to develop them. They began with a large number of observations and then narrowed the range of potential causes to include the fewest assumptions feasible, leaving them with as nearly unbiased facts as they were able to produce.

Machine Learning introduces now, a new reasoning way, the *Inductive reasoning*. An inductive approach is a bottom-up strategy. It begins with the particular data or patterns and generalises from there to draw conclusions that fit what is visible. The proverb "construct the bridge while you cross it" applies well to inductive reasoning. A researcher can start down a road using inductive reasoning, then course-correct as they progress.

The main difference between these two concepts is that Inductive reasoning attempts to develop a theory, whereas Deductive reasoning aims to test an existing theory. In real-life business problems timelines are constrict. It is of high importance that decisions are made expeditiously and that is why Deductive method is not always a suitable solution. Using the data at our disposal to deduce patterns and then testing them out is frequently the most plausible strategy when dealing with extremely complicated or little understood systems like the economy or the human mind. Inductive reasoning, as it is depicted in Fig. 2.2, provides not only pattern recognition but also faster processing and both of them offer the company a competitive advantage.

Subsequently, a detailed introduction of all the fundamental ML methods is taking place.

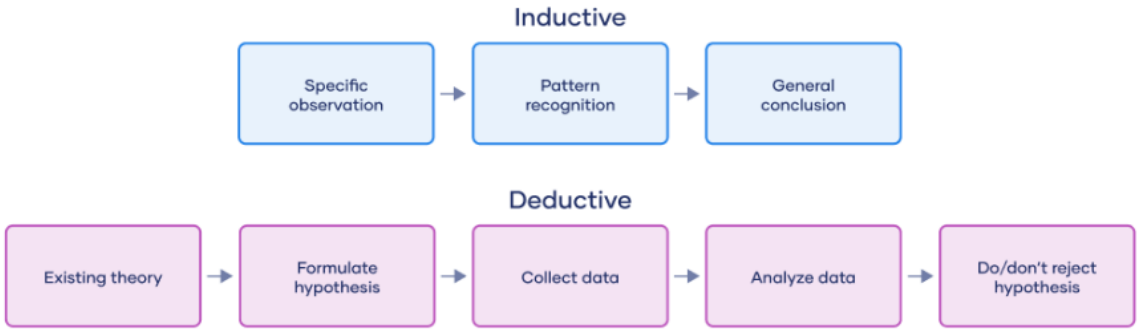

Figure 2.2: Deduction vs. Induction reasoning

In specific, supervised learning algorithms are presented, such as Linear/ Non Linear/ Polynomial/ Logistic Regression. Moreover, K-Nearest Neighbors, Support Vector Machines (SVMs), Random Forests and Decision Trees. Importantly, advanced supervised methods are included, such as ensembled techniques and Neural Networks (NNs). Furthermore, unsupervised learning methods are discussed, including Kmeans, Hierarchical Clustering and Latent Variable Models, while the authors provide a brief theoretical overview of Reinforcement Learning (RL).

Finally, through some strikingly historical examples, the writers demonstrate the value of the proper use of Machine Learning in state-of-art projects. It is essential to notice that the proper usage of ML is defined in three axes.

- Bias.
- Authority.
- Effectiveness.

An indicative example of the ”Effectiveness ax” is the case of Thalidomide. After World War II a company, located in West German, named Chemie Grünenthal was searching and working in order to full fill the demand for further antibiotics. While researching, they found a painkiller medicine, called Thalidomide, that had also anti-emetic effects. In 1957, the drug was strongly advertised, for soothing morning nausea of pregnant women. Over the course of the drug’s marketing campaign, teratogenic birth abnormalities were detected in thousands of newborns. This led to the gradual removal of the drug from the market. The inconsiderate action, of selling this substance to pregnant women only without the vital extensive background research, was not recognized, from the German government as negligent homicide or injury.

The excuse of the company was that they were aware of the ”nothing-crosses-the-placental-barrier”; as such the drug could not influence the infant. However, it is commonly known, since 1957, that alcohol penetrates through the placental barrier, which provides strong indication that any other drug could, potentially, cross it. This case shows a harmful fallacy presented as proof of the mindset ”it just works”. A model or algorithm does not automatically qualify as ”excellent” just because it ”works” in the sense of producing a desirable result, such as easing temporary symptoms like financial limitations. It is important to distinguish between ”it works” and ”it is good,” as doing so would imply that additional critical analysis is not necessary.

2.2.2 A Comparison of Neural Networks and Support Vector Machines Algorithm Performance in Human Resource Analytics

In this paper the authors, namely Hannes Draxl and Ryan Nazareth, highlight the fact that attrition is not only a time-consuming but also an expensive process. Considering that there are a plethora of factors which play a significant role into the decision of an employee to quit the company, the writers attempt to predict this decision, in advance, utilizing robust ML algorithms. Specifically, this study’s primary goal is to evaluate and compare the effectiveness of the Multi-Layer Perceptron (MLP) algorithm and in contrast to Support Vector Machine (SVM) classifier.

The target metric is whether an employee is going to quit. The name of this variable is ”Left” and is divided in two categories:

- Stay.
- Leave.

After exploring the columns of the dataset and indicating the target variable the authors implement Explanatory Analysis. In particular, they use *Heat map* in order to visualize the correlation between the features. They observe that the ”Left” variable (target variable) has a strong correlation (-0.5) with the ”satisfaction_level” feature. So, we conclude that the decision of quitting is affected by their degree of contentedness. Another interesting way of plotting the correlation is used called the *violin plot*. From this plot, it is obvious that the salary feature in combination with the promotion_last_5years can increase the prediction power of the model.

For the training and testing process they split the data into 70% training and 30% test. Then, they used 5-fold stratified cross validation with embedded grid search for hyper parameter tuning. The models with the best HP were retrained using the entire training set after grid search. In order to determine how many samples the models can correctly categorize as being on the left out of all samples that correspond to ”left,” they relied on a confusion matrix and in particular the recall score (the true positive rate).

In conclusion, SVM and MLP performed well in forecasting whether or not a customer will quit the organization, and as a result, is of great importance for human resource departments. SVM and MLP demonstrated both their strengths and shortcomings during the analysis; they are both strong algorithms but take a lot of computational power and time to tweak and assess. Results showed that SVM performed better than MLP in terms of test accuracy overall (F1 score), while MLP had a slightly greater recall rate than SVM, which is consistent with their original premise.

2.2.3 IBM HR Analytics Employee Attrition

In this research, that was initiated by the IBM company, the main purpose is to create a profile (find features) of the employees that are going to quit their job. In order to achieve this, they made a data-set which contains demographic and work related information. Some names of the columns are the follow:

- Age.
- Education.
- Department.
- MonthlyIncome.
- EnvironmentSatisfaction.
- JobInvolvement.

Subsequently, they performed Statistical Analysis and specifically created figures which demonstrate relations between the features. For instance, in the plot where is depicted the ”Attrition” with the ”MaritalStatus” columns, we discern a pattern where the single employees have bigger attrition than those who are married and married workers have greater attrition in

comparison with the divorced. Furthermore, another significant inference can be made from the ”Department” and ”Attrition” plot. We can clearly see that the Sales department has the highest attrition and the HR department has the lowest.

Following, a Decision Tree algorithm is implemented as a way to find and distinguish the *important features*. The weighted decrease in node impurity divided by the likelihood of reaching that node is used to determine a feature's relevance. The node probability can be computed by dividing the total number of samples by the amount of samples that reach the node. The feature is more significant the higher the value. After plotting the importance and the features we notice that the most important columns are the:

- MonthlyIncome.
- Age.
- OverTime.
- DistanceFromHome.
- TotalWorkingYears.
- YearsWithCurrManager.
- NumCompaniesWorked.

After finding the significant variables, the authors, perform the K Means algorithm with 2 clusters. The conclusions are that Cluster 0 characteristics are:

- Lower attrition rate.
- More senior jobs.
- Higher salary.
- Greater number of companies worked for.
- Older age.
- Fewer single persons.
- More YearsAtCompany.
- Fewer YearsSinceLastPromotion.

By the time this research was completed the writers highlight 3 important findings.

2.2.3.1 Finding 1

Despite having a higher rate of promotion and salary than other non-managerial positions, the human resources department has a high attrition rate.

2.2.3.2 Finding 2

People who worked in two to four companies over the course of their careers are less likely to depart. After working for six firms, female attrition is significantly lower than male attrition.

2.2.3.3 Finding 3

Compared to other work levels where doctors nearly always have the lowest attrition rate, Job Level 3 has the greatest attrition rate for medical professionals. One explanation could be that it takes more time for doctors to reach job level 4.

Finally, it is worth mentioning some managerial advice the authors offer. To begin with, they suggest that the business should examine human resource positions more thoroughly to see which aspects of the job are unsatisfactory to employees. It is very advised to have one-on-one conversations regularly. Moreover, they recommend that males who attended more than five companies should still be given attention, even though the corporation is not required to concern much about those who worked for two to four different businesses. Lastly, they draw attention to the fact that doctors are taking longer to go from level 3 to level 4 jobs and are less pleased with their level 3. To ensure employees are adequately compensated, it is advised to carefully analyze the performance evaluation method.

