Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Αγρονόμων και Τοπογράφων Μηχανικών - Μηχανικών Γεωπληροφορικής
Τομέας Τοπογραφίας
Εργαστήριο Φωτογραμμετρίας

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΡΟΒΛΕΨΗ ΤΙΜΩΝ ΕΝΟΙΚΙΑΣΗΣ ΑΚΙΝΗΤΩΝ ΜΕ ΤΗΝ ΧΡΗΣΗ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΧΑΡΑΞΗ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΠΟΛΙΤΙΚΗΣ

Δήμητρα - Αλεξάνδρα Μαλαπάνη

Υπεύθυνος Καθηγητής: Νικόλαος Δουλάμης

Τριμελής επιτροπή

Δουλάμης Νικόλαος	Πότσιου Χρυσή	Κεπαπτσόγλου Κωνσταντίνος
Καθηγητής	Καθηγήτρια	Αναπληρωτής Καθηγητής

National Technical University of Athens
School of Rural, Surveying and Geoinformatics Engineering
Photogrammetry Lab

DIPLOMA THESIS

PREDICTION OF RENTAL PRICES FOR IMMOVABLE PROPERTY USING MACHINE LEARNING FOR THE FORMULATION OF TRANSPORT POLICY

Dimitra - Alexandra Malapani

Supervisor: Nikolaos Doulamis

Three-member committee

Doulamis Nikolaos	Potsiou Chryssy	Kepaptsoglou Konstantinos
Professor	Professor	Associate Professor

Περίληψη

Στην σημερινή εποχή, η αγορά των ακινήτων παρουσιάζει ιδιαίτερη οικονομική αστάθεια ,η οποία επηρεάζεται από πλήθος παραγόντων. Οι παράγοντες αυτοί που επηρεάζουν το κάθε ακίνητο είναι πολυάριθμοι και σχετίζονται όλοι μεταξύ τους. Συνεπώς, γίνεται κατανοητό πως η μοντελοποίηση τους αποτελεί μια ιδιαίτερα σύνθετη διαδικασία . Στην παρούσα διατριβή, μελετάται η πρόβλεψη τιμών ενοικίασης ακινήτων, επικεντρώνοντας το ενδιαφέρον της μελέτης στην περιοχή της Αττικής, λαμβάνοντας υπόψιν σημαντικούς παράγοντες από τους οποίους καθορίζονται οι τιμές ενοικίασης των ακινήτων και χρησιμοποιώντας μεθόδους και τεχνικές της μηχανικής μάθησης για την χάραξη συγκοινωνιακής πολιτικής ταυτόχρονα.Αναλύονται, στη συνέχεια, τα αποτελέσματα που λάβαμε από τις μεθόδους που εφαρμόσαμε και αξιολογούνται η ακρίβεια και η καταλληλότητά αυτών για το παρόν πρόβλημα. Πιο συγκεκριμένα, γίνεται ιστορική αναδρομή σε παλαιότερες μελέτες από τις οποίες μπορούμε να αντλήσουμε πληροφορίες , ενώ στην συνέχεια παρουσιάζεται αναλυτικά το πρόβλημα με το οποίο καταπιάνεται η εργασία. Συνεχίζουμε με την παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε και του τρόπου συλλογής αυτού λεπτομερώς. Έπεται η λεπτομερής ανάλυση και διερευνητική προεπεξεργασία του, έτσι ώστε να είναι δυνατή η πλήρης κατανόησή του και η βέλτιστη αξιοποίησή του. Στη συνέχεια, υλοποιούνται εννέα διαφορετικά μοντέλα μηχανικής μάθησης, τα οποία εκπαιδεύονται με βάση τα δεδομένα. Ακολούθως, παρουσιάζονται και αξιολογούνται τα αποτελέσματα της εφαρμογής τους στα δεδομένα εκπαίδευσης και τα δεδομένα ελέγχου. 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 επερχόμενες μελέτες.

Λέξεις - Κλειδιά

Πρόβλεψη τιμών ενοικίασης ακινήτων, Χάραξη συγκοινωνιακής πολιτικής,Παλινδρόμηση, Μηχανική μάθηση, Τεχνητή Νοημοσύνη

Abstract

In today's era, the real estate market presents particular economic instability, which is influenced by a number of factors. These factors that affect each property are numerous and are all related to each other. It is therefore clear that their modelling is a particularly complex process. In this thesis, we study the prediction of property rental prices, focusing the interest of the study on the region of Attica, taking into account important factors that determine the rental prices of properties and using machine learning methods and techniques for the formulation of transport policy at the same time.We then analyze the results obtained from the methods applied and evaluate their accuracy and suitability for the present problem. In particular, we provide a historical review of previous studies from which we can draw information , and then present in detail the problem addressed in the paper. We continue by presenting the dataset used and the way it was collected in detail. This is followed by a detailed analysis and exploratory pre-processing of it, so that it can be fully understood and optimally used.Nine different machine learning models are then implemented and trained on the data. Subsequently, the results of their implementation on the training and control data are presented and evaluated. Finally, we present some comparisons of our results with results of similar researches, identify the points for improvement in the methodology followed and suggest some future research perspectives for upcoming studies based on them.

Keywords

Property rental price prediction, Transport policy making, Regression, Machine learning, Artificial intelligence

Ευχαριστίες

Με την ευκαιρία ολοκλήρωσης της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω αρχικά τον Καθηγητή κ.Νικόλαο Δουλάμη και τον Αναπληρωτή Καθηγητή κ.Αναστάσιο Δουλάμη , που μου έδωσαν την ευκαιρία να εκπονήσω την εργασία μου στο Εργαστήριο Φωτογραμμετρίας.

Έπειτα, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή κ.Κεπαπτζόγλου Κωνσταντίνο καθώς και την Καθηγήτρια κ.Χρυσή Πότσιου που με την διδασκαλία τους μου παραχώρησαν σημαντικές γνώσεις ώστε να φέρω εις πέρας την προκείμενη εργασία.

Ευχαριστώ επίσης κάθε Καθηγητή της σχολής που μοιράστηκε μαζί μου το πάθος του για το αντικείμενο της τοπογραφίας.

Ευχαριστώ ακόμα όσους φίλους είχα δίπλα μου να με στηρίζουν αυτά τα όμορφα και αλησμόνητα χρόνια.

Μα κυρίως,θα ήθελα να ευχαριστήσω την οικογένειά μου διότι όσα πετυχαίνω καθημερινά τα οφείλω σε εκείνους.

Δήμητρα - Αλεξάνδρα Μαλαπάνη

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων-γραφημάτων

Κεφάλαιο 1ο-Εισαγωγή

1.1 Η σημασία της αυτόματης πρόβλεψης τιμών ενοικίασης

Για τον άνθρωπο, η σημασία της διαβίωσης σε έναν καλό και αξιοπρεπή τόπο δεν μπορεί να υποεκτιμηθεί. Είτε πρόκειται για προσωρινή κατοικία για διακοπές ή εργασία, είτε για μόνιμη κατοικία, το περιβάλλον μέσα στο οποίο ζει έχει σημαντικές επιπτώσεις στην υγεία, την ευημερία και γενικά στη ποιότητα ζωής του. Ένας άνετος, ασφαλής και λειτουργικός χώρος διαβίωσης είναι απαραίτητος για τη σωματική και ψυχική υγεία και μπορεί ακόμη και να επηρεάσει την κοινωνική και συναισθηματική ανάπτυξη αυτού. Στον σημερινό κόσμο που η καθημερινότητα του μέσου ανθρώπου κινείται με τόσο γοργούς ρυθμούς, η σημασία της εύρεσης ενός καλού τόπου διαβίωσης είναι ακόμα πιο σημαντική. Με την αυξανόμενη αστικοποίηση και την άνοδο της τεχνολογίας, πολλοί άνθρωποι βρίσκονται σε μικρά διαμερίσματα ή στενούς χώρους διαβίωσης,κάτι που μπορεί να οδηγήσει σε συναισθήματα απομόνωσης και αποσύνδεσης από τον φυσικό κόσμο γεγονός που μπορεί να έχει αρνητικό αντίκτυπο στην ψυχική και συναισθηματική του υγεία. Από την άλλη, ένα καλό μέρος διαβίωσης, παρέχει μια αίσθηση άνεσης και ασφάλειας, που επιτρέπει στον άνθρωπο, να ηρεμήσει και να χαλαρώσει μετά από μια κουραστική μέρα και ακόμα και να επανασυνδεθεί με τον εαυτό του.

Κατ' επέκταση, η ποιότητα ενός χώρου διαβίωσης επηρεάζει άμεσα την καθημερινή ζωή και τη ρουτίνα του κάθε ανθρώπου. Είναι λοιπόν, ζωτικής σημασίας για τους ανθρώπους να επενδύσουν στον καλύτερο δυνατό χώρο διαβίωσης που μπορούν να υποστηρίξουν οικονομικά (είτε για προσωρινή διαμονή είτε σε μόνιμη κατοικία) για να εξασφαλίσουν τη σωματική, ψυχική και συναισθηματική τους ευεξία.

Σε αυτήν λοιπόν την απαραίτητη ανάγκη που έχει ο άνθρωπος έρχεται να συνεισφέρει η μηχανική μάθηση, η οποία εφαρμόζεται σφόδρα στη σφαίρα των ακινήτων, και πιο συγκεκριμένα στην πρόβλεψη των τιμών των κατοικιών, είτε αγοράς είτε ενοικίασης. Είναι αρκετά ξεκάθαρο το γεγονός ότι η χρήση της μηχανικής μάθησης στην πρόβλεψη των τιμών ενοικίασης των κατοικιών είναι μια σημαντική εξέλιξη στον κλάδο των ακινήτων. Παρέχοντας πιο ακριβείς και ενημερωμένες πληροφορίες σχετικά με τις αξίες των ακινήτων, η μηχανική μάθηση μπορεί να βοηθήσει τους αγοραστές, τους πωλητές και τους επενδυτές να λάβουν πιο ενημερωμένες αποφάσεις και να εξοικονομήσουν χρήματα. Καθώς οι αλγόριθμοι μηχανικής εκμάθησης συνεχίζουν να βελτιώνονται και περισσότερα δεδομένα γίνονται διαθέσιμα, μπορούμε να περιμένουμε να δούμε ακόμη μεγαλύτερες προόδους σε αυτόν τον τομέα στο επερχόμενο μέλλον.

1.2 Στόχος Διπλωματικής

Ο στόχος αυτής της διπλωματικής είναι η δημιουργία ενός μοντέλου μηχανικής μάθησης που να μπορεί να προβλέψει με ακρίβεια τις τιμές ενοικίασης κατοικιών στην ευρύτερη περιοχή της Αττικής. Αυτό το μοντέλο θα βασίζεται σε ένα μεγάλο σύνολο δεδομένων που έχουν συλλεχθεί από τον επίσημο ιστότοπο της Spitogatos. Το πρώτο βήμα για τη δημιουργία αυτού του μοντέλου

θα είναι η συγκέντρωση και ο καθαρισμός του συνόλου δεδομένων. Αυτό θα περιλαμβάνει τον εντοπισμό σχετικών μεταβλητών, όπως η τοποθεσία, το μέγεθος του καταλύματος ,ο αριθμός των λουτρών ,η ύπαρξη χώρου στάθμευσης κτλ., καθώς και η αφαίρεση τυχόν άσχετων ή ελλιπών δεδομένων. Στη συνέχεια, το καθαρισμένο σύνολο δεδομένων θα χωριστεί σε ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου και το σύνολο ελέγχου για την αξιολόγηση της απόδοσής του. Το μοντέλο μηχανικής μάθησης θα αναπτυχθεί χρησιμοποιώντας έναν από του κυρίαρχους αλγορίθμους, όπως ένα δέντρο παλινδρόμησης, ένα τυχαίο δάσος παλινδρόμησης, μία μηχανή διανυσματικής υποστήριξης κτλ. Στην συνέχεια το μοντέλο θα εκπαιδευτεί και θα βελτιστοποιηθεί μέσω μιας διαδικασίας τελειοποίησης.

Αυτό θα περιλαμβάνει την προσαρμογή των παραμέτρων του μοντέλου στα δεδομένα εκπαίδευσης για την αύξηση της απόδοσής του. Στόχος αυτής της διαδικασίας είναι να επιτευχθεί η υψηλότερη δυνατή ακρίβεια στην πρόβλεψη των τιμών των ενοικίων των κατοικιών. Τέλος, το εκπαιδευμένο και τελειοποιημένο μοντέλο θα εξεταστεί σε πραγματικό σενάριο. Τέλος ,συνοψίζοντας τις επιδόσεις όλων των μοντέλων μηχανικής μάθησης που επιλέχθηκαν για την επίλυση του συγκεκριμένου προβλήματος, η διπλωματική εργασία καταλήγει σε ένα συμπέρασμα, καθώς επίσης παρουσιάζονται και μελλοντικές τροποποιήσεις στην αντιμετώπιση του προβλήματος παλινδρόμησης.

Η ανάπτυξη της παρούσας διπλωματικής εργασίας οργανώνεται ως εξής. Στο 1ο κεφάλαιο γίνεται μια εισαγωγή σχετικά με το θέμα και το πρόβλημα που μελετάται. Ορίζεται η σημασία του προβλήματος, οι στόχοι και η συνεισφορά της εργασίας.Στο 2ο κεφάλαιο πραγματοποιείται μια βιβλιογραφική ανασκόπηση σε παλαιότερες μελέτες. Στο 3ο κεφάλαιο, αναλύονται κάποιες βασικές έννοιες σχετικά με το θέμα. Περιγράφονται κάποιοι βασικοί όροι και παρουσιάζεται η απαραίτητη πληροφορία σχετικά με την Μηχανική Μάθηση και την Επεξηγήσιμη Τεχνητή Νοημοσύνη.Επίσης αναλύονται τα εκ φύσεως ερμηνεύσιμα μοντέλα μηχανικής μάθησης, στα οποία δεν χρειάζεται η επεξηγήσιμη τεχνητή νοημοσύνη. Παρουσιάζονται μοντέλα Γραμμικής Παλινδρόμησης, Δέντρων Αποφάσεων και , SHAP. Στο κεφάλαιο 4, παρουσιάζεται η γλώσσα προγραμματισμού Python που μας συνόδευσε σε όλη την διατριβή καθώς επίσης παρουσιάζονται πειραματικές δοκιμές αυτών των μεθόδων μηχανικής μάθησης και επεξηγήσιμης τεχνητής νοημοσύνης, πάνω σε συγκεκριμένα σύνολα δεδομένων. Δείχνεται ο τρόπος που αυτές οι τεχνικές παρέχουν τις επεξηγήσεις στους χρήστες. Το κεφάλαιο 5, αποτελεί μια σύντομη παρουσίαση της πειραματικής διαχείρισης των δεδομένων μας ενώ τέλος, παρουσιάζονται τα συμπεράσματα της παρούσας διπλωματικής εργασίας, οι περιορισμοί της και οι προτεινόμενοι κατευθυντήριοι άξονες για μελλοντική έρευνα και ανάπτυξη.Η πιο σημαντική διαφορά της προκείμενης διπλωματικής με άλλες μελέτες είναι πως παρέχονται ενεργειακά δεδομένα για το κάθε ακίνητο καθώς και στοιχεία συγκοινωνιακής φύσεως,που λείπουν από όσες μελέτες κατάφερα να συμβουλευτώ.

Κεφάλαιο 2ο-Υπάρχουσα Βιβλιογραφία

Στο συγκεκριμένο κεφάλαιο, πραγματοποιείται βιβλιογραφική ανασκόπηση σε παλαιότερες εργασίες που έχουν αντιμετωπίσει με διάφορους τρόπους το πρόβλημα της παλινδρόμησης επιβλεπόμενης μάθησης .Έχοντας ως βάση τις κύριες προκλήσεις της προσέγγισης του προβλήματος από το προηγούμενο κεφάλαιο, αλλά και τη σημασία της μοντελοποίησης της σχέσης μεταξύ της εξαρτώμενης μεταβλητής, στην προκειμένη περίπτωση της τιμής ενοικίασης των καταλυμάτων, και των ανεξάρτητων μεταβλητών, που αποτελούν τα χαρακτηριστικά του κάθε καταλύματος, αξίζει να μελετηθεί η μεθοδολογία αντιμετώπισής του και από άλλους ερευνητές. Οι μελέτες στις οποίες στηρίχθηκε η παρούσα έρευνα παρουσιάζονται στη συνέχεια.

