Στρατιωτική Σχολή Ευέλπιδων

Τμήμα Στρατιωτικών Επιστημών

Πολυτεχνείο Κρήτης

Σχολή Μηχανικών Παραγωγής & Διοίκησης

Διπλωματική εργασία με θέμα:

Πρόβλεψη & Εκτίμηση Τιμών Ακινήτων με Τεχνικές Μηχανικής & Βαθιάς Μάθησης

Prediction & Valuation of Real Estate Prices with Machine & Deep Learning Techniques

Σαραπάνης Ιωάννης

ΧΑΝΙΆ 2023

Η Μεταπτυχιακή Διατριβή του Σαραπάνη Ιωάννη εγκρίνεται:

Καραδήμας Νικόλαος (Επιβλέπων) ,.....NIKOLAOS KARADIMAS....
19/06/2023 23:04

Αναπλ. Καθηγητής

Ματσατσίνης Νικόλαος ,..... Nikolaos Matsatsinis
Digitally signed by Nikolaos Matsatsinis
Date: 2023.06.21 16:56:18 +03'00'

Καθηγητής

Δάρας Νικόλαος ,.....NIKOLAOS DARAS 20/06/2023 13:59.....

Καθηγητής

Ευχαριστίες

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή Καραδήμα Νικόλαο για την βοήθεια του κατά τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας και τον κύριο Δάρα Νικόλαο για την υπευθυνότητα του καθ' όλη την διάρκεια σπουδών.

Ιδιαίτερα θα ήθελα να ευχαριστήσω την οικογένειά μου και τον φίλο Νταβαρίνο Γρηγόριο για την πολύτιμη στήριξη τους καθ' όλη την διάρκεια αυτού του εγχειρήματος.

Μάϊος 2023

Περίληψη

Τα τελευταία χρόνια, όλο και περισσότερο αναπτύσσεται στην ερευνητική κοινότητα ένα αυξημένο ενδιαφέρον για την προσπάθεια εκτίμησης των τιμών ακινήτων, με διάφορες αναπτυγμένες τεχνικές Machine & Deep Learning.

Σε αυτή την διπλωματική εργασία με την αξιοποίηση της επιστήμης της πληροφορικής και συγκεκριμένα της μεθόδου της τεχνητής νοημοσύνης και μηχανικής μάθησης, παρουσιάζεται η διαδικασία σχεδίασης, ανάλυσης και επεξεργασίας μαθηματικών μοντέλων με σκοπό την εκτίμηση αξιών ακινήτων και τελικά στην αξιολόγηση και επιλογή του πιο αξιόπιστου μοντέλου. Πιο συγκεκριμένα, υλοποιούνται διαφορετικά μοντέλα μηχανικής μάθησης, τα οποία εκπαιδεύονται με βάση τα δεδομένα. Η πλειοψηφία των μοντέλων αυτών στηρίζεται σε τεχνικές και αλγορίθμους παλινδρόμησης, όπως είναι η πολλαπλή γραμμική παλινδρόμηση, η οποία αποτελεί τη βάση για την ανάπτυξη πιο σύνθετων και αποδοτικών τεχνικών παλινδρόμησης, όπως είναι η παλινδρόμηση Ridge, η Lasso και η παλινδρόμηση με χρήση της τεχνικής Gradient Boosting. Επίσης, υλοποιήθηκαν μοντέλα που βασίζονται στα Δέντρα Απόφασης, όπως είναι τα Τυχαία Δάση και άλλα μοντέλα όπως τα Νευρωνικά Δίκτυα. Ακολούθως, παρουσιάζονται και αξιολογούνται τα αποτελέσματα της εφαρμογής τους στα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου.

Για αυτό χρησιμοποιήθηκαν δεδομένα από διάσημες διαδικτυακές πλατφόρμες στην Ελλάδα όπως για παράδειγμα xe.gr και plot.gr κ.α. με την βοήθεια των οποίων δίνεται μια σχετική εικόνα της αγοράς ακινήτων. Αναλύονται, στη συνέχεια, τα αποτελέσματα που λάβαμε από τις πλατφόρμες του διαδικτύου και τις μεθόδους που εφαρμόσαμε και αξιολογούνται η ακρίβεια και η καταλληλόλητά τους για το παρόν εγχείρημα.

Τέλος, παρουσιάζονται κάποιες συγκρίσεις των αποτελεσμάτων μας με αποτελέσματα αντίστοιχων ερευνών, εντοπίζονται τα σημεία που επιδέχονται βελτίωση στην μεθοδολογία που ακολουθήθηκε και προτείνονται με βάση αυτά κάποιες μελλοντικές προοπτικές έρευνας για τις επόμενες μελέτες.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Βαθία Μάθηση, Ανάλυση Δεδομένων, Ακίνητα.

Abstract

In recent years, an increased interest in trying to estimate real estate prices, with various developed Machine & Deep Learning techniques, has increasingly developed in the research community.

In this diploma thesis using computer science and specifically the method of artificial intelligence and machine learning, the process of designing, analyzing and processing mathematical models with the estimation of the value of real estate and finally in the evaluation and selection of the most reliable model. More specifically, different machine learning models are implemented, which are trained based on the data. The majority of these models rely on regression techniques and algorithms, such as multiple linear regression, which is the basis for the development of more complex and efficient regression techniques, such as Ridge regression, Lasso, and Gradient regression Boosting. Also, models based on decision trees, such as Random Forests and other models such as Neural Networks, were implemented. Subsequently, the results of their application to the training and control data are presented and evaluated.

For this were used data from famous online platforms in Greece such as xe.gr and spitogatos.gr, with the help of which, a relevant picture of the real estate market is given. The results we obtained from the internet platforms and the methods we applied are then analyzed and their accuracy and appropriateness for the present project are evaluated.

Finally, some comparisons of our results with results of corresponding researches are presented, the points that can be improved in the methodology are identified which was followed and some future research perspectives for the next studies are proposed based on them.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Data Analysis, Real Estate

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1 – Εισαγωγή

1.1 Ακίνητα

Από τα πρώτα χρόνια της ύπαρξης του ανθρώπου στην γη, μία από τις βασικότερες ανάγκες του, αποτέλεσε η ανάγκη για στέγαση. Η κατοικία, ή οικία του ανθρώπου, εμφανίζεται στην πιο απλή μορφή της από την παλαιολιθική εποχή, με τον άνθρωπο να καταφεύγει σε σπήλαια κατά τις μετακινήσεις του, όπου και διέμενε, προκειμένου να προστατευθεί από τις καιρικές συνθήκες ή άλλες απειλές. Οι πρώτες μόνιμες ανθρώπινες κατοικίες παρατηρούνται την νεολιθική εποχή, οπότε και έχουμε την ανάπτυξη των πρώτων μόνιμων οικισμών με πέτρινα ή ξύλινα σπίτια. Στην αρχαία Αθήνα, έχουμε απλές και λιτές κατοικίες από πέτρα, πηλό, ξύλο και κεραμίδια. Πλησιάζοντας το σήμερα, εδραιώνεται η έννοια της ιδιοκτησίας και σε συνδυασμό με την μεγάλη τεχνολογική πρόοδο που βιώνει η ανθρωπότητα, αλλάζει ριζικά η έννοια της κατοικίας.

Η αξία της γης και η εκτίμηση αυτής, αλλά και των ακίνητων κατοικιών, αποτελεί ζήτημα, το οποίο έχει απασχολήσει τον άνθρωπο από τον καιρό της εμφάνισης των πρώτων οικισμών. Σήμερα, η ανάγκη αυτή προκύπτει από την δυνατότητα διάθεσης του ακινήτου και συνεπώς κρίνεται καίριο να εκφραστεί σε χρήματα η αξία του. Οι παράγοντες από τους οποίους αυτή εξαρτάται, είναι τόσο πολυάριθμοι, όσο και περίπλοκοι, εφόσον ενδέχεται να είναι πολεοδομικής, πολιτικής, κοινωνικής, οικονομικής, αλλά και προσωπικής φύσεως. Οι συχνές αλλαγές των παραπάνω παραγόντων, αλλά και η πολυάριθμες έννοιες της αξίας, πολυπλέκουν ακόμα περισσότερο τα δεδομένα μας και έχουν ως συνέπεια την γένεση της ανάγκης του σχηματισμού και της θεμελίωσης ενός «συστήματος» κανόνων εκτίμησης της αξίας των κατοικιών και της γης. Την λύση στο παραπάνω ζήτημα έρχεται να δώσει το Σύνταγμα κάθε χώρας, με την συμβολή του οποίου κατοχυρώνεται η αξία του ακινήτου και προσδιορίζεται επακριβώς μέσω της επιστήμης και της συστηματικής σκέψης. Η γνώση της ακριβούς αυτής αξίας συμβάλλει σε πολλούς σκοπούς και στόχους οι οποίοι είτε προβλέπονται από την νομοθεσία, όπως φορολογικής φύσεως, είτε αποτελούν απόρροια των αναγκών της καθημερινής ζωής.

Στην χώρα μας, η αξία των ακινήτων δεν προσδιορίζεται ενιαία και συνολικά από κάποια νομοθετική ρύθμιση, γεγονός που οδηγεί στην εκτίμηση αυτής με τρόπο διαφορετικό ανά την περίπτωση. Αποτέλεσμα αυτού αποτελεί, η πιθανότητα για ένα ακίνητο να υπάρχει ένα πλήθος υπερτιμημένων ή υποτιμημένων τιμών, οι οποίες ενδέχεται να πλησιάζουν την πραγματική αξία, οι αποκλίσεις ωστόσο μπορεί να είναι πολύ μεγάλες. Η αγορά των ακινήτων και η οικονομία μίας χώρας συνδέονται άρρηκτα μεταξύ τους, αλληλοεπιδρούν και αλληλοεπηρεάζονται. Σε προσωπικό επίπεδο, η αγορά ενός ακινήτου αποτελεί, σημαντική απόφαση στην ζωή ενός πολίτη, καθώς η σωστή επιλογή μπορεί να αποδειχθεί σημαντική επένδυση για το μέλλον του. Γεννιέται, λοιπόν, η αναγκαιότητα δημιουργίας και εδραίωσης ενός συστηματικού και καλά ορισμένου τρόπου προσδιορισμού της αξίας ενός ακινήτου, προκειμένου να καλυφθούν οι σημερινές και μελλοντικές, ολοένα πολλαπλασιαζόμενες ανάγκες της κοινωνίας μας. Το σύστημα αυτό κρίνεται απαραίτητο να είναι δυναμικό, να μεταβάλλεται, δηλαδή, η τιμή των ακινήτων ανάλογα με τις μεταβολές των παραγόντων που την επηρεάζουν.

Δεδομένου ότι κάθε ακίνητο είναι μοναδικό, θεωρητικά οι παράγοντες που επηρεάζουν και διαμορφώνουν την αξία του είναι άπειροι. Οι παράγοντες αυτοί μπορούν να ταξινομηθούν με πολλούς τρόπους και να διαβαθμιστούν σε πολλά επίπεδα. Οι κυριότεροι παράγοντες είναι πολιτικοί, κοινωνικοί, οικονομικοί και άλλοι εξίσου σημαντικοί που αφορούν στην αξία της γης, τον ρυθμό ανάπτυξης μιας περιοχής, την ποιότητα ζωής και την πολεοδομική οργάνωση. Στην προκειμένη περίπτωση για την επίτευξη του σκοπού της εργασίας λαμβάνουμε υπόψιν εκείνους τους παράγοντες που αφορούν την κατοικία αυτή καθαυτή και αφορούν τα χαρακτηριστικά της.

Από αυτά άλλα είναι άμεσα μετρήσιμα όπως για παράδειγμα η επιφάνεια σε τετραγωνικά μέτρα ή το έτος κατασκευής και άλλα περιγραφικά, όπως η περιοχή στην οποία βρίσκεται το ακίνητο (αστική, αγροτική, δασική) ή ο τύπος του ακινήτου (μονοκατοικία, διαμέρισμά, μεζονέτα, βιομηχανική εγκατάσταση, γραφεία κλπ).

Τα ακίνητα έχουν αξία επειδή είναι εκεί που είναι και ως είναι, σε συνδυασμό με τις δυνατότητες μετασχηματισμού τους. Η γνώση της συνταγματικά

κατοχυρωμένης αυτής αξίας μπορεί να προσδιοριστεί κατά τρόπο ακριβή μέσα από την επιστήμη και τη συστηματική σκέψη και όχι πρόχειρα ή εμπειρικά. Η γνώση της αληθούς τιμής ή της αληθούς εικόνας των τιμών εξυπηρετεί πλήθος στόχων και σκοπών που η νομοθεσία και ανάγκες της αγοράς επιβάλλουν σύμφωνα με τον Π.Ζεντέλη (2015).

Σημαντικό ρόλο στην σημερινή αγορά των ακινήτων στην Ελλάδα αλλά και το εξωτερικό, κατέχουν οι μεσίτες, οι οποίοι προσφέρουν κτηματομεσιτικές υπηρεσίες στους πολίτες που ενδιαφέρονται για την αγορά, πώληση ή ενοικίαση κάποιου ακινήτου ή γης. Έχουν την δυνατότητα να εκτιμήσουν την εμπορικότητα, εκμεταλλευσιμότητα, οικοδομησιμότητα, καταλληλόλητα και την αξία των ακινήτων, αλλά και να ρυθμίσουν υπό όρους την αγοραπωλησία τους. Η ραγδαία εξέλιξη της τεχνολογίας και η ανάπτυξη της επιστήμης της Πληροφορικής και της Μηχανικής Μάθησης έχουν πλέον ανοίξει νέους δρόμους σε πολλές πτυχές της ζωής μας, αλλάζοντας σημαντικά τον τρόπο ζωής του σύγχρονου ανθρώπου.

Η Τεχνητή Νοημοσύνη και η Μηχανική Μάθηση έχουν κάνει τα πρώτα τους βήματα, εισχωρώντας εκτός των άλλων και στον τομέα της αγοράς των ακινήτων. Μεγάλες εταιρείες στην περιοχή του Real Estate, όπως είναι η Airbnb και η Zillow, έχουν υλοποιήσει εργαλεία που βασίζονται στην Τεχνητή Νοημοσύνη, βασικές λειτουργίες των οποίων είναι το ταίριασμα ενός αγοραστή ή ενοικιαστή ακινήτου με έναν ιδιοκτήτη, αλλά και η εκτίμηση της τιμής ακινήτων. Για παράδειγμα το πρόγραμμα Zestimate της εταιρείας Zillow αποτελεί ακριβώς αυτό: μία πλατφόρμα μέσω της οποίας κάθε κάτοικος της Αμερικής έχει την δυνατότητα να λάβει μία εκτίμηση της αξίας του ακινήτου του. Οι εκτιμήσεις αυτές βελτιώνονται καθημερινά, εφόσον ο όγκος των δεδομένων τους ολοένα και αυξάνεται.

Στην χώρα μας, δυστυχώς, δεν υπάρχει μέχρι στιγμής κάποια υπηρεσία που να προσφέρει το προαναφερθέν. Παρά την αυξημένη μελέτη γύρω από την περιοχή της αγοράς ακινήτων και της μηχανικής μάθησης, μικρή εφαρμογή των τεχνολογιών αυτών σε προβλήματα κοστολόγησης κατοικιών, φαίνεται να υπάρχει. Το τελευταίο, σε συνδυασμό και με το πόσο σημαντική κρίνεται η ακριβής αποτίμηση της αξίας των ακινήτων, γεννάει την επιθυμία για περαιτέρω

έρευνα και πρακτική δοκιμή των μεθόδων μηχανικής μάθησης, με σκοπό την αυτοματοποίηση της τιμολόγησης των κατοικιών. Η παρούσα διπλωματική εργασία επικεντρώνεται σε αυτήν ακριβώς την προσπάθεια.

1.2 Περιγραφή του προβλήματος

Η αγορά των ακινήτων αποτελεί έναν από τους κυριότερους παράγοντες, εάν όχι τον κυριότερο, παράγοντα κάθε εθνικής οικονομίας, καθώς συνεισφέρει σε μεγάλο ποσοστό στη σύνθεση του ΑΕΠ κάθε χώρας. Συνεπώς, η πληροφόρηση που έρχεται μέσα από τα δεδομένα που παράγονται σε συνάρτηση με την ακριβή εκτίμηση της τιμής κάθε ακίνητου αποτελούν καταλυτικούς παράγοντες στην λήψη αποφάσεων. Στην αγορά ακινήτων συμμετέχουν πολλά μέρη, τα κυριότερα εξ' αυτών είναι οι μεσίτες, οι επενδυτές, η κυβέρνηση, κ.α. Σκοπός των βασικών αυτών συμμετεχόντων είναι η δημιουργία κέρδους καθώς και η εύρεση νέων ευκαιριών. Ας δούμε ένα παράδειγμα εύρεσης τέτοιων ευκαιριών. Λαμβάνοντας υπόψη τα δεδομένα που υπάρχουν για μια κατοικία μπορούμε να κάνουμε εκτίμηση για την πραγματική της αξία. Σε περίπτωση που η αξία πώλησης αποκλίνει από την πραγματική αξία και το ακίνητο δίνεται σε χαμηλότερη τιμή από την εκτιμώμενη, τότε ο επενδυτής μπορεί να αναγνωρίσει την επενδυτική ευκαιρία και να αποκομίσει κέρδος σε ενδεχόμενη μεταπώληση του. Στην πραγματικότητα υπάρχουν διάφοροι λόγοι, οι οποίοι μπορεί να οδηγήσουν τον πωλητή στο να θέσει την τιμή του ακινήτου σε χαμηλότερη τιμή όπως είναι η ανάγκη ρευστότητας, προσωπικοί λόγοι κ.α.. Επενδυτικές ευκαιρίες υπάρχουν παντού γύρω μας ακόμη και σε περιόδους οικονομικής ύφεσης. Σκοπός του επενδυτή ή μεσίτη είναι να μπορεί να αναγνωρίσει τις ευκαιρίες αυτές. Η μέθοδος της Μηχανικής Μάθησης δίνει την ευκαιρία αυτή, καθώς παρέχει γνώση σε αυτόν που την χρησιμοποιεί. Ο επενδυτής δεν μπορεί να βασίζεται στο ένστικτό του όπως γινόταν στο παρελθόν, αλλά οφείλει με βάση την πληροφορία που υπάρχει και διαθέτει να λαμβάνει αποφάσεις με το χαμηλότερο ρίσκο. Νέα άτομα με γνώσεις σε τομείς της Πληροφορικής και τα οποία είναι σε θέση να εφαρμόσουν τις τεχνικές της Τεχνητής Νοημοσύνης θα μπορούσαν να συνεργαστούν με έμπειρους επιχειρηματίες, οι οποίοι έχουν πολυετή εμπειρία στον κλάδο της αγοράς ακινήτων. Τα αποτελέσματα μιας τέτοιας συνεργασίας θα ήταν άμεσα

εμφανή, καθώς η τεχνογνωσία των νέων θα μπορούσε να μεταλαμπαδευτεί και να προσφέρει πολύτιμη γνώση στους επαγγελματίες του κλάδου.

Οικονομικές κρίσεις έχουν υπάρξει πολλές κατά το παρελθόν με την τελευταία να λαμβάνει χώρα στις Η.Π.Α. το 2008, τα αποτελέσματα της οποίας είχαν άμεσο και σοβαρό αντίκτυπο στις τιμές των ακινήτων σε πολλές αναπτυγμένες χώρες. Είναι γεγονός ότι κάποιες σταθερές οικονομίες κατάφεραν να αντιμετωπίσουν την κατάσταση αυτή, όπως συνέβη στην Αυστραλία, στην οποία δεν επηρεάστηκαν σημαντικά οι τιμές των ακινήτων. Εν αντιθέσει με την χώρα μας, όπου η χρηματοπιστωτική κρίση των Η.Π.Α. αποτέλεσε την αφετηρία για την οικονομική κρίση που έπληξε την οικονομία μας, τα αποτελέσματα της οποίας είναι ορατά έως και σήμερα. Η εκτίμηση της αξίας ενός ακινήτου θεωρείται κλειδί για οποιαδήποτε συναλλαγή σχετίζεται με το ακίνητο και είναι πολύ σημαντικό η τιμή να αντανακλά την αξία του, ώστε να αποφευχθούν παρόμοιες κρίσεις κατά το μέλλον.

Οι επιχειρήσεις σήμερα αντιμετωπίζουν μείζονα προβλήματα. Ένα από τα προβλήματα αυτά είναι ο υψηλός ανταγωνισμός που υπάρχει μεταξύ των επιχειρήσεων και οι ολοένα αυξανόμενες απαιτήσεις των πελατών. Για να μπορούν να βρουν λύσεις οι επιχειρήσεις του κλάδου απαιτείται να υπάρχει έγκυρη και σωστή πληροφόρηση, εφαρμογή νέων καινοτόμων τεχνολογιών και υιοθέτηση ενός σύγχρονου τρόπου σκέψης.

Τιμές ακινήτων και COVID-19

Η πανδημία που προκλήθηκε από τον COVID-19 είχε σημαντικό αντίκτυπο σε όλες τις χώρες, επηρεάζοντας όλες τις περιοχές. Η αγορά ακινήτων έχει επηρεαστεί όλα αυτά τα χρόνια από διάφορους οικονομικούς, περιβαλλοντικούς παράγοντες και παράγοντες υγείας. Αυτή η νέα πανδημία είχε επίσης επιπτώσεις στη αγορά ακινήτων.

Ο Mohammed et al. πραγματοποίησε μια ανασκόπηση της πρόσφατης βιβλιογραφίας σχετικά με τις συνέπειες του COVID-19 στην αγορά κατοικίας, παρατηρώντας τόσο αρνητικές όσο και θετικές επιπτώσεις στην τιμές κατοικίας, προσφοράς και ζήτησης. Σε ορισμένες περιπτώσεις, υπήρξε αύξηση της τιμής και

της προμήθειας κατοικιών με υψηλότερες ανέσεις ή που βρίσκονται σε προαστιακές περιοχές. Από την άλλη πλευρά, σε άλλες περιοχές, οι τιμές των κατοικιών, της προσφοράς και της ζήτησης μειώθηκαν για διαφορετικούς λόγους. Επιπλέον, εντοπίστηκαν και άλλες αρνητικές επιπτώσεις, όπως δυσκολίες στη συντήρηση της επιστροφής στεγαστικών δανείων και η καθυστέρηση της νέας κατασκευής λόγω υγειονομικών περιορισμών.

Άλλες μελέτες έχουν αναλύσει τις επιπτώσεις της πανδημίας σε διάφορες περιοχές του κόσμου, όπως οι Ηνωμένες Πολιτείες, η Ευρωζώνη, η Ισπανία, η Πολωνία, Κίνα, Αυστραλία και Τουρκία. Το βασικό συμπέρασμα που μπορεί να εξαχθεί από αυτές είναι ότι η τιμή διέφερε ποικίλος από περιοχή σε περιοχή και οι προτιμήσεις του καταναλωτή άλλαξαν αναζητώντας λιγότερο πυκνοκατοικημένες περιοχές στην περιφέρεια.

Στις Ηνωμένες Πολιτείες, ο COVID-19 έκανε τα νοικοκυριά υψηλού εισοδήματος να αναζητήσουν ολιγοπληθής οικογενειακές κατοικίες με μεγαλύτερη επιφάνεια ορόφων, οδηγώντας σε μείωση της τιμής στις πολυπληθής οικογενειακές πολυκατοικίες. Άλλοι συγγραφείς παρατήρησαν ότι οι τιμές των κατοικιών κυμαίνονταν διαφορετικά σε διαφορετικές περιοχές και ότι η ζήτηση κατοικιών αυξήθηκε στην περιφέρεια με χαμηλότερη πυκνότητα πληθυσμού και σε μικρότερες πόλεις μακριά από αστικά κέντρα με υψηλή πληθυσμιακή πυκνότητα.

Στην Ευρωζώνη, οι Battistini et al. περιέγραψαν ότι αρχικά υπήρξε μείωση της δραστηριότητας σε ζήτηση ακίνητης περιουσίας ως συνέπεια των περιορισμών κινητικότητας, αλλά ότι δεν επηρέασε την ανοδική τάση των τιμών (γ' και δ' τρίμηνο 2020) λόγω των πολιτικών και δημοσιονομικών μέτρων που έλαβαν οι κυβερνήσεις.

Στην Ισπανία, οι Alves Álvarez et al. ανέφεραν ότι η δραστηριότητα στην αγορά ακινήτων μειώθηκε έντονα τους πρώτους μήνες μετά την κήρυξη της πανδημίας, με τη δραστηριότητα να ανακάμπτει καθώς άρθηκαν οι περιορισμοί. Οι τιμές των κατοικιών παρουσίασαν γενικευμένη επιβράδυνση πτώσης ανά περιφέρεια, όντας υψηλότερα σε περιοχές των ακτών της Μεσογείου και των νησιών, κυρίως λόγω της μείωσης των ξένων αγοραστών.

Ακίνητα και Τεχνητή Νοημοσύνη

Σήμερα ο κλάδος της αγοράς των ακινήτων συναντά τον κόσμο της Τεχνητής Νοημοσύνης. Τα ερωτήματα που θέτονται είναι κατά πόσο τα πάντρεμα αυτό θα συμβάλει στην ανάπτυξη και την εξέλιξη του τομέα και κατά πόσο οι μέθοδοι της Τεχνητής Νοημοσύνης που αναπτύσσονται θα αποτελέσουν πραγματικά ≪όπλα≫ στα χέρια των κατόχων τους. Η Τεχνητή Νοημοσύνη γενικότερα έχει σκοπό την δημιουργία ≪μηχανών≫, οι οποίες θα είναι προγραμματισμένες να λειτουργούν σχεδόν παρόμοια με την ανθρώπινη λογική αντιγράφοντας τα πεδία των ανθρώπινων λειτουργιών (Jennifer Conway, 2010).

Ο τεράστιος όγκος των δεδομένων που τείνει να αυξάνεται εκθετικά σε συνάρτηση με την συνεχή ανάπτυξη των τεχνικών ανάλυσης των δεδομένων αυτών έχει ως αποτέλεσμα την κατασκευή αποτελεσματικών εφαρμογών, οι οποίες θα εξελίξουν τον κλάδο της αγοράς ακινήτων. Εκμεταλλευόμενοι τους τεράστιους όγκους δεδομένων που βρίσκονται συνεχώς γύρω μας και εφαρμόζοντας τις τεχνικές της Τεχνητής Νοημοσύνης, μπορεί να προέλθει πολύτιμη πληροφόρηση, η οποία θα μετατραπεί σε γνώση και σε λήψη ορθολογικότερων αποφάσεων. Το μεγάλο ερώτημα είναι πως μπορεί να εφαρμοστεί η μέθοδος αυτή στην Αγορά Ακινήτων (Ping-Fenig Pai & Wen-Chang Wang, 2020).

Όλα τα παραπάνω προκύπτουν μέσω της εκπαίδευσης των δεδομένων – μεταβλητών με σκοπό την εκτίμηση της τιμής στόχου. Λαμβάνοντας υπόψη το σύνολο των μεταβλητών αυτών, ο αλγόριθμος εκπαιδεύεται για να μάθει από παρόμοιες ομάδες.

Τα στοιχεία μαρτυρούν ότι μέχρι την παρούσα χρονική περίοδο το ποσοστό που εφαρμόζει τις νέες αυτές τεχνολογίες στον κλάδο της αγοράς ακινήτων διατηρείται σε χαμηλό επίπεδο, και αυτό λαμβάνει χώρα κυρίως στην χώρα μας. Μελλοντικά οι εκτιμήσεις δείχνουν ότι ο κλάδος της αγοράς των ακινήτων μπορεί να αποκομίσει σημαντικά οφέλη από την εφαρμογή αυτών των νέων τεχνολογιών. Απλώς, όπως συμβαίνει στην ιστορία, για να κερδίσει έδαφος κάτι νέο χρειάζεται αρκετό χρόνο (Alejandro Baldominos, Ivan Blanco, Antonio Jose Moreno, Ruben Iturrarte, Oscar Bernardez and Carlos Afonso, 2018).

Η τεχνολογία συνεχώς εξελίσσεται καλύπτοντας συνεχώς τις αυξανόμενες ανάγκες μας. Η Μηχανική Μάθηση μας επιτρέπει να μαθαίνουμε από το παρελθόν για να προβλέψουμε με ακρίβεια το μέλλον. Οι σύγχρονες ≪μηχανές≫ που αναπτύσσονται και το λογισμικό που τις υποστηρίζει μπορούν να συμβάλουν στην επίλυση πολύπλοκων προκλήσεων. Οι τεράστιες εφαρμογές της Μηχανικής Μάθησης μας φέρνουν εγγύτερα στην πραγματική Τεχνητή Νοημοσύνη και δεν θα μπορούσε να μείνει εκτός από αυτές τις εξελίξεις και ο κλάδος της αγοράς ακινήτων (Jose-Luis Alfaro-Navarro, Emilio L. Cano, Esteban Alfaro-Cortes, Noelia Garcia, Matias Gamez and Beatriz Larraz, 2020).

1.3 Πεδίο Έρευνας και Στόχοι

Όπως αναλύθηκε παραπάνω, η μεταβλητότητα της τιμής των ακινήτων, η οποία διαμορφώνεται από την επίσης μεταβλητή αξία τους, έχει ως απόρροια την ανάγκη της δημιουργίας ενός μοντέλου πρόγνωσης αυτής. Το μοντέλο αυτό, επιθυμούμε, λαμβάνοντας υπόψιν τα χαρακτηριστικά του κάθε ακινήτου να προβλέπει, όσο το δυνατόν ακριβέστερα, την τιμή του. Στο πεδίο έρευνας που έχουμε εστιάσει, ο κλάδος της επιστήμης των οικονομικών, ο οποίος ονομάζεται οικονομετρία, σε συνδυασμό με τις μεθοδολογίες προβλέψεων της μηχανικής μάθησης, χρησιμοποιούνται με σκοπό την κατασκευή ενός τέτοιου μοντέλου. Συλλέγοντας δεδομένα που αφορούν την αγορά των ακινήτων στην περιοχή της Αττικής, της Θεσσαλονίκης και του Πειραιά, στόχος μας είναι η κατασκευή ενός μοντέλου, το οποίο βασιζόμενο στην οικονομετρία και με την χρήση τεχνικών μηχανικής μάθησης, θα προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια τις τιμές των ακινήτων τα οποία παρατίθενται προς πώληση.

Στην προσέγγισή μας, εστιάσαμε στη χρήση διαφόρων τεχνικών παλινδρόμησης όπως της μεθόδου Απόφασης Δέντρων, Τυχαίων Δασών, δοκιμάστηκε η απόδοση ενός Νευρωνικού Δικτύου ωστόσο και αλγόριθμοι Παλινδρόμησης Ενίσχυσης Κλίσης κ.α. Τα αποτελέσματα που μας έδωσαν οι τεχνικές παλινδρόμησης Ενίσχυσης Κλίσης και Απόφασης Δέντρων ήταν πιο ικανοποιητικές σε σύγκριση με τα υπόλοιπα.

Τα δεδομένα μας αφορούν ακίνητα, τα οποία έχουν δημοσιευτεί στις ιστοσελίδες της Χρυσής Ευκαιρίας, του Plot.gr κ.α., την περίοδο 2021-2022 στους νομούς της Αττικής, του Πειραιά και της Θεσσαλονίκης. Ως επιμέρους στόχος που προέκυψε μέσα από την πορεία της εργασίας, τοποθετείται η συλλογή, ο καθαρισμός και η στατιστική ανάλυση των δεδομένων των ακινήτων, καθώς το παραπάνω αποτελεί αναγκαίο βήμα στην προσπάθεια μας. Επομένως, μετά την διερευνητική ανάλυση των συλλεγμένων δεδομένων, κατασκευάζεται το σύνολο δεδομένων, που ως είσοδος στα μοντέλα μηχανικής μάθησης θα μας δώσει, ως αποτέλεσμα, την πρόβλεψη της τιμής κάθε ακινήτου.

1.4 Δομή Διπλωματικής Εργασίας

Στη συνέχεια περιγράφεται κάθε βήμα της πορείας προς την επίτευξη του παραπάνω σκοπού και πώς αυτό σκιαγραφείται σε κάθε ενότητα του παρόντος κειμένου.

Στο Κεφάλαιο 2, περιγράφεται το θεωρητικό υπόβαθρο, το οποίο αποτέλεσε τη βάση της μελέτης μας, μαζί με τις αντίστοιχες αναφορές στα πρωτότυπα κείμενα. Χωρίς αυτό, η κατανόηση της εργασίας δεν θα μπορούσε να είναι πλήρης. Γίνεται εδώ, εκτενής αναφορά στις μεθόδους τις οποίες χρησιμοποιήσαμε στην προσέγγιση της λύσης του προβλήματος. Αρχικά, παρουσιάζονται οι βασικές αρχές της εκτίμησης της αξίας των ακινήτων και η σημασία της στην επίλυση ζητημάτων παρόμοιας φύσεως με αυτό που καταπιάνεται η παρούσα διατριβή. Στη συνέχεια, αναλύονται οι προηγμένες μέθοδοι της αποτίμησης της αξίας των ακινήτων στις οποίες στηριχθήκαμε. Και στο τέλος του κεφαλαίου, ακολουθεί η εισαγωγή στην επιστήμη της μηχανικής μάθησης και η αναλυτικότερη περιγραφή των προηγμένων μεθόδων που χρησιμοποιήθηκαν στα πειράματά μας.

Στο Κεφάλαιο 3, γίνεται ανάλυση της μεθοδολογίας που ακολουθήθηκε για τον σχεδιασμό και την υλοποίηση του στόχου αυτής της εργασίας. Αποτυπώνεται λεπτομερώς, η διαδικασία της συλλογής, διερευνητικής ανάλυσης και επεξεργασίας των δεδομένων, καίριου βήματος προκειμένου να επιτευχθεί εμπεριστατωμένη έρευνα. Αναλύονται, κατ' αυτόν τον τρόπο τα επιμέρους βήματα που ακολουθήθηκαν και παρουσιάζονται τα προβλήματα που

αντιμετωπίσαμε στην προσπάθεια της διαμόρφωσης του τελικού συνόλου δεδομένων, και σκιαγραφείται ο τρόπος επίλυσης καθενός εξ αυτών.

Στο Κεφάλαιο 4, παρουσιάζονται τα τελικά αποτελέσματα των αλγορίθμων μηχανικής και βαθίας μάθησης που εφαρμόσθηκαν και αξιολογείται η επίδοσή τους. Ακόμα, αναλύεται η επίδοση κάθε μία από τις μεθόδους που υλοποιήθηκαν και ερμηνεύονται τα αποτελέσματα κάθε διαφορετικού υποσυνόλου των δεδομένων, το οποίο χρησιμοποιήθηκε ως είσοδος σε κάθε αλγόριθμο, καθώς επίσης και η σημασία κάθε χαρακτηριστικού. Εξηγούνται οι λόγοι για τους οποίους οι αλγόριθμοι είχαν τις συγκεκριμένες επιδόσεις. Τέλος, ακολουθεί μία σύνοψη όσων διαπιστώθηκαν, με στόχο την διευκόλυνση της μελλοντικής επιστημονικής έρευνας.

Στο Κεφάλαιο 5, πραγματοποιείται μία σύνοψη των συμπερασμάτων στα οποία καταλήξαμε κατά την πορεία προς την επίτευξη του στόχου μας και παρουσιάζονται κάποιες προτάσεις για περαιτέρω εξέλιξη της παρούσας μελέτης.

Κεφάλαιο 2 – Θεωρητική Ανάλυση

2.1 Η αξία των ακινήτων

Η αξία ενός ακινήτου αρχικά θα μπορούσε να κατηγοριοποιηθεί στην ποιοτική αξία και την ποσοτική αξία. Η ποιοτική αξία αναφέρεται ως η αισθητική αξία ή αλλιώς κοινωνική αξία που προσδοκά να έχει ο αγοραστής του ακινήτου. Στον αντίποδα, ο ποσοτικός προσδιορισμός της αξίας ενός ακινήτου προσδοκά την ύπαρξη ενός κοινού μέτρου μέτρησης των αξιών κι αυτό θα μπορούσε να είναι το χρήμα (Ζεντέλης Παναγιώτης, 2001). Μελλοντικά εικάζουμε ότι είναι πιθανή η ύπαρξη ενός νέου μέτρου μέτρησης των αξιών καθώς ως μέσο αποδεκτής και νόμιμης ανταλλαγής από τους συμβαλλόμενους θα μπορούσε να ήταν η χρήση εικονικών νομισμάτων ή και κρυπτονομισμάτων. Όπως διαφαίνεται η μέθοδος ανταλλαγής μπορεί να διαφέρει από εποχή σε εποχή. Όμως το σημαντικότερο είναι ότι η αξία του ακινήτου θα πρέπει να προκύπτει από ένα ενιαίο σύστημα προσδιορισμού με την ίδια λογική διαδικασία εκτίμησης λαμβάνοντας υπόψη πάντα τα δεδομένα που διαθέτουμε.

Αξία → Διαδικασία Εκτίμησης→ Τιμή

Είναι εξίσου σημαντικό να κατανοήσουμε ότι κάθε ακίνητο έχει μια αξία ανταλλαγής, η οποία προσδιορίζεται από την αξία χρήσης του. Οι διαφορετικοί τρόποι χρήσης μιας ιδιοκτησίας γης προσδίδουν σε αυτήν και διαφορετική αξία ανταλλαγής. Για το λόγο αυτό επικράτησε ο όρος του Ακινήτου ως το άθροισμα των επιμέρους παραγόντων, τα οποία είναι η Γη, η Εργασία, το Κεφάλαιο και η Επιχειρηματικότητα.

Για να διατηρηθεί σταθερή η οικονομία μιας χώρας θα πρέπει να μπορεί να εκτιμά με ακρίβεια την Αξία των Ακινήτων. Η γνώση της συνταγματικά κατοχυρωμένης αυτής αξίας οφείλει να προσδιορίζεται κατά τρόπο ακριβή μέσα από την επιστήμη, την εφαρμογή νέων τεχνολογιών και την χρήση συστηματικής σκέψης. Συγκεκριμένα στην χώρα μας διαφαίνεται από τα στοιχεία που υπάρχουν έως πρόσφατα, η μη ύπαρξη ενός καθολικού συστήματος πρόβλεψης της Αξίας

των Ακινήτων, καθώς η εκτίμηση γίνεται με διαφορετικό κάθε φορά τρόπο ανάλογα με την περίπτωση. Η αξία ενός ακινήτου μπορεί να προσδιοριστεί από ένα σύστημα αντικειμενικού προσδιορισμού που εφαρμόζεται από διάφορους φορείς, όπως το Υπουργείο Οικονομικών, το Σώμα Ορκωτών Ελεγκτών (ΣΟΕ), το Τεχνικό Επιμελητήριο Ελλάδος (ΤΕΕ), η κτηματική υπηρεσία του δημοσίου, η Εφορία κ.α. Παρατηρήθηκε κατά το παρελθόν ότι για το ίδιο ακίνητο προέκυπταν διαφορετικές τιμές εκτίμησης, υποτιμημένες ή υπερτιμημένες, ενδεικτικά με την πραγματική του αξία και οι αποκλίσεις αυτές διέφεραν ως προς την γεωγραφική περιοχή και ως προς το είδος του ακινήτου. Συμπεραίνουμε ότι στο χρόνια που μεσολάβησαν δεν είχαμε την εφαρμογή μια κοινής ενιαίας νομοθετικής ρύθμισης που θα προέβλεπε την αξία ενός ακινήτου για διάφορους σκοπούς, όπως αγοροπωλησία, φορολογία, απαλλοτριώσεις, κτλ. Αυτό το γεγονός μπορεί να επιφέρει και ανάλογο οικονομικό αντίκτυπο στην οικονομία της χώρας μέσω αποθάρρυνσης των επενδύσεων.