2.2.4 Walking assets: The cost of losing an employee

In this work the author Juan Martin Carriquiry is doing a research on business's expenditures associated with voluntary employee turnover. To calculate the expenses of engineer turnover, evidence from a knowledge-intensive manufacturing company is specifically employed. Data was gathered from company documents, management and HR staff interviews. According to the research, staff turnover is a costly phenomena. The productivity loss is by far the biggest single financial burden for the business. This suggests that turnover is particularly expensive for jobs involving difficult tasks and lengthy learning curves.

The analysis of high attrition costs for the company should help with resource distribution in an effective manner. How much funding should go toward addressing sub-optimal turnover rates is a constant concern for managers, especially in the area of HR decision-making.

The writer implements a model which takes into consideration the expenses of decreased productivity caused by turnover in order to address the issue of the absence of precise information on the growth of employee productivity.

The addition of literature is that we may get a much fuller understanding of the actual effects on the firm by factoring learning processes into the evaluation of staff turnover. These results support Waldman et al(2004) .'s strategy of include learning curves at the individual level to estimate turnover costs and demonstrate that this methodology might be applied to other industries. This method should be used by practitioners to determine the actual financial loss caused by staff turnover and be included in financial statements and balances.

Juan Martin Carriquiry refers to tree different opinions regarding the fact to what extent employee behavior that results in attrition is bad for the business. The first one, mentioned in Baron et al.(2001) [5], considers employee turnover a negative case. Specifically, staff turnover is considered as an event that disrupts organizational patterns and hence poses a performance risk for the company. The second aspect expressed in another published work (Dalton and Todor (1979), Staw (1980), Dalton et al. (1981)[6][7][8]) states that staff turnover is an overestimated fact. They contend that because staff are generally simple to replace, turnover only slightly lowers output. Finally, according to the (Abelson and Baysinger (1984), Glebbeek and Bax (2004) [9]) work the real question is "what percentage of turnover is ideal for the company". Retention techniques have expenses associated with them, just as turnover strategies have costs associated with it. According to this cost-benefit analysis, the "ideal" rate of turnover should occur when the marginal costs of turnover and retention are equal.

This paper suggests that no business should strive for a zero turnover rate. According to them, management should aim to achieve a "optimal" turnover rate, which is defined as the point at which the net costs of attrition are equal to zero. Further efforts to keep employees would be suboptimal and financially ineffective at that point since the costs of retention strategies would equal the expenses of employee retention. This work's main topic is the economic effect of attrition on the company.

By utilizing a learning curve technique, like Waldman et al. $(2004)^2$ [10] did in his examination of turnover in the health care industry, this paper will contribute to the field. This

approach models learning as a continuous, non-linear process rather than a discrete one.

The full accounting of staff turnover expenses contains:

- **Departure of an employee** which includes administrative expenses, departure interviews conducted by the human manager and a representative from human resources, the manager's restructuring of the tasks, and IT expenses. When an employee retires, supervisors may need to rearrange their responsibilities, which might take time. If a temporary successor is needed for the retiring employee, the business may incur additional expenses. If the burden is distributed among the remaining staff, there may be increased overtime costs and lower productivity.

- **Recruiting.**

- **Educating and instructing the new employee.**

- **Costs of decreased productivity as a result of the new worker's poor productivity.**

The early productivity of new personnel is typically lower than that of preceding workers. According to the learning curve idea, the more practice you get, the more proficient you can be at the subject. The type of work will determine the form of the learning curve. The author employs a variant of the well-known parabolic curve model (Hackett, 1983) [11], which is structurally related to the accumulative learning methods (Mazur and Hastie, 1978). It is not universally accepted that learning is a smooth, monotonic process in which learning gradually declines; in reality, learning curves with a wide range of forms have been seen (Mazur and Hastie, 1978). Their model presupposes that learning processes occur at steady rate.

Output loss:

$$y_i = 1 - \frac{1}{ax_i + Exp} \tag{2.1}$$

Then becomes:

$$y_i = -\frac{1}{ax_i + Exp} \tag{2.2}$$

Where:

- x: time function.

- a: learning ability.

- Exp: previous experience.

and

$$Y_i = W_i * \int_i^{i+1} F(x)\,dx \tag{2.3}$$

Where:

- Y_i: productivity loss.
- W_i: salary for a specific period of time.

After, he expands the model to account for various payments throughout numerous time periods.

$$Y(i,N) = \sum_{i=0}^{i=N} W_i * \int_i^{N} F(x)\,dx \tag{2.4}$$

Data:

The HR department submitted firm documents, including information on remuneration, recruitment, commencement, and training expenses, from which attrition figures were produced. Manager statements were used to assess the expenses of additional training and decreased productivity. There was no information available about the productivity of the departing staff. To acquire information that would suggest a skewed attrition of lesser or better employees, managers were indeed asked what the effectiveness level of the engineers who left in the past year compared to those who remained was. In order to calculate the coeffcients for the modeling of learning curves, managers were furthermore requested to judge the attribution of workers at various time periods.

All of the engineers who departed the business between 2010 and 2011 make up the sample. 37 of the 340 engineers departed from the company in the past two years. 24 of these workers left the company voluntarily, 12 completed their short-term contracts, and 1 was fired.

The variables consist of the *departure, hiring, training costs* and the *productivity* which was calculated from the interviews conducted by the managers.

Results –Total costs of employee resignation:

The statistical analysis's findings indicate that, overall, it costs $41,549 to replace an engineer. The expense of decreased productivity, which accounted for over 70% of the overall costs, was by far the greatest factor. The research also demonstrates that, once salary differences are accounted, the value of the productivity loss after 4-5 years is quite comparable

Item	Cost per employee	Total cost
Departure	652	14,345
Hiring	2,546	56,003
Training	8,097	178,134
Loss productivity	30,254	665,592
Total costs	41,549	914,074

Figure 2.3: Total costs.

regardless of whether the new recruit has no experience or much of it. As anticipated, the engineers leaving the firm on general were younger and had worked for the business significantly less time than those staying.

Primarily, a significant portion of the costs of turnover are attributable to the learning process in the highly qualified roles taken into account. This is very pertinent since it emphasizes the financial costs to the company of training new staff.

According to this study, a variety of elements will affect the expenses of staff turnover for the company. First, the firm's recruiting procedures, which may be influenced by cultural and professional conventions. Second, the traits of the new employee, especially if they have prior experience in the field. The expenditures of employee turnover in this scenario do not seem to be much impacted by experience. Third, the technology and expertise needed for the job. These factors demonstrate that the expenses related with training new personnel account for the majority of the expense in the process of turnover.

Chapter 3

Methodology

This section describes the process of developing an intelligent system that is useful for the Human Resources Management department. Specifically, a real dataset is studied and robust models are developed to forecast the probability of an employee to search for a new job or stay in the company. In subsection 3.1, an overview of the dataset is provided. The utilized models are described in subsection 3.3, while 3.2 discusses about the handling of the data.

3.1 Data Collection

The analyzed dataset is the *"HR Analytics: Job Change of Data Scientists"* [13]. This dataset is designed to find the factors that make a person to leave current job by analyzing key elements, such as the current credentials, demographics and experience data. It consists of 14 features (columns) and 19159 rows. The features are:

- ***enrollee_id*** : Unique ID for candidate.
- ***city***: City code.
- ***city_development_index***: Developement index of the city (scaled).
- ***gender***: Gender of candidate.
- ***relevent_experience***: Relevant experience of candidate.
- ***enrolled_university***: Type of University course enrolled if any.
- ***education_level***: Education level of candidate.

- ***major_discipline***: Education major discipline of candidate.
- ***experience***: Candidate total experience in years.
- ***company_size***: No of employees in current employer's company.
- ***company_type*** : Type of current employer.
- ***lastnewjob***: Difference in years between previous job and current job.
- ***training_hours***: training hours completed.
- ***target***: 0 – Not looking for job change, 1 – Looking for a job change.

3.2 Data manipulation

The appropriate data management and pre-processing is of high importance given that the dataset is flawed in a lot of ways:

- The dataset is imbalanced.
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality.
- Missing imputation are a part of the pipeline as well.

To deal with these issues, we follow several important steps. At first, we use the *Label Encoding* technique to transform the labels into numeric form (machine-readable form). This is an essential step as the ML algorithms can only handle data in numeric form. Importantly, we use mapping dictionaries for ordinal features to provide a mapping that takes care of the different levels of a feature. For instance, the *education_level* feature has a lot of values (phd, primary, masters, ...) that should be ordered appropriately. The proper way to do it is with the following mapping:

- 'Primary School' : 0,
- 'High School' : 1,
- 'Graduate' : 2,
- 'Masters' : 3,

- 'Phd' : 4.

Following that, we use k-Nearest Neighbors (k-NN) imputer implementation to fill the missing values in the dataset. In specific, a sample with missing values is selected and the nearest neighbours are found utilising some kind of distance metric; for instance, the Euclidean distance. Then, the mean of all nearest neighbours is calculated and the missing value is filled up.

Furthermore, we utilize the Synthetic Minority Oversampling TEchnique (SMOTE) to increase the data in a balanced manner alleviating the "imbalanced dataset" issue. This is done by oversampling the examples in the minority class synthesizing new examples. SMOTE choses examples close in the feature space, creating a line among them and creating a new sample at a point along that line. This way, we develop a balanced dataset that will be the proper base to build powerful models that will generalize well avoiding the overfitting issue.

