ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΟΝ ΕΝΤΟΠΙΣΜΟ ΡΕΥΜΑΤΟΚΛΟΠΗΣ

Διπλωματική Εργασία

Ελευθερία Πετριανού

Επιβλέπων: Δημήτριος Μπαργιώτας

Βόλος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΕΧΝΙΚΕΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΟΝ ΕΝΤΟΠΙΣΜΟ ΡΕΥΜΑΤΟΚΛΟΠΗΣ

Διπλωματική Εργασία

Ελευθερία Πετριανού

Επιβλέπων: Δημήτριος Μπαργιώτας

Βόλος 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

MACHINE LEARNING TECHNIQUES FOR ELECTRICITY THEFT DETECTION

Diploma Thesis

Eleftheria Petrianou

Supervisor: Dimitrios Bargiotas

Volos 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων	**Δημήτριος Μπαργιώτας** Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Εμμανουήλ Βάβαλης** Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Μιχαήλ Βασιλακόπουλος** Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 15-02-2021

Στην οικογένειά μου.

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε από την ανάγκη μου να συνδυάσω τους τομείς της ενέργειας και της επιστήμης δεδομένων. Πολλά σενάρια διαμορφώθηκαν μέχρι να φτάσουμε στο τελικό θέμα και σε αυτό το κομμάτι θα ήθελα να ευχαριστήσω πολύ τον κύριο Χούστη Ηλία, ομότιμο καθηγητή του τμήματος, για την πολύτιμη βοήθειά του και την καθοδήγησή του σε αυτό το σχεδόν άγνωστο για μένα πεδίο εφαρμογής των τεχνικών μηχανικής μάθησης. Θα ήθελα επίσης να ευχαριστήσω θερμά τον κύριο Βάβαλη Εμμανουήλ, ο οποίος συνέβαλε σημαντικά στην πραγματοποίηση του τεχνικού μέρους της εργασίας, με τις γνώσεις, το χρόνο και την υπομονή του ως δεύτερος επιβλέπων καθηγητής. Επίσης σημαντική ήταν και η εμπιστοσύνη που μου έδειξε από την αρχή ο κύριος Μπαργιώτας Δημήτριος για την εκπόνηση του συγκεκριμένου θέματος. Δεν θα μπορούσα βέβαια να μην αναφέρω πόσο σημαντική ήταν η στήριξη και η βοήθεια της οικογένειάς μου όλο το διάστημα των σπουδών μου, που χωρίς αυτή δεν θα είχα καταφέρει να φτάσω ως εδώ.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Η Δηλούσα

Ελευθερία Πετριανού

15-02-2021

Περίληψη

Η ρευματοκλοπή είναι ένα θέμα που απασχολεί για χρόνια τις εταιρείες παροχής ενέργειας, επηρεάζει την ποιότητα των υπηρεσιών ηλεκτρικής ενέργειας και μειώνει τα λειτουργικά κέρδη. Προκειμένου να βοηθηθούν οι εταιρείες κοινής ωφέλειας να αντιμετωπίσουν την αναποτελεσματική επιθεώρηση ηλεκτρικής ενέργειας και την παράτυπη κατανάλωσή της, μέθοδοι μηχανικής μάθησης, σε συνδυασμό με έξυπνους μετρητές, εφαρμόζονται και προσφέρουν πολλούς και διάφορους τρόπους αντιμετώπισης του φαινομένου. Στην παρούσα διπλωματική εξετάσαμε την απόδοση μη-νευρωνικών αλγορίθμων για τον εντοπισμό ρευματοκλοπής, που περιλαμβάνουν τους XGBoost (eXtreme Gradient Boosting) και autoML (automated Machine Learning) αλγορίθμους. Επιπλέον, συγκρίναμε τους παραπάνω αλγορίθμους με μελέτες βασισμένες σε νευρωνικούς αλγορίθμους για το ίδιο πρόβλημα. Ένα από τα κίνητρα για την χρήση νευρωνικών αλγορίθμων είναι η ικανότητα τους να εκτιμούν τις παραμέτρους (features) αυτόματα. Αυτό επιτυγχάνεται και με τους AutoML που βασίζονται σε σύνολο με γνωστή θεωρητική συμπεριφορά αλγόριθμους. Οι μέθοδοι εφαρμόστηκαν στα δεδομένα από το Irish Social Science Data Archive –ISSDA της Επιτροπής Ρύθμισης Ενέργειας της Ιρλανδίας και από το πρότζεκτ Low Carbon London του UK Power Networks, και εξετάσαμε διάφορα σενάρια χρήσης των αλγορίθμων. Τα αποτελέσματα ήταν αρκετά ικανοποιητικά, καθώς η ακρίβειά τους κινήθηκε πάνω από το 94% για πλήθος δοκιμών.

Abstract

Electricity theft is an issue that has been of concern to electricity companies for years, which affects the quality of the energy supply and reduces operating profits. To help utility companies, to solve the problem of inefficient inspection electricity and irregular energy consumption, methods of machine learning, combined with smart meters, are applied in different ways of dealing with this phenomenon. In this dissertation, we examined the performance of non-neural network algorithms to detect electricity theft, including XGBoost (eXtreme Gradient Boosting) and autoML (automated Machine Learning) algorithms. In addition, we compared them to studies based on neural network algorithms for the same problem. One of the motivations for using neural algorithms is their ability to estimate parameters (features) automatically. This is the main characteristic of AutoML algorithms which are based on algorithms with well-known theoretical behavior and analysis. The considered algorithms applied to various application scenarios based on the datasets from the Irish Social Science Data Archive –ISSDA of Commission for Energy Regulation – CER and the Low Carbon London project of UK Power Networks, and produced significant accuracy above the level of 94%.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Εικόνων

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια η μηχανική μάθηση (ML) έχει καταφέρει να προσφέρει λύσεις σε πολλούς επιστημονικούς και κοινωνικούς τομείς. Παραδείγματα εφαρμογών των τεχνολογιών αυτών μπορεί κανείς να βρει στην οικονομία, υγεία, ρομποτική, εκπαίδευση, και ενέργεια. Συγκεκριμένα στον τομέα της ενέργειας, ένας τομέας που όσο περνάει ο καιρός επηρεάζει κατά πολύ τον τρόπο ζωής μας και την βιωσιμότητα του κόσμου, οι τεχνολογίες αυτές μπορούν να προσφέρουν επίγνωση σε πολλά «ανοιχτά» θέματα. Ένα από αυτά είναι η πρόβλεψη της κατανάλωσης ενέργειας και το αποτύπωμά της, με σκοπό την μείωση του κόστους παραγωγής και των ρύπων για το περιβάλλον.

Η εισαγωγή των ανανεώσιμων πηγών ενέργειας στα υπάρχοντα συστήματα ενέργειας έχει δημιουργήσει μια σειρά νέων προβλημάτων που πρέπει να αντιμετωπιστούν. Επιπλέον, η διαθεσιμότητα των τεχνολογιών IoT (Internet of Things), που επιτρέπουν την σύνδεση δικτυακών δομών, υπολογιστών, και έξυπνων συσκευών και συμβάλουν στην καλύτερη διαχείριση της ενέργειας. Για παράδειγμα, από τη χρήση του IoT σε μια κατοικία, μπορούμε να έχουμε στη διάθεσή μας πληροφορία που αφορά τόσο τη χρήση και την κατανάλωση ενέργειας συνολικά, όσο και ανά συσκευή. Στο πλαίσιο των δικτύων ενέργειας, η εγκατάσταση προηγμένων δομών μέτρησης (Advanced Metering Infrastructure – AMI) έχει προσφέρει καλύτερη διαχείριση και γνώση επί των δικτύων μέσης και υψηλής τάσης.

Οι παραπάνω τεχνολογίες σε συνδυασμό με τεχνολογίες **μηχανικής μάθησης** (ML) για την αξιοποίηση της διαθέσιμης μεγάλου όγκου πληροφορίας έχει σαν αποτέλεσμα την αναβάθμιση των δικτύων ενέργειας σε «έξυπνα» ενεργειακά δίκτυα που προσφέρουν νέες οικονομικές ευκαιρίες για τον καταναλωτή και παραγωγό ενέργειας.

Η ρευματοκλοπή αποτελεί ένα πρόβλημα για τις εταιρείες παροχής ηλεκτρισμού. Όπως ορίζει και η λέξη, η ρευματοκλοπή είναι η κλοπή ρεύματος από κάποιους

χρήστες, ξεγελώντας το σύστημα ενέργειας, που έχει σαν συνέπεια τη μεταφορά του επιπλέον κόστους παραγωγής στους υπόλοιπους «τίμιους» χρήστες. Πέρα από το κόστος για τις εταιρείες και τους καταναλωτές που επωμίζονται τη ζημία, η απαιτούμενη αύξηση της παραγωγής ενέργειας έχει αρνητικές συνέπειες και στο περιβάλλον. Επιπλέον πληροφορίες για τη ρευματοκλοπή, την έκβασή της και τα αποτελέσματά της δίνονται στο Κεφάλαιο 2.

Καθώς μέχρι τώρα γίνεται χρήση μόνο των μηχανικών μετρητών, η κλοπή ρεύματος γίνεται εύκολα για όσους γνωρίζουν πώς να το κάνουν και δεν φοβούνται για τα πρόστιμα που μπορεί να τους επιβληθούν. Οι διάφοροι τρόποι με τους οποίους αυτό μπορεί να γίνει αναλύονται στην ενότητα 2.2. Με την ένταξη όμως του AMI και συγκεκριμένα των έξυπνων μετρητών τα δεδομένα αλλάζουν και ορισμένοι από τους γνωστούς τρόπους κλοπής αποφεύγονται. Επιθέσεις όμως στα ηλεκτρονικά συστήματα των μετρητών μπορούν ακόμα να συμβούν ή ακόμα και κάποιες παρεμβάσεις ώστε να καταγράφεται μικρότερη χρήση ενέργειας από την πραγματική. Ορισμένοι από τους ήδη υπάρχοντες τρόπους εντοπισμού της ρευματοκλοπής αναφέρονται στο Κεφάλαιο 3, όπως και τα χαρακτηριστικά ενός AMI.

Στο Κεφάλαιο 4 γίνεται η περιγραφή των αλγορίθμων μηχανικής μάθησης που μελετήθηκαν σε αυτή την εργασία και των δοκιμών που πραγματοποιήθηκαν. Το κύριο πρόβλημα προς επίλυση για αυτά τα πειράματα ήταν να εξεταστεί το πόσο καλά ένας αλγόριθμος μηχανικής μάθησης μπορεί να αναγνωρίσει ένα ημερήσιο προφίλ κλοπής ενέργειας ενός καταναλωτή. Ο βασικός αλγόριθμος που εξετάστηκε είναι ο XGBoost (eXtreme Gradient Boosting) ο οποίος τα τελευταία χρόνια έχει επιδείξει πολύ καλά αποτελέσματα σε προβλήματα classification και regression. Έγιναν επίσης δοκιμές για την σχετικά καινούργια μέθοδο αυτοματοποιημένης μηχανικής μάθησης (AutoML) για δυο καταναλωτές από δυο διαφορετικά σύνολα δεδομένων.

Συγκεκριμένα, για την αξιολόγηση των παραπάνω αλγορίθμων έγινε χρήση δύο συνόλων δεδομένων από την Ιρλανδία και το Ηνωμένο Βασίλειο, που περιλαμβάνουν

κατανάλωση κατοικιών για ένα διάστημα 1,5 έως 2 χρόνων περίπου. Η περιγραφή της προέλευσης, της δομής τους και του τρόπου επεξεργασίας τους γίνεται στην ενότητα 4.4.

Στην ενότητα 4.7 του Κεφαλαίου 4, γίνεται επίσης και μια σύγκριση των αποτελεσμάτων αλγορίθμων νευρωνικών δικτύων, που χρησιμοποιήθηκαν σε έρευνες της βιβλιογραφίας, τόσο μεταξύ τους όσο και με τα αποτελέσματα που προέκυψαν από την παρούσα εργασία.

Τέλος, γίνονται οι απαραίτητες παρατηρήσεις πάνω στα αποτελέσματα που προέκυψαν μετά την εφαρμογή τους, καθώς και πιθανές επεκτάσεις των ενεργειών που ελήφθησαν κατά τη διάρκεια των δοκιμών.

ΚΕΦΑΛΑΙΟ 2
ΡΕΥΜΑΤΟΚΛΟΠΗ

2.1 Τεχνικές και μη τεχνικές απώλειες (ρευματοκλοπή)

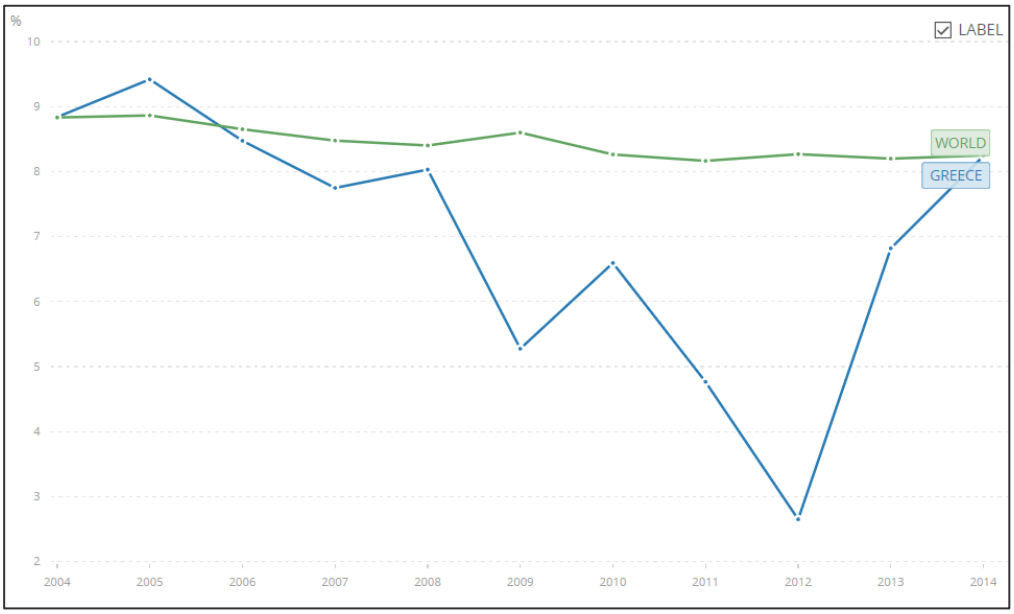
Ως απώλεια ενέργειας σε ένα Σύστημα Ηλεκτρικής Ενέργειας (ΣΗΕ) ορίζεται η διαφορά μεταξύ της ποσότητας ενέργειας που εγχέεται στο σύστημα και της ποσότητας που καταμετρείται στους τελικούς χρήστες. Το ποσοστό απωλειών στο σύστημα μεταφοράς και διανομής μπορεί να κυμαίνεται από 2% έως 50% ανάλογα με τη χώρα και το βιοτικό της επίπεδο. Π.χ. το 2016 στη Γερμανία οι απώλειες έφτασαν το 5%, ενώ στην Ινδία, όπου γίνονται πολλές έρευνες πάνω στο θέμα των απωλειών και της εύρεσης τρόπου αποφυγής τους, βρίσκονταν στο 19% [1].

Σύμφωνα με το report του 2020 από το Συμβούλιο Ευρωπαϊκών Ρυθμιστικών Αρχών Ενέργειας (Council of European Energy Regulators-CEER) σε έρευνα για τις απώλειες ενέργειας στην οποία πήραν μέρος 35 Ευρωπαϊκές χώρες απαντώντας σε ερωτηματολόγια που στάλθηκαν στις Εθνικές Ρυθμιστικές Αρχές κάθε χώρας, ως απώλειες θεωρούν τη διαφορά στη συνολική ενέργεια που εγχέεται σε ένα σύστημα μεταφοράς και κατ' επέκταση διανομής (ενέργεια όχι μόνο από την παραγωγή, αλλά και από την εισαγόμενη άλλων χωρών), και της συνολικής αφαιρούμενης ενέργειας από το σύστημα (όχι μόνο λόγω κατανάλωσης από τους χρήστες αλλά και λόγω εξαγωγής σε άλλες χώρες). Στην Ελλάδα οι απώλειες μεταφοράς υπολογίστηκαν ως περίπου το 2.8% το 2015, μειώθηκαν στο 2.35% το 2017 και αυξήθηκαν ξανά στο 2.65% το 2018. Αντίθετα οι απώλειες διανομής από το 9% το 2015, αυξήθηκαν το 2016 στο περίπου 10% και μειώθηκαν ξανά το 2017 στο 9.5% [2].

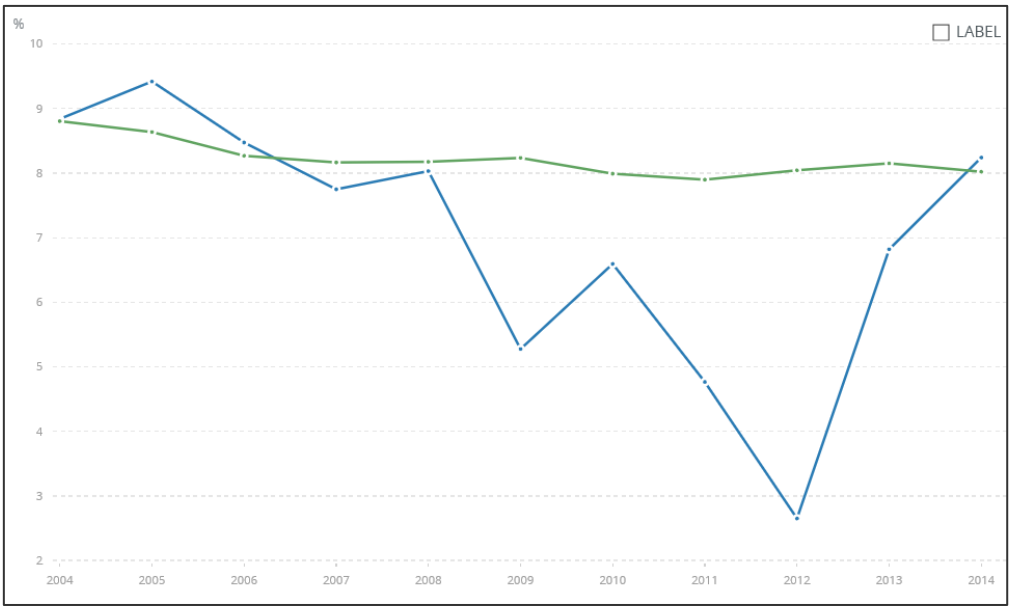
Στα Σχήματα 2.1 και 2.2 [3] φαίνεται η θέση της Ελλάδας την δεκαετία 2004 με 2014, τόσο σε σχέση με τον υπόλοιπο πλανήτη (Σχήμα 2.1), όσο και με τις Ευρωπαϊκές χώρες και χώρες της κεντρικής Ασίας (Σχήμα 2.2), όσον αφορά το ποσοστό απωλειών στο σύστημα μεταφοράς και διανομής ενέργειας σε σχέση με τη συνολική παραγόμενη ενέργεια σε GWh των σταθμών ηλεκτροπαραγωγής. Οι παρακάτω απώλειες περιλαμβάνουν απώλειες τόσο στη μετάδοση μεταξύ πηγών τροφοδοσίας

και σημείων διανομής, όσο και στη διανομή στους καταναλωτές, συμπεριλαμβανομένης και της κλοπής ενέργειας.

Όπως φαίνεται από τα Σχήματα 2.1 και 2.2, η Ελλάδα είχε σημαντική πτώση των απωλειών το 2012, ενώ μέχρι το 2014 έφτασε το παγκόσμιο ποσοστό απωλειών (8.2%) και ξεπέρασε για δύο δέκατα το ποσοστό της Ευρώπης και κεντρικής Ασίας.



Σχήμα 2.1: Απώλειες ηλεκτρικής ενέργειας (% της παραγωγής) – Ελλάδα και παγκόσμια



Σχήμα 2.2: Απώλειες ηλεκτρικής ενέργειας (% της παραγωγής) – Ελλάδα (μπλε) και Ευρώπη/Κεντρ. Ασία (πράσινο)

Οι απώλειες χωρίζονται σε δύο κατηγορίες, τεχνικές (technical losses) και μη τεχνικές απώλειες (non-technical losses)[4][5].

Οι τεχνικές απώλειες σχετίζονται με τις γραμμές μεταφοράς, με γεννήτριες και τους μετασχηματιστές του συστήματος, και σε αυτές περιλαμβάνονται:

- Ωμικές απώλειες (απώλειες joule σε γραμμή μεταφοράς)
- Απώλειες φαινομένου κορώνας σε γραμμή μεταφοράς
- Απώλειες σιδήρου σε μετασχηματιστές ισχύος (κενού φορτίου)
- Απώλειες χαλκού σε μετασχηματιστές ισχύος (υπό φορτίο)
- Άεργος ισχύς

Στις μη τεχνικές απώλειες υπάγονται τα παρακάτω:

- Κλοπή ρεύματος με παράκαμψη, παραβίαση μετρητή κ.ά.
- Σφάλμα στον υπολογισμό των τεχνικών απωλειών
- Μη μετρούμενη κατανάλωση
- Προβληματικός μετρητής
- Λάθος στην κοστολόγηση (σφάλμα στην καταγραφή και καθυστέρηση στην χρέωση)
- Καθυστερημένες ή μη πραγματοποιημένες πληρωμές από πελάτες
- «Κρυμμένες» απώλειες, όπως η επιπλέον χρήση ενέργειας για ψύξη μετασχηματιστών

Σε παγκόσμιο επίπεδο οι μη τεχνικές απώλειες κοστίζουν περίπου 96 δισεκατομμύρια δολάρια στις εταιρείες παροχής ενέργειας σύμφωνα με έρευνα του 2017 [6]. Μόνο στην Ινδία οι μη τεχνικές απώλειες κοστίζουν περίπου 23 δισεκατομμύρια δολάρια. Στην Ευρώπη το κόστος την ίδια περίοδο ανερχόταν σε 3.7 δισεκατομμύρια ευρώ [7]. Οι απώλειες αυτές μπορούν να οδηγήσουν σε παράλυση τις εταιρείες, να ανεβάσουν κατά πολύ τις τιμές του ρεύματος για τους πελάτες που είναι συνεπείς και πληρώνουν, να οδηγήσουν σε αυξανόμενη χρήση των περιορισμένων φυσικών πόρων άρα και μεγαλύτερη μόλυνση του περιβάλλοντος, να προκαλέσουν έλλειψη

πόρων για μελλοντικές επενδύσεις κεφαλαίου ή να εξαντλήσουν κυβερνητικές επιδοτήσεις που θα μπορούσαν να χρησιμοποιηθούν για εκσυγχρονισμό των υποδομών μιας χώρας.

Αναφορικά με το προαναφερθέν report του CEER το 2020 στην περίπτωση των μη τεχνικών απωλειών η Ελλάδα έκανε αναφορά για κρυμμένες απώλειες και άλλου είδους απώλειες στο σύστημα μεταφοράς, ενώ έκανε αναφορά για κρυμμένες και απώλειες λόγω κλοπής στο σύστημα μετάδοσης, και καθόλου για μη μετρημένη κατανάλωση. Στα πλαίσια βελτιώσεων στο σύστημα μέτρησης απωλειών, η Ελλάδα αύξησε τον αριθμό συνδεδεμένων φορτίων χαμηλής τάσης και των εγκαταστάσεων παραγωγής που χρησιμοποιούν προηγμένο σύστημα μέτρησης, οδηγώντας σε μια συνολική αύξηση της ποιότητας των δεδομένων που χρησιμοποιούνται για τον υπολογισμό των απωλειών.

Η ρευματοκλοπή όπως φαίνεται και από την παραπάνω λίστα αποτελεί κομμάτι των μη τεχνικών απωλειών και το μεγαλύτερο ποσοστό αυτών, στα συστήματα ηλεκτρικής ενέργειας, είναι η πιο «ακριβή» μη τεχνική απώλεια του συστήματος και μπορεί να έχει διάφορες μορφές. Αυτές αναφέρονται στο παρακάτω κεφάλαιο. Επίσημα και σύμφωνα με το Εγχειρίδιο Ρευματοκλοπών [8] ως ρευματοκλοπή ορίζεται εν γένει η αυθαίρετη και με δόλο επέμβαση σε εξοπλισμό ή / και εγκαταστάσεις του Δικτύου, με σκοπό την κατανάλωση ηλεκτρικής ενέργειας χωρίς αυτή να καταγράφεται, ή χωρίς να αντιστοιχίζεται με Εκπρόσωπο Φορτίου, και να μην τιμολογείται.

Στην Ελλάδα σύμφωνα με αναφορά του 2017 της Ρυθμιστικής Αρχής Ενέργειας (ΡΑΕ) για το έτος 2016 οι μη τεχνικές απώλειες ανέρχονταν στο 4.2%, ενώ σε αναφορά του 2018 για το έτος 2017 υπήρξε πτώση των μη τεχνικών απωλειών στο 3.2% συνολικής εισερχόμενης ενέργειας στο Διασυνδεδεμένο Δίκτυο. Όπως υπολογίστηκε από τη ΔΕΔΔΗΕ το κόστος των ρευματοκλοπών για το 2017 ήταν στα 80 εκατ. Ευρώ [9], ενώ με τα στοιχεία του 2018 το ποσοστό των ρευματοκλοπών σε σχέση με τη συνολική εισερχόμενη ενέργεια στο δίκτυο διανομής ήταν 4.1% με κόστος 139 εκατ. Ευρώ [10].

2.2 Είδη ρευματοκλοπής

Η ρευματοκλοπή μπορεί να έχει πολλές μορφές και όπως αναφέρεται και σε άρθρο του Εγχειριδίου Ρευματοκλοπών, έχει ως συνηθέστερη περίπτωση την επέμβαση στο μετρητή ή άλλο στοιχείο της μετρητικής διάταξης που αποσκοπεί στην αλλοίωση της καταγραφόμενης ενέργειας (καταγραφή μικρότερων ποσοτήτων έναντι της πραγματικής κατανάλωσης της εγκατάστασης). Στην περίπτωση αυτή υπάγεται η πλειονότητα των εντοπιζόμενων ρευματοκλοπών.