Αξίζει να αναφέρουμε πως η χώρα μας κινείται προς μια κοινή ευρωπαϊκή κατεύθυνση εφαρμογής των νέων τεχνολογιών και καινοτομιών με απώτερο σκοπό τον εκσυγχρονισμό της οικονομίας, συμπεριλαμβανομένης αυτής και του κλάδου των ακινήτων. Σύμφωνα με την Βίβλο Ψηφιακού Μετασχηματισμού 2021 - 2025 η χώρα μας προσανατολίζεται στην εφαρμογή νέων μέτρων με τα οποία θα μπορέσει να ξεπεράσει αποτελεσματικά τα εμπόδια του παρελθόντος και τα οποία ταλάνιζαν τον κλάδο των ακινήτων και της οικονομίας γενικότερα. Έτσι με την υιοθέτηση μιας αναπτυξιακής εθνικής στρατηγικής για τα δεδομένα, η οποία θα είναι στοιχισμένη με την Ευρωπαϊκή πολιτική θα προσβλέπει στην συγκέντρωση ενός μεγάλου όγκου δεδομένων τα οποία θα είναι ποιοτικά και τα οποία θα επιτρέπουν την αποτελεσματική τους χρήση. Η ανοιχτή διάθεση, η επαναχρησιμοποίηση και η μέγιστη αξιοποίηση των δεδομένων του δημοσίου τομέα αφενός θα ενισχύσει την συμμετοχή, την διαφάνεια καθώς και τον έλεγχο και αφετέρου θα προωθήσει την επιχειρηματικότητα, καθώς θα εξασφαλιστεί ότι οι επιχειρήσεις που δραστηριοποιούνται στον τομέα των νέων τεχνολογιών θα έχουν και τα ανάλογα οφέλη και την πρώτη ύλη για να αναπτύξουν υπηρεσίες προστιθέμενης αξίας. Η στρατηγική επικεντρώνεται στην ανάπτυξη δράσεων για την σταδιακή αξιοποίηση των δεδομένων ως θεμελιώδους υποδομής. Τα υψηλής αξίας δεδομένα δίνουν πλέον την εκκίνηση που μπορούν να φέρουν εν τέλη την

αξιοποίηση δημόσιων, επιχειρηματικών και επιστημονικών δεδομένων μαζικά. Ο βασικός στόχος που θα πραγματωθεί με την αξιοποίηση των ψηφιακών δεδομένων και συστημάτων είναι ο εξ ορθολογισμός του συστήματος αντικειμενικού προσδιορισμού της αξίας των ακινήτων καθώς και η τυποποίηση της διαδικασίας των εκτιμήσεων ώστε να αποφέρει καλύτερα, πιο αξιοποιήσιμα και πιο αντικειμενικά αποτελέσματα με σκοπό την συμβολή σε ένα πιο σταθερό οικονομικό περιβάλλον. Η δημιουργία μιας Βάσης Δεδομένων στοιχείων των Ακινήτων και των αξιών τους, καθώς και η υλοποίηση ενός ολοκληρωμένου γεωπληροφορικού συστήματος μαζικής εκτίμησης των αξιών των ακινήτων (CAMA) έχουν τεθεί ως πρωταρχικοί στόχοι.

Αναφορικά με την δημιουργία του Μητρώου ακινήτων και της αξίας τους, αναφέρεται ότι θα συγκεντρώνει όλα τα στοιχεία που επηρεάζουν την Αγορά των ακινήτων και θα ενημερώνεται με στοιχεία που συγκεντρώνονται από όλες τις διαθέσιμες πηγές, όπως υπηρεσίες και φορείς του ευρύτερου δημοσίου φορέα (με έμφαση στοιχεία του Κτηματολογίου και της Εφορίας (ΑΑΔΕ), καθώς και ανοιχτά δεδομένα. Ενδεικτικά δεδομένα που θα έχει σκοπό να συγκεντρώνει το σύστημα είναι:

1) Δεδομένα αξιών των ακινήτων (όπως αξίες συμβολαιογραφικών πράξεων, αξίες μισθολογικών συμβολαίων, εκτιμηθείσες αξίες και αξίες απαλλοτρίωσης),
2) Δεδομένα χαρακτηριστικών ακινήτων (όπως αξίες χρήσης, επιφάνειας ακινήτων, όροφος, εγγύτητας σε σημαντικούς κόμβους και υπηρεσίες, όπως νοσοκομεία, στάσεις του μετρό, σχολεία, πανεπιστήμια, κτλ.).

Το Μητρώο ακινήτων θα αποτελεί ένα πολυχρηστικό και ένα ολοκληρωμένο Πληροφορικό Σύστημα (ΟΠΣ) της Γενικής Γραμματείας Οικονομικής Πολιτικής (ΓΓΟΠ) του Υπουργείου Οικονομικών και θα συμβάλει στον προσδιορισμό της αντικειμενικής αξίας των ακινήτων, καθώς θα παρέχει τα δεδομένα των ακινήτων καθώς και των αξιών τους αντίστοιχα εντοπισμένα γεωχωρικά. Αναφορικά με το ολοκληρωμένου γεωπληροφορικού συστήματος μαζικής εκτίμησης των αξιών των ακινήτων (CAMA), το οποίο θα τεθεί σε λειτουργία, θα είναι η δημιουργία ενός πολυχρηστικού και ολοκληρωμένου Πληροφορικού Συστήματος (ΟΠΣ)

κεντρικής γεωχωρικής υποδομής το οποίο θα ενημερώνεται και θα ελέγχεται από την Γενική Γραμματεία Οικονομικής Πολιτικής (ΓΓΟΠ) του Υπουργείου Οικονομικών και στο οποίο θα πραγματοποιείται συγκεντρωτικά μαζική εκτίμηση των αξιών των ακινήτων (Computer Assisted Mass Appraisal – CAMA), δημοσίων και ιδιωτικών, κατά τρόπο ομοιόμορφο, δίκαιο, νομικά κατοχυρωμένο και πλήρως αυτοματοποιημένο.

Έτσι από την μεριά του χρήστη θα παρέχεται η δυνατότητα εντοπισμού του ακινήτου χαρτογραφικά ενημερωμένο με τα δεδομένα που συλλέγονται και υπάρχουν σε πραγματικό χρόνο. Με τον τρόπο αυτό δίνεται η δυνατότητα να βλέπει τα πολεοδομικά τετράγωνα, το υπάρχον οδικό δίκτυο, την πολεοδομική διαρρύθμιση, την διοικητική διαίρεση και τα σημεία ενδιαφέροντος, τα οποία θα απεικονίζονται σε ένα γεωαναφερόμενο υπόβαθρο ορθοφωτοχαρτών του Εθνικού Κτηματολογίου.

Επιπλέον θα του δίνεται η δυνατότητα πλοήγησης μέσω δια δραστικών χαρτών, να εισάγει δεδομένα τα οποία τον ενδιαφέρουν με σκοπό την απόκτηση πληροφορίας που να του δίνει την δυνατότητα να μαθαίνει την εκτιμώμενη αξία του ακινήτου που τον ενδιαφέρει. Το σύστημα θα παρέχει διαδικτυακές υπηρεσίες προς τρίτα συστήματα, θα διαθέτει στατιστικά στοιχεία προς αξιοποίηση, γεγονός που θα δίνει την δυνατότητα σε όλους τους φορείς και οργανισμούς να ενημερώνονται βάση αυτού του Ολοκληρωμένου Πληροφοριακού Συστήματος. Το σύστημα θα ανταποκρίνεται άμεσα και θα υποστηρίζει διοικητικές διαδικασίες ενστάσεων και προσφυγών, καθώς θα προσβλέπει στην λήψη μιας δίκαιης απόφασης εκτίμησης των αξιών κάθε ακινήτου με βάση την δυναμική της Κτηματαγοράς, αξιοποιώντας τα στοιχεία του Μητρώο Ακινήτων και Αξιών. Είναι σημαντικό τέλος να αναφέρουμε ότι θα τεκμηριώνεται σαφώς η μέθοδος υπολογισμού της αξίας κάθε ακινήτου με γνώμονα την διαφάνεια, την αμεροληψία και την εγκυρότητα. Η αξιοποίηση των δεδομένων αποτελεί την κοινή συνισταμένη για την ορθή λήψη των αποφάσεων. Έτσι οι αποφάσεις, όπως στην περίπτωση την διαμόρφωσης της αξίας των ακινήτων, θα πρέπει να λαμβάνονται με βάση μια κοινή επιστημονικά τεκμηριωμένη μέθοδο, η οποία αξιοποιεί αποτελεσματικά την χρήση των δεδομένων (Βίβλος Ψηφιακού Μετασχηματισμού 2020 – 2025).

Αρχικά είδαμε ότι οι παράγοντες που επηρεάζουν την τιμή ενός ακινήτου μπορεί να κατηγοριοποιούνται ποιοτικά ή ποσοτικά. Γενικότερα, η εκτίμηση της τιμής ενός ακινήτου αποτελεί μια πολύπλοκη και δύσκολη διαδικασία, η οποία εξαρτάται από πολλούς έμμεσους καθώς και άμεσους παραμέτρους, οι οποίοι αναπόφευκτα ασκούν επιρροή στην ακρίβεια της πρόβλεψης. Στην επόμενη ενότητα θα αναφερθούμε στους κύριους παράγοντες που παίζουν σημαντικό ρόλο στην διαμόρφωση της αξίας κάθε ακινήτου και οι οποίοι λαμβάνονται υπόψη στην μέθοδο της ανάλυσης των δεδομένων με την χρήση των εργαλείων της Μηχανικής Μάθησης.

2.2 Παράγοντες που επηρεάζουν την Αξία των Ακινήτων

Στην παρούσα ενότητα παρουσιάζονται οι κυριότεροι παράγοντες που διαμορφώνουν την τιμή ενός ακινήτου. Σε γενικές γραμμές οι παράγοντες αυτοί μπορεί να κατηγοριοποιούνται ποιοτικά ή ποσοτικά. Η εκτίμηση της τιμής ενός ακινήτου, όπως προαναφέρθηκε, αποτελεί μια πολύπλοκη και δύσκολη διαδικασία, η οποία εξαρτάται από πολλούς άμεσους και έμμεσους παραμέτρους, οι οποίοι ασκούν αναπόφευκτα επιρροή στην ακρίβεια της πρόβλεψης. Η μοναδικότητα που έχει κάθε ακίνητο, στον γεωγραφικό χώρο τον οποίο βρίσκεται καθώς και για την χρήση για την οποία προδιαθέτετε, σημαίνει ότι θεωρητικά υπάρχει ένας μη πεπερασμένος αριθμός παραγόντων που διαφοροποιεί την αξία των δεδομένων ανομοιογενώς.

Για να μπορέσουμε να ομαλοποιήσουμε την υπάρχουσα κατάσταση θα χρησιμοποιήσουμε μια διαβάθμιση κλιμακωτά που θα μελετά αρχικά την κατάσταση σε επίπεδο χώρας καταλήγοντας εν τέλει στο επίπεδο του τμήματος του συγκεκριμένου ακινήτου. Στο επίπεδο χώρας η επίδραση που ασκούν οι πολιτικοί, κοινωνικοί και οικονομικοί παράγοντες παίζουν σαφώς καθοριστικό ρόλο στην διαμόρφωση της τιμής του. Ξεκινώντας με τους πολιτικούς παράγοντες, η εφαρμογή μιας ενιαίας και προκαθορισμένης πολιτική γης και η τεχνική νομοθεσία που εφαρμόζεται θα πρέπει να είναι διαχρονικά σταθερή. Επίσης η στάση, η οποία διαμορφώνει η κυβέρνηση και η άσκηση γενικής

πολιτικής που εφαρμόζει έχει σημαντικό αντίκτυπο στην Αγορά των Ακινήτων (Ζεντέλης Παναγιώτης, 2001).

Στον Ελλαδικό χώρο κατά το πρόσφατο παρελθόν παρατηρήθηκε το φαινόμενο της κερδοσκοπίας στην εμπορία της Γης, κυρίως σε τουριστικές περιοχές. Αυτό σημαίνει ότι οι μεσίτες και επενδυτές αγόραζαν ένα ακίνητο με σκοπό να το μεταπωλήσουν αργότερα σε υψηλότερη τιμή. Ακόμα πιο συγκεκριμένα παρατηρήθηκε το φαινόμενο του πολυτεμαχισμού (οικοπεδοποίηση), δηλαδή η μεγάλη ζήτηση οδήγησε στην διαίρεση μεγάλων οικοπέδων σε μικρότερα τμήματα και στην μεταπώλησή τους. Το αποτέλεσμα ήταν η δημιουργία μιας άναρχης ανάπτυξης και η υποβάθμιση του περιβάλλοντος χώρου. Μακροπρόθεσμα, όπως εκτιμάται, η άναρχη δόμηση της Γης δεν θα μπορούσε να έχει θετικά αποτελέσματα, καθώς η μικρότερης έκτασης ιδιοκτησία παρουσιάζει απώλεια αξίας (όπως περιορισμός κατασκευής, γειτνίαση με πολλούς, περιορισμός εκμετάλλευσης χώρου, κ.α.). Για το λόγο αυτό όπως διαπιστώνεται, η στάση της κυβέρνησης των χωρών και η πολιτική που ασκεί στον χώρο επηρεάζει σε μέγιστο βαθμό και τους παράγοντες της αγοράς ακινήτων.

Εν συνεχεία αναφερόμενοι στους κοινωνικούς παράγοντες, το υπάρχον κοινωνικό σύστημα με τις ιδέες, τις ιδεολογίες και τα ιδεώδη που διαμορφώνονται δημιουργούν μια αλυσίδα αποτελεσμάτων. Η κοινωνική ισχύ και οι κοινωνικές σχέσεις που αναπτύσσουν μεταξύ τους οι ιδιοκτήτες ακινήτων διαμορφώνουν μια κοινωνική δυναμική που ασκεί επιρροή και στις αξίες της ιδιοκτησίας. Οι δημογραφικές εξελίξεις και διαφοροποιήσεις καθώς και η γενικότερη διάρθρωση του πληθυσμού (όπως η ηλικιακή ομάδα, το φύλο, η πυκνότητα και η κινητικότητα του πληθυσμού) παίζουν επίσης καθοριστικό ρόλο στην διαμόρφωση της τιμής των ακινήτων.

Ακολούθως, η οικονομική πολιτική που ασκείται από μια χώρα, δημιουργεί και την ανάλογη πίεση στις αξίες των ακινήτων. Οι τιμές των ακινήτων θα πρέπει να προσδιορίζονται με μεγάλη ακρίβεια. Συγκεκριμένα αυτό που αξίζει πάντα να επισημαίνεται είναι ότι η τιμή του ακινήτου θα πρέπει να αντιπροσωπεύει σε μεγάλο βαθμό την αξία του. Σήμερα οι αποφάσεις θα πρέπει να λαμβάνονται με γνώμονα την συλλογή επαρκής πληροφόρησης και δεν θα πρέπει να βασίζονται κυρίως στις ανθρώπινες εκτιμήσεις (Jordan & Ellen, 2009). Οι οικονομικοί

παράγοντες ασκούν πολύ μεγάλη επιρροή στην διαμόρφωση της τιμής των ακινήτων. Η οικονομική σταθερότητα και η άσκηση γενικότερα μιας σταθερής οικονομικής πολιτικής παίζουν σημαντικό ρόλο και έχουν καθοριστικό αντίκτυπο και στον κλάδο της αγοράς ακινήτων.

Παράλληλα, εξετάζονται οι παράμετροι που διαμορφώνουν τις τιμές των κατοικιών σε επίπεδο τμήματος μια; πόλης. Γενικότερα οι βασικές ιδέες περιλαμβάνουν τον διαχωρισμό σε ζώνες ομοιόμορφης συμπεριφοράς που με βάση κάποια κριτήρια μπορεί να γίνει η ταξινόμηση ανάλογα με:

- την διοικητική δομή της πόλης (δημοτική ενότητα)
- την διαβάθμιση της ανάπτυξης (πυκνότητα των ακίνητων, δια δραστικοί χώροι, χώροι αναψυχής)
- τον τύπο χρήσεως (οικιστική, εμπορική, βιομηχανική ζώνη)
- την διαβάθμιση της υποβάθμισης του περιβάλλοντος
- το διαχωρισμό των κοινωνικών ομάδων
- την ακριβή τοποθεσία (απόσταση από το κέντρο, θάλασσα, τις δημόσιες συγκοινωνίες, κτλ.)
- την ακριβή περιοχή (βόρεια, νότια, ανατολικά & δυτικά της πόλης)

Είναι φανερό ότι τα κριτήρια αυτά μπορούν να επιδράσουν με διαφορετικό τρόπο στις αξίες της Αγοράς Ακινήτων. Τα δεδομένα που λήφθηκαν υπόψη στην δημιουργία του μοντέλου κατηγοριοποιούνται με βάση την παραπάνω ταξινόμηση, καθώς οι βασικές μεταβλητές που χρησιμοποιούνται και προσδιορίζουν τα χαρακτηριστικά των κατοικιών είναι:

- Το προάστιο ή η γειτονιά στο οποίο βρίσκεται το σπίτι
- Η διεύθυνση στην οποία ανήκει
- Ο τύπος του σπιτιού
- Το μέγεθος του σπιτιού
- Η απόσταση από το κέντρο
- Η δημοτική ενότητα που ανήκει

- Το γεωγραφικό πλάτος & μήκος
- Η περιοχή γενικά
- Το πλήθος των ακίνητων που υπάρχει σε μια ζώνη

Η ροή ανάπτυξης αποτελεί πρωτεύον παράγοντα ανάπτυξης μιας πόλης.

Μακροπρόθεσμα θα πρέπει να εφαρμόζονται πολιτικές που να κατευθύνουν μια ομοιόμορφή ανάπτυξη στα διάφορα τμήματα της πόλεως, καθώς η ισόρροπη ανάπτυξη θεωρητικά είναι πιο αποτελεσματική. Στην πραγματικότητα όμως δημιουργούνται ζώνες, όπου η κάθε ζώνη έχει διαφορετική ζήτηση και συνεπώς αυτό συντελεί κι στην διαμόρφωση διαφορετικής ζώνης τιμών. Γενικότερα, εξετάζονται διαφορετικοί συντελεστές, οι οποίοι επηρεάζουν σημαντικά την ανάπτυξη μιας περιοχής και διαμορφώνουν άλλες ζώνες τιμών στην ίδια πόλη. Αρχικά αναφέρονται τα δίκτυα υποδομής. Όταν γίνονται σημαντικά έργα ανάπλασης και μεγάλα έργα υποδομής σε τμήματα μιας πόλης, αυτό δημιουργεί και τις ανάλογες διαφοροποιήσεις, όπου οι μικροοικονομικές και μακροοικονομικές επιπτώσεις αναμένεται να επιδράσουν στην αξία της αγοράς ακινήτων. Παραδείγματος χάρη, η κατασκευή ενός νοσοκομείου, ενός καινούργιου αεροδρομίου ή σταθμού του μετρό, επηρεάζει και διαμορφώνει τις επιλογές των αγοραστών με τελικό αντίκτυπο την τιμή του ακινήτου, όπου διαμορφώνεται σε κάθε περίπτωση ανάλογα. Επισημαίνεται ότι το είδος δραστηριότητας, το οποίο αναπτύσσει κάθε γεωγραφική ζώνη (όπως εμπορική, τουριστική, βιομηχανική, οικιστική, κ.α.) σε συνδυασμό με την υποβάθμιση του περιβάλλοντος δημιουργεί κι τον ανάλογο χωρικό διαχωρισμό. Έτσι, αναφορικά με αυτές τις συγκυρίες γίνεται και μια αντίστοιχη κατανομή του πληθυσμού και των δραστηριοτήτων. Άτομα με υψηλά εισοδηματικά κριτήρια και καλύτερο βιοτικό επίπεδο μετακινούνται και κατ' επέκταση συγκεντρώνονται σε συγκεκριμένες περιοχές, με αποτέλεσμα να επηρεάζουν ακόμα θετικότερα την τιμή των ακινήτων της συγκεκριμένης ζώνης. Σαφώς στην επιλογή για την ανεύρεση της κατάλληλης ιδιοκτησίας, λαμβάνονται πάντα υπόψη, τα χαρακτηριστικά του τμήματος της περιοχής καθώς και τα χαρακτηριστικά του συγκεκριμένου ακινήτου (Ζεντέλης Παναγιώτης, 2001).

Τα κυριότερα εξ' αυτών των χαρακτηριστικών είναι τα φυσικά, όπως ο προσανατολισμός, η κλίση, το υψόμετρο, η ποιότητα του κλίματος, ο περιβάλλον χώρος του (δέντρα, θάλασσα, θέα, κτλ) και η γενικότερη θέση του, όπου προσδίδουν στο ακίνητο την ιδιαιτερότητα που έχει. Αναφερόμενοι συνοπτικά στις έννοιες αυτές ο περιβάλλον χώρος, περιλαμβάνει εξίσου σημαντικά ποιοτικά χαρακτηριστικά που ασκούν θετική είτε αρνητική επιρροή στην αξία της αγοράς ακινήτων. Για το λόγο αυτό ένα ακίνητο, το οποίο ανήκει στο ίδιο τμήμα της πόλης και στην ίδια ζώνη μπορεί να διαφέρει αρκετά ως προς την γειτνίαση του και στο περιβάλλον χώρο και για τον λόγο αυτό να παρουσιάζει διαφορετική ζήτηση. Στα θετικά συγκαταλέγεται η γειτνίαση με πάρκα και γενικά πράσινους χώρους καθώς και η αισθητική των γύρω οικοδομημάτων. Επίσης η θέα σε κάποιο βουνό, θάλασσα ή ωραίο φυσικό τοπίο, αναμενόμενα δημιουργεί πλεονέκτημα. Στον αντίποδα, η γειτνίαση του ακινήτου σε χώρο, όχι ιδιαίτερα καλής αισθητικής, (όπως νοσοκομεία, νεκροταφείο, χωματερές, κτλ.) έχει τεράστια αρνητική επίδραση στην τιμή του. Επίσης ακίνητα, τα οποία γειτνιάζουν σε μεγάλους δρόμους, βιομηχανικές περιοχές και κοντά σε μεγάλα αεροδρόμια συναντούν σημαντικούς περιορισμούς και μειωμένη ζήτηση, τα οποία συνεπώς αμβλύνουν την τιμή του (Ζεντέλης Παναγιώτης, 2001).

Ανακεφαλαιώνοντας, κάθε ακίνητο είναι ξεχωριστό, διότι έχει τις δικές του ιδιαιτερότητες και χαρακτηριστικά. Αξίζει να αναφερθεί ότι οι παράγοντες που αναφέρθηκαν, ξεκινώντας από το επίπεδο της κλίμακας χωρών και καταλήγοντας σε επίπεδο του τμήματος του ακινήτου, είναι ένα μέρος αυτών. Σαφώς υπάρχουν και άλλοι παράγοντες, όπου επηρεάζουν και διαμορφώνουν την τιμή των ακινήτων. Στην παρούσα εργασία προσπαθούμε να προσεγγίσουμε τους κυριότερους εξ' αυτών, καθώς όπως αναφέραμε αρχικά η εκτίμηση της αξίας κάθε ακινήτου αποτελεί μια πολύπλοκη και δύσκολη διαδικασία που εξαρτάται από πάρα πολλές μεταβλητές. Στο επόμενο Κεφάλαιο αναφερόμαστε στην έννοια της Τεχνικής Νοημοσύνης και Μηχανικής Μάθησης, όπου θα δούμε ότι η εφαρμογή των νέων μεθόδων και τεχνολογιών θα μπορέσει να συμβάλει καθοριστικά στην λύση αυτών των προβλημάτων.

2.3 Μέθοδοι αποτίμησης

Κάθε χώρα έχει διαφορετική κουλτούρα και εμπειρία, η οποία συντελεί στο να καθορίσει τις μεθόδους που θα υιοθετηθούν για κάθε συγκεκριμένη αποτίμηση. Είναι χαρακτηριστικό ότι η πλειονότητα όλων των μεθόδων θα βασίζεται σε κάποια μορφή σύγκρισης για την εκτίμηση της αγοραίας αξίας. Αυτό θα μπορούσε στην απλούστερη μορφή του να γίνει με άμεση σύγκριση κεφαλαίων ή μπορεί να βασίζεται σε μια σειρά παρατηρήσεων που επιτρέπουν στον εκτιμητή να προσδιορίσει ένα μοντέλο παλινδρόμησης. Οποιαδήποτε τέτοια μέθοδος θα την αναφέρουμε ως παραδοσιακή.

Άλλα μοντέλα ή μέθοδοι εκτίμησης προσπαθούν να αναλύσουν την αγορά μιμούμενοι άμεσα τις διαδικασίες σκέψης των παικτών στην αγορά σε μια προσπάθεια να εκτιμήσουν το σημείο ανταλλαγής. Αυτά τα μοντέλα τείνουν να είναι πιο ποσοτικά στη μέθοδο και θα αναφέρονται ως προηγμένα. Θεωρητικά, η αξία που προσδίδεται σε ένα ακίνητο είναι συνάρτηση ποιότητας και ποσότητας. Η ποσότητα αφορά τη φυσική διάσταση του ακινήτου (χώρος), ενώ η ποιότητα είναι μάλλον υποκειμενική.

Οι μέθοδοι εκτίμησης διακρίνονται σε δύο μεγάλες κατηγορίες, όπου είναι οι παραδοσιακές μέθοδοι αποτίμησης και οι προηγμένες μέθοδοι αποτίμησης.

Παραδοσιακές μέθοδοι αποτίμησης

1. Συγκριτική μέθοδος
2. Προσέγγιση κόστους
3. Προσέγγιση εισοδήματος
4. Μέθοδος κερδών
5. Μέθοδος ανάπτυξης / υπολειμματικής
6. Μέθοδος πολλαπλής παλινδρόμησης
7. Ηδονικές μέθοδοι πώλησης
8. Μέθοδος επανάληψης πωλήσεων
9. Μοντέλο ισοδύναμης απόδοσης
10. Μοντέλο επιλογών

Προηγμένες μέθοδοι αποτίμησης

1. Τεχνητά νευρωνικά δίκτυα (ANN)
2. Συνελικτικά νευρωνικά δίκτυα (ConvNets)
3. Μοντέλα SVC
4. Μοντέλο που βασίζεται σε Τεχνητή Νοημοσύνη
5. Μοντέλα Decision Tree & Random Forest
6. Ιεραρχικό μοντέλο
7. Ανάλυση συστάδων
8. Ακατέργαστη θεωρία συνόλων και θεωρία ασαφών συνόλων
9. Μοντέλο που βασίζεται σε GIS
10. Μοντέλο που βασίζεται σε MIX
11. Άλλα μοντέλα

Η παρούσα διπλωματική εργασία βασίζεται στην εφαρμογή προηγμένων μεθόδων αποτίμησης. Προηγμένοι μέθοδοι αποτίμησης αναπτύχθηκαν ευρέως τα τελευταία χρόνια και η εφαρμογή τους, όπως προκύπτει, φαίνεται να έχει θετική έκβαση στην εκτίμηση της αξίας των ακινήτων.

Γενικότερα, όπως θα δούμε παρακάτω, η παρούσα εργασία στοχεύει στην εφαρμογή προηγμένων μοντέλων και στην εφαρμογή της μαζικής αποτίμησης. Λαμβάνοντας υπόψη την τεράστια ανάπτυξη της επιστήμης της πληροφορικής σε συνδυασμό με τον μεγάλο όγκο των δεδομένων που υπάρχουν, θα προκύψει καλύτερη πληροφόρηση, τόσο για τους εκτιμητές όσο και για τα μέρη τα οποία συναλλάσσονται. Στην παρούσα εργασία θα προσπαθήσουμε να εκτιμήσουμε την τιμή των ακινήτων εφαρμόζοντας τις μεθόδου; του Decision Tree & Random Forest, Νευρωνικών Δικτύων και παλινδρόμησης MLR (Multiple Linear Regression) κ.α. από τις οποίες συμπερασματικά προκύπτουν αξιόπιστα αποτελέσματα.

2.3.1 Προηγμένες μέθοδοι αποτίμησης

Τεχνητά νευρωνικά δίκτυα (ANN)

Τα μοντέλα τεχνητών νευρωνικών δικτύων έχουν προσφερθεί ως πιθανή λύση σε πολλά προβλήματα στην αποτίμηση ακινήτων. Ένα μοντέλο τεχνητού νευρωνικού δικτύου πρέπει πρώτα να εκπαιδευτεί από ένα σύνολο δεδομένων και το μοντέλο στη συνέχεια να χρησιμοποιηθεί για την εκτίμηση των τιμών των νέων ακινήτων από την ίδια αγορά. Τα νευρωνικά δίκτυα είναι μοντέλα τεχνητής νοημοσύνης που αρχικά σχεδιάστηκαν για να αναπαράγουν τις διαδικασίες μάθησης του ανθρώπινου εγκεφάλου.

Αυτά τα μοντέλα έχουν τρία κύρια στοιχεία:

1. το επίπεδο δεδομένων εισόδου
2. τα κρυφά στρώματα, που συνήθως αναφέρονται ως ≪μαύρο κουτί≫
3. το επίπεδο μέτρησης εξόδου, οι εκτιμώμενες αξίες ιδιότητας

Τα κρυφά στρώματα περιέχουν δύο διεργασίες: τις σταθμισμένες συναρτήσεις άθροισης και τις συναρτήσεις μετασχηματισμού. Και οι δύο αυτές συναρτήσεις συσχετίζουν τις τιμές από τα δεδομένα εισόδου (π.χ. τα χαρακτηριστικά ιδιοκτησίας: αριθμός δωματίων, ηλικία σπιτιού, μέγεθος οικοπέδου, επιφάνεια υπογείου, συνολική επιφάνεια, αριθμός γκαράζ) με τα μέτρα παραγωγής (την τιμή πώλησης).

Ένα σημαντικό πλεονέκτημα των ANN στη μοντελοποίηση συστημάτων είναι ότι δεν υπάρχει ανάγκη επιβεβαίωσης του μοντέλου εκ των προτέρων. Εκπαιδεύοντας τα δείγματα δεδομένων εισόδου, το ANN προσαρμόζεται για να αναπαράγει την έξοδο. Μία από τις πιο δημοφιλείς δομές ANN είναι το πολυστρωματικό (MLP). Το ANN αποδίδει καλά στη μοντελοποίηση της μη γραμμικής σχέσης λόγω των χαρακτηριστικών της ημιπαραμετρικής παλινδρόμησης και εξακολουθεί να είναι το πιο δημοφιλές μοντέλο που χρησιμοποιείται σε μοντέλα που βασίζονται σε Τεχνητή Νοημοσύνη.

Συνελικτικά νευρωνικά δίκτυα (ConvNets)

Τα συνελικτικά νευρωνικά δίκτυα (ConvNets) έχουν επιτύχει επιδόσεις αιχμής σε εργασίες όπως η αναγνώριση εικόνας, η τμηματοποίηση, η ανίχνευση αντικειμένων και το γενετικό μοντέλο.

Ένα από τα πρώιμα σημαντικά ConvNets ήταν το LeNet, το οποίο χρησιμοποιήθηκε για την ταξινόμηση χειρόγραφων ψηφίων. Ωστόσο, μόλις πρόσφατα τα ConvNets απέδωσαν ανώτερη απόδοση σε εργασίες ταξινόμησης και ανίχνευσης αντικειμένων σε σύγκριση με προηγούμενες μεθόδους, όπως χειροποίητα χαρακτηριστικά και μηχανές υποστήριξης διανυσμάτων. Η επιτυχία του AlexNet στην ταξινόμηση εικόνων στο ImageNet Large-Scale Visual Recognition Challenge (ILSVRC 2012) έχει αναζωπυρώσει το ενδιαφέρον για τα ConvNets. Οι αρχιτεκτονικές αιχμής όπως τα ResNets, τα Highway Networks και το DenseNet επιτρέπουν καλύτερες πληροφορίες και ροή κλίσης χρησιμοποιώντας επίπεδα παράλειψης ή συνδέοντας διαφορετικά επίπεδα εντός του δικτύου.

Μία από τις κύριες προκλήσεις για την ανάθεση ενός συγκεκριμένου στυλ σε μια εικόνα είναι ότι το στυλ είναι δύσκολο να καθοριστεί αυστηρά, καθώς η ερμηνεία του μπορεί να διαφέρει από άτομο σε άτομο. Στην προκειμένη περίπτωση μας ενδιαφέρει να κωδικοποιήσουμε πληροφορίες σχετικές με το επίπεδο πολυτέλειας των φωτογραφιών ακινήτων. Μια προσέγγιση για την πρόβλεψη του στυλ των εικόνων ορίζει διαφορετικούς τύπους στυλ εικόνας και συγκεντρώνει ένα σύνολο δεδομένων μεγάλης κλίμακας από φωτογραφίες με σχολιασμό στυλ που περιλαμβάνει πολλές διαφορετικές πτυχές του οπτικού στυλ. Επίσης, συγκρίνει διαφορετικά χαρακτηριστικά εικόνας για την εργασία πρόβλεψης στυλ και δείχνει ότι τα χαρακτηριστικά που λαμβάνονται από βαθιά συνελικτικά νευρωνικά δίκτυα (ConvNets) υπερτερούν των άλλων χαρακτηριστικών.

Μοντέλα SVC

Τα μοντέλα με διακριτικά μεταβαλλόμενους συντελεστές (SVC) επιτρέπουν στα οριακά φαινόμενα να είναι μη ακίνητα στο χώρο και έτσι προσφέρουν υψηλότερο βαθμό ευελιξίας. Ταυτόχρονα, τα μοντέλα SVC έχουν το πλεονέκτημα ότι είναι

εύκολα ερμηνεύσιμα. Έχουν δημοσιευτεί αρκετές μεθοδολογίες και εφαρμογές με μοντέλα SVC. Για να αναφέρουμε δύο, γεωγραφικά σταθμισμένη παλινδρόμηση (GWR) από τους Fotheringham et al. (2002) και ένα Bayesian πλαίσιο με διαδικασίες SVC από τους Gelfand et al. (2003) αποτελούν κυρίως εξέχοντα παραδείγματα. Σε μια μελέτη προσομοίωσης οι Wheeler και Calder (2007) καταλήγουν στο συμπέρασμα ότι οι διαδικασίες SVC παρέχουν πιο ακριβείς εκτιμήσεις συντελεστών παλινδρόμησης από το GWR. Μια περαιτέρω σύγκριση των διεργασιών GWR και SVC δίνεται από τον Finley (2010) σχετικά με οικολογικά δεδομένα. Αποδεικνύεται ότι οι διαδικασίες SVC έχουν γενικά καλύτερη προγνωστική απόδοση. Ωστόσο, όταν πρόκειται για την εκτίμηση μοντέλων SVC σε μεγάλα δεδομένα, οι περισσότερες από τις καθιερωμένες μεθοδολογίες αντιμετωπίζουν προβλήματα. Επί του παρόντος διαθέσιμες υλοποιήσεις Μπεϋζιανών προσεγγίσεων, όπως οι προγνωστικές διεργασίες Gauss, που παρουσιάζονται στους Banerjee et al. (2008), είτε περιορίζονται από τον αριθμό των SVC σε ένα μοντέλο, είτε από τον αριθμό των παρατηρήσεων. Επομένως, απαιτείται μια γεωστατιστική μέθοδος εκτίμησης και πρόβλεψης που, αφενός, μπορεί να αντιμετωπίσει μεγάλο αριθμό παρατηρήσεων και, αφετέρου, να μπορεί να εφαρμοστεί σε μοντέλα που περιλαμβάνουν πολλά SVC. Αυτοματοποιημένες μέθοδοι αποτίμησης Ο στόχος των αυτοματοποιημένων μεθόδων αποτίμησης είναι η αυτόματη εκτίμηση της αγοραίας αξίας ενός σπιτιού με βάση τις διαθέσιμες πληροφορίες του. Με βάση τον ορισμό της Επιτροπής Διεθνών Προτύπων Αποτίμησης (IVSC), η αγοραία αξία είναι μια αντιπροσώπευση της αξίας σε αντάλλαγμα ή του ποσού που θα απέφερε ένα ακίνητο εάν προσφερόταν προς πώληση στην ανοιχτή αγορά κατά την ημερομηνία αποτίμησης.

Η μαζική εκτίμηση είναι η διαδικασία αποτίμησης μιας ομάδας ιδιοτήτων από μια δεδομένη ημερομηνία και με την χρήση κοινών δεδομένων, τυποποιημένων μεθόδων και στατιστικών δοκιμών. Ο παραπάνω ορισμός προέρχεται από το SMARP (Standard of Mass Appraisal of Real Property). Πριν από τον 21ο αιώνα, τα σχετικά ιδρύματα και οι μελετητές είχαν κάνει γόνιμη εργασία σχετικά με τη θεωρητική κατασκευή και το πρότυπο καθορισμό της μαζικής εκτίμησης ακινήτων. Με την ανάπτυξη της υποβοηθούμενης από υπολογιστή μαζικής αξιολόγησης (CAMA), τόσο τα μοντέλα όσο και τα πρότυπα υιοθετούν σταδιακά

μια αυτοματοποιημένη μεθοδολογία αποτίμησης (AVM) για τη μαζική αξιολόγηση. Γενικά, οι ερευνητές εφαρμόζουν τις ιδέες, τα μοντέλα και τις μεθόδους άλλων τομέων, όπως η στατιστική, η επιστήμη των υπολογιστών ή η γεωγραφική επιστήμη, στο πεδίο της μαζικής εκτίμησης ακινήτων.

Μοντέλο που βασίζεται σε Τεχνητή Νοημοσύνη

Επειδή ο στόχος της μαζικής εκτίμησης είναι ένας μεγάλος αριθμός ακινήτων και τα αποτελέσματα της αποτίμησης πρέπει να εξηγηθούν στο κοινό, οι βασικές ανάγκες είναι η λειτουργία και η απλή κατανόηση. Μόλις συλλέξουμε τα σχετικά δεδομένα του στόχου αξιολόγησης, η άμεση μέθοδος είναι η ανάλυση της σχέσης μεταξύ των σχετικών χαρακτηριστικών (ηλικία δόμησης, επιφάνεια, όροφος, ύψος, τοποθεσία κ.λπ.) και της σχετικής αξίας εκτίμησης του ακινήτου. Μέσω της ποσοτικής ανάλυσης υπολογίζεται η μαθηματική σχέση μεταξύ εξαρτημένης μεταβλητής και ανεξάρτητης μεταβλητής. Η μαζική εκτίμηση της ακίνητης περιουσίας με παρόμοια χαρακτηριστικά θα εκτιμηθεί χρησιμοποιώντας τη γνωστή μαθηματική σχέση. Η ανάλυση πολλαπλής παλινδρόμησης (MRA) είναι μια στατιστική μέθοδος για την πρόβλεψη της αξίας της ακίνητης περιουσίας (εξαρτημένη μεταβλητή) με βάση δύο ή περισσότερα άλλα σχετικά χαρακτηριστικά (ανεξάρτητες μεταβλητές).

Οι εγγενείς αξίες των οικιστικών ακινήτων αποκλίνουν λόγω διαφόρων παραγόντων που απαιτούν εξέταση κατά τη διαδικασία αποτίμησης. Η πρόκληση είναι να αναπτυχθεί ένα έμπειρο σύστημα που να είναι προσαρμόσιμο στην πραγματική ζωή για πραγματικά προβλήματα αξιολόγησης. Το έμπειρο σύστημα δεν θα ≪διδάξει μόνο του≫, αλλά ο εκτιμητής θα χρησιμοποιήσει τα δεδομένα, με τρόπο ασαφούς λογικής, για να αναπτύξει παράγοντες προσαρμογής. Η πρακτική επίλυσης προβλημάτων των ειδικών είναι απλώς η φύση της διαδικασίας αποτίμησης.

Μοντέλα Decision Tree & Random Forest

Το μοντέλο που βασίζεται σε δέντρα έχει καλή απόδοση τόσο στην ταξινόμηση όσο και στην παλινδρόμηση με καλή ακρίβεια, σταθερότητα και

διαλειτουργικότητα. Τα τυπικά μοντέλα περιλαμβάνουν ένα δέντρο αποφάσεων, ένα τυχαίο δάσος και ένα ενισχυμένο δέντρο.