3.3 Machine Learning Algorithms

A plethora of machine learning and deep learning techniques is utilized in order to identify the advantages and disadvantages of each method. In the end, our goal is to come up with the best algorithm that combines high accuracy with minimal overhead. We utilize nine algorithms, namely XGBoost, LightGBM, CatBoost, Random Forest, FNN, LSTM, GRU, CNN and CNN-LSTM. There are some of the best state-of-the-art algorithms existing today.

3.3.1 XGBoost

Extreme Gradient Boosting, or XGBoost, is a distributed gradient boosting toolkit that has been tuned for speed and adaptability. It makes use of the gradient boosting framework, which creates a strong learner out of a group of weak learners (usually decision trees). The objective function that needs to be optimized is

$$J(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.1}$$

, where l is the loss function and Ω is the regularization term.

The technique finds the optimal splits more quickly by approximating the loss function with a second-order Taylor expansion. Because of this, XGBoost is very scalable and useful for a variety of issues.

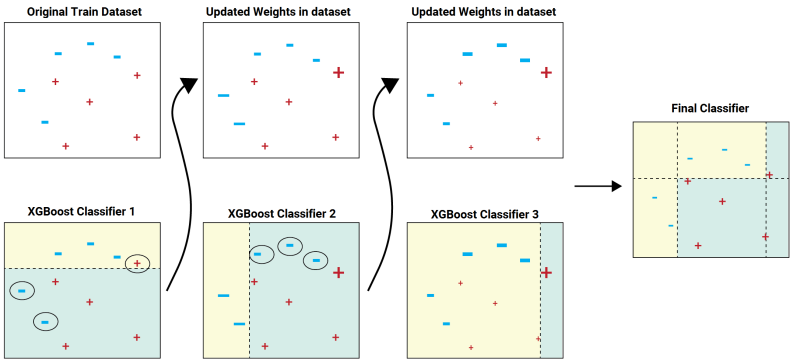

Figure 3.1: XGBoost classifier

3.3.2 LightGBM

Because of its focus on efficiency and speed, LightGBM, also known as the Light Gradient Boosting Machine, stands out among gradient boosting frameworks. LightGBM uses a unique histogram-based learning approach to generate decision trees leaf-wise, in contrast to conventional gradient boosting algorithms that build decision trees level by level. A more accurate model is frequently produced as a result of this divergence from the traditional method. While LightGBM's objective function is similar to XGBoost's, it uses distinct techniques for building trees.

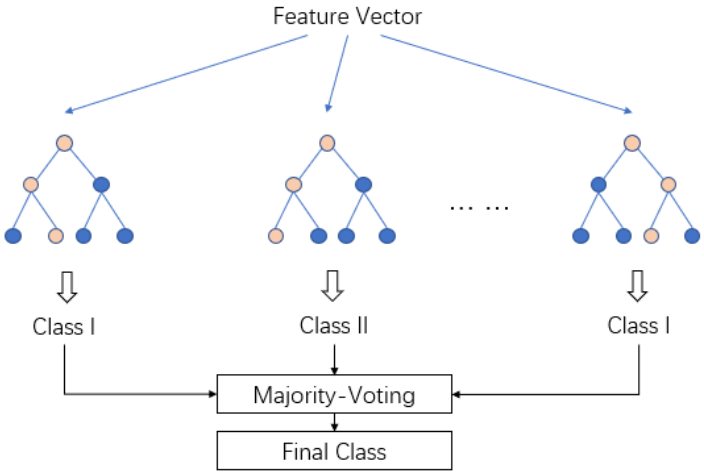

Figure 3.2: LightGBM classifier

In comparison to the level-wise technique, the leaf-wise growth strategy has the advantage of minimizing loss more, which enables the model to discover complex patterns in the data. It does have a possible negative, though, in that it is more prone to overfitting, in which

the model fits the training data too closely and finds it difficult to generalize to new data. LightGBM uses early stopping criteria in addition to other regularization methods to deal with this problem. LightGBM is a reliable option that is especially well-suited for large-scale data applications where speed, efficiency, and model accuracy are crucial. These methods assist prevent overfitting. It is a useful tool in machine learning and data science for tasks like classification, regression, and ranking since it may achieve a balance between these elements.

3.3.3 CatBoost

CatBoost, also known as categorical boosting, is a state-of-the-art gradient boosting algorithm created to excel at managing categorical characteristics inside datasets, doing away with the need for laborious preprocessing processes. CatBoost stands out due to its innovative ”ordered boosting” methodology, which deviates from conventional gradient boosting by using permutations to successfully counteract overfitting. It has a thorough comprehension of categorical variables and uses an extra term in the objective function to directly incorporate categorical variables into its learning process. This improves CatBoost’s capacity to uncover significant patterns and associations from such variables and enables it to make informed decisions when faced with categorical input.

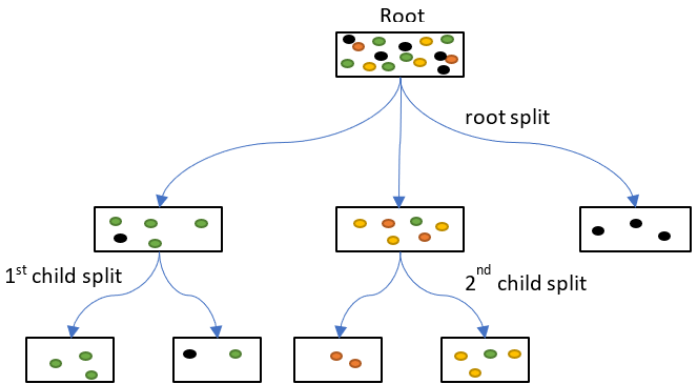

Figure 3.3: CatBoost classifier

CatBoost’s creativity extends to the model structure, where it introduces ”oblivious trees,” a decision tree variant renowned for its interpretability and effective inference skills. CatBoost excels in fields where both model accuracy and interpretability are essential due to its unique combination of characteristics. Whether it’s improving recommendation systems, enabling accurate medical diagnoses, or optimizing financial risk assessments, CatBoost stands

out as a potent tool that not only produces accurate predictions but also offers insights that are simply understandable, bridging the gap between machine learning complexity and practical usability for data scientists and machine learning practitioners.

3.3.4 Random Forest

In the area of machine learning, Random Forest is a powerful and adaptable ensemble learning technique that has gained great appeal. For classification tasks, it aggregates their outputs to calculate the mode of classes, and for regression tasks, it generates the mean prediction. Its essential idea entails the development of numerous decision trees throughout the training phase. Several significant benefits are introduced by this ensemble approach.

The ability of Random Forest to handle high-dimensional data is one of its distinguishing characteristics. When faced with datasets that include a large number of features, Random Forest excels at spotting important trends and connections while reducing the chance of overfitting. This tolerance to overfitting is a direct result of the ensemble method, which combines the knowledge of various decision trees. The final prediction becomes more reliable and less susceptible to the peculiarities of any one tree by averaging or collecting the majority vote from these trees. This improves the model's capacity to generalize effectively to new data and yields more accurate results.

Additionally, the parallelization potential of Random Forest's design makes it a highly effective user of distributed computing resources. Because of its capacity to utilize the power of numerous cores or distributed systems, Random Forest is a good choice for cloud-native applications and large-scale data processing, where the algorithm's ability to speed up training and increase scalability can be very useful. Since data dimensionality and model resilience are important considerations in a variety of applications, Random Forest remains a significant tool in the machine learning toolbox.

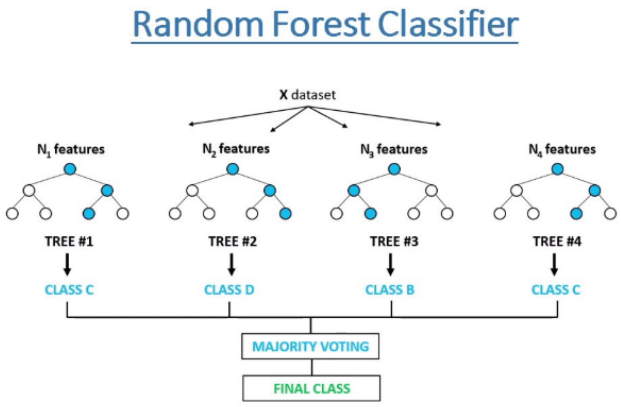

Figure 3.4: Random Forest classifier

3.3.5 Feed-forward Neural Network (FNN)

FNN is the abbreviation for ”feed-forward neural network.” One of the most basic and popular artificial neural network architectures is this one. A feed-forward neural network only allows information to flow in one direction, without creating any loops or cycles, from the input layer through one or more hidden layers to the output layer. The connections between each neuron (or node) in one layer and every neuron in the layer below have associated weights and biases that are modified during training.