[1] Η θεματική που θίγεται σε αυτήν την εργασία, αφορά υποκατηγορία της επιστήμης των υπολογιστών που ονομάζεται μηχανική μάθηση. Η μηχανική μάθηση, στηρίζεται σε μοντέλα της στατιστικής, κλάδου της επιστήμης των μαθηματικών και χρησιμοποιείται για την μελέτη και υλοποίηση αλγορίθμων που έχουν την δυνατότητα να «μαθαίνουν» από τα δεδομένα, να μοντελοποιούν την φύση τους και με βάση αυτήν να κάνουν τις εκάστοτε προβλέψεις. Το πρόβλημα που μελετάται στην παρούσα εργασία, αποτελεί ένα πρόβλημα παλινδρόμησης επιβλεπόμενης μάθησης. Η μοντελοποίηση της σχέσης μεταξύ της εξαρτώμενης μεταβλητής, στην προκειμένη περίπτωση της τιμής ενοικίασης των καταλυμάτων, και των ανεξάρτητων μεταβλητών, που αποτελούν τα χαρακτηριστικά του κάθε καταλύματος, συνιστά τον απώτερο σκοπό της. Εδώ και αρκετά χρόνια, πραγματοποιούνται προσπάθειες και γίνονται μελέτες σχετικά με το παραπάνω πρόβλημα επιστρατεύοντας τεχνικές μηχανικής μάθησης για την επίλυση τού. Πολλοί μελετητές και επιστήμονες, των οποίων το έργο υπάρχει δημοσιευμένο έχουν προσεγγίσει το παραπάνω πρόβλημα. Οι μελέτες στις οποίες στηρίχθηκε η παρούσα έρευνα παρουσιάζονται στη συνέχεια.

Τα ξενοδοχεία κυριαρχούσαν εδώ και χρόνια στον κλάδο της φιλοξενίας μέχρι που εμφανίστηκαν επιχειρήσεις κοινής χρήσης, όπως η Airbnb. Η Airbnb, που ιδρύθηκε το 2008, παρουσίασε ένα νέο επιχειρηματικό μοντέλο οικονομίας διαμοιρασμού που έφερε επανάσταση στον κλάδο της φιλοξενίας και συνδέει τους ιδιοκτήτες ελεύθερων καταλυμάτων με ταξιδιώτες που αναζητούν προσωρινή διαμονή [2]. Από το 2019, υπάρχουν πάνω από 6 εκατομμύρια καταχωρήσεις στον ιστότοπο του Airbnb σε περίπου 220 χώρες και περιοχές, πραγματοποιώντας κατά μέσο όρο πάνω από 1 εκατομμύρια διαμονές ανά βραδιά [3].

Ο [4] εξηγεί ότι τα εργαλεία δυναμικής τιμολόγησης καταλυμάτων του Airbnb βασίζονται σε τεχνητή νοημοσύνη. Αναφέρει μάλιστα, πώς το αρχικό εργαλείο που κυκλοφόρησε το 2015 βασιζόταν σε τεχνικές παλινδρόμησης και χρησιμοποιούσε ως είσοδο τις ανέσεις (amenities) ενός καταλύματος και πληροφορίες σχετικά με τα γειτονικά ακίνητα. Πλέον, έχει αντικατασταθεί με εργαλείο της εταιρίας που βασίζεται στην ενισχυτική μάθηση (Reinforcement Learning). Παρόλα αυτά έχει ιδιαίτερο ενδιαφέρον ότι και η ίδια εταιρία που διαθέτει τον πλήρη όγκο δεδομένων στην

αρχή επέλεξε να δημιουργεί προτάσεις τιμών χρησιμοποιώντας παραδοσιακές τεχνικές μηχανικής μάθησης.

Το 2019 οι [5], με τη χρήση των δεδομένων που παρέχονται από την πλατφόρμα Inside Airbnb, μελετούν την πρόβλεψη ενοικίασης των καταλυμάτων Airbnb στην Μελβούρνη της Αυστραλίας με την χρήση μοντέλων μηχανικής μάθησης. Η έρευνά τους περιλαμβάνει την σύγκριση διάφορων μοντέλων πρόβλεψης τιμής, όπως τα νευρωνικά δίκτυα, καθώς και παραδοσιακών μεθόδων μηχανικής μάθησης, όπως είναι τεχνικές παλινδρόμησης, τα τυχαία δάση (Random Forest) και η ενισχυτική κλίση (Gradient Boosting). Υπολογίζοντας το Μέσο Τετραγωνικό Σφάλμα (RMS) και τον συντελεστή προσδιορισμού R 2 , αξιολογούν τα παραπάνω μοντέλα και διαπιστώνουν ότι καλύτερη απόδοση έχει η μέθοδος παλινδρόμησης με ενισχυτική κλίση, ενώ αμέσως μετά έρχεται η μέθοδος των τυχαίων δασών, η οποία ενδεχομένως να είχε βελτιωμένη απόδοση με αυστηρότερη επιλογή χαρακτηριστικών.

Ακόμα, για να μπορέσουν να καθορίσουν με ακρίβεια την τιμή ενοικίασης ενός καταλύματος για έναν host, πολλοί ερευνητές χρησιμοποιούν διάφορες μεθόδους με τρία κύρια στοιχεία:

i) Ένα δυαδικό μοντέλο ταξινόμησης προβλέπει την πιθανότητα κράτησης κάθε διανυκτέρευσης

ii) Ένα μοντέλο παλινδρόμησης προβλέπει το ιδανικό κόστος για κάθε διανυκτέρευση Μεταπτυχιακή Διατριβή Γεώργιος Σερβετάς Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ

iii) Εξατομικευμένο συλλογισμό πάνω στην πρόβλεψη του μοντέλου παλινδρόμησης για την παροχή των τελευταίων προτάσεων τιμής ανάλογα με τους στόχους του host [6]

Την παραπάνω προσέγγιση επιβεβαιώνει προγενέστερη έρευνα των [7], που χρησιμοποιούν μοντέλα παλινδρόμησης για την μεγιστοποίηση των εσόδων ενός host. Πιο συγκεκριμένα, εφαρμόζουν τον αλγόριθμο Gradient Boosting Machine (GBM) για να προβλέψουν την πιθανότητα κράτησης ενός καταλύματος. Στη συνέχεια, δημιούργησαν ένα μοντέλο παλινδρόμησης σχετικό με την πρόβλεψη τιμής για το εκάστοτε βράδυ και τέλος, προσάρμοζαν τις προτάσεις του μοντέλου παλινδρόμησης στους τους προσωπικούς στόχους που εκάστοτε host.

Οι [8] σύγκριναν μοντέλα όπως τυχαία δάση (Random Forest), μετεξέλιξη της ενισχυτικής κλίσης (XGBoost) και νευρωνικά δίκτυα (Neural Networks) στην πρόβλεψη τιμών Airbnb πάνω σε δεδομένα της Νέας Υόρκης και του Παρισιού. Αξίζει να σημειωθεί ότι στο στάδιο της προεπεξεργασίας, κατάργησαν πολλά χαρακτηριστικά για να μειώσουν το θόρυβο και να δώσουν έμφαση σε χαρακτηριστικά όπως country_code και τον αριθμό υπνοδωματίων, τα οποία και θεώρησαν ως πιο ουσιώδη. Ακόμα, δεδομένα κειμένου, όπως η περιγραφή του καταλύματος, θεωρούνται επίσης χρήσιμα χαρακτηριστικά. Παρόλα αυτά όμως τέτοια χαρακτηριστικά απαιτούν ιδιαίτερη προεπεξεργασία καθώς είναι σε αδόμητη μορφή. Τέλος, μετά την εκτέλεση εκτεταμένης προεπεξεργασίας και καθαρισμού δεδομένων, το μοντέλο XGBoost πέτυχε την καλύτερη απόδοση. Ως μετρικές αξιολόγησης των μοντέλων χρησιμοποιήθηκαν οι: R 2 και το μέσο τετραγωνικό σφάλμα (MSE).

Οι ερευνητές στο Πανεπιστημίου του Στάνφορντ πραγματοποιούν παρόμοια ερεύνα, συγκρίνουν επίσης πολλαπλές μεθόδους μηχανικής μάθησης που περιλαμβάνουν: Γραμμική Παλινδρόμηση, μοντέλα που βασίζονται σε Δέντρα Παλινδρόμησης, Διανυσματικές Μηχανές Υποστήριξης για Παλινδρόμηση (SVR) και K-means (KMC). Οι κύριες συνεισφορές αυτής της εργασίας είναι ότι χρησιμοποιούν τεχνικές επιλογής χαρακτηριστικών (feature selection), οι οποίες και δίνουν τα 22 καλύτερα χαρακτηριστικά για την πρόβλεψη της τιμής ενοικίασης. Επίσης, πρόσθεσαν ανάλυση συναισθήματος για να εξετάσουν τις κριτικές των πελατών [9]. Ωστόσο, αυτή η έρευνα δεν προσέφερε τελικά ένα ολοκληρωμένο μοντέλο πρόβλεψης τιμών. Η Laura Lewis στην δουλειά της το 2019 μελετά επίσης τον καθορισμό τιμών σε καταχωρήσεις της Airbnb στο Λονδίνο κυρίως μέσο της μεθόδου XGBoost. Στην προσέγγιση της δίνει ιδιαίτερη έμφαση στις εργασίες προεπεξεργασίας, καθαρισμού και ανάλυση των δεδομένων [10].

Όπως γίνεται αντιληπτό, πολλές μέθοδοι μηχανικής μάθησης έχουν εφαρμοστεί για την πρόβλεψη της αξίας ενός καταλύματος. Παρόλα αυτά, το σημαντικότερο στοιχείο στην επίτευξη ανταγωνιστικών αποτελεσμάτων αποτελεί η σωστή επιλογή χαρακτηριστικών [11]. Κατ' επέκταση, πολλές έρευνες εργάστηκαν σχετικά με το ποια χαρακτηριστικά επηρεάζουν την τιμή ενοικίασης του καταλύματος. Ορισμένες από αυτές, επιλέγουν να χωρίζουν τα χαρακτηριστικά ενός καταλύματος σε δύο κατηγορίες. Στα host-controlled χαρακτηριστικά, που είναι χαρακτηριστικά που παρέχονται από τον host είτε αφορούν πληροφορίες σχετικά με τον ίδιο τον host. Και στα out_of_host_controlled χαρακτηριστικά, που αφορούν πληροφορίες που δεν μπορεί να επηρεάσει ο host. Σύμφωνα με τον Brando MaNeil, τόσο τα host_controlled όσο και τα out_of_host_controlled χαρακτηριστικά είναι εξίσου σημαντικά για τον καθορισμό της τιμής ενοικίασης ενός καταλύματος στην Airbnb. Διαπίστωσε επίσης, ότι όταν συνδυάζονται χαρακτηριστικά host_controlled και χαρακτηριστικά out_of_host_controlled προβλέπεται με μεγαλύτερη ακρίβεια η τιμή ενοικίασης του εκάστοτε καταλύματος. Ωστόσο, τα host_controlled χαρακτηριστικά εξακολουθούν να έχουν μεγαλύτερη σημασία για την τιμή καταχώρισης από ό,τι τα out_of_host_controlled χαρακτηριστικά [12].

Προηγούμενη έρευνα έδειξε ότι η ενοικίαση ενός ολόκληρου σπιτιού θα έχει υψηλότερα μέσα έσοδα σε σχέση με την ενοικίαση του κάθε δωματίου ξεχωριστά. Επίσης, πιο επαγγελματίες hosts Μεταπτυχιακή Διατριβή Γεώργιος Σερβετάς Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ 15 αποκτούν κατά μέσο όρο περισσότερες κριτικές και τείνουν να έχουν υψηλότερα μηνιαία έσοδα. Στην Airbnb, αυτό το χαρακτηριστικό εμφανίζεται ως "super-host" [13]. Ακόμα, οι [14] επιβεβαίωσαν ότι χαρακτηριστικά που σχετίζονται με τοποθεσίες, τις υπηρεσίες καθώς και τις κριτικές πελατών θα μπορούσαν να επηρεάσουν τις τιμές των ενοικιαζόμενων καταλυμάτων. Προσδιόρισαν τα χαρακτηριστικά του host (host_controlled) ως σημαντικούς καθοριστικούς παράγοντες της τιμής.

Κεφάλαιο 3ο-Βασικές έννοιες Τεχνητής Νοημοσύνης

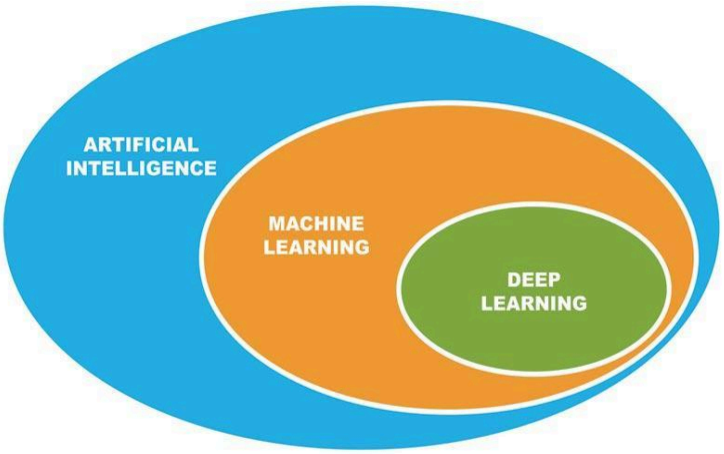
Η Τεχνητή Νοημοσύνη (ΤΝ) αναφέρεται στην ανάπτυξη υπολογιστικών συστημάτων ή μηχανών που μπορούν να εκτελούν εργασίες που συνήθως απαιτούν ανθρώπινη νοημοσύνη. Τα καθήκοντα αυτά περιλαμβάνουν τη μάθηση, τη συλλογιστική, την επίλυση προβλημάτων, την κατανόηση της φυσικής γλώσσας, την αντίληψη (μέσω της όρασης, της ομιλίας ή άλλων αισθητηριακών εισροών) και την αλληλεπίδραση με το περιβάλλον με ουσιαστικό τρόπο. Ο απώτερος στόχος της τεχνητής νοημοσύνης είναι η δημιουργία μηχανών που μπορούν να μιμούνται, να προσομοιώνουν ή να αναπαράγουν τις γνωστικές ικανότητες του ανθρώπου.

Η τεχνητή νοημοσύνη μπορεί σε γενικές γραμμές να κατηγοριοποιηθεί σε δύο τύπους:

- Αδύναμη τεχνητή νοημοσύνη (Weak AI): Αυτός ο τύπος ΤΝ έχει σχεδιαστεί και εκπαιδευτεί για μια συγκεκριμένη εργασία. Είναι άριστη στην εκτέλεση της συγκεκριμένης εργασίας, αλλά δεν διαθέτει τις γενικές γνωστικές ικανότητες ενός ανθρώπου. Παραδείγματα αποτελούν οι εικονικοί προσωπικοί βοηθοί, τα συστήματα συστάσεων και τα συστήματα αναγνώρισης εικόνων.
- Γενική τεχνητή νοημοσύνη (ισχυρή τεχνητή νοημοσύνη): Πρόκειται για μια προηγμένη μορφή ΤΝ που διαθέτει την ικανότητα να κατανοεί, να μαθαίνει και να εφαρμόζει τη γνώση σε ένα ευρύ φάσμα εργασιών, παρόμοια με την ανθρώπινη νοημοσύνη. Η επίτευξη της γενικής ΤΝ παραμένει θεωρητικός στόχος και αποτελεί αντικείμενο συνεχούς έρευνας και εξερεύνησης.

Υπάρχουν διάφορες προσεγγίσεις για την εφαρμογή της τεχνητής νοημοσύνης, και μερικές από τις σημαντικότερες περιλαμβάνουν:

- Μηχανική μάθηση (ML): ΜΑ: Ένα υποσύνολο της ΤΝ που περιλαμβάνει την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να βελτιώνουν τις επιδόσεις τους με την πάροδο του χρόνου χωρίς να προγραμματίζονται ρητά.
- Βαθιά μάθηση (Deep Learning): Ένας συγκεκριμένος τύπος μηχανικής μάθησης που χρησιμοποιεί νευρωνικά δίκτυα με πολλά επίπεδα (βαθιά νευρωνικά δίκτυα) για τη μοντελοποίηση και την επίλυση σύνθετων προβλημάτων. Η βαθιά μάθηση έχει σημειώσει ιδιαίτερη επιτυχία σε εργασίες όπως η αναγνώριση εικόνας και ομιλίας.
- Επεξεργασία φυσικής γλώσσας (NLP): Ένας τομέας της τεχνητής νοημοσύνης που επικεντρώνεται στο να επιτρέπει στις μηχανές να κατανοούν, να ερμηνεύουν και να παράγουν την ανθρώπινη γλώσσα. Αυτό είναι ζωτικής σημασίας για εφαρμογές όπως τα chatbots, η γλωσσική μετάφραση και η ανάλυση συναισθήματος.
- Computer Vision: Κλάδος της ΤΝ που επιτρέπει στις μηχανές να ερμηνεύουν και να λαμβάνουν αποφάσεις με βάση οπτικά δεδομένα. Χρησιμοποιείται στην αναγνώριση προσώπου, στην ανίχνευση αντικειμένων και στα αυτόνομα οχήματα.