Υπάρχουν δύο συνηθισμένοι τύποι μοντέλων δέντρων αποφάσεων: M5 και MARS (πολυμεταβλητές προσαρμοστικές στροφές παλινδρόμησης). Το M5 είναι ένας αλγόριθμος μοντέλων δέντρων που προβλέπει συνεχείς μεταβλητές για παλινδρόμηση. Στη συνέχεια βελτιστοποιείται σε M5P που συνδυάζει το δέντρο αποφάσεων με γραμμική παλινδρόμηση στους κόμβους. Η κατασκευή δέντρων, το κλάδεμα και η εξομάλυνση είναι τα τρία βασικά βήματα κατά την εφαρμογή της μεθοδολογίας M5P. Ένα άλλο δέντρο αποφάσεων είναι το MARS που είναι μια μη παραμετρική παλινδρόμηση. Τα μοντέλα MARS χωρίζονται σε τρία στάδια: τη διαδικασία προς τα εμπρός, τη διαδικασία κλαδέματος προς τα πίσω και τη διαδικασία επιλογής μοντέλου. Οι Reyes-Bueno et al. (2018) περιγράφουν την κύρια διαφορά μεταξύ M5P και MARS. Στα όρια των χωρισμένων περιοχών, το M5P είναι διακριτό ενώ το MARS είναι συνεχές. Ένα τυχαίο δάσος είναι ένα είδος συνόλου που μαθαίνει να ενσωματώνει πολλά δέντρα απόφασης σε ένα ≪δάσος≫. Το μοντέλο μπορεί να εκτελεστεί αποτελεσματικά σε ένα μεγάλο σύνολο δεδομένων ιδιοτήτων και να αντιμετωπίζει μεταβλητές εισόδου χωρίς διαγραφή. Οι Antipov και Pokryshevskaya (2012) προσπαθούν να χρησιμοποιήσουν το τυχαίο μοντέλο δασών στη μαζική αξιολόγηση για πρώτη φορά και βρίσκουν ότι έχει την καλύτερη απόδοση μεταξύ άλλων μοντέλων. Σε σύγκριση με ένα τυχαίο δάσος, ένα ενισχυμένο μοντέλο δέντρου μπορεί να επιτύχει μεγαλύτερη ακρίβεια και μεγαλύτερη ταχύτητα τρεξίματος. Αυτά τα πλεονεκτήματα απαιτούνται επειγόντως για μια μαζική αξιολόγηση με μεγάλο αριθμό δεδομένων και ένα χρονικό κόμβο αξιολόγησης. Οι McCluskey et al. (2014) εφαρμόζουν το δέντρο ενισχυμένης παλινδρόμησης για τη μαζική αξιολόγηση κατοικιών στη Μαλαισία. Κατόπιν έρευνας διαπίστωσαν ότι το ενισχυμένο δέντρο είναι καλύτερο από το μοντέλο MRA ως προς τον συντελεστή διασποράς και το μέσο απόλυτο ποσοστό σφάλματος.

Ιεραρχικό Μοντέλο

Το παραδοσιακό οικονομετρικό μοντέλο, όπως τα συνηθισμένα ελάχιστα τετράγωνα (OLS), δεν λαμβάνει υπόψη την ιεραρχική δομή των δεδομένων. Η

χρήση του ιεραρχικού μοντέλου μπορεί να ξεπεράσει αυτό το μειονέκτημα. Οι ερευνητές χρησιμοποιούν αυτό το πλαίσιο για την εφαρμογή αποτίμησης ακινήτων, όπως η ιεραρχική Μπεϋζιανή προσέγγιση και η αναλυτική διαδικασία ιεραρχίας. Το ιεραρχικό μοντέλο υπολογίζει επίσης το ποσοστό του σφάλματος διακύμανσης που προκαλείται από κάθε επίπεδο. Δύο τύποι ιεραρχικών μοντέλων έχουν χρησιμοποιηθεί στην αξιολόγηση ακινήτων, το ιεραρχικό γραμμικό μοντέλο (HLM) και το μοντέλο ιεραρχικής τάσης (HTM). Οι Arribas et al. (2016) χρησιμοποιούν το HLM για να ταξινομήσουν τις μεταβλητές σε επίπεδα διαμερισμάτων και γειτονιών. Διαπιστώνουν ότι οι παράμετροι HTM έχουν χαμηλότερη εκτιμώμενη διακύμανση από το OLS. Αναφέρεται ότι το HTM μπορεί να θεωρηθεί ως επέκταση ενός μοντέλου εικονικής μεταβλητής με χρονικά μεταβαλλόμενες σταθερές για τα διαφορετικά συμπλέγματα.

Ανάλυση συστάδων

Η ετερογένεια και η ομοιογένεια των δεδομένων ιδιοκτησίας κατέχουν σημαντική θέση στη μοντελοποίηση μαζικής εκτίμησης. Η ανάλυση συστάδων είναι μια διαδικασία ταξινόμησης δεδομένων σε διαφορετικές κλάσεις ή συστάδες, έτσι ώστε οι στόχοι στο ίδιο σύμπλεγμα να είναι παρόμοιοι, ενώ οι στόχοι σε ένα σύμπλεγμα είναι διαφορετικοί από εκείνους σε άλλα συμπλέγματα. Με βάση τα δείγματα δεδομένων, η ανάλυση συμπλέγματος μπορεί να ταξινομήσει αυτόματα όλη τη βάση δεδομένων. Αυτή η διαδικασία εξόρυξης δεδομένων ή προεπεξεργασίας δεδομένων μπορεί να μεταφέρει την αγορά ακινήτων με ετερογένεια σε μια υποαγορά ακινήτων με ομοιογένεια. Η προσέγγιση συμπλέγματος μπορεί να ταξινομηθεί σε διάφορους τύπους: ιεραρχική ομαδοποίηση, ομαδοποίηση κατάτμησης, ομαδοποίηση βάσει πλέγματος, ομαδοποίηση με βάση την πυκνότητα, ομαδοποίηση βάσει ασαφούς και ομαδοποίηση βάσει μοντέλου. Μετά την ανάλυση συστάδων, είναι απαραίτητο να εξηγηθεί η πρακτική έννοια των διαφορετικών συστάδων. Η χωρίς νόημα ομαδοποίηση θα καθοδηγήσει εκ νέου τη ρύθμιση της ανάλυσης συστάδων.

Ακατέργαστη Θεωρία Συνόλων και Θεωρία Ασαφών Συνόλων

Η αβεβαιότητα είναι ένα αντικειμενικό φαινόμενο στη μαζική αξιολόγηση. Η αβεβαιότητα της εκτίμησης μάζας επηρεάζει τη σταθερότητα του μοντέλου και την ακρίβεια των αποτελεσμάτων. Για την κατάσταση του ανακριβούς συνόλου δεδομένων, που εμφανίζεται στην αναδυόμενη ή αδύναμη αγορά κατοικίας στην πληροφόρηση, η ακατέργαστη θεωρία συνόλων και η θεωρία ασαφών συνόλων παρέχουν έναν διαθέσιμο τρόπο για μαζική εκτίμηση της ακίνητης περιουσίας. Επιπλέον, η εφαρμογή της θεωρίας ακατέργαστων συνόλων (RST) στον τομέα των ακινήτων υπογραμμίζει τις δυνατότητές της για μοντελοποίηση μαζικής εκτίμησης.

Το RST δημιουργεί έναν τρόπο εκτέλεσης του μοντέλου αξιολόγησης ακινήτων χωρίς να λαμβάνονται υπόψη οι σχετικοί δείκτες που επηρεάζουν την αξία της ακίνητης περιουσίας. Τελευταίο αλλά όχι λιγότερο σημαντικό, με την εισαγωγή του ασαφούς συνόλου, η κρίση και η διαδικασία σκέψης των ανθρώπων μπορεί να εκφραστεί άμεσα σε μια σχετικά απλή μαθηματική μορφή, η οποία καθιστά δυνατή την αντιμετώπιση πολύπλοκων συστημάτων με πρακτικό και ανθρώπινο τρόπο σκέψης. Η θεωρία των ασαφών συνόλων μπορεί να λύσει την εγγύτητα μεταξύ διαφορετικών δειγμάτων ιδιοτήτων που πρόκειται να αξιολογηθούν. Χρησιμοποιείται για την αποτελεσματική διόρθωση του βάρους των δεδομένων, ακόμη και αν ο βαθμός εγγύτητας είναι χαμηλός.

Μοντέλο που βασίζεται σε GIS

Το GIS, γνωστό ως σύστημα/ επιστήμη γεωπληροφοριών, εστιάζει σε χωρικά ή γεωγραφικά δεδομένα. Κάθε ακίνητο έχει τις δικές του πληροφορίες χωρικών χαρακτηριστικών. Τα χωρικά χαρακτηριστικά μαζί με τις μη χωρικές πληροφορίες συμβάλλουν στην αξία του ακινήτου. Πολλοί μελετητές έχουν δώσει προσοχή στα χαρακτηριστικά GIS των ακινήτων και έχουν εξετάσει τις επιπτώσεις τους στην αξιολόγηση. Εν τω μεταξύ, ορισμένοι μελετητές όπως οι Bourassa et al. (2007), McCluskey and Borst (2007) έχουν κάνει βελτιώσεις στην ταξινόμηση και τη συλλογή μοντέλων που βασίζονται σε GIS.

Μοντέλο που βασίζεται σε MIX

Σε αυτό το μέρος, ο κύριος σκοπός είναι να εξηγηθεί η έμφαση και η σκέψη μιας εφαρμογής μοντέλου με βάση το μείγμα στη μαζική αξιολόγηση. Οι Guo et al. (2014) ενσωματώνουν ορισμένα στοιχεία από μια προσέγγιση σύγκρισης πωλήσεων και προσέγγιση εισοδήματος στην προσέγγιση κόστους για να βελτιώσουν την ακρίβεια της αποτίμησης της ακίνητης περιουσίας. Τα σχετικά μοντέλα βασίζονται στα παραδοσιακά ή υπάρχοντα μοντέλα, σε συνδυασμό με μεθόδους AI και GIS και στη συνέχεια επισημαίνουν την εφαρμογή και την ανάλυση. Επιπλέον, ορισμένα μοντέλα μπορούν να συνδυαστούν μεταξύ τους σε ένα καλύτερο, όπως η ασαφής ομαδοποίηση, το γεωστατιστικό μοντέλο και η ομαδοποίηση, η ανάλυση πολλαπλών κριτηρίων και οι γενετικοί αλγόριθμοι, το ANN και το GIS, το διάνυσμα υποστήριξης μηχανή και σύστημα υποστήριξης αποφάσεων και ούτω καθεξής. Ο τρίτος τύπος είναι η ανάμειξη με καινοτόμες ιδέες και μοναδικές προοπτικές. Για παράδειγμα, οι Chen et al. (2017) εφαρμόζουν όχι μόνο τις πληροφορίες των παραδοσιακών δεδομένων ακίνητης περιουσίας, αλλά και τις επιπλέον πληροφορίες αγοράς σε πραγματικό χρόνο από την ηλεκτρονική ανάδραση του crowdsourcing, γεγονός που καθιστά το εκτιμώμενο αποτέλεσμα κοντά στην αγορά.

Άλλα μοντέλα

Υπάρχουν πολλά άλλα είδη κλασικών μοντέλων που έχουν εφαρμοστεί στον τομέα της εκτίμησης μάζας, π.χ., γενετικός αλγόριθμος, μηχανή, ανάλυση περιβλήματος δεδομένων και σύμμορφοι προγνωστικοί παράγοντες. Αν και μόνο λίγοι μελετητές έχουν δοκιμάσει αυτά τα μοντέλα, τα αποτελέσματα έχουν καλή τιμή αναφοράς. Ο γενετικός αλγόριθμος (GA) είναι ένα υπολογιστικό μοντέλο που προσομοιώνει τη φυσική επιλογή και τον γενετικό μηχανισμό της δαρβινικής βιολογικής εξέλιξης και είναι μια μέθοδος αναζήτησης της βέλτιστης λύσης προσομοιώνοντας τη φυσική εξέλιξη. Οι Morano et al. (2018) συνδυάζουν την εξελικτική πολυωνυμική παλινδρόμηση με γενετικούς αλγόριθμους για την αναζήτηση αυτών των μοντέλων με ακρίβεια μεγιστοποίησης των δεδομένων και φειδωλότητα των μαθηματικών συναρτήσεων. Οι Ahn et al. (2012) χρησιμοποιούν παλινδρόμηση κορυφογραμμής σε συνδυασμό με έναν γενετικό

αλγόριθμο (GA-Ridge) για να δοκιμάσουν την απόδοση στην κορεατική αγορά ακινήτων.

Η ανάλυση περιβλήματος δεδομένων (DEA) είναι ένα ερευνητικό πεδίο επιχειρησιακής έρευνας, επιστήμης διαχείρισης και μαθηματικών οικονομικών. Μπορεί να χρησιμοποιηθεί για την αξιολόγηση του εύρους αξίας για μονάδες ακινήτων. Η αβεβαιότητα στην αξία της μονάδας που προκύπτει από συναλλαγές στην αγορά εξετάστηκε αντιπροσωπεύοντας ρητά τους οικονομικούς παράγοντες που εμπλέκονται σε μια συναλλαγή, δηλαδή τον αγοραστή και τον πωλητή, των οποίων οι ενέργειες καθορίζουν ένα σύνολο πραγματοποιημένων συναλλαγών. Οι Conformal Predictors (CP) είναι ένας κλασικός αλγόριθμος μηχανικής μάθησης που μπορεί να παρέχει αξιόπιστες προβλέψεις με τη μορφή περιοχών. Για την παλινδρόμηση, ένα διάστημα πρόβλεψης θα σχηματίζεται τυπικά από τις περιοχές. Οι περιοχές είναι αξιόπιστες στο επίπεδο εμπιστοσύνης που ορίζει ο χρήστη.

2.4 Αλγόριθμοι παλινδρόμησης

Η παλινδρόμηση είναι ένα σύνολο στατιστικών διαδικασιών που στοχεύουν στην εκτίμηση των σχέσεων μεταξύ των μεταβλητών. Πιο συγκεκριμένα, η ανάλυση παλινδρόμησης μοντελοποιεί την επίδραση μιας ή περισσότερων αριθμητικών μεταβλητών σε μια μεμονωμένη αριθμητική μεταβλητή. Η μεταβλητή της οποίας η τιμή εκτιμάται ονομάζεται *εξαρτημένη ή ανταποκρίνον* ενώ οι μεταβλητές που χρησιμοποιούνται για την εκτίμηση της εξαρτημένης ονομάζονται *ανεξάρτητες ή επεξηγηματικές*. Ο στόχος της παλινδρόμησης είναι να προβλέψει μια μεταβλητή στόχου (εξαρτώμενη) όσο το δυνατόν ακριβέστερα, επομένως υποθέτει ότι η μεταβλητή στόχος ταιριάζει σε κάποιο γνωστό τύπο συνάρτησης (γραμμική, λογιστική, κ.λπ.) και στη συνέχεια καθορίζει την καλύτερη συνάρτηση αυτού του τύπου που μοντελοποιεί τις δεδομένες μεταβλητές (ανεξάρτητες) (DUNHAM, n.d.). Στη συνέχεια, θα παρουσιάσουμε όλους τους αλγόριθμους μηχανικής μάθησης που πρόκειται να εφαρμόσουμε, ξεκινώντας από τους γραμμικούς αλγόριθμους και στη συνέχεια θα αναφέρουμε κάποια boosting καθώς και ένα bagging.

2.4.1 Γραμμική Παλινδρόμηση

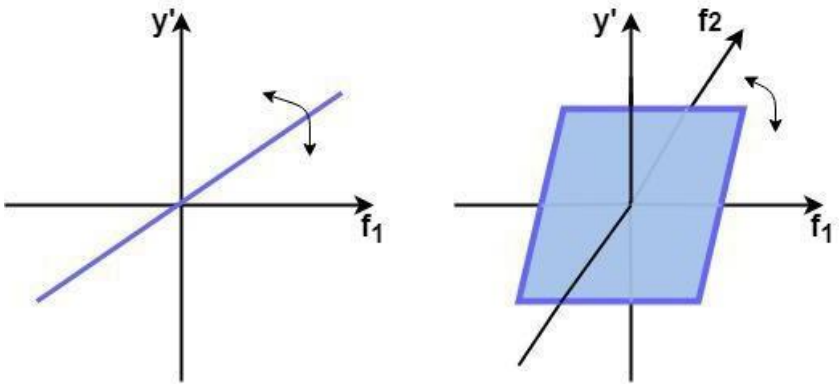
Το απλούστερο μοντέλο παλινδρόμησης ονομάζεται γραμμική παλινδρόμηση και περιγράφει μια μεταβλητή στόχο με έναν γραμμικό συνδυασμό χαρακτηριστικών που είναι οι ανεξάρτητες μεταβλητές (Beyeler, 2017). Ας πούμε ότι έχουμε δύο χαρακτηριστικά, f1 και f2, τότε ο στόχος της γραμμικής παλινδρόμησης είναι να βρούμε δύο συντελεστές βάρους, w1 και w2, προκειμένου να προβλέψουμε τη μεταβλητή στόχο y και η εξίσωση μπορεί να γραφτεί ως εξής:

$$y' = w_1 f_1 + w_2 f_2$$

Εδώ, y' είναι η πρόβλεψη της τιμής y. Αν έχουμε μόνο ένα χαρακτηριστικό, τότε η παλινδρόμηση ονομάζεται απλή γραμμική παλινδρόμηση. Στην περίπτωση που έχουμε περισσότερα από δύο χαρακτηριστικά, ας πούμε n, τότε η προηγούμενη εξίσωση γράφεται ως:

$$y' = w_1 f_1 + w_2 f_2 + \cdots + w_n f_n = \sum_{i=1}^{n} wifi$$

Το παρακάτω σχήμα απεικονίζει τις εξισώσεις με ένα χαρακτηριστικό και δύο χαρακτηριστικά αντίστοιχα. Στην πρώτη περίπτωση η εξίσωση είναι μια ευθεία γραμμή ενώ στη δεύτερη είναι ένα επίπεδο. Αν είχαμε n χαρακτηριστικά, τότε η εξίσωση θα ήταν υπερεπίπεδο.



***Εικόνα 1**: Απεικονίζεται η πρόβλεψη στοχευμένων τιμών σε δύο και τρείς διαστάσεις χρησιμοποιώντας την γραμμική παλινδρόμηση*

Είναι συνηθισμένο να προσθέτουμε έναν ακόμη συντελεστή βάρους που δεν εξαρτάται από κανένα χαρακτηριστικό, προκειμένου να λειτουργήσει ως όρος μεροληψίας. Έτσι, η τελική εξίσωση, για την πρόβλεψη της μεταβλητής στόχου y, με την προκατάληψη w_0, μπορεί να περιγραφεί ως εξής (Beyeler, 2017):

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i f_i$$

Ο τελικός στόχος της γραμμικής παλινδρόμησης είναι να βρεθεί ένα σύνολο συντελεστών βάρους, προκειμένου να προβλεφθεί η μεταβλητή στόχος όσο το δυνατόν ακριβέστερα. Η απόδοση του μοντέλου παλινδρόμησης μετριέται με μια συνάρτηση βαθμολόγησης που παίρνει τη μορφή συνάρτησης κόστους ή ζημίας, σε προβλήματα παλινδρόμησης. Υπάρχουν πολλές συναρτήσεις απώλειας που μπορούμε να χρησιμοποιήσουμε, αλλά η πιο συχνά χρησιμοποιούμενη είναι το μέσο τετράγωνο σφάλμα. Η συνάρτηση απώλειας είναι:

$$L = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

Τέλος, ο στόχος είναι να βρεθεί το σύνολο των βαρών που ελαχιστοποιεί την παραπάνω συνάρτηση.

2.4.2 Πολλαπλή Γραμμική Παλινδρόμηση

Το μοντέλο της απλής γραμμικής παλινδρόμησης $y = b_0 + b_1x + u$ και υπολογίσθηκαν οι παράμετροι και μέσω των εκτιμητριών των ελαχίστων τετραγώνων. Σε αυτήν την ενότητα θα επεκτείνουμε το παραπάνω για περισσότερες εκ της μία εξόδου, και επίσης περισσότερες εισόδους. Έστω τώρα ότι το μοντέλο δίνεται από την σχέση: $Y_i = X_i\beta + \varepsilon_\iota$, όπου το Y_0 αποτελεί την μεταβλητή που επιθυμούμε να προβλέψουμε ως έξοδο και Y_i είναι η τιμή της μεταβλητής στην i παρατήρηση. Το Y_0 καλείται εξαρτημένη ή μεταβλητή απόκρισης και αποτελεί την έξοδο του μοντέλου μας. Υποθέτουμε ότι το X_0 είναι ένα $1 \times n$ διάνυσμα, που αποτελείται από τις παρατηρούμενες, τυχαίες μεταβλητές εισόδου και X_i αποτελεί το διάνυσμα των εισόδων κατά την i παρατήρηση. Ακόμα, το β είναι ένα διάνυσμα παραμέτρων μεγέθους $n \times 1$, οι οποίες αρχικά είναι άγνωστες και στη συνέχεια εκτιμούνται με την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων, με βάση τα δεδομένα. Κάθε μία από τις παραμέτρους εκφράζει την μεταβολή του Y, όταν μεταβάλλεται μόνο η συγκεκριμένη μεταβλητή X, στην οποία αναφέρεται το αντίστοιχο β.

Οι όροι ε_ι είναι ανεξάρτητοι και όμοια κατανεμημένοι, με μηδενική μέση τιμή και διακύμανση σ^2, δηλαδή $\varepsilon_\iota \sim N(0, \sigma^2)$. Για παράδειγμα για την πρώτη παρατήρηση η εξίσωση της πολλαπλής γραμμικής παλινδρόμησης γράφεται:

$$Y_1 = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon_\iota$$

Η εξίσωση της γραμμικής παλινδρόμησης ορίζει ευθεία γραμμή όταν σε αυτήν υπάρχει μόνο μία ανεξάρτητη μεταβλητή X, επίπεδο όταν συμμετέχουν δύο μεταβλητές X και υπερεπίπεδο όταν έχουμε περισσότερες εκ των δύο μεταβλητές

X. Με βάση την μέθοδο των ελαχίστων τετραγώνων η οποία θα εφαρμοστεί για την εκτίμηση των παραμέτρων β, ελαχιστοποιείται το σφάλμα του αθροίσματος των σφαλμάτων, τα οποία αφορούν τις αποστάσεις των τιμών Y που υπολογίσθηκαν από το υπερεπίπεδο προσαρμογής. Η πολλαπλή γραμμική παλινδρόμηση υποθέτει γραμμική σχέση μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών, όχι μεγάλη συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών, σταθερή διακύμανση των υπολειμμάτων, όπως αναφέρθηκε και παραπάνω, ανεξαρτησία των παρατηρήσεων, δηλαδή ανεξαρτησία των τιμών των σφαλμάτων και κανονική κατανομή των σφαλμάτων.

2.4.3 Παλινδρόμηση κορυφογραμμής - Ridge Regression

Ένα κοινό πρόβλημα στη μηχανική μάθηση είναι η υπερπροσαρμογή και ακόμη και η γραμμική παλινδρόμηση μπορεί να επηρεαστεί από αυτό το φαινόμενο. Αυτό που ακριβώς κάνει η παλινδρόμηση κορυφογραμμής είναι να χρησιμοποιεί την κανονικοποίηση για να ελέγξει την υπερπροσαρμογή εφαρμόζοντας μια ποινή στους συντελεστές. Αυτή η ποινή προστίθεται στη συνήθη συνάρτηση ελάχιστου τετραγώνου απώλειας και η συνάρτηση κόστους που προκύπτει από την παλινδρόμηση κορυφογραμμής είναι η ακόλουθη εξίσωση:

$$L = \|y - X\bar{w}\|_2^2 + a\|\bar{w}\|_2^2$$

όπου ο όρος $\bar{w}$ αντιπροσωπεύει το διάνυσμα βάρους και το X είναι ένας πίνακας που περιέχει όλα τα δείγματα ως στήλες. Αυτός ο πρόσθετος όρος, μέσω του συντελεστή a, αναγκάζει τη συνάρτηση απώλειας να μην επιτρέπει μια άπειρη αύξηση του διανύσματος βάρους (Bonaccorso, 2017b). Έτσι, η παλινδρόμηση κορυφογραμμής προσπαθεί να ελαχιστοποιήσει το μέγεθος των συντελεστών.

2.4.4 Lasso Παλινδρόμηση - Lasso Regression

Στην παλινδρόμηση λάσο προστίθεται και μια ποινή, αλλά αυτή τη φορά εφαρμόζεται στις απόλυτες τιμές των συντελεστών και βασίζεται στον κανόνα L1 του διανύσματος βάρους, ενώ στην παλινδρόμηση κορυφογραμμής βασίζεται στον κανόνα L2 του διανύσματος βάρους (Beyeler, 2017). Σε αυτήν την περίπτωση η συνάρτηση απώλειας είναι:

$$L = \|y - X\bar{w}\|_2^2 + a\|\bar{w}\|1$$

Η μέθοδος Lasso ξεπερνά το μειονέκτημα της παλινδρόμησης της κορυφογραμμής όχι μόνο τιμωρώντας τις υψηλές τιμές των συντελεστών βάρους, αλλά ουσιαστικά μηδενίζοντας τις αν δεν είναι σχετικές. Έτσι, μπορεί να καταλήξουμε με λιγότερα χαρακτηριστικά που περιλαμβάνονται στο μοντέλο από ό,τι ξεκινήσαμε, κάτι που είναι τεράστιο πλεονέκτημα.

2.5 Μηχανική Μάθηση

Εισαγωγή στην Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένα πεδίο της Επιστήμης Υπολογιστών στο οποίο μελετώνται αλγόριθμοι, οι οποίοι επιτρέπουν σε ένα υπολογιστικό σύστημα να πραγματοποιεί προβλέψεις ή να λαμβάνει αποφάσεις χωρίς να έχει προγραμματιστεί εξαρχής η ακριβής συλλογιστική πορεία που πρέπει να ακολουθήσει για να τις λάβει. Συγκεκριμένα, ο αλγόριθμος μηχανικής μάθησης δημιουργεί ένα μαθηματικό μοντέλο που βασίζεται σε δεδομένα – δείγματα, ή αλλιώς δεδομένα εκπαίδευσης με τα οποία τροφοδοτείται. Μοντέλα μηχανικής μάθησης, χρησιμοποιούνται σε διάφορους τομείς, όπως η πρόβλεψη των τιμών μίας μετοχής, η αυτόματη αναγνώριση ηλεκτρονικών μηνυμάτων απάτης, η

επεξεργασία φυσικής γλώσσας κ.α. Ένα από τα πεδία αυτά είναι και η πρόβλεψη τιμών ακινήτων.

Η μηχανική μάθηση επιτρέπει την ανάλυση τεράστιων ποσοτήτων δεδομένων. Αν και γενικά παρέχει ταχύτερα πιο ακριβή αποτελέσματα για τον εντοπισμό κερδοφόρων ευκαιριών ή κινδύνων, μπορεί επίσης να απαιτήσει επιπλέον χρόνο και πόρους για να εκπαιδευτεί σωστά. Ο συνδυασμός της μηχανικής μάθησης με την τεχνητή νοημοσύνη και τις γνωστικές τεχνολογίες μπορεί να την κάνει ακόμη πιο αποτελεσματική στην επεξεργασία μεγάλου όγκου πληροφοριών.

2.5.1 Κατηγορίες Μηχανικής Μάθησης

Οι αλγόριθμοι μηχανικής μάθησης κατηγοριοποιούνται συχνά ως με επίβλεψη ή χωρίς επίβλεψη.

Οι με επίβλεψη (supervised) αλγόριθμοι μηχανικής μάθησης μπορούν να εφαρμόσουν ό,τι έχουν μάθει στο παρελθόν σε νέα δεδομένα χρησιμοποιώντας επισημασμένα παραδείγματα για την πρόβλεψη μελλοντικών γεγονότων. Ξεκινώντας από την ανάλυση ενός γνωστού συνόλου δεδομένων, ο αλγόριθμος εκμάθησης παράγει προβλέψεις σχετικά με τις τιμές εξόδου. Το σύστημα είναι σε θέση να παρέχει προβλέψεις για οποιαδήποτε νέα είσοδο μετά από επαρκή εκπαίδευση. Ο αλγόριθμος εκμάθησης μπορεί επίσης να συγκρίνει την έξοδο του με τη σωστή έξοδο και να βρει σφάλματα προκειμένου να τροποποιήσει ανάλογα το εσωτερικό του μοντέλο.

Αντίθετα, οι μη επιτηρούμενοι (unsupervised – χωρίς επίβλεψη) αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται όταν οι πληροφορίες που χρησιμοποιούνται για την εκπαίδευση δεν ταξινομούνται ούτε επισημαίνονται με κάποιο τρόπο από τον άνθρωπο. Η μη επιτηρούμενη μάθηση μελετά πώς τα συστήματα μπορούν να συμπεράνουν μια συνάρτηση για να περιγράψουν μια κρυφή δομή από δεδομένα χωρίς ετικέτα. Το σύστημα δεν καταλαβαίνει τη σωστή έξοδο, αλλά διερευνά τα δεδομένα και μπορεί να εξαγάγει συμπεράσματα από σύνολα δεδομένων για να περιγράψει κρυφές δομές από δεδομένα χωρίς ετικέτα.

Οι ημι-επιβλεπόμενοι (semi-supervised) αλγόριθμοι μηχανικής μάθησης εμπίπτουν κάπου μεταξύ της εποπτευόμενης και της μη εποπτευόμενης μάθησης, δεδομένου ότι χρησιμοποιούν τόσο δεδομένα με ετικέτα όσο και χωρίς για την εκπαίδευση – συνήθως μια μικρή ποσότητα δεδομένων με ετικέτες και μια μεγάλη ποσότητα δεδομένων χωρίς. Τα συστήματα που χρησιμοποιούν αυτήν τη μέθοδο είναι σε θέση να βελτιώσουν σημαντικά την ακρίβεια της μάθησης. Συνήθως, η ημι-εποπτευόμενη μάθηση επιλέγεται όταν τα αποκτηθέντα επισημασμένα δεδομένα απαιτούν εξιδεικευμένους και σχετικούς πόρους για να οδηγήσουν σε ένα καλό μοντέλο.

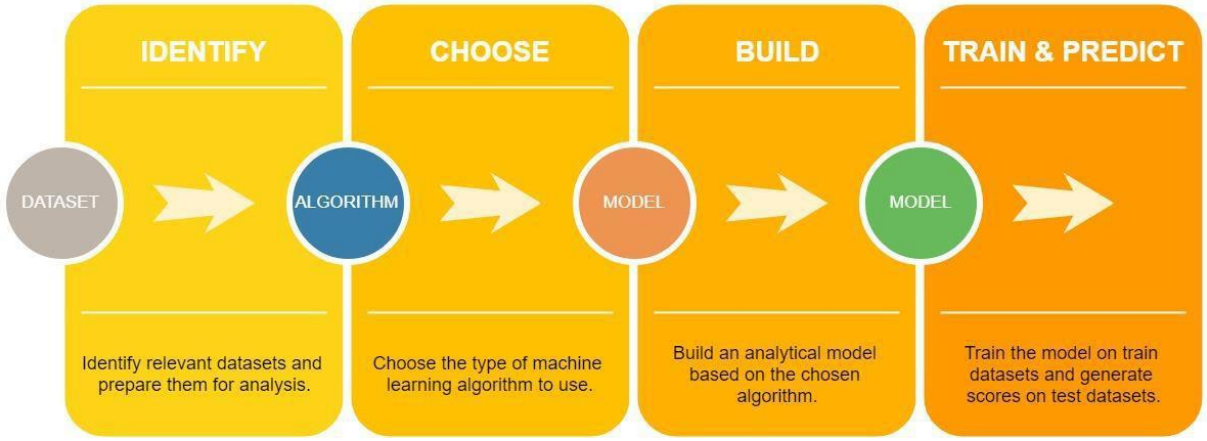
Οι αλγόριθμοι ενισχυμένης μάθησης (reinforshment learning) είναι μια μέθοδος μάθησης που αλληλοεπιδρά με το περιβάλλον της παράγοντας ενέργειες και ανακαλύπτει σφάλματα ή ανταμοιβές. Η αναζήτηση δοκιμών και σφαλμάτων και η καθυστερημένη ανταμοιβή είναι τα πιο σχετικά χαρακτηριστικά της μάθησης ενίσχυσης. Αυτή η μέθοδος επιτρέπει σε μηχανές κι πράκτορες λογισμικού να προσδιορίζουν αυτόματα την ιδανική συμπεριφορά σε ένα συγκεκριμένο πλαίσιο, προκειμένου να μεγιστοποιήσουν την απόδοση της. Απαιτείται απλή ανατροφοδότηση ανταμοιβής για να μάθει ο πράκτορας ποια ενέργεια είναι καλύτερη. Αυτό είναι γνωστό ως σήμα ενίσχυσης.

Στα πλαίσια των επιβλεπόμενων αλγορίθμων που αφορούν την παρούσα εργασία, ένα άλλο κριτήριο κατηγοριοποίησης των αλγορίθμων μηχανικής μάθησης είναι το είδος προβλήματος που αυτοί επιλύουν. Δύο μεγάλες ομάδες προβλημάτων είναι τα προβλήματα ταξινόμησης (classification) και τα προβλήματα παλινδρόμησης (regression).

Οι αλγόριθμοι ταξινόμησης επιχειρούν να εκτιμήσουν τη διακριτή ομάδα στην οποία ανήκει ένας στόχος ψ με βάση τις μεταβλητές εισόδου χ. Για παράδειγμα, όταν παρέχεται ένα σύνολο δεδομένων για ακίνητα, ένας αλγόριθμος ταξινόμησης μπορεί να προσπαθήσει να προβλέψει εάν οι τιμές για τα σπίτια πωλούν περισσότερο ή λιγότερο από τη συνιστάμενη λιανική τιμή.

Στη μηχανική μάθηση, οι αλγόριθμοι παλινδρόμησης προσπαθούν να εκτιμήσουν τη συνάρτηση χαρτογράφησης (mapping function) φ που οδηγεί από τις μεταβλητές εισόδου ξ σε αριθμητικές ή συνεχείς μεταβλητές εξόδου ψ. Σε αυτήν την περίπτωση, το ψ είναι μια πραγματική τιμή, η οποία μπορεί να είναι

ακέραιος ή δεκαδική τιμή. Επομένως, τα προβλήματα πρόβλεψης παλινδρόμησης είναι συνήθως ποσότητες ή μεγέθη. Για παράδειγμα, όταν παρέχεται ένα σύνολο δεδομένων για ακίνητα και ζητείται να προβλεφθούν οι τιμές τους, αυτό είναι μια εργασία παλινδρόμησης επειδή η τιμή θα είναι μια συνεχής έξοδος. Παραδείγματα των κοινών αλγορίθμων παλινδρόμησης περιλαμβάνουν την γραμμική παλινδρόμηση και τα δέντρα παλινδρόμησης. Ορισμένοι αλγόριθμοι, όπως η λογιστική παλινδρόμηση, έχουν τη λέξη παλινδρόμηση στα ονόματά τους, αλλά δεν είναι αλγόριθμοι παλινδρόμησης.



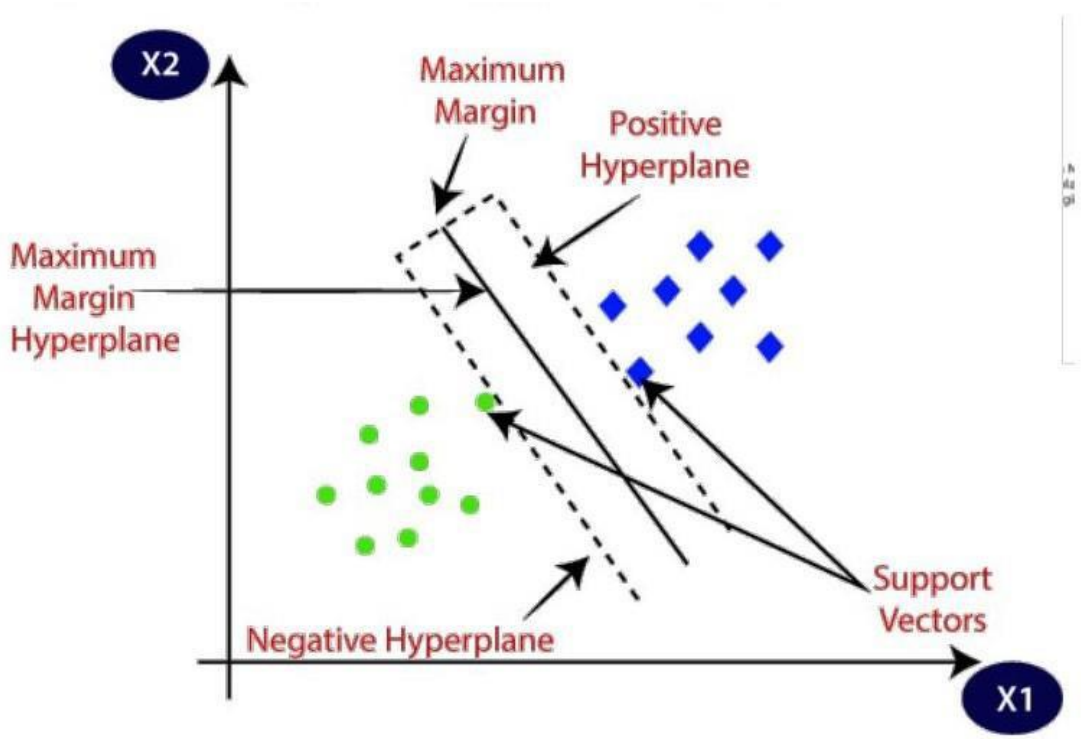
***Εικόνα 2**: Διαδικασία Μηχανικής Μάθησης*

2.5.2 Support Vector Machines (SVM)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι ένας αλγόριθμος εποπτευόμενης μηχανικής μάθησης (supervised learning) που μπορεί να χρησιμοποιηθεί σε προβλήματα ταξινόμησης (classification) ή παλινδρόμησης (regression). Ωστόσο, χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης. Στον αλγόριθμο αυτό απεικονίζουμε κάθε στοιχείο δεδομένων ως ένα σημείο στον n-διάστατο χώρο (όπου n είναι ο αριθμός των χαρακτηριστικών που περιέχουν τα δεδομένα) με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Κύρια λειτουργία του είναι η εύρεση ενός υπερεπιπέδου που

διαχωρίζει βέλτιστα τις τάξεις (2 ή περισσότερες) των δεδομένων (στην ταξινόμηση) (Analytics Vidhya, Ray S., 2017).

Η παλινδρόμηση διανυσμάτων υποστήριξης (Support Vectors Regression) χρησιμοποιεί τις ίδιες αρχές με το μοντέλο ταξινόμησης, με μερικές μόνο μικρές διαφορές. Επειδή η έξοδος είναι ένας πραγματικός αριθμός καθίσταται πολύ δύσκολη η πρόβλεψη της συγκεκριμένης πληροφορίας (διαχωρισμός του ύπερ - επιπέδου), η οποία έχει άπειρες δυνατότητες. Στην περίπτωση της παλινδρόμησης, τίθεται ένα περιθώριο ανοχής (epsilon) κατά προσέγγιση, το οποίο πρέπει να δίνεται (ή να ζητιέται) από το πρόβλημα (Ritchie N., 2021). Γενικά, αλγόριθμος είναι αρκετά πολύπλοκος στα προβλήματα παλινδρόμησης και δεν προτιμάται.



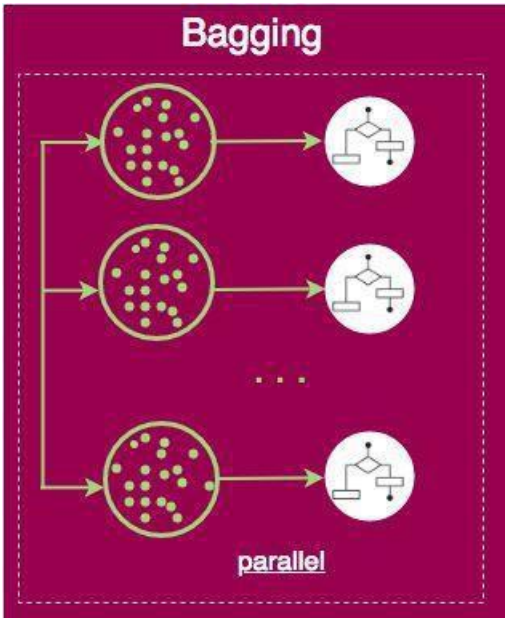
***Εικόνα 3**: Support Vector Machine Αλγόριθμος*

Σημείωση:

Στα μοντέλα που κατασκευάζουμε για τον σκοπό της Διπλωματικής μας χρησιμοποιούμε τους αλγορίθμους Multiple Linear Regression, Decision Tree Regressor και Random Forest Regressor, οι οποίοι είναι παραδοσιακοί και διαδεδομένοι για μοντελοποίηση και προβλέψεις σε προβλήματα παλινδρόμησης, ενώ ταυτόχρονα είναι αξιόπιστοι και αποδοτικοί.