In order to produce an output, a neuron in a FNN must first compute the weighted sum of its inputs, add a bias term, and then apply an activation function. The neurons in the following layer are then given access to this output. FNNs are frequently employed for several machine learning applications, such as function approximation, regression, and classification, among others. On the other hand, some neural network architectures like recurrent neural networks (RNNs) or transformers are more ideal for tasks involving sequential data or those needing recall of previous inputs. According to mathematics, a neuron’s output y is a function of the weighted sum of its inputs:

$$y = f(\sum_{i} w_i x_i + b) \tag{3.2}$$

, where f is an activation function, w_i are the weights, and b is the bias.

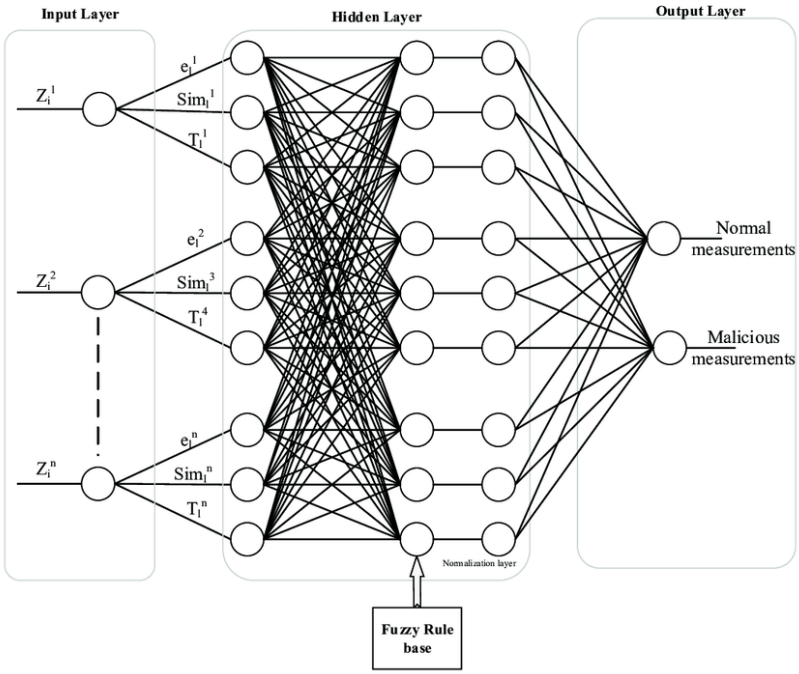

Figure 3.5: FNN classifier

3.3.6 Long Short-Term Memory (LSTM)

In the field of artificial neural networks, Long Short-Term Memory (LSTM) networks are a potent and significant advancement that are particularly designed for tackling problems involving sequential input. LSTMs were created primarily to get around the drawbacks of conventional recurrent neural networks (RNNs), which have trouble capturing long-range relationships and frequently experience the vanishing gradient problem during training. An LSTM's memory cell, which can store data across long sequences, is located at the center of the device. The forget gate, input gate, and output gate are three of the gating mechanisms this cell interacts with; each is in charge of regulating the information flow into and out of the cell state.

The output gate filters information from the cell state to generate the hidden state or output at each time step. The forget gate selects what information should be maintained or deleted, the input gate decides what new information should be introduced, and the output gate filters information from the cell state. LSTMs are skilled at identifying long-term dependencies and patterns in sequential data due to their ability to selectively recall and discard information provided by these sophisticated gating mechanisms.

The uses of LSTMs are wide-ranging and include time-series forecasting, speech recog-

nition, language translation, sentiment analysis, and other natural language processing tasks. Their ability to represent intricate temporal linkages has enabled advancements in a variety of fields, from the comprehension and production of human language to the accurate forecasting of financial markets. As a result, LSTMs have transformed the field of machine learning and artificial intelligence by making it possible to model complex dependencies and memory in dynamic data sequences. They are now an essential tool for researchers, data scientists, and engineers working on sequential data tasks. The following equations define the LSTM cell:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
\tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
h_t &= o_t * \tanh(C_t)
\end{aligned}
$$

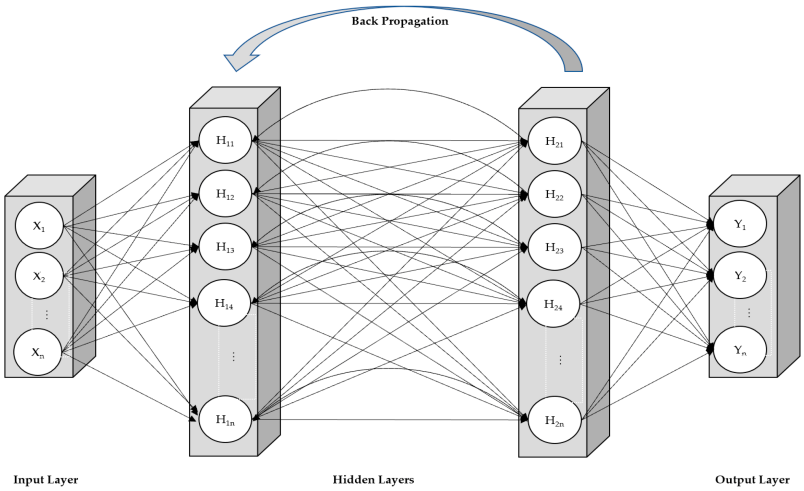

Figure 3.6: LSTM classifier

3.3.7 Gated Recurrent Unit (GRU)

Similar to Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs) are a sort of recurrent neural network (RNN) architecture that was created to overcome the difficulties of modeling sequential data. Cho et al. introduced GRUs in 2014, and they have grown in popularity due to their usefulness and efficiency in a variety of applications. They

resemble LSTMs in some ways but have a simpler architecture and fewer gating mechanisms.

The update gate and the reset gate are the two main parts of a GRU. The reset gate decides which information from the past should be ignored, while the update gate controls which information from the previous time step should be carried over to the present time step. Together, these gates selectively update the hidden state, enabling GRUs—which are computationally more effective than LSTMs—to capture pertinent temporal connections.

GRUs have found applications in a variety of fields, including time-series analysis, speech recognition, and natural language processing, where a balance between model complexity and performance is crucial. They have become an important tool in the toolbox of deep learning practitioners due to their capacity to model sequential data accurately and rapidly. A GRU cell's governing equations are as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z)$$
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r)$$
$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t] + b)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

GRUs offer similar performance to LSTMs but are computationally more efficient due to their reduced complexity.

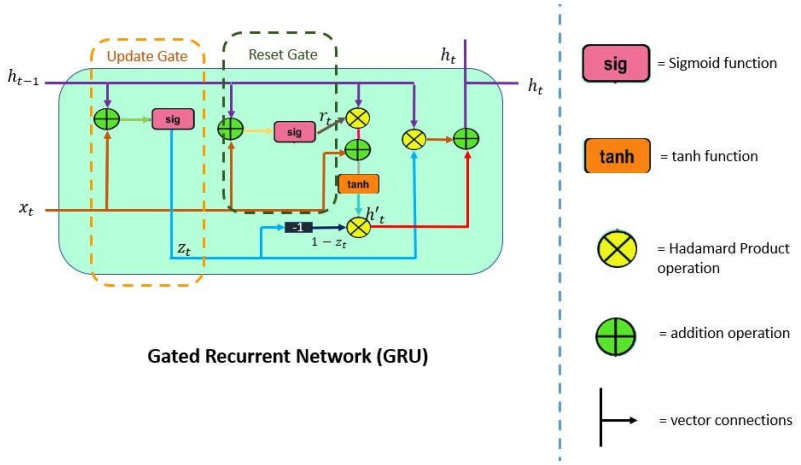

Figure 3.7: GRU classifier

3.3.8 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are a subset of artificial neural networks that are particularly adept in processing and analyzing visual data, especially pictures and videos.

They were motivated by the human visual system's capacity to identify hierarchical patterns, ranging from straightforward elements like edges to intricate ones like shapes and objects. Convolutional neural networks (CNNs) are distinguished by their distinctive design, which consists of pooling, convolutional, and fully connected layers. CNNs have shown to be incredibly efficient in a variety of computer vision tasks.

Convolutional layers, which perform convolutional operations on input data, are the foundation of CNNs. Small filters or kernels are slid across the input throughout these operations, allowing the network to automatically learn and extract features like edges, textures, and corners from the unprocessed pixel values. Following convolutional layers, pooling layers downsample the feature maps to reduce the dimensionality of the input while keeping the key details. For classification or regression tasks, the collected features are subsequently routed via fully linked layers.

By obtaining cutting-edge results in image classification, object detection, facial recognition, and many other applications, CNNs have transformed computer vision. They are an essential component of contemporary artificial intelligence, enabling innovations like self-driving cars, medical image analysis, and image-based recommendation systems. This is due to their capacity to automatically learn and represent hierarchical characteristics. A convolution operation is described mathematically as:

$$(f * g)(t) = \sum_{\tau} f(\tau) g(t - \tau) \tag{3.3}$$

CNNs are highly effective in tasks related to image recognition, object detection, and even some natural language processing tasks. Their ability to automatically and adaptively learn spatial hierarchies of features makes them highly efficient and accurate.