Εικόνα 1:Υποσύνολα Τεχνητής Νοημοσύνης

3.1 Μηχανική Μάθηση

3.1.1 Ορισμός και Βασικές Αρχές Μηχανικής Μάθησης

Ο ανθρώπινος νους επιχειρεί να κατανοήσει το περιβάλλον του μέσω της παρατήρησης και δημιουργώντας μία απλοποιημένη εκδοχή που ονομάζεται "μοντέλο". Η δημιουργία ενός τέτοιου μοντέλου, ορίζεται ως "επαγωγική μάθηση", ενώ γενικότερα η διαδικασία ονομάζεται "επαγωγή". Επιπλέον, ο άνθρωπος χαρακτηρίζεται από την ικανότητα που έχει να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις που αποκομίζει , δημιουργώντας ολοκαίνουριες δομές που ονομάζονται "πρότυπα". Η δημιουργία, συνεπώς, μοντέλων ή προτύπων από ένα σύνολο δεδομένων ονομάζεται Μηχανική Μάθηση. Η μηχανική μάθηση είναι μια μέθοδος ανάλυσης δεδομένων που αυτοματοποιεί την ανάπτυξη αναλυτικών μοντέλων. Είναι ένας κλάδος της τεχνητής νοημοσύνης που βασίζεται στην ιδέα ότι τα συστήματα μπορούν να μάθουν από τα δεδομένα, να εντοπίσουν μοτίβα και να λάβουν αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση.

Ο Tom M. Mitchell πρότεινε σαν ορισμό της μάθησης τον εξής : «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία Ε ως προς μια κλάση εργασιών Τ και ένα μέτρο επίδοσης Ρ, αν η επίδοσή του σε εργασίες της κλάσης Τ, όπως αποτιμάται από το μέτρο Ρ, βελτιώνεται με την εμπειρία Ε»

Οι εργασίες μηχανικής μάθησης ταξινομούνται κυρίως σε τρεις μεγάλες κατηγορίες, σύμφωνα με τη φύση του εκπαιδευτικού «σήματος » που είναι διαθέσιμο σε ένα σύστημα εκμάθησης .Αυτές είναι:

- **Επιτηρούμενη μάθηση** (είτε επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη)(supervised learning): Το πρόγραμμα (υπολογιστικό) δέχεται τις παραδειγματικές εισόδους και τα επιθυμητά αποτελέσματα από έναν <<δάσκαλο>> και στόχος αυτού είναι να μάθει έναν γενικό κανόνα ώστε να αντιστοιχίσει τις εισόδους κατάλληλα με τα αποτελέσματα .

- **Μη επιτηρούμενη μάθηση**(αλλιώς επίβλεπτη μάθηση ή μάθηση χωρίς επίβλεψη)(unsupervised learning):Πρέπει να βρεθεί η δομή των δεδομένων εισόδου χωρίς την παροχή εμπειρίας στον αλγόριθμο μάθησης .Η παραπάνω μάθηση μπορεί να είναι αυτοσκοπός(ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή μέσο για ένα τέλος (χαρακτηριστικό της μάθησης).

- **Ενισχυτική μάθηση** : Ένα υπολογιστικό πρόγραμμα στην προκειμένη αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας στόχος χωρίς κάποιος να του λέει αν έχει φτάσει κοντά σε αυτόν.

Για ένα σύνολο δεδομένων εισόδου, ένας αλγόριθμος μηχανικής μάθησης μπορεί να εκτελέσει τις ακόλουθες βασικές εργασίες (Τ):

Ταξινόμηση (Classification): Όπου τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχεί τα δεδομένα σε μία ή περισσότερες κλάσεις, κάτι που συνήθως εμπίπτει στην επιτηρούμενη μάθηση που αναλύσαμε παραπάνω .

Παλινδρόμηση (Regression): Ο αλγόριθμος μηχανικής μάθησης προσπαθεί να γραμμικοποιήσει τις τιμές εισόδου. Επίσης πρόβλημα επιτηρούμενης μάθησης, με τα αποτελέσματα α είναι συνεχή και όχι διακριτά.

Συσταδοποίηση (Clustering): Ο αλγόριθμος μηχανικής μάθησης ζητείται χωρίσει σε ομάδες ένα σύνολο εισόδων. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εξ αρχής, καθιστώντας έτσι τον διαχωρισμό αυτό εργασία μη επιτηρούμενης μάθησης.

Εκτίμηση πυκνότητας (**Density estimation):** Όπου ο αλγόριθμος βρίσκει την κατανομή των δεδομένων εισόδου σε κάποιο χώρο .

Προβλήματα **μείωσης διαστασιμότητας (Dimensionality reduction) :** Τα δεδομένα απλοποιούνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων . Το στατιστικό μοντέλο θεμάτων είναι ένα σχετικό πρόβλημα , όπου ο υπολογιστής καλείται να εντοπίσει έγγραφα που καλύπτουν άλλα παρόμοια θέματα από ένα σύνολο εγγράφων που είναι γραμμένα σε φυσική γλώσσα .[15]

3.2 Επεξηγήσιμη Τεχνητή Νοημοσύνη

Η ραγδαία εξέλιξη της τεχνητής νοημοσύνης (AI) και της μηχανικής μάθησης έχει επιφέρει επαναστατικές αλλαγές σε πολλές βιομηχανίες και οργανισμούς, καθιστώντας την ως ένα αναπόσπαστο κομμάτι της καθημερινότητας. Πλέον, τα συστήματα τεχνητής νοημοσύνης εμπεριέχονται σε όλο και περισσότερες εφαρμογές, έχοντας μεταμορφώσει τον τρόπο με τον οποίο οι άνθρωποι αλληλεπιδρούν με την τεχνολογία. Ωστόσο, μια σημαντική πρόκληση που δημιουργείται, είναι η έλλειψη διαφάνειας και ερμηνευσιμότητας στα σύγχρονα μοντέλα μηχανικής μάθησης, καθώς τα περισσότερα και πιο χρήσιμα από αυτά αποτελούν ‘’μαύρα κουτιά’’, στα οποία ούτε οι ίδιοι οι δημιουργοί τους δεν γνωρίζουν πως ακριβώς λειτουργούν. Ιδιαίτερα σε συστήματα που αφορούν κρίσιμους τομείς, αξίες όπως η εμπιστοσύνη, η ακεραιότητα και η λογοδοσία είναι ύψιστης σημασίας.

Η Επεξηγήσιμη Τεχνητή Νοημοσύνη (Explainable Artificial Intelligence, XAI) αναφέρεται στην προσπάθεια να γίνουν τα μοντέλα και τα συστήματα τεχνητής νοημοσύνης πιο κατανοητά, διαφανή και ερμηνεύσιμα από τον άνθρωπο. Ο στόχος της XAI είναι να παράσχει πληροφορίες σχετικά με τον τρόπο με τον οποίο τα μοντέλα τεχνητής νοημοσύνης λαμβάνουν αποφάσεις, ιδίως σε πολύπλοκες ή κρίσιμες εφαρμογές, όπου η κατανόηση της λογικής πίσω από τις προβλέψεις ή τις ενέργειες της τεχνητής νοημοσύνης είναι απαραίτητη.

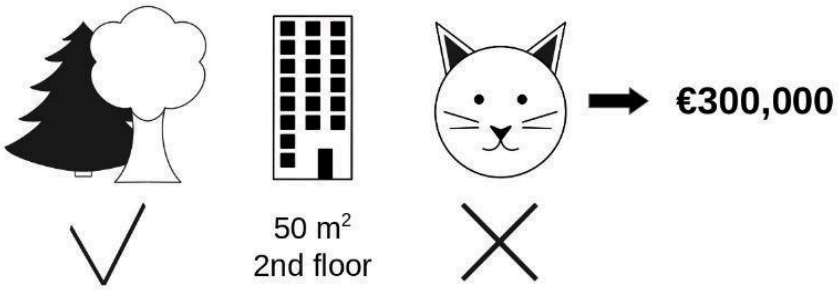
3.2.1 Shapley Values

[16]Μια πρόβλεψη μπορεί να εξηγηθεί υποθέτοντας ότι κάθε τιμή χαρακτηριστικού της περίπτωσης είναι ένας "παίκτης" σε ένα παιχνίδι όπου η πρόβλεψη είναι η πληρωμή. Οι τιμές Shapley - μια μέθοδος από τη θεωρία συμμαχικών παιγνίων - μας λέει πώς να κατανέμουμε δίκαια την "πληρωμή" μεταξύ των χαρακτηριστικών

Ας υποθέσουμε το ακόλουθο σενάριο:

Έχετε εκπαιδεύσει ένα μοντέλο μηχανικής μάθησης για την πρόβλεψη των τιμών των διαμερισμάτων. Για ένα συγκεκριμένο διαμέρισμα προβλέπει 300.000 ευρώ και πρέπει να εξηγήσετε

αυτή την πρόβλεψη. Το διαμέρισμα έχει εμβαδόν 50 m2, βρίσκεται στον 2ο όροφο, έχει ένα πάρκο κοντά και οι γάτες απαγορεύονται:



Εικόνα 2 :Shapley values

Η μέση πρόβλεψη για όλα τα διαμερίσματα είναι 310.000 ευρώ. Πόσο έχει συνεισφέρει κάθε τιμή χαρακτηριστικού στην πρόβλεψη σε σύγκριση με τη μέση πρόβλεψη;

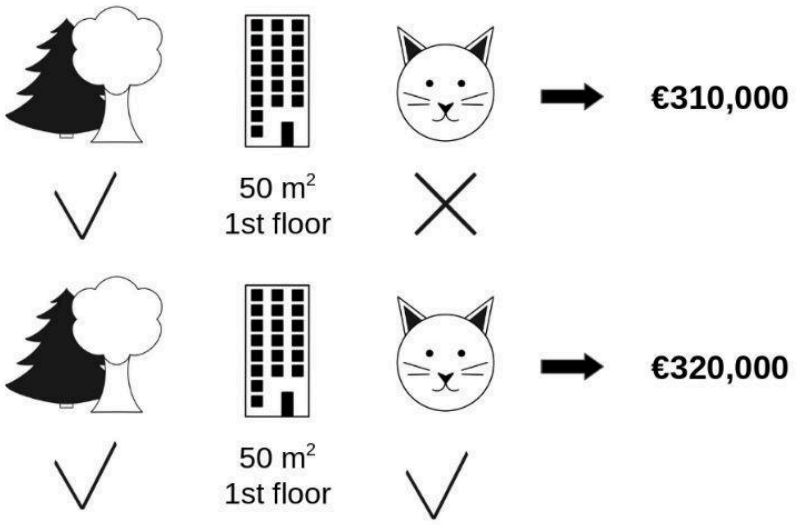
Η απάντηση είναι απλή για τα μοντέλα γραμμικής παλινδρόμησης. Η επίδραση κάθε χαρακτηριστικού είναι το βάρος του χαρακτηριστικού επί την τιμή του χαρακτηριστικού. Αυτό λειτουργεί μόνο λόγω της γραμμικότητας του μοντέλου. Για πιο σύνθετα μοντέλα, χρειαζόμαστε μια διαφορετική λύση. Για παράδειγμα, το LIME προτείνει τοπικά μοντέλα για την εκτίμηση των επιδράσεων. Μια άλλη λύση προέρχεται από τη θεωρία συνεργατικών παιγνίων: Η αξία Shapley, που επινοήθηκε από τον Shapley (1953)63 , είναι μια μέθοδος για την ανάθεση πληρωμών στους παίκτες ανάλογα με τη συμβολή τους στη συνολική πληρωμή. Οι παίκτες συνεργάζονται σε έναν συνασπισμό και λαμβάνουν ένα ορισμένο κέρδος από αυτή τη συνεργασία.

Οι παίκτες; Παιχνίδι; Πληρωμή; Ποια είναι η σύνδεση με τις προβλέψεις μηχανικής μάθησης και την ερμηνευσιμότητα; Το "παιχνίδι" είναι η εργασία πρόβλεψης για μια μοναδική περίπτωση του συνόλου δεδομένων. Το "κέρδος" είναι η πραγματική πρόβλεψη για αυτή την περίπτωση μείον τη μέση πρόβλεψη για όλες τις περιπτώσεις. Οι "παίκτες" είναι οι τιμές των χαρακτηριστικών της περίπτωσης που συνεργάζονται για να λάβουν το κέρδος (= προβλέπουν μια συγκεκριμένη τιμή). Στο παράδειγμά μας για το διαμέρισμα, οι τιμές χαρακτηριστικών park-nearby, cat-banned, area-50 και floor-2nd συνεργάστηκαν για να επιτευχθεί η πρόβλεψη των 300.000 €. Στόχος μας είναι να εξηγήσουμε τη διαφορά μεταξύ της πραγματικής πρόβλεψης (300.000 €) και της μέσης πρόβλεψης (310.000 €): μια διαφορά -10.000 €.
Η απάντηση θα μπορούσε να είναι: Η περιοχή 50 συνεισέφερε 10.000 ευρώ, ο 2ος όροφος συνεισέφερε 0 ευρώ, η απαγόρευση γάτας συνεισέφερε -50.000 ευρώ. Οι συνεισφορές αθροίζονται σε -€10.000, την τελική πρόβλεψη μείον τη μέση προβλεπόμενη τιμή διαμερίσματος.

Πώς υπολογίζουμε την τιμή Shapley για ένα χαρακτηριστικό;

Η τιμή Shapley είναι η μέση οριακή συνεισφορά της τιμής ενός χαρακτηριστικού σε όλους τους δυνατούς συνασπισμούς. Όλα κατανοητά τώρα;

Στο ακόλουθο σχήμα αξιολογούμε τη συνεισφορά της αξίας του χαρακτηριστικού cat-banned όταν προστίθεται σε έναν συνασπισμό των park-nearby και area-50. Προσομοιώνουμε ότι μόνο τα park-nearby, cat-banned και area-50 είναι σε συνασπισμό, αντλώντας τυχαία ένα άλλο διαμέρισμα από τα δεδομένα και χρησιμοποιώντας την τιμή του για το χαρακτηριστικό floor. Η τιμή floor-2nd αντικαταστάθηκε από την τυχαία αντληθείσα τιμή floor-1st. Στη συνέχεια προβλέπουμε την τιμή του διαμερίσματος με αυτόν τον συνδυασμό (310.000 ευρώ). Σε ένα δεύτερο βήμα, αφαιρούμε το cat-banned από τον συνασπισμό αντικαθιστώντας το με μια τυχαία τιμή του χαρακτηριστικού cat allowed/banned από το τυχαία κληρωθέν διαμέρισμα. Στο παράδειγμα ήταν cat-allowed, αλλά θα μπορούσε να είναι και πάλι cat-banned. Προβλέπουμε την τιμή του διαμερίσματος για τον συνασπισμό park-nearby και area-50 (320.000 €). Η συνεισφορά της απαγόρευσης γάτας ήταν 310.000 € - 320.000 € = -10.000 €. Αυτή η εκτίμηση εξαρτάται από τις τιμές του τυχαία κληρωθέντος διαμερίσματος που χρησίμευσε ως "δότης" για τη γάτα και τις τιμές των χαρακτηριστικών του ορόφου. Θα έχουμε καλύτερες εκτιμήσεις αν επαναλάβουμε αυτό το βήμα δειγματοληψίας και υπολογίσουμε τον μέσο όρο της συνεισφοράς



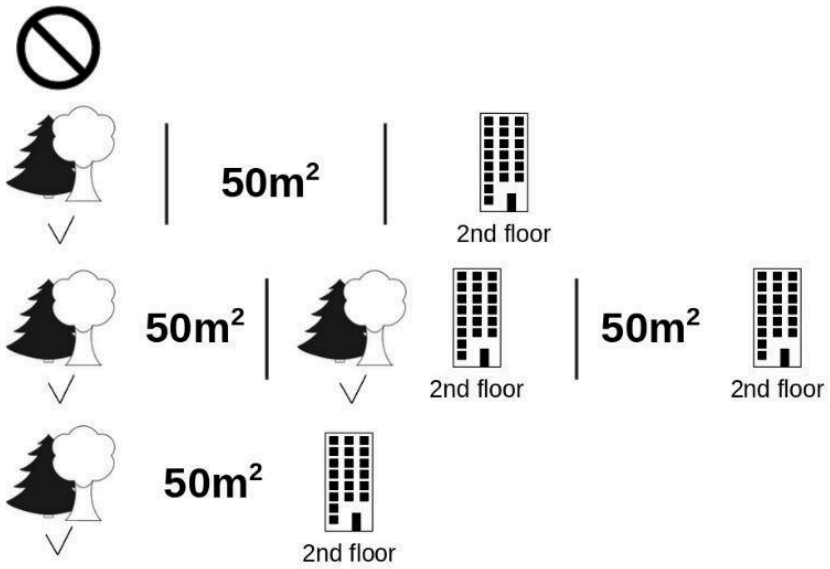
Επαναλαμβάνουμε αυτόν τον υπολογισμό για όλους τους πιθανούς συνασπισμούς. Η τιμή Shapley είναι ο μέσος όρος όλων των οριακών συνεισφορών σε όλους τους πιθανούς συνασπισμούς. Ο χρόνος υπολογισμού αυξάνεται εκθετικά με τον αριθμό των χαρακτηριστικών. Μια λύση για να διατηρηθεί ο χρόνος υπολογισμού διαχειρίσιμος είναι ο υπολογισμός των συνεισφορών για λίγα μόνο δείγματα των πιθανών συνασπισμών.