2.5.3 Bagging

Το Bagging είναι μια τεχνική συνόλου στην οποία χτίζουμε πολλούς ανεξάρτητους εκπαιδευόμενους (μοντέλα/προγνωστικά) και τους συνδυάζουμε χρησιμοποιώντας ορισμένες τεχνικές υπολογισμού του μέσου όρου μοντέλων. Λαμβάνουμε τυχαίο υπόδειγμα δεδομένων για κάθε μοντέλο, έτσι ώστε όλα τα μοντέλα να διαφέρουν ελάχιστα μεταξύ τους. Κάθε παρατήρηση επιλέγεται με αντικατάσταση που θα χρησιμοποιηθεί ως είσοδος για κάθε μοντέλο. Έτσι, κάθε μοντέλο θα έχει διαφορετικές παρατηρήσεις με βάση τη διαδικασία του bootstrap. Το Bagging είναι μια τεχνική που εφαρμόζεται σε δέντρα απόφασης προκειμένου να ελαχιστοποιηθεί το σφάλμα διακύμανσης και ταυτόχρονα να μην αυξηθεί η συνιστώσα σφάλματος λόγω μεροληψίας και πετυχαίνει επειδή αυτή η τεχνική απαιτεί πολλούς ασυσχέτιστους μαθητές για να φτιάξουν ένα τελικό μοντέλο (Dangeti, 2017). Παράδειγμα συνόλου bagging είναι το μοντέλο Random Forest το οποίο θα αναλυθεί στη συνέχεια. Το στάδιο εκπαίδευσης είναι παράλληλο για το bagging και κάθε μοντέλο κατασκευάζεται ανεξάρτητα, και αυτό φαίνεται στην εικόνα.



Εικόνα 4: *Ο Διαδοχικός Τρόπος που η Ενίσχυση χτίζει τον νέο Μαθητή*

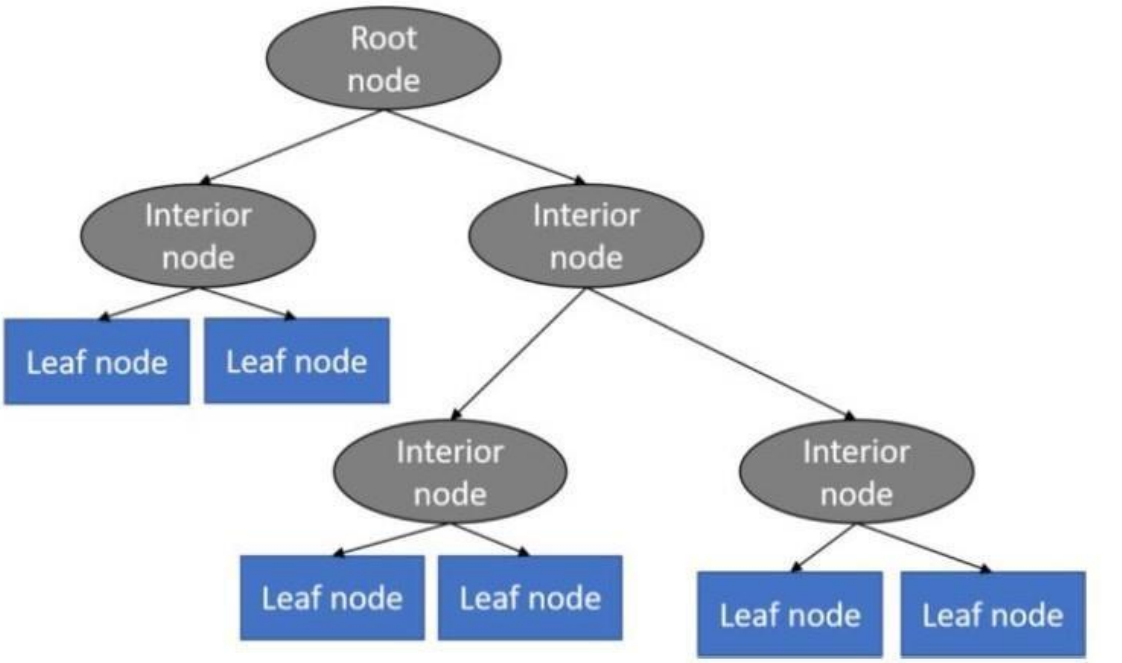
2.5.4 Δέντρα Απόφασης - Decision Trees

Τα δέντρα απόφασης χρησιμοποιούνται τόσο σε προβλήματα ταξινόμησης (classification machine learning problems), όσο και σε προβλήματα παλινδρόμησης (regression machine learning problems). Στα προβλήματα παλινδρόμησης ο αλγόριθμος των δέντρων απόφασης έχει ως εξής:

Βασίζεται στην δομή ενός δέντρου που ταξινομεί, δηλαδή που απαντάει διαδοχικά στην ερώτηση True or False. Αποτελείται από τρία είδη κόμβων: τον αρχικό κόμβο – ρίζα (root node), τους εσωτερικούς κόμβους (interior nodes) και τους τελικούς κόμβους – κόμβους φύλλα (leaf nodes). Ο ριζικός κόμβος αποτελεί τον αρχικό κόμβο και αντιπροσωπεύει το αρχικό ολόκληρο δείγμα με αποτέλεσμα να χωρίζεται διαδοχικά σε περισσότερους κόμβους. Οι εσωτερικοί κόμβοι αντιπροσωπεύουν συγκεκριμένα χαρακτηριστικά – μεταβλητές του συνόλου δεδομένων, ενώ τα κλαδιά που δημιουργούνται από τις συνδέσεις μεταξύ των διάφορων διαδοχικών κόμβων αντιπροσωπεύουν τις επιλογές ή κανόνες απόφασης. Οι τελικοί κόμβοι (φύλλα) απαντούν στο ερώτημα για το ποια είναι η τελική απόφαση ή με άλλα λόγια, την τελική απόφαση του μοντέλου. Η τελική πρόβλεψη του μοντέλου, (δηλαδή η τελική απόφαση του) αποτελεί τον μέσο όρο της τιμής της εξαρτώμενης μεταβλητής στο εκάστοτε φύλλο που παρουσιάζεται.

Η παραπάνω διαδικασία επαναλαμβάνεται πολλαπλά με σκοπό ο αλγόριθμος να εξοικειωθεί με τα δεδομένα και να προβλέψει μια κατάλληλη τιμή για κάθε παρατήρηση που θέλουμε (Drakos G., 2019).

Συνοπτικά, ο αλγόριθμος του δέντρου απόφασης παλινδρόμησης παρουσιάζεται στην παρακάτω εικόνα:



***Εικόνα 5**: Δέντρο Απόφασης*

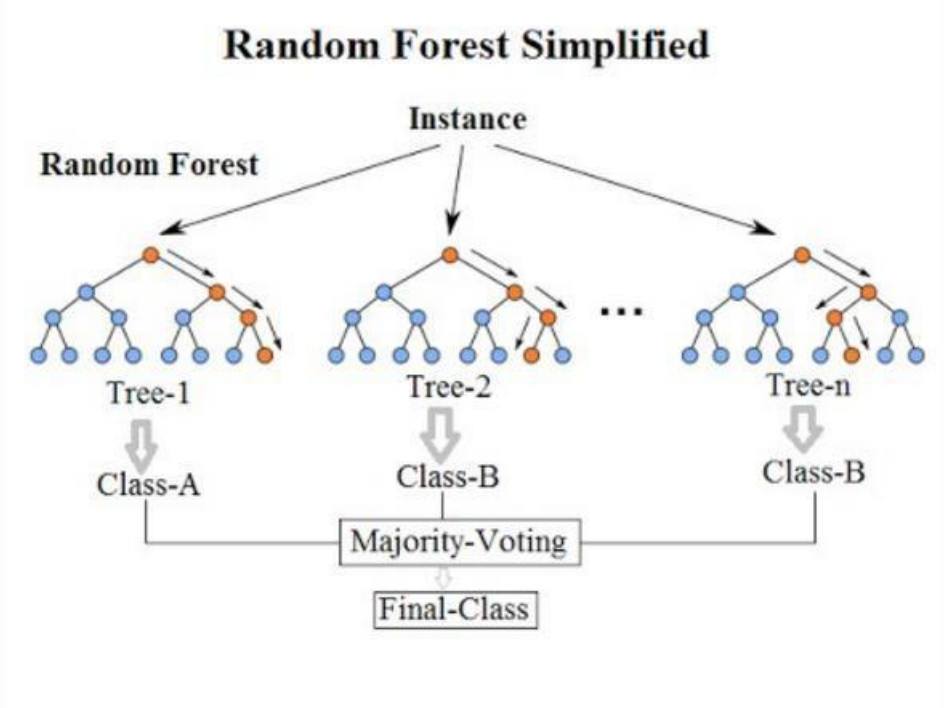
2.5.5 Τυχαία Δάση – Random Forests

Όπως και τα δέντρα απόφασης έτσι και ο αλγόριθμος Random Forest είναι εξαιρετικά διαδεδομένος τόσο για προβλήματα ταξινόμησης μηχανικής μάθησης, όσο και για προβλήματα παλινδρόμησης. Αποτελεί, όπως είναι γνωστό, μια εξέλιξη και επέκταση του αλγόριθμου Decision Tree. Και αυτό γιατί πρακτικά και θεωρητικά αποτελείται και μοντελοποιείται από πολλά δέντρα απόφασης μαζί (στη σειρά).

Ο Random Forest παρουσιάζει δύο βασικές ιδιότητες που το χαρακτηρίζουν: το Aggregation και το Bootstrap.

Το Aggregation έχει να κάνει με το γεγονός πως ο αλγόριθμος αυτός εφόσον συνδυάζει και χρησιμοποιεί πολλά δέντρα απόφασης μαζί, τότε η διακύμανση του αντίστοιχου μοντέλου είναι χαμηλή σε αντίθεση γενικά με ένα δέντρο απόφασης μόνο του, καθώς όλα τα δέντρα απόφασης του μοντέλου εκπαιδεύονται ιδανικά στα δείγματα που τους αντιστοιχούν και η τελική πρόβλεψη - απόφαση του μοντέλου είναι πολλαπλή. Ως εκ τούτου, η τελική πρόβλεψη του αλγορίθμου θα είναι ο μέσος όρος όλων των επιμέρους προβλέψεων - αποφάσεων.

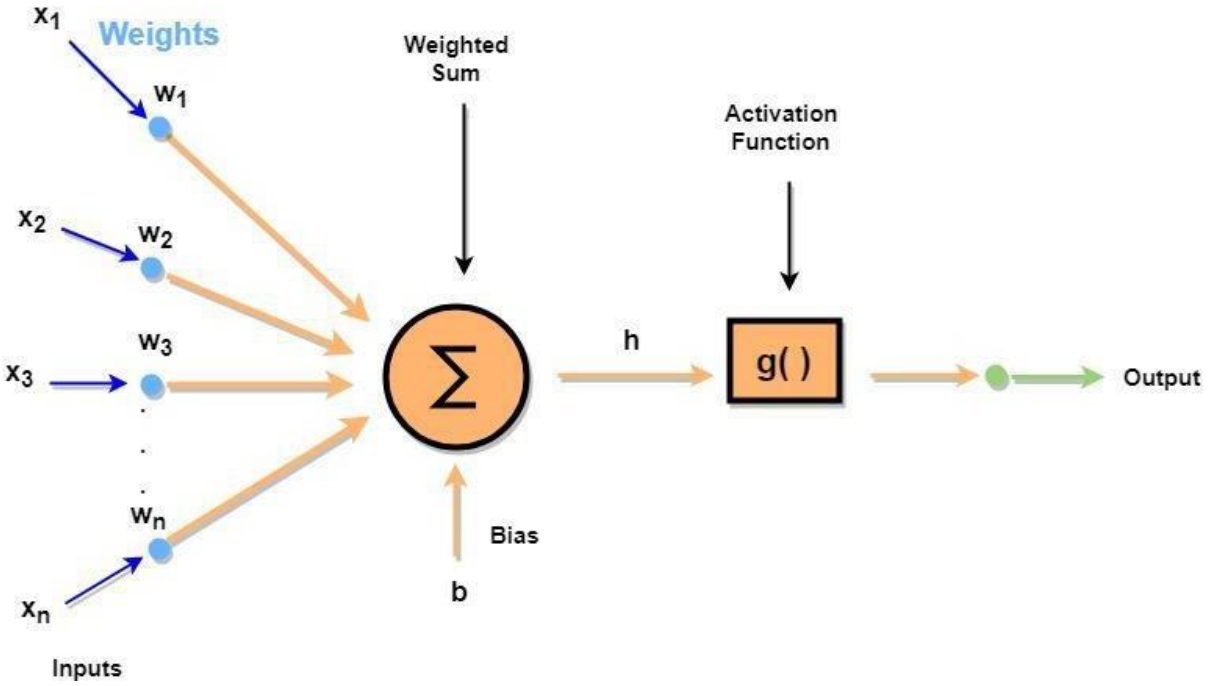
Το Bootstrap έχει να κάνει με το γεγονός πως ο αλγόριθμος πραγματοποιεί τυχαία δειγματοληψία για την σειρά των δέντρων απόφασης, αλλά και για τα χαρακτηριστικά – μεταβλητές του συνόλου δεδομένων με αποτέλεσμα να σχηματίζονται νέα δείγματα συνόλων δεδομένων για κάθε πιθανό μοντέλο (Geek for Geeks, Dutta A., 2022).



Εικόνα 6: *Τυχαίο Δάσος*

2.5.6 Νευρωνικά Δίκτυα

Ένα τεχνητό νευρωνικό δίκτυο (ANN) είναι ένα υπολογιστικό μοντέλο που μιμείται ένα βιολογικό νευρωνικό δίκτυο και βασίζεται σε μια συλλογή συνδεδεμένων μονάδων που ονομάζονται τεχνητοί νευρώνες. Κάθε σύνδεση μεταδίδει ένα σήμα από έναν τεχνητό νευρώνα στον άλλο. Ο τεχνητός νευρώνας που λαμβάνει το σήμα μπορεί να το επεξεργαστεί και στη συνέχεια να το μεταδώσει σε έναν άλλο νευρώνα που είναι συνδεδεμένος με αυτόν. Το παρακάτω σχήμα απεικονίζει τη μορφή ενός τεχνητού νευρώνα.



***Εικόνα 7**: Δομή ενός τεχνητού νευρώνα*

Από το σχήμα η δομή ενός τεχνητού νευρώνα, μπορούμε να δούμε ότι ένας τεχνητός νευρώνας αποτελείται από εισόδους xi. Κάθε είσοδος σταθμίζεται και η έξοδος είναι ένας πραγματικός αριθμός w που ονομάζεται βάρος. Στη συνέχεια, υπάρχει μια συνάρτηση που κάνει το σταθμισμένο άθροισμα όλων των εισόδων και έχει την ακόλουθη μορφή:

$$h = \sum_{i=1}^{n} x_i w_i$$

Ο τεχνητός νευρώνας επιστρέφει μια έξοδο μέσω της συνάρτησης ενεργοποίησης g, εάν το σταθμισμένο άθροισμα υπερβεί ένα κατώφλι b που ονομάζεται προκατάληψη:

$$g\left(\sum_{i=1}^{n} x_i w_i\right) - b > 0$$

Μερικές από τις πιο γνωστές λειτουργίες ενεργοποίησης που χρησιμοποιούνται σε τεχνητά νευρωνικά δίκτυα δίνονται παρακάτω:

- Λειτουργία μεταφοράς βημάτων

$$g(h) = \begin{cases} 1, if\ h \geq b \\ 0, if\ h < b \end{cases}$$

Η έξοδος της συνάρτησης ενεργοποίησης g θα είναι μηδέν, εάν ο νευρώνας είναι απενεργοποιημένος και ένα, εάν είναι ενεργοποιημένος.

- Γραμμική συνάρτηση

$$g(h) = h$$

Η συνάρτηση γραμμικής μεταφοράς επιστρέφει ως έξοδο του νευρώνα το σταθμισμένο άθροισμα h.

- Σιγμοειδής συνάρτηση

$$g(h) = \text{sigma(h)} = \frac{1}{1 - e^{-h}}$$

Μπορεί να θεωρηθεί ως μια οριοθετημένη συνάρτηση, αυστηρά αύξουσα και θετική που συνθλίβει τις τιμές μεταξύ 0 και 1 (Kamath & Choppella, 2017).

- Υπερβολική συνάρτηση εφαπτομένης (“tanh”)

$$g(h) = \tanh(\text{h}) = \frac{e^{h} - e^{-h}}{e^{h} + e^{-h}}$$

Η πιο χρησιμοποιούμενη συνάρτηση ενεργοποίησης και μπορεί να θεωρηθεί ως μια οριοθετημένη, αυστηρά αυξανόμενη αλλά ως θετική ή αρνητική συνάρτηση που συνθλίβει τις τιμές μεταξύ -1 και 1 (Kamath & Choppella, 2017).

Αρχιτεκτονική ANN

Ο τρόπος με τον οποίο συνδέονται οι νευρώνες σε ένα τεχνητό νευρωνικό δίκτυο ονομάζεται αρχιτεκτονική του δικτύου. Ένα ANN αποτελείται από πολλούς νευρώνες που είναι ομαδοποιημένοι σε πολλά στρώματα. Συνολικά, όλοι οι νευρώνες που βρίσκονται στο ίδιο στρώμα τείνουν να έχουν παρόμοια συμπεριφορά. Κάθε ANN έχει δύο βασικά επίπεδα, το επίπεδο εισόδου που αποτελείται από όλες τις εισόδους του δικτύου και το επίπεδο εξόδου που παρέχει την τελική έξοδο(ους). Ένα ANN μπορεί να αποτελείται από περισσότερα από δύο επίπεδα και στην περίπτωση αυτή περιλαμβάνει κρυφά στρώματα. Εδώ, είναι σημαντικό να αναφέρουμε, ότι όσο περισσότερα κρυφά επίπεδα έχουμε, τόσο καλύτερη λύση μπορεί να λάβουμε, αλλά χρειάζεται περισσότερος χρόνος εκπαίδευσης.

Η απόφαση για τον αριθμό των νευρώνων στα κρυφά στρώματα είναι επίσης ένα πολύ σημαντικό μέρος της οικοδόμησης της συνολικής αρχιτεκτονικής νευρωνικών δικτύων μας. Αν και τα στρώματα δεν αλληλοεπιδρούν άμεσα με το εξωτερικό περιβάλλον, έχουν μεγάλη επιρροή στην τελική έξοδο. Τόσο ο αριθμός

των κρυφών επιπέδων όσο και ο αριθμός των νευρώνων σε κάθε κρυφό στρώμα θα πρέπει να ληφθούν προσεκτικά υπόψη.

Από τη μία πλευρά, η χρήση πολύ λίγων νευρώνων στα κρυφά στρώματα θα έχει ως αποτέλεσμα κάτι που ονομάζεται underfitting. Από την άλλη πλευρά, η χρήση πάρα πολλών νευρώνων στα κρυφά στρώματα μπορεί να οδηγήσει σε πολλά προβλήματα. Πρώτον, πάρα πολλοί νευρώνες στα κρυφά στρώματα μπορεί να οδηγήσουν σε υπερπροσαρμογή. Η υπερπροσαρμογή συμβαίνει όταν το νευρωνικό δίκτυο έχει τόση ικανότητα επεξεργασίας πληροφοριών που η περιορισμένη ποσότητα πληροφοριών που περιέχεται στο σετ εκπαίδευσης δεν είναι αρκετή για να εκπαιδεύσει όλους τους νευρώνες στα κρυφά επίπεδα. Ένα δεύτερο πρόβλημα μπορεί να παρουσιαστεί ακόμα και όταν τα δεδομένα εκπαίδευσης είναι αρκετά. Ένας υπερβολικά μεγάλος αριθμός νευρώνων στα κρυφά στρώματα μπορεί να αυξήσει τον χρόνο που απαιτείται για την εκπαίδευση του δικτύου.

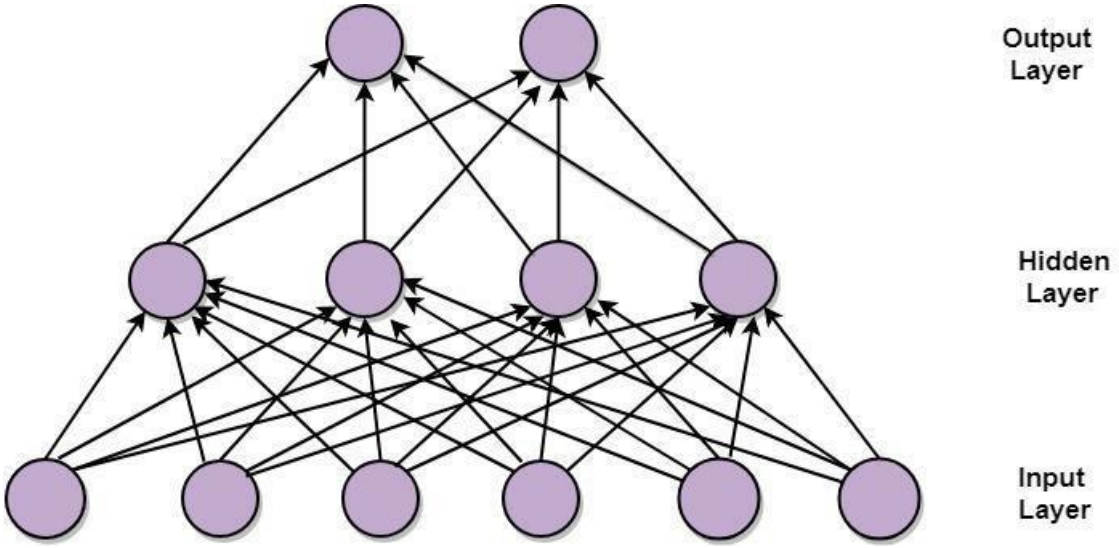
Ένας άλλος τρόπος κατηγοριοποίησης του ANN είναι μέσω του τρόπου με τον οποίο οι πληροφορίες μεταφέρονται μεταξύ των νευρώνων. Έτσι, υπάρχουν τα τροφοδοτικά και τα επαναλαμβανόμενα νευρωνικά δίκτυα.

Νευρωνικό Δίκτυο Feedforward

Στα Feedforward Neural Networks οι νευρώνες είναι οργανωμένοι με τέτοιο τρόπο ώστε να μεταδίδουν πληροφορίες από το ένα επίπεδο στο άλλο. Κάθε νευρώνας ενός στρώματος σχετίζεται με όλους τους νευρώνες του επόμενου στρώματος. Έτσι, δεν υπάρχει νευρώνας όπου η έξοδος του είναι είσοδος για νευρώνα του ίδιου ή προηγούμενου επιπέδου. Ένα παράδειγμα αυτού του είδους δικτύου είναι η backpropagation. Πιο συγκεκριμένα, η διαδικασία μετάδοσης πληροφοριών έχει ως εξής.

Το πρώτο στρώμα νευρώνων είναι το στρώμα εισόδου, από το οποίο οι πληροφορίες μεταφέρονται στο επόμενο στρώμα, το οποίο είναι το πρώτο κρυφό στρώμα. Αυτό δέχεται ως είσοδο την έξοδο του προηγούμενου επιπέδου, το αποτέλεσμα περνά ως είσοδο στο επόμενο κρυφό στρώμα και ούτω καθεξής, έως ότου οι πληροφορίες φτάσουν στο επίπεδο εξόδου, το οποίο είναι και το τελικό

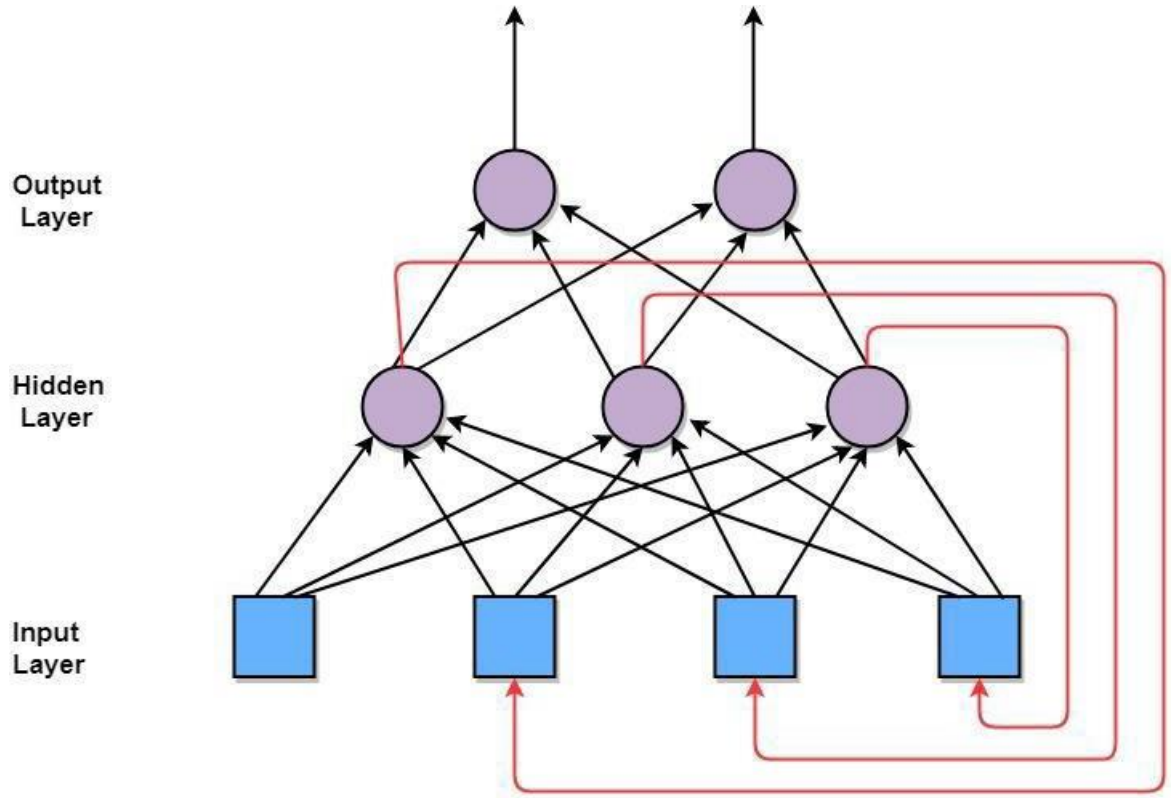
στρώμα του νευρωνικού δικτύου. Γενικά όλα οι νευρώνες του ίδιου επιπέδου έχουν την ίδια λειτουργία ενεργοποίησης, αλλά δεν είναι απαραίτητο να έχουμε τον ίδιο αριθμό νευρώνων σε κάθε επίπεδο.



***Εικόνα 8**: Προοδευτικό νευρωνικό δίκτυο*

Επαναλαμβανόμενο νευρωνικό δίκτυο

Στα επαναλαμβανόμενα νευρωνικά δίκτυα, σε αντίθεση με τα δίκτυα προώθησης, οι πληροφορίες μπορούν να μεταφερθούν από τους νευρώνες ενός στρώματος στους νευρώνες του ίδιου ή του προηγούμενου επιπέδου. Επίσης, είναι δυναμικά, που σημαίνει ότι η κατάστασή τους δεν είναι σταθερή, αλλά αλλάζουν συνεχώς μέχρι να φτάσουν στην επιθυμητή κατάσταση.



Εικόνα 9: *Επαναλαμβανόμενο νευρωνικό δίκτυο*

Αφού γίνει το μαθηματικό υπόβαθρο και αναλυθούν σωστά όλοι οι αλγόριθμοι που πρόκειται να χρησιμοποιηθούν, είμαστε έτοιμοι να προχωρήσουμε στην εφαρμογή αυτών των αλγορίθμων.

2.5.7 Βαθία Μάθηση – Deep Learning

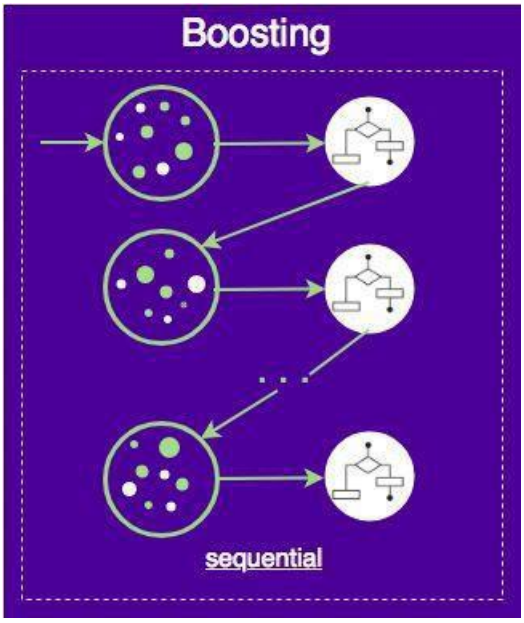
Το Deep Learning είναι μια συλλογή τεχνικών από το τεχνητό νευρωνικό δίκτυο (ANN), το οποίο είναι κλάδος της μηχανικής μάθησης. Τα ANN διαμορφώνονται με βάση τον ανθρώπινο εγκέφαλο και υπάρχουν κόμβοι συνδεδεμένοι μεταξύ τους που περνούν πληροφορίες ο ένας στον άλλο.

Ένα τυπικό νευρωνικό δίκτυο (NN) αποτελείται από πολλούς απλούς, συνδεδεμένους επεξεργαστές που ονομάζονται νευρώνες, ο καθένας από τους οποίους παράγει μια ακολουθία ενεργοποιήσεων πραγματικής αξίας. Οι νευρώνες εισόδου ενεργοποιούνται μέσω αισθητήρων που αντιλαμβάνονται το περιβάλλον, άλλοι νευρώνες ενεργοποιούνται μέσω σταθμισμένων συνδέσεων από

προηγουμένως ενεργούς νευρώνες. Ορισμένοι νευρώνες μπορεί να επηρεάσουν το περιβάλλον ενεργοποιώντας ενέργειες. Η εκμάθηση ή η ανάθεση εργασίας αφορά την εύρεση βαρών που κάνουν το NN να εκδηλώνει επιθυμητή συμπεριφορά, όπως η οδήγηση αυτοκινήτου. Ανάλογα με το πρόβλημα και τον τρόπο σύνδεσης των νευρώνων, μια τέτοια συμπεριφορά μπορεί να απαιτεί μεγάλες αιτιακές αλυσίδες υπολογιστικών σταδίων, όπου κάθε στάδιο μετασχηματίζει τη συνολική ενεργοποίηση του δικτύου. Το Deep Learning αφορά την ακριβή ανάθεση εργασιών σε πολλά τέτοια στάδια (Ramachandran, Rajeev, Krishnan, & Subathra, 2015).

2.5.8 Ενίσχυση - Boosting

Το Boosting είναι ένας διαδοχικός αλγόριθμος που εφαρμόζεται σε ασθενείς ταξινομητές προκειμένου να δημιουργηθεί ένας ισχυρός ταξινομητής συναθροίζοντας τα αποτελέσματα. Με άλλα λόγια, το boosting είναι μια τεχνική συνόλου στην οποία οι προγνωστικοί παράγοντες σημειώνονται ανεξάρτητα, αλλά διαδοχικά. Ο αλγόριθμος ξεκινά με ίσα βάρη που αποδίδονται σε όλες τις παρατηρήσεις, ακολουθούμενες από επόμενες επαναλήψεις όπου δόθηκε περισσότερη έμφαση σε εσφαλμένες παρατηρήσεις αυξάνοντας το βάρος τους και μειώνοντας το βάρος των σωστά ταξινομημένων παρατηρήσεων. Στο τέλος, όλοι οι μεμονωμένοι ταξινομητές συνδυάστηκαν για να δημιουργήσουν, όπως αναφέραμε προηγουμένως, έναν ισχυρό ταξινομητή. Η ενίσχυση μπορεί να έχει πρόβλημα υπερπροσαρμογής, αλλά ρυθμίζοντας προσεκτικά τις παραμέτρους, μπορούμε να πάρουμε το καλύτερο από το μοντέλο εκμάθησης αυτο-μηχανής (Dangeti, 2017). Το παρακάτω σχήμα (Εικόνα), απεικονίζει τον διαδοχικό τρόπο με τον οποίο η ενίσχυση χτίζει τον νέο μαθητή.

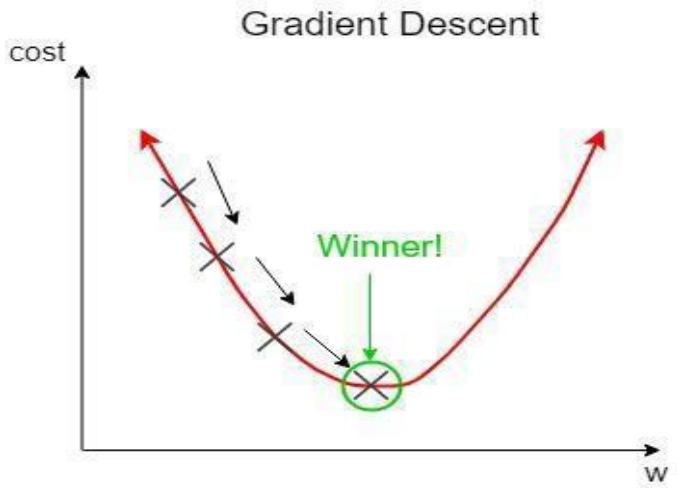


***Εικόνα 10**: Ο διαδοχικός τρόπος που το Boosting χτίζει τον νέο Μαθητή*

2.5.9 Ενίσχυση κλίσης – Gradient Boosting

Στο Gradient Boosting, που ανήκει στους αλγόριθμους ενίσχυσης, η διαδικασία εκμάθησης ταιριάζει διαδοχικά σε νέα μοντέλα για να παρέχει πιο ακριβή εκτίμηση της εξαρτημένης μεταβλητής. Ονομάζεται gradient boosting επειδή χρησιμοποιεί έναν αλγόριθμο gradient descent για να ελαχιστοποιήσει την απώλεια κατά την προσθήκη νέων μοντέλων. Η κύρια ιδέα πίσω από τον αλγόριθμο είναι να κατασκευαστούν οι νέοι μαθητές ώστε να συσχετίζονται στο μέγιστο βαθμό με την αρνητική κλίση της συνάρτησης απώλειας που σχετίζεται με ολόκληρο το σύνολο. Η συνάρτηση απώλειας που εφαρμόζεται μπορεί να είναι αυθαίρετη, αλλά εάν η συνάρτηση σφάλματος είναι η κλασική απώλεια τετραγώνου σφάλματος, η διαδικασία εκμάθησης θα οδηγήσει σε διαδοχική προσαρμογή σφαλμάτων (Natekin & Knoll, 2013). Με άλλα λόγια, η ενίσχυση κλίσης είναι μια τεχνική μηχανικής μάθησης για προβλήματα παλινδρόμησης και ταξινόμησης, η οποία παράγει ένα μοντέλο πρόβλεψης με τη μορφή ενός συνόλου αδύναμων μοντέλων πρόβλεψης. Ο στόχος κάθε εποπτευόμενου αλγορίθμου μάθησης είναι να ορίσει μια συνάρτηση απώλειας και να την ελαχιστοποιήσει και υπάρχει μια πλούσια ποικιλία συναρτήσεων απώλειας, αλλά εναπόκειται στον

ερευνητή να επιλέξει μία από αυτές. Ο τρόπος ελαχιστοποίησης της συνάρτησης απώλειας μπορεί να περιγραφεί στο Σχήμα . Ξεκινώντας από την κορυφή του βουνού, κάνουμε το πρώτο μας βήμα κατηφορικά προς την κατεύθυνση που ορίζει η αρνητική κλίση. Στη συνέχεια, υπολογίζουμε ξανά την αρνητική κλίση και παίρνουμε άλλο ένα βήμα προς την κατεύθυνση που ορίζει. Συνεχίζουμε αυτή τη διαδικασία επαναληπτικά μέχρι να φτάσουμε στο κάτω μέρος του γραφήματος μας, ή σε ένα σημείο, όπου δεν μπορούμε πλέον να κινηθούμε κατηφορικά - ένα τοπικό ελάχιστο.



***Εικόνα 11**: Σύγκλιση κλίσης καθόδου*

2.5.10 CatBoost

Το CatBoost είναι ένας πρόσφατα ανοιχτός αλγόριθμος μηχανικής μάθησης που αναπτύχθηκε από την Yandex και παρέχει κορυφαία ακρίβεια για τις πιο συνηθισμένες επιχειρηματικές περιπτώσεις μηχανικής μάθησης. Το CatBoost βασίζεται στην ενίσχυση κλίσης, μια τεχνική μηχανικής εκμάθησης που λειτουργεί εξαιρετικά εάν έχετε δεδομένα από διαφορετικές πηγές.

Πολλά άλλα εργαλεία απαιτούν τον λεγόμενο συντονισμό παραμέτρων, πρέπει να εκτελέσουμε τον αλγόριθμο πολλές φορές και έχουμε καλό αποτέλεσμα μόνο μετά από πολλούς γύρους, αλλά το CatBoost παρέχει εξαιρετικά αποτελέσματα

μετά τον πρώτο γύρο. Μειώνει την ανάγκη για εκτεταμένο συντονισμό υπερπαραμέτρων και μειώνει τις πιθανότητες υπερβολικής προσαρμογής, γεγονός που οδηγεί σε πιο γενικευμένο μοντέλο. Ωστόσο, ο αλγόριθμος CatBoost έχει πολλαπλές παραμέτρους για συντονισμό και περιέχει παραμέτρους όπως ο αριθμός των δέντρων, ο ρυθμός εκμάθησης, η τακτοποίηση, το βάθος δέντρου, το μέγεθος διπλώματος, η θερμοκρασία σακουλών και άλλες.

Ένα άλλο πλεονέκτημα του αλγόριθμου CatBoost είναι ότι μπορεί να χειριστεί αυτόματα κατηγορικά δεδομένα. Έτσι, μπορούμε να το χρησιμοποιήσουμε χωρίς καμία ξεχωριστή προεπεξεργασία για να μετατρέψουμε κατηγορίες σε αριθμούς. Για να το πετύχει αυτό, το CatBoost χρησιμοποιεί πολλά στατιστικά στοιχεία για συνδυασμούς κατηγορικών χαρακτηριστικών και συνδυασμούς κατηγορικών και αριθμητικών χαρακτηριστικών. Επιπλέον, το CatBoost είναι επεκτάσιμο, και με αυτό εννοούμε ότι επιτρέπει τον καθορισμό προσαρμοσμένων συναρτήσεων απώλειας, καθώς και εύχρηστο από τη γραμμή εντολών, χρησιμοποιώντας ένα φιλικό προς το χρήστη API τόσο για την Python όσο και για το R.

2.5.11 XGBoost

Ο XGBoost είναι ένας νέος αλγόριθμος που αναπτύχθηκε το 2014 από τον Tianqi Chen με βάση τις αρχές ενίσχυσης Gradient (Dangeti, 2017). Είναι μια από τις πιο δημοφιλείς και αποτελεσματικές υλοποιήσεις του αλγόριθμου Gradient Boosted Trees, μιας εποπτευόμενης μεθόδου εκμάθησης που βασίζεται στην προσέγγιση συναρτήσεων βελτιστοποιώντας συγκεκριμένες συναρτήσεις απώλειας και εφαρμόζοντας πολλές τεχνικές τακτοποίησης. Ο στόχος του αλγορίθμου είναι να ωθήσει τα άκρα των ορίων υπολογισμού των μηχανών για να παρέχει κλιμακούμενα, φορητά και ακριβή αποτελέσματα.

Το πιο σημαντικό πλεονέκτημα του αλγορίθμου XGBoost έναντι της ενίσχυσης διαβάθμισης είναι όσον αφορά την απόδοση και τις διαθέσιμες επιλογές για τον έλεγχο του συντονισμού του μοντέλου. Αλλάζοντας μερικά από αυτά, το XGBoost κερδίζει ακόμη και την ενίσχυση κλίσης (Dangeti, 2017). Ένα

μειονέκτημα του αλγορίθμου είναι ότι δεν μπορεί να χειριστεί κατηγορηματικά χαρακτηριστικά από μόνος του και δέχεται μόνο αριθμητικές τιμές.