Άλλες περιπτώσεις ρευματοκλοπής όπως καταγράφονται στο Εγχειρίδιο [8] αποτελούν οι παρακάτω:

- Η παράκαμψη υφιστάμενου μετρητή (απ' ευθείας σύνδεση της γραμμής πίνακα-μετρητή με το καλώδιο παροχής της εγκατάστασης), οπότε και το σύνολο της καταναλισκόμενης ενέργειας δεν καταγράφεται (bypassing).
- Η απευθείας σύνδεση της εσωτερικής εγκατάστασης με το Δίκτυο, απουσία μετρητικού εξοπλισμού (σε περιπτώσεις που ο μετρητής έχει αποξηλωθεί ή και ουδέποτε εγκαταστάθηκε).
- Η απευθείας σύνδεση με αγκίστρωση στους αγωγούς του εναερίου δικτύου, απουσία μετρητή ή/και παροχής ή/και νομίμως υφιστάμενου κτίσματος (περιπτώσεις καταυλισμών κ.λπ.) (hooking).
- Η αυθαίρετη επανενεργοποίηση παροχών που έχουν απενεργοποιηθεί αλλά υπάρχει σύμβαση προμήθειας σε ισχύ (π.χ. απενεργοποίηση μετά από αίτημα Προμηθευτή λόγω υπερημερίας, παραβίαση όρων σύμβασης σύνδεσης από πελάτη, μη ανανέωση εργοταξιακής παροχής), με ή χωρίς αλλοίωση της μέτρησης.
- Η αυθαίρετη επανασύνδεση παροχών που έχουν διακοπεί κατόπιν αίτησης οικειοθελούς διακοπής από τον τελευταίο χρήστη ή κατόπιν υποβολής δήλωσης παύσης εκπροσώπησης από τον τελευταίο προμηθευτή (χωρίς νέα δήλωση εκπροσώπησης και δυνατότητα υπαγωγής στο καθεστώς προμήθειας καθολικής υπηρεσίας), με ή χωρίς παράκαμψη του μετρητή, οπότε και η

καταναλισκόμενη ενέργεια δεν τιμολογείται (παροχές χωρίς χρήστη και προμηθευτή), είτε αυτή καταγράφεται είτε όχι.

Σε επόμενο άρθρο του εγχειριδίου γίνεται διάκριση των ρευματοκλοπών σε δύο κατηγορίες, διαπιστωμένες και πιθανολογούμενες κλοπές.

Στις *διαπιστωμένες* συγκαταλέγονται οι παρακάτω περιπτώσεις:

1. Απευθείας σύνδεση στο Δίκτυο με παράκαμψη του εξοπλισμού μέτρησης ή και απουσία αυτού.
2. Αφαίρεση των σφραγίδων του κελύφους του μετρητή και παρεμπόδιση της περιστροφής του δίσκου μέσω παρεμβολής ξένου σώματος.
3. Απομόνωση πηνίων τάσεως-εντάσεως.
4. Διακοπή μιας ή δύο εκ των τριών διεγέρσεων σε τριφασικούς μετρητές
5. Ορατή επέμβαση στον μηχανισμό μέτρησης ή τον απαριθμητή του μετρητή
6. Παρέμβαση στους ακροδέκτες του μετρητή ή/και τοποθέτηση συσκευής βραχυκύκλωσής τους
7. Ανοικτές επαφές στο κιβώτιο δοκιμών
8. Μονωτικό υλικό στις επαφές του κιβωτίου δοκιμών
9. Διακοπές επαφών στις γέφυρες των μετασχηματιστών μέτρησης
10. Σύνδεση ουδετέρου στη γείωση
11. Παρέμβαση στη συνδεσμολογία της μέτρησης
12. Αλλαγή σχέσης μετασχηματιστών έντασης

Ως *πιθανολογούμενες* χαρακτηρίζονται οι περιπτώσεις κατά τις οποίες υπάρχουν μεν ευρήματα, ωστόσο αυτά αποτελούν ενδείξεις αλλά όχι αποδείξεις αλλοίωσης της μέτρησης, όπως ενδεικτικά η παραβίαση σφραγίδας του κελύφους του μετρητή, ενέργεια για την οποία δεν υπάρχει άλλος λόγος διάπραξής της, πέραν της επέμβασης στον μετρητή, σε αντίθεση με την κοπή της σφραγίδας του κιβωτίου του μετρητή που πιθανώς δικαιολογείται λόγω επανοπλισμού του μικρο-αυτόματου διακόπτη, τα οποία ευρήματα ενισχύονται και από δυσεξήγητη μεταβολή της καταναλωτικής συμπεριφοράς (μείωση της κατανάλωσης).

Στην ξένη βιβλιογραφία συναντάμε πιο συχνά τη μορφή του line hooking, δηλαδή την άμεση σύνδεση/αγκίστρωση στους αγωγούς του εναέριου δικτύου καθώς και της παράκαμψης του μετρητή με άμεση σύνδεση στο δίκτυο, το λεγόμενο bypassing. Στα υπόλοιπα είδη του meter tampering, δηλαδή της παραβίασης του μετρητή με διάφορες μεθόδους ώστε να γίνεται καταμέτρηση λιγότερης κατανάλωσης, συγκαταλέγονται οι παραπάνω περιπτώσεις 2-12 [11]. Επιπλέον, όπως θα δούμε και παρακάτω πιο λεπτομερώς, σε αυτές προστίθεται και η δημιουργία τρυπών στο κέλυφος του μετρητή, όπως και η χρήση μαγνήτη ή φωτογραφικής μεμβράνης (φιλμ) για τη μείωση της ταχύτητας του δίσκου του μετρητή. Ακόμη, έχει εντοπιστεί και η χρήση επιπλέον κυκλώματος, το οποίο ενεργοποιείται απομακρυσμένα για να εμποδίσει ή να καθυστερήσει την καταγραφή της κατανάλωσης.

Σε μια συγκεντρωτική και αρκετά λεπτομερή έρευνα για τα είδη ρευματοκλοπής [12], τα τελευταία χωρίζονται σε τρεις κατηγορίες: α) κλασσική απ' ευθείας σύνδεση με τη γραμμή, β) τροποποίηση της λειτουργίας με βραχυκύκλωμα ή παραβίαση των φυσικών μηχανισμών ασφάλειας του μετρητή, γ) τροποποίηση της λειτουργίας του μετρητή με αλλοίωση της εσωτερικής μνήμης του τσιπ της μητρικής του πλακέτας. Στις παραπάνω κατηγορίες βέβαια όπως αναφέρουν και οι ίδιοι, αλλά έχει γίνει αναφορά και σε άλλες έρευνες, μπορεί να προστεθεί και η «εκ των έσω απάτη», δηλαδή η συνεργασία με κάποιον υπάλληλο της εταιρείας παροχής ηλεκτρικής ενέργειας που έχει πρόσβαση και μπορεί να τροποποιήσει τα δεδομένα κατανάλωσης.

Σύμφωνα με την έρευνα οι πιο συνήθεις τρόποι είναι οι παρακάτω και περιλαμβάνουν τους περισσότερους από τους ήδη αναφερθέντες τρόπους κλοπής:

1. Κρυφή πρόσθετη εγκατάσταση στη γραμμή παρακάμπτοντας το σύστημα μέτρησης. Είναι η πιο συνήθης μορφή όπως αναφέρουν, και συναντάται κυρίως σε περιοχές κατοίκων χαμηλού εισοδήματος όπου οι πολυκατοικίες και οι λοιπές κατοικίες είναι αρκετά παλιές και επιτρέπουν τις παράνομες τροποποιήσεις. Στη βασική εκδοχή αυτού του είδους κλοπής γίνεται μόνο χρήση μιας επιπλέον πηγής παροχής ηλεκτρικής ενέργειας που είναι όμως

αρκετά εύκολο να εντοπιστεί. Στις πιο εξελιγμένες μεθόδους γίνεται μερική χρήση της εξωτερικής πηγής οπότε και ο εντοπισμός γίνεται πιο δύσκολος. Η κλοπή αυτή συμβαίνει κυρίως εποχιακά, όπως το χειμώνα για θέρμανση.

2. Άμεση παραβίαση των ακροδεκτών του μετρητή ή αλλιώς γεφύρωση. Στην περίπτωση αυτή κάποιοι που είναι γνώστες των ηλεκτρολογικών κυκλωμάτων και της μηχανικής μπορούν να αποσυνδέσουν το καλώδιο του ουδετέρου (N), προσέχοντας να μην επιτραπεί η γαλβανική σύνδεση του ουδέτερου του μετρητή με αυτόν της εγκατάστασης. Έτσι μπορεί να υπάρξει ασύμμετρη κατανάλωση ενέργειας, εφόσον η τιμή του ρεύματος στις τρεις φάσεις (για τριφασική εγκατάσταση) θα είναι διαφορετική από αυτή του ουδετέρου (δηλ. της επιστροφής) και άρα υποτίμηση της τιμής του ρεύματος που καταναλώνεται.
3. Φυσική παραβίαση του μηχανισμού ενός αναλογικού μετρητή. Αυτό μπορεί να υλοποιηθεί με τρεις τρόπους:
    a. Απομαγνητισμό με τη χρήση μαγνητών νεοδημίου, δημιουργώντας ισχυρό ηλεκτρομαγνητικό πεδίο με αποτέλεσμα την επιβράδυνση του δίσκου του μετρητή μέχρις ότου να σταματήσει.
    b. Χρήση φωτογραφικής μεμβράνης. Σε αυτή τη μέθοδο εισάγεται μεταξύ του πίσω καλύμματος του μετρητή και του μπροστινού γυαλιού φωτογραφικό φιλμ, που λόγω των ιδιοτήτων του (είναι στενό και ευέλικτο) μπορεί να περνάει εύκολα μέσα από στενά και καμπυλωτά σημεία του μετρητή, με αποτέλεσμα να τυλίγεται μέσα σε αυτόν και να προκαλεί μέχρι και τελειωτικό σταματημό της κίνησης του δίσκου. Ο συγκεκριμένος τρόπος κλοπής δεν χρειάζεται το άνοιγμα του μετρητή και έτσι δεν εντοπίζεται εύκολα καθώς δεν υπάρχουν εμφανείς αποδείξεις.
    c. Τρύπημα περιβλήματος, όπου ουσιαστικά δημιουργείται μια πολύ μικρή διακριτική τρύπα, που δεν είναι εύκολα ορατή από τον ελεγκτή κατανάλωσης, με διάμετρο λιγότερο από 1 χιλιοστό είτε με κλασσικό

τρυπάνι είτε με πυρωμένη καρφίτσα, και έπειτα μπορεί να εισαχθεί από κει κάποιο ελαστικό ή άλλου είδους αντικείμενο που θα εμποδίζει την κίνηση του μετρητικού δίσκου.

4. Παραβίαση του λογισμικού ψηφιακού μετρητή. Σε αυτό το είδος κλοπής, γίνονται αλλαγές σε μεταβλητές του μετρητή που αποθηκεύουν τιμές που σχετίζονται με τον τρόπο που γίνεται η κοστολόγηση από τον προμηθευτή ηλεκτρικής ενέργειας. Η μέθοδος αυτή βέβαια λόγω του ότι απαιτεί παραπάνω χρόνο και πιο ειδικές γνώσεις, συναντάται λιγότερο συχνά από τις παραπάνω.

2.3 Τεχνικές αποφυγής ρευματοκλοπής

Η ρευματοκλοπή αν και αποτελεί ποινικό αδίκημα φαίνεται να μην εμποδίζει τις προσπάθειες των επιτήδειων. Παραπάνω φάνηκε πως οι τρόποι με τους οποίους μπορεί να επιτύχει κάποιος κλοπή ρεύματος είναι πολλοί και δίνουν πολλές επιλογές σε αυτούς που επιθυμούν να ξεγελάσουν το σύστημα μετρήσεων.

Οι εταιρείες παροχής ενέργειας δεν έχουν βρει ακόμα τρόπους που να εμποδίζουν αποτελεσματικά και τελειωτικά την εφαρμογή των παραπάνω μεθόδων. Μπορούν να αποτρέψουν πιθανές κλοπές από το να συμβούν, με αύξηση των προστίμων και ποινές φυλάκισης αν εντοπιστεί και αποδειχθεί κλοπή, αυτό δεν σημαίνει όμως ότι δίνεται μια τελική λύση.

Οι αναλογικοί μετρητές φαίνεται να είναι το σημαντικό εμπόδιο πλέον. Τα τελευταία χρόνια έχει αρχίσει να γίνεται χρήση ψηφιακών/ηλεκτρονικών «έξυπνων» μετρητών, οι οποίοι κάνουν συνεχή ηλεκτρονική καταγραφή της κατανάλωσης και επιτρέπουν τόσο την παρακολούθηση των καταγραφών 24 ώρες το 24ωρο καθώς και τον έλεγχο της σύνδεσης, απομακρυσμένα. Σε αυτή την περίπτωση μπορεί να αποφευχθεί η ρευματοκλοπή είτε με το φόβο της συνεχούς παρακολούθησης η οποία μπορεί να την αποκαλύψει γρηγορότερα, είτε με συστήματα πρόληψης, τα οποία χρησιμοποιώντας ενσωματωμένα κυκλώματα με τη δυνατότητα απομακρυσμένου ελέγχου των μετρητών και των δεδομένων πραγματικού χρόνου (real time data), διακόπτουν την ηλεκτροδότηση σε περίπτωση εντοπισμού ύποπτης κατανάλωσης.

Η γενική ιδέα στην αποφυγή ρευματοκλοπής είναι συστήματα τα οποία αποτελούνται κυρίως από μονάδες αισθητήρων (sensors), ελεγκτών (controllers) και επικοινωνίας (communication units). Οι αισθητήρες είναι αυτοί που λαμβάνουν την οποιαδήποτε ανωμαλία ή παραβίαση, στέλνουν την αντίστοιχη πληροφορία στη μονάδα ελέγχου κι από κει, ανάλογα το σύστημα, είτε στέλνεται απευθείας εντολή για διακοπή της ροής ρεύματος και μετά ενημερώνεται η εταιρεία παροχής για την κλοπή, είτε πρώτα ενημερώνεται μέσω της μονάδας επικοινωνίας η εταιρεία για την ανωμαλία κι έπειτα δίνεται η εντολή για απομακρυσμένη διακοπή. Η μονάδα ελέγχου μπορεί να είναι για παράδειγμα μια πλακέτα Arduino που πραγματοποιεί όλους τους απαραίτητους υπολογισμούς με βάση τις τιμές που λαμβάνει και από τους αισθητήρες, και ύστερα δίνει εντολές στις υπόλοιπες μονάδες [13].

Μερικά παραδείγματα για τους παραπάνω τρόπους αποφυγής της κλοπής αποτελούν προτάσεις, κυρίως από χώρες όπως η Ινδία όπου το πρόβλημα είναι ιδιαίτερα έντονο:

Στην περίπτωση των αναλογικών/ηλεκτρομαγνητικών μετρητών, μια πρόταση είναι η τοποθέτηση ενός συστήματος φωτοδιόδου με ένα λαμπτήρα IR Led (infrared Led) [14]. Η φωτοδίοδος τοποθετείται στον άξονα περιστροφής του δίσκου και φωτίζεται με IR φως από το λαμπτήρα, στέλνοντας ένα χαμηλό λογικό σήμα σε ένα μικροελεγκτή. Αν υπάρξει παραβίαση στην κίνηση του δίσκου ή αφαιρεθεί το κάλυμμα του μετρητή δημιουργείται ένα εμπόδιο μεταξύ του λαμπτήρα και της διόδου και το σήμα που λαμβάνει ο μικροελεγκτής είναι υψηλό, οπότε λόγω της απότομης αλλαγής του σήματος γίνεται ο εντοπισμός της ανωμαλίας, που στέλνεται μέσω ενός GSM (global system for mobile communication) modem ενημερώνοντας την εταιρεία για την παραβίαση και λαμβάνεται η απόφαση για διακοπή της σύνδεσης.

Όσον αφορά στους ηλεκτρονικούς ή έξυπνους μετρητές φαίνεται να υπάρχουν παραπάνω λύσεις. Μια από αυτές αποτελεί μια απλή περίπτωση όπου οι καταγραφές του φορτίου, όπως γίνονται από τον μετρητή, μεταφέρονται μέσω ασύρματης σύνδεσης σε αποδέκτη που βρίσκεται στην πλευρά του συστήματος διανομής. Εκεί

γίνεται η σύγκριση μεταξύ των τιμών του φορτίου όπως έχει αποσταλεί από το μετρητή και του φορτίου που έχει διανεμηθεί από το σύστημα. Αν υπάρχει μια διαφορά στα φορτία πάνω από το ποσοστό που οριοθετούν οι λογικές απώλειες του συστήματος, τότε εξετάζεται η περίπτωση ρευματοκλοπής και ακολουθεί διακοπή της ροής ρεύματος στη συγκεκριμένη γραμμή. Η παραπάνω on-time ειδοποίηση μπορεί να συμβεί και με την τοποθέτηση αντίστοιχων αισθητήρων όπως και στην προηγούμενη περίπτωση, όταν γίνει κάποια φυσική παρεμβολή στη λειτουργία του μετρητή [15]. Σε άλλες περιπτώσεις το παραπάνω σύστημα μπορεί να διαθέτει λογισμικό που αναλύει τα δεδομένα προηγούμενης κατανάλωσης, λαμβάνοντας υπόψη και τις απώλειες της γραμμής, προειδοποιώντας την εταιρεία ώστε να ακολουθήσει ο απαραίτητος έλεγχος.

Μια άλλη πρόταση έχει ως αρχή την παραπάνω διαδικασία, δηλαδή τη συλλογή δεδομένων κατανάλωσης και μετρούμενων τιμών όπως ενεργό τάση και ένταση, και αποστολή τους σε απομακρυσμένη μονάδα/δέκτη [16]. Στην περίπτωση όμως που εντοπιστεί ανωμαλία στο σύστημα εν συγκρίσει με τις απώλειες διανομής, δεν διακόπτεται απλά η ροή του ρεύματος στην εντοπισμένη γραμμή. Δίνεται ένα σήμα στους μετρητές των καταναλωτών που έχουν αναγνωριστεί νωρίτερα ως νόμιμοι και αποκλείονται από το σύστημα για ένα διάστημα, όπου μόνο οι δράστες είναι συνδεδεμένοι. Στην μικρή αυτή περίοδο εγχέεται στο σύστημα μια αρμονική διαταραχή από μια γεννήτρια αρμονικών (harmonic generator: τέσσερα thyristors και μια γεννήτρια παλμών για την πυροδότηση των πυλών των thyristors) η οποία μπορεί να προκαλέσει βλάβη στις συσκευές των παραβατών. Μετά από αυτό το διάστημα δίνεται σήμα από την απομακρυσμένη μονάδα επικοινωνίας για αποκατάσταση της παροχής του ρεύματος σε όλη τη γειτονιά. Για την υλοποίηση αυτή χρησιμοποιείται ένας αρμονικός αισθητήρας (harmonic sensor) από την πλευρά του έξυπνου μετρητή για την ανίχνευση επιπλέον αρμονικής συνιστώσας του ρεύματος, επιπρόσθετα στην συνολική αρμονική διαταραχή της παρεχόμενης ηλεκτρικής ενέργειας. Αυτή η πρόσθετη τιμή της αρμονικής συνιστώσας υπολογίζεται από τη διαφορά μεταξύ του συνολικού τροφοδοτούμενου ρεύματος φάσης και της

θεμελιώδους συνιστώσας του ρεύματος με βάση τη θεωρία της στιγμιαίας άεργης ισχύος.

Μια επιπλέον λύση παρουσιάζεται στο [17]. Στην πλευρά του μετασχηματιστή διανομής ισχύος της παροχής ενέργειας ενσωματώνεται σύστημα μεταγωγής (switching system) με ημιαγωγούς υψηλής ισχύος, ώστε οι τρεις φάσεις του δικτύου και ο ουδέτερος να περνούν μέσα από αυτό. Ο μικροελεγκτής που υπάρχει στο σύστημα μεταγωγής παράγει ανά διαστήματα μια ακολουθία bit η οποία προκαλεί μια εναλλαγή των 4 αγωγών, με τέτοιο τρόπο ώστε σε κάθε χρονικό διάστημα η διάταξη των αγωγών να μας είναι άγνωστη. Η ίδια ακολουθία είναι απαραίτητο να παραχθεί και στην πλευρά του μετρητή-δέκτη μέσω επικοινωνίας πομπού-δέκτη RF. Οπότε όποιος προσπαθήσει να παρεμβάλει τους αγωγούς ή γειώσει τον ουδέτερο μπορεί να προκαλέσει βλάβη, δημιουργώντας βραχυκύκλωμα ή γειώνοντας μια φάση κ.ά. Το σύστημα αυτό, πέρα από τους ηλεκτρονικούς μετρητές, προτείνεται να εφαρμοστεί με μικρές προσθήκες κυκλωμάτων και στο ήδη υπάρχον συμβατικό σύστημα της Ινδίας.

Οι παραπάνω λύσεις αποτελούν τόσο μεθόδους αποφυγής αλλά και εντοπισμού της ρευματοκλοπής ταυτόχρονα. Αυτό που προσφέρουν είναι η προστασία των καταναλωτών που ακολουθούν τη νόμιμη και διαφανή οδό και αποθαρρύνουν τους παραβάτες, υιοθετώντας μεθόδους που κάνουν χρήση όλων των δεδομένων και τιμών στο σύστημα διανομής για υπολογισμούς σε πραγματικό χρόνο για πιο άμεση αντίδραση. Στο επόμενο κεφάλαιο θα αναφερθούν κι άλλοι μέθοδοι εντοπισμού στα πλαίσια αξιοποίησης των έξυπνων μετρητών.

ΚΕΦΑΛΑΙΟ 3
ΕΝΤΟΠΙΣΜΟΣ ΡΕΥΜΑΤΟΚΛΟΠΗΣ

Στο ήδη υπάρχον δίκτυο ηλεκτρικής ενέργειας που αποτελείται από συμβατικούς αναλογικούς-μηχανικούς μετρητές, ο εντοπισμός της ρευματοκλοπής μπορεί να γίνει με περιορισμένες τεχνικές. Σύμφωνα και με το Εγχειρίδιο Ρευματοκλοπών όπως έχει δημοσιευτεί στο ΦΕΚ, αυτοί οι μέθοδοι περιλαμβάνουν κυρίως προγραμματισμένους ή μη, ελέγχους του ειδικού τεχνικού προσωπικού του ΔΕΔΔΗΕ σε περιπτώσεις τακτικών καταμετρήσεων ή έκτακτων μετρήσεων όταν π.χ. γίνει αλλαγή χρήστη ή ζητηθεί διακοπή της παροχής [8]. Επίσης όταν παρατηρηθεί παραβίαση της εγκατάστασης του μετρητή ή σημαντική μεταβολή στην κατανάλωση ενέργειας χωρίς την αλλαγή χρήστη, πραγματοποιούνται στοχευμένοι έλεγχοι πάλι από το τεχνικό προσωπικό του ΔΕΔΔΗΕ, όπως ακόμη και μετά από εργασίες συντήρησης του εξοπλισμού ή από τεχνικούς ή από συνεργαζόμενους εργολάβους, ή αν γίνει κάποια καταγγελία από τρίτους.

Γίνεται λοιπόν αντιληπτό ότι οι παραπάνω τρόποι χρειάζονται τόσο το τεχνικό προσωπικό να είναι σε ετοιμότητα σε κάθε περίπτωση ελέγχου και να διενεργούν λεπτομερή καταγραφή κατά τη σύνταξη της αναφοράς τους, αλλά επίσης μπορεί να απαιτηθεί και ένα σημαντικό χρονικό διάστημα μέχρι να γίνει ο εντοπισμός, η αξιολόγηση και αργότερα αντιμετώπιση της κάθε πιθανής περίπτωσης κλοπής, με αποτέλεσμα τόσο τις οικονομικές συνέπειες για τους παρόχους ενέργειας και τους καταναλωτές, όταν οι τελευταίοι είναι οι ίδιοι τα θύματα της κλοπής ή όταν υφίστανται το επιπλέον κόστος στην τιμολόγηση λόγω εξισορρόπησης της ζημίας της εταιρείας, όσο και τις περιβαλλοντικές, λόγω της επιπλέον καύσης για παραγωγή όταν η ενέργεια δεν προέρχεται από ΑΠΕ. Οι ηλεκτρονικοί-έξυπνοι μετρητές όπως έχει αναφερθεί και νωρίτερα έρχονται για να αλλάξουν το τοπίο και να φέρουν επιπλέον λύσεις.

3.1 Ηλεκτρονικοί-Έξυπνοι Μετρητές

Ο τελευταίος τρόπος εντοπισμού που αναφέρεται και στο εγχειρίδιο εντοπισμού ρευματοκλοπής αφορά συμβάντα που καταγράφονται και δεδομένα που συλλέγονται μέσω τηλεμέτρησης που προσφέρουν οι ηλεκτρονικοί μετρητές και εν συνεχεία εξετάζονται με τεχνικό έλεγχο από το προσωπικό.

Σε συνέντευξη του κύριου Χατζηαργυρίου στο ΑΠΕ-ΜΠΕ τον Ιούνιο του 2018 [18], μετά κι από τις πρώτες εγκαταστάσεις έξυπνων μετρητών στην Ελλάδα που έγιναν το 2017, σημείωσε πως δεν είχαν εντοπιστεί δείγματα ρευματοκλοπών μετά από ελέγχους που έγιναν στους καταναλωτές που έχει γίνει η εγκατάσταση των έξυπνων μετρητών. Η πρώτη φάση των δοκιμών των μετρητών περιλάμβανε την εγκατάσταση μετρητών από το ΔΕΔΔΗΕ σε 13.000 πελάτες μέσης τάσης καθώς και σε 74.000 μεγάλους πελάτες χαμηλής τάσης που αντιπροσώπευαν από κοινού το 36% της συνολικής κατανάλωσης ενέργειας. Το επόμενο βήμα αποτελούσε η προκήρυξη εγκατάστασης του νέου Συστήματος Τηλεμέτρησης και Επεξεργασίας Μετρητικών Δεδομένων (Κύριου και Εφεδρικού) με δυναμικότητα επικοινωνίας 7.500.000 μετρητικών σημείων, που θα καλύπτει δηλαδή το σύνολο των καταναλωτών ρεύματος. Αυτό το έργο από το 2018 έχει κολλήσει στη διαδικασία και πλέον όπως αναφέρεται σε άρθρο του Σεπτεμβρίου 2020 της energy press [19], ο ΔΕΔΔΗΕ προχωρεί σε διαγωνισμό του πρότζεκτ για την προμήθεια κι εγκατάσταση των «ευφυών» συστημάτων καταμέτρησης της κατανάλωσης ηλεκτρικής ενέργειας που περιλαμβάνει την αντικατάσταση των 7.5 εκατομμυρίων μετρητών ανοίγοντας το δρόμο στους προμηθευτές ρεύματος αλλά και τον Διαχειριστή για την παροχή σύγχρονων και αυτόματων υπηρεσιών προς τους καταναλωτές.