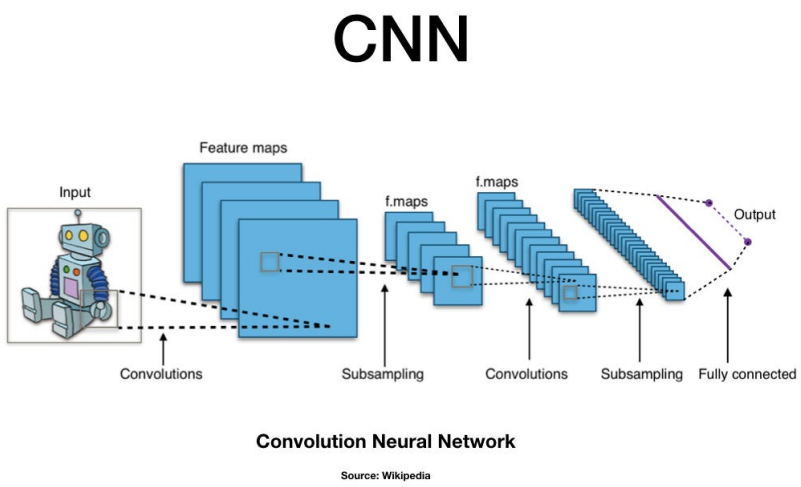

Figure 3.8: CNN classifier

3.3.9 CNN-LSTM

In order to understand both spatial and temporal aspects inside data, CNN-LSTM architectures combine Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). The advantages of these two different types of neural networks are combined in this hybrid model to produce a powerful tool for several applications. In this design, feature extraction from individual frames or data segments is typically carried out by CNN layers. These layers are excellent in finding spatial patterns in photos, movies, or other data with grid-like features, such as edges, textures, or object sections. CNNs offer a rich representation of spatial data by separately processing each frame.

However, LSTM layers excel at identifying temporal connections within data sequences. They are highly suited for jobs requiring time-series data, natural language processing, and any situation where understanding the order and context of input is vital because they are specifically intended to remember and learn from previous knowledge. LSTM layers work on the outputs of CNN layers in CNN-LSTM architectures, allowing them to simulate the temporal relationships between frames or segments.

This combination is especially useful for jobs like video classification, where it's important to consider both the sequence of frames and their temporal evolution in addition to the content of individual frames. Similar to how it can efficiently learn patterns and trends for

time series prediction while taking historical context into account. In essence, CNN-LSTM architectures combine the strengths of CNNs and LSTMs to provide a strong response for tasks requiring a thorough comprehension of both spatial and temporal information, advancing research in areas such as video analysis, gesture recognition, and more.

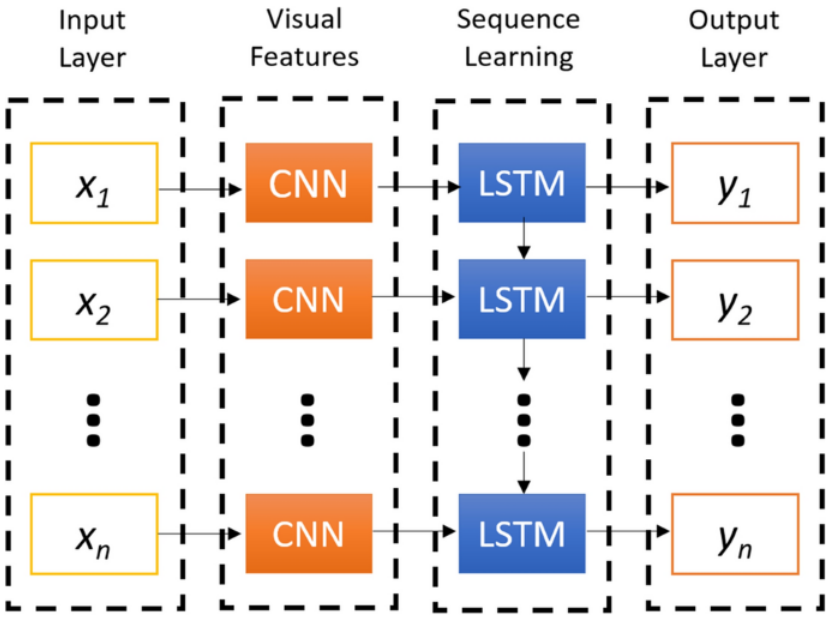

Figure 3.9: CNN - LSTM classifier

Chapter 4

Results

4.1 Evaluation Metrics

We use the ROC curve and the AUC score to assess our models. The Receiver Operating Characteristics (ROC) curve is a graph that shows a classification model's effectiveness across all thresholds. Plotting two parameters—the true positive rate (TPR) and the false positive rate (FPR)—leads to the creation of this curve. The Area Under Curve (AUC) score reveals how separable something is. Higher AUC models are more accurate at predicting True Positives and True Negatives. The total area below the ROC curve is determined by the AUC score. Scale and threshold invariance are both characteristics of AUC. AUC score, in terms of probability, is the likelihood that a classifier would rank a randomly selected positive instance higher than a randomly selected negative one.

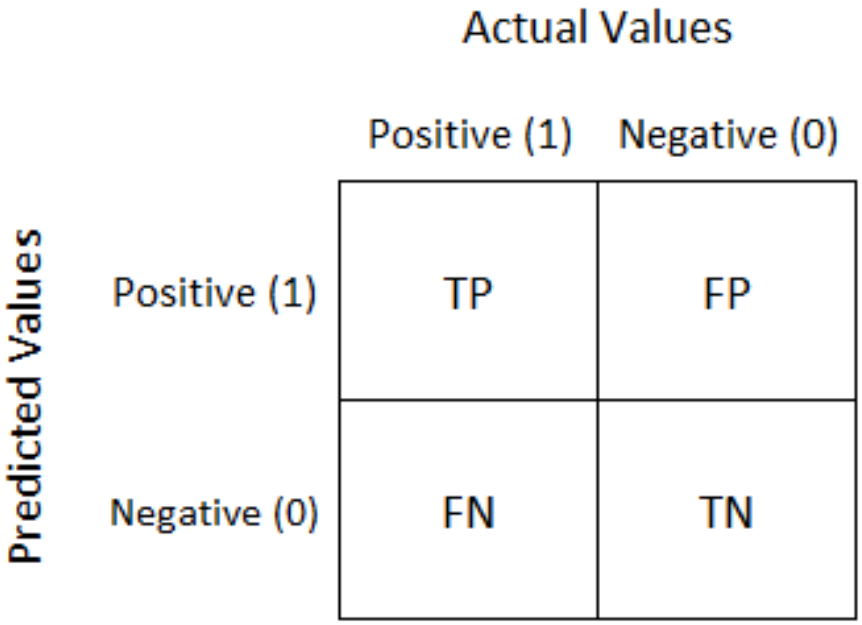

Figure 4.1: True Positive Rate - False Positive Rate.

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{4.1}$$

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4.2}$$

4.2 Hyper-parameter Tuning

The process of fine-tuning hyperparameters is a crucial step in the development of a reliable machine learning model. These hyperparameters are unique parameters that control the model's training process; they are configured beforehand rather than being learned from the data. They have a substantial impact on the model's performance, therefore determining the best set requires careful consideration.

The process of hyperparameter tuning is compared in the document to cracking a three-dial combination lock, where each dial stands for a parameter. Finding the ”best” combination is a difficult but necessary task because the model admits numerous combinations as correct, unlike the lock which only accepts one.

The theory suggests assessing all possible parameter combinations to speed up this process; this strategy potentially requires 1000 evaluations for a model with three parameters,

each with a range of 1 to 10. In light of the potential time and computing resources that this method may require, we support more effective techniques, such as:

- **Random Search:** This approach entails defining a defined range for hyperparameter values and picking random locations to assess within this range. When compared to other methods, this one is praised for its ability to identify the ideal parameters in a shorter amount of time.

- **Grid Search:** A grid of hyperparameter values is created in this case, and each spot on the grid is assessed to determine the optimal values. Although laborious, this approach guarantees a comprehensive search across all potential combinations inside the specified grid.

- **Bayesian Optimization** To find the ideal hyperparameters, this advanced technique minimizes a particular function, usually the loss function. In order to choose the most promising hyperparameters for assessment in the real objective function, it builds a probabilistic model of the objective function.

As a conclusion, we outline a thorough methodology for hyperparameter tuning that combines time-tested approaches like Grid and Random Search with cutting-edge methods like Bayesian Optimization. This method not only guarantees careful optimization but also encourages the growth of a model that is both accurate and reliable, opening the door for effective machine learning implementations.