Το ακόλουθο σχήμα δείχνει όλους τους συνασπισμούς των τιμών των χαρακτηριστικών που απαιτούνται για τον προσδιορισμό της τιμής Shapley για το cat-banned. Η πρώτη σειρά δείχνει τον συνασπισμό χωρίς καμία τιμή χαρακτηριστικών. Η δεύτερη, η τρίτη και η τέταρτη σειρά

δείχνουν διαφορετικούς συνασπισμούς με αυξανόμενο μέγεθος συνασπισμού, που χωρίζονται με "|". Συνολικά, είναι δυνατοί οι ακόλουθοι συνασπισμοί:

- Χωρίς τιμές χαρακτηριστικών
- park-nearby
- area-50
- όροφος-2ος
- park-nearby+area-50
- park-nearby+floor-2nd
- area-50+floor-2nd
- park-nearby+area-50+floor-2nd.

Για κάθε έναν από αυτούς τους συνασπισμούς υπολογίζουμε την προβλεπόμενη τιμή διαμερίσματος με και χωρίς την τιμή του χαρακτηριστικού cat-banned και παίρνουμε τη διαφορά για να πάρουμε την οριακή συνεισφορά. Η τιμή Shapley είναι ο (σταθμισμένος) μέσος όρος των οριακών συνεισφορών. Αντικαθιστούμε τις τιμές χαρακτηριστικών των χαρακτηριστικών που δεν ανήκουν σε συνασπισμό με τυχαίες τιμές χαρακτηριστικών από το σύνολο δεδομένων διαμερισμάτων για να λάβουμε μια πρόβλεψη από το μοντέλο μηχανικής μάθησης.



Εάν εκτιμήσουμε τις τιμές Shapley για όλες τις τιμές των χαρακτηριστικών, έχουμε την πλήρη κατανομή της πρόβλεψης (μείον το μέσο όρο) μεταξύ των τιμών των χαρακτηριστικών.

Advantages

Η διαφορά μεταξύ της πρόβλεψης και της μέσης πρόβλεψης κατανέμεται δίκαια μεταξύ των τιμών χαρακτηριστικών της περίπτωσης - η ιδιότητα της αποδοτικότητας των τιμών Shapley. Αυτή η ιδιότητα διακρίνει την τιμή Shapley από άλλες μεθόδους όπως η LIME. Η LIME δεν εγγυάται ότι η πρόβλεψη είναι δίκαια κατανεμημένη μεταξύ των χαρακτηριστικών. Η τιμή Shapley μπορεί να είναι η μόνη μέθοδος που παρέχει μια πλήρη εξήγηση. Σε περιπτώσεις όπου ο νόμος απαιτεί επεξηγηματικότητα - όπως το "δικαίωμα στην παροχή εξηγήσεων" της ΕΕ - η τιμή Shapley μπορεί να είναι η μόνη νομικά συμβατή μέθοδος, επειδή βασίζεται σε μια σταθερή θεωρία και κατανέμει δίκαια τα αποτελέσματα. Δεν είμαι δικηγόρος, οπότε αυτό αντανακλά μόνο τη διαίσθησή μου σχετικά με τις απαιτήσεις.

Η αξία Shapley επιτρέπει αντιθετικές εξηγήσεις. Αντί να συγκρίνετε μια πρόβλεψη με τη μέση πρόβλεψη ολόκληρου του συνόλου δεδομένων, μπορείτε να τη συγκρίνετε με ένα υποσύνολο ή ακόμη και με ένα μεμονωμένο σημείο δεδομένων. Αυτή η αντιθετικότητα είναι επίσης κάτι που δεν διαθέτουν τα τοπικά μοντέλα όπως το LIME.

Η τιμή Shapley είναι η μόνη μέθοδος εξήγησης με σταθερή θεωρία. Τα αξιώματα - αποτελεσματικότητα, συμμετρία, ομοίωμα, προσθετικότητα - δίνουν στην εξήγηση μια λογική βάση. Μέθοδοι όπως το LIME υποθέτουν γραμμική συμπεριφορά του μοντέλου μηχανικής μάθησης τοπικά, αλλά δεν υπάρχει θεωρία για το γιατί αυτό πρέπει να λειτουργεί.

Είναι εντυπωσιακό να εξηγείται μια πρόβλεψη ως ένα παιχνίδι που παίζεται από τις τιμές των χαρακτηριστικών.

Disadvantages

Η τιμή Shapley απαιτεί πολύ χρόνο υπολογισμού. Στο 99,9% των προβλημάτων του πραγματικού κόσμου, μόνο η προσεγγιστική λύση είναι εφικτή. Ο ακριβής υπολογισμός της τιμής Shapley είναι υπολογιστικά δαπανηρός επειδή υπάρχουν 2k δυνατοί συνασπισμοί των τιμών των χαρακτηριστικών και η "απουσία" ενός χαρακτηριστικού πρέπει να προσομοιωθεί με την κλήρωση τυχαίων περιπτώσεων, γεγονός που αυξάνει τη διακύμανση για την εκτίμηση της εκτίμησης των τιμών Shapley. Ο εκθετικός αριθμός των συνασπισμών αντιμετωπίζεται με τη δειγματοληψία συνασπισμών και τον περιορισμό του αριθμού των επαναλήψεων M. Η μείωση του M μειώνει τον χρόνο υπολογισμού, αλλά αυξάνει τη διακύμανση της τιμής Shapley. Δεν υπάρχει καλός κανόνας για τον αριθμό των επαναλήψεων M. Το M πρέπει να είναι αρκετά μεγάλο ώστε να εκτιμώνται με ακρίβεια οι τιμές Shapley, αλλά αρκετά μικρό ώστε να ολοκληρώνεται ο υπολογισμός σε εύλογο χρόνο. Θα πρέπει να είναι δυνατή η επιλογή του M με βάση τα όρια Chernoff, αλλά δεν έχω δει κάποια εργασία για να γίνει αυτό για τις τιμές Shapley για προβλέψεις μηχανικής μάθησης.

Η τιμή Shapley μπορεί να παρερμηνευθεί. Η τιμή Shapley μιας τιμής χαρακτηριστικού δεν είναι η διαφορά της προβλεπόμενης τιμής μετά την αφαίρεση του χαρακτηριστικού από την εκπαίδευση του μοντέλου. Η ερμηνεία της τιμής Shapley είναι η εξής: Δεδομένου του τρέχοντος συνόλου τιμών χαρακτηριστικών, η συμβολή μιας τιμής χαρακτηριστικού στη διαφορά μεταξύ της πραγματικής πρόβλεψης και της μέσης πρόβλεψης είναι η εκτιμώμενη τιμή Shapley.
Η τιμή Shapley είναι η λάθος μέθοδος εξήγησης αν αναζητάτε αραιές εξηγήσεις (εξηγήσεις που περιέχουν λίγα χαρακτηριστικά). Οι εξηγήσεις που δημιουργούνται με τη μέθοδο της τιμής Shapley χρησιμοποιούν πάντα όλα τα χαρακτηριστικά. Οι άνθρωποι προτιμούν τις επιλεκτικές εξηγήσεις, όπως αυτές που παράγονται από τη LIME. Η LIME μπορεί να είναι η καλύτερη επιλογή για εξηγήσεις με τις οποίες έχουν να κάνουν οι απλοί άνθρωποι. Μια άλλη λύση είναι η SHAP που εισήγαγαν οι Lundberg και Lee (2016)[18] , η οποία βασίζεται στην αξία Shapley, αλλά μπορεί επίσης να παρέχει εξηγήσεις με λίγα χαρακτηριστικά.

Η τιμή Shapley επιστρέφει μια απλή τιμή ανά χαρακτηριστικό, αλλά όχι μοντέλο πρόβλεψης όπως το LIME. Αυτό σημαίνει ότι δεν μπορεί να χρησιμοποιηθεί για να γίνουν δηλώσεις σχετικά με τις αλλαγές στην πρόβλεψη για αλλαγές στην είσοδο, όπως π.χ: "Αν κέρδιζα 300 ευρώ περισσότερα το χρόνο, η βαθμολογία της πιστοληπτικής μου ικανότητας θα αυξανόταν κατά 5 μονάδες".

Ένα άλλο μειονέκτημα είναι ότι χρειάζεστε πρόσβαση στα δεδομένα αν θέλετε να υπολογίσετε την τιμή Shapley για μια νέα περίπτωση δεδομένων. Δεν αρκεί η πρόσβαση στη συνάρτηση πρόβλεψης, διότι χρειάζεστε τα δεδομένα για να αντικαταστήσετε τμήματα της περίπτωσης ενδιαφέροντος με τιμές από τυχαία αντλημένες περιπτώσεις των δεδομένων. Αυτό μπορεί να αποφευχθεί μόνο αν μπορείτε να δημιουργήσετε περιπτώσεις δεδομένων που μοιάζουν με πραγματικές περιπτώσεις δεδομένων αλλά δεν είναι πραγματικές περιπτώσεις από τα δεδομένα εκπαίδευσης.
Όπως και πολλές άλλες μέθοδοι ερμηνείας με βάση την αντιμετάθεση, η μέθοδος της τιμής Shapley πάσχει από τη συμπερίληψη μη ρεαλιστικών περιπτώσεων δεδομένων όταν τα χαρακτηριστικά συσχετίζονται. Για να προσομοιώσουμε ότι μια τιμή χαρακτηριστικού λείπει από έναν συνασπισμό, περιθωριοποιούμε το χαρακτηριστικό. Αυτό επιτυγχάνεται με τη δειγματοληψία τιμών από την οριακή κατανομή του χαρακτηριστικού. Αυτό είναι καλό εφόσον τα χαρακτηριστικά είναι ανεξάρτητα. Όταν τα χαρακτηριστικά είναι εξαρτημένα, τότε μπορεί να δειγματοληπτήσουμε τιμές χαρακτηριστικών που δεν έχουν νόημα για τη συγκεκριμένη περίπτωση. Θα τις χρησιμοποιούσαμε όμως για να υπολογίσουμε την τιμή Shapley του χαρακτηριστικού. Μια λύση θα μπορούσε να είναι να αντιμεταθέσουμε συσχετιζόμενα χαρακτηριστικά μαζί και να πάρουμε μια αμοιβαία τιμή Shapley για αυτά. Μια άλλη προσαρμογή είναι η δειγματοληψία υπό όρους: Η δειγματοληψία των χαρακτηριστικών εξαρτάται από τα χαρακτηριστικά που βρίσκονται ήδη στην ομάδα. Ενώ η υπό όρους δειγματοληψία διορθώνει το ζήτημα των μη ρεαλιστικών σημείων δεδομένων, εισάγεται ένα νέο ζήτημα: Οι τιμές που προκύπτουν δεν είναι πλέον οι τιμές Shapley για το παιχνίδι μας, καθώς παραβιάζουν το αξίωμα συμμετρίας, όπως διαπίστωσαν οι Sundararajan et al. (2019)[19] και συζήτησαν περαιτέρω οι Janzing et al. (2020)[20] .

3.3 SHAP (SHapley Additive exPlanations)

[16]Η SHAP (SHapley Additive exPlanations) των Lundberg και Lee (2017)69 είναι μια μέθοδος για την εξήγηση μεμονωμένων προβλέψεων. Η SHAP βασίζεται στις θεωρητικά βέλτιστες από πλευράς παιγνίων τιμές Shapley.

Υπάρχουν δύο λόγοι για τους οποίους η SHAP απέκτησε το δικό της κεφάλαιο και δεν αποτελεί υποκεφάλαιο των αξιών Shapley. Πρώτον, οι συγγραφείς του SHAP πρότειναν το KernelSHAP, μια εναλλακτική, βασισμένη στον πυρήνα προσέγγιση εκτίμησης για τις τιμές Shapley εμπνευσμένη από τοπικά υποκατάστατα μοντέλα. Και πρότειναν το TreeSHAP, μια αποδοτική προσέγγιση εκτίμησης για μοντέλα που βασίζονται σε δέντρα. Δεύτερον, το SHAP συνοδεύεται από πολλές συνολικές μεθόδους ερμηνείας που βασίζονται σε συσσωρεύσεις των τιμών Shapley. Το παρόν κεφάλαιο εξηγεί τόσο τις νέες προσεγγίσεις εκτίμησης όσο και τις μεθόδους συνολικής ερμηνείας.

3.3.1 Ορισμός SHAP

Ο στόχος του SHAP είναι να εξηγήσει την πρόβλεψη μιας περίπτωσης x υπολογίζοντας τη συμβολή κάθε χαρακτηριστικού στην πρόβλεψη. Η μέθοδος εξήγησης SHAP υπολογίζει τις τιμές Shapley από τη θεωρία συμμαχικών παιγνίων. Οι τιμές των χαρακτηριστικών μιας περίπτωσης δεδομένων λειτουργούν ως παίκτες σε έναν συνασπισμό. Οι τιμές Shapley μας λένε πώς να κατανείμουμε δίκαια την "πληρωμή" (= την πρόβλεψη) μεταξύ των χαρακτηριστικών. Ένας παίκτης μπορεί να είναι μια μεμονωμένη τιμή γνωρίσματος, π.χ. για δεδομένα σε πίνακες. Ένας παίκτης μπορεί επίσης να είναι μια ομάδα τιμών χαρακτηριστικών. Για παράδειγμα, για την εξήγηση μιας εικόνας, τα εικονοστοιχεία μπορούν να ομαδοποιηθούν σε superpixels και η πρόβλεψη να κατανεμηθεί μεταξύ τους. Μια καινοτομία που φέρνει στο τραπέζι το SHAP είναι ότι η εξήγηση των τιμών Shapley αναπαρίσταται ως μια προσθετική μέθοδος απόδοσης χαρακτηριστικών, ένα γραμμικό μοντέλο. Αυτή η άποψη συνδέει τις τιμές LIME και Shapley. Το SHAP προσδιορίζει την εξήγηση ως εξής:

$$g(z') = \varphi 0 + \sum_{j=1}^{M} \varphi j z j' \qquad (1)$$

Όπου g είναι το μοντέλο επεξήγησης, z e $\{0,1\}$ είναι το διάνυσμα του συνασπισμού, M είναι το μέγιστο μέγεθος του συνασπισμού και και

ϕ

j

∈

R

είναι η απόδοση χαρακτηριστικού για ένα χαρακτηριστικό j,τις τιμές Shapley. Αυτό που ονομάζω "διάνυσμα συνασπισμού" ονομάζεται "απλοποιημένα χαρακτηριστικά" στο έγγραφο SHAP. Νομίζω ότι επιλέχθηκε αυτή η ονομασία, επειδή, π.χ. για τα δεδομένα εικόνας, οι εικόνες δεν αναπαρίστανται σε επίπεδο εικονοστοιχείου, αλλά αθροίζονται σε υπερ-εικονοστοιχεία. Πιστεύω ότι είναι χρήσιμο να σκεφτούμε ότι τα z περιγράφουν συνασπισμούς: Στο διάνυσμα συνασπισμού, μια καταχώρηση 1 σημαίνει ότι η αντίστοιχη τιμή χαρακτηριστικού είναι "παρούσα" και 0 ότι είναι "απούσα". Για να υπολογίσουμε τις τιμές Shapley, προσομοιώνουμε ότι μόνο κάποιες τιμές χαρακτηριστικών είναι ("παρούσες") και κάποιες όχι ("απούσες"). Η αναπαράσταση ως γραμμικό μοντέλο συνασπισμών είναι ένα τέχνασμα για τον υπολογισμό των ϕ 's. Για το x, την περίπτωση ενδιαφέροντος, το διάνυσμα συνασπισμού x' είναι ένα διάνυσμα όλων των 1, δηλαδή όλες οι τιμές χαρακτηριστικών είναι "παρούσες". Ο τύπος απλοποιείται ως εξής:

$$g(x') = \varphi 0 + \sum_{j=1}^{M} \varphi j \qquad (2)$$

Οι τιμές Shapley είναι η μόνη λύση που ικανοποιεί τις ιδιότητες της Αποδοτικότητας, της Συμμετρίας, του Ομοιώματος και της Προσθετικότητας. Το SHAP τις ικανοποιεί επίσης, αφού υπολογίζει τις τιμές Shapley.