3.1.1 Λειτουργία Έξυπνων Μετρητών

Οι έξυπνοι μετρητές πρώτα από όλα προσφέρουν την συνεχή καταγραφή της ενέργειας που καταναλώνεται από το χρήστη η οποία έπειτα διατίθεται μέσω ανάλογων πλατφορμών τόσο στο χρήστη για να μπορεί να έχει πρόσβαση στα δεδομένα κατανάλωσής του όποτε χρειαστεί, όσο και στην εταιρεία παροχής [20].

Στην περίπτωση του χρήστη όλα τα στοιχεία που προσφέρονται από το μετρητή γίνονται διαθέσιμα μέσω μια ασύρματης οθόνης στο σπίτι (In Home Display). Για τη δεύτερη περίπτωση, π.χ. στο Ηνωμένο Βασίλειο όλες οι πληροφορίες αποστέλλονται σε ένα κόμβο επικοινωνίας εσωτερικά του μετρητή κι από κει διανέμονται στην εταιρεία επικοινωνίας που είναι η Data and Communications Company (DCC) και τα δεδομένα γίνονται διαθέσιμα για οποιονδήποτε από τους χρήστες της υπηρεσίας DCC (πάροχοι ενέργειας και διαχειριστές του δικτύου διανομής) [21]. Ο τελικός χρήστης (καταναλωτής) μπορεί σε πραγματικό χρόνο να γνωρίζει πόση ενέργεια καταναλώθηκε τόσο συνολικά μέχρι εκείνη τη χρονική στιγμή όσο και από ποιες συσκευές, όπως επίσης και ποια είναι η μέχρι τότε κοστολόγηση του ρεύματος. Επίσης από την πλευρά της η εταιρεία μπορεί να στείλει απομακρυσμένα δεδομένα και εντολές ελέγχου στο μετρητή, όπως αναφέρθηκε και νωρίτερα, χτίζοντας έτσι ένα αμφίδρομο τρόπο επικοινωνίας (bi-directional communication) μεταξύ του μετρητή του χρήστη και ενός κεντρικού υπολογιστή από την πλευρά της εταιρείας [22].

Αν στην ίδια κατοικία γίνεται και χρήση αερίου για κεντρική θέρμανση (κυρίως στη Μεγάλη Βρετανία), τότε υπάρχει και ένας ακόμα έξυπνος μετρητής για την μέτρηση των αντίστοιχων τιμών. Οι δυο αυτοί μετρητές επικοινωνούν με τον κόμβο επικοινωνίας που είναι συνήθως εγκατεστημένος στον μετρητή ηλεκτρικής ενέργειας, ή μπορεί να είναι και διαφορετικός για τους δυο μετρητές, αν η εγκατάσταση αυτού του αερίου γίνει νωρίτερα από αυτόν του ηλεκτρισμού. Παρόλα αυτά όλα τα δεδομένα για τους δυο μετρητές είναι διαθέσιμα μέσω του in home display.

Οι βασικές τιμές που παρακολουθούνται και καταγράφονται από τον έξυπνο μετρητή είναι πρώτα από όλα η τάση, το ρεύμα, η ενεργός και άεργος ισχύς, ο συντελεστής ισχύος και η συχνότητα. Επιπλέον γίνεται παρακολούθηση των απωλειών τάσης ή υπέρταση, αρμονικές τάσεων και ρεύματος, ασυμμετρία τάσεων, απώλειες φάσης, απώλειες ρεύματος, ανισορροπία τάσης και ρεύματος, παλιρροιακό ρεύμα αναστροφής, ανάποδη μέτρηση ουδετέρου [22][23].

Παρακάτω στην επόμενη ενότητα θα αναφερθούν οι πιθανοί τρόποι συνδέσεων των μετρητών σε δίκτυα σε τοπικό επίπεδο όσο και σε ευρύτερες περιοχές.

Τα πλεονεκτήματα και τα οφέλη χρήσης των έξυπνων μετρητών όπως αναφέρονται και σε πηγές της βιβλιογραφίας μπορούν να είναι τα παρακάτω [20][24][25]:

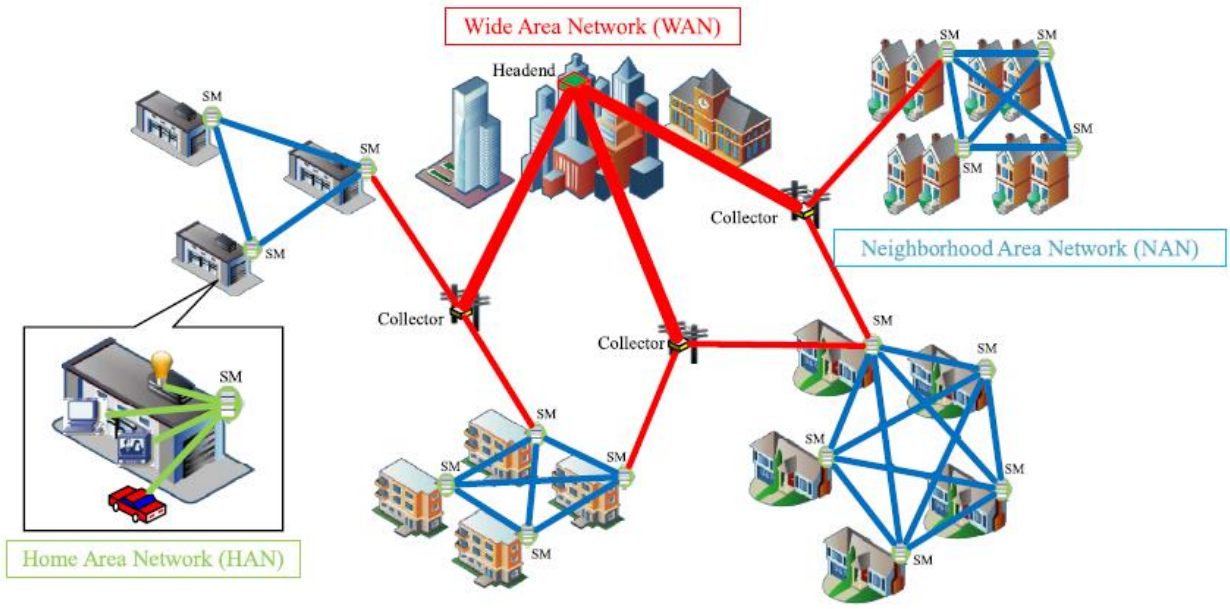
- Γίνεται πιο ακριβής μέτρηση, χωρίς να χρειάζεται η «έναντι» τιμολόγηση
- Δεν απαιτείται η παρουσία ειδικού προσωπικού για τη διακοπή της ηλεκτροδότησης σε οποιαδήποτε περίπτωση (είτε λόγω αλλαγής χρήστη, ή ακύρωσης λογαριασμού, ή ύποπτης κατανάλωσης)
- Δεν υπάρχει δυνατότητα επιβράδυνσης του δίσκου (δεν υπάρχει μηχανικό στέλεχος στο μετρητή) ή διακοπή της μέτρησης
- Συνεχής παρακολούθηση της κατανάλωσης και της λειτουργίας του μετρητή, οπότε υπάρχει και άμεση αντιμετώπιση οποιουδήποτε τεχνικού προβλήματος ή εξωτερικής παρέμβασης
- Θα μπορούν οι προμηθευτές ενέργειας να παρέχουν προγράμματα κοστολόγησης βάσει του χρόνου χρήσης (Time of Use) με ειδικά χρονομεταβλητά τιμολόγια
- Οι καταναλωτές θα μπορούν με βάση το πρόγραμμα τιμολόγησής τους και της παρακολούθησης της κατανάλωσής τους να προγραμματίζουν τη χρήση συσκευών σε ώρες που το κόστος είναι χαμηλότερο αλλά να γίνεται και εξοικονόμηση ενέργειας καθώς αυτό συμφέρει και τον προμηθευτή αφού μειώνεται το φορτίο σε ώρες αιχμής όπως και οι εκπομπές διοξειδίου
- Μπορεί να υπάρξει και η δυνατότητα για προπληρωμένη «κάρτα» ενέργειας

3.1.2 Προηγμένη Δομή Μέτρησης

Η προηγμένη δομή μέτρησης (Advanced Metering Infrastructure – AMI) αποτελεί σημαντικό κομμάτι των ευφυών συστημάτων (smart grids) ως σύστημα μέτρησης, συλλογής και ανάλυσης των δεδομένων κατανάλωσης ενέργειας. Βασικό τμήμα του AMI αποτελούν οι έξυπνοι μετρητές μέσω των οποίων πραγματοποιούνται οι παραπάνω λειτουργίες. Αυτό που διαφοροποιεί το AMI από το AMR (automatic

meter reading) είναι η bidirectional επικοινωνία μεταξύ των μετρητών των καταναλωτών και των παρόχων ενέργειας όπως περιγράφηκε στην προηγούμενη ενότητα. Σε ένα AMI μπορούν να περιλαμβάνονται επίσης μετρητές αερίου, ηλιακής ενέργειας, θερμότητας ή νερού. Στην προηγμένη δομή ανήκουν στην ουσία τα συστήματα λογισμικού, hardware, επικοινωνιών, ελέγχου, οθονών πληροφόρησης, διαχείρισης δεδομένων και επιχειρηματικών πλάνων των προμηθευτών [11][26].

Όσον αφορά τους διάφορους τρόπους επικοινωνίας, οι έξυπνοι μετρητές σε πρώτο στάδιο συνδέονται σε ένα δίκτυο περιοχής κατοικίας (Home Area Network – HAN) σχετικά μικρής εμβέλειας και σχετίζεται με την επικοινωνία των διάφορων μετρητών μεταξύ τους όπως επίσης και με τον κεντρικό κόμβο επικοινωνίας. Οι τρόποι σύνδεσης σε αυτό το δίκτυο μπορεί να είναι είτε ενσύρματοι (π.χ. Ethernet, επικοινωνίας γραμμής ισχύος – power line communication – PLC) είτε ασύρματοι (π.χ. Wi-Fi, Zig-Bee, wireless ad hoc network). Σε επόμενη φάση κάθε δίκτυο HAN συνδέεται σε ένα δίκτυο περιοχής γειτονίας (Neighborhood Area Network – NAN). Στο NAN συνδέονται γειτονικοί μετρητές και εκεί υπάρχει επίσης ένας συλλέκτης δεδομένων που τα ενώνει και τα διανέμει στο επόμενο επίπεδο επικοινωνίας. Οι κύριοι τρόποι σύνδεσης σε ένα NAN είναι ασύρματοι όπως το Wi-Fi, το WiMAX, ή το σύστημα τηλεφωνίας 3G/4G. Στο ιεραρχικό σύστημα επικοινωνίας επόμενο είναι το δίκτυο ευρείας περιοχής (Wide Area Network – WAN). Σε ένα WAN γίνεται η σύνδεση όλων των επιμέρους συλλεκτών δεδομένων των NAN ή απευθείας κάποιων έξυπνων μετρητών και η μεταφορά των δεδομένων στα κεντρικά συστήματα της εταιρείας παροχής ενέργειας. Για το λόγο αυτό πρώτη επιλογή στον τρόπο σύνδεσης αποτελούν οι οπτικές ίνες για σταθερότητα, μεγαλύτερο bandwidth και αξιοπιστία, και έπειτα προτείνονται τα δίκτυα τηλεπικοινωνιών, PLC, ραδιοσυχνοτήτων ή δορυφορικά. Ένα απλό AMI απεικονίζεται στην Εικόνα 3.1 [27].



Εικόνα 3.1: Μια απλή προηγμένη δομή μέτρησης (AMI)

3.2 Μέθοδοι εντοπισμού

Ο εντοπισμός της ρευματοκλοπής έχει προβληματίσει αρκετά και για πολύ καιρό τις εταιρείες παροχής ενέργειας. Οι μέθοδοι που προτείνονται κάθε χρόνο από την επιστημονική κοινότητα πληθαίνουν συνεχώς [28]. Με βάση και την κατηγοριοποίηση που έκαναν σε έρευνά τους οι κύριοι Μεσσίνης και Χατζηαργυρίου [29], για την παρούσα εργασία χωρίζουμε σε πρώτο στάδιο τις μεθόδους σε τρία βασικά ήδη και από κει και πέρα αναφέρουμε παραδείγματα από τη βιβλιογραφία που έχουμε βρει. Τις μεθόδους που είναι προσανατολισμένες στα δεδομένα (data oriented methods), αυτές που είναι στο δίκτυο (network oriented methods) και τις υβριδικές (hybrid methods) που συνδυάζουν χαρακτηριστικά και από τις δυο παραπάνω κατηγορίες.

Data Oriented Methods: Στην περίπτωση των data oriented μεθόδων γίνεται χρήση δεδομένων που σχετίζονται με τον καταναλωτή, όπως η κατανάλωση ενέργειας, ο τύπος του καταναλωτή, ο τρόπος της κατανάλωσης και άλλα.

Στις data oriented μεθόδους μπορούμε να συμπεριλάβουμε τη χρήση αλγορίθμων μηχανικής μάθησης (machine learning) καθώς και αυτούς της θεωρίας παιγνίων

(game theory). Οι αλγόριθμοι μηχανικής μάθησης που εντοπίστηκαν στα πλαίσια της εργασίας θα αναφερθούν σε παρακάτω ενότητα πιο λεπτομερώς.

Στις μεθόδους θεωρίας παιγνίων υπονοείται ότι υπάρχει ένα είδους παιχνίδι μεταξύ του παραβάτη και του παρόχου ενέργειας ή των καταναλωτών. Στην ουσία ο παραβάτης προσπαθεί να κλέψει μια προκαθορισμένη ποσότητα ενέργειας καθώς μειώνει την πιθανότητα εντοπιστεί, ενώ η εταιρεία παροχής προσπαθεί να αυξήσει την πιθανότητα εντοπισμού του παραβάτη αλλά και του λειτουργικού κόστους που θα επιβαρυνθεί εκείνος για τη διαχείριση του μηχανισμού εντοπισμού [27].

Network Oriented Methods: Οι μέθοδοι που προσανατολίζονται στο δίκτυο ασχολούνται με δεδομένα που λαμβάνουν από αισθητήρες και αφορούν το δίκτυο, όπως τις διάφορες τιμές που χαρακτηρίζουν το δίκτυο (την τοπολογία, τάση, ρεύμα, άεργος και ενεργός ισχύς, αντίσταση κ.λπ.) και βασίζονται στην ανάλυση του δικτύου και των φυσικών κανόνων που περιγράφουν ένα σύστημα ενέργειας.

Μια από τις βασικές τεχνικές αυτής της κατηγορίας είναι ο έλεγχος της ισορροπίας ενέργειας στο δίκτυο, συγκρίνοντας την ενέργεια που έχει καταναλωθεί σε σχέση με αυτή που έχει μετρηθεί στην πλευρά ενός μετασχηματιστή διανομής ή αλλιώς observer meter (μετρητή παρατηρητή). Με βάση ένα δοσμένο ποσοστό τεχνικών απωλειών, μπορεί να υπολογιστεί αν υπάρχουν διαφορές στις αναμενόμενες από τις πραγματικές τιμές, ικανές ώστε να πούμε αν υπάρχει αρκετά μεγάλη πιθανότητα κλοπής και τι δράσεις θα πρέπει να συμβούν για να μπορέσει να αντιμετωπιστεί το πρόβλημα.

Σε αυτή την περίπτωση εμπίπτει και μια έρευνα που έγινε από την Mitsubishi Electric [30], όπου αφού υπολογιστούν οι απώλειες του δικτύου, υπολογίζονται τόσο οι τεχνικές όσο και οι μη τεχνικές απώλειες, λαμβάνοντας υπόψη την τοπολογία του δικτύου, τις αντιστάσεις των γραμμών, τη σειρά και το σημείο ένωσης των μετρητών στο δίκτυο, τη στιγμιαία κατανάλωση ισχύος, όπως επίσης την άεργο ισχύ και τη στιγμιαία τιμή της τάσης και του ρεύματος.

Σε άλλη έρευνα [31] γίνεται η πρόταση ενός συστήματος πρόληψης και εντοπισμού ρευματοκλοπής σε πραγματικό χρόνο χωρίς να χρειαστεί η διακοπή της παροχής, όπου παρέχεται ένα διαφορετικό εύρος τάσης τροφοδοσίας που δεν είναι κατάλληλη για τυχαία χρονικά διαστήματα στα παράνομα συνδεδεμένα φορτία ανάλογα με τη ρύθμιση της τάσης (voltage regulation) του δικτύου διανομής. Για να μην υπάρξει κάποιο τεχνικό πρόβλημα για τους νόμιμους πελάτες του δικτύου, τους παρέχεται μια μονάδα παρακολούθησης (Consumer Supervision Unit), μέσω της οποίας γίνεται βελτιωμένη ρύθμιση της τάσης.

Τέτοια παραδείγματα αποτελούν επίσης οι μέθοδοι αποφυγής ρευματοκλοπής που περιεγράφηκαν στην ενότητα 2.3 οι οποίες αυτόματα αποτελούν και μεθόδους εντοπισμού.

Hybrid Methods: Οι υβριδικές μέθοδοι αποτελούν ένα συνδυασμό των μεθόδων που βασίζονται στα δεδομένα και των μεθόδων με βάση τα χαρακτηριστικά των δικτύων. Για παράδειγμα με τις μεθόδους δικτύου μπορούν να εντοπιστούν οι μη τεχνικές απώλειες σε επίπεδο μετασχηματιστών και έπειτα με ένα αλγόριθμο μηχανικής μάθησης να γίνει κατηγοριοποίηση ώστε να γίνει εντοπισμός των μη-τεχνικών απωλειών στο επίπεδο των καταναλωτών.

Σε μία από τις βασικές έρευνες που μελετήθηκαν [32] το αποτέλεσμα του αλγόριθμου μηχανικής μάθησης συγκρίνεται με αυτό των αλγορίθμων εξισορρόπησης ενέργειας. Στον αλγόριθμο αυτό γίνεται εκτίμηση των τεχνικών απωλειών του δικτύου και έπειτα αξιολόγηση της ισορροπίας ενεργούς ισχύος. Αν υπάρχει διαφορά που να ξεπερνά ένα κατώφλι και ο αλγόριθμος μηχανικής μάθησης κατηγοριοποιήσει το ημερήσιο πρότυπο κατανάλωσης ως θετικό σε ρευματοκλοπή, τότε ο καταναλωτής θεωρείται υπεύθυνος κλοπής και διερευνάται.

Ακόμη μια έρευνα που συνδυάζει μεθόδους όπως παραπάνω αποτελεί μια από τη Σερβία [33], όπου ο εντοπισμός γίνεται σε δυο φάσεις. Στην πρώτη φάση χρησιμοποιείται ένας αλγόριθμος μηχανικής μάθησης ώστε να δημιουργηθεί ένα πρώτο σετ από ύποπτους καταναλωτές. Όποιος καταναλωτής βρεθεί στην περιοχή

με αυξημένες συνολικές τεχνικές απώλειες αναλύεται. Μια χρονοσειρά κοστολογούμενων τιμών κατανάλωσης παράγεται και χρησιμοποιείται μαζί με τις αντίστοιχες σχέσεις μεταξύ τους για τη δημιουργία των σετ ύποπτων καταναλωτών. Έπειτα υπολογίζεται ο βαθμός υποψίας κάθε καταναλωτή μέσω του αλγορίθμου μηχανικής μάθησης. Με βάση τις εκτιμώμενες συνολικές και τεχνικές απώλειες ενέργειας στην περιοχή του καταναλωτή (περιοχή που παρέχεται από έναν ή περισσότερους σταθμούς μετασχηματιστών ΜΤ/ΧΤ) και το υπόλοιπο της συνολικής, τιμολογημένης ενέργειας των απωλειών, καθορίζεται μια οριακή τιμή του ποσοστού υποψίας. Όλοι οι πελάτες, των οποίων η τιμή υποψίας είναι μεγαλύτερη από αυτή την τιμή, δηλώνονται ύποπτοι και ακολουθεί η διερεύνησή τους.

3.3 Μέθοδοι μηχανικής μάθησης που βασίζονται σε δεδομένα

Οι data oriented μέθοδοι βασίζονται κυρίως σε ανάλυση δεδομένων και τεχνικές μηχανικής μάθησης (machine learning) και μπορούν να διακριθούν σε δύο βασικές κατηγορίες, τις supervised και unsupervised μεθόδους με βάση τη φύση των διαθέσιμων δεδομένων.

Στους supervised αλγορίθμους ως είσοδος χρησιμοποιούνται δεδομένα που έχουν τη μορφή μιας ακολουθίας σημειωμένης με ετικέτες (labeled data), δηλαδή ένας αλγόριθμος δέχεται μεταβλητές εισόδου (variables or features) όπως επίσης και μεταβλητές εξόδου, και πρέπει να μάθει να παράγει το σωστό αποτέλεσμα αν του δοθεί ως είσοδος κάτι που δεν είναι επισημασμένο με ετικέτα (unlabeled data). Τα δεδομένα με τα οποία τροφοδοτείται ο αλγόριθμος ονομάζονται σύνολο δεδομένων εκπαίδευσης (training dataset).

Όσον αφορά στους unsupervised αλγορίθμους, τα δεδομένα εκπαίδευσης δεν διαθέτουν κάποια ετικέτα και αυτό που χρειάζεται να κάνουν είναι να μπορέσουν να βρουν μια «κρυμμένη» δομή στο σύνολο των δεδομένων [34].

Οι παραπάνω διαδικασίες αφορούν τη δημιουργία των μοντέλων μάθησης και διαφέρουν όπως περιγράφηκε. Κοινό των δύο κατηγοριών είναι το στάδιο πριν από

αυτό, όπου γίνεται η επεξεργασία των δεδομένων πριν επιλεχθεί μοντέλο και δοθούν ως τροφοδοσία σε αυτό.

Παρακάτω αναφέρονται οι αλγόριθμοι που μελετήθηκαν στη βιβλιογραφία της εργασίας, χωρισμένοι στις δυο κατηγορίες.

3.3.1 Supervised μέθοδοι

Ένας από τους βασικούς supervised αλγόριθμους που χρησιμοποιήθηκαν στον εντοπισμό ρευματοκλοπής μέχρι πρόσφατα είναι ο SVM ή Support Vector Machine [28]. Ο SVM βασίζεται στο διαχωρισμό των κλάσεων με όσο το δυνατόν πιο ευδιάκριτο τρόπο. Γι' αυτό το λόγο μπορούμε να φανταστούμε μια νοητή γραμμή που ονομάζεται hyperplane, όπου στόχος είναι τα σημεία (vectors) των κλάσεων που βρίσκονται πιο κοντά σε αυτή να δημιουργούν το μεγαλύτερο δυνατό περιθώριο (margin) από τη γραμμή. Όταν τα δεδομένα είναι αρκετά περίπλοκα, μπορεί να γίνει χρήση μιας συνάρτησης πυρήνα (kernel function) με την οποία μπορούν να αποτυπωθούν σε έναν άλλο χώρο υψηλότερων διαστάσεων, οπότε να μπορεί να γίνει πιο ξεκάθαρος ο διαχωρισμός τους. Στην έρευνα των Jokar και Arianpoo [32], αφού πρώτα τα πραγματικά δεδομένα χωρίστηκαν σε περισσότερες κλάσεις και ενώθηκαν με τα δεδομένα κλοπής, χρησιμοποιήθηκε ο SVM για να λύσει ένα πρόβλημα πολλαπλής κατηγοριοποίησης (multiclass classification), με χρήση του πυρήνα rbf (radial basis function).

Η γραμμική παλινδρόμηση ή αλλιώς Linear Regression που χρησιμοποιήθηκε από τους Yip, Wong [35], είναι μια μέθοδος που προσπαθεί να λύσει ένα γραμμικό πρόβλημα της μορφής a×x + b = y, βρίσκοντας ποια είναι η σχέση μεταξύ του y και των εξαρτημένων μεταβλητών x. Στην έρευνα αυτή, οι συγγραφείς καλούνταν να επιλύσουν ένα σύστημα γραμμικών εξισώσεων σαν αυτό της Σχέσης 3.1.

$$\begin{cases} a_1 p_{t_{1,1}} + a_2 p_{t_{1,2}} + \cdots + a_N p_{t_{1,N}} = y_{t_1} \\ \vdots \\ a_1 p_{t_{T,1}} + a_2 p_{t_{T,2}} + \cdots + a_N p_{t_{T,N}} = y_{t_T} \end{cases} \tag{3.1}$$

Τα $p_{t,N}$ αποτελούν την ενέργεια που έχει δηλωθεί από τους καταναλωτές 1 έως N κατά τα χρονικά διαστήματα 1 έως T, ενώ y_t είναι η ενέργεια που έχει μετρηθεί από την πλευρά του συστήματος διανομής. Οι συντελεστές α είναι οι δείκτες για την ύπαρξη κάποιας ανωμαλίας. Οπότε, αν προκύψει ότι $\alpha_n = 0$, ο καταναλωτής n δεν διαπράττει κλοπή, αν $\alpha_n > 0$ τότε αναφέρεται λιγότερη ενέργεια από ότι καταναλώνεται και αν $\alpha_n < 0$, τότε το N-στός μετρητής καταγράφει περισσότερο από ότι έχει καταναλωθεί (για την έρευνά τους, αυτή η περίπτωση εμπίπτει σε μη λειτουργικό μετρητή).

Τα νευρωνικά δίκτυα (neural networks) είναι μια μέθοδος μηχανικής μάθησης που στην περίπτωση του εντοπισμού ρευματοκλοπής έχουν δοκιμαστεί αρκετά και με διάφορες μεθόδους όπως τα convolutional neural networks (CNN), recurrent neural networks (RNN), αλλά σε πολλές περιπτώσεις συνδυάζονται και με άλλους αλγόριθμους μηχανικής μάθησης, όπως SVM, kNN, Random Forests, Adaptive Boosting. Στο παράδειγμα της μεθόδου που προτείνουν οι Jeyaranjani και Devaraj [36], αφού τα προφίλ κατανάλωσης διακριθούν σε πραγματικά και κλοπής, εκπαιδεύεται ένας artificial neural network classifier για να λύσει ένα πρόβλημα πολλαπλής κατηγοριοποίησης (multiple classification) και τεστάρεται ώστε να αξιολογηθεί αν μπορεί να αναγνωρίσει τα πραγματικά (κλάση 0) από τους 3 τύπους δεδομένων κλοπής (οι υπόλοιπες κλάσεις). Σε άλλη έρευνα [37], έγινε χρήση του deep Recurrent Neural Network, με hidden recurrent layers (κρυφά διαδοχικά επίπεδα), όπου στην ουσία η πληροφορία των χρονοσειρών κατανάλωσης μεταφέρεται από επίπεδο σε επίπεδο «μαθαίνοντας» όλες τις απαραίτητες παραμέτρους κάθε επιπέδου. Παράδειγμα συνδυασμού ενός νευρωνικού δικτύου και ενός άλλου machine learning αλγόριθμου αποτελεί η έρευνα των Li, Han, Yao [38], όπου χρησιμοποίησαν ένα convolutional neural network, για να εντοπίσουν στοιχεία μεταξύ των διαφορετικών ημερών και των ωρών της ημέρας (feature selection process) και έπειτα εκπαίδευσαν με τα παραπάνω στοιχεία τον αλγόριθμο Random Forests (classification αλγόριθμος) για να αποφασιστεί η κλοπή ενέργειας. Κατά τον CNN σε κάθε στάδιο εφαρμόζεται σε ένα επίπεδο η μέθοδος convolution (μια μέθοδος που

χρησιμοποιείται και στην επεξεργασία εικόνας) ακολουθούμενη από ένα επίπεδο downsampling για σμίκρυνση των δεδομένων.