4.3 Evaluation with Cross Validation – Kfold

We must evaluate the model after configuring the hyper-parameters. We employ the well-known method of K-Fold Cross Validation (CV) to effectively evaluate our methods.

k-Fold CV is a resampling method used to assess how well a machine learning model is working. By dividing the original dataset into k folds of equal (or nearly equal) size, the main goal is to achieve a more accurate evaluation of the model's performance. The remaining $k-1$ folds are utilized for training, and one fold is kept as the validation set. This procedure is carried out k times, using a different fold as the validation set each time. The average error across all k trials is calculated mathematically as:

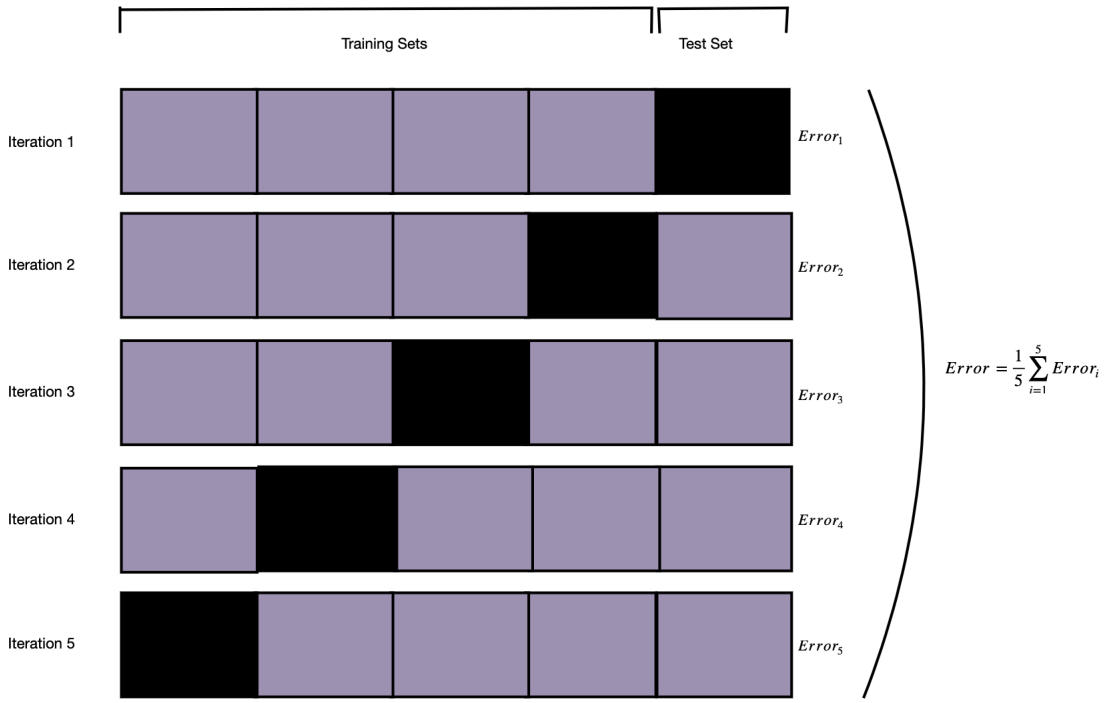

Figure 4.2: 5-Fold Cross Validation

$$\text{Average Error} = \frac{1}{k}\sum_{i=1}^{k} E_i$$

where E_i is the error rate on the i-th fold.

This method maximizes both the test and training data, making it very helpful for small datasets. Additionally, it offers a thorough analysis of a model's performance, providing information on the stability and dependability of the model. k-Fold is widely utilized in machine learning tasks like classification, regression, and time-series forecasting.

4.3.1 XGBoost

XGBoost performed remarkably well with the approrpiate hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is:

- $n\_estimators$: 114,
- min_child_weight: 14.3,
- $gamma$: 4.5,
- $colsample\_bytree$: 0.90.

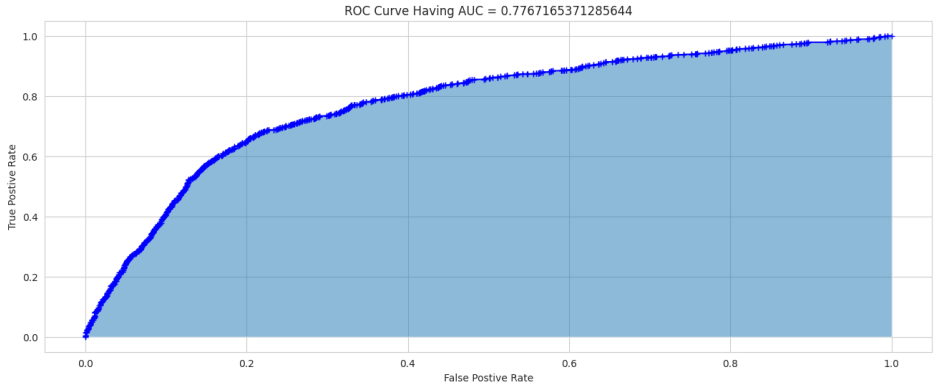

Figure 4.3: XGBoost Classifier ROC Curve

It managed to achieve an AUC score of 0.78. The respective ROC curve is illustrated in figure 4.3.

4.3.2 LightGBM

LightGBM has a great performance with the approrpiate hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is:

- $n\_estimators$: 100,
- min_child_weight: 1.0.

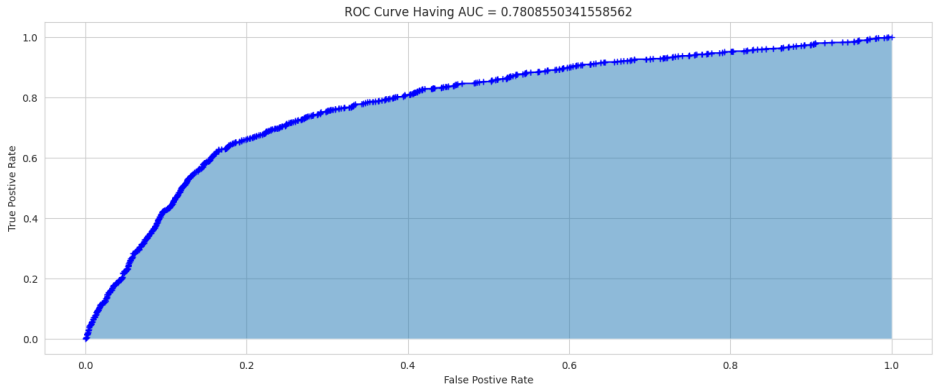

Figure 4.4: LightGBM Classifier ROC Curve

It managed to achieve an AUC score of 0.78. The respective ROC curve is illustrated in figure 4.4.

4.3.3 CatBoost

CatBoost was also very efficient with the proper hyper-parameter tuning. Specifically, the optimal hyper-parameter configuration is with $n\_estimators$ equal to 979. It achieves an AUC score of 0.77. The respective ROC curve is shown in figure 4.5.

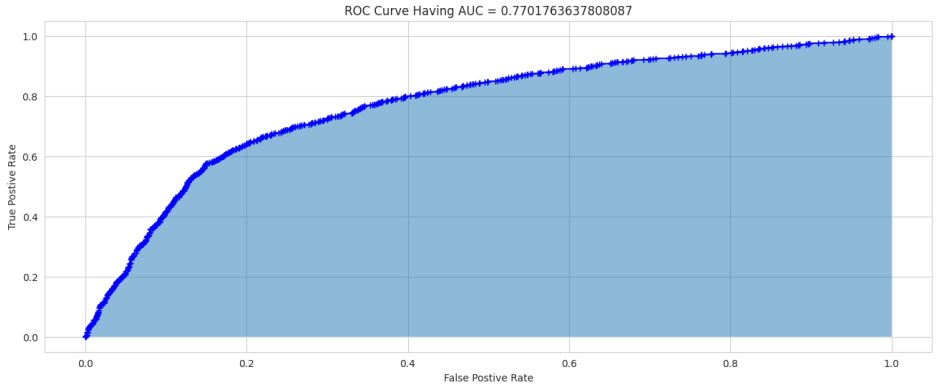

Figure 4.5: Catboost Classifier ROC Curve

4.3.4 Random Forest

Random Forest performed well using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is with $n\_estimators$ equal to 174. It achieves an AUC score of 0.74. The respective ROC curve is shown in figure 4.6.

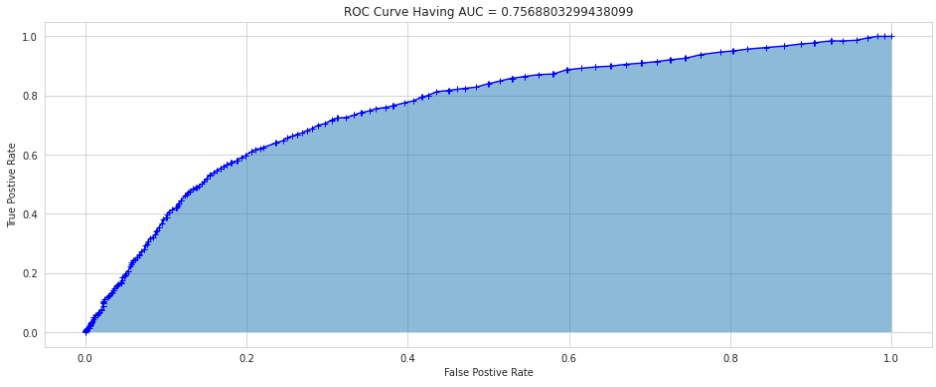

Figure 4.6: Random Forest Classifier ROC Curve

.

4.3.5 Feed-forward Neural Network

FNN performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is shown in 4.2.

Layer	Type	Neurons	Activation Func.
1	$Dense$	64	$Relu$
2	$Dense$	64	$Relu$
3	$Dense$	1	$Sigmoid$

Table 4.1: FNN Hyper-parameter Configuration.