Η συνάρτηση απώλειας και η τακτοποίηση του αλγορίθμου κατά την επανάληψη t που πρέπει να ελαχιστοποιήσουμε είναι τα εξής:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

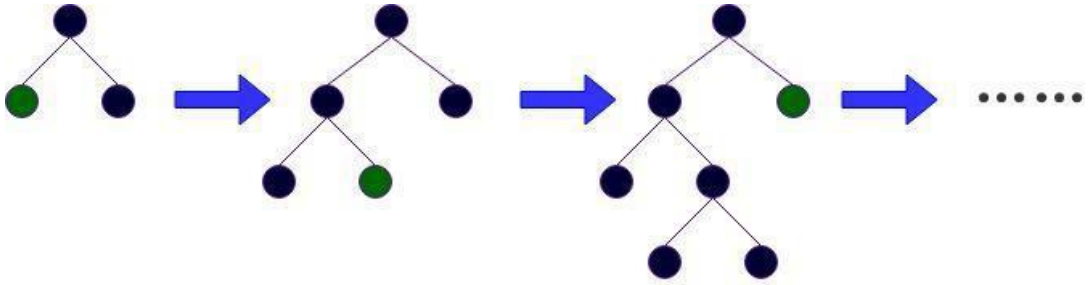
όπου το l είναι μια διαφοροποιήσιμη κυρτή συνάρτηση απώλειας που μετρά τη διαφορά μεταξύ της πρόβλεψης και του στόχου yi. Ο δεύτερος όρος Ω τιμωρεί την πολυπλοκότητα του μοντέλου.

3.5.12 Light GBM

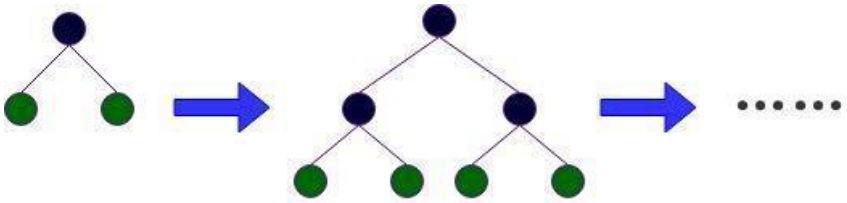
Το Light GBM είναι ένα πλαίσιο ενίσχυσης κλίσης που χρησιμοποιεί αλγόριθμο μάθησης βάσει δέντρων. Αναπτύσσεται σε δέντρο κατακόρυφα ενώ άλλοι αλγόριθμοι μεγαλώνουν σε δέντρο οριζόντια και με αυτό εννοούμε ότι το Light GBM μεγαλώνει κατά φύλλα (Εικόνα), ενώ άλλοι αλγόριθμοι αναπτύσσονται σε επίπεδο (Εικόνα) και η διαφορά τους φαίνεται στα δύο παρακάτω σχήματα . Θα επιλέξει το φύλλο με τη μέγιστη απώλεια δέλτα για την ανάπτυξη. Οι αλγόριθμοι με γνώμονα τα φύλλα, όπως το Light GBM, όταν μεγαλώνουν το ίδιο φύλλο, μπορούν να μειώσουν περισσότερες απώλειες από έναν αλγόριθμο σε επίπεδο.

Ένα πλεονέκτημα του αλγόριθμου Light GBM είναι ότι, όπως και το CatBoost, μπορεί να χειριστεί κατηγορίες χαρακτηριστικών λαμβάνοντας την εισαγωγή ονομάτων χαρακτηριστικών. Αυτό που κάνει αυτόν τον αλγόριθμο δημοφιλή είναι ότι εστιάζει στην ακρίβεια των αποτελεσμάτων. Επιπλέον, έχει υψηλή ταχύτητα, μπορεί να χειριστεί το μεγάλο μέγεθος δεδομένων και χρειάζεται μικρότερη μνήμη για να τρέξει.

Το Light GBM είναι ευαίσθητο στην υπερπροσαρμογή και μπορεί εύκολα να υπερπροσαρμόσει μικρά δεδομένα και επομένως δεν συνιστάται η χρήση αυτού του αλγόριθμου σε μικρά σύνολα δεδομένων.



***Εικόνα 12**: Εξήγηση του πώς λειτουργεί το LGBM*



***Εικόνα 13**: Πώς λειτουργούν άλλοι αλγόριθμοι Ενίσχυσης*

Κεφάλαιο 3 – Μεθοδολογία

Σχεδιασμός και Υλοποίηση

Στο παρόν κεφάλαιο θα αναπτυχθεί αναλυτικά η μεθοδολογία που ακολουθήθηκε για την εκτέλεση της παρούσας διπλωματικής εργασίας και την εξαγωγή των επιθυμητών αποτελεσμάτων. Τα βήματα που εκτελέστηκαν είναι:

1) η συλλογή των δεδομένων - τεχνική web scraping,
2) ο καθορισμός και η επεξεργασία των δεδομένων
3) η στατιστική ανάλυση των δεδομένων

Στη συνέχεια περιγράφεται αναλυτικά η ανωτέρω διαδικασία ως τη βάση για την υλοποίηση αυτής της εργασίας.

3.1 Δεδομένα

Στα πλαίσια του πρακτικού μέρους της διπλωματικής μας εργασίας και των αντίστοιχων μοντέλων - παραδειγμάτων χρησιμοποιήσαμε δεδομένα που έχουν να κάνουν με ποικίλες πληροφορίες για 20996 ακίνητα στις πόλεις της Αθήνας, του Πειραιά και της Θεσσαλονίκης αλλά και στην ευρύτερη περιοχή τους. Επιπρόσθετα, οι πληροφορίες αυτές κατανέμονται δηλαδή εκφράζονται από τις 10 βασικές στήλες - μεταβλητές που έχει το πλαίσιο δεδομένων πριν ξεκινήσουμε την επεξεργασία αυτών. Αυτές οι μεταβλητές προσδιορίζουν συγκεκριμένα χαρακτηριστικά των ακινήτων και είναι οι εξής:

- Η τιμή (ευρώ) – ‘Price’
- Το εμβαδόν (μ2) – ‘Area’
- Ο τύπος ακινήτου (πχ διαμέρισμα, μονοκατοικία, ρετιρέ κτλ) – ‘H_type’
- Το έτος κατασκευής του ακινήτου – ‘Con_year’
- Σε ποιο όροφο βρίσκεται – ‘Levels’
- Πόσα μπάνια έχει – ‘Baths’

- Πόσα κρεβάτια έχει – ‘Beds’
- Το μέσον θέρμανσης – ‘Heating’
- Η πόλη του ακινήτου – ‘City’
- Η γειτονιά του ακινήτου – ‘Neighborhood’

3.2 Συλλογή Δεδομένων

3.2.1 Εισαγωγή στο Web Scraping

Το Web Scraping είναι η διαδικασία συλλογής δεδομένων ιστού με αυτοματοποιημένο τρόπο. Ονομάζεται επίσης εξαγωγή δεδομένων ιστού. Μερικές από τις κύριες περιπτώσεις χρήσης του web scraping περιλαμβάνουν παρακολούθηση τιμών, παρακολούθηση ειδήσεων και έρευνα αγοράς μεταξύ πολλών άλλων.

Σε γενικές γραμμές, η εξαγωγή δεδομένων ιστού χρησιμοποιείται από άτομα και επιχειρήσεις που θέλουν να κάνουν χρήση του τεράστιου όγκου διαθέσιμων στο κοινό δεδομένων ιστού για τη λήψη πιο έξυπνων αποφάσεων.

Σε αντίθεση με τη συνηθισμένη, χειροκίνητη διαδικασία εξαγωγής δεδομένων με μη αυτόματο τρόπο, το web scraping χρησιμοποιεί έξυπνο αυτοματισμό για να ανακτήσει εκατοντάδες εκατομμύρια ή ακόμα και δισεκατομμύρια στοιχεία δεδομένων από το Διαδίκτυο.

Οι web scrapers λειτουργούν με έναν φαινομενικά απλό τρόπο. Καταρχάς, στον web scraper θα δοθούν μία ή περισσότερες διευθύνσεις URL για φόρτωση πριν από τη δημιουργία. Στη συνέχεια, αυτός φορτώνει ολόκληρο τον κώδικα HTML για την εν λόγω σελίδα. Οι πιο προηγμένοι web scrapers θα φορτώσουν ολόκληρο τον ιστότοπο, συμπεριλαμβανομένων στοιχείων CSS και Javascript.

Στη συνέχεια, ο scraper θα εξάγει όλα τα δεδομένα στη σελίδα ή συγκεκριμένα δεδομένα που θα επιλέξει ο χρήστης πριν από τη εκτέλεση της εργασίας. Στην ιδανική περίπτωση, ο χρήστης θα περάσει από τη διαδικασία επιλογής των συγκεκριμένων δεδομένων που θέλει από τη σελίδα.

Τέλος, ο web scraper θα εξάγει όλα τα δεδομένα που έχουν συλλεχθεί σε μορφή που είναι πιο χρήσιμη για τον χρήστη.

Ο λόγος που χρησιμοποιείται web scraper στα πλαίσια της εργασίας είναι ότι τα δεδομένα που αφορούν τα χαρακτηριστικά και τις τιμές ακινήτων δεν βρίσκονται έτοιμα σε δομημένη μορφή στην οποία να έχει οποιοσδήποτε πρόσβαση.

Επειδή το web scraping μπορεί να είναι επιβαρυντικό για τον ιστότοπο στον οποίο δρα (λόγω του πλήθους των αιτημάτων που μπορεί να στείλει σε μικρό χρονικό διάστημα), είναι σημαντικό να ρυθμιστεί έτσι ώστε να μην παρεμποδίζει την ομαλή λειτουργία του ιστοτόπου όσο λειτουργεί.

3.2.2 Μέθοδοι Web Scraping

Θα παρουσιαστούν 4 άξονες που διαφοροποιούν τους web scrapers. Φυσικά υπάρχουν ακόμα περισσότερες δυνατότητες διαχωρισμού αλλά εδώ παρουσιάζονται οι 4 αυτές βασικές.

- **Αυτό-κατασκευασμένο ή Προ-κατασκευασμένο:** Η δημιουργία από το μηδέν ενός web scraper απαιτεί κάποιες προηγμένες γνώσεις προγραμματισμού. Το εύρος αυτής της γνώσης αυξάνεται επίσης με τον αριθμό των δυνατοτήτων που είναι επιθυμητές να έχει ο web scraper. Για τον λόγο αυτό υπάρχει πλήθος από προκατασκευασμένες λύσεις αλλά και πλατφόρμες που προσφέρουν έτοιμα δομικά στοιχεία για την δημιουργία ενός web scraper. Οι περισσότερες από τις πλατφόρμες αυτές προσφέρουν και επιπλέον υπηρεσίες όπως χρήση proxies για την δημιουργία αιτημάτων από τους scrapers.
- **Επέκταση προγράμματος περιήγησης ή λογισμικό:** Σε γενικές γραμμές, οι web scrapers διατίθενται σε δύο μορφές. Επεκτάσεις προγράμματος περιήγησης ή λογισμικό υπολογιστή. Οι επεκτάσεις προγράμματος περιήγησης είναι προγράμματα που μοιάζουν με εφαρμογές και μπορούν να προστεθούν σε ένα πρόγραμμα περιήγησης. Οι επεκτάσεις προγράμματος περιήγησης έχουν το πλεονέκτημα της απλούστερης εκτέλεσης και της ενσωμάτωσης απευθείας στο πρόγραμμα

περιήγησης. Ωστόσο, αυτές οι επεκτάσεις περιορίζονται συνήθως από τη διανομή τους σε αυτό. Αυτό σημαίνει ότι κάθε προηγμένη δυνατότητα του θα έπρεπε να εμφανιστεί εκτός αυτού θ ήταν αδύνατο να εφαρμοστεί. Για παράδειγμα, οι περιστροφές IP δεν θα ήταν δυνατές. Από την άλλη, το λογισμικό υπολογιστή είναι λιγότερο βολικό από τις επεκτάσεις του προγράμματος περιήγησης αλλά προσφέρει συνήθως περισσότερο προηγμένες λειτουργίες που δεν περιορίζονται από αυτό που μπορεί και δεν μπορεί να κάνει ένα πρόγραμμα περιήγησης.

- **Διεπαφή χρήστη:** Η διεπαφή χρήστη μεταξύ των web scrapers μπορεί να ποικίλει πολύ. Για παράδειγμα, ορισμένα εργαλεία web scraping εκτελούνται με ένα ελάχιστο περιβάλλον εργασίας χρήστη και μια γραμμή εντολών. Από την άλλη πλευρά, ορισμένοι web scrapers έχουν ένα πλήρες περιβάλλον εργασίας χρήστη όπου ο ιστότοπος είναι ορατός ώστε ο χρήστης να επιλέγει εύκολα ποια δεδομένα θέλει να διαγράψει και ποια να κρατήσει. Αυτοί οι web scrapers είναι συνήθως πιο εύκολο να χρησιμοποιηθούν για τα περισσότερα άτομα με περιορισμένες τεχνικές γνώσεις.
- **Στο Cloud ή τοπικό:** Σε περίπτωση που οι πόροι ενός προσωπικού υπολογιστή ή ενός απλού εταιρικού συστήματος δεν αρκούν για την ικανοποιητική απόδοση ενός web scraper, μία συνήθης λύση είναι η μεταφορά του στο Cloud. Λύνει επίσης προβλήματα όπως η δυναμική απόδοση IP καθώς συνήθως οι υπηρεσίες που προσφέρουν web scraping υπηρεσίες βασισμένες στο Cloud προσφέρουν και τη δυνατότητα εναλλαγής IP σε κάθε κλήση του web scraper. Με τον τρόπο αυτό αποφεύγεται το μπλοκάρισμά του από την σελίδα την οποία επισκέπτεται. Συνήθως οι Cloud πλατφόρμες οι οποίες προσφέρουν δυνατότητες φιλοξενίας web scrapers παρέχουν και δικές τους αρχιτεκτονικές για την κατασκευή τους και τον έλεγχο της ομαλής λειτουργίας τους.

3.2.3 Διαδικασία εξαγωγής Δεδομένων

Η εξαγωγή δεδομένων έγινε χρησιμοποιώντας τεχνικές web scraping μέσω της χρήσης της γλώσσας προγραμματισμού Python.

Για την εξαγωγή δεδομένων από τις πλατφόρμες προγραμματίστηκε web scraper μέσω της γλώσσας προγραμματισμού Python. Συγκεκριμένα, δημιουργήθηκε ένας Actor ο οποίος επέλεγε πληροφορίες μόνο για ακίνητα στις πόλεις της Θεσσαλονίκης, του Πειραιά και της Αττικής. Η δομή του βασίστηκε στο να κάνει scrape πληροφορίες για μη πωληθέντα ακίνητα. Στη συνέχεια οι πληροφορίες των ακινήτων αυτών αποθηκεύονται από τον scraper σε μορφή .ipynb και στη συνέχεια σε αρχείο της μορφής ecxel.

Τα δεδομένα συλλέχθηκαν με τις εξής παραμέτρους:

- *Τύπος ακινήτων:* Πωλήσεις
- *Μέγιστη ηλικία αγγελίας:* 30 ημέρες (στην πραγματικότητα όμως, διατηρούνται πολύ παλαιότερες αγγελίες, μέχρι και πάνω από 3 χρόνια, οι οποίες ανανεώνονται)
- *Τιμή:* Μεταξύ 20.000 και 1.000.000 ευρώ (το εύρος αυτό μικραίνει λόγω ύπαρξης ακραίων τιμών)
- *Τετραγωνικά μέτρα:* Μεταξύ 15 και 2500 τ.μ. (το εύρος αυτό μικραίνει λόγω ύπαρξης ακραίων τιμών)
- Για κάθε ιστοσελίδα ορίστηκαν συναρτήσεις επιλογής για το ποια χαρακτηριστικά θα συλλέξει ο scraper (π.χ. H_type, area, price, city κτλ.)
- Για την συλλογή των χαρακτηριστικών του περιβάλλοντος του ακινήτου (πχ εάν είναι πλησίον μεταφορικών μέσων, σούπερ μάρκετ και σχολεία), η διαδικασία ήταν αδύνατη, λόγω της έλλειψης ή της μη σωστής καταγραφής τους στη παραγόμενη λίστα δεδομένων ή ακόμα και μετά την συλλογή τους, θα διαγράφονταν οι μεταβλητές αυτές λόγω αστοχίας των τελικών αποτελεσμάτων.

❖ *Η Ιστοσελίδα Χρυσή Ευκαιρία (ΧΕ)*

	H_type	Price	Area	City	Beds	Baths	Floors	Con_year	Heating
0	[Διαμέρισμα]	[\n, [120.000 €], \n]	[90 τ.μ.]	[[Αθήνα (Ηπείρου)]]	[2]	[1]	[1]	[Ανακαινισμένο]	[Διαθέσιμο από:]
1	[Διαμέρισμα]	[\n, [280.000 €], \n]	[74 τ.μ.]	[[Αθήνα (Ιπποκράτειο)]]	[2]	[2]	[3ος]	[1971]	[Κατάλληλο για επένδυση]
2	[Κτίριο]	[\n, [1.000.000 €], \n]	[600 τ.μ.]	[[Αθήνα (Πολύγωνο)]]	[15]	[12]	[4ος, 3ος, 2ος, 1ος, Ισόγειο, Υπόγειο]	[1985]	[Μέσο θέρμανσης:]
3	[Διαμέρισμα]	[\n, [130.000 €], \n]	[67 τ.μ.]	[[Αθήνα (Καλλιρρόης)]]	[2]	[1]	[4ος]	[1976]	[Κλήση]
4	[Διαμέρισμα]	[\n, [75.000 €], \n]	[30 τ.μ.]	[[Αθήνα (Κουντουριώτικα)]]	[1]	[1]	[4ος]	[1978]	[Αποστολή μηνύματος]

Πίνακας 1: *Δείγμα δεδομένων που συλλέχθηκαν από την ιστοσελίδα* Χρυσή Ευκαιρία (ΧΕ)

❖ *Η Ιστοσελίδα Plot*

	H_type	Price	Container
0	[\n Διαμέρισμα 60 τ.μ. για πώληση, Πειραι...	[140.000 €]	[[[40567232]], [[<a d...
1	[\n Διαμέρισμα 140 τ.μ. για πώληση, Αθήνα...	[400.000 €]	[[[40148050]], [[090]], [[<div class="tw-inlin...
2	[\n Διαμέρισμα 120 τ.μ. για πώληση, Αθήνα...	[290.000 €]	[[[40679909]], [[<a d...
3	[\n Διαμέρισμα 110 τ.μ. για πώληση, Αθήνα...	[420.000 €]	[[[40071930]], [[<a d...
4	[\n Διαμέρισμα 86 τ.μ. για πώληση, Υπόλοι...	[460.000 €]	[[[40679081]], [[<a d...

Πίνακας 2: *Δείγμα δεδομένων που συλλέχθηκαν από την ιστοσελίδα* Plot

❖ *Η Ιστοσελίδα Spiti360*

	H_type	Price	Area	City	Neighbourhood	Beds	Baths	Floor	Heating	Con_year
0	[Διαμέρισμα]	[210.000 €]	[120 τ.μ.]	[Δημος Πειραιά]	[Ευαγγελίστρια]	[3]	[2]	[5ος]	[Ατομικό ηλεκτρικό]	[1967]
1	[Διαμέρισμα]	[400.000 €]	[152 τ.μ.]	[Δημος Πειραιά]	[Καστέλα]	[3]	[1]	[1]	[5ος]	[Κεντρικό πετρέλαιο]
2	[Διαμέρισμα]	[115.000 €]	[64 τ.μ.]	[Δημος Πειραιά]	[Καλλίπολη]	[2]	[1]	[3ος]	[Απροσδιόριστο]	[1970]
3	[Μονοκατοικία]	[440.000 €]	[200 τ.μ.]	[Δημος Πειραιά]	[Χατζηκυριάκειο]	[4]	[3]	[Ισόγειο]	[Αυτόνομο πετρέλαιο]	[2001]
4	[Διαμέρισμα]	[80.000 €]	[68 τ.μ.]	[Δημος Πειραιά]	[Καστέλα]	[2]	[1]	[1ος]	[1972]	[03/04/2023]

Πίνακας 3: *Δείγμα δεδομένων που συλλέχθηκαν από την ιστοσελίδα* Spiti360

3.3 Επεξεργασία δεδομένων

3.3.1 Προετοιμασία - Καθαρισμός δεδομένων

Προτού προχωρήσουμε στην δημιουργία μοντέλων μηχανικής μάθησης πρέπει να ελέγξουμε αν υπάρχουν NAN values, δηλαδή τιμές που λείπουν είναι σύμβολα (όπως π.χ. το άπειρο) ή ακόμα και τιμές οι οποίες δεν συνάδουν με το χαρακτηριστικό και την τιμή που υποδηλώνει η αντίστοιχη μεταβλητή. Βρίσκουμε, λοιπόν, ότι υπάρχουν 68 στην μεταβλητή ‘’H_type’’, 11 στην ‘’Price’’, 3122 στην ‘’Area’’, 1061 στην ‘’City’’, 20408 στην ‘’Neighborhood’’, 4863 στην “Baths”, 1528 στην ‘’Beds’’, 5923 στην ‘’Levels’’, 11938 στην ‘’Con_year’’ και 12803 στην ‘’Heating’’.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20996 entries, 0 to 20995
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   H_type        20928 non-null  category
 1   Price         20985 non-null  float64
 2   Area          17874 non-null  float64
 3   City          19935 non-null  category
 4   Neighborhood  588 non-null    category
 5   Beds          19468 non-null  float64
 6   Baths         16133 non-null  float64
 7   Levels        15073 non-null  float64
 8   Heating       8193 non-null   category
 9   Con_year      9058 non-null   float64
dtypes: category(4), float64(6)
memory usage: 1.1 MB
```

***Πίνακας 4**: Δείγμα συνόλου δεδομένων με μη μηδενικές τιμές των μεταβλητών*

Υπάρχουν 2 τρόποι για να διαχειριστούμε τις εκλειπόμενες τιμές. Ο 1ος είναι απλά να τις διαγράψουμε, άρα να διαγράψουμε και τις αντίστοιχες γραμμές, εμείς αποφασίσαμε να διαγράψουμε μόνο την μεταβλητή ‘’Neighborhood’’ αφού το

97% των τιμών ήταν ελλιπής και ο 2ος να αντικαταστήσουμε την εκάστοτε τέτοια τιμή με τον μέσο όρο της μεταβλητής – χαρακτηριστικού (Ironhardt B., 2020). Αυτός ο τρόπος πραγματοποιείται για όλες τις υπόλοιπες μεταβλητές.

Επιλέξαμε να μην διαγράψουμε τις εκλειπόμενες τιμές και τις αντίστοιχες παρατηρήσεις των μεταβλητών ‘’Heating’’ και ‘’Con_year, καθώς αν συνέβαινε κάτι τέτοιο το σύνολο δεδομένων θα αποτελούνταν από μόλις 8193 ή από 9058 ακίνητα αντίστοιχα, αντί 20996 που ήταν αρχικά. Όπως, γίνεται κατανοητό κάτι τέτοιο είναι λάθος, καθώς μια έρευνα πρέπει να βασίζεται σε πραγματικά δεδομένα και να μην γίνεται ποσοστιαία τόσο μεγάλη αλλαγή στο πλήθος τους. Η αντικατάσταση των Missing Values γίνεται είτε με τον μέσο όρο (mean) της αριθμητικής στήλης που ανήκει η εκλειπόμενη τιμή είτε με την επικρατούσα τιμή (mode) της ή διαφορετικά <<γεμίζουμε>> τις εκλειπόμενες μεταβλητές με αυτές που βρίσκονται αμέσως μετά της άνω και κάτω μεταβλητής στην εκάστοτε στήλη. Εμείς, επιλέξαμε την 1η και 3η επιλογή, καθώς αποτελούν τις πιο διαδεδομένες και σίγουρες λύσεις, καθώς με την επικρατούσα τιμή υπάρχει η περίπτωση δημιουργίας στατιστικών λαθών κατά την μετατροπή, όπως για παράδειγμα στην κατανομή της μεταβλητής. Έτσι, λοιπόν, ≪γεμίζουμε≫ τις τιμές που λείπουν από τις στήλες: ‘Price’, ‘Area’, ‘Beds’, ‘Baths’, ‘Levels’ και ‘Con_year’ με την μέση τιμή τους, ενώ στις κατηγορικές μεταβλητές: ‘H_type’, ‘City’ και ‘Heating’ αντικαθιστούμε τις τιμές με αυτές που βρίσκονται αμέσως μετά πάνω και κάτω από τις κενές μεταβλητές της εκάστοτε στήλης.

Το επόμενο βήμα κατά την προετοιμασία των δεδομένων είναι να μετατρέψουμε τον τύπο των μεταβλητών σε όσες δεν ανταποκρίνεται η πραγματική τους σημασία στον σωστό τύπο. Συγκεκριμένα, όσες μεταβλητές είναι κατηγορικές μετατρέπονται σε category ως είθισται στην γλώσσα Python. Επιπλέον, τις αριθμητικές μεταβλητές τις μετατρέπουμε από ακέραιες αριθμητικές σε συνεχείς, ενώ εκφράζουν ακέραια ποσότητα (όλα τα ακίνητα κοστολογούνται με κάποια ακέραια τιμή αλλά για λόγους στατιστικούς και ενοποίηση τιμών επιλέγουμε τη μορφοποίηση αυτή).

```
print(df.dtypes)
```

```
H_type       category
Price         float64
Area          float64
City         category
Beds          float64
Baths         float64
Levels        float64
Heating      category
Con_year      float64
```

***Πίνακας 5**: Απεικόνιση του τύπου των μεταβλητών*

Ακραίες Τιμές & Αφαίρεση Διπλοτύπων

Σε αυτήν την φάση, θα διευθετήσουμε τις ακραίες τιμές που υπάρχουν στο σύνολο δεδομένων. Οι ακραίες τιμές έχουν την δυνατότητα να μας προσφέρουν σημαντικές πληροφορίες για την συμπεριφορά του δείγματος και επηρεάζουν σημαντικά την εκπαίδευση του μοντέλου. Ξεκινώντας θα διορθώσουμε για κάθε ένα από τα πεδία των μεταβλητών τις ακραίες τιμές που ενδέχεται να υπάρχουν.

Έπειτα, από τις παραπάνω ενέργειες κοιτάμε να δούμε αν υπάρχουν διπλότυπα δεδομένα και να τα διαγράψουμε, κάτι πολύ σημαντικό σε αυτό το στάδιο για δυο λόγους: 1) Δεν θέλουμε το dataset μας να αποτελείται από διπλά και τριπλά δεδομένα καθώς αυξάνει την αναξιοπιστία του αποτελέσματος και 2) να αποφύγουμε στατιστικά σφάλματα.

Να σημειωθεί εδώ όμως, ότι θα πρέπει κάθε ερευνητής να φέρει μια ισορροπία μεταξύ του όγκου των δεδομένων, δηλαδή τον αριθμό των γραμμών και τον αριθμό των μεταβλητών που θα χρησιμοποιηθούν για την αποφυγή των στατιστικών σφαλμάτων.

Προσθήκη νέων μεταβλητών - Feature Engineering

Στη συνέχεια, θα ασχοληθούμε με την συμπλήρωση κάποιων νέων σημαντικών μεταβλητών ποιοτικού χαρακτήρα για να προσδώσουμε στο τελικό dataset μας

επιπλέον χαρακτηριστικά που προέρχονται, μετά από μετασχηματισμούς στην γλώσσα Python, από τις μεταβλητές που έχουμε συλλέξει μέχρι τώρα. Με αυτήν την διαδικασία η εκτίμηση των προβλέψεων τιμών των ακινήτων θα γίνει πιο αποδοτική.

Παρακάτω παρουσιάζονται αυτές οι νέες μεταβλητές οι οποίες είναι:

1. Η 1η μεταβλητή είναι ο Φόρος Ένφια που αναλογεί σε κάθε ακίνητο, με βάση λοιπόν την αξία του ακινήτου και το ποσοστό φόρου που καθορίζεται από την αρμόδια αρχή (Α.Α.Δ.Ε.), ορίσαμε την μεταβλητή 'Enfia'.
2. Η 2η μεταβλητή είναι ο πληθυσμός κάθε πόλης, των προαστίων αυτών και τις υπόλοιπες γύρω περιοχές τους, με βάση των κριτηρίων που κάναμε όπως προαναφέρθηκε στην διαδικασία συλλογής των μεταβλητών. Την ορίσαμε ως 'Population'.
3. Η 3η μεταβλητή είναι το ποσοστό εγκληματικότητας της κάθε περιοχής του ακινήτου. Καθορίστηκε με βάση το ποσοστό εγκληματικότητας της κάθε πόλης και των υπό περιοχών τους (γειτονιές). Την ορίσαμε ως 'Crime_rate'.

Ως αποτέλεσμα, το καινούριο σύνολο δεδομένων που προκύπτει αποτελείται από 20996 σειρές -παρατηρήσεις, οι οποίες αναφέρονται στο κάθε ακίνητο ξεχωριστά και σε 12 στήλες – μεταβλητές, οι οποίες εκφράζουν τα 12 διαφορετικά χαρακτηριστικά των ακινήτων αυτών.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20996 entries, 0 to 20995
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   H_type      20996 non-null  category
 1   Price       20996 non-null  float64
 2   Area        20996 non-null  float64
 3   City        20996 non-null  category
 4   Beds        20996 non-null  float64
 5   Baths       20996 non-null  float64
 6   Levels      20996 non-null  float64
 7   Heating     20996 non-null  category
 8   Con_year    20996 non-null  float64
 9   Enfia       20996 non-null  float64
 10  Population  20996 non-null  float64
 11  Crime_rate  20996 non-null  float64
dtypes: category(3), float64(9)
memory usage: 1.5 MB
```

***Πίνακας 6**: Δείγμα συνόλου δεδομένων με μη μηδενικές τιμές των μεταβλητών μετά την επεξεργασία*

3.3.2 Δομή Δεδομένων

Ακολουθεί η παρουσίαση ενός συλλογικού δείγματος των ακινήτων που έγιναν συλλογή και επεξεργασία αυτών.

	H_type	Price	Area	City	Beds	Baths	Levels	Heating	Con_year	Enfia	Population	Crime_rate
0	Μονοκατοικία	460000.0	165.000000	Αθήνα - Ανατολικά Προάστια	5.000000	2.000000	1.000000	Κεντρικό πετρέλαιο	1983.734378	483.285000	168151.0	7.600195
1	Διαμέρισμα	120000.0	85.000000	Αθήνα	5.000000	1.000000	2.237179	Κεντρικό πετρέλαιο	1975.000000	248.965000	1002212.0	14.683246
2	Διαμέρισμα	280000.0	140.000000	Αθήνα	2.000000	1.000000	1.000000	Αυτόνομο αέριο	1983.734378	410.060000	1002212.0	14.683246
3	Διαμέρισμα	350000.0	102.000000	Αθήνα	3.000000	1.000000	1.000000	Αυτόνομο αέριο	1983.734378	298.758000	1002212.0	14.683246
4	Γκαρσονιέρα	140000.0	60.000000	Θεσσαλονίκη	1.000000	1.000000	7.000000	Αυτόνομο αέριο	2010.000000	175.740000	319045.0	15.475094
5	Διαμέρισμα	121500.0	111.000000	Αθήνα	2.000000	1.000000	1.000000	Κεντρικό αέριο	1983.734378	325.119000	1002212.0	14.683246
6	Διαμέρισμα	172000.0	123.937619	Θεσσαλονίκη Περ/κοι Δήμοι	1.000000	5.000000	2.237179	Κεντρικό αέριο	1983.734378	363.013286	784978.0	38.074906
7	Μεζονέτα	385000.0	123.937619	Αθήνα - Ανατολικά Προάστια	3.000000	1.000000	2.237179	Κεντρικό αέριο	2018.000000	363.013286	168151.0	7.600195
8	Διαμέρισμα	505000.0	130.000000	Αθήνα - Βόρεια Προάστια	5.000000	2.000000	1.000000	Κεντρικό αέριο	1983.734378	380.770000	168151.0	8.807542
9	Διαμέρισμα	100000.0	69.000000	Αθήνα	2.000000	1.000000	3.000000	Κεντρικό αέριο	1970.000000	202.101000	1002212.0	14.683246
10	Μονοκατοικία	285000.0	123.937619	Αθήνα - Ανατολικά Προάστια	3.000000	1.000000	2.237179	Αυτόνομο πετρέλαιο	2005.000000	363.013286	168151.0	7.600195
11	Διαμέρισμα	113000.0	70.000000	Θεσσαλονίκη	2.000000	1.000000	1.000000	Αυτόνομο αέριο	1983.734378	205.030000	319045.0	15.475094
12	Διαμέρισμα	230000.0	115.000000	Θεσσαλονίκη	2.000000	1.000000	2.000000	Αυτόνομο αέριο	1990.000000	336.835000	319045.0	15.475094
13	Διαμέρισμα	130000.0	85.000000	Αθήνα	2.000000	1.000000	1.000000	Κεντρικό πετρέλαιο	1965.000000	248.965000	1002212.0	14.683246
14	Διαμέρισμα	140000.0	72.000000	Αθήνα	1.000000	1.000000	3.000000	Κεντρικό πετρέλαιο	1975.000000	210.888000	1002212.0	14.683246
15	Διαμέρισμα	250000.0	76.000000	Αθήνα - Νότια Προάστια	2.000000	1.000000	5.000000	Αυτόνομο πετρέλαιο	2010.000000	222.604000	168151.0	7.756959
16	Διαμέρισμα	170000.0	50.000000	Θεσσαλονίκη	1.000000	1.000000	2.000000	Αυτόνομο αέριο	1972.000000	146.450000	319045.0	15.475094
17	Μεζονέτα	850000.0	123.937619	Αθήνα - Βόρεια Προάστια	4.000000	1.433769	2.237179	Αυτόνομο πετρέλαιο	1983.734378	363.013286	168151.0	8.807542
18	Μονοκατοικία	500000.0	123.937619	Αθήνα - Ανατολικά Προάστια	2.000000	1.433769	2.237179	Αυτόνομο πετρέλαιο	1983.734378	363.013286	168151.0	7.600195
19	Μεζονέτα	420000.0	123.937619	Αθήνα - Ανατολικά Προάστια	3.000000	1.000000	2.237179	Αυτόνομο πετρέλαιο	1994.000000	363.013286	168151.0	7.600195
20	Διαμέρισμα	200000.0	57.000000	Θεσσαλονίκη	2.000000	1.000000	2.000000	Αυτόνομο αέριο	1960.000000	166.953000	319045.0	15.475094
21	Διαμέρισμα	950000.0	190.000000	Αθήνα - Βόρεια Προάστια	3.000000	2.000000	1.000000	Κεντρικό πετρέλαιο	1990.000000	556.510000	168151.0	8.807542
22	Διαμέρισμα	180000.0	85.000000	Θεσσαλονίκη	2.000000	1.000000	3.000000	Αυτόνομο αέριο	1975.000000	248.965000	319045.0	15.475094
23	Διαμέρισμα	97000.0	123.937619	Θεσσαλονίκη Περ/κοι Δήμοι	1.000000	1.433769	2.237179	Αυτόνομο αέριο	1983.734378	363.013286	784978.0	38.074906
24	Διαμέρισμα	87000.0	53.000000	Αθήνα	1.000000	1.000000	2.237179	Αυτόνομο αέριο	2017.000000	155.237000	1002212.0	14.683246
25	Μεζονέτα	490000.0	260.000000	Αθήνα - Ανατολικά Προάστια	4.000000	2.000000	1.000000	Αυτόνομο αέριο	1983.734378	761.540000	168151.0	7.600195
26	Μεζονέτα	355000.0	123.937619	Αθήνα - Δυτικά Προάστια	2.000000	1.000000	5.000000	Αυτόνομο αέριο	1983.734378	363.013286	168151.0	7.016037
27	Μεζονέτα	510000.0	147.000000	Αθήνα - Ανατολικά Προάστια	3.000000	2.000000	1.000000	Αυτόνομο αέριο	1983.734378	430.563000	168151.0	7.600195
28	Μονοκατοικία	260000.0	123.937619	Αθήνα - Ανατολικά Προάστια	2.198839	1.433769	2.237179	Αυτόνομο αέριο	1983.734378	363.013286	168151.0	7.600195

***Πίνακας 7**: Ενδεικτική δομή δεδομένων τελικού dataset ακινήτων*

Διαχωρισμός στοιχείων ακινήτου και στοιχείων Γειτονιάς ακινήτου

Αυτά είναι όλα τα στοιχεία τα οποία συλλέχθηκαν και κατέληξαν για να χρησιμοποιηθούν στα τελικά μοντέλα που θα εκπαιδεύσουμε για την πρόβλεψη τιμών ακινήτων.

Από τα δεδομένα τα οποία διατηρήθηκαν κάποια χαρακτηρίζουν το ίδιο το ακίνητο και κάποια την περιοχή στην οποία βρίσκεται. Κάθε ένας από τους άξονες αυτούς αποτελεί ένα σημαντικό κομμάτι της τιμής. Δεν αρκεί όμως από μόνο του για να προσδιορίσει ικανοποιητικά την τιμή.

Αυτός είναι και ένας λόγος για τον οποίο η συλλογή ακόμα περισσότερων δεδομένων για το περιβάλλον ενός ακινήτου, έχει διαρκώς μικρότερη αξία. Έχοντας ήδη μια εικόνα για τις παροχές κοντά στο ακίνητο, την τιμή των άλλων ακινήτων στην περιοχή, το πληθυσμό της πόλης του ακινήτου κλπ. Έχουμε μια καλή προσέγγιση της αξίας περιοχής του ακινήτου. Παραπάνω εξωτερικά στοιχεία δεν θα συμβάλλουν καθοριστικά στον προσδιορισμό της τιμής του η οποία εξαρτάται πλέον κυρίως από εσωτερικά του χαρακτηριστικά όπως τα κατοικήσιμα τετραγωνικά μέτρα και τα δωμάτια.

Δυσκολίες που αντιμετωπίσαμε

Στη συνέχεια, θα αναφερθούμε στα προβλήματα τα οποία αντιμετωπίσαμε, κατά την πραγματοποίηση της παρούσας εργασίας, και συγκεκριμένα σε προβλήματα που αφορούν την συλλογή των δεδομένων και την κατασκευή του συνόλου δεδομένων. Το κύριο πρόβλημα που συναντήσαμε, αφορούσε την «ποιότητα» των δεδομένων, τα οποία είχαμε την δυνατότητα να εντοπίσουμε και να συλλέξουμε. Αρχικά, τα ακίνητα και οι περιγραφές τους στον ιστότοπο της Χρυσής Ευκαιρίας, καταχωρούνται και ενημερώνονται από τους χρήστες, χωρίς αυστηρούς κανόνες και μορφοποιήσεις, αλλά και χωρίς να γίνεται κάποιος έλεγχος για την εγκυρότητα των στοιχείων των ακινήτων. Αυτό σημαίνει πως ο καθένας μπορεί να δημοσιεύσει κάποιο ακίνητο προς πώληση, με οποιαδήποτε στοιχεία αυτός επιθυμεί, ακόμα και αν δεν συμπίπτουν με την ιδιοκτησία. Τέτοιες περιπτώσεις ακινήτων των οποίων η περιγραφή είναι σε ελλιπής, συναντήσαμε σε αρκετά μεγάλο αριθμό καταχωρήσεων. Τα χαρακτηριστικά των

καταχωρήσεων αυτών, όπου ήταν δυνατόν διορθώσαμε, παρόλα αυτά υπήρξαν πολλές εγγραφές που διαγράφηκαν από το σύνολο δεδομένων εξαιτίας των ελλιπών τους στοιχείων. Ακόμα, κάθε χρήστης μπορεί να δημοσιεύσει το ακίνητο που επιθυμεί, περισσότερες εκ της μίας φορές, γεγονός που οδήγησε στο να συλλέξουμε μεγάλο ποσοστό διπλοτύπων, κάτι που καθυστέρησε ακόμα περισσότερο την δημιουργία του συνόλου δεδομένων. Επίσης, η μη αναφορά σημαντικών μεταβλητών στις αγγελίες των ακινήτων, οι οποίες έχουν υψηλό ποσοστό επιρροής στην τιμή τους, όπως είναι το μέσον θέρμανσης του ακινήτου, το έτος κατασκευής του ακινήτου επιδρούν αρνητικά στην ακρίβεια των προβλέψεων. Εδώ θα αναφερθούμε για ακόμα μία φορά στην δυσκολία συλλογής των δεδομένων που χρησιμοποιήθηκαν, διαδικασία η οποία κρίθηκε πολύ χρονοβόρα και για το λόγο αυτό ο όγκος των δεδομένων μας δεν είναι ο επιθυμητός. Αντί αυτού, είναι σημαντικά λιγότερες, από ότι αρχικά υπολογίζαμε, οι είσοδοι του συνόλου δεδομένων μας. Το γεγονός αυτό έχει σημαντική επίπτωση στην απόδοση των μοντέλων που χρησιμοποιήσαμε και στον αριθμό των χαρακτηριστικών που τελικά λήφθηκαν υπόψιν κατά την εκπαίδευση των μοντέλων.