Οι gradient boosting αλγόριθμοι βασίζονται σε decision trees (δέντρα αποφάσεων) και λειτουργούν με τη μέθοδο λήψης αποφάσεων με βάση τα δεδομένα εισόδου και τη λογική του συνδυασμού των αποτελεσμάτων των δέντρων που παράγονται κατά την εκπαίδευση, προσπαθώντας πάντα να βελτιώσουν τα ρεκόρ των δέντρων με τα λιγότερο καλά αποτελέσματα. Περισσότερα για αυτούς τους αλγορίθμους θα παρουσιαστούν στο επόμενο κεφάλαιο, όπως και για την έρευνα στην οποία βασίστηκε η παρούσα εργασία. Μια ακόμη αξιόλογη έρευνα όπου έγινε χρήση του gradient boosting αλγόριθμου για τον εντοπισμό ρευματοκλοπής είναι αυτή των Razavi, Gharipour [39], οι οποίοι εφάρμοσαν πρώτα feature engineering αλγόριθμους ώστε να παράγουν νέα δεδομένα που μπορούν να προσφέρουν περισσότερη πληροφορία στο μοντέλο και ύστερα εκπαίδευσαν ένα GBM αλγόριθμο με τα δεδομένα χρονοσειράς και τα νέα στοιχεία, για την ταξινόμηση ενός καταναλωτή ως παραβάτη ή όχι.

3.3.2 Unsupervised μέθοδοι

Από τους βασικούς αλγόριθμους μηχανικής μάθησης που δεν κάνει χρήση ετικετών είναι ο k-Nearest Neighbors (kNN). Η διαδικασία που ακολουθείται είναι η επιλογή της κλάσης στην οποία ανήκει ένα νέο δείγμα με βάση τους k πιο κοντινούς γείτονες, τους οποίους συνήθως βρίσκει με τη χρήση της ευκλείδειας απόστασης. Στον εντοπισμό ρευματοκλοπής οι Sowndaraya, Latha [40], χρησιμοποίησαν τον kNN για να κατηγοριοποιήσουν τα προφίλ των καταναλωτών ως πραγματικά ή κλοπής και έπειτα τα έδωσαν ως είσοδο σε ένα network state αλγόριθμο ο οποίος έκανε σύγκριση των δεδομένων των μετρητών με αυτά που είχαν καταγραφεί από την υπηρεσία παροχής.

Μια έρευνα που αφορά τον εντοπισμό ανωμαλιών στην κατανάλωση είναι αυτή των Yeckle και Tang [41] οι οποίοι χρησιμοποιούν και αξιολογούν εφτά outlier detection αλγόριθμους. Η βάση των outlier detection αλγόριθμων είναι ο εντοπισμός των

σημείων που φαίνεται να διαφέρουν πολύ από άλλα σημεία. Αφού πρώτα μειώσουν τα ημερήσια δεδομένα των καταναλωτών με ομαδοποίηση χρησιμοποιώντας τον kNN, αξιολογούν τους παρακάτω εφτά αλγόριθμους: Local Outlier Factor, Local Density Factor, Flexble Kernel Density Estimates, Influenced Outlierness, Relative Density-based Outlier Score, Mutual k-nearest neighbor, Indegree Number.

ΚΕΦΑΛΑΙΟ 4
ΕΝΤΟΠΙΣΜΟΣ ΡΕΥΜΑΤΟΚΛΟΠΗΣ ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ GRADIENT BOOSTING ΚΑΙ AUTOML

4.1 Ορισμός του προβλήματος – Μεθοδολογία

Το πρόβλημα που καλούμαστε να αντιμετωπίσουμε σε αυτή την εργασία είναι η κατηγοριοποίηση (classification) των ημερήσιων μοτίβων κατανάλωσης ενός καταναλωτή σε δείγμα κλοπής (theft sample) ή όχι. Αποτελεί λοιπόν ένα πρόβλημα δυαδικής κατηγοριοποίησης (binary classification), όπου οι δυο πιθανές ετικέτες (labels) δείχνουν κλοπή (=1) ή όχι (=0).

Τα δεδομένα που έχουμε λάβει είναι σε μορφή χρονοσειράς (time series) και δέχονται κάποια επεξεργασία ώστε μεταξύ άλλων να έρθουν σε μορφή όπου οι γραμμές να παρουσιάζουν τις μέρες και οι στήλες τις χρονικές στιγμές μέτρησης μέσα στη μέρα.

Επειδή τα αρχικά δεδομένα θεωρούνται ως γνήσια και υπάρχει η παραδοχή ότι δεν περιλαμβάνουν δείγματα κλοπής, χρειάζεται να δημιουργήσουμε δεδομένα με χαρακτηριστικά ρευματοκλοπής και να τα εισάγουμε στο σετ δεδομένων.

Βασικό κομμάτι της μεθοδολογίας αποτελεί η εφαρμογή του eXtreme Gradient Boosting (XGBoost) αλγορίθμου αλλά και η δοκιμή του automated Machine Learning (autoML) σε δυο σετ δεδομένων που περιλαμβάνουν τις καταγραφές των δεδομένων κατανάλωσης σε ημερήσια βάση όπως διαμορφώθηκαν κατά το προηγούμενο στάδιο. Παρακάτω αναφέρονται τα βασικά χαρακτηριστικά τόσο του XGBoost όσο και της AutoML και παρουσιάζονται τα βήματα που ακολουθήθηκαν κατά τη διάρκεια των εφαρμογών.

Η χρήση του XGBoost αποτελούσε μέρος της ερευνητικής αναφοράς των Punmiya και Choe [42], όπου γινόταν σύγκρισή του με τους δυο gradient boosting αλγόριθμους, τους AdaBoost (Adaptive Boosting) και LightGBM. Με βάση την παραπάνω έρευνα αλλά και αυτή των Jokar και Arianpoo [32], στην οποία βασίστηκαν και οι Punmiya και Choe, θα παρουσιαστούν σε επόμενη ενότητα τα βήματα που ακολουθήθηκαν κατά τις δοκιμές.

Η δημιουργία των τεχνητών δεδομένων κλοπής στην έρευνα των Punmiya και Choe βασίζεται σε αυτή των Jokar και Arianpoo με κάποιες μικρές παραλλαγές οι οποίες θα αναφερθούν στην ανάλογη ενότητα.

4.2 Ο αλγόριθμος XGBoost

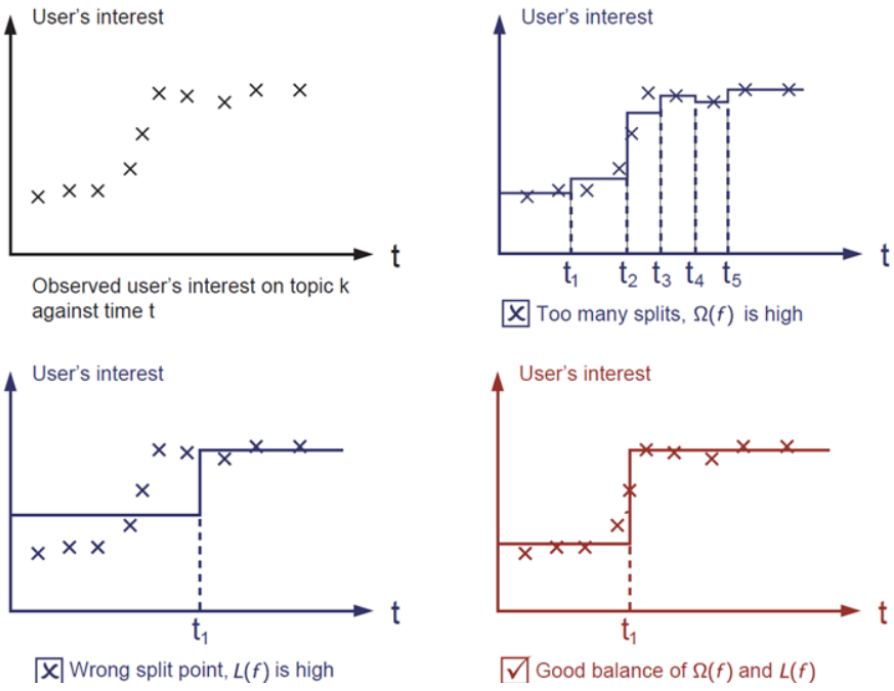
Ο XGBoost [43][44] ανήκει στους gradient boosting αλγόριθμους οι οποίοι έχουν ως κύριο χαρακτηριστικό τους ότι συνδυάζουν μοντέλα πιο αδύναμα στη μάθηση ώστε να δημιουργήσουν ένα πιο δυνατό μοντέλο προβλέψεων δουλεύοντας με τη λογική της αρχής συνόλου (ensemble). Τα μοντέλα που συνδυάζονται είναι συνήθως τα δέντρα αποφάσεων (decision trees). Σε κάθε χρονική στιγμή t τα αποτελέσματα του μοντέλου ζυγίζονται με βάση τα αποτελέσματα της χρονικής στιγμής t-1. Στα αποτελέσματα που έχουν προβλεφθεί σωστά δίνεται μεγαλύτερη βαρύτητα, ενώ αυτά που δεν εμπίπτουν στη σωστή κατηγορία παίρνουν μεγαλύτερα βάρη. Στην ουσία σε κάθε βήμα δημιουργείται ένα αδύναμο μοντέλο (ένα τέτοιο μοντέλο μπορεί να είναι κάποιο που προβλέπει σωστά λίγο παραπάνω από το 50% των περιπτώσεων), προκύπτουν κάποια συμπεράσματα που αφορούν τις παραμέτρους του αλγορίθμου και τη σημασία/βάρος των στοιχείων των δεδομένων και αυτά χρησιμοποιούνται για να δημιουργηθεί ένα νέο και πιο δυνατό μοντέλο που εστιάζει στη μείωση του σφάλματος της λανθασμένης κατηγοριοποίησης. Οι gradient boosting αλγόριθμοι στηρίζονται ως επί το πλείστον στη συνάρτηση του σφάλματος, η οποία για τους αλγόριθμους κατηγοριοποίησης (classification) είναι κυρίως το λογαριθμικό σφάλμα ενώ για τους αλγορίθμους παλινδρόμησης (regression) είναι το τετραγωνικό σφάλμα, όπως μπορεί επίσης να είναι και μια προσαρμοσμένη συνάρτηση.

Τα τελευταία χρόνια ο XGBoost έχει χρησιμοποιηθεί σε πολλούς διαγωνισμούς του Kaggle και τα αποτελέσματά του είναι πολύ ικανοποιητικά ειδικά σε διαδικασίες κατηγοριοποίησης. Τα tree ensembles στα οποία βασίζεται ο XGBoost αποτελούν ένα σετ από Classification και Regression Trees τα λεγόμενα CART [45]. Τα δέντρα δημιουργούνται το ένα μετά το άλλο, και γίνονται προσπάθειες μείωσης του ποσοστού εσφαλμένης ταξινόμησης σε επόμενες επαναλήψεις. Το σημαντικό για το

μοντέλο που δημιουργείται κάθε φορά είναι η βελτιστοποίηση της αντικειμενικής συνάρτησης (objective function), που δηλώνει πόσο καλά ταιριάζει το μοντέλο στα δεδομένα εκπαίδευσης (fitting training data) και η οποία αποτελείται από δυο όρους, το σφάλμα εκπαίδευσης (training loss) L(θ) και το ρυθμιστικό όρο (regularization term) Ω(θ) σύμφωνα με τη Σχέση 4.1 (objective function).

$$obj(\theta) = L(\theta) + \Omega(\theta) \tag{4.1}$$

Το training loss αφορά το πόσο καλά γίνεται η πρόγνωση σε σχέση με τα training data, ενώ ο regularization term σχετίζεται με το πόσο πολύπλοκο γίνεται ένα μοντέλο και βοηθά στο να αποφύγουμε το υπερβολικό ταίριασμα (overfitting) των δεδομένων στο μοντέλο. Αυτό σημαίνει ότι δεν θέλουμε το μοντέλο που θα παραχθεί να έχει άριστη γνώση των δεδομένων στο training set σε βαθμό που να μπορεί να προβλέψει καλά μόνο τα συγκεκριμένα δεδομένα. Χρειάζεται λοιπόν μια ισορροπία μεταξύ των δυο όρων της συνάρτησης ώστε τα δεδομένα να ταιριάζουν καλά και παράλληλα να μην αυξάνεται το σφάλμα, όπως φαίνεται και Σχήμα 4.1.



Σχήμα 4.1: Ισορροπία Ω και L

4.2.1 Παράμετροι του XGBoost και Python API

Ο κώδικας που χρησιμοποιήθηκε για την εργασία είναι σε Python και γράφτηκε σε περιβάλλον Spyder κατά την επεξεργασία των δεδομένων και σε Jupyter Notebook στις υπόλοιπες φάσεις.

Ο XGBoost δίνει τη δυνατότητα χρήσης του δικού του (native) API αλλά και μέσω της βιβλιοθήκης scikit-learn [43]. Η βασική βιβλιοθήκη του XGBoost λαμβάνει ως είσοδο δεδομένα με μια δική του εσωτερική δομή, οπότε και χρειάζεται να γίνει μετατροπή των δεδομένων σε μορφή Dmatrix που προσφέρει καλύτερη απόδοση σε ταχύτητα και χρήση της μνήμης. Στην περίπτωση χρήσης της βιβλιοθήκης scikit-learn δεν χρειάζεται να γίνει κάποια μετατροπή των δεδομένων, οπότε χρησιμοποιούμε τη δομή ενός data frame από τη βιβλιοθήκη pandas. Το native API του XGBoost μεταξύ άλλων διαθέτει και δική του μέθοδο για cross-validation, την οποία θα εξηγήσουμε σε επόμενη ενότητα πώς τη χρησιμοποιήσαμε στη δική μας περίπτωση.

Όσον αφορά τις παραμέτρους του αλγορίθμου, αυτές μπορούν να διαχωριστούν ως εξής:

- Γενικοί παράμετροι: σχετίζονται με το είδος του «ενισχυτή» (booster) που θα χρησιμοποιηθεί, συνηθώς για δενδρικό ή γραμμικό μοντέλο
- Booster παράμετροι: ανάλογοι του ενισχυτή που έχει επιλεχθεί
- Παράμετροι διαδικασίας μάθησης (learning task parameters): καθορίζουν το σενάριο μάθησης

Από τις γενικές παραμέτρους αναφέρουμε την “booster”, που διαλέγουμε να είναι εξ ορισμού στην επιλογή gbtree για δενδρικές μορφές, και όχι γραμμικές όπως με την επιλογή gblinear, και την “nthread”, που εξ ορισμού θέτει το μέγιστο αριθμό παράλληλων νημάτων που διαθέτει το σύστημα κατά τη διάρκεια εκτέλεσης.

Σχετικά με τις booster παραμέτρους αναφέρουμε τις εξής:

- Learning rate: αφορά το βαθμό στον οποίο συρρικνώνονται τα βάρη των στοιχείων (features) σε κάθε βήμα ώστε να αποφευχθεί το overfitting, συνήθεις τιμές μεταξύ 0.01 και 0.2
- Min_child_weight: το ελάχιστο άθροισμα των βαρών όλων των δειγμάτων (samples/observations) σε ένα παιδί του δέντρου που επιτρέπει τον περαιτέρω διαχωρισμό του κόμβου
- Max_depth: το μέγιστο βάθος του δέντρου
- Max_leaf_nodes: ο μέγιστος αριθμός των τερματικών κόμβων και είναι ανάλογο του max_depth
- Gamma: η ελάχιστη μείωση σφάλματος ώστε να γίνει περαιτέρω διαχωρισμός ενός κόμβου-φύλλου, δηλαδή ο διαχωρισμός θα συμβεί αν υπάρχει θετική μείωση της συνάρτησης σφάλματος. Εξ ορισμού θέτεται στο 0 και όσο μεγαλώνει ο αλγόριθμος γίνεται πιο συντηρητικός
- Subsample: το κλάσμα των γραμμών-δειγμάτων (samples/observations) που διαλέγονται τυχαία για την εκπαίδευση
- Colsample_bytree/bylevel/bynode: το κλάσμα/ποσοστό των στηλών-στοιχείων (features) που θα χρησιμοποιηθεί για τη δημιουργία ενός ολόκληρου δέντρου/ενός επιπέδου του δέντρου/ενός κόμβου
- Lambda: ρυθμιστικός όρος L2 για τα βάρη για να αποφευχθεί το overfitting και παίρνει τιμές κοντά στο 1
- Alpha: ρυθμιστικός όρος L1 για τα βάρη, χρησιμοποιείται κυρίως σε προβλήματα μεγάλων διαστάσεων και παίρνει τιμές κοντά στο 0

Από τις παραμέτρους εκμάθησης ξεχωρίζουμε τις παρακάτω:

- Objective: ορίζει τη συνάρτηση σφάλματος που χρειάζεται να ελαχιστοποιηθεί και στην περίπτωση της δυαδικής κατηγοριοποίησης (binary classification) ορίζεται ως binary:logistic

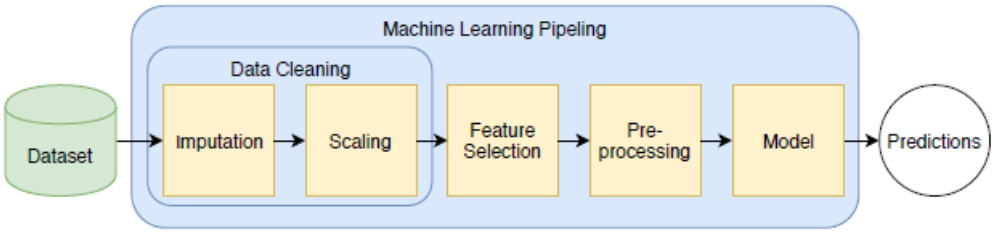
- Eval_metric: καθορίζει το μέτρο με το οποίο γίνεται η αξιολόγηση των δεδομένων επικύρωσης (validation data), επιλέγεται ως επί το πλείστον το error για classification ή auc (Area Under roc Curve)
- Seed: δίνει το «σπόρο» για να ρυθμιστεί το πόσο τυχαία θα είναι μια εκτέλεση (randomness), ορίζεται σε έναν αριθμό ώστε να έχουμε παραγωγή των ίδιων αποτελεσμάτων αν το επιθυμούμε

Για να υπάρξει βελτιστοποίηση της απόδοσης του αλγορίθμου, ακολουθείται μια διαδικασία ρύθμισης των παραπάνω παραμέτρων ή υπερ-παραμέτρων (hyper parameters) όπως αναφέρονται στη βιβλιογραφία. Σε στάδιο εφαρμογής του αλγορίθμου πραγματοποιούμε ρύθμιση ορισμένων από τις παραπάνω παραμέτρους με διάφορους τρόπους για βελτιστοποίηση (hyper parameter optimization).

4.3 Αυτοματοποιημένη Μηχανική Μάθηση

Η αυτοματοποιημένη μηχανική μάθηση (Automated Machine Learning) είναι στη ουσία αυτό που ορίζει και το όνομά της. Προσφέρει τη δυνατότητα σε επιστήμονες που δεν έχουν ιδιαίτερη γνώση μαθηματικών μεθόδων και στατιστικής την ευκαιρία να χρησιμοποιήσουν αλγορίθμους μηχανικής μάθησης ρυθμίζοντας μόνο ελάχιστες παραμέτρους και να μην εμπλέκονται περαιτέρω στη διαδικασία, ειδικά στην περίπτωση της ρύθμισης των υπερ-παραμέτρων ορισμένων αλγόριθμων που χρειάζονται γνώση και εξοικείωση. Η διαδικασία επιλογής του κατάλληλου μοντέλου συνδυάζεται επίσης και με τις απαραίτητες μεθόδους προ επεξεργασίας των δεδομένων, ανάμεσα σε όλους τους συνδυασμούς που θα δοκιμαστούν από την autoML [46].

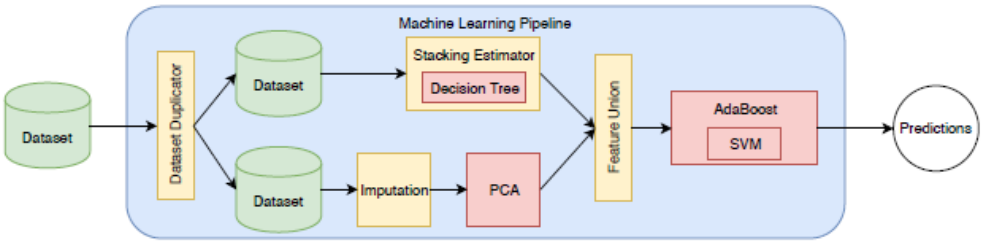
Η διαδικασία που ακολουθεί η autoML αποτελείται από διάφορα στάδια, τα οποία όλα μαζί ορίζουν ένα pipeline [47]. Σε αυτό περιλαμβάνονται τα: specialized data preprocessing, domain-driven meaningful feature engineering and fine-tuned models. Οι παραπάνω λειτουργίες χρειάζονται πολύ χρόνο για την επιλογή και τη βελτιστοποίησή τους, κάτι που με την autoML γίνεται αυτόματα. Παρακάτω μπορούμε να δούμε τη βασική δομή ενός pipeline (Σχήμα 4.2).



Σχήμα 4.2: AutoML pipeline που χρησιμοποιείται από τα περισσότερα frameworks

Αυτό που μπορούμε να δούμε είναι ένα data cleaning στάδιο όπου οι πιο συνήθεις λειτουργίες καθαρισμού δεδομένων είναι το imputation (για missing values) και το scaling (όταν θέλουμε να φέρουμε τα δεδομένα όλα στην ίδια κλίμακα αριθμών). Έπειτα έχουμε το στάδιο του feature selection κατά το οποίο διαλέγονται κάποια από τα στοιχεία με βάση το κατά πόσο γίνεται πιο αποδοτικός ένας αλγόριθμος όταν συμμετέχουν στη διαδικασία μόνο τα πιο σχετικά ή πιο χρήσιμα στοιχεία του σετ. Στη συνέχεια υπάρχει ένα ή περισσότερα preprocessing στάδια, κατά τα οποία εφαρμόζονται τεχνικές όπως η PCA για αλλαγή του άξονα αναπαράστασης των σημείων, και τέλος έχουμε το στάδιο επιλογής του καλύτερου μοντέλου για classification ή regression.

Για να μπορέσουν τα δημιουργούμενα pipelines να είναι ευέλικτα σε διάφορες διεργασίες που θα τους ανατεθούν, έχουν ορισμένα βασικά χαρακτηριστικά. Αυτά είναι δύο operators, το data set duplicator και το feature union. Με το πρώτο δημιουργείται ένας κλώνος των δεδομένων και έτσι δημιουργούνται δυο διαφορετικά μονοπάτια για κάθε σετ, και με το δεύτερο γίνεται η συνένωση των δύο μονοπατιών. Ένα τέτοιο παράδειγμα είναι αυτό του Σχήματος 4.3.

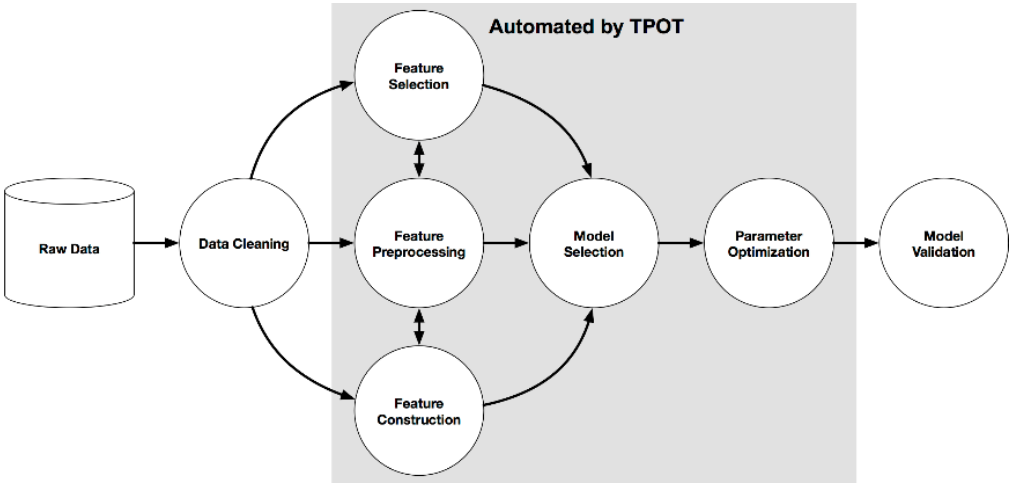


Σχήμα 4.3: Ειδικό pipeline για συγκεκριμένη ML διαδικασία

Όσον αφορά τις διάφορες δομές/βιβλιοθήκες που είναι διαθέσιμες για να μπορέσουμε να αξιοποιήσουμε τις λειτουργίες της autoML, μερικές από τις πιο γνωστές είναι οι παρακάτω:

Auto-sklearn [48]: όπως υποδηλώνει και το όνομά της, η auto-sklearn βασίζεται στη βιβλιοθήκη της python, scikit-learn, και περιλαμβάνει πολλούς από τους πιο γνωστούς αλγόριθμους τους οποίους χρησιμοποιούμε και με εκείνη. Δημιουργεί ένα pipeline και το βελτιστοποιεί χρησιμοποιώντας αναζήτηση Bayes. Διαθέτει ένα data cleaning στάδιο με τις παρακάτω λειτουργίες: categorical encoding, imputation, removing variables with low variance scaling και προαιρετικό στάδιο preprocessing. Για τη ρύθμιση των υπερ-παραμέτρων με την Bayes λογική έχουν προστεθεί δύο κομμάτια, εκ των οποίων το ένα είναι το meta-learning, για την αρχικοποίηση των παραμέτρων βελτιστοποίησης. Η επιλογή των μοντέλων γίνεται μεταξύ 15 classification αλγορίθμων και 14 preprocessing.