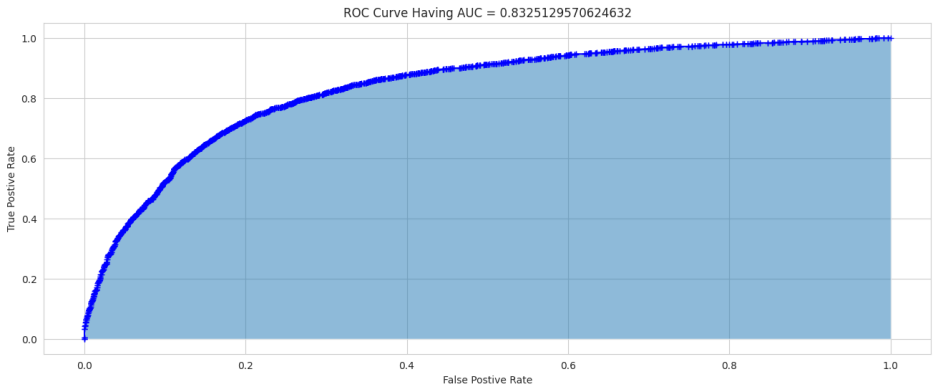

Figure 4.7: FNN Classifier ROC Curve.

We use *binary crossentropy* loss and $Adam$ Optimizer. It achieves an AUC score of 0.83. The respective ROC curve is shown in figure 4.7.

4.3.6 LSTM Neural Network

LSTM performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is shown in 4.2.

We use *binary crossentropy* loss and $Adam$ Optimizer. It achieves an AUC score of 0.81. The respective ROC curve is shown in figure 4.8.

Layer	Type	Neurons	Activation Func.
1	$LSTM$	64	$Relu$
2	$Dense$	64	$Relu$
3	$Dense$	1	$Sigmoid$

Table 4.2: LSTM Hyper-parameter Configuration

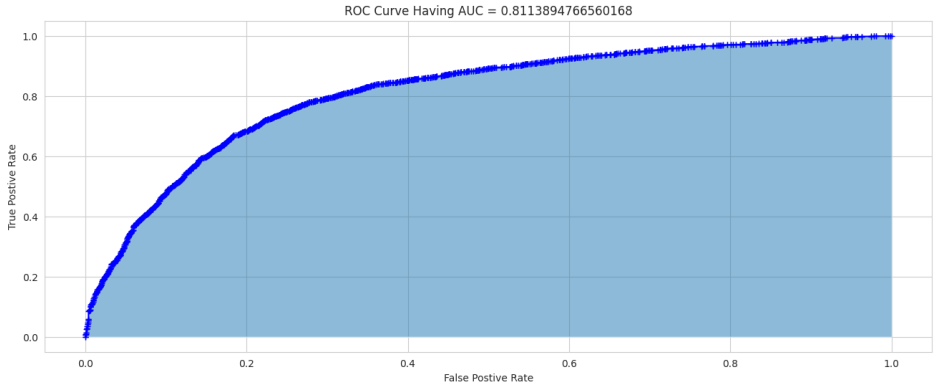

Figure 4.8: LSTM Classifier ROC Curve.

Layer	Type	Neurons	Activation Func.
1	GRU	64	$Relu$
2	$Dense$	64	$Relu$
3	$Dense$	1	$Sigmoid$

Table 4.3: GRU Hyper-parameter Configuration.

4.3.7 GRU Neural Network

GRU performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is shown in 4.3.

We use *binary crossentropy* loss and $Adam$ Optimizer. It achieves an AUC score of 0.82. The respective ROC curve is shown in figure 4.9.

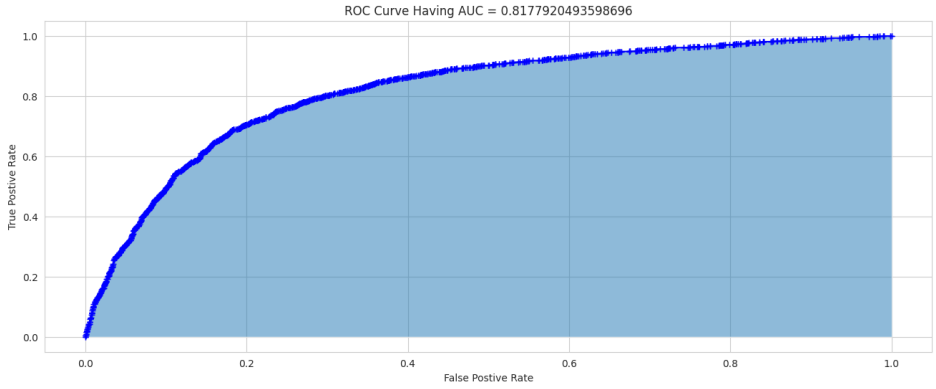

Figure 4.9: GRU Classifier ROC Curve.

4.3.8 Convolutional Neural Network

CNN performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is shown in 4.4.

Layer	Type	Neurons	Activation Func.	Kernel Size	Pool Size
1	$Conv1D$	64	$Relu$	3	-
2	$MaxPooling1D$	-	-	-	2
3	$Conv1D$	64	$Relu$	3	
4	$MaxPooling1D$	-	-	-	2
5	$Flatten$	-	-	-	-
6	$Dense$	64	$Relu$	-	-
7	$Dense$	1	$Sigmoid$	-	-

Table 4.4: CNN Hyper-parameter Configuration.

We use *binary crossentropy* loss and $Adam$ Optimizer. It achieves an AUC score of 0.85. The respective ROC curve is shown in figure 4.10.

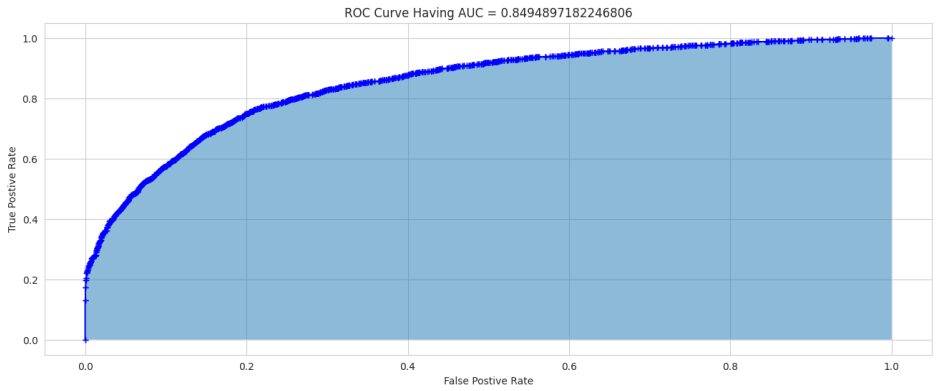

Figure 4.10: CNN Classifier ROC Curve.

4.3.9 CNN-LSTM

CNN-LSTM performed excellently using the right hyper-parameters. Specifically, the optimal hyper-parameter configuration is shown in 4.5.

We use *binary crossentropy* loss and $Adam$ Optimizer. It achieves an AUC score of 0.85. The respective ROC curve is shown in figure 4.11.

Layer	Type	Neurons	Activation Func.	Kernel Size	Pool Size
1	$Conv1D$	64	$Relu$	3	-
2	$MaxPooling1D$	-	-	-	2
3	$Conv1D$	64	$Relu$	3	
4	$MaxPooling1D$	-	-	-	2
5	$LSTM$	64	$Relu$	-	-
6	$Dense$	64	$Relu$	-	-
7	$Dense$	1	$Sigmoid$	-	-

Table 4.5: CNN-LSTM Hyper-parameter Configuration.

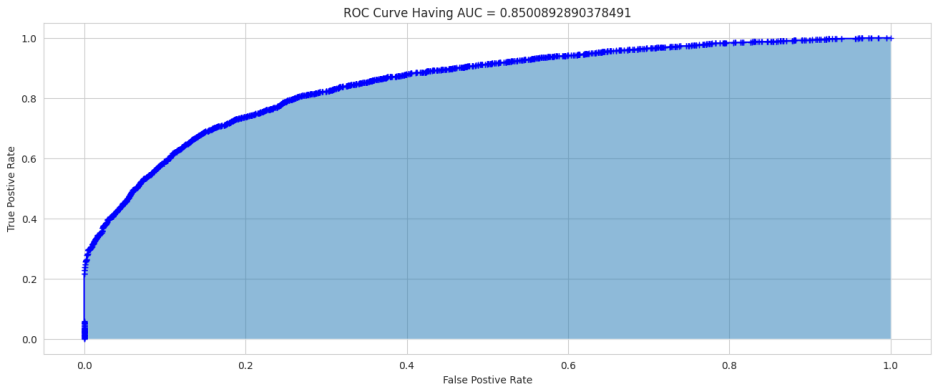

Figure 4.11: CNN-LSTM Classifier ROC Curve.

4.3.10 Summary

The results are summarized in figure 4.12 depicting the models' great generalization performance. Generally, the models avoided over-fitting and reached excellent classification scores. Thus, we recommend all of these models for the studied problem as they combine high accuracy with little training time. Specifically, outstanding performance is noticed when utilizing neural networks and specifically, FNN, CNN and CNN-LSTM with AUC scores of 0.83, 0.85 and 0.85 respectively.