Ορισμένα από τα αναφερθέντα προβλήματα θα μπορούσαν να βελτιωθούν με την καλύτερη οργάνωση των ακινήτων που δημοσιεύονται, την εισαγωγή πιο αυστηρών κανόνων για τους χρήστες της ιστοσελίδας και την απαίτηση ενός σεβαστού αριθμού χαρακτηριστικών τα οποία θα αποτελούν απαραίτητη προϋπόθεση, προκειμένου να δημοσιευθεί μία ιδιοκτησία προς πώληση.

3.3.3 Μετασχηματισμοί για συμβατότητα με τα μοντέλα μηχανικής μάθησης

Η βιβλιοθήκη scikit-learn προσφέρει ορισμένες κλάσεις οι οποίες καθιστούν σημαντικά ευκολότερη την επεξεργασία στηλών οι οποίες έχουν κατηγορικά δεδομένα ή δεδομένα σε μορφή string. Επιπλέον, διαθέτει κλάσεις για το αυτόματο scaling των στηλών ώστε μία στήλη με πολύ μεγάλες τιμές να μην επισκιάσει τις υπόλοιπες που έχουν μικρότερες τιμές.

Οι μετασχηματισμοί αυτοί δεν απαιτείται να εφαρμοστούν σε όλα τα μοντέλα (για παράδειγμα τα τυχαία δάση λειτουργούν καλά και χωρίς feature scaling), ωστόσο, προκειμένου να υπάρχει ομοιομορφία στις εισόδους κάθε μοντέλου θα χρησιμοποιηθούν για όλα τα μοντέλα.

Robust Scaler

Ένα άλλο πρόβλημα που μπορεί να προκύψει σε προβλήματα μηχανικής μάθησης είναι όταν είσοδοι με μεγάλο εύρος τιμών επισκιάζουν μεταβλητές με μικρότερο εύρος. Προκειμένου να αποφευχθεί αυτό, τα αριθμητικά δεδομένα περνούν από μια διαδικασία κλιμάκωσης ώστε να έχουν όλα παρόμοια εύρη.

Υπάρχουν διάφοροι τρόποι να γίνει η διαδικασία αυτή. Μία από τις πλέον χρήσιμες κλάσεις για αυτό τον είναι η κλάση Robust Scaler της βιβλιοθήκης skikit-learn. Κύριο χαρακτηριστικό της είναι η ανθεκτικότητα σε ακραίες τιμές.

Λειτουργεί καταργώντας τη διάμεση τιμή και κλιμακώνοντας τα δεδομένα σύμφωνα με το εύρος μεταξύ του 1^{ου} τεταρτημόριου (25ο quantile) και του 3^{ου} τεταρτημόριου (75ο quantile).

Η τυποποίηση ενός συνόλου δεδομένων είναι μια κοινή απαίτηση για πολλούς εκτιμητές μηχανικής μάθησης. Συνήθως αυτό γίνεται με την αφαίρεση του μέσου όρου και την κλιμάκωση σε διακύμανση μονάδας. Ωστόσο, οι ακραίες τιμές μπορούν συχνά να επηρεάσουν τον μέσο όρο/διακύμανση του δείγματος με αρνητικό τρόπο. Σε τέτοιες περιπτώσεις, η διάμεσος και το εύρος μεταξύ των παραπάνω quantiiles συχνά δίνουν καλύτερα αποτελέσματα.

One-Hot – (Dummy) Encoding

Το one-hot encoding είναι μία τεχνική κωδικοποίησης η οποία χρησιμοποιείται στην μηχανική μάθηση για κατηγορικές μεταβλητές. Επιτρέπει στις μεταβλητές να κωδικοποιηθούν με τρόπο ο οποίος δεν υπονοεί την ύπαρξη κάποιας διάταξης στις τιμές των μεταβλητών. Για παράδειγμα, εάν οι τιμές μιας στήλης ήταν ‘black’, ‘white’, ‘rose’ το πιθανότερο είναι πως δεν υπάρχει κάποια σχέση διάταξης μεταξύ τους ώστε να κωδικοποιηθούν ως 1,2,3.

Ο τρόπος με τον οποίο κωδικοποιεί το one-hot encoding μια μεταβλητή είναι μέσω της κατασκευής ενός αραιού ή πυκνού πίνακα (ανάλογα με τους περιορισμούς μνήμης) στον οποίο κάθε πιθανή τιμή της κατηγορικής μεταβλητής έχει και από μία ξεχωριστή στήλη. Έτσι κάθε εγγραφή έχει τιμή 1 σε μία από τις παραγόμενες αυτές στήλες και 0 στις υπόλοιπες επιτρέποντας την εύκολη, μολονότι χρονοβόρα σε μνήμη, επεξεργασία από ένα μοντέλο μηχανικής μάθησης.

Στα παρόντα δεδομένα οι κατηγορικές μεταβλητές είναι πολύ περισσότερες από τις βασικές μεταβλητές όπου στην περίπτωσή μας είναι (37) και κάποιες από αυτές έχουν μεγάλο εύρος στηλών (π.χ. Heating = 16), επομένως το one-hot encoding αποτελεί χρήσιμή λύση στο πρόβλημα της αναπαράστασης τους. Για την κωδικοποίηση χρησιμοποιείται ο Dummy encoder της βιβλιοθήκης skikit-learn.

Column Transformer

Προκειμένου να εφαρμοστούν οι παραπάνω μετασχηματισμοί στα δεδομένα με ευκολία και ταχύτητα, γίνεται χρήση της κλάσης Column Transformer της βιβλιοθήκης skikit-learn. Η κλάση αυτή επιτρέπει την αλυσιδωτή εφαρμογή μετασχηματιστών σε συγκεκριμένες στήλες ενός dataset.

Με την χρήση της εφαρμόζονται οι παραπάνω μετασχηματισμοί στις συμβατές στήλες του training set.

Hypertuning – Εύρεση υπερ-παραμέτρων

Το Hypertuning αποτελεί μια μέθοδο εύρεσης υπερ-παραμέτρων, δηλαδή εύρεσης των παραμέτρων εκείνων οι οποίες βελτιστοποιούν την απόδοση και την ικανότητα πρόβλεψης ενός μοντέλου μηχανικής μάθησης. Στην γλώσσα προγραμματισμού Python, με την οποία ασχολούμαστε υπάρχουν διάφορες τεχνικές εύρεσης βέλτιστων παραμέτρων, από τις οποίες οι σημαντικότερες είναι η μέθοδος GridSearchCV, αλλά και η μέθοδος RandomizedSearchCV. Εμείς στα πλαίσια της εργασίας μας χρησιμοποιήσαμε τον αλγόριθμο RandomizedSearchCV. Η κεντρική ιδέα αυτού του αλγορίθμου – τεχνικής

εύρεσης υπερ-παραμέτρων είναι να βρούμε τις παραμέτρους εκείνους ενός μοντέλου μηχανικής μάθησης που επηρεάζουν σε μεγαλύτερο βαθμό και με ποιον τρόπο τις προβλέψεις του, παραθέτοντας τυχαία και όχι με συγκεκριμένη σειρά τα σύνολα των παραμέτρων αυτών στον αλγόριθμο. Έτσι, ο αλγόριθμος υπολογίζει τα σκορ των διαφορετικών συνόλων υπερ-παραμέτρων και τελικά μας επιστρέφει τις παραμέτρους που μας δίνουν τα καλύτερα αποτελέσματα και έχουν το καλύτερο σκορ (Kouate P.M., 2020). Τέλος, ο χρήστης δύναται να επιλέξει τον τρόπο με τον οποίο θα εισάγει τις παραμέτρους με τι εύρος και το πώς τελικά θα διαμορφώσει τα σύνολα από τα οποία θα εργαστεί εν συνεχεία ο αλγόριθμος.

3.4 Στατιστική - Ανάλυση δεδομένων

Στόχος του παρόντος κεφαλαίου αποτελεί η διερευνητική ανάλυση του συνόλου δεδομένων, προκειμένου να ληφθούν όλες οι δυνατές πληροφορίες που μπορούν να εξαχθούν από αυτό. Το συγκεκριμένο κρίνεται καίριο βήμα στην διαδικασία της έρευνας, και προτού μοντελοποιηθούν τα δεδομένα, καθώς, όπως αναφέρει και ο Tukey (1977), μεγάλης σημασίας αποτελεί η κατανόηση των όσων έχουμε πραγματικά την δυνατότητα να καταφέρουμε, πριν αξιολογήσουμε το πόσο καλά τα καταφέραμε. Η ανάλυση του συγκεκριμένου συνόλου δεδομένων έγινε σε γλώσσα προγραμματισμού Python. Στη συνέχεια ακολουθούνται τεχνικές και μέθοδοι μονομεταβλητής και πολυμεταβλητής ανάλυσης και απεικόνισης των δεδομένων και παρουσιάζονται τα αποτελέσματα.

3.4.1 Έλεγχος Στατιστικών Υποθέσεων

Τα προγενέστερα βήματα που ακολουθήθηκαν, της διαχείρισης των κενών, των διπλότυπων και ακραίων τιμών, επιχείρησαν να «καθαρίσουν» τα δεδομένα που έχουμε στη διάθεσή μας και να τα μετατρέψουν σε μία μορφή περισσότερο κατάλληλη για πολυμεταβλητή ανάλυση. Ο έλεγχος, τώρα, των δεδομένων,

προκειμένου να εξεταστούν οι στατιστικές υποθέσεις που αφορούν τις τεχνικές πολυμεταβλητής ανάλυσης, ασχολείται με τις βάσεις που πρέπει να υπάρχουν και στις οποίες θα στηριχτεί η στατιστική ανάλυση. Όπως αναφέρει και ο Hair (2009), η ανάγκη του ελέγχου των στατιστικών υποθέσεων στις περιπτώσεις όπου στην συνέχεια θα εφαρμοστεί πολυμεταβλητή ανάλυση, όπως στην περίπτωση του πλαισίου της συγκεκριμένης μελέτης, αυξάνεται εξαιτίας δύο χαρακτηριστικών της. Πρώτον, η πολυπλοκότητα των σχέσεων μεταξύ των μεταβλητών, η οποία αποτελεί απόρροια της, τυπικά, χρήσης πολλών μεταβλητών, αυξάνει την πιθανότητα ύπαρξης στρεβλώσεων και προκαταλήψεων μεταξύ των δεδομένων. Δεύτερον, η πολυπλοκότητα της ανάλυσης και των αποτελεσμάτων, ενδέχεται να υποκρύψει τους δείκτες οι οποίοι προμηνύουν παραβιάσεις των υποθέσεων, οι οποίες είναι πιθανότατα πιο εμφανείς στην απλούστερη μονομεταβλητή ανάλυση. Εν πάση περιπτώσει, τα μοντέλα πολυμεταβλητής ανάλυσης θα υπολογίσουν αποτελέσματα ακόμα και όταν οι στατιστικές υποθέσεις οι οποίες αφορούν τα δεδομένα δεν ικανοποιούνται. Για το λόγο αυτό, ο ερευνητής είναι σημαντικό να λαμβάνει υπόψιν του αυτές τις παραβιάσεις και τις επιπτώσεις που μπορεί να έχουν στην διαδικασία εκτίμησης και την ερμηνεία των αποτελεσμάτων. Στην συνέχεια του κεφαλαίου θα εξεταστούν οι στατιστικές υποθέσεις της γραμμικότητας και της απουσίας συσχετισμένων σφαλμάτων κ.α.

3.4.2 Στατιστική Ανάλυση & Απεικόνιση Μεταβλητής Στόχου

Η μεταβλητή που μας ενδιαφέρει και την οποία θέλουμε αρχικά να διερευνήσουμε, είναι η μεταβλητή ‘Price’, η οποία μας δίνει την τιμή του κάθε ακινήτου και είναι αυτή που θα μελετηθεί αρχικά. Παρακάτω φαίνονται τα χαρακτηριστικά και το ιστόγραμμα της μεταβλητής.

```
count     20996.000000
mean     303292.809899
std      214246.925402
min       20000.000000
25%      135000.000000
50%      250000.000000
75%      410000.000000
max      998000.000000
Name: Price, dtype: float64
```

***Πίνακας 8**: Περιγραφική Στατιστική Περίληψη Μεταβλητής Τιμών*

Ο παραπάνω πίνακας απεικονίζει μια περιγραφή της τιμής στόχου μας καθώς και ότι η ελάχιστη τιμή είναι μεγαλύτερη από το μηδέν. Αυτό είναι εξαιρετικό! Ας συνεχίσουμε με ένα απλό ιστόγραμμα.

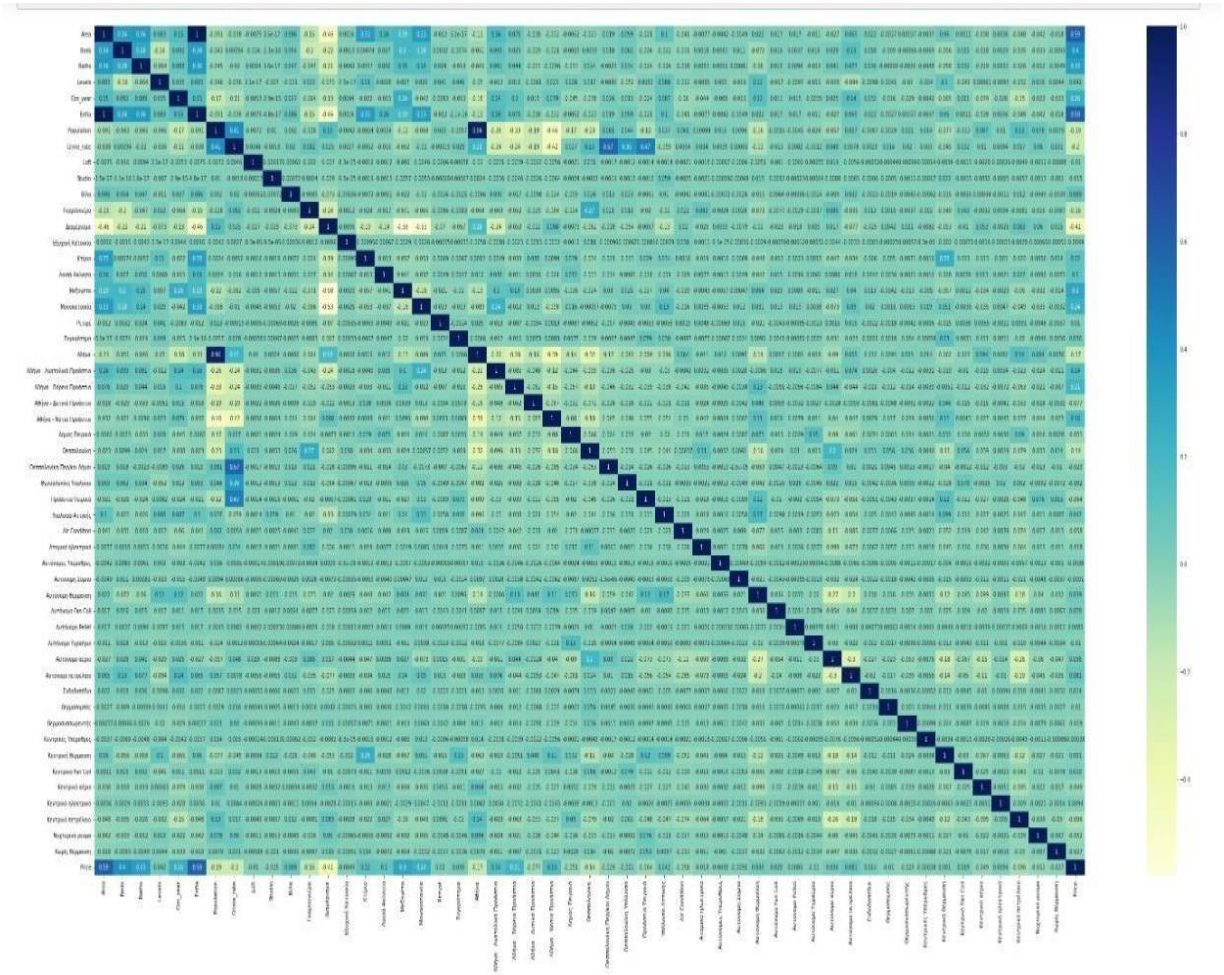
Ιστογράμματα

Τα ιστογράμματα τα κατασκευάζουμε για τις μεταβλητές του πλαισίου δεδομένων, στην περίπτωσή μας, μόνο για την τιμή (Price) και μας παρέχουν πληροφορίες σχετικά με όλα τα βασικά στατιστικά τους όπως το εύρος τιμών και η επικρατούσα τιμή, αλλά και το σημαντικότερο την κατανομή τους. Επίσης, διαφαίνονται πιθανές ακραίες τιμές αλλά και μεγάλες συγκεντρώσεις τιμών. Ουσιαστικά, όλες οι τιμές που παίρνει η μεταβλητή ομαδοποιούνται σε ορθογώνιες μπάρες ανάλογα με ένα καθορισμένο εύρος τιμών το οποίο παρουσιάζεται από το ύψος της κάθε μπάρας (Math is Fun, 2019).

Εικόνα 14: *Ιστόγραμμα Μεταβλητής Τιμής*

Heatmap

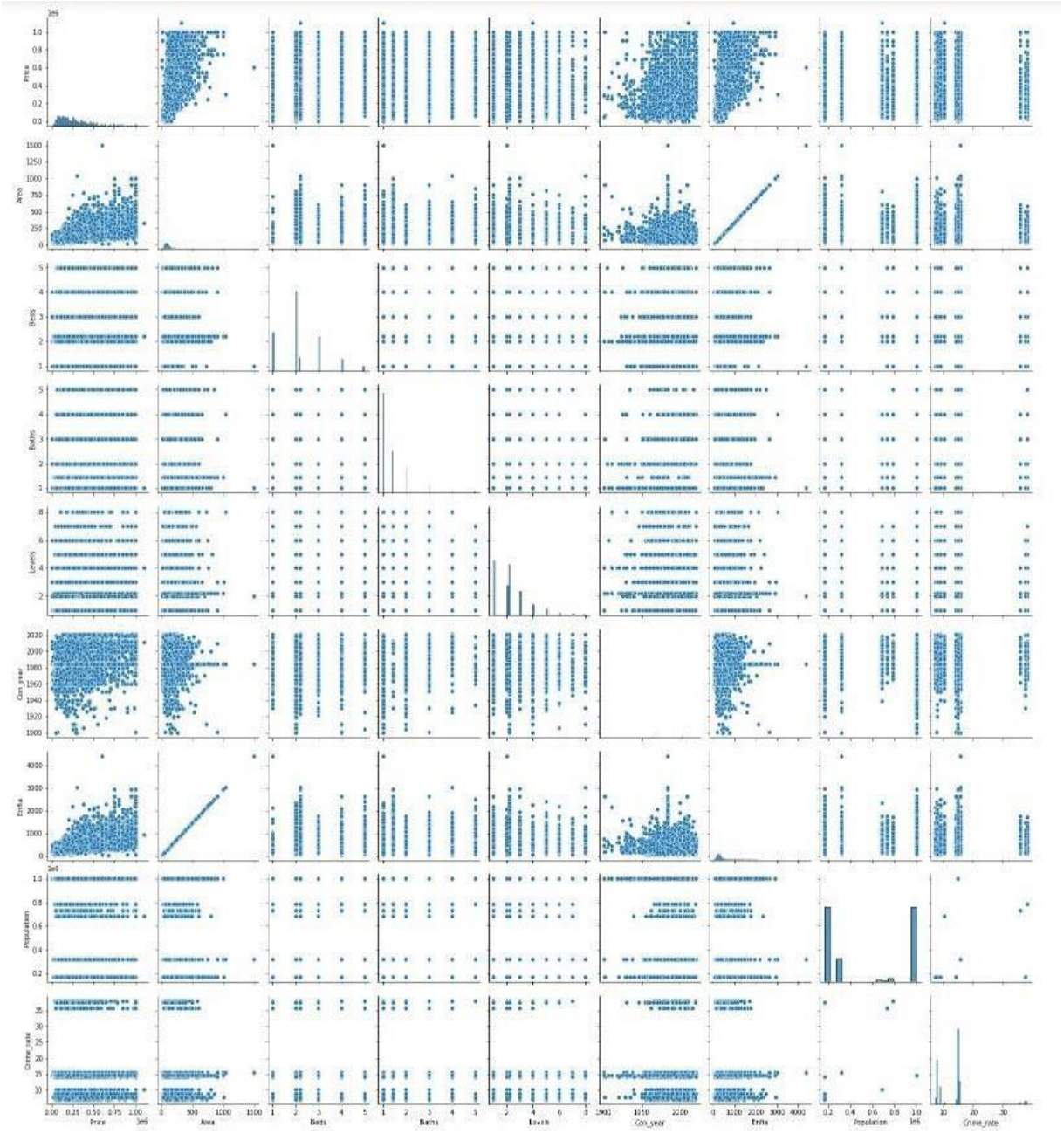
Ο Heatmap αποτελεί έναν πολύ σημαντικό πίνακα ή χάρτη ο οποίος χρησιμοποιείται για να υπολογίσουμε σε ποσοστά τις συσχετίσεις μεταξύ διαφορετικών μεταβλητών και τις αναπαριστά με γραφικό και ζωντανό τρόπο (Wikipedia, 2022). Τα θετικά ποσοστά (ή δεκαδικοί αριθμοί) μεταφράζονται ως θετικές συσχετίσεις, που σημαίνει ότι υπάρχει εξάρτηση - σχέση μεταξύ των δύο χαρακτηριστικών και είναι θετικό αποτέλεσμα. Αν η συσχέτιση είναι ίση με 0 τότε σημαίνει πως δεν υπάρχει καθόλου. Ενώ, αν είναι αρνητικό το ποσοστό, τότε, η αρνητική συσχέτιση έχει αρνητικό αποτέλεσμα και οι 2 μεταβλητές δεν έχουν πολύ εξάρτηση μεταξύ τους. Όπου υπάρχει μεγάλη θετική συσχέτιση υπάρχει και περισσότερη πληροφορία η οποία παρέχεται από τις εμπλεκόμενες μεταβλητές (Sanat S., 2018).



Εικόνα 15: *Heatmap*

Διάγραμμα ζευγών συσχετίσεων - Pairplot

Παρόμοια με τον heatmap το pairplot δημιουργείται για την ένδειξη συσχετίσεων μεταξύ ζευγών μεταβλητών, αλλά αυτή τη φορά με την αναπαράσταση σημείων και όχι ποσοστών. Δηλαδή, δημιουργούνται διαγράμματα σημείων - scatterplot σε καρτεσιανό σύστημα συντεταγμένων για κάθε πιθανό ζεύγος μεταβλητών όπου στους 2 άξονες βρίσκονται οι 2 μεταβλητές. Και εδώ μπορούμε να παρατηρήσουμε το γράφημα και να εξάγουμε και άλλα συμπεράσματα.

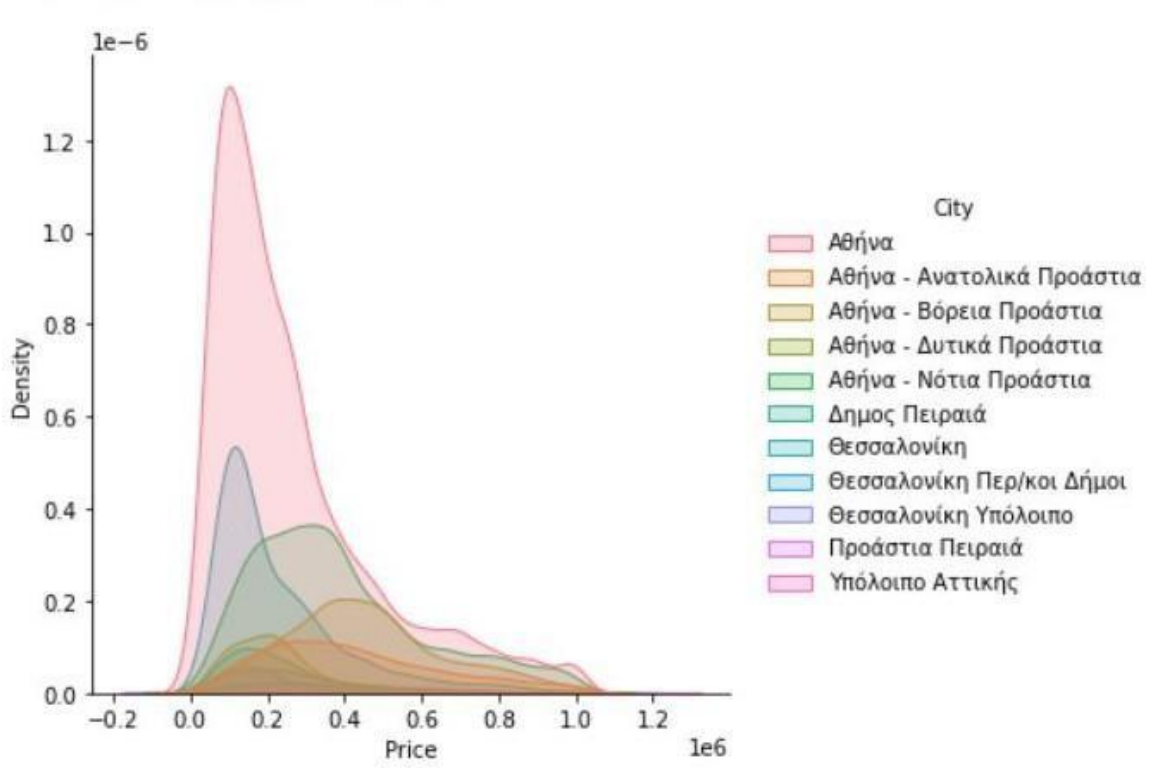


***Εικόνα 16**: Pairplot*

Γράφημα πυκνότητας (Density plot)

Το γράφημα πυκνότητας αποτελεί ένα στατιστικό γράφημα που μας παρέχει πληροφορίες σχετικά με την κατανομή μιας αριθμητικής μεταβλητής πάνω σε ένα (άλλο) συνεχές αριθμητικό διάστημα (π.χ. χρόνος). Παρουσιάζει, επίσης, το πού υπάρχει μεγάλη συγκέντρωση τιμών και πού χαμηλότερη, ενώ καθορίζει και τα όρια – εύρος αυτών των τιμών (Wilke C., 2021). Πολλές φορές τα γραφήματα

πυκνότητας είναι χρήσιμα, όσον αφορά την σύγκριση των τιμών μια αριθμητικής μεταβλητής στις διάφορες κατηγορίες μιας ποιοτικής μεταβλητής. Τέλος, ένα πλεονέκτημα των density plots σε σχέση με τα ιστογράμματα αποτελεί το γεγονός ότι το σχήμα των πρώτων παρέχει περισσότερες πληροφορίες σχετικά με την κατανομή που ακολουθείται σε κάθε μέλος. Στο παρακάτω γράφημα έχουμε ένα συγκριτικό γράφημα πυκνότητας των τιμών των ακινήτων (Price) ανάλογα με την περιοχή στην οποία βρίσκονται (City).



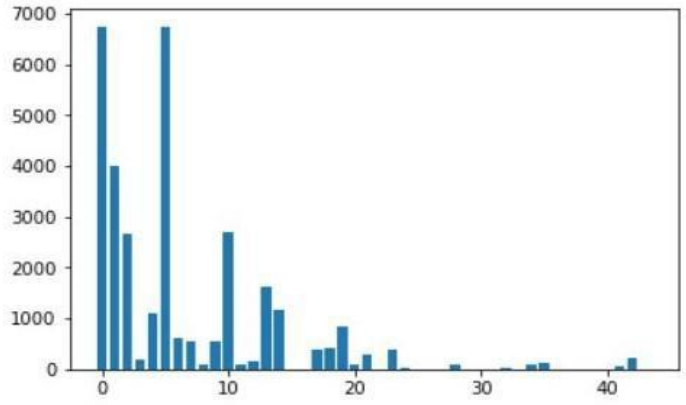
***Εικόνα 17**: Density Plot*

Επιλογή καλύτερων μεταβλητών – Feature Selection

Στους έξι αλγορίθμους παλινδρόμησης που εφαρμόζουμε, κατασκευάζουμε δύο μοντέλα για κάθε αλγόριθμο αντίστοιχα. Το 1ο μοντέλο εκπαιδεύεται και αξιολογείται στα τελικά δεδομένα, τα οποία προκύπτουν με τον τρόπο που παρουσιάσαμε προηγουμένως. Ωστόσο, το 2ο μοντέλο που κατασκευάζεται για κάθε αλγόριθμο εκπαιδεύεται και αξιολογείται σε δεδομένα που προκύπτουν από την τεχνική της εύρεσης των καλύτερων, δηλαδή των πιο αποδοτικών μεταβλητών. Η τεχνική που χρησιμοποιήσαμε είναι η μέθοδος της εύρεσης των καλύτερων μεταβλητών βάσει των καλύτερων συσχετίσεων μεταξύ τους (correlation feature importance method) (Towards Data Science, Bex T., 2021). Η μέθοδος αυτή χρησιμοποιεί, αρχικά, τον αλγόριθμο Select K Best της βιβλιοθήκης sklearn της Python, και στη συνέχεια υπολογίζει τα scores (τις αποδόσεις) των μεταβλητών ανάλογα με το πόσο υψηλή συσχέτιση έχει η κάθε μια μεταβλητή με τις υπόλοιπες και έτσι ελέγχονται όλα τα πιθανά ζεύγη και οι αντίστοιχες συσχετίσεις τους. Εφαρμόζοντας την παραπάνω τεχνική βρίσκουμε ότι οι καλύτερες 20 μεταβλητές είναι οι: Area, Beds, H_type (1.*Γκαρσονιέρα*, 2.*Διαμέρισμα*, 3.*Λοιπά Ακίνητα,* 4. *Μεζονέτα*, 5.*Μονοκατοικία*), Baths, Levels, Con_year, Enfia, Population, Crime_rate, City (1.*Αθήνα*, 2. *Αθήνα - Ανατολικά Προάστια,* 3.*Αθήνα – Βόρεια Προάστια*, 4. *Αθήνα - Νότια Προάστια, 5.Θεσσαλονίκη*) και Heating (1.*Αυτόνομο πετρέλαιο*, 2.*Κεντρικό πετρέλαιο*) οι οποίες μαζί με την μεταβλητή Price δημιουργούν το 2ο σύνολο δεδομένων πάνω στο οποίο κατασκευάζονται τα επόμενα μοντέλα για κάθε αλγόριθμο. Το παραπάνω αποτέλεσμα παρουσιάζεται και από την παρακάτω εικόνα, όπου βρίσκονται οι μεταβλητές και τα αντίστοιχα σκορ τους (Brownlee J., 2019).

```
Feature 0: 6753.097190
Feature 1: 3998.823517
Feature 2: 2662.091800
Feature 3: 173.258628
Feature 4: 1116.689452
Feature 5: 6753.097190
Feature 6: 625.305149
Feature 7: 560.194777
Feature 8: 100.825249
Feature 9: 533.457135
Feature 10: 2711.821855
Feature 11: 81.176744
Feature 12: 163.532761
Feature 13: 1613.219240
Feature 14: 1165.789673
Feature 15: 0.177612
Feature 16: nan
Feature 17: 373.496014
Feature 18: 406.365133
Feature 19: 827.787140
Feature 20: 82.612251
Feature 21: 291.419006
Feature 22: nan
Feature 23: 392.987473
Feature 24: 19.237666
Feature 25: 0.359276
Feature 26: nan
Feature 27: nan
Feature 28: 80.812234
Feature 29: 8.169975
Feature 30: 1.213025
Feature 31: nan
Feature 32: 16.434264
Feature 33: 2.499367
Feature 34: 79.957848

Feature 35: 135.150335
Feature 36: 2.871374
Feature 37: 2.222181
Feature 38: 5.547501
Feature 39: nan
Feature 40: 8.039037
Feature 41: 47.998981
Feature 42: 219.795963
Feature 43: 4.999259
```




```
Best features are: Area, Beds, Baths, Levels, Con_year, Enfia, Population, Crime_rate,
H_type(1.Γκαρσονιέρα, 2.Διαμέρισμα, 3.Λοιπά Ακίνητα 4.Μεζονέτα, 5.Μονοκατοικία, City(1.Αθήνα,
2.Αθήνα - Ανατολικά Προάστια, 3.Αθήνα - Βόρεια Προάστια, 4.Αθήνα - Νότια Προάστια
5.Θεσσαλονίκη, Heating(1.Αυτόνομο πετρέλαιο, 2.Κεντρικό πετρέλαιο
```

***Εικόνα 18**: Feature Selection Plot*

Το νέο dataset, λοιπόν, που προκύπτει έπειτα από την εφαρμογή της τεχνικής εύρεσης καλύτερων μεταβλητών ανάλογα με την συσχέτιση που έχουν, αποτελείται από 20996 παρατηρήσεις – γραμμές και από 21 στήλες – μεταβλητές, που είναι οι εξής: 'Area', 'Beds', 'Baths', 'Levels', 'Con_year', 'Enfia', 'Population', 'Crime_rate', 'Γκαρσονιέρα', 'Διαμέρισμα', 'Λοιπά Ακίνητα', 'Μεζονέτα', 'Μονοκατοικία', 'Αθήνα', 'Αθήνα - Ανατολικά Προάστια', 'Αθήνα – Βόρεια Προάστια', 'Αθήνα - Νότια Προάστια', 'Θεσσαλονίκη', 'Αυτόνομο πετρέλαιο', 'Κεντρικό πετρέλαιο' και 'Price'.

Το 3ο και το 4ο μοντέλο προκύπτουν μετά την εύρεση υπερ-παραμέτρων για τον κάθε αλγόριθμο (Hypertuning) και στα δεδομένα του 1ου και του 2ου μοντέλου αντίστοιχα.

Να σημειώσουμε, εδώ, πως μετά από λεπτομερή μελέτη των γραφημάτων που κατασκευάσαμε στην Στατιστική ανάλυση δεδομένων, και συγκεκριμένα στα γραφήματα Pairplot και Heatmap, μπορούμε να εξάγουμε το συμπέρασμα πως τα αποτελέσματα αυτών συμφωνούν με τα αποτελέσματα της μεθόδου επιλογής μεταβλητών.

Γραμμικότητα

Η σιωπηρή υπόθεση όλων των πολυμεταβλητών μεθόδων, οι οποίες βασίζονται σε συσχετιστικά μέτρα σύνδεσης, συμπεριλαμβανομένης και της πολλαπλής παλινδρόμησης, είναι η γραμμικότητα. Οι συσχετίσεις των μεταβλητών αντικατοπτρίζουν μόνο τη γραμμική συσχέτιση μεταξύ των μεταβλητών και συνεπώς οποιαδήποτε μη γραμμική σχέση δεν εξηγείται από αυτές. Αυτή η παράλειψη οδηγεί σε μη – ακριβή εκτίμηση, και μάλιστα υποεκτίμηση, της πραγματικής ισχύος των μεταξύ τους σχέσεων. Είναι συνετό, να εξετασθούν όλες οι σχέσεις μεταξύ των μεταβλητών, προκειμένου να αναγνωριστούν αποκλίσεις από τη γραμμικότητα, που ενδεχομένως να επηρεάσουν το αποτέλεσμα.

Στην παρούσα εργασία, όπως φαίνεται και από τα διαγράμματα διασποράς που έχουν παρατεθεί παραπάνω, οι σχέσεις μεταξύ των μεταβλητών μας παρουσιάζουν γραμμική συμπεριφορά, και για το λόγο αυτό δεν θα χρειαστεί να προχωρήσουμε σε περαιτέρω μετασχηματισμούς των δεδομένων μας,

προκειμένου να ικανοποιήσουμε την υπόθεση της γραμμικότητας. Συνεχίζουμε με την μελέτη της απουσίας συσχετισμένων σφαλμάτων.

Απουσία Συσχετισμένων Σφαλμάτων

Οι προβλέψεις που προκύπτουν από την εφαρμογή οποιασδήποτε μεθόδου, η οποία στηρίζεται στην εξάρτηση των δεδομένων, σε σπάνιες περιπτώσεις είναι αλάνθαστες, παρόλα αυτά οφείλουμε να προσπαθήσουμε για την εξασφάλιση της μη συσχέτισης των σφαλμάτων που θα προκύψουν. Εδώ, αναζητούμε μοτίβα που μπορεί να παρουσιάζουν τα σφάλματα, τα οποία μαρτυρούν συστηματικές σχέσεις μεταξύ των μεταβλητών, τις οποίες δεν έχουμε ακόμα εντοπίσει. Στην παρούσα διπλωματική εργασία τα δεδομένα συλλέχθηκαν με τέτοιον τρόπο, έτσι ώστε να μπορούμε με αυτοπεποίθηση να ισχυριστούμε ότι αυτός δεν συντέλεσε στην προσθήκη κάποιου συστηματικού σφάλματος, σε κάποιο τμήμα των δεδομένων. Επιπλέον, το γεγονός ότι το σύνολο δεδομένων δεν αποτελείται από δεδομένα μορφής χρονοσειρών, έχει ως συνέπεια την ανεξαρτησία μεταξύ των δεδομένων και άρα την εξασφάλιση της ανεξαρτησίας των σφαλμάτων τους.

Κεφάλαιο 4 - Μοντελοποίηση & Αξιολόγηση Αποτελεσμάτων

4.1 Μοντελοποίηση

4.1.1 Διαχωρισμός των Δεδομένων σε Train, Test και Validation sets

Η μεθοδολογία που χρησιμοποιήθηκε κατά την εκπαίδευση των υποψήφιων μοντέλων μηχανικής μάθησης είναι οι εξής: Καταρχάς, ένα υποσύνολο των δεδομένων (70%) απομακρύνθηκε και φυλάχτηκε ώστε να λειτουργήσει σαν train set, δηλαδή πρώτο έλεγχο της ευστοχίας των αρχικών μοντέλων.

Στη συνέχεια, έγινε ένα δοκιμαστικό run για κάθε μοντέλο προκειμένου να κριθεί η γενική του επίδοση. Ακολούθησε Hypertuning για κάθε μοντέλο προκειμένου να βρεθούν οι βέλτιστες υπερπαράμετροι με τις οποίες θα εκπαιδευτεί πάνω στο validation set (15%) που δημιουργήσαμε ως ενδιάμεσο έλεγχο.

Αφού εκπαιδευτούν όλα τα μοντέλα, επιλέγονται τα πιο αποτελεσματικά και συγκρίνεται η απόδοση τους στο τελικό test set (15%) το οποίο έχει φυλαχτεί μέχρι τη στιγμή για τον σκοπό αυτό στα τελικά μοντέλα.

Η δημιουργία ενός test set το οποίο θα μείνει κρυφό ως το τέλος πριν από οποιαδήποτε άλλη δράση στην εκπαίδευση μοντέλων είναι σημαντική για δύο λόγους:

- ✓ Ο πρώτος είναι ο προφανής λόγος πως τα μοντέλα θα χρειαστεί να αξιολογηθούν από τον τρόπο με τον οποίο μπορούν να γενικεύσουν πάνω σε πραγματικά , άγνωστα κατά την εκπαίδευση δεδομένα. Με τον τρόπο αυτό κρίνεται κατά πόσο εμφανίζεται το σύνηθες πρόβλημα του overfitting, το οποίο συμβαίνει όταν ένα μοντέλο μαθαίνει να αποδίδει πολύ καλά στο training set συγκεκριμένα και χάνει την δυνατότητα γενίκευσης σε νέα δεδομένα.