TPOT [49]: Tree-Based Pipeline Optimization Tool. Με το TPOT δημιουργούνται ευέλικτα pipelines χρησιμοποιώντας γενετικό προγραμματισμό (genetic programming) για τη βελτιστοποίησή τους με βάση τη δενδρική δομή. Βασίζεται στη scikit-learn βιβλιοθήκη και έχει ενσωματωμένους δικούς του αλγόριθμους για classification και regression. Ένα παράδειγμα του πώς δημιουργείται ένα pipeline στο TPOT είναι το παρακάτω (Σχήμα 4.4).



Σχήμα 4.4: TPOT pipeline

MLBox [50]: Άλλη μια βιβλιοθήκη σε python, με τρία βασικά στάδια (preprocessing, optimization, prediction) που συγκεκριμένα προσφέρει: data preprocessing, cleaning και formatting. Επίσης feature selection και leak detection, hyper-parameter optimization σε high-dimensional space και από τα πιο γνωστά μοντέλα προβλέψεων για classification και regression.

H2O autoML [51]: Με το H2O παρέχονται οι γνωστές λειτουργίες ενός pipeline, data-preprocessing, feature engineering και model deployment. Βασικό χαρακτηριστικό της συγκεκριμένης βιβλιοθήκης είναι ότι διαμορφώνει και αξιολογεί pipelines που χρησιμοποιούν πολλά από τα πιο γνωστά μοντέλα για classification και regression, αλλά πάντα υπάρχει και ένα pipeline που περιλαμβάνει ένα ensemble μοντέλο, συνδυάζοντας όλα τα μοντέλα με τα καλύτερα σκορ που προέκυψαν από τη διαδικασία της εκπαίδευσης.

Auto-Keras [52]: η Auto-Keras είναι η αυτοματοποιημένη εκδοχή της βιβλιοθήκης Keras για την εύρεση ενός deep learning μοντέλου, αυτοματοποιώντας τη διαδικασία ρύθμισης των υπερ-παραμέτρων και της κατάλληλης αρχιτεκτονικής. Η δομή της είναι βασισμένη στη scikit-learn, γι' αυτό και είναι αρκετά εύκολη στη χρήση.

4.4 Σετ δεδομένων και προ επεξεργασία

Για την αξιολόγηση του XGBoost και την εφαρμογή της autoML έγινε χρήση δύο διαφορετικών σετ δεδομένων. Το πρώτο προέρχεται από την αποθήκη δεδομένων του Ιρλανδικού Κοινωνικού Επιστημονικού Αρχείου Δεδομένων (Irish Social Science Data Archive –ISSDA) [53]. Τα δεδομένα αυτά αποτελούν μέρος ενός smart metering πρότζεκτ της Επιτροπής Ρύθμισης Ενέργειας (Commission for Energy Regulation – CER) της Ιρλανδίας. Το δεύτερο προέρχεται από το πρότζεκτ Low Carbon London του UK Power Networks [54].

4.4.1 Σετ δεδομένων από το ISSDA

Το smart metering πρότζεκτ ξεκίνησε το 2007 με σκοπό να γίνει αξιολόγηση της απόδοσης και δοκιμή των έξυπνων μετρητών, όπως και του αντικτύπου που αυτοί μπορεί να έχουν στην κατανάλωση ενέργειας των πελατών, γι' αυτό το λόγο και το

πρότζεκτ περιλαμβάνει το Electricity Customer Behaviour Trial (δοκιμές συμπεριφοράς καταναλωτών ενέργειας). Οι δοκιμές αυτές πραγματοποιήθηκαν κατά την περίοδο μεταξύ 2009 και 2010, με συμμετέχοντες λίγο παραπάνω από 5000 Ιρλανδικά σπίτια και μικρομεσαίες επιχειρήσεις.

Οι συμμετέχοντες ήταν χωρισμένοι σε δυο ομάδες, τις test και control. Αυτοί που ανήκαν στο test group τους ζητήθηκε να δοκιμάσουν διαφορετικούς τρόπους κοστολόγησης βάσει του χρόνου χρήσης (Time of Use – ToU), ενώ στο control group οι καταναλωτές είχαν σταθερή τιμή ανά κιλοβατώρα.

Τα δεδομένα περιλαμβάνουν τις τιμές κατανάλωσης ενέργειας ανά μισάωρο σε kWh και ήταν αρχικά σε μορφή χρονοσειράς με τρεις διακριτές στήλες, το meter ID, ένα 5ψήφιο κωδικό που αντιπροσωπεύει την ημέρα και τον αριθμό της μέτρησης μέσα στη μέρα, όπου τα πρώτα τρία ψηφία είναι ο αριθμός της μέρας και τα δυο τελευταία ο αριθμός μέτρησης, και το ύψος της κιλοβατώρας που είχε καταναλωθεί στο διάστημα ενός μισαώρου. Ένα παράδειγμα φαίνεται στον Πίνακα 4.1.

Πίνακας 4.1: Μορφή αρχικών δεδομένων (ISSDA)

```
2113 19501 0.189
2113 19502 0.139
2113 19503 0.149
2113 19504 0.039
2113 19505 0.039
2113 19506 0.142
2113 19507 0.133
2113 19508 0.039
2113 19509 0.039
2113 19510 0.152
```

Μαζί με τα txt αρχεία που περιλαμβάνουν τη χρονοσειρά κατανάλωσης, τα δεδομένα του ISSDA περιέχουν τόσο ένα αρχείο excel με τον τρόπο που είναι χωρισμένοι οι καταναλωτές σε κατοικίες αλλά και το είδος κοστολόγησης στο οποίο εμπίπτουν, αλλά και τα pre- και post-trial ερωτηματολόγια που έχουν συμπληρωθεί από τους ίδιους, όσον αφορά κάποια οικογενειακά και οικονομικά χαρακτηριστικά. Τα τελευταία δεν χρησιμοποιούνται σε αυτή την εργασία. Από το αρχείο excel ωστόσο

μπορέσαμε να μετρήσουμε τον αριθμό των αναγνωριστικών (ids) για τις κατοικίες που ήταν 4225 και για τις μικρομεσαίες επιχειρήσεις που ήταν 485. Τα υπόλοιπα ids δεν υπάγονταν σε καμία ομάδα από τις παραπάνω. Επίσης, στο control group ανήκουν από τις κατοικίες οι 929 και από τις επιχειρήσεις οι 197.

Ξεκινώντας την αρχική επεξεργασία των δεδομένων στα αρχεία txt, χρησιμοποιήθηκε το αρχείο excel μέσω του οποίου μπορέσαμε να κρατήσουμε μόνο τα αναγνωριστικά των κατοικιών και όχι των επιχειρήσεων. Έπειτα μετρήσαμε τις εμφανίσεις κάθε id. Εφόσον τα δεδομένα ξεκινάνε από τη μέρα 195 και τελειώνουν στη μέρα 730 και γνωρίζοντας ότι οι μετρήσεις έγιναν από το 2009 μέχρι το 2010, τότε ξέρουμε επίσης ότι οι μέρες καταγραφής είναι από τις 17/7/2009 έως τις 31/12/2010, άρα 536 μέρες. Επομένως για κάθε id θα πρέπει να υπάρχουν 536*48=25728 εγγραφές χρονοσειράς. Άρα όσοι καταναλωτές έχουν λιγότερες από αυτές τις εγγραφές, έχουν δεδομένα που λείπουν (missing values), και δεν τους χρησιμοποιούμε. Παρόλα αυτά παρατηρήθηκε ότι για όλα τα υπόλοιπα ids υπήρχαν δυο παραπάνω τιμές, οπότε κατά τη δημιουργία κάθε ξεχωριστού πίνακα για κάθε αναγνωριστικό παραλήφθηκαν οι δυο τελευταίες τιμές. Τέλος, για κάθε καταναλωτή δημιουργήθηκαν πίνακες μεγέθους 536 γραμμών, όσες δηλαδή και οι μέρες, και 48 στηλών, όσες και οι μετρήσεις μέσα στη μέρα.

4.4.2 Σετ δεδομένων από το UK Power Networks

Το dataset του Low Carbon London πρότζεκτ περιλαμβάνει δεδομένα μετρήσεων κατανάλωσης (meter readings) από ένα δείγμα 5567 νοικοκυριών του Λονδίνου στο χρονικό διάστημα από Νοέμβριο του 2011 μέχρι Φεβρουάριο του 2014. Οι μετρήσεις πραγματοποιήθηκαν σε διαστήματα μισής ώρας, όπως και στο προηγούμενο σετ, ένας λόγος για τον οποίο και το επιλέξαμε.

Τα δεδομένα περιλαμβάνουν την κατανάλωση ενέργειας σε kWh ανά μισάωρο, το ξεχωριστό id για κάθε νοικοκυριό, την ώρα και την ημερομηνία και το CACI Acorn group στο οποίο ανήκουν.

Το σετ περιλαμβάνει πελάτες που ανήκουν σε δυο κατηγορίες. Στην πρώτη ομάδα ανήκουν 1100 πελάτες που έχουν υποβληθεί σε δυναμική χρέωση ενέργειας ανάλογα με τη χρήση (dynamic Time of Use – dToU) για το έτος 2013. Στους πελάτες δόθηκαν σήματα υψηλής (67,20 p/kWh), χαμηλής (3,99 p/kWh) ή κανονικής τιμής (11,76 p/kWh) και οι ώρες της ημέρας που εφαρμόστηκαν. Οι υπόλοιποι περίπου 4500 πελάτες δεν υποβλήθηκαν σε δυναμική χρέωση και είχαν μια σταθερή κοστολόγηση στα 14,228p/kWh.

Ένα παράδειγμα της μορφής των δεδομένων πριν την επεξεργασία τους φαίνεται στον Πίνακα 4.2.

Πίνακας 4.2: Μορφή αρχικού αρχείου δεδομένων (UKPN)

LCLid	stdorToU	DateTime	KWH/hh (per half hour)	Acorn	Acorn_grouped
MAC003718	Std	17/10/2012 13:00:00	0.09	ACORN-A	Affluent
MAC003718	Std	17/10/2012 13:30:00	0.16	ACORN-A	Affluent
MAC003718	Std	17/10/2012 14:00:00	0.212	ACORN-A	Affluent
MAC003718	Std	17/10/2012 14:30:00	0.145	ACORN-A	Affluent
MAC003718	Std	17/10/2012 15:00:00	0.104	ACORN-A	Affluent
MAC003718	Std	17/10/2012 15:30:00	0.122	ACORN-A	Affluent
MAC003718	Std	17/10/2012 16:00:00	0.184	ACORN-A	Affluent
MAC003718	Std	17/10/2012 16:30:00	0.171	ACORN-A	Affluent

Για την περίπτωση αυτής της εργασίας οι στήλες που μας ενδιέφεραν ήταν οι παρακάτω: LCLid, DateTime, KWH/hh, ενώ όλες οι άλλες αφαιρέθηκαν και έτσι η μορφή του σετ έγινε ίδια με αυτή του προηγούμενου σετ.

Σε επόμενο στάδιο η στήλη ημερομηνίας και ώρας μετατράπηκε σε datetime object της βιβλιοθήκης pandas της Python, ώστε να μπορούμε να τη διαχειριστούμε καλύτερα. Έτσι, τυπώνοντας την μικρότερη και τη μεγαλύτερη ημερομηνία και ώρα θέσαμε ως μια αρχική ημερομηνία την 1/1/2012 και ώρα 00:00 και τελική ημερομηνία 31/12/2013 και ώρα 23:30, ώστε να έχουμε στη διάθεσή της δεδομένα 2 ετών και κρατήσαμε τα αντίστοιχα ids που είχαν όλες τις μετρήσεις διαθέσιμες σε αυτό το διάστημα. Με αυτό τον τρόπο όσα ids είχαν λιγότερες από τις επιθυμητές μετρήσεις δεν συμπεριλήφθηκαν. Επίσης όσες εγγραφές παραπάνω υπήρχαν, αλλά χωρίς να περιλαμβάνουν κάποια τιμή στη στήλη kWh, δηλαδή null, αφαιρέθηκαν για να καθαριστούν τα δεδομένα. Στο τέλος κρατήσαμε 3 ids που περιλάμβαναν όλες τις

τιμές που ζητούσαμε και δημιουργήθηκε ένας πίνακας μεγέθους 366+365=731 γραμμών που αντιπροσωπεύουν τις ημέρες και 48 στηλών που αντιπροσωπεύουν τις 48 μετρήσεις κατανάλωσης που έγιναν σε κάθε μέρα με διάστημα μισάωρου.

4.4.3 Δημιουργία δεδομένων φαινομένου κλοπής

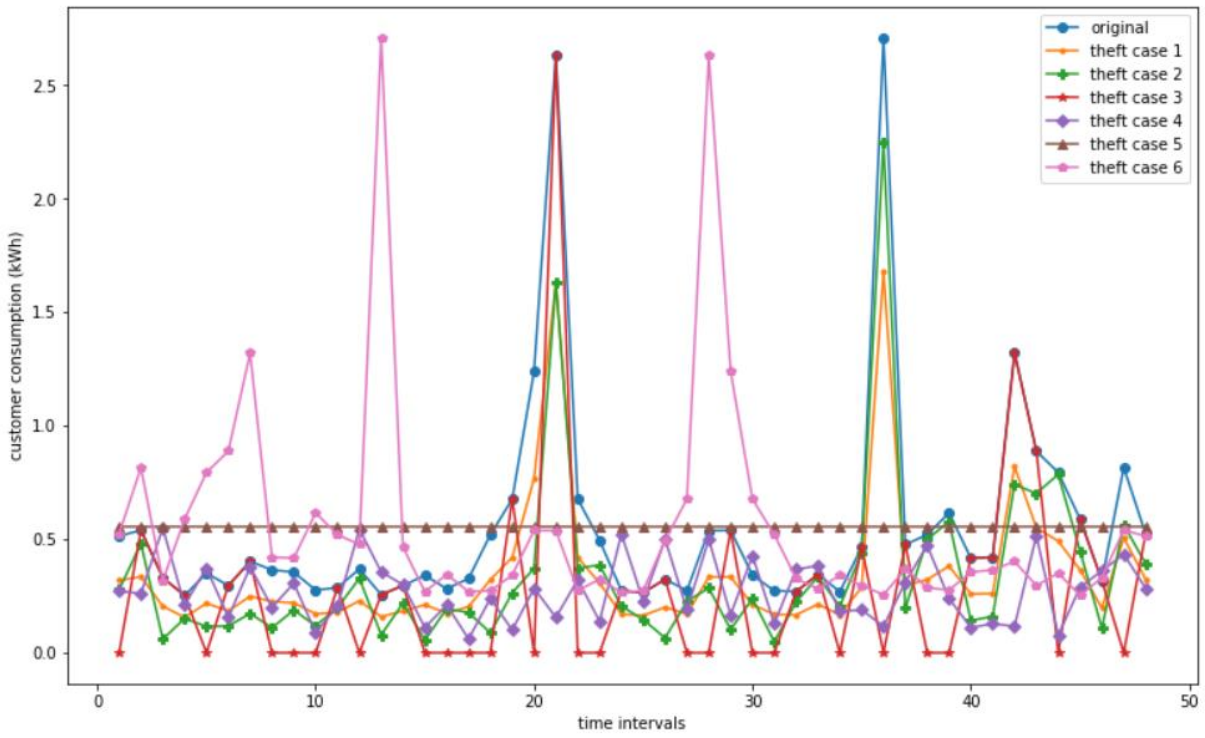
Σύμφωνα με την έρευνα των Jokar και Arianpoo [32], επειδή τα δεδομένα κατανάλωσης έχουν ληφθεί με τη συγκατάθεση των συμμετεχόντων στις δοκιμές, θεωρείται ότι αυτά είναι αληθή και δεν περιλαμβάνουν δεδομένα κλοπής (theft case data). Επομένως τα τελευταία θα πρέπει να δημιουργηθούν με τεχνητό τρόπο βασιζόμενα σε γνωστές συμπεριφορές κλοπής. Οι Punmiya και Choe [42], βασιζόμενοι στην έρευνα των παραπάνω, προτείνουν έξι διαφορετικές περιπτώσεις δεδομένων ρευματοκλοπής που προκύπτουν από τα πραγματικά δεδομένα που έχουμε στη διάθεσή μας και οι οποίες χρησιμοποιούνται και σε αυτή την εργασία. Αυτές είναι οι παρακάτω, όπου x_t είναι η πραγματική τιμή κατανάλωσης με $t \in [1, 48]$:

- Theft case 1: $h_1(x_t) = x_t \times a$, a=random(0.1, 0.9)
- Theft case 2: $h_2(x_t) = x_t \times b_t$, b_t=random(0.1, 1]
- Theft case 3: $h_3(x_t) = x_t \times c_t$, c_t=random 0 or 1
- Theft case 4: $h_4(x_t) = \text{mean}(x) \times d_t$, d_t=random(0.1, 1]
- Theft case 5: $h_5(x_t) = \text{mean}(x)$
- Theft case 6: $h_6(x_t) = x_{T-t} = x_{48-t}$

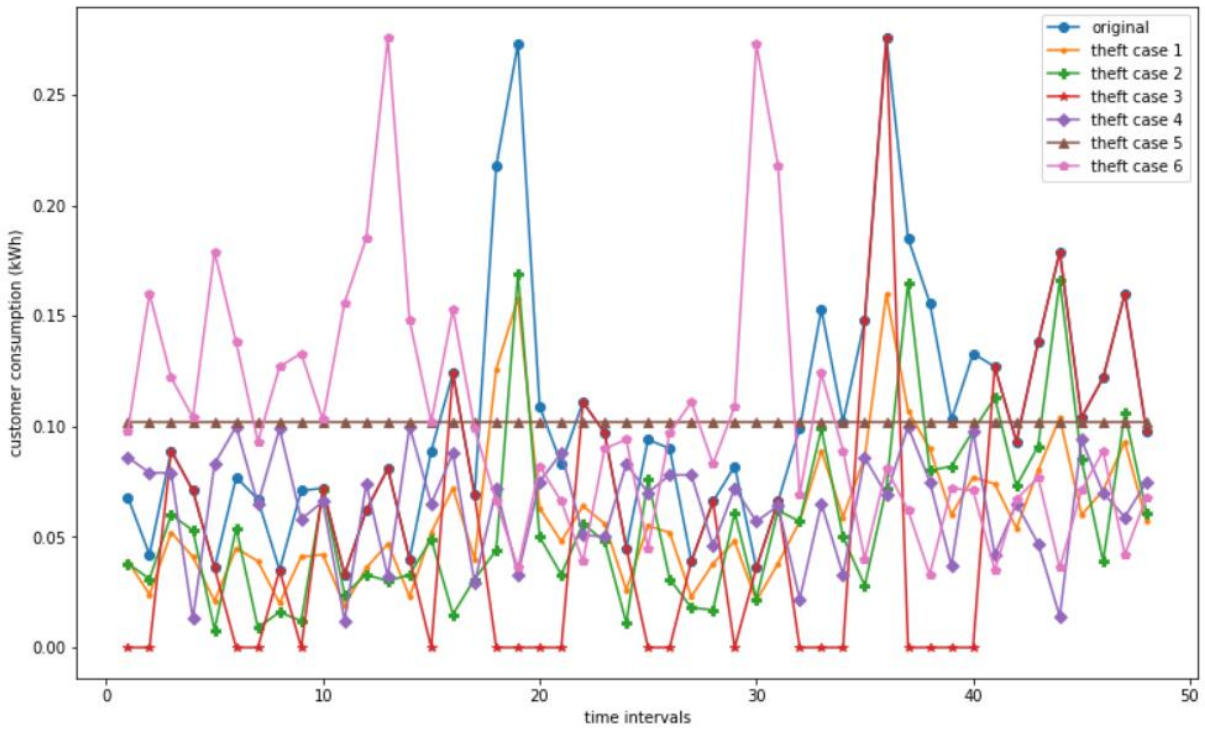
Σε κάθε περίπτωση αυτό που γίνεται ξεκάθαρο είναι ότι δηλώνεται πάντα ένα μέρος της ενέργειας που καταναλώνεται. Στην περίπτωση h_1 όλες οι μετρήσεις της ημέρας πολλαπλασιάζονται με την ίδια τυχαία μεταβλητή, οπότε και δηλώνεται ένα κλάσμα της καταναλισκόμενης ισχύος. Στην h_2 κάθε μέτρηση πολλαπλασιάζεται με ένα διαφορετικό τυχαίο ποσοστό. Η h_3 είναι η περίπτωση όπου ο χρήστης δεν φαίνεται να καταναλώνει καθόλου ενέργεια σε τυχαίες χρονικές στιγμές ή δηλώνει την πραγματική κατανάλωση. Στις περιπτώσεις h_5 και h_4 δηλώνεται είτε η μέση τιμή της ημερήσιας κατανάλωσης, είτε η μέση τιμή πολλαπλασιασμένη με μια τυχαία μεταβλητή σε κάθε χρονική στιγμή. Και τέλος στην h_6 έχουμε αντίστροφη δήλωση

των μετρήσεων, ειδικά για τις περιπτώσεις όπου γίνεται ειδική, πιο χαμηλή, τιμολόγηση σε ώρες εκτός αιχμής, κυρίως δηλαδή τις βραδινές ώρες.

Στη συνέχεια, για κάθε πίνακα δεδομένων κάθε καταναλωτή, και από τα δυο σετ, δημιουργήθηκαν άλλοι έξι πίνακες, μεταμορφωμένοι με βάση τις παραπάνω έξι περιπτώσεις ρευματοκλοπής. Έτσι για κάθε καταναλωτή έχουμε συνολικά 7 πίνακες, ένα με τα πραγματικά δεδομένα και έξι με τις περιπτώσεις κλοπής. Παρακάτω φαίνονται δυο διαγράμματα πραγματικής κατανάλωσης μαζί με αυτές της ρευματοκλοπής, ένας καταναλωτής από κάθε σετ (ISSDA και UK Power Networks) σε μια τυχαία μέρα (Σχήματα 4.5 και 4.6).



Σχήμα 4.5: Ημερήσια κατανάλωση και διαμόρφωση σε περιπτώσεις κλοπής (ISSDA dataset)



Σχήμα 4.6: Ημερήσια κατανάλωση και διαμόρφωση σε περιπτώσεις κλοπής (UK Power Networks dataset)

4.5 Εφαρμογή του αλγορίθμου XGBoost

Εφόσον αποκτήθηκαν και οι 7 πίνακες για κάθε καταναλωτή, όπως περιγράφηκε παραπάνω, το επόμενο βήμα είναι η συνένωσή τους σε ένα ενιαίο σετ και η προσθήκη ετικετών (labels). Για τα πραγματικά δεδομένα θέτουμε την ετικέτα 0 και για τα τεχνητά δεδομένα κλοπής την ετικέτα 1. Έτσι ο τελικός πίνακας σε μορφή pandas dataframe μοιάζει όπως στον Πίνακα 4.3.

Πίνακας 4.3: Τελικός πίνακας που περιλαμβάνει τα labeled πραγματικά και τεχνητά δεδομένα κλοπής

	0	1	2	3	4	5	6	7	8	9	...	39	40	41	42	43	44	45	46	47	class
0	0.160	0.054	0.045	0.069	0.084	0.016	0.078	0.075	0.045	0.035	...	0.182	0.135	0.125	0.190	0.155	0.119	0.153	0.188	0.103	0
1	0.091	0.074	0.082	0.014	0.086	0.069	0.047	0.033	0.096	0.055	...	0.156	0.141	0.194	0.179	0.133	0.149	0.190	0.108	0.123	0
2	0.151	0.060	0.035	0.083	0.066	0.036	0.066	0.095	0.036	0.043	...	0.163	0.132	0.106	0.185	0.168	0.139	0.151	0.164	0.109	0
3	0.065	0.090	0.054	0.034	0.080	0.075	0.033	0.063	0.078	0.043	...	0.164	0.140	0.190	0.168	0.135	0.161	0.189	0.113	0.091	0
4	0.093	0.065	0.034	0.057	0.088	0.040	0.034	0.082	0.071	0.033	...	0.192	0.202	0.150	0.134	0.182	0.157	0.134	0.144	0.117	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5112	0.061	0.062	0.060	0.092	0.073	0.078	0.123	0.115	0.144	0.115	...	0.088	0.082	0.111	0.085	0.098	0.044	0.042	0.060	0.025	1
5113	0.104	0.186	0.274	0.196	0.215	0.194	0.206	0.171	0.184	0.218	...	0.054	0.043	0.042	0.061	0.034	0.110	0.086	0.102	0.103	1
5114	0.093	0.118	0.157	0.140	0.217	0.178	0.214	0.216	0.238	0.192	...	0.087	0.071	0.049	0.021	0.057	0.027	0.045	0.087	0.084	1
5115	0.101	0.180	0.198	0.207	0.197	0.162	0.130	0.167	0.136	0.217	...	0.080	0.094	0.041	0.032	0.056	0.063	0.111	0.081	0.100	1
5116	0.220	0.194	0.211	0.221	0.190	0.218	0.219	0.218	0.197	0.206	...	0.061	0.093	0.085	0.106	0.070	0.060	0.028	0.053	0.075	1

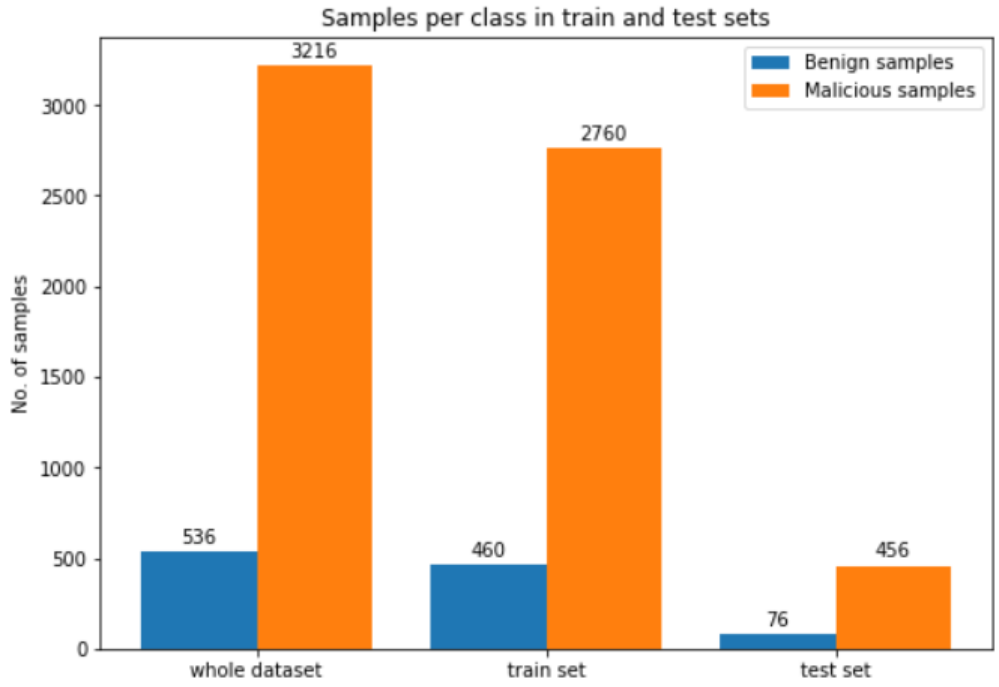
5117 rows × 49 columns

Για τα ISSDA δεδομένα θα έχουμε λοιπόν 536*7=3752 γραμμές ενώ για τα UK Power Networks (UKPN) δεδομένα θα έχουμε 731*7=5117 γραμμές.