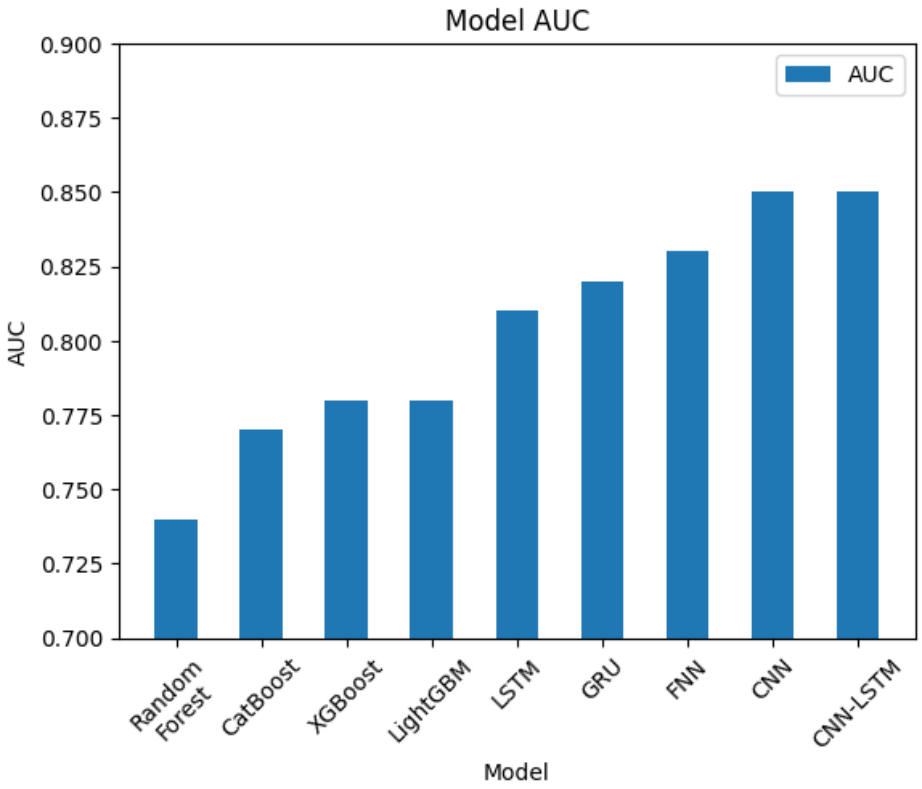

Figure 4.12: Model's Performance.

Table 4.6: AUC for each Classifier.

Classifiers	AUC
Random Forest	0.74
CatBoost	0.77
XGBoost	0.78
LightGBM	0.78
LSTM	0.81
GRU	0.82
FNN	**0.83**
CNN	**0.85**
CNN-LSTM	**0.85**

Chapter 5

Conclusions

The goal of the current work is to provide a thorough overview of the developing topic of Human Resources (HR) Analytics by analyzing three important publications and a renowned book. These chosen sources provide insight into the history, development, and prospective uses of HR Analytics in addition to encapsulating its fundamental ideas and approaches. For HR specialists, data scientists, and researchers interested in comprehending the complex environment of HR Analytics, this literature review is a helpful resource.

This study explores a real-world HR dataset in addition to the theoretical underpinnings to apply knowledge to practice. We have created a sophisticated predictive system that can anticipate the likelihood that an employee would decide to remain with the company or quit using cutting-edge artificial intelligence techniques. This prediction tool is a prime example of how HR Analytics is applied practically in today's data-driven enterprises, providing HR departments with invaluable insights for talent retention and workforce management.

Our research goes a step further by suggesting and carefully analyzing nine artificial intelligence techniques specifically designed for HR Analytics. These techniques have shown excellent success in projecting employee retention because they combine the benefits of numerous models to increase predictive accuracy. This empirical analysis not only highlights the potential of AI techniques in HR Analytics, but it also demonstrates our dedication to the field's advancement by providing workable solutions that go beyond theoretical insights. In summary, our work aims to close the knowledge gap between HR theory and data-driven decision-making, enabling organizations to fully realize the transformative potential of HR Analytics.

5.1 Future Work

In the future, there is a lot more we can do to advance our work in HR Analytics. There is still a lot of room for growth and improvement in the use of artificial intelligence to help businesses maintain their employees' happiness and engagement.

First off, by including more variables in the mix, we can improve our prediction tool even further. We may be able to better understand what would cause employees to stay or go if we know things like how content they are at work, how they interact with their managers, and even their personal hobbies. Additionally, utilizing real-time data could assist businesses in making quick adjustments to their plans in order to stay competitive.

Next, we can look into fresh applications for AI in HR. The approaches we've seen so far have been quite promising, but what if we merged them or changed them to make them even more effective? These tools could also be used in other HR areas, such as assisting in the search for fresh talent or determining the most effective means of assisting employees in furthering their careers.

It's also very important to consider the ethical side of things as we move forward. We must be careful to respect people's privacy and use data in a responsible manner. One of the main focuses of our future work will be developing standards and best practices for this.

Finally, collaborating with other subject-matter experts from the academic and business worlds could significantly speed up the process. Working together and exchanging ideas allows us to develop ground-breaking solutions that have a significant impact on the HR industry.

In summary, HR Analytics has a promising future. There are countless opportunities available, and we are only beginning to explore them. We can help create a future where HR isn't only about managing people, but also about fostering their success by building on what we've learned thus far.

Bibliography

[1] AI & Automation, [online] https://www.gsma.com/futurenetworks/wiki/aiautomation-anoverview/?fbclid=IwAR1gvlMewcrYvXYFrLwW_120E5oreF9zjzGX9CCJVNuaFlPcPleBX5XTpQ . Access Date: 5/07/2022.

[2] How HR Analytics Are Changing Business, [online] https://lesley.edu/article/howhr-analyticsarechangingbusiness . Access Date: 17/07/2022.

[3] Human Resource Management Review, Volume 32, Issue 2, June 2022, 100795 Human Resource Management Review Volume 32, Issue 2, June 2022, 100795, Human resources analytics: A systematization of research topics and directions for future research.

[4] Deductive vs Inductive Reasoning, Inductive vs. Deductive Research Approach (with Examples), [online] https://www.scribbr.com/methodology/inductive-deductive-reasoning/.

[5] Baron, J., M. Hannan, and M. Burton (2001). Labor pains: Change in organizational models and employee turnover in young, high-tech rms. American Journal of Sociology 106 (4), 960-1012.

[6] Dalton, D. and W. Todor (1979). Turnover turned over: An expanded and positive perspective. Academy of Management Review 4 (2), 225-235.

[7] Staw, B. (1980). The consequences of turnover. Journal of Occupational Behaviour 1 (4), 253-273.

[8] Dalton, D., W. Todor, and D. Krackhardt (1982). Turnover overstated: The functional taxonomy. Academy of management Review 7 (1), 117-123.

[9] Abelson, M. and B. Baysinger (1984). Optimal and dysfunctional turnover: Toward an organizational level model. Academy of Management Review 9 (2), 331-341.

[10] Waldman, J., F. Kelly, S. Arora, and H. Smith (2004). The shocking cost of turnover in health care. Health Care Management Review 29 (1),2.

[11] Mazur, J. and R. Hastie (1978). Learning as accumulation: A reexamination of the learning curve. Psychological Bulletin 85 (6), 1256.

[12] Juan Martin Carriquiry. Walking assets: The cost of losing an employee. Aalborg University January 1, 2013.

[13] HR Analytics: Job Change of Data Scientists, [online] https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?resource=download&select=aug_train.csv . Access Date: 5/07/2022.

[14] Christopher M. Rosett Austin Hagerty. Introducing HR Analytics with Machine Learning.

[15] Andy Liaw and Matthew C. Wiener. Classification and regression by randomforest. 2007.

[16] Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR Analytics. The Oxford Handbook of Personnel Assessment and Selection.

[17] Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. The MIT Press.

[18] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep Learning. MIT press Cambridge.

[19] Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. Harvard Business Review, 88(10), 52-58.

[20] Buhl, A., Müller, S., & Matzler, K. (2019). Towards a Unified Model of Machine Learning in Business and Society. Journal of Business Research.

[21] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[22] XGBoost: A Scalable Tree Boosting System" (Chen & Guestrin, 2016).

[23] XGBoost: A Fast Distributed Gradient Boosting System" (Chen et al., 2016).

[24] ImageNet Classification with Deep Convolutional Neural Networks.

[25] Long short-term memory (Hochreiter & Schmidhuber, 1997).

[26] Understanding LSTM Networks(Greff, Srivastava, Koutník, Steunebrink,& Schmidhuber, 2017).

[27] People Analytics in the Era of Big Data (Davenport, Harris, & Shapiro, 2010).

[28] A Roadmap for HR Analytics: From Data to Value (Rasmussen, Ulrich, & Beatty, 2011).

[29] The Future of HR: Embrace Data-Driven Decision-Making(Bersin, 2013).

[30] Predictive HR Analytics: Mastering the HR Metric(Lawler III, Levenson, & Boudreau, 2004).

[31] People Analytics: Driving Business Performance with People Data (Van Den Heuvel, Bondarouk, & Strohmeier, 2021).

APPENDICES

Appendix A

Code

The code which was created for this thesis is available in the following link: https://drive.google.com/drive/folders/118W9HRn8fVYLhHhThkxz_jJ9rG9QMaqn?usp=drive_link