- ✓ Ο δεύτερος λόγος είναι πως το ανθρώπινο μυαλό σχηματίζει και αυτό μοτίβα με μεγάλη ευκολία και η αξιολόγηση ενός μοντέλου πάνω στο validation set προτού ολοκληρωθεί η διαδικασία εκπαίδευσης μπορεί να οδηγήσει στην απόφαση τροποποίησης του μοντέλου απλά και μόνο για να αποδίδει και στο test set. Αυτό μπορεί να γίνει ακόμα και ακούσια, με τη σκέψη πώς το validation set αποκάλυψε κάποια σημαντική λεπτομέρεια των δεδομένων η οποία δεν ήταν σαφής από το training set. Έτσι όμως, χάνεται το νόημα του validation set, το οποίο είναι ο ενδιάμεσος έλεγχος της απόδοσης των μοντέλων σε πραγματικά τελικό test set. Όταν το μοντέλο βγει σε παραγωγικό περιβάλλον, η απόδοση του θα είναι χαμηλότερη από αυτήν στο test set. Το φαινόμενο αυτό ονομάζεται data snooping bias.

4.1.2 Επίπεδα (layers), Κρυμμένα Επίπεδα (hidden layers) και Παράμετροι του Compiler

Τα επίπεδα και κρυμμένα επίπεδα που χρησιμοποιήθηκαν αρχικά για την δημιουργία του μοντέλου φαίνονται στον παρακάτω πίνακα, στην συνέχεια μετά την παραμετροποίηση (Hypertune) γίνονται οι απαραίτητες αλλαγές στα επίπεδα layers και τα κρυμμένα επίπεδα:

```
simple_nn.add(InputLayer((44,)))
simple_nn.add(Dense(128, 'relu'))
simple_nn.add(Dense(64, 'relu'))
simple_nn.add(Dense(32, 'relu'))
simple_nn.add(Dense(1, 'linear'))
```

Πίνακας 9: *Χαρακτηριστικά Layers*

Οι παράμετροι για τον Compiler είναι:

- ✓ Optimizer → Adam
- ✓ Losses → Mean Squared Error

- ✓ Metrics → Root Mean Squared Error

Προκειμένου να δημιουργηθεί το μοντέλο για κάθε σύνολο δεδομένων, ο αλγόριθμός εκτελέστηκε αρχικά σε 30 epochs.

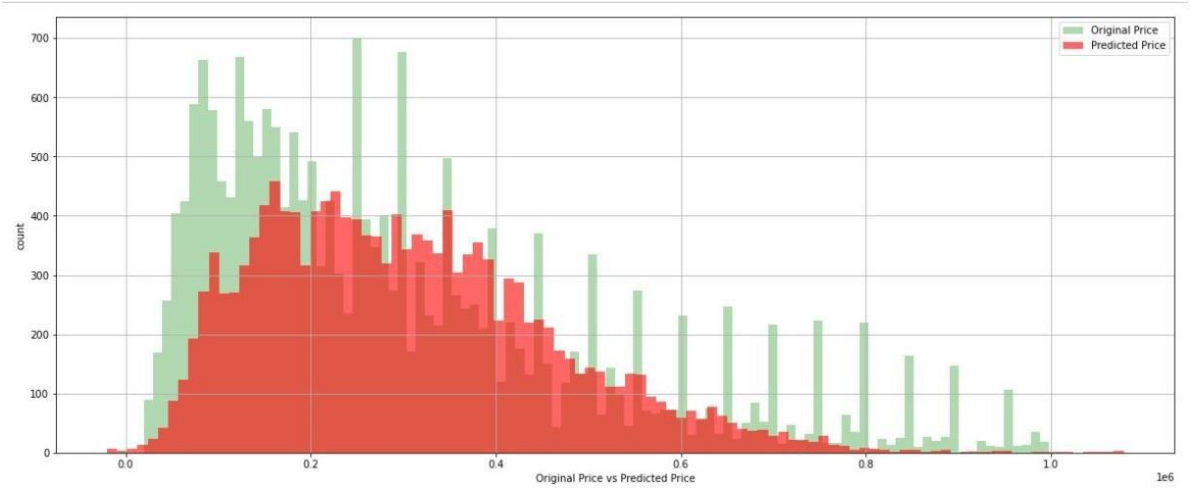
4.2 Αποτελέσματα Αλγορίθμων

Σε αυτήν την ενότητα θα παρουσιαστούν τα αποτελέσματα των μετρικών που αναφέρθηκαν, για τους αλγόριθμους μηχανικής μάθησης που χρησιμοποιήθηκαν για την κατασκευή των μοντέλων, που μας έδωσαν τα βέλτιστα αποτελέσματα. Αυτά τα αποτελέσματα προέκυψαν από την εκπαίδευση των μοντέλων, με το 70% των συνολικών δεδομένων να αποτελεί το σύνολο εκπαίδευσης και το 15% το σύνολο ελέγχου. Στα σύνολα αυτά εφαρμόστηκε μείωση των χαρακτηριστικών, προκειμένου να ληφθούν υπόψιν μόνο οι βέλτιστες παράμετροι και δεν εισήχθησαν ως είσοδοι των μοντέλων όλες οι ανεξάρτητες μεταβλητές. Τα σύνολα που προέκυψαν μετά την περαιτέρω επεξεργασία, όπως και ο τρόπος που οδηγηθήκαμε σε αυτά, παρουσιάζονται, επίσης, στη συνέχεια.

4.2.1 Πολλαπλή Γραμμική Παλινδρόμηση – M.L.R.

Αποτελέσματα 1ου μοντέλου:

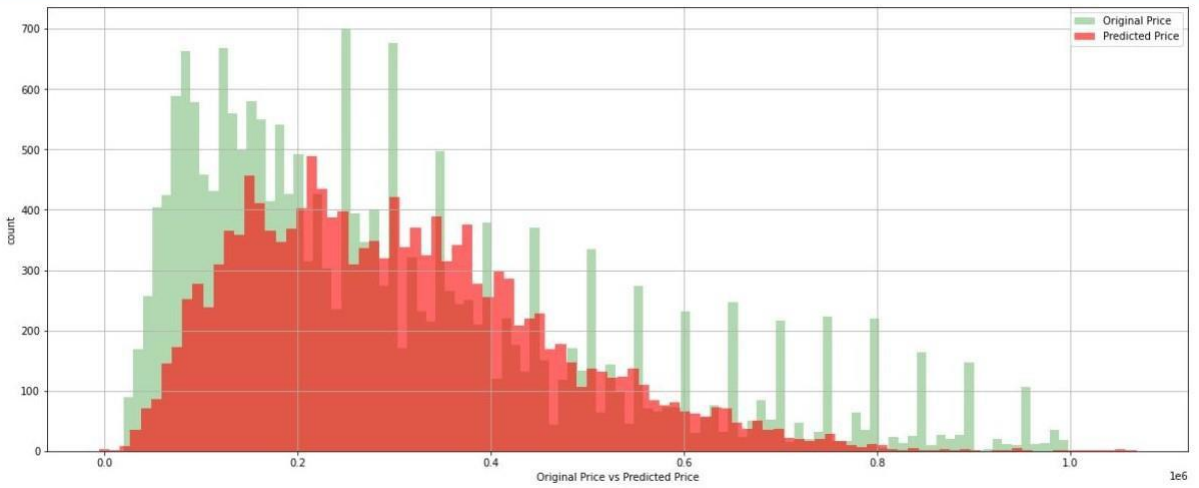
***Εικόνα 19**: Predicted Price vs Original Price (ROC Curve) – MLR1*



***Εικόνα 20**: Predicted Price vs Original Price – MLR1*

Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:

***Εικόνα 21**: Predicted Price vs Original Price (ROC Curve) – MLR2*



***Εικόνα 22**: Predicted Price vs Original Price – MLR2*

Αποτελέσματα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:

```
Fitting 10 folds for each of 48 candidates, totalling 480 fits
{'normalize': True, 'n_jobs': 1, 'fit_intercept': False, 'copy_X': True}
0.5210106502716407
```

***Εικόνα 23**: Το αποτέλεσμα υπολογισμένο με R-squared μετρική – MLR3*

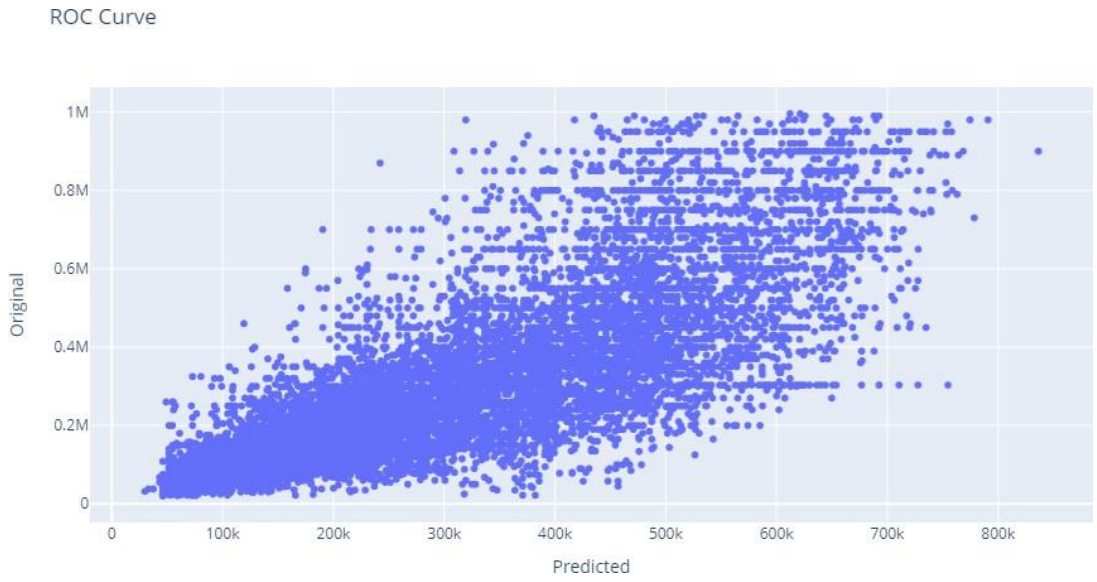
Αποτελέσματα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:

```
Fitting 10 folds for each of 48 candidates, totalling 480 fits
{'normalize': False, 'n_jobs': 1, 'fit_intercept': True, 'copy_X': True}
0.526474041846638
```

***Εικόνα 24**: Το αποτέλεσμα υπολογισμένο με R-squared μετρική – MLR4*

4.2.2 Παλινδρόμηση Χ Ενίσχυσης Κλίσης – X.G.Boosting

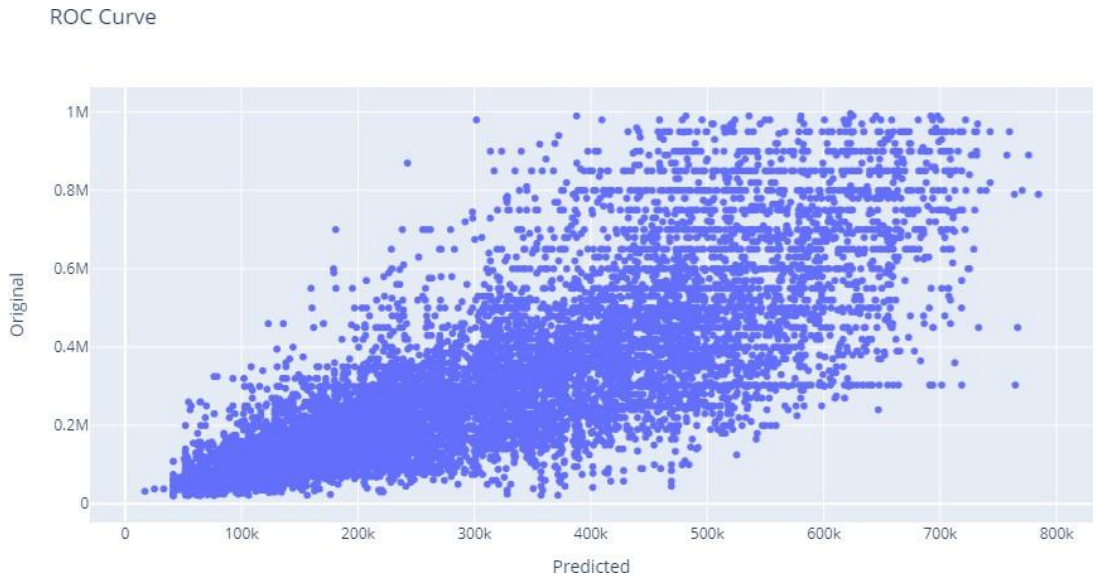
Αποτελέσματα 1ου μοντέλου:



***Εικόνα 25**: Predicted Price vs Original Price (ROC Curve) – Xboosting1*

***Εικόνα 26**: Predicted Price vs Original Price – Xboosting1*

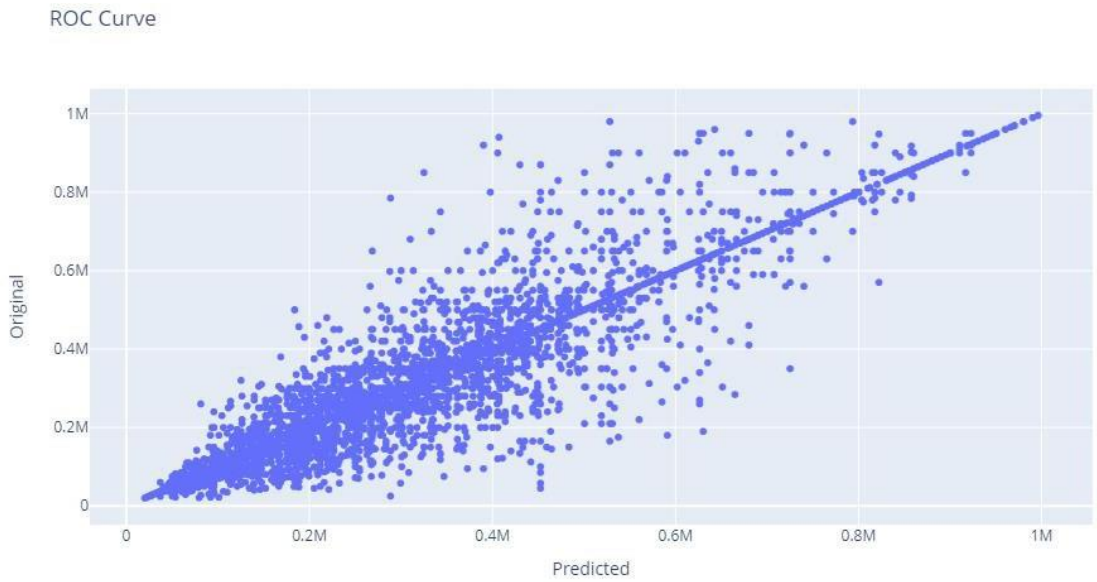
Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:



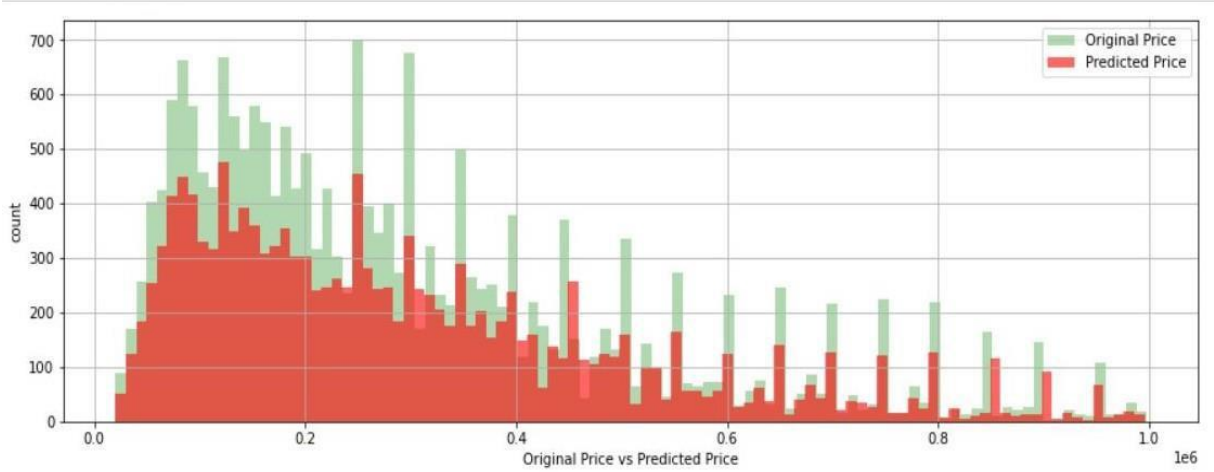
***Εικόνα 27**: Predicted Price vs Original Price (ROC Curve) – Xboosting2*

***Εικόνα 28**: Predicted Price vs Original Price – Xboosting2*

Αποτελέσματα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:

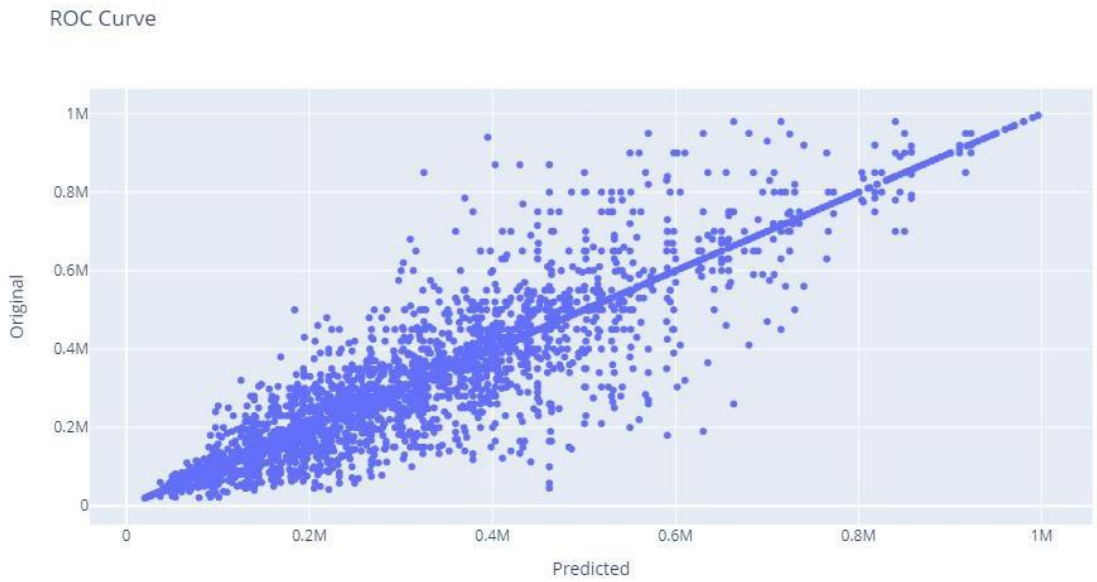


***Εικόνα 29**: Predicted Price vs Original Price (ROC Curve) – Xboosting3*

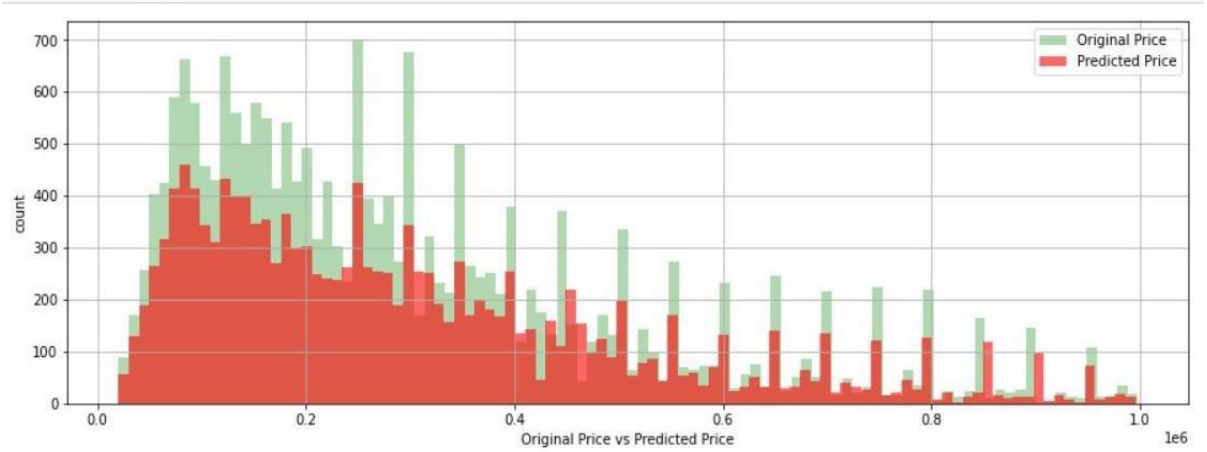


***Εικόνα 30**: Predicted Price vs Original Price – Xboosting3*

Αποτελέσματα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:



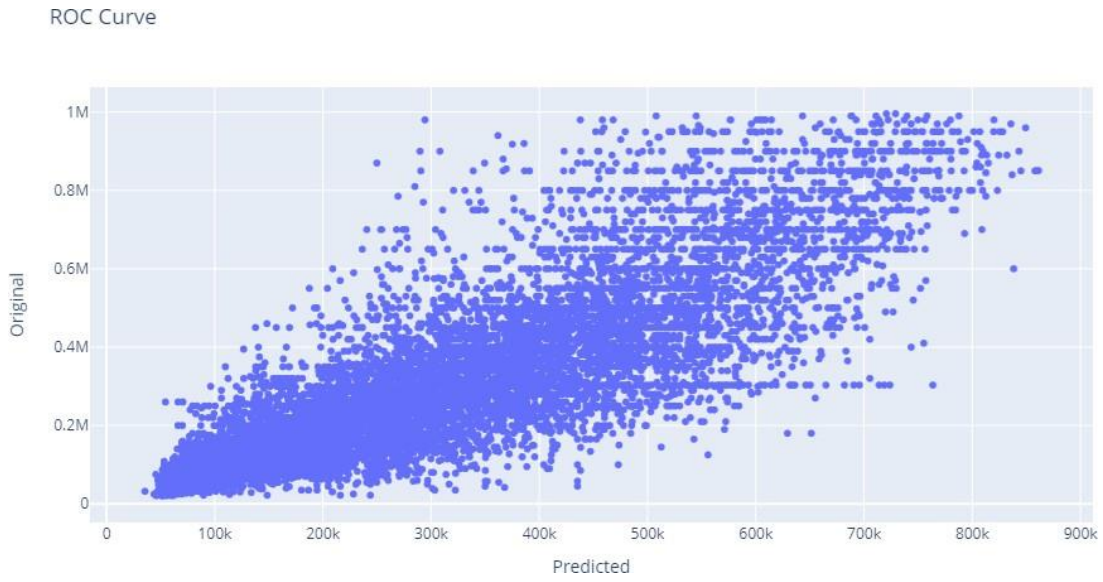
***Εικόνα 31**: Predicted Price vs Original Price (ROC Curve) – Xboosting4*



***Εικόνα 32**: Predicted Price vs Original Price - XBoosting4*

4.2.3 Παλινδρόμηση Ελαφριάς Μηχανής Ενίσχυσης – Light G.B.M.

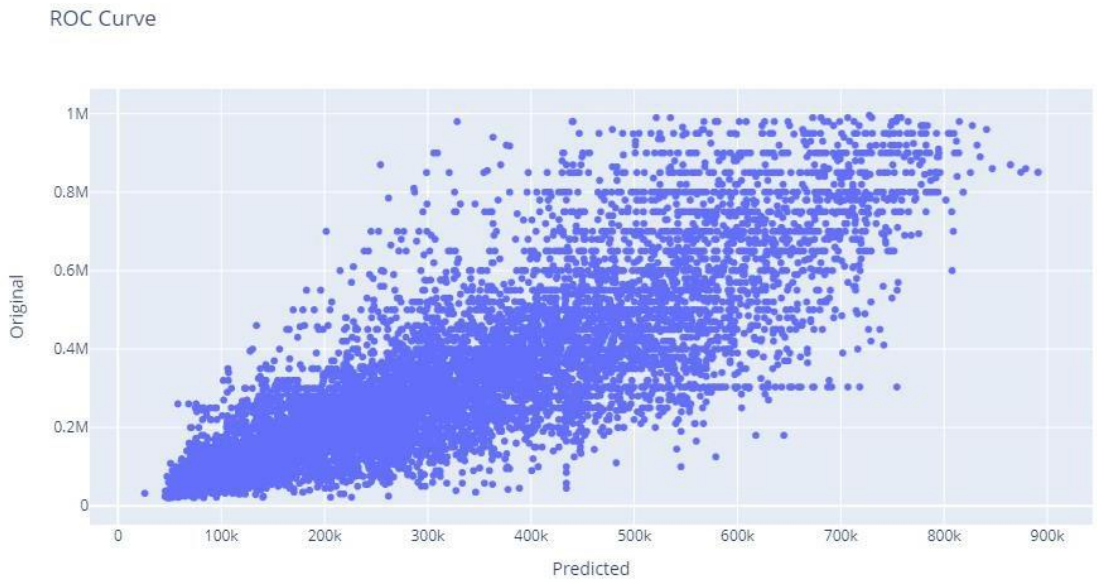
Αποτελέσματα 1ου μοντέλου:



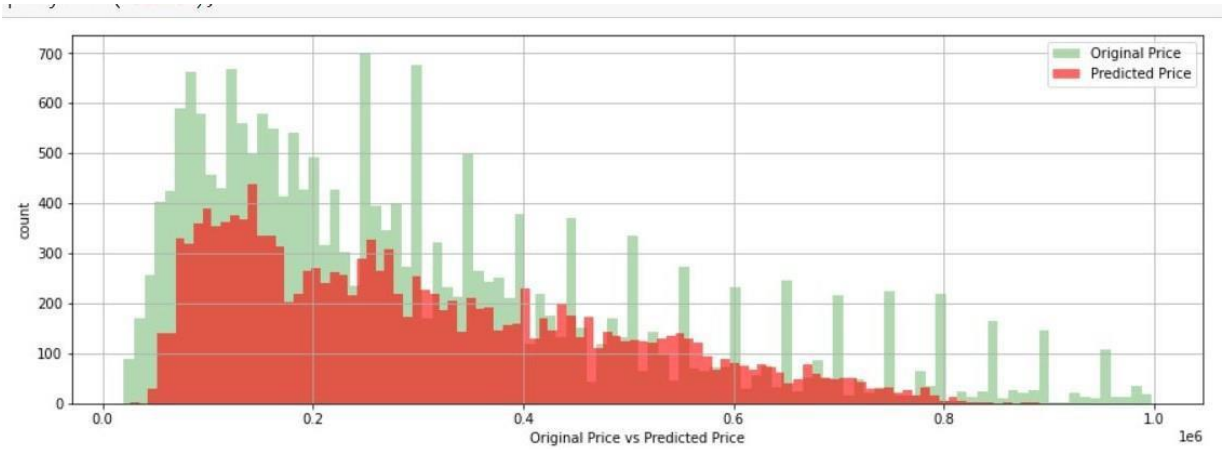
Εικόνα 33: *Predicted Price vs Original Price (ROC Curve) -* ***Light G.B.M.1***

Εικόνα 34*: Predicted Price vs Original Price -* ***Light G.B.M.1***

Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:

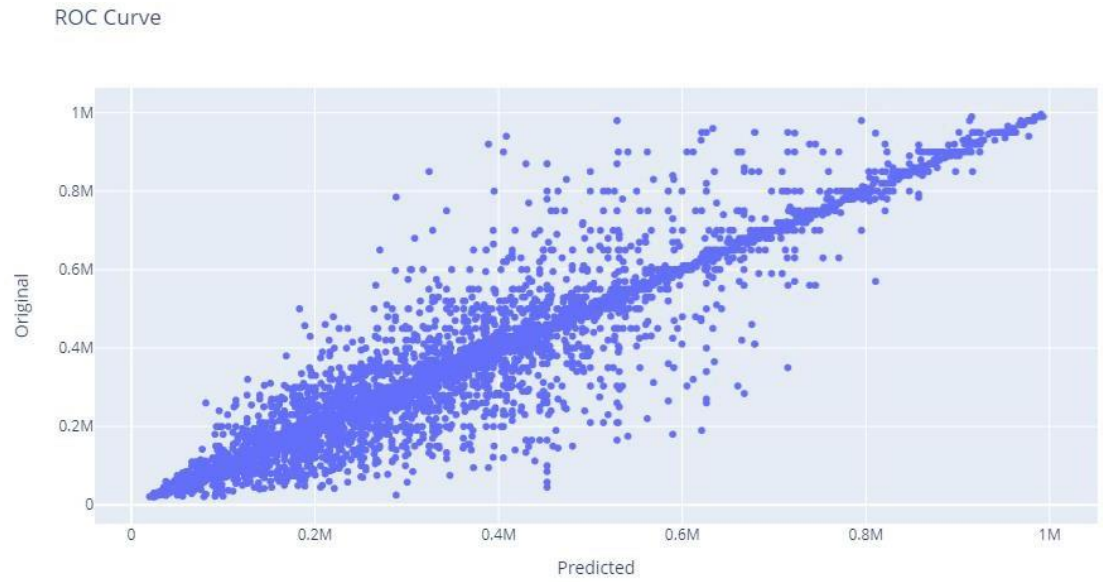


***Εικόνα 35**: Predicted Price vs Original Price (ROC Curve) - Light G.B.M.2*



***Εικόνα 36:** Predicted Price vs Original Price - Light G.B.M.2*

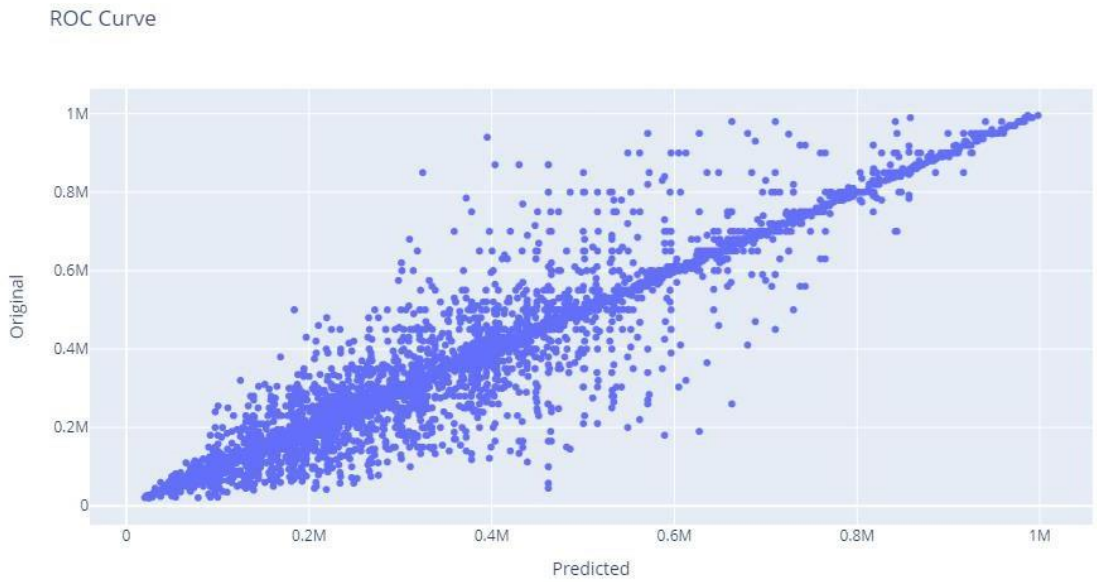
Αποτελέσματα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:



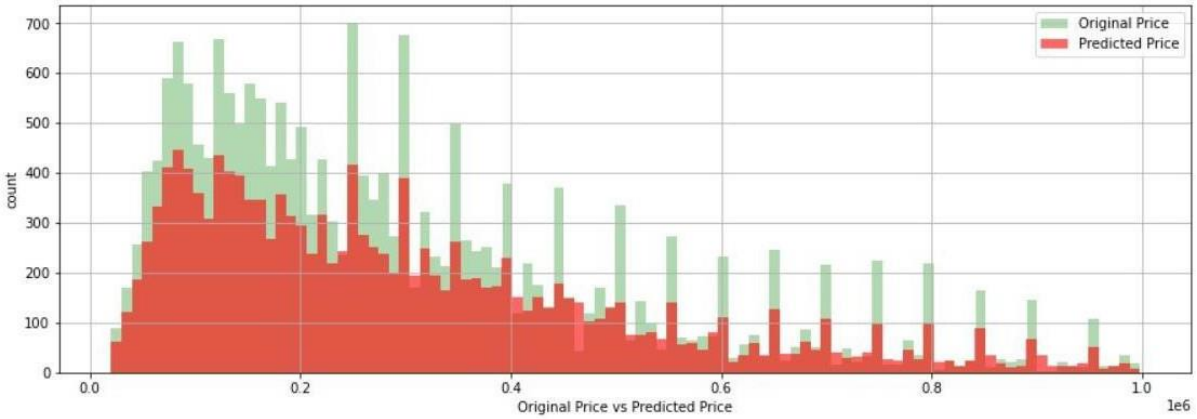
***Εικόνα 37**: Predicted Price vs Original Price (ROC Curve) - Light G.B.M.3*

***Εικόνα 38**: Predicted Price vs Original Price - Light G.B.M.3*

Αποτελέσματα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:



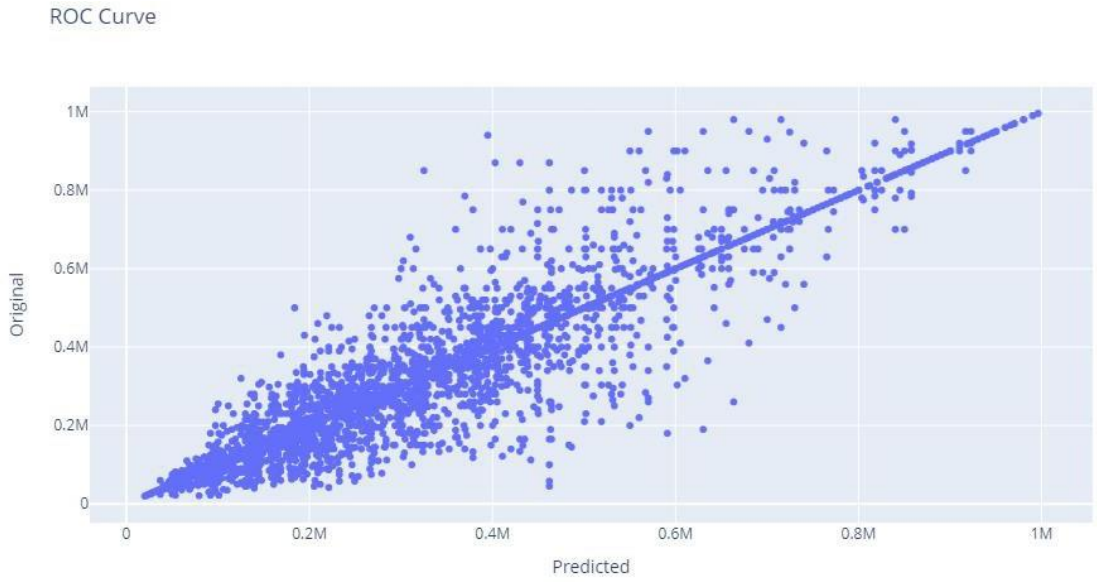
***Εικόνα 39**: Predicted Price vs Original Price (ROC Curve) - Light G.B.M.4*



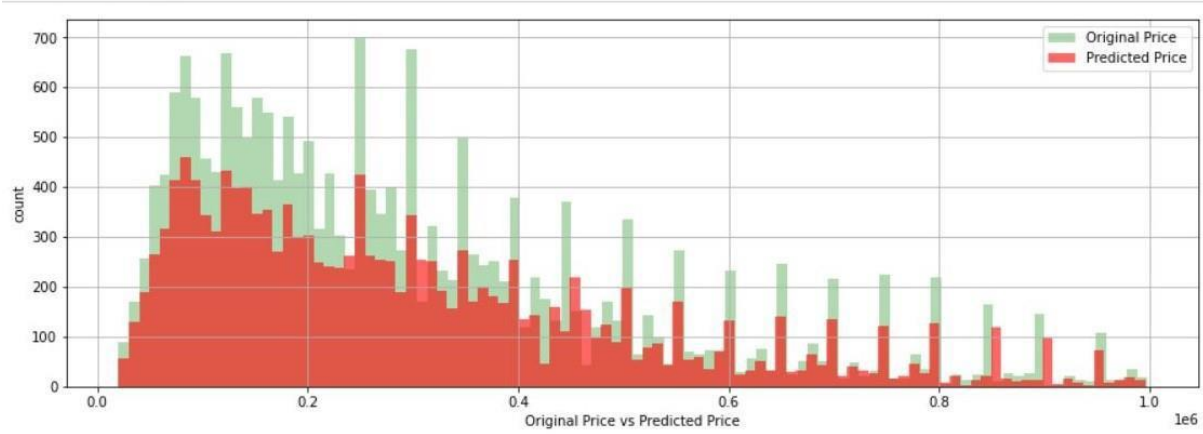
***Εικόνα 40**: Predicted Price vs Original Price - Light G.B.M.4*

4.2.4 Δέντρα Απόφασης – D.T.

Αποτελέσματα 1ου μοντέλου:

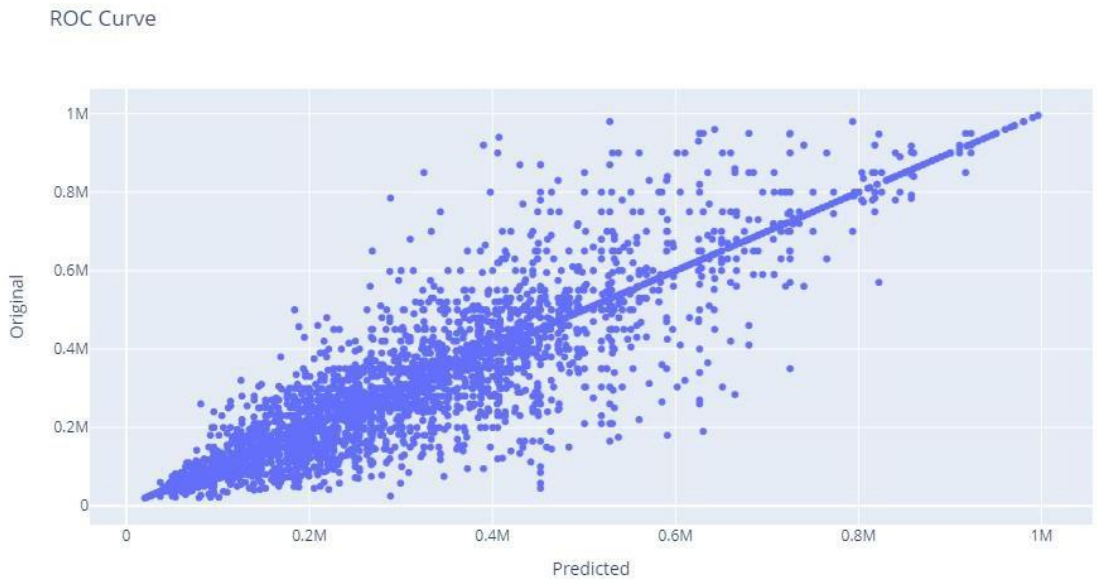


***Εικόνα 41**: Predicted Price vs Original Price (ROC Curve) – D.T.1*



***Εικόνα 42:** Predicted Price vs Original Price – D.T.1*

Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:



***Εικόνα 43**: Predicted Price vs Original Price (ROC Curve) – D.T.2*

***Εικόνα 44:** Predicted Price vs Original Price – D.T.2*

Αποτέλεσμα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:

```
[CV 10/10] END max_depth=5000, max_features=log2, max_leaf_nodes=3000, min_samples_leaf=10, splitter=best;, score=(train=
0.682, test=0.492) total time=   0.0s
{'splitter': 'best', 'min_samples_leaf': 10, 'max_leaf_nodes': 1500, 'max_features': 'auto', 'max_depth': 5000}
0.747920240751656
```

Εικόνα 45: *Το αποτέλεσμα υπολογισμένο με R-squared μετρική – D.T.3*

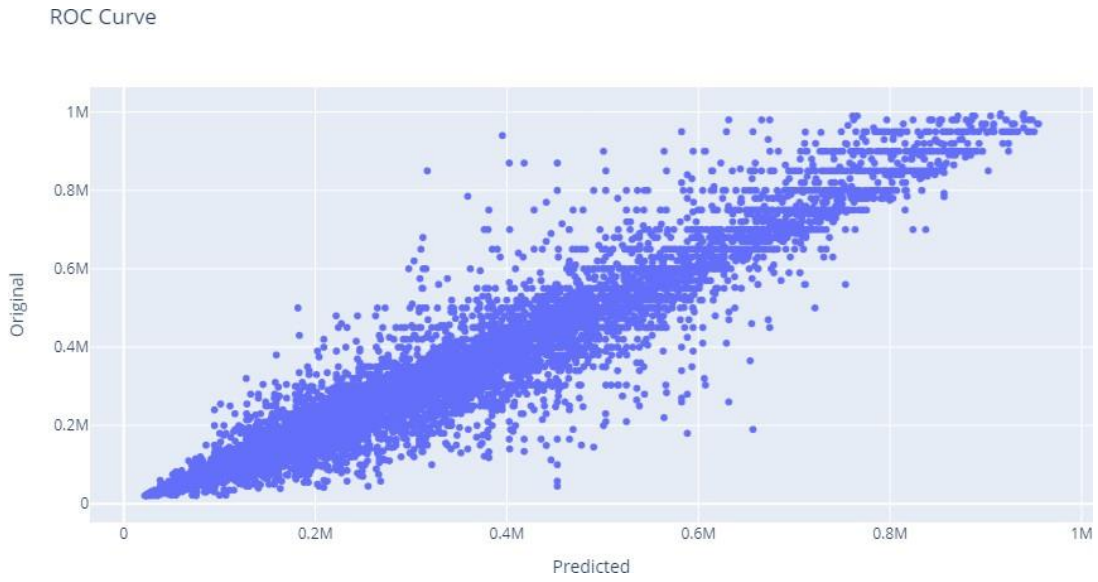
Αποτέλεσμα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:

```
[CV 4/5] END max_depth=750, max_features=log2, max_leaf_nodes=800, min_samples_leaf=10, splitter=best;, score=(train=0.579,
test=0.519) total time=   0.0s
[CV 5/5] END max_depth=750, max_features=log2, max_leaf_nodes=800, min_samples_leaf=10, splitter=best;, score=(train=0.594,
test=0.412) total time=   0.0s
{'splitter': 'random', 'min_samples_leaf': 10, 'max_leaf_nodes': 3000, 'max_features': 'auto', 'max_depth': 750}
0.6978998661173716
```

Εικόνα 46*: Το αποτέλεσμα υπολογισμένο με R-squared μετρική – D.T.4*

4.2.5 Τυχαία Δάση – R.F.