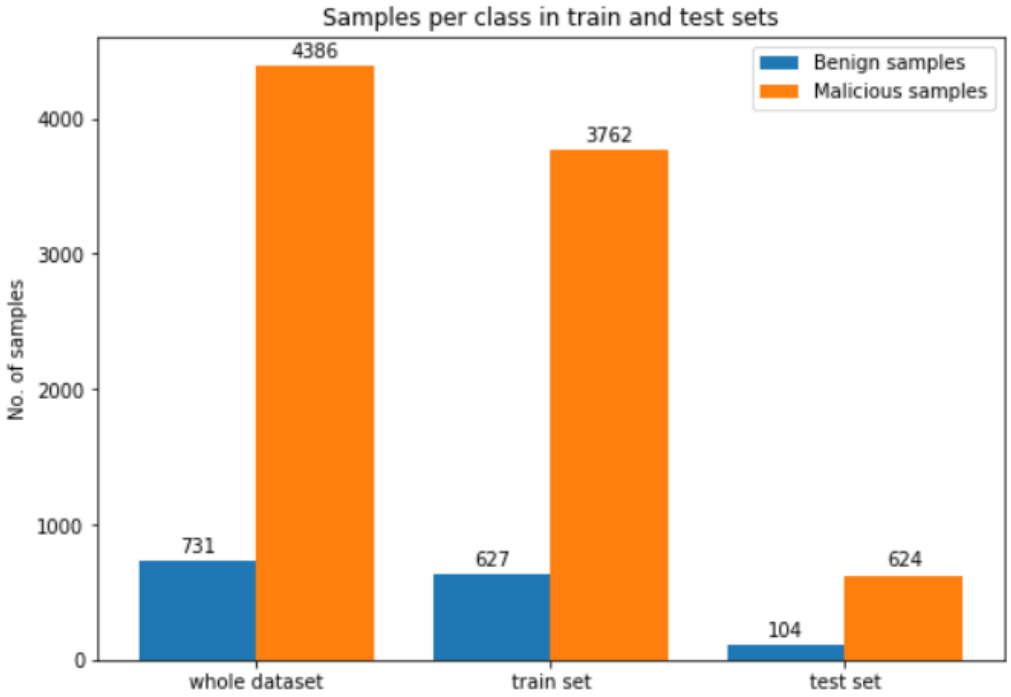
Το επόμενο βήμα είναι ο διαχωρισμός σε training (εκπαίδευση) και test sets.

4.5.1 Training και Test σετ

Οι Punmiya και Choe έχουν ορίσει ως το μέγεθος του testing set στο 14% περίπου, καθώς θέτουν 59 στις 420 εγγραφές ως test set και εφόσον γνωρίζουμε ότι βασίστηκαν στην έρευνα των Jokar και Arianpoo, χωρίσαμε τα δεδομένα με αντίστοιχο τρόπο. Οι τελευταίοι ξεχώρισαν μια μέρα της εβδομάδας για το test set για κάθε βδομάδα και τις υπόλοιπες έξι τις συμπεριλαμβάνουν στο σετ εκπαίδευσης. Εφαρμόζοντας αυτή τη λογική, για κάθε εφτά μέρες του dataset παίρνουμε μια μέρα με τυχαίο τρόπο για το test set. Έτσι στο ISSDA σετ έχουμε 460 μέρες στο training set και 76 στο test set, ενώ στο UKPN έχουμε 4389 samples στο training set και 728 samples στο test set. Επίσης στο ISSDA training set υπάρχουν 2760 μέρες κλοπής και στο test set 456 μέρες, ενώ στο UKPN έχουμε 3762 μέρες κλοπής στο σετ εκπαίδευσης και 624 μέρες στο test set. Αυτό μπορεί να γίνει πιο ξεκάθαρο με τα Σχήματα 4.7 και 4.8. Επίσης μετά το διαχωρισμό έχουμε τα training και test sets στη μορφή πινάκων, όπου τα features ή μεταβλητές βρίσκονται σε ένα πίνακα, τον X_train και X_test, και οι στήλες των ετικετών των κλάσεων σε μορφή διανυσμάτων y_train και y_test.



Σχήμα 4.7: Εγγραφές ανά κλάση για το ISSDA data set



Σχήμα 4.8: Εγγραφές ανά κλάση για το UK Power Networks data set

Από τα παραπάνω σχήματα γίνεται αντιληπτό ότι υπάρχει μεγάλη αναλογία πραγματικών δεδομένων προς αυτά της κλοπής, συγκεκριμένα 1:6. Αυτό σημαίνει ότι

για μια γραμμή/μέρα πραγματικής κατανάλωσης αντιστοιχούν στο σετ 6 μέρες κλοπής. Αυτή η αναλογία μπορεί να κάνει τον αλγόριθμο μεροληπτικό ως προς την κλάση με τα περισσότερα δείγματα και στη συγκεκριμένη περίπτωση τα δείγματα κλοπής. Γι' αυτό το λόγο και χρησιμοποιούνται αλγόριθμοι εξισορρόπησης των κλάσεων σε μια επιθυμητή αναλογία. Ένας από αυτούς είναι και ο SMOTE (Synthetic Minority Over-sampling Technique) [55].

4.5.2 Χρήση του SMOTE για εξισορρόπηση των κλάσεων στο training σετ

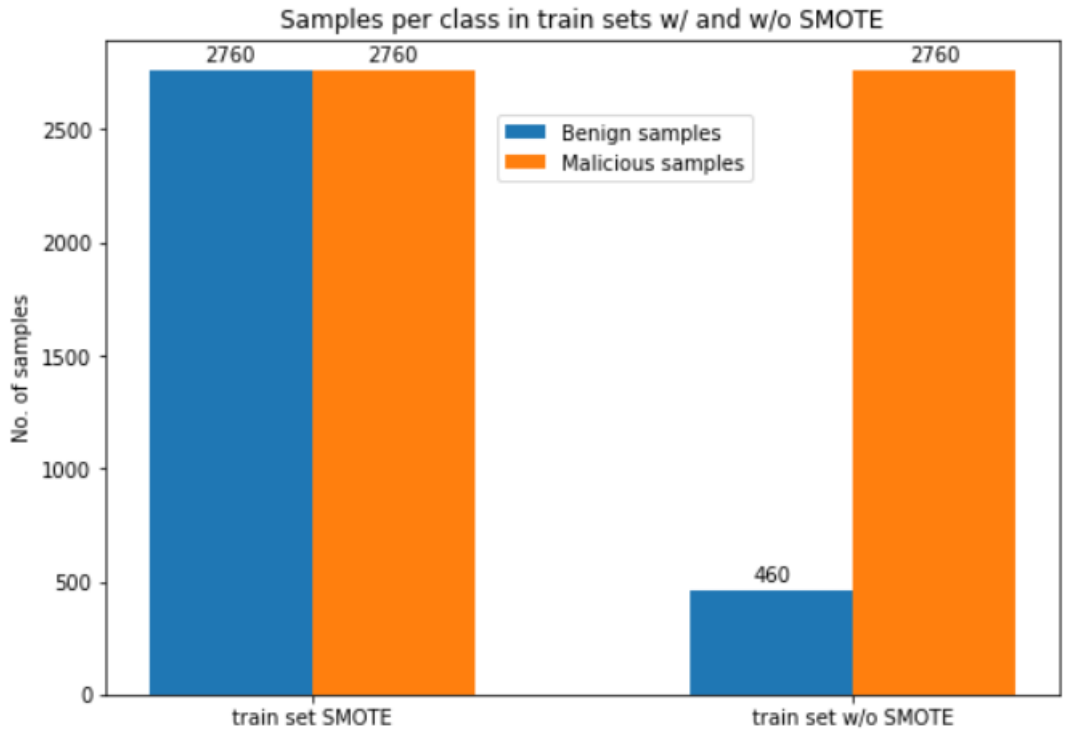
Το πρόβλημα των μη ισορροπημένων κλάσεων μπορεί να γίνει αισθητό σε περιπτώσεις όπως παραπάνω και να δημιουργήσει εμπόδιο στη σωστή κατηγοριοποίηση της κλάσης με τα λιγότερα στοιχεία (minority class). Όταν παρατηρήσουμε ότι πέρα από το ποσοστό των προβλέψεων, το οποίο μπορεί να είναι υψηλό, τα στοιχεία της minority κλάσης δεν γίνονται σωστά classify σημαίνει ότι πρέπει να ληφθούν παραπάνω μέτρα. Υπάρχουν διάφορες τεχνικές [56] για την εξισορρόπηση των κλάσεων, όπως για παράδειγμα η αντιγραφή των ήδη υπαρχόντων στοιχείων στη minority κλάση (replicating/resampling) τόσες φορές όσες να εξαλειφθεί η μεγάλη αναλογία. Επίσης, μπορεί να γίνει μείωση των στοιχείων της τάξης που έχει πλεονασμό (majority class), ώστε να φτάσει ένα επίπεδο που η αναλογία θα έρθει πιο κοντά στο 50%. Παρόλα αυτά τέτοιου είδους τεχνικές έχουν τα μειονεκτήματά τους. Στην περίπτωση του resampling υπάρχει κίνδυνος για overfitting, καθώς ο αλγόριθμος που θα εφαρμοστεί σε αυτά εκπαιδεύεται στην ουσία με τα ίδια και τα ίδια σημεία. Στη δεύτερη περίπτωση, όπως είναι λογικό, υπάρχει απώλεια πληροφορίας όταν απλά μειώνουμε τις εγγραφές του σετ δεδομένων.

Με τον SMOTE δημιουργούνται στην ουσία νέα συνθετικά στοιχεία της minority κλάσης, τα οποία «μοιάζουν» με γειτονικά στοιχεία της κλάσης στην οποία ανήκουν. Για τη λειτουργία του SMOTE αρκεί να καταλάβουμε ότι τα νέα στοιχεία βρίσκονται πάνω στη νοητή γραμμή που δημιουργείται μεταξύ των k κοντινότερων γειτόνων (nearest neighbors) των στοιχείων της minority κλάσης και διαλέγονται τυχαία μεταξύ των k κοντινότερων γειτόνων. Ένα από τα μειονεκτήματα της χρήσης αυτού

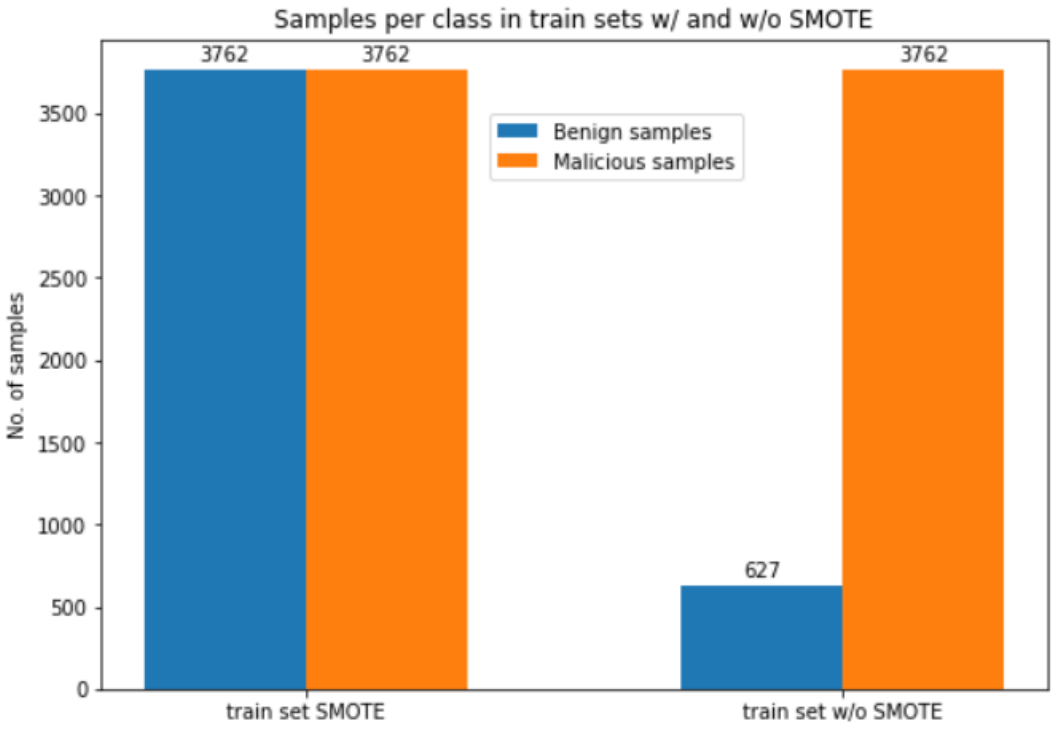
του αλγόριθμου είναι ότι μπορεί να δημιουργηθούν δείγματα που μοιάζουν πολύ ή συμπίπτουν με δείγματα της majority κλάσης (overlapping).

Επίσης καλό είναι να σημειώσουμε ότι η εφαρμογή του SMOTE και οποιασδήποτε άλλης τεχνικής oversampling πρέπει να γίνει στο training set και όχι σε όλο το dataset, καθώς έτσι έχουμε διαρροή των συνθετικών δεδομένων στο test set. Έτσι κι αλλιώς η ανισορροπία των κλάσεων μας επηρεάζει στη διαδικασία εκπαίδευσης του αλγόριθμου και όχι στην αξιολόγησή του, που γίνεται με το test set.

Με τη χρήση του SMOTE με τις default παραμέτρους (k=5 για τον αριθμό των κοντινότερων γειτόνων) έχουμε τα παρακάτω αποτελέσματα στη διαμόρφωση των δεδομένων όπως φαίνονται στα Σχήματα 4.9 και 4.10 για τα δύο datasets, δίνοντας ως είσοδο τα X_train και y_train.



Σχήμα 4.9: Αριθμός δειγμάτων ανά κλάση στο train set με και χωρίς τη χρήση του SMOTE (ISSDA set)



Σχήμα 4.10: Αριθμός δειγμάτων ανά κλάση στο train set με και χωρίς τη χρήση του SMOTE (UKPN set)

4.5.3 Μετρικές αξιολόγησης

Για την αξιολόγηση ενός αλγορίθμου είναι απαραίτητο να θέσουμε ένα τρόπο μέτρησης ή μια σειρά από μετρικές (Evaluation metrics) που μπορούν να μας προμηθεύσουν με τις απαραίτητες πληροφορίες για να καταλάβουμε πόσο καλά δουλεύει ένα μοντέλο [57][58]. Στην προκειμένη περίπτωση θέλουμε να αξιολογήσουμε ένα μοντέλο κατηγοριοποίησης και οι μέθοδοι μέτρησης που πρέπει να γνωρίζουμε θα αναφερθούν παρακάτω αφού πρώτα αναλύσουμε τα στοιχεία ενός confusion matrix.

Κάνοντας χρήση της συνάρτησης metrics.confusion_matrix της βιβλιοθήκης scikit-learn, δίνοντας ως είσοδο τις τιμές των κλάσεων του test set (y_test) και των προβλεπόμενων τιμών μετά από τη χρήση ενός αλγορίθμου κατηγοριοποίησης (y_predictions) λαμβάνουμε έναν πίνακα της μορφής του Πίνακα 4.4.

Πίνακας 4.4: Confusion Matrix as produced from scikit-learn library

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Τα στοιχεία του πίνακα προκύπτουν από το ποια είναι η πραγματική κλάση στην οποία ανήκει μια εγγραφή του test set και ποια είναι η κλάση στην οποία τελικά κατηγοριοποιήθηκαν μετά την εκτέλεση του αλγόριθμου:

- TN (True Negatives): Τα στοιχεία που ανήκουν στην κλάση 0 και κατηγοριοποιήθηκαν στην ίδια κλάση
- TP (True Positives): Τα στοιχεία που ανήκουν στην κλάση 0 και κατηγοριοποιήθηκαν στην ίδια κλάση
- FP (False Positives): Τα στοιχεία που ανήκουν στην κλάση 0 και κατηγοριοποιήθηκαν στην κλάση 1
- FN (False Negatives): Τα στοιχεία που ανήκουν στην κλάση 1 και κατηγοριοποιήθηκαν στην κλάση 0

Με βάση τα χαρακτηριστικά του πίνακα προκύπτουν οι παρακάτω μετρικές αξιολόγησης:

- Accuracy: είναι ο λόγος των σωστών προβλέψεων προς όλες τις προβλέψεις και είναι ίση με $\frac{TP+TN}{TP+TN+FP+FN}$
- Recall/Sensitivity/True Positive Rate/Detection Rate: είναι λόγος των σωστών προβλέψεων της θετικής κλάσης προς όλα τα στοιχεία που ανήκουν πραγματικά στη θετική κλάση, δηλαδή $\frac{TP}{TP+FN}$ και στην ουσία συνοψίζει το πόσο καλά έχει γίνει ταξινόμηση της θετικής κλάσης

- False Positive Rate (FPR): είναι ο λόγος των λανθασμένα ταξινομημένων στοιχείων της αρνητικής κλάσης (FP) προς όλων των πραγματικά αρνητικών στοιχείων, δηλαδή $\frac{FP}{TN+FP}$
- False Negative Rate (FNR): είναι ο λόγος των λανθασμένα ταξινομημένων στοιχείων της θετικής κλάσης (FN) προς όλων των πραγματικά θετικών στοιχείων, δηλαδή $\frac{FN}{FN+TP}$
- Precision: είναι ο λόγος των στοιχείων που σωστά προβλέφθηκαν ότι ανήκουν στη θετική κλάση προς τον αριθμό όλων των στοιχείων που προβλέφθηκαν ότι ανήκουν στη θετική κλάση, δηλαδή $\frac{TP}{TP+FP}$. Συνοψίζει το πόσο ακριβές είναι το μοντέλο όσον αφορά τα στοιχεία που προβλέφθηκαν ως θετικά, κατά πόσο είναι και στην πραγματικότητα θετικά
- F-score/F1-score: είναι η αρμονική μέση τιμή των precision και recall και χρησιμοποιείται όταν θέλουμε να δούμε πόσο καλή ισορροπία υπάρχει μεταξύ των δυο
- Receiver Operating Characteristic (ROC) curve: είναι ένα διαγνωστικό διάγραμμα που οπτικοποιεί τη συμπεριφορά ενός binary classifier μοντέλου υπολογίζοντας το FPR και το TPR αλλάζοντας το κατώφλι κατηγοριοποίησης των στοιχείων. Είναι πρακτικά ένα διάγραμμα σήματος (TPR)-θορύβου (FPR). Το κατώφλι αφορά την τιμή της πιθανότητας ότι ένα στοιχείο θα προβλεφθεί ότι ανήκει στη θετική ή την αρνητική κλάση. Η εξ ορισμού τιμή του κατωφλίου είναι 0.5, οπότε αν π.χ. η πιθανότητα προκύψει μεγαλύτερη από αυτή την τιμή η πρόβλεψη θα είναι ότι το στοιχείο ταξινομείται στη θετική κλάση. Όταν ένας αλγόριθμος ταξινόμησης κάνει σωστή διάκριση μεταξύ των δυο κλάσεων η καμπύλη ROC θα πλησιάζει στην πάνω αριστερή γωνία
- ROC Area Under Curve (AUC): υπολογίζει την περιοχή κάτω από την καμπύλη ROC και μας δίνει ένα ποσοστό για το πόσο μεγάλη είναι η περιοχή που καλύπτεται κάτω από την καμπύλη. Όσο πιο κοντά στο 1 τόσο πιο καλό είναι το μοντέλο

4.5.4 Δημιουργία βασικών μοντέλων και ρύθμιση υπερ-παραμέτρων

Τα βήματα που ακολουθούνται για την εύρεση του κατάλληλου μοντέλου για κάθε καταναλωτή είναι:

- Δημιουργία ενός classifier XGBoost με τις επιθυμητές σταθερές παραμέτρους
- Αναζήτηση των τιμών ορισμένων παραμέτρων για βελτιστοποίηση της απόδοσης του μοντέλου με αναζήτηση πλέγματος (grid search)
- Χρήση της native cross validation μεθόδου για εύρεση του βέλτιστου αριθμού estimators
- Εκπαίδευση (training) του επιλεγμένου μοντέλου με τα δεδομένα εκπαίδευσης (fitting του μοντέλου στα δεδομένα)
- Πρόβλεψη της κλάσης των στοιχείων του test set με το εκπαιδευμένο μοντέλο

Σε πρώτο στάδιο, πριν αρχίσουμε να πειραματιζόμαστε με τις τιμές των παραμέτρων του αλγόριθμου XGBoost, δημιουργούμε ένα μοντέλο classification της scikit-learn βιβλιοθήκης ως βάση (βασικό μοντέλο-base model) με τις default τιμές εκτός από το βάθος του δέντρου που το θέτουμε στο 5 αντί για 3 (πολύ ρηχό δέντρο) που είναι η εξ ορισμού τιμή, μιας και είναι μια καλή προτεινόμενη τιμή για αρχή, και με n_estimators ίσο με 500 αντί για 100 που είναι η default τιμή. Οι estimators αφορούν τον αριθμό των δέντρων που δημιουργούνται κατά τη διάρκεια της εκπαίδευσης.

Κάνουμε χρήση της native cross validation (CV) μεθόδου από τη βασική βιβλιοθήκη του XGBoost (για την οποία χρειάζεται τα δεδομένα του training set να βρίσκονται στην εσωτερική δομή που χρησιμοποιεί η βιβλιοθήκη και τα έχουμε μετατρέψει πιο μπροστά σε dmatrix), θέτοντας τις παραμέτρους του μοντέλου που είχαμε πριν, και μεταξύ των αποτελεσμάτων του CV μπορούμε να βρούμε το βέλτιστο αριθμό estimators που χρειάστηκαν κατά τη διάρκεια αυτού. Θέτοντας αυτό τον αριθμό στην αντίστοιχη παράμετρο του μοντέλου, ακολουθεί η εκπαίδευσή του με τα δεδομένα του σετ εκπαίδευσης, όπως έχουν διαμορφωθεί μετά τη χρήση του SMOTE.

Να σημειώσουμε εδώ πως κατά τη διάρκεια της εκπαίδευσης του μοντέλου αλλά και κατά το CV, η τεχνική που χρησιμοποιείται για τη μέτρηση της απόδοσης είναι το AUC.

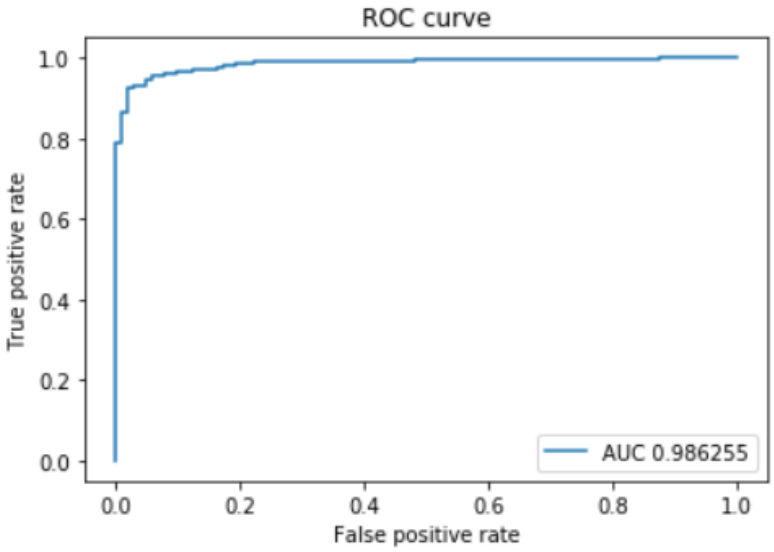
Επίσης, η native CV μέθοδος του XGBoost μας δίνει τη δυνατότητα να χρησιμοποιήσουμε μια μεταβλητή που λέγεται early_stopping_rounds και με την οποία το CV σταματά αν δεν υπάρχει βελτίωση του AUC για τον αριθμό γύρων που έχουμε θέσει μέσω αυτής, πριν φτάσουμε τον αρχικό αριθμό estimators που είχαμε θέσει στο μοντέλο. Αν βέβαια υπάρχει βελτίωση, το CV θα σταματήσει στον αριθμό των estimators και γι' αυτό στα αποτελέσματα θα δούμε ότι ο βέλτιστος αριθμός είναι αυτός που είχαμε θέσει από την αρχή.

Επόμενο στάδιο αποτελεί η πρόβλεψη των κλάσεων των στοιχείων του test set με το μοντέλο που έχει διαμορφωθεί και η αξιολόγησή του με ορισμένες από τις μετρικές που αναφέρθηκαν στην αντίστοιχη ενότητα: confusion matrix, Accuracy, Recall, Precision, FPR, FNR και το ROC curve με την αντίστοιχη τιμή για το AUC. Τα αποτελέσματα των base models για όλους τους πελάτες που εξετάστηκαν φαίνονται στον Πίνακα 4.5.

Πίνακας 4.5: Αριθμητικά αποτελέσματα των base models (με μπεζ χρώμα είναι οι καταναλωτές του UKPN

	Κατ. 1	*Κατ. 2*	*Κατ. 3*	*Κατ. 4*	*Κατ. 5*	*Κατ. 6*	*Κατ. 7*
Accuracy	93.05%	95.43%	98.31%	93.8%	95.88%	95.33%	95.6%
AUC	94.42%	98.7%	99.6%	97.1%	98.84%	98.55%	98.63%
Recall	94.3%	96.93%	98.25%	95.83%	96.47%	96.63%	96.15%
Precision	97.51%	98.88%	99.78%	96.9%	98.69%	97.89%	98.68%
FPR	14.47%	6.58%	1.32%	18.42%	7.69%	12.5%	7.69%
FNR	5.7%	3.07%	1.75%	4.17%	3.53%	3.37%	3.85%

Ένα παράδειγμα του ROC curve ενός καταναλωτή στο base model είναι το Σχήμα 4.11.



Σχήμα 4.11: ROC curve και AUC score ενός από τους καταναλωτές (καταναλωτής 7)

Το confusion matrix του ίδιου καταναλωτή φαίνεται στον Πίνακα 4.6.