Το SHAP περιγράφει τις ακόλουθες τρεις επιθυμητές ιδιότητες:

1) Local accuracy

$$f(x) = g(x') = \varphi 0 + \sum_{j=1}^{M} \varphi j x j' \qquad (3)$$

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_X(\hat{f}(X)) + \sum_{j=1}^{M} \phi_j$$

2) Missingness

$$x'_j = 0 \Rightarrow \phi_j = 0$$

3) Consistency

Let $\hat{f}_x(z') = \hat{f}(h_x(z'))$ and z'_{-j} indicate that $z'_j = 0$. For any two models f and f' that satisfy:

$$\hat{f}'_x(z') - \hat{f}'_x(z'_{-j}) \geq \hat{f}_x(z') - \hat{f}_x(z'_{-j})$$

for all inputs $z' \in \{0,1\}^M$, then:

$$\phi_j(\hat{f}', x) \geq \phi_j(\hat{f}, x)$$

Η ιδιότητα της συνέπειας λέει ότι αν ένα μοντέλο αλλάξει έτσι ώστε η οριακή συνεισφορά της τιμής ενός χαρακτηριστικού να αυξηθεί ή να παραμείνει η ίδια (ανεξάρτητα από άλλα χαρακτηριστικά), η τιμή Shapley επίσης αυξάνεται ή παραμένει η ίδια.

3.3.2 TreeSHAP

Οι Lundberg κ.ά. (2018)[21]πρότειναν το TreeSHAP, μια παραλλαγή του SHAP για μοντέλα μηχανικής μάθησης που βασίζονται σε δέντρα, όπως δέντρα απόφασης, τυχαία δάση και δέντρα με ενισχυμένη κλίση. Το TreeSHAP εισήχθη ως μια γρήγορη, ειδική για κάθε μοντέλο εναλλακτική λύση στο KernelSHAP, αλλά αποδείχθηκε ότι μπορεί να παράγει μη διαισθητικές αποδόσεις χαρακτηριστικών.
Το TreeSHAP ορίζει τη συνάρτηση αξίας χρησιμοποιώντας την υπό συνθήκη προσδοκία $E_{x_j|x_j}($
για την εκτίμηση των αποτελεσμάτων. Θα σας δώσω κάποια διαίσθηση για το πώς μπορούμε να υπολογίσουμε την αναμενόμενη πρόβλεψη για ένα μόνο δέντρο, μια περίπτωση x και ένα υποσύνολο χαρακτηριστικών S. Αν εξαρτούσαμε την πρόβλεψη από όλα τα χαρακτηριστικά - αν το S ήταν το σύνολο όλων των χαρακτηριστικών - τότε η πρόβλεψη από τον κόμβο στον οποίο εμπίπτει η περίπτωση x θα ήταν η αναμενόμενη πρόβλεψη. Εάν δεν εξαρτούσαμε την πρόβλεψη από κανένα χαρακτηριστικό - εάν το S ήταν κενό - θα χρησιμοποιούσαμε τον σταθμισμένο μέσο όρο των προβλέψεων όλων των τερματικών κόμβων. Εάν το S περιέχει ορισμένα, αλλά όχι όλα, χαρακτηριστικά, αγνοούμε τις προβλέψεις των μη προσβάσιμων κόμβων. Μη προσβάσιμος σημαίνει ότι το μονοπάτι απόφασης που οδηγεί σε αυτόν τον κόμβο έρχεται σε αντίθεση με τις τιμές στο x_s.

Από τους υπόλοιπους τερματικούς κόμβους, λαμβάνουμε το μέσο όρο των προβλέψεων σταθμισμένων με βάση το μέγεθος του κόμβου (δηλαδή τον αριθμό των δειγμάτων εκπαίδευσης στον συγκεκριμένο κόμβο). Ο μέσος όρος των εναπομεινάντων τερματικών κόμβων, σταθμισμένος με τον αριθμό των παραδειγμάτων ανά κόμβο, είναι η αναμενόμενη πρόβλεψη για το x δεδομένης της S. Το πρόβλημα είναι ότι πρέπει να εφαρμόσουμε αυτή τη διαδικασία για κάθε πιθανό υποσύνολο S των τιμών των χαρακτηριστικών. Το TreeSHAP υπολογίζει σε πολυωνυμικό χρόνο αντί για εκθετικό. Η βασική ιδέα είναι να σπρώχνουμε όλα τα πιθανά υποσύνολα S προς τα κάτω στο δέντρο ταυτόχρονα. Για κάθε κόμβο απόφασης πρέπει να παρακολουθούμε τον αριθμό

των υποσυνόλων. Αυτό εξαρτάται από τα υποσύνολα στο γονικό κόμβο και το χαρακτηριστικό διάσπασης. Για παράδειγμα, όταν η πρώτη διάσπαση σε ένα δέντρο είναι στο χαρακτηριστικό x3, τότε όλα τα υποσύνολα που περιέχουν το χαρακτηριστικό x3 θα πάνε σε έναν κόμβο (αυτόν που πηγαίνει το x). Τα υποσύνολα που δεν περιέχουν το χαρακτηριστικό x3 πηγαίνουν και στους δύο κόμβους με μειωμένο βάρος. Δυστυχώς, υποσύνολα διαφορετικού μεγέθους έχουν διαφορετικά βάρη. Ο αλγόριθμος πρέπει να παρακολουθεί το συνολικό βάρος των υποσυνόλων σε κάθε κόμβο. Αυτό περιπλέκει τον αλγόριθμο. Για λεπτομέρειες σχετικά με το TreeSHAP παραπέμπω στην αρχική εργασία. Ο υπολογισμός μπορεί να επεκταθεί σε περισσότερα δέντρα: Χάρη στην ιδιότητα προσθετικότητας των τιμών Shapley, οι τιμές Shapley ενός συνόλου δέντρων είναι ο (σταθμισμένος) μέσος όρος των τιμών Shapley των μεμονωμένων δέντρων.

3.3.3 SHAP Plots

Η ιδέα πίσω από τη σημασία του χαρακτηριστικού SHAP είναι απλή:

Χαρακτηριστικά με μεγάλες απόλυτες τιμές Shapley είναι σημαντικά.

Εφόσον θέλουμε τη συνολική σημασία, υπολογίζουμε το μέσο όρο των απόλυτων τιμών Shapley ανά χαρακτηριστικό σε όλα τα δεδομένα:

$$Ij = \frac{1}{n}\sum_{i=1}^{n} |\varphi j^{(i)}| \qquad (4)$$

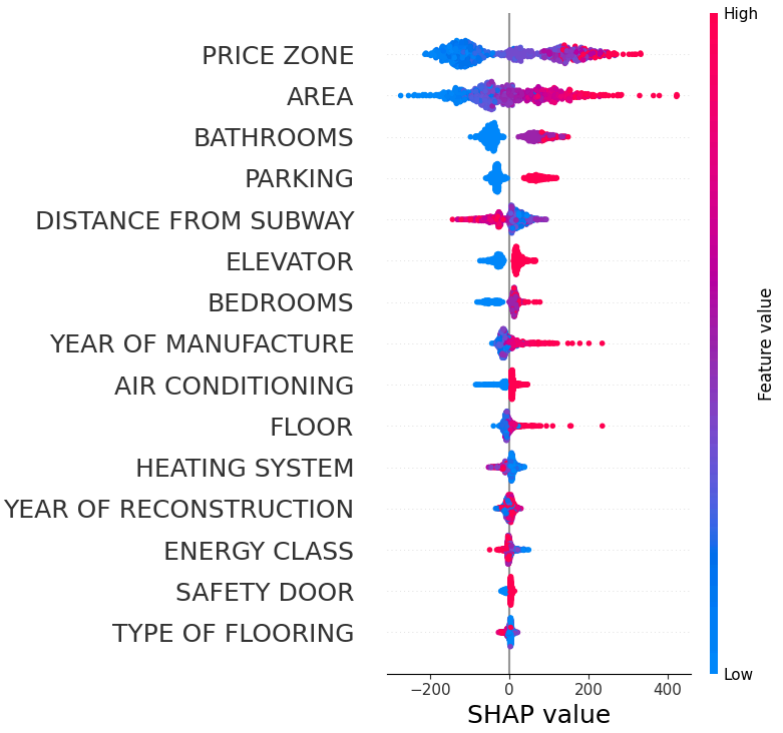
Στη συνέχεια, ταξινομούμε τα χαρακτηριστικά κατά φθίνουσα σημασία και τα σχεδιάζουμε.Η σημασία χαρακτηριστικών SHAP είναι μια εναλλακτική λύση στη σημασία χαρακτηριστικών μετατροπής. Υπάρχει μεγάλη διαφορά μεταξύ των δύο μέτρων σημαντικότητας: Η σημασία των χαρακτηριστικών μεταβολής βασίζεται στη μείωση της απόδοσης του μοντέλου. Το SHAP βασίζεται στο μέγεθος των χαρακτηριστικών που αποδίδονται.

Το διάγραμμα σπουδαιότητας χαρακτηριστικών είναι χρήσιμο, αλλά δεν περιέχει καμία πληροφορία πέραν των εισαγωγών. Για μια πιο κατατοπιστική γραφική παράσταση, θα εξετάσουμε στη συνέχεια τη συνοπτική γραφική παράσταση.

Το συνοπτικό διάγραμμα συνδυάζει τη σημασία των χαρακτηριστικών με τις επιδράσεις των χαρακτηριστικών. Κάθε σημείο στο συνοπτικό διάγραμμα είναι μια τιμή Shapley για ένα χαρακτηριστικό και μια περίπτωση. Η θέση στον άξονα y καθορίζεται από το χαρακτηριστικό και στον άξονα x από την τιμή Shapley. Το χρώμα αντιπροσωπεύει την τιμή του χαρακτηριστικού από χαμηλή προς υψηλή. Τα επικαλυπτόμενα σημεία είναι jittered στην κατεύθυνση του άξονα y, έτσι ώστε να έχουμε μια αίσθηση της κατανομής των τιμών Shapley ανά χαρακτηριστικό. Τα χαρακτηριστικά είναι ταξινομημένα σύμφωνα με τη σημασία τους.Στο συνοπτικό διάγραμμα, βλέπουμε τις πρώτες ενδείξεις της σχέσης μεταξύ της τιμής ενός χαρακτηριστικού και της επίδρασης στην πρόβλεψη. Για να δούμε όμως την ακριβή μορφή της σχέσης, πρέπει να εξετάσουμε τα διαγράμματα εξάρτησης SHAP.

Summary Plot

Το συνοπτικό διάγραμμα συνδυάζει τη σημασία των χαρακτηριστικών με τις επιδράσεις των χαρακτηριστικών. Κάθε σημείο στο συνοπτικό διάγραμμα είναι μια τιμή Shapley για ένα χαρακτηριστικό και μια περίπτωση. Η θέση στον άξονα y καθορίζεται από το χαρακτηριστικό και στον άξονα x από την τιμή Shapley. Το χρώμα αντιπροσωπεύει την τιμή του χαρακτηριστικού από χαμηλά προς υψηλά. Τα επικαλυπτόμενα σημεία είναι jittered στην κατεύθυνση του άξονα y, έτσι ώστε να έχουμε μια αίσθηση της κατανομής των τιμών Shapley ανά χαρακτηριστικό. Τα χαρακτηριστικά είναι ταξινομημένα σύμφωνα με τη σημασία τους.



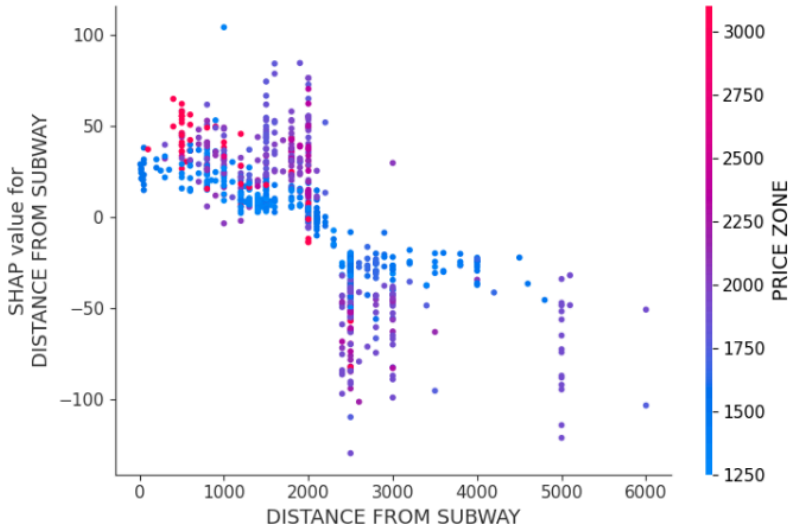
Πίνακας 1:Summary plot

Dependence Plot

Η εξάρτηση των χαρακτηριστικών SHAP μπορεί να είναι το απλούστερο διάγραμμα συνολικής ερμηνείας: 1) Διαλέξτε ένα χαρακτηριστικό. 2) Για κάθε περίπτωση δεδομένων, σχεδιάστε ένα σημείο με την τιμή του χαρακτηριστικού στον άξονα x και την αντίστοιχη τιμή Shapley στον άξονα y. 3) Έγινε.
Μαθηματικά, η γραφική παράσταση περιλαμβάνει τα ακόλουθα σημεία :

$$\left\{\left(xj^{(i)}, \varphi j^{(i)}\right)\right\}^{n} i = 1 \qquad (5)$$

Το ακόλουθο σχήμα δείχνει την εξάρτηση του χαρακτηριστικού SHAP για απόσταση από το μετρό:



Πίνακας 2: Dependence plot

Τα διαγράμματα εξάρτησης SHAP είναι μια εναλλακτική λύση για τα διαγράμματα μερικής εξάρτησης και τα συσσωρευμένα τοπικά αποτελέσματα. Ενώ τα διαγράμματα PDP και ALE δείχνουν τις μέσες επιδράσεις, η εξάρτηση SHAP δείχνει επίσης τη διακύμανση στον άξονα y. Ειδικά στην περίπτωση αλληλεπιδράσεων, το διάγραμμα εξάρτησης SHAP θα είναι πολύ πιο διασκορπισμένο στον άξονα y. Το διάγραμμα εξάρτησης μπορεί να βελτιωθεί με την επισήμανση αυτών των αλληλεπιδράσεων χαρακτηριστικών.

3.3.4 SHAP Disadvantages

Το **TreeSHAP** μπορεί να παράγει μη διαισθητικές αποδόσεις χαρακτηριστικών. Ενώ το TreeSHAP επιλύει το πρόβλημα της προεκβολής σε απίθανα σημεία δεδομένων, το κάνει αλλάζοντας τη συνάρτηση αξίας και επομένως αλλάζει ελαφρώς το παιχνίδι. Το TreeSHAP αλλάζει τη συνάρτηση αξίας βασιζόμενο στην υπό όρους αναμενόμενη πρόβλεψη. Με την αλλαγή στη συνάρτηση αξίας, τα χαρακτηριστικά που δεν έχουν καμία επιρροή στην πρόβλεψη μπορούν να λάβουν μια τιμή TreeSHAP διαφορετική από το μηδέν.

Τα μειονεκτήματα των τιμών Shapley ισχύουν και για το SHAP: Οι τιμές Shapley μπορούν να **παρερμηνευθούν** και απαιτείται πρόσβαση σε δεδομένα για τον υπολογισμό τους για νέα δεδομένα (εκτός από την TreeSHAP).

Είναι δυνατόν να δημιουργηθούν **σκόπιμα παραπλανητικές ερμηνείες με το SHAP**, οι οποίες μπορούν να κρύψουν προκαταλήψεις. Εάν είστε ο επιστήμονας δεδομένων που δημιουργεί τις ερμηνείες, αυτό δεν αποτελεί πραγματικό πρόβλημα (θα ήταν μάλιστα πλεονέκτημα εάν είστε ο κακός επιστήμονας δεδομένων που θέλει να δημιουργήσει παραπλανητικές ερμηνείες). Για τους αποδέκτες μιας εξήγησης SHAP, είναι μειονέκτημα: δεν μπορούν να είναι σίγουροι για την ειλικρίνεια της εξήγησης.

3.4 Μοντέλα Πρόβλεψης (Predictive models)

Στο προκείμενο κεφάλαιο θα αναλυθεί η λειτουργικότητα των μοντέλων μηχανικής μάθησης που θα αξιοποιηθούν για την μοντελοποίηση του συγκεκριμένου προβλήματος που μας αφορά. Τα μοντέλα παλινδρόμησης χρησιμοποιούνται στα προβλήματα επιτηρούμενης μάθησης όπου η μεταβλητή εξόδου λαμβάνει συνεχόμενες τιμές. Στόχος των μοντέλων είναι η εύρεση μιας συνάρτησης f, η οποία να αντιστοιχίζει τα δεδομένα εισόδου x σε μία συνεχόμενη μεταβλητή εξόδου y.

3.4.1 Γραμμική Παλινδρόμηση

Ένα μοντέλο γραμμικής παλινδρόμησης (linear regression) είναι ένας τρόπος να εκφράσουμε τα δύο βασικά συστατικά μιας στατιστικής σχέσης.

- Την τάση της μεταβλητής εξόδου Y (εξαρτημένη μεταβλητή) να σχετίζεται με την ανεξάρτητη μεταβλητή X συστηματικά.
- Τον τρόπο με τον οποίο οι παρατηρήσεις είναι διασκορπισμένες γύρω από την καμπύλη της στατιστικής σχέσης.

Με την χρήση της μεθόδου των ελαχίστων τετραγώνων υπολογίζουμε τους εκτιμητές (estimators), σύμφωνα με τους οποίους θα προκύψει η καμπύλη παλινδρόμησης .