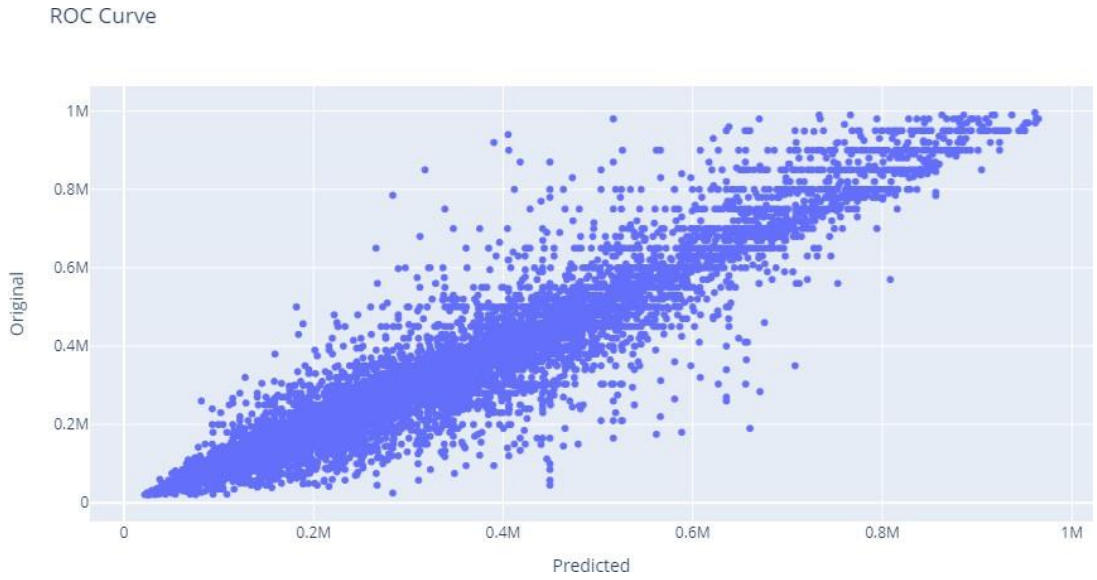
Αποτελέσματα 1ου μοντέλου:



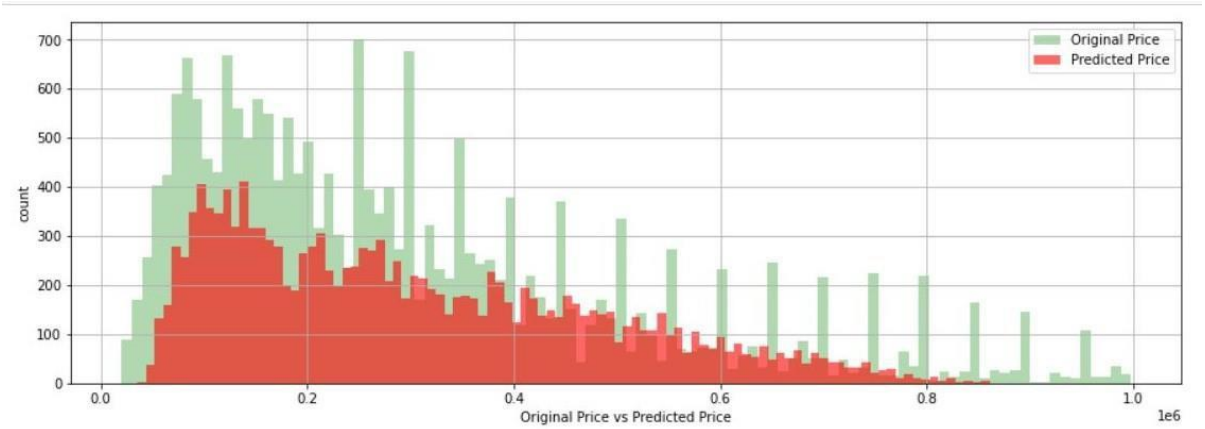
Εικόνα 47**: Predicted Price vs Original Price (ROC Curve) – **R.F.1

Εικόνα 48**: Predicted Price vs Original Price – **R.F.1

Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:



Εικόνα 49: *Predicted Price vs Original Price (ROC Curve – R.F.2*



Εικόνα 50: *Predicted Price vs Original Price – R.F.2*

Αποτελέσματα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
{'n_estimators': 150, 'min_samples_leaf': 2, 'max_leaf_nodes': 5000, 'max_features': 'auto', 'max_depth': 40}
0.8863051657215004
```

Εικόνα 51: *Το αποτέλεσμα υπολογισμένο με R-squared μετρική –* ***R.F.3***

Αποτελέσματα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:

```
Fitting 5 folds for each of 10 candidates, totalling 50 fits
{'n_estimators': 50, 'min_samples_leaf': 1, 'max_leaf_nodes': 5000, 'max_features': 'auto', 'max_depth': 40}
0.9336899584856952
```

Εικόνα 52: *Το αποτέλεσμα υπολογισμένο με R-squared μετρική –* ***R.F.4***

4.2.6 Τεχνητά Νευρωνικά Δίκτυα – A.N.N.

Αποτελέσματα 1ου μοντέλου:

***Εικόνα 53**: Predicted Price vs Original Price – A.N.N.1*

Αποτελέσματα 2ου μοντέλου μετά την επιλογή των καλύτερων μεταβλητών:

***Εικόνα 54**: Predicted Price vs Original Price – A.N.N.2*

Αποτελέσματα 3ου μοντέλου με Hypertuning στο 2ο μοντέλο:

Εικόνα 55: *Predicted Price vs Original Price – A.N.N.3*

Αποτελέσματα 4ου μοντέλου με Hypertuning στο 1ο μοντέλο:

Εικόνα 56: *Predicted Price vs Original Price – A.N.N.4*

4.3 Αξιολόγηση μοντέλων

4.3.1 Μετρικές Σύγκρισης

Για την αξιολόγηση των μοντέλων παλινδρόμησης μηχανικής μάθησης χρησιμοποιούμε τέσσερις μετρικές αξιολόγησης, οι οποίες αποτελούν συγκεκριμένα σκορ. Οι μετρικές αυτές αποτελούν και τα ποσοστά ακρίβειας του κάθε μοντέλου.

MAE - Mean Absolute Error (Μέσο Απόλυτο Σφάλμα)

Το μέσο απόλυτο σφάλμα (Mean Absolute Error) αναπαριστά τον μέσο όρο της απόλυτης τιμής της διαφοράς μεταξύ των πραγματικών και των προβλεπόμενων τιμών στο σύνολο δεδομένων. Είναι μετρική υπολογισμού μέσης τιμής.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Όπου yi είναι οι πραγματικές τιμές του συνόλου δεδομένων και y είναι οι προβλεπόμενες τιμές.

MSE – Mean Squared Error (Μέσο Τετραγωνικό Σφάλμα)

Το μέσο τετραγωνικό σφάλμα (Mean Squared Error) αναπαριστά τον μέσο όρο της τετραγωνικής διαφοράς ανάμεσα στις πραγματικές τιμές του συνόλου δεδομένων και τις προβλεπόμενες τιμές του μοντέλου. Είναι μετρική υπολογισμού διακύμανσης.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

Όπου yi είναι οι πραγματικές τιμές του συνόλου δεδομένων και y είναι οι προβλεπόμενες τιμές.

RMSE – Root Mean Squared Error (Ριζικό Μέσο Τετραγωνικό Σφάλμα)

Το ριζικό μέσο τετραγωνικό σφάλμα (Root Mean Squared Error) αποτελεί την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος και είναι μετρική υπολογισμού τυπικής απόκλισης.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

Όπου yi είναι οι πραγματικές τιμές του συνόλου δεδομένων και y είναι οι προβλεπόμενες τιμές

R2 – R – squared (τετραγωνικό R)

Το τετραγωνικό R – squared αναπαριστά το ποσοστό της διακύμανσης στην εξαρτημένη μεταβλητή, το οποίο μπορεί να εξηγηθεί – να κατανοηθεί από το μοντέλο γραμμικής παλινδρόμησης (linear regression). Οι τιμές αυτής της μετρικής είναι μικρότερες από τη μονάδα πάντα.

$$R^2 = 1 - \frac{\Sigma\ (yi\ - \hat{y})^2}{\Sigma\ (y_i\ - \bar{y})^2}$$

Γενικά, οι τρεις πρώτες μετρικές αξιολόγησης θέλουμε να είναι όσο γίνεται πιο μικρές αριθμητικά, ενώ η τέταρτη μετρική θέλουμε να είναι όσο γίνεται πιο μεγάλη, δηλαδή κοντά στο ένα (παίρνει τιμές $0 \leq x \leq 1$). (Akhilendra P.S., 2019)

Υπενθυμίζουμε ότι: 1ο μοντέλο (αρχικό σε όλο το σύνολο δεδομένων), 2ο μοντέλο (βελτιστοποιημένο σε όλο το σύνολο δεδομένων), 3ο μοντέλο (αρχικό μοντέλο στο 2ο σύνολο μετά την επιλογή λιγότερων μεταβλητών), 4ο μοντέλο (βελτιστοποιημένο μοντέλο στο 2ο σύνολο μετά την επιλογή λιγότερων μεταβλητών).

Η μετρική που επιλέχθηκε για την αξιολόγηση των μοντέλων μας είναι το Root Mean Squared Error (*RMSE*) καθώς έχει τις ίδιες μονάδες μέτρησης με τις αρχικές τιμές αλλά τιμωρεί περισσότερο τις μεγάλες αποκλίσεις από το απλό απόλυτο σφάλμα.

4.3.2 Σύγκριση αποτελεσμάτων

Στους παρακάτω πίνακες παρουσιάζονται τα αποτελέσματα των μετρικών αξιολόγησης και για τα 6 μοντέλα μηχανικής μάθησης που κατασκευάστηκαν:

1ο Μοντέλο απεικονίζει τα αποτελέσματα του πρώτου τρεξίματος στα δεδομένα

	MAE	MSE	RMSE	R-squared
MLR	108065.2143	149288.7119	386.3790	0.5265
XGB	89709.4588	127117.6523	356.5356	0.6567
LGBM	76544.7655	109227.4655	330.4958	0.7465
DT	**12843.1387**	**42253.8247**	**205.5554**	**0.9621**
RF	31693.8313	55189.5284	234.9245	0.9353
ANN	106060.1764	148095.7834	384.8321	0.5340

***Πίνακας 10**: Αποτελέσματα 1ου Μοντέλου*

2ο Μοντέλο απεικονίζει τα αποτελέσματα του τρεξίματος στα δεδομένα μετά την επιλογή καλύτερων μεταβλητών (Feature Selection)

	MAE	MSE	RMSE	R-squared
MLR	109133.0778	130103.07255	387.4314	0.5213
XGB	89637.4988	126830.5512	356.1328	0.6583
LGBM	76735.0050	109626.0139	331.0982	0.7447
DT	**16095.5892**	**47915.4186**	**218.8959**	**0.9512**
RF	33233.5063	58702.0792	242.2851	0.9268
ANN	106380.0928	148552.6311	385.4253	0.5311

***Πίνακας 11**: Αποτελέσματα 2ου Μοντέλου*

3ο Μοντέλο απεικονίζει τα αποτελέσματα του τρεξίματος στα δεδομένα μετά την επιλογή καλύτερων μεταβλητών (Feature Selection) και την παραμετροποίηση των αλγορίθμων (Hypertuning).

Να σημειωθεί εδώ πως για τους αλγόριθμους: 'MLR', 'LGBM', 'RF' και 'DT' η επιλογή καλύτερων μεταβλητών των αλγορίθμων δυσχέραινε το αποτέλεσμα τους και για αυτό επιλέξαμε να μην την κάνουμε στο 4ο μοντέλο, εκτός του αλγόριθμου 'XGB' ο οποίος ήταν πιο αποδοτικός με το συγκεκριμένο μοντέλο, χωρίς να ξεπερνάει ακόμη σε τελικό σκορ τον 'DT'.

Ο αλγόριθμος 'ANN' έδειξε καλύτερο σκορ με την επιλογή καλύτερων μεταβλητών, αλλά χωρίς τη μέθοδο αυτήν θα είχε ακόμα καλύτερα αποτελέσματα γιατί αυτός ο αλγόριθμος δουλεύει καλύτερα με πολλά δεδομένα, η καλύτερη απόδοση του οφείλεται μόνο στην παραμετροποίηση που έγινε.

Για τον αλγόριθμο 'MLR' συνεχίσαμε τον έλεγχο για τυπικούς λόγους, γιατί φανερά είχε τα μικρότερο σκορ ήδη κατά το 1ο , 2ο και 3ο μοντέλο.

	MAE	**MSE**	**RMSE**	**R-squared**
MLR	nan	nan	nan	0.5210
XGB	**16095.5892**	**47915.4186**	**218.8959**	**0.9512**
LGBM	17994.0992	48310.2291	219.7959	0.9504
DT	nan	nan	nan	0.7479
RF	nan	nan	nan	0.8863
ANN	81913.1492	114095.2155	337.7798	0.7234

***Πίνακας 12**: Αποτελέσματα 3ου Μοντέλου*

4ο Μοντέλο απεικονίζει τα αποτελέσματα του τρεξίματος στα δεδομένα χωρίς να γίνει επιλογή καλύτερων μεταβλητών (Feature Selection) εκτός του αλγόριθμου 'XGB'. Η παραμετροποίηση των αλγορίθμων (Hypertuning) έγινε για όλα τα μοντέλα.

Να σημειωθεί εδώ πως ο αλγόριθμος 'XGB' εφόσον έδειξε να ανταποκρίνονται στην επιλογή καλύτερων μεταβλητών και την παραμετροποίησή του, χρησιμοποιήθηκε για τις προβλέψεις στα validation και test dataset με όλες τις αλλαγές του. Για τους αλγόριθμους 'LGBM', 'DT', 'RF' και 'ANN' δεν έγινε η επιλογή καλύτερων μεταβλητών (Feature Selection), μόνο η παραμετροποίηση (Hypertuning), καθώς η απόδοση των συγκεκριμένων αλγόριθμων θα αυξανόταν ακόμα περισσότερο στις προβλέψεις των validation και test sets.

	MAE	**MSE**	**RMSE**	**R-squared**
MLR	nan	nan	nan	0.5265
XGB	**12843.1387**	**42253.0247**	**205.5554**	**0.9621**
LGBM	14446.5821	42575.3145	206.3379	0.9615
DT	nan	nan	nan	0.6979
RF	nan	nan	nan	0.9337
ANN	57075.1731	81959.6834	286.2860	0.8573

***Πίνακας 13**: Αποτελέσματα 4ου Μοντέλου*

Κεφάλαιο 5 – Συμπεράσματα & Προτάσεις

5.1 Σύνοψη

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της μοντελοποίησης των τιμών των ακινήτων. Ένα πρόβλημα εξέχουσας σημασίας για ένα σεβαστά μεγάλο τμήμα του Ελληνικού και βέβαια όχι μόνο, πληθυσμού, εφόσον η αγορά των ακινήτων επηρεάζει άμεσα την οικονομία μας. Στην συγκεκριμένη μελέτη, επικεντρωθήκαμε στην αγορά των ακινήτων του νομού της Αττικής, του νομού της Θεσσαλονίκης, του Πειραιά και των υπολοίπων γύρω περιοχών τους. Πραγματοποιήθηκε μία προσπάθεια υλοποίησης ενός μοντέλου πρόβλεψης των τιμών των ακινήτων, ακολουθώντας τις μεθοδολογίες και τη λογική της μηχανικής μάθησης, που προβλέπονται μέσω της βιβλιογραφίας.

Αρχικά, συλλέξαμε δεδομένα από την ιστοσελίδα της Χρυσής Ευκαιρίας (ΧΕ), (Spiti360) και (Plot) τα οποία αφορούσαν ακίνητα προς πώληση, και προχωρήσαμε στην λεπτομερή επεξεργασία τους. Η αναλυτική διερεύνηση των δεδομένων, μας επέτρεψε να κατανοήσουμε σε βάθος την φύση των δεδομένων και τα χαρακτηριστικά του, γεγονός που μας οδήγησε στην βέλτιστη χρήση του και στην επιλογή του ορθότερου τρόπου επεξεργασίας του. Η επιλογή των καταλληλότερων μεθόδων μηχανικής μάθησης, οι οποίες χρησιμοποιήθηκαν για την μοντελοποίηση της μεταβλητής στόχου, αποτέλεσε απόρροια της βαθύτερης ανάλυσης των δεδομένων μας. Για να φτάσουμε στην βέλτιστη μορφή του συνόλου δεδομένων, κληθήκαμε να εντοπίσουμε έναν αποτελεσματικό τρόπο διαχείρισης της μέτριας, έως και κακής, ποιότητάς του, την υψηλή τυπική του απόκλισή, όσον αφορά την μεταβλητή στόχο και τις πολυάριθμες κενές τιμές του, λαμβάνοντας, επίσης, υπόψιν μας τις σχέσεις μεταξύ των γνωρισμάτων. Το μικρό μέγεθος του dataset, μας οδήγησε στην μείωση και απλοποίηση των χαρακτηριστικών εισόδου, προκειμένου να αντιμετωπιστεί υπερεκπαίδευση των μοντέλων μας.

Δοκιμάστηκαν, έτσι, ποικίλοι αλγόριθμοι, για την εκπαίδευση των μοντέλων με είσοδο το σύνολο δεδομένων, από τους οποίους βέλτιστα αποτελέσματα μας έδωσαν οι τεχνικές παλινδρόμησης LGBM, η παλινδρόμηση με χρήση της τεχνικής X Gradient Boosting, τα Τυχαία Δάση (RF) και τα Δέντρα Απόφασης

(DT). Με βάση αυτά, διακρίνουμε και αναλύουμε τη σπουδαιότητα των χαρακτηριστικών και κατανοούμε τον ρόλο που κατείχε το καθένα από αυτά στην όψη των αποτελεσμάτων. Όλοι οι αλγόριθμοι έκριναν ότι το εμβαδόν των ακινήτων, η πόλη του ακινήτου, το έτος κτίσης, ο αριθμός των υπνοδωματίων, το μέσον θέρμανσης, ο φόρος Ένφια, ο πληθυσμός του κάθε νομού και το ποσοστό εγκληματικότητας αντίστοιχα αποτελούν τα κρισιμότερα γνωρίσματα, επηρεάζοντας σε σημαντικό βαθμό την τιμή πώλησης.

Τα αποτελέσματα των πειραμάτων μας, υποδεικνύουν την πολυπλοκότητα του προβλήματος και του πλήθος των παραγόντων από τους οποίους αυτό εξαρτάται. Εξετάζοντας τις τιμές των μετρικών που λάβαμε από τα μοντέλα μας, αναρωτιόμαστε ποιοι είναι οι λόγοι εξαιτίας των οποίων αυτές δεν ήταν δυνατόν να βελτιωθούν. Ως φυσικό επακόλουθο λοιπόν, έρχεται η ερμηνεία και αξιολόγηση των αποτελεσμάτων. Καταλήγουμε, έτσι, σε ένα σύνολο αιτιών, οι οποίες κρίνουμε ότι κατέχουν ρόλο εξέχουσας σημασίας στην διαμόρφωση της επίδοσης των μοντέλων μας. Εδώ πρέπει να σημειωθεί, ότι σύμφωνα με τα πειράματα που διεξήχθησαν, αλλά και με τη μελέτη πολυάριθμων ερευνών ίδιου θέματος με αυτό που καταπιανόμαστε, μπορούμε να καταλήξουμε με ασφάλεια στην ορθότητα της επιλογής των συγκεκριμένων μεθόδων και μοντέλων πρόβλεψης και να αποκλείσουμε την πιθανότητα να ευθύνονται για τα επίπεδα των σφαλμάτων. Αντιθέτως, το περιορισμένο μέγεθος του dataset και η απουσία χαρακτηριστικών που εκφράζουν σημαντικές λεπτομέρειες, οι οποίες διαχωρίζουν τα υψηλής με τα χαμηλής τιμής ακίνητα, σε συνδυασμό με την μεγάλη διακύμανση των τιμών της τιμής πώλησής τους, ακόμα και για ακίνητα στην ίδια περιοχή και με παραπλήσιο εμβαδόν καθώς και η έλλειψη καταγραφής του μέσου θέρμανσης αλλά και η παραπλάνηση του έτους κατασκευής του ακινήτου αφού για την σωστή χρονολογία κατασκευής αναγραφόταν η χρονολογία μερικής ή και της ελάχιστης ανακαίνισης του ακινήτου, αποτελούν από τους κυριότερους παράγοντες και πιστεύουμε πως συνέβαλαν καθοριστικά στην ποιότητα των αποτελεσμάτων.

Στη συνέχεια, προχωράμε σε κάποιες συγκρίσεις που διεξήγαμε, μεταξύ παρόμοιων δημοσιευμένων ερευνών, οι οποίες ακολούθησαν παραπλήσια σειρά βημάτων με αυτήν που ακολουθήθηκε και στην παρούσα μελέτη. Ακόμα, ενδιαφέρουσα κρίθηκε η σύγκριση των τιμών, οι οποίες προκύπτουν με βάση την

αντικειμενική αξία των ακινήτων, αρχικά με τις πραγματικές τιμές πώλησης του dataset, αξιολογώντας τις τιμές των μετρικών της μεταξύ τους σύγκρισης και έπειτα με τις τιμές πώλησης που προέβλεψαν τα μοντέλα μας.

Συμπερασματικά, η πρόβλεψη των τιμών πώλησης των ακινήτων κρίνεται ιδιαίτερα απαιτητική και πολύπλοκη διαδικασία, καθώς η τιμή ενός ακινήτου επηρεάζεται από πλήθος παραγόντων, ποικίλων φύσεων, εφόσον συνδέεται άρρηκτα με την κατάσταση και την οικονομία κάθε κοινωνίας, τις αλλαγές που υφίστανται και τους παράγοντες που τις επηρεάζουν. Προτείνεται, όμως, εδώ, μία βασική μεθοδολογία αντιμετώπισης του συγκεκριμένου ζητήματος, κατά την οποία γίνεται προσπάθεια αξιοποίησης των υπαρχόντων πόρων, με σκοπό να ληφθούν όσο το δυνατόν ακριβέστερα αποτελέσματα.

5.2 Μελλοντικές Προτάσεις

Η παρούσα διπλωματική εργασία ανέδειξε τα προβλήματα που προκύπτουν στην προσπάθεια πρόβλεψης της τιμής πώλησης των ακινήτων με τεχνικές μηχανικής μάθησης. Στα πλαίσια της μελέτης αυτής διερευνήθηκαν διάφορες προσεγγίσεις και μεθοδολογίες του ζητήματος με το οποίο καταπιάνεται και παράλληλα διαπιστώθηκαν ορισμένα σημεία στα οποία θεωρούμε ότι διαφορετική αντιμετώπιση θα ωφελούσε. Ακόμα, κρίνοντας πως στο μέλλον οι μελέτες και εφαρμογές γύρω από το συγκεκριμένο ζήτημα θα πληθύνουν, παρουσιάζουμε κάποιες προτάσεις που θα μπορούσαν να συμβάλλουν το μελλοντικό έργο.

Αρχικά, για μελλοντικές και πιο ολοκληρωμένες μελέτες, περισσότερα και πιο έγκυρα δεδομένα απαιτούνται για την επιτυχία τους. Η μελέτη ενός μεγαλύτερου συνόλου δεδομένων θα έδινε μια πιο γενικευμένη και πραγματική εικόνα, η οποία θα μας έδινε εγκυρότερα αποτελέσματα. Ακόμα, απαιτούνται περισσότερες μεταβλητές, οι οποίες θα περιγράφουν πιο λεπτομερείς και σημαντικούς παράγοντες και χαρακτηριστικά, που επηρεάζουν άμεσα την τιμή πώλησης ενός ακινήτου. Χαρακτηριστικά τέτοιου είδους αποτελούν το επίπεδο πολυτέλειας του ακινήτου και των υλικών από τα οποία είναι δομημένο, η κατάσταση των βασικών ηλεκτρικών συσκευών και εγκαταστάσεων του, η λεπτομερέστερη περιγραφή του πιθανού κήπου του. Επιπλέον, εξίσου σημαντικό κρίνεται να ληφθούν υπόψιν περιβαλλοντικοί παράγοντες που επηρεάζουν ένα ακίνητο, όπως

τα επίπεδα μόλυνσης της ατμόσφαιρας κάθε περιοχής. Τα δεδομένα μας συγκεντρώθηκαν κατά μία περίοδο διάρκειας αρκετών μηνών, γεγονός που σημαίνει ότι δεν περιέχουν σημαντικά στοιχεία εποχικότητας ή οικονομικά στοιχεία σεβαστής σημασίας. Σε μελλοντικές, λοιπόν, μελέτες προτείνεται να δοθεί προσοχή σε μακροοικονομικούς παράγοντες, αλλά και σε στοιχεία των πωλήσεων ακινήτων που έχουν ήδη ολοκληρωθεί στο παρελθόν. Γενικότερα, η καλύτερη ποιότητα των δεδομένων που θα μπορούσε να εξασφαλιστεί εάν αυτά παρέχονται με περισσότερη ευκολία από τους αρμόδιους φορείς θα βελτίωνε σημαντικά τις μοντελοποιήσεις.

Ένα ακόμη ενδιαφέρον σημείο των μελλοντικών κατευθύνσεων, θα μπορούσε να αποτελέσει η αξιοποίηση των δεδομένων εικόνας των φωτογραφιών των ακινήτων, αλλά και δορυφορικών δεδομένων, καθώς επίσης και η αξιοποίηση των γεωχωρικών δεδομένων τους. Οι τεχνικές αυτές φαίνεται να συμβάλουν στην βελτίωση της ακρίβειας των αποτελεσμάτων και η ποιότητα τους στο μέλλον μόνο θα αναβαθμίζεται.

Επιπροσθέτως, θα μπορούσαμε να προτείνουμε την κατασκευή και εκπαίδευση διαφορετικών μοντέλων για διαφορετικές περιοχές ή σύνολα περιοχών με παρόμοια χαρακτηριστικά. Εάν τα δεδομένα είναι επαρκή, το παραπάνω μόνο να βελτίωνε θα μπορούσε την εγκυρότητα των προβλέψεων. Παράλληλα, ενθαρρύνεται η προσπάθεια υλοποίησης μοντέλων στηριζόμενα σε διαφορετικές τεχνικές, όπως είναι η βαθιά μάθηση και τα νευρωνικά δίκτυα, με τη βοήθεια των οποίων θα μοντελοποιούταν ευκολότερα και ακριβέστερα η σχέση της τιμής του ακινήτου με το πλήθος των μεταβλητών από τις οποίες αυτή εξαρτάται.

Επιπλέον, ενδιαφέρουσα κρίνεται η δημιουργία ενός δυναμικού συστήματος πρόβλεψης των τιμών των ακινήτων, προτείνεται ως πιθανός στόχος μελλοντικών ερευνών. Με τον όρο δυναμικό, εννοούμε ένα σύστημα το οποίο θα μεταβάλλεται ζωντανά, με βάση τους παράγοντες του παρόντος από τους οποίους επηρεάζεται. Έτσι, θα μπορούσε να καταγράφεται και να μελετάται σε βάθος χρόνου η συμπεριφορά και η πορεία της αγοράς των ακινήτων, αλλά και να προτείνονται σε αληθινό χρόνο οι τιμές των ακινήτων στους ενδιαφερόμενους.

Συνιστάται, ακόμα, η έρευνα της εφαρμογής παρόμοιων μεθόδων πρόβλεψης σε προβλήματα διαφόρων τομέων, όπως είναι η μελλοντική κατεύθυνση της αγοράς

των ακινήτων, η πρόβλεψη των τιμών πετρελαίου και η πρόβλεψη των τιμών των μετοχών στο χρηματιστήριο, προς γενίκευση της παρούσας μελέτης.

Τέλος, η πιθανότητα πώλησης ενός ακινήτου, όπως αυτή εξελίσσεται σε μία συγκεκριμένη χρονική περίοδο από την ημέρα δημοσίευσής της αγγελίας του, ανάλογα με την τιμή του, αποτελεί μία δυνατότητα που θα μπορούσε να προστεθεί σε μελλοντικές εφαρμογές.

Βιβλιογραφία

1) Ζέντελης, Π., (2001), Real Estate. Αξία. Εκτιμήσεις. Ανάπτυξη. Επενδύσεις. Διαχείριση. Εκδόσεις Παπασωτηρίου.

2) Franz, H. (2020), Deutsch-Hellenische Schutzgemeinschaft fur Auslandsgrundbesitz.

3) Καρανικόλας, Ν. (2010), Η Εκτίμηση των Ακινήτων, Εκδόσεις Δισίγμα.

4) Κοχιός Π. (2007), Εισαγωγή στην εκτίμηση των Ακινήτων και μέθοδοι αποτίμησης της αξίας αυτών, Ιδιωτική Έκδοση, Αθήνα.

5) Ελληνική Δημοκρατία, Υπουργείο Ψηφιακής Διακυβέρνησης, ≪Βίβλος Ψηφιακού Μετασχηματισμού 2020-2025≫.

6) Συντακτική ομάδα ert.gr, "Μειώθηκε κατά 8% ο πληθυσμός στο Λονδίνο λόγω κορονοϊού", Αναστασία Καντζάβελου, 22 Ιανουαρίου 2021.

7) Μαρινάκης, Κ., "Η παγκόσμια Χρηματοοικονομική κρίση του 2007", Greekonomics, 07 Νοεμβρίου 2020.

8) Reuters, "UBS : Στη Φρανκφούρτη ο μεγαλύτερος παγκοσμίως κίνδυνος φούσκας στην αγορά των ακινήτων", 13 Οκτωβρίου 2021, ΑΠΕ-ΜΠΕ, ertnews.gr.

9) Wikipedia, "Australia", wikipedia, 01 Ιανουαρίου 2022.

10) Salam, F., Walid, A, Ammar, E, Bushra, A., Maisa, A. (2021), "Business Intelligence Framework Design and Implementation: A Real - estate Market Case study", Journal of Data and Information Quality, Vol. 13, Issue 2, No 10, 1-15.

11) Conway, D., Jennifer, E., (2018), "Artificial intelligence and machine learning: current application in real estate", Massachusetts Institute of Technology, Center for Real Estate, Programs in Real Estate Development, Vol. 3, 113-117.

12) Byeonghwa, P., Jae Kwon, B. (2015), “Using machine learning algorithms for housing price prediction: The case of Fairfax Country housing data”, Expert Systems with applications, Vol. 42, No. 6, 2928-2934.

13) Jose-Luis, A., Cano, E., Esteban, A., Noelia, G., Gamez. M. (2020), “A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems”, Computational Methods Applied to Data Analysis for Modeling Complex Real Estate Systems, Vol. 2020, No 5287263.

14) Tengxiang, S., Haijiang, L., Yi A. (2021), “A BIM and machine learning integration framework for automated property valuation”, Journal of Building Engineering, Vol. 44, No. 102636.

15) Ozlan, I., Terry, F., Işıl, E., Uwe, D., (2021), “Market news and credibility cues improve house price predictions: An experiment on bounded rationality in real estate”, Journal of Behavioral and Experimental Finance, Vol 31, No. 100550.

16) Shen, L., Ross, St., (2021), “Information value of property description: A Machine Learning approach”, Journal of Urban Economics, Vol. 121, No. 103299.

17) Quang, T., Nguyen, M., Dank, H., Mei, B., (2020), “Housing price prediction via improved Machine Learning Teqniques”, Vol. 174, 433-442.

18) Kang, Y., Zhang, F., Peng, W., Jinmeng, R., Duarte, F., Ratti, C. (2021), “Understanding house price appreciation using multi-source big geo-data and machine learning”, Land of Policy, Vol. 111, No. 104919.

19) Ping-Feng, P., Wen-Chang, W., (2020), “Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices”, Department of Information Management, Nation Chi Nan University.

20) Baldominos, A., Blanco, I., Jose Moreno, A., Iturrarte, R., Bernarde, O., Afonso,

C., (2018), “Identifying Real Estate Opportunities Using Machine Learning”, Applied Sciences, Vol. 8, Issue 11, No 10.3390.

21) Forbes, “11 Predictions For The Future of The Real Estate Market”, 13 Σεπτεμβρίου 2021.

22) Rehman, F., Hira, M., Adnan, U., (2019), “An intelligent Context Aware Recommender System for Real Estate”, Mediterranean Conference on Pattern Recognition and Artificial Intelligence, Vol. 1144, No. 32, 177-191.

23) Dimopoulos, T., Bakas, N., (2019), “Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate”, Remote Sensing in Applications of Geoinformation, School of Architecture, Journals, Remote sensing, Vol. 11, Issue 24, No. 3047.

24) Pinter, G., Mosavl, A., Felde, I., (2020), “Artificial Intelligence for Modeling Real

Estate Price Using Call Detail Records and Hybrid Machine Learning Approach”, Entropy 2020, Vol. 22, No. 1421.

25) Centric Consulting., “Machine Learning: A Quick Introduction and Five Core Steps”. 10 Απριλίου 2019.

26) Gurucharan M., “Machine Learning Basics: Decision Tree Regression”, 15 Ιουλίου 2020.

27) Drakos G., “Decision Tree Regressor explained in depth”, 23 Μαΐου 2019.

28) Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

29) Gurucharan M., “Machine Learning Basics: Random Forest Regression”, Towards Data Science, 18 Ιουλίου 2020.

30) Dutta, A., “Random Forest Regression in Python”, Geek for Geeks, 18 Ιανουαρίου 2022.

31) Esmalifalak, M., “Anomaly detection with KNN”, 25 Ιουνίου 2020.

32) Teixeira-Pinto, A., “Machine Learning for Biostatistics”, Bio Statistics Collaboration of Australia, 02 Αυγούστου 2021.

33) Javatpoint, “Classification: Support Vector Machine Algorithm”

34) Ray, S., “Understanding Support Vector Machine(SVM) algorithm from examples (along with code)”, Analytics Vidhya, 13 Σεπτεμβρίου 2017.

35) Ritchie, N., “Supervised Learning Theory: Support Vector Machines (SVMs)”, 11 Δεκεμβρίου 2021.

36) Ironhardt, B., “Machine Learning - Handling Missing Data”, Towards Data Science, 28 Ιουλίου 2020. 105

37) Wikipedia, “Heat Map”, wikipedia, 10 Ιανουαρίου 2022.

38) Sanat, S., “Data Visualization using Python for Machine Learning and Data science:” Towards Data Science, 15 Δεκεμβρίου 2018.

39) Math is Fun, (2019), “Histograms”.

40) Wilke, C., “Fundamentals of Data Visualization: Visualizing distributions: Histograms and density plots”, O’Reilly, 3 Σεπτεμβρίου 2021.

41) Science Direct, (2018), “Density Plot”, Journals and Books.

42) Chug, A., “ML | Label Encoding of datasets in Python”, Geek for Geeks, 24 Σεπτεμβρίου 2021.

43) Wikipedia, “Feature Selection”, wikipedia, 30 Ιανουαρίου 2022.

44) Bex, T., “How to Use Pairwise Correlation For Robust Feature Selection”, Towards Data Science, 13 Απριλίου 2021.

45) Brownlee, J., “How to Choose a Feature Selection Method For Machine Learning”, Machine Learning Mastery, 20 Αυγούστου 2020.

46) Kouate, P.M., “Machine Learning: GridSearchCV & RandomizedSearchCV”, Towards Data Science, 11 Σεπτεμβρίου 2020.

47) Akhilendra, P.S., “Evaluation Metrics for Regression models- MAE Vs MSE Vs RMSE vs RMSLE”, 20 Μαρτίου 2019.

48) Skyscraper center, (2021), “Frankfurt am Main”, Council on Tall Buildings and Urban Habitat.

49) Wilson-Powell, G., “London’s Top 10 Iconic Buildings”, The Culture Trip, 25 Ιουνίου 2019.

50) Cowan, J., “Melbourne CBD”, ABC News, 31 Ιουλίου 2019.

51) Dorsch, J., “Making Buildings Smarter”, 04 Οκτωβρίου 2018.

52) Mburugu, C., “10 Factors That Affect Property Value”, 17 Ιουλίου 2019.

53) Yadav, V., (2021), “Emerging Technology In Real Estate 2021”, Analytics Online.

54) Revenue-Hub, (2016), “Next Generation Hotel Property Management Systems”.

55) Kummerow, M., (2002), “Theory for Real Estate Valuation : An Alternative Way to Teach Real Estate Price Estimation Methods”, Department of Property Studies, Curtin University.

56) Pagourtzsi, E., Assimakopoulos, V., Hatzichristos, T., French, N., (2003), “Real estate appraisal: a review of valuation methods”, Journal of Property Investments & Finance, Vol. 21, No 4, Pages 383-401.

57) Poursaeed, O., Matera, T., Belongie, S., (2018), “Vision – based real estate price estimation”, Machine Vision and Applications, Vol. 34, No 5, Pages 667 – 676.

58) Dambon, J., Sigrist, F., Furrer, R., (2021), “Maximum likelihood estimation of spatially varying coefficient models for large data with an application to real estate price prediction”, Vol. 41, No 100470,

59) Adetiloye, K., Omoruyi, P., (2014), “A review of real estate valuation and optimal pricing techniques”, Asian Economic and Financial Review, Vol. 4, Pages 1878 – 1893.

60) Yu, Y., Lu, J., Shen, D., Chen, B., (2020), Research on real estate pricing methods based on data mining and machine learning", Neural Computing and Applications, Vol. 33, Pages 3925 – 3937.

61) Shen, J., Pretorius, F., (2013), "Binomial option pricing models for real estate development", Journal of Property Investment & Finance, Vol. 31, No 5.

62) Yeh, C., Hsu, T., (2018), "Building real estate valuation models with comparative approach through case – based reasoning", Applied Soft Computing, Vol. 65, Pages 260 – 271.

63) Crosby, N. Jackson, C., Orr, A., (2016), "Refining the real estate pricing model", Journal of Property Research, Vol.33, No. 4, Pages 332 – 358.

64) Wang, D., Jing, V., (2019), "Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review", Vol. 11, Issue 24, No. 7006.

64) Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair - "Housing Price Prediction Using Machine Learning and Neural Networks" 2018, IEEE.

65) G.Naga Satish, Ch.V.Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu "House Price Prediction Using Machine Learning". IJITEE, 2019.

66) CH. Raga Madhuri, G. Anuradha, M. Vani Pujitha -" House Price Prediction Using