Πίνακας 4.6: Confusion matrix (καταναλωτής 7)

	PREDICTED 0	**PREDICTED 1**
ACTUAL 0	96	8
ACTUAL 1	24	600

Βλέποντας τα παραπάνω νούμερα μπορούμε να διακρίνουμε πως ακόμα και με τις default τιμές των παραμέτρων η απόδοση του αλγόριθμου είναι παραπάνω από ικανοποιητική. Οι μετρικές στις οποίες δίνουμε περισσότερη βάση είναι κυρίως το AUC και το recall. Ένα πολύ καλό AUC score υποδεικνύει ότι το ποσοστό των λανθασμένα ταξινομημένων στοιχείων στη θετική κλάση (FPR) σε αναλογία με το TPR/recall είναι αρκετά μικρό. Όσον αφορά το recall, δίνουμε βάση γιατί εξαρτάται από το πλήθος των False Negatives. Τα FN έχουν μεγάλη σημασία, γιατί σημαίνει ότι κάποιος καταναλωτής που στην πραγματικότητα εκείνη την ημέρα είχε παρουσιάσει παράνομη συμπεριφορά, άρα ανήκει στην κλάση 1, προβλέπεται ότι ανήκει στην κλάση 0, που υποδηλώνει μη κλοπή ρεύματος. Επομένως ένας τέτοιος χρήστης συνεχίζει να κλέβει όσο κατηγοριοποιείται στην κλάση 0 και το κόστος για την

εταιρεία παροχής ενέργειας είναι στην ουσία μεγαλύτερο. Άρα κύριο μέλημά μας είναι να μειωθούν τα FN όσο δυνατόν περισσότερο και κατά συνέπεια το recall να αυξηθεί, χωρίς βέβαια να έχουμε overfitting.

4.5.4.1 Hyperparameter tuning

Στην περίπτωση που θέλουμε να βελτιώσουμε την απόδοση ενός αλγορίθμου μπορούμε να ακολουθήσουμε μια διαδικασία ρύθμισης των παραμέτρων (hyperparameter tuning) του μοντέλου με αναζήτηση πλέγματος (grid search). Ακολουθώντας τη βασική μεθοδολογία από το tutorial του Aarshay Jain στο Analytics Vidhya [59], αλλά και από άλλα αντίστοιχα tutorials ([60], [61], [45]) σε δοκιμές στους πρώτους τρεις πελάτες ακολουθήσαμε τα εξής βήματα:

- Ρύθμιση των max_depth και min_child_weight, καθώς αυτοί συνδέονται με την πολυπλοκότητα των δέντρων και ορίζουν την ισορροπία μεταξύ του bias και της διακύμανσης (variance) που σχετίζονται με το over- και underfitting. Το πεδίο τιμών στην αναζήτηση είναι: max_depth (3,9), min_child_weight (1,5).
- Ρύθμιση των subsample και colsample_bytree, για να ελέγξουμε τι ποσοστό των δειγμάτων (samples) και των στοιχείων (features) μπορεί να χρησιμοποιηθεί στο χτίσιμο ενός δέντρου. Το πεδίο τιμών στην αναζήτηση είναι: subsample (0.7, 1), colsample_bytree (0.6, 0.9).
- Ρύθμιση του gamma που είναι χρήσιμο στην περίπτωση που τα δέντρα είναι σχετικά ρηχά. Το πεδίο τιμών στην αναζήτηση είναι: gamma (0, 0.08).
- Ρύθμιση των reg_alpha και reg_lambda. Το πεδίο τιμών στην αναζήτηση είναι: reg_alpha [0, 0.01, 0.05, 0.1], reg_lambda (1, 1.3).
- Ρύθμιση του βαθμού εκπαίδευσης learning_rate και των estimators, οι οποίοι συνδέονται με σχεδόν αντίστροφη αναλογία, δηλαδή αν μειωθεί ο βαθμός στον οποίο το μοντέλο μαθαίνει τα δεδομένα τότε χρειάζεται να δημιουργηθούν περισσότερα δέντρα για να βελτιωθεί η απόδοση του μοντέλου. Για τους πρώτους 3 πελάτες η ρύθμιση δεν έγινε με grid search, αλλά σε ζεύγη τιμών [learning_rate, n_estimators] με το learning_rate στο μισό και στο 1/10 από το

αρχικό, που ήταν στο 0.1: [0.05, 1500], [0.01, 5000] για τον πρώτο πελάτη, [0.05, 2000], [0.01, 10000] για τον δεύτερο και [0.05, 600], [0.01, 900] για τον τρίτο.

Σε κάθε στάδιο της ρύθμισης θέταμε στο μοντέλο που ρυθμίζαμε τις τιμές των παραμέτρων που είχαν προκύψει από τα προηγούμενα βήματα. Μόλις ολοκληρωνόταν μια ρύθμιση αξιολογούσαμε το μοντέλο όπως είχε διαμορφωθεί μέχρι τότε. Τα μοντέλα με τα καλύτερα αποτελέσματα για κάθε πελάτη θα αναφερθούν στην επόμενη ενότητα.

Στους υπόλοιπους τέσσερεις πελάτες προσπαθήσαμε αρχικά να δημιουργήσουμε ένα ενιαίο grid search με όλες τις παραμέτρους στο πλέγμα όμως ο χρόνος που απαιτούνταν ήταν μεγάλος για την υπολογιστική δύναμη που διαθέτουμε. Έτσι δημιουργήσαμε μια μέθοδο ρύθμισης τριών σταδίων με βάση την ομαδοποίηση των παραμέτρων και σε κάθε στάδιο ανανεώναμε τις τιμές των παραμέτρων του μοντέλου με αυτές του προηγούμενου σταδίου. Η ομαδοποίηση είχε ως εξής:

- Οι παράμετροι του πρώτου σετ και το πεδίο τιμών τους είναι:
    - max_depth (3, 7)
    - min_child_weight (1, 4)
    - subsample (0.6, 0.9)
    - colsample_bytree (0.7, 1)
- Οι παράμετροι του δεύτερου σετ και το πεδίο τιμών τους είναι:
    - Gamma (0, 0.04)
    - reg_alpha [0, 0.01, 0.05, 0.1, 0.15]
    - reg_lambda (1, 1.2)
- Αφού γίνουν οι παραπάνω δύο ρυθμίσεις, αξιολογούμε το μοντέλο και έπειτα εκτελούμε την τελευταία αναζήτηση για τις παρακάτω παραμέτρους στα παρακάτω πεδία τιμών:
    - learning_rate [0.01, 0.05],
    - n_estimators [800, 1000, 1500, 2000]

Στα πρώτα δύο σετ το learning_rate και οι estimators έχουν τις σταθερές τιμές 0.1 και 1100 αντίστοιχα.

Με την παραπάνω μέθοδο χρειάστηκαν κατά μέσο όρο 57 λεπτά για το πρώτο σετ παραμέτρων, 17 λεπτά για το δεύτερο και 6.5 λεπτά για το τρίτο.

4.5.5 Αριθμητικά αποτελέσματα από την εφαρμογή του XGBoost

Παρακάτω βλέπουμε στους δύο Πίνακες 4.7 και 4.8 τα αριθμητικά αποτελέσματα των καλύτερων μοντέλων για κάθε καταναλωτή όπως προέκυψαν και μετά από το hyperparameter tuning.

Πίνακας 4.7: Αριθμητικά αποτελέσματα των best models (με μπεζ χρώμα είναι οι καταναλωτές του UKPN)

	Κατ. 1	*Κατ. 2*	*Κατ. 3*	*Κατ. 4*	*Κατ. 5*	*Κατ. 6*	*Κατ. 7*	*Μ.Ο.*
Accuracy (%)	93.23	96.62	98.12	94.74	96.29	96.43	95.6	95.86
AUC (%)	97.36	98.79	99.7	97.27	98.95	98.65	98.63	98.48
Recall (%)	94.52	97.15	98.25	96.27	96.96	97.44	96.15	96.68
Precision (%)	97.51	98.88	99.56	97.56	98.69	98.38	98.68	98.47
FPR (%)	14.47	6.57	2.63	14.47	7.69	9.62	7.69	9.02
FNR (%)	5.48	2.85	1.75	3.73	3.05	2.56	3.85	3.32

Πίνακας 4.8: Τιμές των παραμέτρων των best models (με μπεζ χρώμα είναι οι καταναλωτές του UKPN)

	Κατ. 1	*Κατ. 2*	*Κατ. 3*	*Κατ. 4*	*Κατ. 5*	*Κατ. 6*	*Κατ. 7*
Max_depth	5	3	5	5	7	5	5
Min_child_weight	1	1	3	1	1	1	1
Subsample	0.9	0.8	0.6	0.8	0.6	0.7	1
Colsample_bytree	1	0.8	0.8	0.7	1	0.9	1
Gamma	0.03	0.05	0.02	0.04	0	0.01	0
Alpha/lambda	0/1	0.1/1.1	0.15/1.1	0/1	0/1	0.1/1.1	0/1
Learning_rate	0.01	0.1	0.1	0.1	0.1	0.1	0.1
N_estimators	2070	399	149	234	325	401	300

Αξίζει να σημειωθεί ότι οι τιμές από τα βασικά μοντέλα δεν διαφέρουν ιδιαίτερα, αλλά είναι λογικό αν υπολογίσουμε ότι ήδη η απόδοση ήταν πολύ καλή. Η επιλογή των καλύτερων μοντέλων έγινε κυρίως με κριτήριο τους confusion matrices και των AUC και recall από τις μετρικές, όπως περιγράφηκε στην προηγούμενη ενότητα. Το γεγονός επίσης ότι έχουμε το ίδιο καλά αποτελέσματα σε δύο διαφορετικά σετ δεδομένων είναι εξίσου ενθαρρυντικό, αλλά μας κάνει επίσης να αναρωτηθούμε αν υπάρχει τρόπος να δούμε σημαντική πτώση των σκορ. Οι πιθανοί τρόποι που μελετήθηκαν αναφέρονται στην επόμενη ενότητα.

4.5.6 Πιθανά σενάρια διαμόρφωσης δεδομένων

Παίρνοντας ως βάση τον καταναλωτή με τα καλύτερα αποτελέσματα (καταναλωτής 3) θέσαμε δύο πιθανά θέματα που θα μπορούσαν να οδηγήσουν σε πτώση των σκορ και μελετήσαμε τα αποτελέσματα που αυτά μπορεί να έχουν στα βασικά μοντέλα (base models).

Σενάριο 1: το training set να περιλαμβάνει τον πίνακα των πραγματικών δεδομένων και 3 πίνακες από τις περιπτώσεις κλοπής (συμπεριλήφθηκαν οι 1, 3 και 5), ενώ το test set περιελάμβανε όλες τις περιπτώσεις κλοπής, επομένως οι 2, 4, και 6 ήταν άγνωστες στο μοντέλο όταν εκτελέστηκε η πρόβλεψη. Επίσης στο training set εφαρμόστηκε ο SMOTE καθώς υπήρχε ανισορροπία κλάσεων με αναλογία 1/3 (πραγματικά/δεδομένα κλοπής). Οπότε στο training set έχουμε 1840 samples εκ των οποίων τα 1380 ανήκουν στην κλάση 1 και 460 στην κλάση 0.

Σενάριο 2: το training set να περιλαμβάνει τον πίνακα των πραγματικών δεδομένων και 2 πίνακες από τις περιπτώσεις κλοπής (συμπεριλήφθηκαν οι 1 και 5), ενώ το test set περιελάμβανε όλες τις περιπτώσεις κλοπής. Έγινε και εδώ εφαρμογή του SMOTE καθώς υπήρχε ανισορροπία κλάσεων με αναλογία ½ (πραγματικά/δεδομένα κλοπής) στο σετ εκπαίδευσης. Άρα στο training set έχουμε 1380 samples εκ των οποίων τα 920 ανήκουν στην κλάση 1 και 460 στην κλάση 0.

Σενάριο 3: όπως το σενάριο 2 ως προς τα είδη και το πλήθος των περιπτώσεων κλοπής στο training set, αλλά χωρίς τη χρήση SMOTE για ισορροπία κλάσεων.

Σενάριο 4: θέλαμε να δοκιμάσουμε την απόδοση του αλγόριθμου αν η αναλογία πραγματικών δειγμάτων και κλοπής ήταν αντίστροφη στο training set, δηλαδή ανά 6 πραγματικά δείγματα αντιστοιχεί 1 κλοπής, παίρνοντας ένα τυχαίο δείγμα 15 ημερών από κάθε σετ δεδομένων κλοπής ως δείγματα κλοπής για το σετ εκπαίδευσης, και χρησιμοποιήσαμε τον SMOTE για εξισορρόπηση των κλάσεων. Στο training set πριν τη χρήση του SMOTE έχουμε 536 πραγματικά δείγματα και 90 κλοπής.

Σενάριο 5: το σενάριο αυτό είναι ίδιο με το 4ο μόνο που δεν γίνεται χρήση του SMOTE.

Τα αποτελέσματα συνοψίζονται στον Πίνακα 4.9.

Πίνακας 4.9: Αποτελέσματα απόδοσης για τα 5 διαφορετικά σενάρια

	Σενάριο 1	*Σενάριο 2*	*Σενάριο 3*	*Σενάριο 4*	*Σενάριο 5*
Accuracy (%)	95.45	92	91.8	98.86	98.86
AUC (%)	99	97.7	97.6	1	99.67
Recall (%)	95.48	92	91.7	91.67	91.67

4.6 Εφαρμογή της AutoML

Για την εφαρμογή της AutoML επιλέχθηκε η χρήση του TPOT (Tree-Based Pipeline Optimization Tool) καθώς είναι πιο εύκολα διαχειρίσιμο λόγω της βιβλιοθήκης του που είναι βασισμένη στη scikit-learn και της διαθεσιμότητας του XGBoost μεταξύ των αλγορίθμων που δοκιμάζονται.

Τα δεδομένα που χρησιμοποιήθηκαν είναι ένα σετ ενός πελάτη από το ISSDA dataset και ενός πελάτη από UKPN. Έχει εφαρμοστεί η ίδια διαδικασία για το διαχωρισμό σε training και test sets, όπως και η χρήση του SMOTE.

Οι βασικές παράμετροι του αλγόριθμου που τροποποιούνταν στις δοκιμές είναι:

- Generations: αφορά τον αριθμό των «γενεών» ή αλλιώς επαναλήψεων που θα εκτελεστεί η διαδικασία βελτίωσης των pipelines
- Population_size: το πλήθος των ατόμων (pipelines) που θα διατηρηθούν σε κάθε γενιά (που θα περάσουν από την προηγούμενη γενιά στην επόμενη)

- Scoring: η μετρική για την επιλογή των καλύτερων pipelines κατά τη διαδικασία αξιολόγησης με cross validation και η οποία είναι το roc_auc
- Config_dict: το block στο οποίο περιλαμβάνουμε τις μεθόδους που επιθυμούμε να δοκιμαστούν από τον αλγόριθμο ακολουθούμενες από το block των παραμέτρων τους. Αν δεν θέσουμε κάποια συγκεκριμένη επιλογή τότε δοκιμάζονται όλα τα διαθέσιμα μοντέλα

Τα βήματα που ακολουθήθηκαν είναι τα παρακάτω:

- Δημιουργία ενός TPOT classifier όπου οι παράμετροι διαμορφώθηκαν ως εξής:
    - Generations: ο αλγόριθμος δοκιμάστηκε με τιμές 5, 10, 15 ή 20.
    - Population_size: το πλήθος ορίστηκε στο 50, οπότε είχαμε ανάλογα και με τα generations τη δημιουργία 300, 550, 800 ή 1050 pipelines
    - Config_dict: έγινε δοκιμή με τα παρακάτω blocks μοντέλων:
        - Config_dict 1: DecisionTreeClassifier, ExtraTreesClassifier, RandomForestClassifier, GradientBoostingClassifier, LogisticRegression, XGBClassifier
        - Config_dict 2: ExtraTreesClassifier, GradientBoostingClassifier, XGBClassifier
        - Config_dict 3: όλα τα διαθέσιμα μοντέλα
- Εκπαίδευση (fitting) του classifier στα εξισορροπημένα δεδομένα εκπαίδευσης (training set)
- Εξαγωγή του επιλεγμένου pipeline, μια λειτουργία που διατίθεται από το TPOT ώστε να μπορεί να χρησιμοποιηθεί και αργότερα σε ένα test set
- Πρόβλεψη της κλάσης των στοιχείων του test set με το εκπαιδευμένο μοντέλο (pipeline)
- Αξιολόγηση του μοντέλου με τις μετρικές που χρησιμοποιήθηκαν και για τον XGBoost

4.6.1 Αποτελέσματα των δοκιμών

Για κάθε πελάτη έχουμε ένα πίνακα για τις αντίστοιχες δοκιμές σε διάφορους συνδυασμούς των configuration blocks και του αριθμού των generations (Πίνακες 4.10 και 4.11). Σε κάθε εκτέλεση του TPOT το pipeline περιλάμβανε ως βασικό μοντέλο τα Extra Trees ή αλλιώς Extremely Randomized Trees, μια εξέλιξη του αλγορίθμου Random Forests, με διαφορετικούς συνδυασμούς των παραμέτρων του και μόνο σε μια περίπτωση υπήρχε preprocessing στάδιο με τη μέθοδο των Polynomial Features.

Πίνακας 4.10: Αριθμητικά αποτελέσματα από τους διάφορους συνδυασμούς παραμέτρων του TPOT (πελάτης 1)

	Configuration 1			Config 2	Config 3	M.O.
generations	5	10	20	10	10	
Accuracy (%)	93.61	93.05	92.29	92.86	92.67	92.9
AUC (%)	97.05	96.69	96.33	96.9	96.62	96.72
Recall (%)	94.96	94.3	94.08	94.96	94.74	94.61
Precision (%)	97.52	97.51	96.84	96.65	96.64	97.03
FPR (%)	14.47	14.47	18.42	19.74	19.74	17.37
FNR (%)	5.04	5.7	5.92	5.04	5.26	5.39

Πίνακας 4.11: Αριθμητικά αποτελέσματα από τους διάφορους συνδυασμούς παραμέτρων του TPOT (πελάτης 5)

	Configuration 1		Config 2	Config 3	M.O.
generations	5	10	20	15	
Accuracy (%)	95.96	95.96	95.05	94.37	95.34
AUC (%)	98.7	98.7	98.47	98.19	98.52
Recall (%)	95.99	95.99	95.83	94.71	95.63
Precision (%)	98.84	98.84	98.36	98.66	98.68
FPR (%)	6.73	6.73	9.6	7.69	7.69
FNR (%)	4	4	4.17	5.28	4.36

Αρχικά να σημειώσουμε πως σε κάθε εκτέλεση οι συνδυασμοί των παραμέτρων και των μεθόδων κατά τη δημιουργία των pipelines γίνεται με τυχαίο τρόπο οπότε είναι και λογικό τα αποτελέσματα (των προτεινόμενων αλγόριθμων) να είναι διαφορετικά. Παρόλα αυτά, ενώ σε κάθε εκτέλεση είχαμε διαφορετική παραμετροποίηση του Extra Trees, παρατηρούμε ότι τα score metrics είναι πολύ κοντά μεταξύ τους. Επίσης, όπως φαίνεται, παρόλο που γνωρίζουμε ότι όσο αυξάνονται οι επαναλήψεις (γενιές) δίνονται περισσότερες ευκαιρίες σε ένα βελτιστοποιημένο pipeline, η αύξηση των γενεών δεν αύξησε κάποια από τις μετρικές, αντιθέτως λαμβάνουμε τις καλύτερες τιμές στις περιπτώσεις 5 γενεών.

Επιπλέον, με την autoML και με χρόνο εκτέλεσης το περισσότερο 2,5 ώρες (στην περίπτωση των 20 γενεών), μπορούμε να λάβουμε μια ολοκληρωμένη λύση με πολύ ικανοποιητικά αποτελέσματα. Αν δώσουμε παραπάνω χρόνο της τάξης των 50-100 γενεών είναι πιθανό να δούμε ακόμα καλύτερα αποτελέσματα.

4.7 Αποτελέσματα ερευνών με χρήση νευρωνικών δικτύων και σύγκριση με χρησιμοποιούμενους αλγόριθμους της εργασίας

Η χρήση νευρωνικών δικτύων (neural networks) στον εντοπισμό της ρευματοκλοπής είναι αρκετά συχνή στη βιβλιογραφία που μελετήθηκε. Οι πιο χαρακτηριστικοί αλγόριθμοι είναι τα Recurrent Neural Networks (RNN) και τα Convolution Neural Networks (CNN). Παρακάτω παραθέτονται οι αλγόριθμοι και τα αποτελέσματα των ερευνών (Πίνακας 4.12) από τη βιβλιογραφία της παρούσας εργασίας.

Στην έρευνα των Chatterjee, Archana [62], χρησιμοποιήθηκαν τα RNN για τη δημιουργία προφίλ κατανάλωσης κάθε χρήστη έχοντας ως στόχο τη μείωση των False Positives. Στο μοντέλο που αναπτύχθηκε, μεγαλύτερη ή μικρότερη κατανάλωση ενέργειας από το συνηθισμένο μπορεί να είναι ένδειξη ανωμαλίας και κατηγοριοποιείται ανάλογα. Τα RNN μπορούν να χρησιμεύσουν στη διαχείριση σειριακών δεδομένων καθώς έχουν τη δυνατότητα αποθήκευσης προηγούμενων υπολογισμών (διαθέτουν δηλαδή μνήμη) ώστε να χρησιμοποιηθούν σε επόμενους υπολογισμούς, κωδικοποιώντας τις εξαρτήσεις μεταξύ των εισόδων του δικτύου. Τα

Long Short Term Memory (LSTM) είναι μια υποκατηγορία των RNN τα οποία έχουν τη δυνατότητα να μαθαίνουν εξαρτήσεις μεταξύ εισόδων μακροπρόθεσμα. Αυτό είναι ιδιαίτερα χρήσιμο στην περίπτωση των σειριακών δεδομένων καθώς δέχονται την παραδοχή ότι η ενέργεια που καταναλώνεται στο παρόν εξαρτάται από αυτή που έχει καταναλωθεί στο παρελθόν. Στη μεθοδολογία που ακολούθησαν δημιουργήθηκαν ζευγάρια εισόδου-εξόδου όπου η έξοδος υπολογίζεται από τις προηγούμενες έξι εισόδους. Τα δεδομένα είχαν τη μορφή πινάκων «μέρες x χρονικά διαστήματα», με μετρήσεις ανά μισάωρο κάθε μέρα. Επομένως η πρόβλεψη μιας εξόδου υπολογίζεται βάσει των καταγεγραμμένων τιμών των προηγούμενων 3 ωρών (6 μισάωρα). Τα αποτελέσματα του αλγορίθμου φαίνονται στον Πίνακα 4.12.

Οι Nabil και Ismaily στο [37] χρησιμοποίησαν επίσης Recurrent Neural Networks εκμεταλλευόμενοι και αυτοί τη χρονικά σειριακή μορφή των δεδομένων, και παρουσίασαν τον τρόπο με τον οποίο πραγματοποίησαν τη ρύθμιση των υπερ-παραμέτρων. Σκοπός τους ήταν να βελτιώσουν την απόδοση του αλγόριθμου όταν εφαρμόζεται σε σετ πραγματικών δεδομένων κατανάλωσης, του χρόνου εκπαίδευσής του αλλά και να διερευνήσουν τον πιο αποδοτικό τρόπο ρύθμισης των παραμέτρων των RNN. Το δίκτυο στη μέθοδο που ακολούθησαν αποτελούνταν από L επίπεδα κρυμμένων Gated Recurrent Units (GRU), όπου το καθένα έχει N νευρώνες. Η είσοδος ήταν ένα διάνυσμα ακολουθίας, το ίδιο και η έξοδος του δικτύου. Τα επαναλαμβανόμενα επίπεδα (recurrent layers) είναι πιο αποτελεσματικά στο να εκμεταλλεύονται μοτίβα από διαδοχικές πληροφορίες. Στο τελευταίο επίπεδο που διέθετε δύο νευρώνες έβγαινε το αποτέλεσμα για το αν ένα ημερήσιο δείγμα ήταν ειλικρινές (honest) ή κακόβουλο (malicious). Έπειτα κι από τη χρήση random tuning για τη ρύθμιση των παραμέτρων που αποδείχτηκε γρηγορότερο από το grid search, το μοντέλο αξιολογήθηκε για τρία διαφορετικά L (=2, 3, 4) και στον Πίνακα 4.12 καταγράφονται οι μέσοι όροι των μετρικών αξιολόγησης.

Μια άλλη χρήση NN έκαναν οι Wu, Wang και Hu στο [63], καθώς συνδύασαν ένα Adaptive Boosting (AdaBoost) αλγόριθμο με ένα General Regression NN, ώστε να εντοπίσουν την περίπτωση ρευματοκλοπής αλλά και το χρονικό διάστημα στο οποίο

αυτή συνέβη. Ο AdaBoost είναι επίσης ensemble αλγόριθμος και στην έρευνά τους οι Wu, Wang χρησιμοποίησαν τα SVM ως weak classifiers, για τα οποία αναφέρουν ότι έχουν καλύτερη απόδοση σε μη ισορροπημένα σετ δεδομένων, ρυθμίζοντας σε κάθε επανάληψη τα βάρη των στοιχείων μέχρι να φτάσουν τα αποτελέσματα ένα συγκεκριμένο κατώφλι. Αφού αποφασιστεί ποιοι καταναλωτές είναι παράνομοι ή όχι, χρησιμοποιούν τα GRNN για να προβλέψουν την κατανάλωση των «μη κανονικών» χρηστών, να τη συγκρίνουν με την πραγματική τιμή της κατανάλωσης και έπειτα υπολογίζουν το σχετικό σφάλμα. Έτσι, σχεδιάζοντας το σχετικό σφάλμα σε σχέση με τις δυο τιμές, εντοπίζουν συγκεκριμένα διαστήματα κατά τα οποία συμβαίνει η ρευματοκλοπή. Στο σημείο αυτό η πρόβλεψη έγινε για τρεις διαφορετικές περιόδους (Ιούλιο – Αύγουστο, Σεπτέμβριο – Οκτώβριο, Νοέμβριο – Δεκέμβριο) με βάση την κατανάλωση που είχε καταγραφεί μέχρι και τον Ιούνιο. Τα αποτελέσματα του πρώτου μέρους όπου έγινε χρήση του AdaBoost-SVM ήταν στο 85% περίπου η ακρίβεια για ποσοστό 10% του σετ δεδομένων να είναι καταναλωτές που έκλεβαν και στο 76.5% περίπου το F-measure για την ίδια κατηγορία. Όσον αφορά στα αποτελέσματα του NN, όπως αναφέρουν, το absolute regression error είναι αρκετά υψηλό, για το οποίο παραθέτουν και αντίστοιχους λόγους που συμβαίνει, παρόλα αυτά όμως με τον έλεγχο του σχετικού σφάλματος οι περίοδοι κλοπής φαίνεται να εντοπίζονται αρκετά αποτελεσματικά.