Το μοντέλο της χρησιμοποιείται με σκοπό να επιτευχθούν ποσοτικές εκτιμήσεις οικονομικών σχέσεων, οι οποίες είχαν διατυπωθεί έως τότε μόνο θεωρητικά. Η παλινδρόμηση είναι μια στατιστική τεχνική η οποία προσπαθεί να εξηγήσει τον τρόπο με τον οποίο μεταβάλλεται

Μία απλή εξίσωση, που ορίζει την απλή γραμμική παλινδρόμηση και συνδέει δύο μεταβλητές είναι: (2.1). Το ονομάζεται εξαρτημένη μεταβλητή, το ανεξάρτητη, ο όρος αποτελεί το σφάλμα ή τη διαταραχή. Αν αυτό είναι σταθερό, δηλαδή , τότε η μεταβολή της μεταβλητής εξαρτάται γραμμικά από εκείνη της μεταβλητής : Αυτό σημαίνει ότι η παράμετρος είναι η παράμετρος της κλίσης της σχέσης μεταξύ και και έχει τραβήξει το ενδιαφέρον στην εφαρμοσμένη οικονομετρία. Η παράμετρος , είναι ο σταθερός όρος. Η γραμμικότητα της σχέσης υποδεικνύει ότι αλλαγή μίας μονάδας του θα έχει τα ίδια αποτελέσματα στην μεταβλητή Αυτή η σχέση αποτελεί μία βάση, αλλά προφανώς δεν είναι ρεαλιστική η εφαρμογή της σε πιο σύνθετες εφαρμογές. Η εξίσωση 2.1 μπορεί να θεωρηθεί ότι αποτελείται από δύο συνιστώσες, την ντετερμινιστική συνιστώσα και την στοχαστική . Υποθέτουμε ότι το σφάλμα έχει μηδενική μέση τιμή, δηλαδή , προκειμένου να μπορούμε να εξάγουμε αξιόπιστα συμπεράσματα για τη σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής. Μία επίσης βασική υπόθεση είναι ότι η μέση τιμή του δεν εξαρτάται από την μεταβλητή : ή (2.2). Έτσι με βάση την εξίσωση 2.2, το ντετερμινιστικό κομμάτι της εξίσωσης 2.1 μπορεί να γραφτεί: (2.3), που δηλώνει ότι η μέση τιμή του δεδομένου του , είναι μία γραμμική συνάρτηση της ανεξάρτητης μεταβλητής. Βέβαια η τιμή του η οποία παρατηρείται στον πραγματικό κόσμο, είναι μάλλον απίθανο να ισούται με την 2.3 και κατά συνέπεια η εισαγωγή ενός όρου σφάλματος κρίνεται απαραίτητη [3]. Άρα: . (2.4)

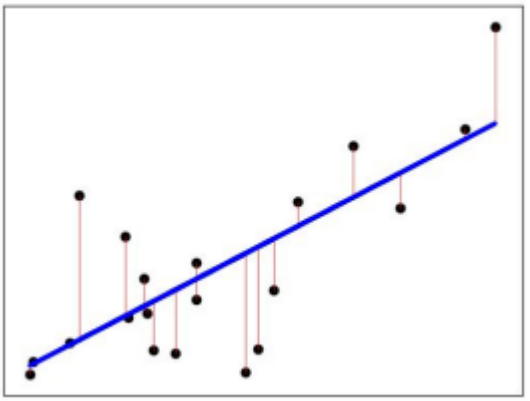
Ένα μοντέλο γραμμικής παλινδρόμησης (linear regression) είναι ένας τρόπος να εκφράσουμε τα δύο βασικά συστατικά μιας στατιστικής σχέσης.

1. Την τάση της μεταβλητής εξόδου Y (εξαρτημένη μεταβλητή) να σχετίζεται με την ανεξάρτητη μεταβλητή X συστηματικά.
2. Τον τρόπο με τον οποίο οι παρατηρήσεις είναι διασκορπισμένες γύρω από την καμπύλη της 30 στατιστικής σχέσης.

Με τη μέθοδο των ελαχίστων τετραγώνων προσπαθούμε να υπολογίσουμε τους εκτιμητές (estimators), βάσει των οποίων προκύπτει η καμπύλη της παλινδρόμησης. Στην Εικόνα 1 έχουμε με μαύρο χρώμα τις παρατηρήσεις, με μπλε χρώμα την ευθεία της παλινδρόμησης και με κόκκινο χρώμα την απόσταση κάθε σημείου από την ευθεία αυτή. Οι κόκκινες ευθείες απεικονίζουν την απόκλιση της πραγματικής τιμής της εξαρτημένης μεταβλητής Y , από την τιμή της που true Y υπολογίζεται έπειτα από τη μέθοδο της παλινδρόμησης. Στόχος μας είναι η καμπύλη παλινδρόμησης που προκύπτει να ελαχιστοποιεί το άθροισμα όλων των αποκλίσεων. Στην παρακάτω σχέση οι εκτιμητές β_0 και β_1 και υπολογίζονται έτσι ώστε να ελαχιστοποιείται η τιμή του Q.

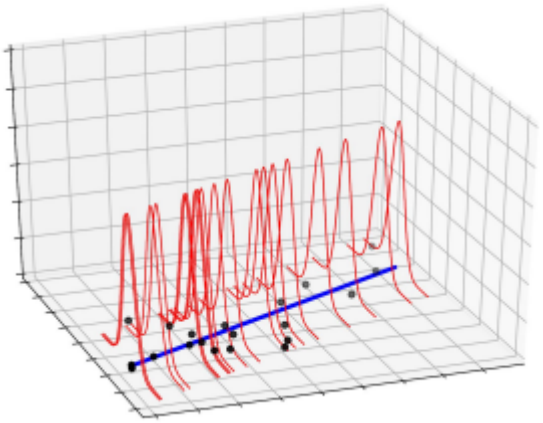
$$Q = \sum_{i=1}^{n} (Yi - \beta o - \beta i Xi)^2 \qquad (6)$$

Έτσι σχηματίζεται μία ευθεία γραμμή η οποία αντιπροσωπεύει το σύνολο των εξόδων των παρατηρήσεων.



Εικόνα 3:Ευθεία γραμμή που προκύπτει έπειτα από εφαρμογή της μεθόδου ελαχίστων τετραγώνων σε ένα σύνολο δεδομένων.

Αυτό που μας δείχνει η καμπύλη παλινδρόμησης, είναι ότι σε κάθε σημείο της καμπύλης υπάρχει μία συνάρτηση πυκνότητας πιθανότητας με μέση τιμή την τιμή του σημείου και διακύμανση που καθορίζεται από την τιμή του σφάλματος,



Εικόνα 4:Σχέση καμπύλης παλινδρόμησης με κατανομή πυκνότητας πιθανότητας

Η παραπάνω προσέγγιση είναι καθαρά στατιστική .Παρόλα' αυτά η χρήση της έχει μεγάλη σημασία στην δημιουργία μοντέλων πρόβλεψης. Έστω ότι έχουμε ένα σύνολο δεδομένων που αποτελείται από τις παρατηρήσεις μας και την τιμή της μεταβλητής εξόδου. Χωρίζουμε τις παρατηρήσεις μας σε δύο ομάδες (train set, test set) όπως αναφέραμε προηγουμένως. Με τη μέθοδο των ελαχίστων τετραγώνων φτιάχνουμε την καμπύλη παλινδρόμησης βάσει των δεδομένων που περιέχονται στο train set (προσαρμόζουμε την κλιση της ευθείας έτσι ώστε να αντιπροσωπεύει όσο το δυνατό καλύτερα το σύνολο των δεδομένων) και βάσει αυτής προβλέπουμε την έξοδο των παρατηρήσεων του test set. Έτσι εξετάζουμε την απόδοση του μοντέλου μας και το κατά πόσο μπορεί να προβλέψει την έξοδο μελλοντικών παρατηρήσεων. Στην περίπτωση που η απόδοση του μοντέλου μας είναι αρκετά υψηλή, πλέον θα έχουμε τη δυνατότητα να αντιστοιχίσουμε μελλοντικές εισόδους X , στην τιμή που i Y i τους αναλογεί, στηριζόμενοι στην καμπύλη παλινδρόμησης που έχουμε δημιουργήσει.

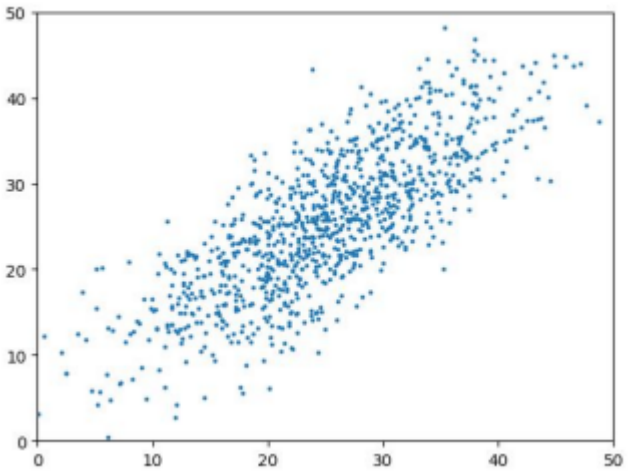
Από τα παραπάνω παρατηρούμε ότι η σημασία των μοντέλων γραμμικής παλινδρόμησης είναι αρκετά μεγάλη, τόσο στην επιστήμη της στατιστικής, όσο και στην επιστήμη των υπολογιστών. Παρόλα' αυτά, κάποια βασικά μειονεκτήματά τους μας οδηγούν στη χρήση πιο πολύπλοκων μοντέλων.

Η γραμμική παλινδρόμηση περιορίζεται σε γραμμικά προβλήματα. Από τη φύση της η γραμμική παλινδρόμηση κοιτάζει τις γραμμικές σχέσεις μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών. Συνεπώς υποθέτει ότι η σχέση τους μπορεί να απεικονιστεί με τη μορφή μιας ευθείας γραμμής. Αυτό φυσικά δεν είναι πάντα σωστό, καθώς στα περισσότερα προβλήματα οι ευθείες που συσχετίζουν τις μεταβλητές δεν είναι γνησίως μονότονες.

Η γραμμική παλινδρόμηση εστιάζει μόνο στη μέση τιμή της εξαρτημένης μεταβλητής. Όπως η μέση τιμή δεν αντιπροσωπεύει πάντα πλήρως μία μεταβλητή, έτσι και η γραμμική παλινδρόμηση δεν αντιπροσωπεύει πλήρως τις σχέσεις μεταξύ των μεταβλητών. Η γραμμική παλινδρόμηση επηρεάζεται από ακραίες τιμές παρατηρήσεων. Στα περισσότερα δεδομένα παρατηρούνται πολλές φορές ακραίες τιμές εξόδου σε ορισμένες παρατηρήσεις. Στην περίπτωση της γραμμικής παλινδρόμησης οι τιμές αυτές επηρεάζουν σε μεγάλο 32 βαθμό την κλίση της ευθείας. Οι παρατηρήσεις αυτές ονομάζονται παρατηρήσεις με ισχυρή επιρροή (influential observations) και η διαγραφή τους μπορεί να αυξήσει την απόδοση του μοντέλου μας. Αυτό συμβαίνει διότι δεν επιθυμούμε η καμπύλη μας να επηρεάζεται από τις ακραίες τιμές, αλλά από τιμές οι οποίες είναι αντιπροσωπευτικές του συνόλου που μελετάμε. Τα δεδομένα πρέπει να είναι ανεξάρτητα μεταξύ τους.

Η γραμμική παλινδρόμηση υποθέτει ότι τα δεδομένα είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει ότι η τιμή μιας ανεξάρτητης μεταβλητής δε θα έπρεπε να επηρεάζει την τιμή μιας άλλης. Αυτό φυσικά δεν είναι πάντα σωστό. Πολλές φορές οι ανεξάρτητες μεταβλητές σε ένα πρόβλημα συσχετίζονται, γεγονός που επιφέρει αρνητικές συνέπειες στα αποτελέσματά μας. Με την ανάλυση συσχέτισης (correlation analysis) μπορούμε να μετρήσουμε και να ερμηνεύσουμε το κατά πόσο η γραμμική ή μη γραμμική σχέση μεταξύ δύο συνεχόμενων μεταβλητών είναι ισχυρή.

Στην Εικόνα 5 παρατηρούμε το διάγραμμα συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών. Στο διάγραμμα αυτό φαίνεται πως η επηρεάζεται η τιμή της μιας μεταβλητής συναρτήσει της άλλης. Παρατηρούμε ότι τα στίγματα σχηματίζονται γύρω από την ευθεία x = y, γεγονός που μας οδηγεί στο συμπέρασμα ότι οι δύο αυτές μεταβλητές δεν είναι τελείως ανεξάρτητες μεταξύ τους. Συνεπώς η ταυτόχρονη χρήση τους στην εκτίμηση της ευθείας παλινδρόμησης θα οδηγούσε σε λανθασμένα αποτελέσματα.[17]



Εικόνα 5: Διάγραμμα συσχέτισης μεταξύ δύο ανεξάρτητων μεταβλητών

3.4.2 Μηχανές Διανυσμάτων Υποστήριξης(Support Vector Machine-SVM)

Όπως είδαμε παραπάνω, στόχος της γραμμικής παλινδρόμησης είναι η ελαχιστοποίηση του σφάλματος ανάμεσα στην προβλεπόμενη και την αναμενόμενη τιμή της εξόδου. Στην παλινδρόμηση διανυσμάτων υποστήριξης (support vector regression - SVR) θεωρούμε ένα κατώφλι σφάλματος ε , εντός του οποίου οι τιμές σφάλματος είναι αποδεκτές, και εκτιμούμε με τη χρήση του δείκτη C > 0 κατά πόσο "ανεχόμαστε" αποκλίσεις μεγαλύτερες του ε .

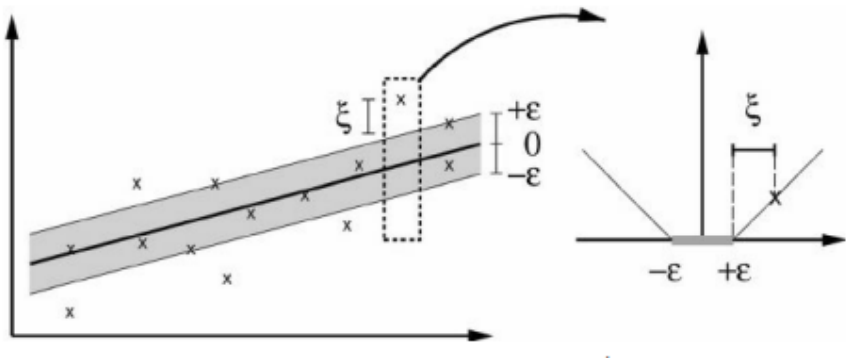
Στόχος είναι η ελαχιστοποίηση της συνάρτησης:

$$\frac{1}{2}||w||^2 + c\sum_{i=1}^{l}(\xi i + \xi i^*) \qquad (7)$$

η οποία υπόκειται στη σχέση:

$$\begin{cases} y_i - \langle w, x \rangle - b \le \varepsilon + \xi_i \\ \langle w, x \rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \qquad\qquad \ge 0 \end{cases} \tag{8}$$

Στην παραπάνω σχέση ορίζονται οι δύο ευθείες της εικόνας 6 οι οποίες ορίζουν το γκρι πλαίσιο, εντός του οποίου η τιμή του σφάλματος είναι μικρότερη ή ίση της τιμής ε .



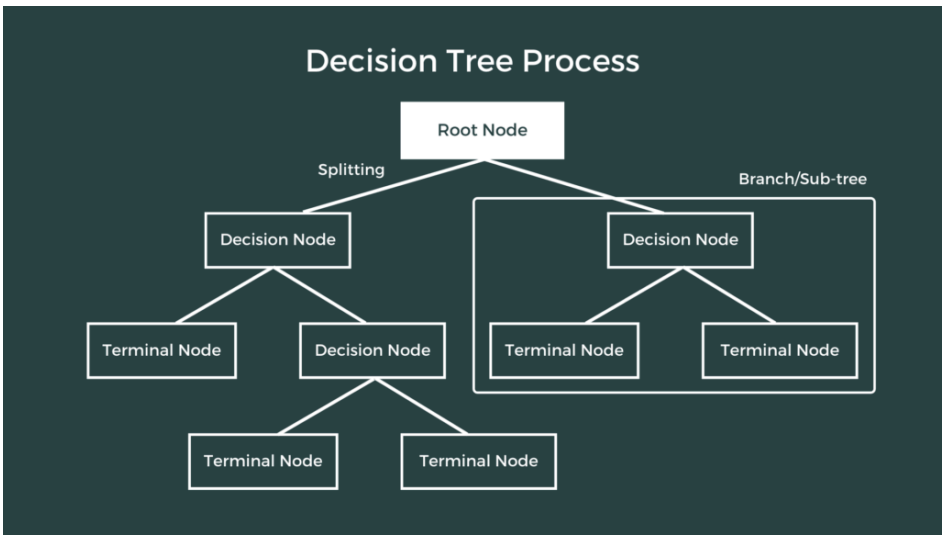
Εικόνα 6

Η συνάρτηση ξ ορίζεται ως εξής:

$$|\xi|_\varepsilon := \begin{cases} 0 & , |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & \text{ειδάλλως} \end{cases}$$

3.4.3 Δέντρα Παλινδρόμησης(Decision tree)

Ένα δέντρο παλινδρόμησης είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που μπορεί να προβλέψει αριθμητικές τιμές σε μια ποικιλία εφαρμογών. Ανήκει στην οικογένεια των δέντρων αποφάσεων και έχει σχεδιαστεί ειδικά για εργασίες παλινδρόμησης, όπου ο στόχος είναι να εκτιμηθεί μια μεταβλητή συνεχούς εξόδου με βάση ένα σύνολο χαρακτηριστικών εισόδου.



Εικόνα 7: Διαδικασία δέντρου αποφάσεων

Στον πυρήνα αυτού, ένα δέντρο παλινδρόμησης ακολουθεί μια διαδοχική προσέγγιση βασισμένη σε κανόνες, η οποία έχει οργανωθεί σε δομή δέντρου, το οποίο και αποτελείται από τρεις κύριους τύπους κόμβων: τον κόμβο ρίζας, τους εσωτερικούς κόμβους και τους κόμβους φύλλων. Ο κόμβος ρίζας αντιπροσωπεύει την αρχική κατάσταση του αλγορίθμου, που περιλαμβάνει ολόκληρο το σύνολο δεδομένων. Οι εσωτερικοί κόμβοι αντιστοιχούν σε συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων και περιέχουν κανόνες απόφασης, ενώ τέλος οι κόμβοι φύλλων αντιπροσωπεύουν τις τελικές προβλέψεις.

Το στάδιο εκπαίδευσης ενός δέντρου παλινδρόμησης περιλαμβάνει τη διαίρεση του πολυδιάστατου χώρου εισόδου σε υποσύνολα με ιεραρχικό τρόπο. Αυτή η διαδικασία καθοδηγείται από μια σειρά ερωτήσεων που τίθενται στα δεδομένα, με κάθε ερώτηση να στοχεύει στη λήψη μιας δυαδικής απόφασης με βάση ένα συγκεκριμένο χαρακτηριστικό. Οι απαντήσεις σε αυτές τις ερωτήσεις, που συνήθως αντιπροσωπεύονται ως "Ναι" ή "Όχι", καθοδηγούν τα δεδομένα στη δομή του δέντρου μέχρι να φτάσουν σε έναν κόμβο φύλλου. Κάθε κόμβος φύλλου περιέχει μια προκαθορισμένη αριθμητική τιμή, η οποία επιλέγεται κατά τη διαδικασία εκπαίδευσης.

Για την κατασκευή ενός δέντρου παλινδρόμησης, ο αλγόριθμος επαναλαμβάνει τα χαρακτηριστικά του συνόλου δεδομένων. Για κάθε χαρακτηριστικό, αξιολογεί διαφορετικά σημεία διαχωρισμού για να καθορίσει το καταλληλότερο σημείο τομής. Η επιλογή του σημείου διαίρεσης βασίζεται σε μια μέτρηση διαχωρισμού, όπως η ελαχιστοποίηση του σφάλματος ή της απόστασης μεταξύ των προβλεπόμενων και των πραγματικών τιμών σε κάθε υποσύνολο. Επιλέγεται το χαρακτηριστικό που επιτυγχάνει τον καλύτερο διαχωρισμό και τα δεδομένα χωρίζονται ανάλογα σε δύο διακριτές ομάδες.

Με τη διαίρεση των δεδομένων σε υποσύνολα, ο αλγόριθμος προχωρά στον υπολογισμό της μέσης τιμής της μεταβλητής στόχου σε κάθε υποσύνολο. Αυτή η μέση τιμή χρησιμεύει ως η

προβλεπόμενη αριθμητική τιμή για τυχόν μελλοντικά δεδομένα που εμπίπτουν στο ίδιο υποσύνολο. Ο στόχος είναι να βρεθεί το χαρακτηριστικό που μεγιστοποιεί την ανομοιότητα μεταξύ των υποσυνόλων ως προς τις μέσες τιμές τους, δημιουργώντας έτσι διακριτές περιοχές για πρόβλεψη. Με την επανάληψη αυτής της διαδικασίας επαναληπτικά, το δέντρο παλινδρόμησης προσθέτει περισσότερους κανόνες και αυξάνει το βάθος του, διαιρώντας περαιτέρω το σύνολο δεδομένων σε μικρότερες υποπεριοχές. Αυτή η επαναληπτική κατάτμηση και ανάθεση πρόβλεψης επιτρέπει στο δέντρο να καταγράφει σύνθετες σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου. Το βάθος του δέντρου καθορίζει τον αριθμό των ερωτήσεων που απαιτούνται για την επίτευξη μιας πρόβλεψης και μπορεί να ελεγχθεί για να αποφευχθεί η υπερβολική προσαρμογή θέτοντας ένα κριτήριο διακοπής.

Κατά τη φάση της πρόβλεψης, ένα νέο σημείο δεδομένων τροφοδοτείται στο εκπαιδευμένο δέντρο παλινδρόμησης. Ξεκινά από τον ριζικό κόμβο και ακολουθεί τους κανόνες απόφασης σε κάθε εσωτερικό κόμβο, με βάση τις τιμές των χαρακτηριστικών. Η διαδρομή μέσα από το δέντρο οδηγεί τα δεδομένα σε έναν συγκεκριμένο κόμβο φύλλου, όπου η προκαθορισμένη αριθμητική τιμή που έχει εκχωρηθεί σε αυτόν τον κόμβο φύλλου γίνεται η τελική πρόβλεψη για την είσοδο. Τα δέντρα παλινδρόμησης προσφέρουν πολλά πλεονεκτήματα στη μηχανική μάθηση. Είναι εύκολο να κατανοηθούν και να ερμηνευτούν, παρέχοντας σαφείς γνώσεις σχετικά με τη διαδικασία λήψης αποφάσεων. Επιπλέον, μπορούν να χειριστούν τόσο κατηγορίες όσο και αριθμητικά χαρακτηριστικά χωρίς να απαιτείται εκτεταμένη προεπεξεργασία. Επιπλέον, τα δέντρα παλινδρόμησης μπορούν να συλλάβουν μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου, καθιστώντας τα καταλληλα για ένα ευρύ φάσμα εφαρμογών. Ωστόσο, είναι σημαντικό να σημειωθεί ότι τα δέντρα παλινδρόμησης είναι επιρρεπή σε υπερβολική προσαρμογή, ειδικά όταν το δέντρο γίνεται πολύ βαθύ και πολύπλοκο. Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο ταιριάζει υπερβολικά στα δεδομένα εκπαίδευσης, με αποτέλεσμα κακή γενίκευση σε μη ορατά δεδομένα. Τεχνικές όπως το κλάδεμα, το οποίο περιλαμβάνει την αφαίρεση ή την κατάρρευση ορισμένων κόμβων στο δέντρο ή τη χρήση μεθόδων συνόλου όπως τα τυχαία δάση, μπορούν να βοηθήσουν στον μετριασμό της υπερβολικής προσαρμογής και στη βελτίωση της προγνωστικής απόδοσης των δέντρων παλινδρόμησης.

Συνοπτικά, ένα δέντρο παλινδρόμησης είναι ένας ευέλικτος αλγόριθμος μηχανικής μάθησης για την πρόβλεψη αριθμητικών τιμών. Έχει τη δυνατότητα να διαχωρίζει αναδρομικά το σύνολο δεδομένων με βάση τις τιμές χαρακτηριστικών εκχωρώντας μέσες προβλέψεις σε κάθε υποσύνολο. Έτσι το δέντρο παλινδρόμησης καταγράφει σχέσεις και μοτίβα στα δεδομένα. Ιδιαίτερη προσοχή θα πρέπει να δοθεί στον έλεγχο του βάθους και της πολυπλοκότητας του δέντρου για να αποφευχθεί η υπερβολική προσαρμογή και να ενισχυθούν οι δυνατότητες γενίκευσης.

3.4.4 Τυχαία Δάση(Random forest)

Το Random Forest είναι ένας αλγόριθμος μάθησης με επίβλεψη που χρησιμοποιεί τη μέθοδο μάθησης συνόλου για ταξινόμηση και παλινδρόμηση. Εκτελεί n αριθμό δέντρων παλινδρόμησης και τα συνδυάζει σε ένα ενιαίο μοντέλο για να κάνει ακριβέστερη πρόβλεψη από ένα μεμονωμένο δέντρο. Ο RF κατασκευάζει πολλά δέντρα απόφασης κατά την εκπαίδευση και οι προβλέψεις από

όλα τα δέντρα συνδυάζονται για να γίνει η τελική πρόβλεψη. Χρησιμοποιώντας τυχαία δειγματοληψία με αντικατάσταση (bagging στην ορολογία της μηχανικής μάθησης), το RF βοηθά τους επιστήμονες δεδομένων να μειώσουν τη διακύμανση που σχετίζεται με τους αλγορίθμους που έχουν υψηλή διακύμανση, συνήθως τα δέντρα αποφάσεων. Δεδομένου ενός συνόλου εκπαίδευσης με χαρακτηριστικό X και έξοδο Y, η μέθοδος bagging επιλέγει επανειλημμένα ένα τυχαίο δείγμα του συνόλου εκπαίδευσης για m φορές (m=1,2,...,m) και προσαρμόζει τα δέντρα σε αυτά τα δείγματα

Για κάθε δέντρο, λαμβάνουμε μια ακολουθία περιπτώσεων που αντικαθίστανται με τυχαία δειγματοληψία από το σύνολο εκπαίδευσης. Κάθε ακολουθία περιπτώσεων αντιστοιχεί σε ένα τυχαίο διάνυσμα ⌀k που σχηματίζει ένα συγκεκριμένο δέντρο. Δεδομένου ότι όλες οι ακολουθίες δεν θα είναι ακριβώς ίδιες, τα δέντρα απόφασης που κατασκευάζονται από αυτές θα είναι επίσης ελαφρώς παραλλαγμένα. Οι De Aquino Afonso κ.ά. (Citation2020) προτείνουν ότι η πρόβλεψη του K-οστού δέντρου για μια είσοδο X μπορεί να αναπαρασταθεί από την εξίσωση (9).

$$hk(X) = h(X, 0K, \forall k \epsilon \{1, 2, \ldots, K\} \qquad (9)$$

Όπου K είναι ο αριθμός των δέντρων. Καθώς ένα δέντρο διασπάται, καθένα από τα οποία επιλέγει τυχαία τα χαρακτηριστικά για να αποφευχθούν οι συσχετίσεις μεταξύ των χαρακτηριστικών. Ο Alpaydin (Citation2009) επισημαίνει ότι ένας κόμβος S μπορεί να χωριστεί σε δύο υποσύνολα, S1 και S2, επιλέγοντας ένα κατώφλι c που ελαχιστοποιεί τη διαφορά στο άθροισμα των τετραγωνικών σφαλμάτων.

$$SSE = \left(\sum i \epsilon Si(vi - \frac{1}{|Si|}\sum i \epsilon S1\ vi\right)^2 + \sum i{:}\ i \epsilon Si(vi - \frac{1}{|S2|}\sum i \epsilon S2 vi)^2) \qquad (10)$$

Ακολουθώντας τους ίδιους κανόνες απόφασης, μπορούμε να προβλέψουμε οποιοδήποτε υποδέντρο ως τη μέση τιμή ή τη διάμεση τιμή εξόδου των περιπτώσεων. Τέλος, μπορούμε να λάβουμε την τελική πρόβλεψη ως μέσο όρο της εξόδου κάθε δέντρου, όπως αναφέρεται στην Εξίσωση Εξίσωση (11)

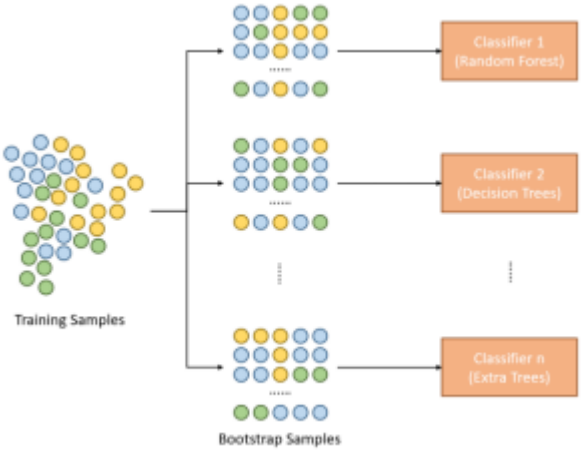
$$\text{h(X)}=\frac{1}{K}\sum_{i=1}^{K} hk(X) \qquad (11)$$

3.4.5 XGBoost

Η παλινδρόμηση XGBoost είναι η συντομογραφία για την παλινδρόμηση ακραίας κλίσης (extreme gradient boost regression). Ο XGBoost είναι ένας από τους καλύτερους αλγορίθμους μάθησης με επίβλεψη, ο οποίος μπορεί να συναχθεί από τον τρόπο που ρέει καθώς αποτελείται από την αντικειμενική συνάρτηση και τους βασικούς εκπαιδευόμενους. Η συνάρτηση απώλειας είναι παρούσα στην αντικειμενική συνάρτηση, και δείχνει τη διαφορά μεταξύ των πραγματικών

τιμών με αυτή των προβλεπόμενων τιμών, ενώ ο όρος κανονικοποίησης χρησιμοποιείται για να δείξει πόσο απέχει η πραγματική τιμή από την προβλεπόμενη τιμή. Η μάθηση συνόλου που χρησιμοποιείται στο XGBoost θεωρεί πολλά μοντέλα που είναι γνωστά ως εκπαιδευόμενοι βάσης για την πρόβλεψη μιας ενιαίας τιμής.

Ένα μοντέλο πρόβλεψης που προκύπτει από Μηχανική Μάθηση δεν είναι πάντα τέλειο. Η απόδοση του κυμαίνεται ανάλογα με τον αριθμό και την ποιότητα των δεδομένων και την καταλληλότητα του αλγορίθμου που χρησιμοποιήθηκε. Μια προφανής πρόταση σχετικά με την βελτίωση των προβλέψεων πάνω σε ένα συγκεκριμένο σύνολο δεδομένων είναι ο συνδυασμός των αποφάσεων πολλών διαφορετικών μοντέλων πρόβλεψης (ensemble techniques). Χαρακτηριστική μέθοδο για την πραγματοποίηση της παραπάνω πρότασης είναι αυτή της ενθυλάκωσης (Bagging) και ενίσχυσης (Boosting) που εφαρμόζει τον ίδιο αλγόριθμο σε διαφορετικά υποσύνολα δεδομένων. Για να μπορέσει να αποφέρει βελτιωμένα αποτελέσματα η παραπάνω μέθοδος, απαιτείται να εφαρμοστεί σε αλγορίθμους μάθησης οι οποίοι είναι ασταθείς (δηλ. έχουν μεγάλη διακύμανση (variance)). Αυτό σημαίνει, πως μικρές αλλαγές στα δεδομένα εκπαίδευσης προκαλούν αλλαγές και στο μοντέλο με αποτέλεσμα αυτό να βγάζει άλλες αποφάσεις. Ειδάλλως, η εκπαίδευση του κάθε αλγορίθμου σε διαφορετικό υποσύνολο, ενδεχομένως να μην προσέφερε και διαφορετικά μοντέλα.



Εικόνα 8: Bagging

Το Bagging αρχικά δημιουργεί (τυχαία) πολλά διαφορετικά σύνολα δεδομένων από το αρχικό μέσω "Δειγματοληψίας με Επανατοποθέτηση". Το κάθε νέο σύνολο δεδομένων έχει ίδιο αριθμό δεδομένων με το αρχικό, αλλά κάποια δεδομένα έχουν επαναληφθεί ενώ κάποια δεν έχουν συμπεριληφθεί καθόλου. Στη συνέχεια εφαρμόζει τον ασταθή αλγόριθμο μάθησης σε όλα τα νέα σύνολα δεδομένων και παράγει αντίστοιχα μοντέλα πρόβλεψης. Για τη διαδικασία πρόβλεψης λαμβάνονται υπόψιν οι αποφάσεις όλων των μοντέλων και η τελική τιμή είναι είτε η κατηγορία που συγκεντρώνει τις περισσότερες αποφάσεις μοντέλων (voting) είτε ο μέσος όρος των αριθμητικών προβλέψεων των διαφορετικών μοντέλων στην περίπτωση παλινδρόμησης.