Στο [64] οι Chandel και Thakur, συνδύασαν ένα bidirectional LSTM RNN, δηλαδή ένα αμφίδρομο LSTM δίκτυο, με ένα Convolution NN (CNN), ώστε να εκμεταλλευτούν τη δυνατότητα αποθήκευσης και ανάκτησης πληροφορίας ενός LSTM και την περιοδικότητα των δεδομένων κατανάλωσης με το CNN. Συγκεκριμένα με το CNN έχοντας στη διάθεσή τους δεδομένα 1D μπορούσαν να ταυτοποιήσουν την περιοδικότητα ή μη των 2D καταγραφών κατανάλωσης, μεταμορφώνοντας τα δεδομένα από 1D σε 2D, μειώνοντας τον αριθμό των στοιχείων. Όπως και σε προηγούμενες αναφορές, η χρήση των RNN-LSTM προσέφεραν τη δυνατότητα εκμάθησης των διάφορων εξαρτήσεων μεταξύ των δεδομένων, ενώ σε αυτή την περίπτωση η δυνατότητα αποθήκευσης και διαμοιρασμού πληροφορίας μεταξύ των

επιπέδων του δικτύου ενισχύεται λόγω της αμφίδρομης αρχιτεκτονικής του. Το RNN-LSTM έπαιρνε ως είσοδο το αποτέλεσμα του CNN με το μειωμένο αριθμό των στοιχείων του πίνακα εισόδου και στο τελευταίο επίπεδο το αποτέλεσμα ήταν μία από τις 6 κλάσεις που αντιστοιχεί σε ένα από τα 6 σενάρια κλοπής, με βάση τις τιμές κατωφλίου που έχουν βρεθεί για στοιχεία όπως της τάσης και του ρεύματος φάσης, της ενεργού ισχύος ή του συντελεστή ισχύος. Τα αποτελέσματα του μοντέλου παρουσιάζονται στον Πίνακα 4.12.

Ένα άλλο είδος νευρωνικού δικτύου παρουσιάστηκε από τις Ghasemi και Gitizadeh για τον εντοπισμό ρευματοκλοπής στο [65]. Η μέθοδός τους περιλάμβανε ένα πιθανοτικό νευρωνικό δίκτυο (Probabilistic Neural Network – PNN) και ένα μαθηματικό μοντέλο που βασίζεται στη μέθοδο Levenberg – Marquardt. Η βασική παραδοχή του μοντέλου τους ήταν ότι οι παράνομοι χρήστες μπορούν να χωριστούν σε δυο ομάδες, α) αυτούς που κλέβουν όλη τη ποσότητα κατανάλωσης σε μια περίοδο της ημέρας και β) αυτούς που κλέβουν μόνο μια ποσότητα ενέργειας και όχι όλη. Χρησιμοποιώντας πρώτα έναν αλγόριθμο κωδικοποίησης (χαρτογράφηση των προφίλ κατανάλωσης σε έναν από 7 βαθμούς παρατυπίας) για feature extracting ετοίμασαν τα δεδομένα εισόδου για το PNN, μέσω του οποίου εντόπισαν τους καταναλωτές της πρώτης ομάδας. Ο λόγος χρήσης του συγκεκριμένου τύπου δικτύου όπως αναφέρουν είναι ότι προσφέρει γρήγορη εκπαίδευση, ένα γενικό και απλό σε χαρακτηριστικά μοντέλο και αρκετά ισχυρό. Σε επόμενο στάδιο έκαναν χρήση της μεθόδου Levenberg – Marquardt για τον εντοπισμό της δεύτερης ομάδας καταναλωτών αφού πρώτα αφαίρεσαν από το σύνολο αυτούς που είχαν εντοπιστεί στο προηγούμενο στάδιο.

Μια ακόμα περίπτωση χρήσης των CNN συναντήσαμε στην έρευνα των Li, Han, Yao [38] όπου συνδύασαν το νευρωνικό δίκτυο με τον αλγόριθμο Random Forests για να αντιμετωπίσουν το θέμα μη αποδοτικής επιθεώρησης και της «μη κανονικής» κατανάλωσης ενέργειας. Η χρήση του CNN έγινε για να εντοπιστούν/μαθευτούν στοιχεία μεταξύ διαφορετικών ωρών της ημέρας όπως και διαφορετικών ημερών μεταξύ τους, ανάμεσα σε πολλά και διαφορετικά δεδομένα έξυπνου μετρητή. Με τα

δεδομένα που έχουν εξαχθεί μετά το convolution και το downsampling στο CNN, ο αλγόριθμος RF εκπαιδεύεται ώστε να αποφασιστεί αν υπάρχει κλοπή ή όχι. Κατά τη χρήση του CNN στο convolution επίπεδο το μοντέλο μαθαίνει τα στοιχεία που αντιπροσωπεύουν τα δεδομένα εισόδου και μειώνουν την επίδραση του θορύβου, χρησιμοποιώντας κάποια φίλτρα για να υπολογίσουν τους διαφορετικούς χάρτες στοιχείων. Στο downsampling επίπεδο που υπάρχει μεταξύ δυο convolution επιπέδων μειώνεται ο αριθμός των παραμέτρων και των διαστάσεων των πινάκων, ενώ στο fully connected επίπεδο τα στοιχεία που έχουν εξαχθεί χρησιμοποιούνται για να μετατρέψει το feature map (χάρτη στοιχείων) σε ένα διάνυσμα. Τα αποτελέσματα των μετρικών καταγράφονται στον Πίνακα 4.12.

Μια ακόμα προσέγγιση για τον εντοπισμό ρευματοκλοπής με νευρωνικά δίκτυα παρουσιάστηκε στο [66] όπου συνδυάστηκαν δυο είδη νευρωνικών, τα CNN και LSTM. Η χρήση των CNN έγινε και εδώ για να αυτοματοποιηθεί η εξαγωγή στοιχείων (feature extracting), ενώ με τα LSTM εκμεταλλεύονται και πάλι τη σειριακή μορφή των δεδομένων. Το δίκτυο που σχεδίασαν περιλαμβάνει 7 hidden επίπεδα, από τα οποία τα 4 χρησιμοποιούνται για convolution και τα 3 πραγματοποιούν τη διαδικασία των LSTM. Κι εδώ μετά την εφαρμογή του convolution στην είσοδο μέσω κατάλληλων φίλτρων δημιουργούνται χάρτες των στοιχείων και στο pooling επίπεδο γίνεται downsampling του κάθε χάρτη για μείωση των διαστάσεων. Για να πετύχουν μείωση του overfitting και πιο γρήγορο training η διαδικασία γίνεται με max pooling. Στο fully connected layer πραγματοποιείται η κατηγοριοποίηση με βάση τα στοιχεία που έχουν εξαχθεί προηγουμένως. Το αποτέλεσμα έπειτα περνάει στο LSTM των τριών επιπέδων, που διαθέτουν τη δυνατότητα αποθήκευσης της πληροφορίας από τα αρχικά έως τα τελικά στάδια όπως έχει ήδη αναφερθεί. Το μοντέλο τους εκπαιδεύτηκε με εβδομαδιαία, 14-ημερών ή μηνιαία patterns, με τα τελευταία να έχουν σχετικά καλύτερη απόδοση. Τα αποτελέσματα κατηγοριοποίησης των παράνομων χρηστών σε ένα ισορροπημένο σετ δεδομένων φαίνεται στον Πίνακα 4.12.

Πίνακας 4.12: Αποτελέσματα των αλγορίθμων NN της βιβλιογραφίας

	Accuracy (%)	*Recall (%)*	*Precision (%)*	*AUC (%)*	*FPR (%)*	*F1(%)*
RNN-LSTM [62]	72.93	-	-	-	-	-
GRU-RNN [37]	95.47	92.87	-		5.73	
AdaBoost-SVM + GRNN [63]	-	-	-	-	-	-
CNN-RNN-biLSTM [64]	97.12	99.9	96.4	-	0.71	-
PNN-Levenberg–Marquardt [65]	96.11	-	-	-	-	-
CNN-RF [38]	-	97	97	98	-	97
CNN-LSTM [66]	88	91	87	-	-	89

Όπως παρατηρούμε από τον Πίνακα 4.12, οι αλγόριθμοι νευρωνικών δικτύων υπόσχονται πολύ καλά αριθμητικά αποτελέσματα. Στις περιπτώσεις ειδικά που αποτελούν και το βασικό αλγόριθμο κατηγοριοποίησης ([62], [37], [64], [66]) από τις μετρικές των αποτελεσμάτων η ακρίβεια κυμαίνεται κατά μέσο όρο κοντά στο 88% (με χαμηλότερη αυτή των απλών RNN-LSTM) και το Recall στο 94.6% ([37], [64], [66]), ενώ την ίδια ώρα στη δική μας προσπάθεια με τον XGBoost η ακρίβεια ήταν κατά μέσο όρο κοντά στο 96% και το recall στο 97%, ενώ με την autoML στο 94% και 95% αντίστοιχα. Παρόλα αυτά ιδιαίτερα καλή απόδοση είχαν και αλγόριθμοι που χρησιμοποίησαν τα NN ως ένα τρόπο φιλτραρίσματος των αρχικών δεδομένων (με χρήση των PNN) ή για feature engineering (με χρήση των CNN). Δυστυχώς δεν ήταν διαθέσιμα σε όλες τις περιπτώσεις τα αριθμητικά αποτελέσματα για να μπορέσει να γίνει μια πιο ολοκληρωμένη σύγκριση.

ΚΕΦΑΛΑΙΟ 5
ΣΥΝΟΨΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Το πρόβλημα της ρευματοκλοπής απασχολεί αρκετά την επιστημονική κοινότητα ως προς τους διάφορους τρόπους με τους οποίους μπορεί να εντοπιστεί. Οι αλγόριθμοι μηχανικής μάθησης έχουν πλέον κύριο ρόλο στη διαχείριση της ενέργειας με την εισαγωγή των Advanced Metering Infrastructures και των έξυπνων μετρητών στη ζωή μας.

Στην εργασία αυτή έγινε η χρήση δυο μεθόδων για τον εντοπισμό των ημερών που ένας καταναλωτής παρουσιάζει παράνομη συμπεριφορά κατανάλωσης, του XGBoost, ενός αλγόριθμου που βασίζεται στη δημιουργία δεντρικών δομών για την πρόβλεψη μιας κλάσης, και της autoML μέσω του TPOT, έναν αυτοματοποιημένο τρόπο για την εύρεση του βέλτιστου pipeline μοντέλου που βασίζεται στο γενετικό προγραμματισμό και μπορεί να περιλαμβάνει επίσης στάδια προεπεξεργασίας των δεδομένων αν κριθεί αναγκαίο.

Η προσέγγισή μας βασίστηκε στην έρευνα των Punmiya και Choe, οι οποίοι έκαναν σύγκριση 3 διαφορετικών gradient boosting αλγορίθμων, και ακολουθήσαμε την πρώτη φάση του πειράματός τους, που περιλάμβανε την εφαρμογή του αλγόριθμου στα δεδομένα από το Irish Social Science Data Archive –ISSDA. Τα δεδομένα αυτά χρειάστηκε επίσης να χρησιμοποιηθούν ως βάση για να παραχθούν δεδομένα υπό μορφή κλοπής, κάτι το οποίο οδήγησε σε ανισορροπία των δύο κλάσεων στο training set, οπότε και έγινε χρήση του αλγόριθμου SMOTE συνθετικών νέων δεδομένων και την εξισορρόπηση των κλάσεων.

Η διαφοροποίησή μας ήταν ότι πραγματοποιήσαμε hyper parameter tuning για τα σετ κάθε καταναλωτή, κάτι που δεν αναφέρεται στη δουλειά τους και δοκιμάσαμε τον αλγόριθμο σε ακόμα ένα σετ δεδομένων από το UK Power Network. Δεν έγινε δοκιμή της διαδικασίας feature engineering, κάτι που μπορεί να φανεί χρήσιμο αν θέλουμε να ξεχωρίσουμε ορισμένα χαρακτηριστικά/μεταβλητές του σετ δεδομένων

με βάση τη σημασία τους (feature importance), μια λειτουργία που διατίθεται πολύ εύκολα από τη βιβλιοθήκη του XGBoost.

Τα αποτελέσματα που προέκυψαν είναι πολύ ικανοποιητικά ακόμα και πριν το hyper parameter tuning και μας έκανε να αναρωτηθούμε αν θα υπήρχε τρόπος ο αλγόριθμος να μην έχει τόσο καλή απόδοση. Γι' αυτό το λόγο και γίνανε δοκιμές δυο διαφορετικών τύπων. Η πρώτη αφορούσε την προσθήκη μόνο ορισμένων περιπτώσεων κλοπής στο σετ εκπαίδευσης και όλων των τύπων στο test set (κάτι που εφαρμόστηκε κι από τους Jokar και Arianpoo), και η δεύτερη δοκιμή αφορούσε την αντιστροφή του λόγου των πραγματικών δειγμάτων ως προς τα δείγματα κλοπής με χρήση του SMOTE ή χωρίς. Τα αποτελέσματα δεν διέφεραν ιδιαίτερα όπως μπορεί να περιμέναμε, οπότε εγείρεται το ερώτημα αν ο αλγόριθμος κάνει overfitting, δηλαδή μαθαίνει πολύ καλά τα δεδομένα εκπαίδευσης, και αν η μορφή των δεδομένων, όπως έχει διαμορφωθεί, επηρεάζει την απόδοση.

Επομένως η χρήση του feature engineering και feature selection, ή άλλων μεθόδων προεπεξεργασίας των δεδομένων θα μπορούσε να είναι η επόμενη φάση των δοκιμών. Επιπλέον τα δεδομένα θα μπορούσαν να διαμορφωθούν και να μελετηθούν οργανώνοντάς τα σε κατανάλωση ανά βδομάδα, μήνα ή εποχή, όπου οι καταναλωτές παρουσιάζουν κάποια patterns ως προς την κατανάλωση, όπως ότι τους καλοκαιρινούς μήνες έχουμε αυξημένη κατανάλωση ρεύματος ενώ όχι τόσο στους χειμερινούς, καθώς επίσης και ότι παρουσιάζουν ένα μοτίβο της χρήσης βάσει της ημέρας της εβδομάδας.

Μια επίσης πρόταση είναι η μελέτη των δεδομένων των χρηστών ομαδικά με βάση την ομοιότητά τους ή αλλιώς την ομοιότητα της διανομής των δεδομένων με ανάλογους αλγόριθμους.

Μια ακόμα παρατήρηση ή υπόθεση είναι ότι υπάρχει ίσως μια πιθανότητα, παρά την παραδοχή ότι τα δεδομένα είναι πραγματικά και «ειλικρινή», να υπάρχουν ημερήσια patterns τα οποία «ταιριάζουν» με κάποια μορφή από τα τεχνητά δεδομένα κλοπής, γι' αυτό ενώ ανήκουν στην κλάση 0 να κατηγοριοποιούνται στην κλάση 1 (κλοπή) και

έχουμε κάποια FPs παραπάνω (αν και σε κάθε περίπτωση πάντα τα FPs ήταν σχετικά λίγα).

Τέλος, στα σενάρια κλοπής περιλαμβάνονται μόνο αυτά που επιδεικνύουν μειωμένη κατανάλωση και δεν μελετώνται οι περιπτώσεις που ένας καταναλωτής είναι θύμα κλοπής, οπότε και έχουμε αρκετά μεγαλύτερη κατανάλωση ισχύος από ότι θα όριζε το μοτίβο κατανάλωσης του χρήστη.

Όσον αφορά στην autoML, από τα αριθμητικά αποτελέσματα μπορούμε να συμπεράνουμε πως αποτελεί μια μέθοδο με πολύ ικανοποιητικά σκορ και σε χρόνο πολύ λιγότερο από αυτό που απαιτεί ένα exhaustive grid search και το hyper parameter tuning, καθώς επίσης χρειάζεται και ελάχιστη παρέμβαση από το χρήστη.

Επιπλέον, αν και συμπεριλάβαμε τον XGBoost στο μπλοκ δοκιμών του TPOT, σε καμία από τις περιπτώσεις δεν προέκυψε σε κάποιο pipeline και ακόμη κι όταν μελετήσαμε τα 10-15 pipelines με τα καλύτερα σκορ δεν βρισκόταν ανάμεσά τους, αλλά υπερίσχυσε ο Extra Trees. Επομένως θα μπορούσε σε επόμενη φάση να μελετηθεί και η απόδοση αυτού του αλγορίθμου σε περιβάλλον εκτός της autoML.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] ”We calculated emissions due to electricity loss on the power grid globally, it’s a lot”, (άρθρο), S. M. Jordaan, K. Surana, 2019, https://theconversation.com/we-calculated-emissions-due-to-electricity-loss-on-the-power-grid-globally-its-a-lot-128296

[2] ”2nd CEER Report on Power Losses”, Council of European Energy Regulators, Ref: C19-EQS-101-03, 2020

[3] Data World Bank, https://data.worldbank.org/indicator/EG.ELC.LOSS.ZS

[4] ”Δίκτυα διανομής ηλεκτρικής ενέργειας –Εξοικονόμηση ενέργειας μέσω έξυπνων δικτύων (smartgrids)”, Α. Τσικόγιας, 2011

[5] ”The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey”, P. Glauner, J. Meira, P. Valtchev, R. State, Fr. Bettinger, 2017

[6] ”Electricity Theft and Non-Technical Losses: Global Markets, Solutions, and Vendors”, Northeast Group Llc, 2017

[7] ”Revenue protection for smart utilities: dig value from Big Data”, Atos Codex, 2017

[8] “Ρευματοκλοπή. Ορισμός, διαδικασία εντοπισμού και συνέπειες”, (Εγχειρίδιο ρευματοκλοπών σε εφαρμογή της παραγράφου 23 του άρθρου 95 του Κώδικα Διαχείρισης Δικτύου Διαχείρισης Διανομής Ηλεκτρικής Ενέργειας), 2017, http://www.odigostoupoliti.eu/reymatoklopi-orismos-diadikasia-entopismou-synepeies/

[9] ”Μειώθηκαν οι ρευματοκλοπές για πρώτη φορά μετά από 5 χρόνια”, (άρθρο), Κων/νος Φιλίππου, 2019, https://energypress.gr/news/meiothikan-oi-reymatoklopes-gia-proti-fora-meta-apo-5-hronia-sto-32-oi-mi-tehnikes-apoleies-apo

[10] ”ΡΑΕ: Με καινοτομίες το νέο ρυθμιστικό πλαίσιο για το δίκτυο διανομής”, (άρθρο), Κων/νος Φιλίππου, 2020, https://energypress.gr/news/rae-kinitro-sto-deddie-gia-safari-ton-reymatoklopon-me-kainotomies-neo-rythmistiko-plaisio-gia

[11] ”Non-technical loss analysis and prevention using smart meters”, T. Ahmad, 2017

[12] “The Most Frequent Energy Theft Techniques and Hazards in Present Power Energy Consumption”, R. Czechowski, A. M. Kosek, 2016

[13] ”An IoT Based Tamper Prevention System for Electricity Meter”, R.E. Ogu, Prof. G. A. Chukwudebe, I. A. Ezenugu, 2016

[14] ”Power Theft Prevention Techniques”, EL-PRO-CUS, https://www.elprocus.com/power-theft-prevention-techniques/

[15] ”AMI Smart Meter Reading Solution: Electricity theft prevention”, Huawei, https://support.huawei.com/enterprise/en/doc/EDOC1100069580/46062d12/electricity-theft-prevention

[16] ”Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft”, Sh. Depuru, L. Wang, V. Devabhaktuni, 2011

[17] ”Electricity Theft Prevention in Distribution System with Distribution Generation”, Sh. Mohammad, A. Dar, 2018

[18] ”Οι έξυπνοι μετρητές ρεύματος εξουδετερώνουν τις ρευματοκλοπές”, (άρθρο), Κ. Βουτσαδάκης, 2018, https://ecopress.gr/i-exypni-metrites-revmatos-exoudet/

[19] ”ΔΕΔΔΗΕ: Ετοιμάζονται οι προδιαγραφές για τους 7,5 εκ. «έξυπνους» μετρητές”, (άρθρο), Κων/νος Φιλίππου, 2020, https://energypress.gr/news/deddie-etoimazontai-oi-prodiagrafes-gia-toys-75-ek-exypnoys-metrites-ta-senaria-gia-ti

[20] ”Τί είναι οι έξυπνοι μετρητές ηλεκτρισμού;”, (άρθρο), Α. Πουλικκάς, 2020, link

[21] ”How Smart Meters work”, last update 2020, https://www.smartme.co.uk/how-they-work.html#gsc.tab=0

[22] ”AMI Smart Meter Reading Solution: Basic functions of smart meters”, Huawei, https://support.huawei.com/enterprise/fr/doc/EDOC1100069580/4213e84b/basic-functions-of-smart-meters

[23] ”Έξυπνα Ενεργειακά Δίκτυα: Διαχείριση και Εφαρμογές”, (διπλωματική εργασία), Παντίσκας Ν., 2016

[24] ”Smart meters explained”, (άρθρο), Kasey Cassells, 2020, https://www.uswitch.com/gas-electricity/guides/smart-meters-explained/

[25] ”Έξυπνοι μετρητές ηλεκτρικής ενέργειας”, Energy Lab, http://www.energylab.gr/products/energy-monitoring/

[26] ”Smart meter”, Wikipedia, https://en.wikipedia.org/wiki/Smart_meter#cite_note-7

[27] ”Energy-Theft Detection Issues for Advanced Metering Infrastructure in Smart Grid”, R. Jiang, R. Lu, Y. Wang, 2014

[28] ”Review of various modeling techniques for the detection of electricity theft in smart grid environment”, T. Ahmad, H. Chen, 2018

[29] ”Review of non-technical loss detection methods”, G. M. Messinis, N. D. Hatziargyriou, 2018

[30] ”Smart Meter Data Analysis for Power Theft Detection”, Nikovski D., Wang Z., 2013

[31] ”A novel approach to detection and prevention of electricity pilferage over power distribution network”, M. Aryanezhad, 2019

[32] ”Electricity Theft Detection in AMI Using Customers’ Consumption Patterns”, P. Jokar, N. Arianpoo, 2016

[33] ”Identification of suspicious electricity customers”, J. V. Spirića, S. S. Stankovićb, M. B. Dočićb, 2018

[34] “Machine learning Techniques for Energy Theft Detection in AMI”, A. Maamar, K. Benahmed, 2018

[35] ”Detection of energy theft and defective smart meters in smart grids using linear regression”, S. Yip, K. Wong, W-P. Hewa, 2017

[36] ”Machine Learning Algorithm for Efficient Power Theft Detection using Smart Meter Data”, J. Jeyaranjani, D. Devaraj, 2018

[37] ”Deep Recurrent Electricity Theft Detection in AMI Networks with Random Tuning of Hyper-parameters”, M. Nabil, M. Ismaily, 2018

[38] ”Electricity Theft Detection in Power Grids with Deep Learning and Random Forests”, S. Li, Y. Han, Xu Yao, 2019

[39] ”A practical feature-engineering framework for electricity theft detection in smart grids”, R. Razavia, A. Gharipour, 2019

[40] ”An Artificial Intelligent Algorithm for Electricity Theft Detection in AMI”, R. Sowndarya, Dr. P. Latha, 2017

[41] ”Detection of Electricity Theft in Customer Consumption using Outlier Detection Algorithms”, J. Yeckle, B. Tang, 2018

[42] ”Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing”, R. Punmiya and S. Choe, 2019

[43] XGBoost Documentation, https://xgboost.readthedocs.io/en/latest/

[44] ”Using XGBoost in Python”, (tutorial), Manish Pathak, Datacamp, 2019, https://www.datacamp.com/community/tutorials/xgboost-in-python

[45] ”Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python”, (tutorial), Aarshay Jain, Analytics Vidhya, 2016, https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

[46] ”Survey on Automated Machine Learning”, M. Zoller, M. F. Huber, 2019

[47] https://www.automl.org/

[48] Auto-sklearn Documentation, https://automl.github.io/auto-sklearn/master/index.html

[49] MLBox Documentation, https://mlbox.readthedocs.io/en/latest/introduction.html

[50] H2O AutoML Documentaion, http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

[51] TPOT Documentation, https://epistasislab.github.io/tpot/

[52] Auto-Keras Documentation, https://autokeras.com/

[53] Irish Social Science Data Archive – ISSDA, https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[54] Low Carbon London - UK Power Networks, https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[55] SMOTE Documentation, https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

[56] ”Imbalanced Data : How to handle Imbalanced Classification Problems”, (tutorial), Analytics Vidhya, 2017, https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

[57] ”Classification Model Evaluation Metrics in Scikit-Learn”, (tutorial), Data Courses, https://www.datacourses.com/classification-model-evaluation-metrics-in-scikit-learn-924/

[58] ”Evaluation of binary classifiers”, Wikipedia, https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

[59] ”Complete Guide to Parameter Tuning in XGBoost with codes in Python”, (tutorial), Aarshay Jain, Analytics Vidhya, 2016, https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

[60] ”Hyperparameter tuning in XGBoost”, (tutorial), Cambridge Spark, 2017, https://blog.cambridgespark.com/hyperparameter-tuning-in-xgboost-4ff9100a3b2f

[61] ”Use decision trees and XGBoost to classify tumor data”, (notebook), IBM data platform, https://dataplatform.cloud.ibm.com/exchange/public/entry/preview?url=https://raw.githubusercontent.com/IBMDataScience/sample-notebooks/master/Cloud/HTML/Use%20decision%20trees%20and%20XGBoost%20to%20classify%20tumors.html#bullet-17

[62] “Detection of Non-Technical Losses using Advanced Metering Infrastructure and Deep Recurrent Neural Networks”, S. Chatterjee, V. Archana, K. Suresh, 2017

[63] “AdaBoost-SVM for Electrical Theft Detection and GRNN for Stealing Time Periods Identification”, R. Wu, L. Wang, T. Hu , 2018

[64] “Smart Meter Data Analysis for Electricity Theft Detection using Neural Networks”, P. Chandel, T. Thakur, 2019

[65] “Detection of illegal consumers using pattern classification approach combined with Levenberg-Marquardt method in smart grid”, A. Ghasemi, M. Gitizadeh, 2018

[66] “Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach”, N. Hasan, R. N. Toma, A. Nahid, M. M. Islam, J. Kim, 2019