Επέκταση της ενθυλάκωσης (Bagging), αποτελεί η ενίσχυση (Boosting). Είναι μια επαναληπτική διαδικασία που μετατρέπει όπως και παραπάνω, μετατρέπει ασταθείς αλγορίθμους σε ισχυρούς ελαττώνοντας το bias και το variance. Στηρίζεται στην ανάθεση βαρών (θετικών αριθμών) στα δεδομένα, έτσι ώστε ο αλγόριθμος μάθησης να επικεντρωθεί σε δεδομένα που συνήθως ταξινομούνται λάθος. Η πιθανότητα επιλογής ενός δεδομένου κατά τη δειγματοληψία είναι ανάλογη του βάρους του. Έτσι δεδομένα με μεγαλύτερο βάρος εμφανίζονται περισσότερες φορές ενώ δεδομένα με μικρότερο βάρος μπορεί να μην εμφανιστούν καθόλου. Ακόμα, τα βάρη χρησιμοποιούνται και για τον τρόπο υπολογισμού της απόδοσης ενός αλγορίθμου. Χωρίς βάρη, το ποσοστό λάθους είναι ο αριθμός των δεδομένων ελέγχου που ταξινομούνται λάθος προς το συνολικό αριθμό των δεδομένων ελέγχου. Με βάρη, είναι το άθροισμα των βαρών των δεδομένων ελέγχου που ταξινομούνται λάθος προς το συνολικό άθροισμα των βαρών του συνόλου των δεδομένων ελέγχου (σταθμισμένο σφάλμα). Έτσι δεδομένα που έχουν μεγάλο βάρος, είναι και πιο πιθανό να επιλεχθούν στο υποσύνολο εκπαίδευσης και κατ' επέκταση να τα μάθει ο αλγόριθμος. Αλλά προκαλούν και μεγαλύτερο σφάλμα όταν δεν προβλεφθούν σωστά.

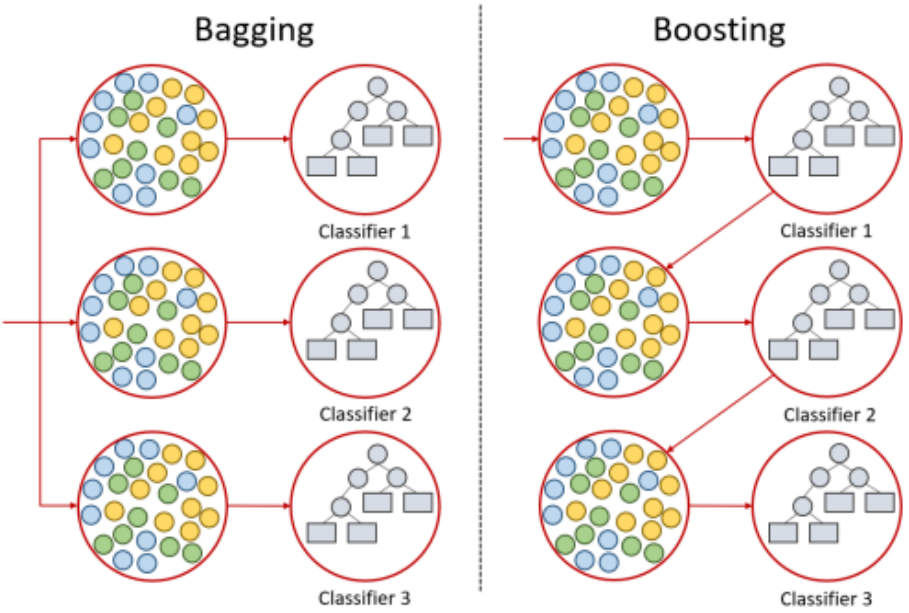
Η ενίσχυση (Boosting) είναι μια επαναληπτική διαδικασία όπου, τα διαφορετικά μοντέλα κατασκευάζονται το ένα μετά το άλλο και η απόδοση του προηγούμενου μοντέλου επηρεάζει την κατασκευή του επόμενου. Συγκεκριμένα προσπαθεί να κατασκευάσει το επόμενο μοντέλο έτσι ώστε να μην πραγματοποιεί τα ίδια λάθη με αυτά που έκανε το προηγούμενο (χρήση βαρών).

Διαδικασία Boosting:

1. Το πρώτο μοντέλο παράγεται από το αρχικό σύνολο δεδομένων. Θέτονται σε όλα τα δεδομένα του αρχικού συνόλου N ίσα βάρη ($1/N$).

2. Έπειτα τα δεδομένα ταξινομούνται από το μοντέλο. Αν το σταθμισμένο σφάλμα e είναι πάνω από 50% ($e > 0.5$), το μοντέλο απορρίπτεται, τα βάρη τίθενται στην τιμή $1/N$ και η διαδικασία επιστρέφει στο προηγούμενο βήμα. Ακόμα, αν η απόφαση του μοντέλου για κάποιο δεδομένο είναι λάθος τότε το βάρος του αυξάνεται ενώ αν είναι σωστή μειώνεται.

3. Η διαδικασία επαναλαμβάνεται από το βήμα 2 για τη μάθηση του επόμενου μοντέλου έως ότου επιτευχθεί ένα επιθυμητό όριο σφάλματος ή για προκαθορισμένο πλήθος κύκλων ενίσχυσης.



Εικόνα 9: Bagging VS Boosting

Αναφορικά με το κομμάτι ενημέρωσης των βαρών, πραγματοποιείται η παρακάτω διαδικασία. Για τα δεδομένα που ταξινομούνται λάθος το βάρος παραμένει όσο ήταν αρχικά, ενώ για αυτά που ταξινομούνται σωστά το βάρος μειώνεται αντιστρόφως ανάλογα με το ποσοστό λαθών e του ταξινομητή στα δεδομένα:

$$weight = weight\frac{e}{1-e} \qquad (12)$$

Το σταθμισμένο σφάλμα ισούται με το άθροισμα των βαρών των δεδομένων που ο ταξινομητής ταξινομεί λάθος δια N, δηλαδή:

$$ei = \frac{1}{N}\sum_{J=1}^{N} w \qquad (13)$$

Στη συνέχεια τα βάρη κανονικοποιούνται έτσι ώστε το άθροισμα τους να παραμείνει όσο και πριν. Κάθε βάρος διαιρείται με το άθροισμα των νέων βαρών και πολλαπλασιάζεται με το άθροισμα

των παλιών. Έτσι αυτόματα, κατά την διαδικασία δημιουργίας του συνόλου δεδομένων στο οποίο θα εκπαιδευτεί το επόμενο μοντέλο, αυξάνεται το βάρος των δεδομένων που ταξινομούνται λάθος και μειώνεται αυτών που ταξινομούνται σωστά . Ακόμα, στο στάδιο της πρόβλεψης άγνωστων δεδομένων, συνδυάζονται οι αποφάσεις όλων των μοντέλων μέσω ψηφοφορίας με βάρη. Το βάρος της απόφασης κάθε μοντέλου είναι αντίστοιχο του ποσοστού λαθών e στα δεδομένα από τα οποία εκπαιδεύτηκε:

$$weight = -log\frac{e}{1-e}$$

Το boosting είναι ένας μετα-αλγόριθμος που συνδυάζει μοντέλα μηχανικής μάθησης με σκοπό κυρίως τη μείωση της μεροληψίας αλλά και της διακύμανσης στην επιβλεπόμενη μάθηση μέσω μιας διαδικασία ελαχιστοποίησης μιας κυρτής συνάρτησης κόστους. Η τεχνική του boosting προτάθηκε στο πλαίσιο του αλγορίθμου Adaboost (adaptive boosting) ο οποίος χρησιμοποιεί δέντρα ταξινόμησης σαν βασικό αλγόριθμο. Νεότεροι αλγόριθμοι βασισμένοι στο boosting πετυχαίνουν καλύτερα αποτελέσματα, όπως οι XGBoost, LPBoost, Totalboost, BrownBoost, LogitBoost, MadaBoost και άλλοι. Η διαφορά μεταξύ τους είναι κυρίως στον τρόπο υπολογισμού των βαρών στα δεδομένα . Ο XGboost (Extreme Gradient boosting) προτάθηκε το 2016 από τους Tianqi Chen και Carlos Guestrin [37]. Είναι ένας αλγόριθμος ενίσχυσης ενός δέντρου απόφασης (ταξινόμησης/παρεμβολής) μέσω της επικλινούς ενίσχυσης (gradient boosting). Η επικλινής ενίσχυση είναι μια τεχνική ενίσχυσης που αντί να παράγει μοντέλα μεταβάλλοντας τα βάρη του συνόλου εκπαίδευσης, ενισχύει ένα αδύναμο μοντέλο μέσω μιας διαδικασίας βελτιστοποίησης επικλινούς καθόδου (gradient descent optimization procedure) ελαχιστοποιώντας μια κατάλληλη συνάρτηση κόστους. Πρόκειται για έναν εξαιρετικά δυνατό αλγόριθμο μηχανικής μάθησης που έχει αποκτήσει μεγάλη δημοσιότητα καθώς χρησιμοποιήθηκε από πολλές ερευνητικές ομάδες που κατάφεραν διακρίθηκαν σε διαγωνισμούς μηχανικής μάθησης.

3.4.6 LightGBM

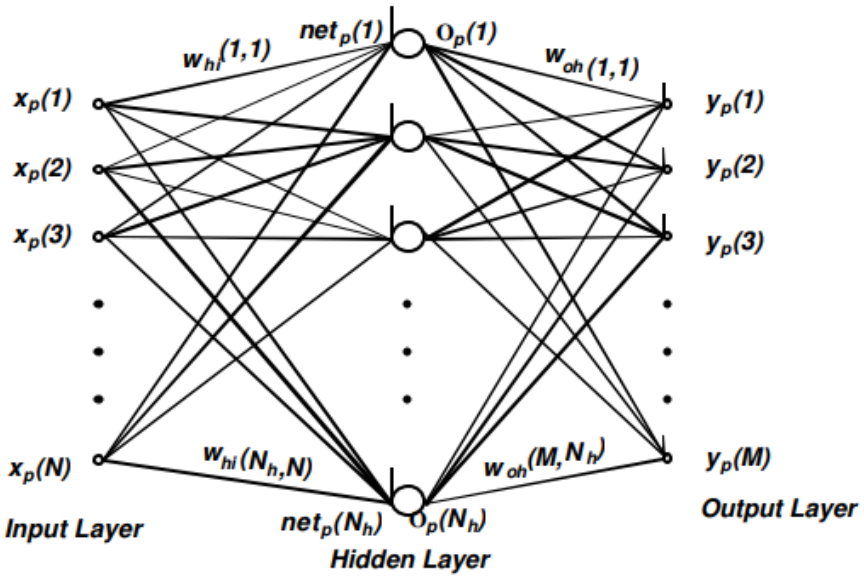
Η Light Gradient Boosting Machine (LightGBM) [23] είναι ένα δέντρο απόφασης που χρησιμοποιεί το πλαίσιο gradient boosting. Εισήχθη από τους ερευνητές της Microsoft, καθώς θεώρησαν ότι η ήδη υπάρχουσα μέθοδος XGBoost ήταν ανεπαρκής όσον αφορά την αποτελεσματικότητα και την επεκτασιμότητα της μεθόδου όταν το χαρακτηριστικό
διάσταση είναι υψηλή και το μέγεθος των δεδομένων είναι μεγάλο.

Η εύρεση των καλύτερων σημείων διάσπασης κατά τη διαδικασία μάθησης της ανάπτυξης ενός δέντρου αποφάσεων είναι το πιο χρονοβόρο κομμάτι [23]. Τα περισσότερα δέντρα αποφάσεων με ενίσχυση κλίσης (gradient boosting decision tree) (GBDT) μέθοδοι χρησιμοποιούν τον άπληστο αλγόριθμο, ο οποίος απαριθμεί κάθε πιθανό διάσπαση σε όλα τα χαρακτηριστικά. Ο αλγόριθμος εκτελεί το έργο της εύρεσης του πιο βέλτιστων διαχωρισμών, αλλά καταναλώνει πολλή μνήμη και είναι αναποτελεσματικός κατά την εκπαίδευση
διαδικασία.

Νευρωνικά δίκτυα

3.4.7 Multi Layer Perceptron

Τα νευρωνικά δίκτυα MLP αποτελούνται από μονάδες τοποθετημένες σε επίπεδα. Κάθε στρώμα αποτελείται από κόμβους και στα πλήρως συνδεδεμένα δίκτυα που εξετάζονται στην παρούσα εργασία κάθε κόμβος συνδέεται με κάθε κόμβο στα επόμενα στρώματα. Κάθε MLP αποτελείται από τουλάχιστον τρία στρώματα που αποτελείται από ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Ο παραπάνω ορισμός αγνοεί το εκφυλισμένο γραμμικό πολυστρωματικό perceptron που αποτελείται μόνο από ένα στρώμα εισόδου στρώμα και ένα στρώμα εξόδου. Το στρώμα εισόδου διανέμει τις εισόδους στα επόμενα στρώματα. Οι κόμβοι εισόδου έχουν γραμμικές συναρτήσεις ενεργοποίησης και δεν έχουν κατώφλια. Κάθε κόμβος κρυφής μονάδας και κάθε κόμβος εξόδου έχουν κατώτατα όρια που συνδέονται με αυτούς εκτός από τα βάρη. Οι κόμβοι της κρυφής μονάδας έχουν μη γραμμικές συναρτήσεις ενεργοποίησης και οι έξοδοι έχουν γραμμικές συναρτήσεις ενεργοποίησης. Ως εκ τούτου, κάθε σήμα που τροφοδοτεί έναν κόμβο σε ένα επόμενο επίπεδο έχει την αρχική είσοδο πολλαπλασιάζεται με ένα βάρος με προστιθέμενο κατώφλι και στη συνέχεια περνάει μέσα από μια συνάρτηση ενεργοποίησης που μπορεί να είναι γραμμική ή μη γραμμική (κρυφές μονάδες). Ένα τυπικό τριών επιπέδων δίκτυο παρουσιάζεται στο σχήμα 1. Για το πραγματικό MLP τριών στρωμάτων, όλες οι είσοδοι συνδέονται επίσης απευθείας με όλες τις εξόδους. Αυτές οι συνδέσεις δεν παρουσιάζονται στο σχήμα 1 για να απλοποιηθεί το διάγραμμα.[22]



Σχήμα 1: Multi layer

3.4.8 Gaussian Process Regression

Η εξαγωγή συμπερασμάτων για μια συνεχή συνάρτηση από ένα σύνολο μεμονωμένων (παρατηρημένων ή υπολογισμένων) σημείων δεδομένων είναι ένα συνηθισμένο έργο στην επιστημονική έρευνα. Ανάλογα με την προηγούμενη γνώση της διαδικασίας που διέπει τις παρατηρήσεις, διατίθεται ένα ευρύ φάσμα προσεγγίσεων. Εάν υπάρχει ένα εύλογο μοντέλο που μπορεί να μεταφραστεί σε έναν κλειστό λειτουργικό τύπο, η παραμετρική προσαρμογή είναι η πλέον κατάλληλη, καθώς τα περιορισμένα δεδομένα είναι συχνά επαρκή για την εκτίμηση των άγνωστων παραμέτρων. Παραδείγματα περιλαμβάνουν την αλληλεπίδραση πραγματικών (μη ιδανικών) σωματιδίων αερίου, την εξίσωση Arrhenius, ή, πιο κοντά στο θέμα της παρούσας ανασκόπησης, τη διάσπαση της ουράς μεγάλης εμβέλειας της αλληλεπίδρασης διασποράς van der Waals.

Στην πράξη, δεν μπορούν να μοντελοποιηθούν όλες οι διαδικασίες με απλές εκφράσεις. Οι σχέσεις δομής-ιδιοτήτων, η κινητική των βιομοριακών αντιδράσεων και οι κβαντικές αλληλεπιδράσεις πολλών σωμάτων είναι παραδείγματα παρατηρήσιμων αποτελεσμάτων που εξαρτώνται από τις μεταβλητές εισόδου με πολύπλοκο, μη εύκολα διαχωρίσιμο τρόπο, λόγω της παρουσίας κρυφών μεταβλητών. Αντί να προσπαθήσει κανείς να κατανοήσει αυτή την εξάρτηση αναλυτικά, μπορεί να προσπαθήσει να την περιγράψει με βάση τα υπάρχοντα δεδομένα και παρατηρήσεις. Οι τεχνικές παρεμβολής και παλινδρόμησης παρέχουν εργαλεία για τη συμπλήρωση του χώρου μεταξύ των σημείων δεδομένων, με αποτέλεσμα την αναπαράσταση

μιας συνεχούς συνάρτησης, η οποία, αφού καθιερωθεί, μπορεί να χρησιμοποιηθεί σε περαιτέρω εργασίες. Η γραμμική παρεμβολή και οι κυβικές γραμμές είναι ευρέως χρησιμοποιούμενα παραδείγματα αυτών των μεθόδων, αλλά περιορίζονται σε δεδομένα χαμηλών διαστάσεων και σε περιπτώσεις όπου υπάρχει μικρός θόρυβος στις παρατηρήσεις. Με περισσότερες από λίγες μεταβλητές, καθίσταται εκθετικά πιο δύσκολο να συλλεχθούν επαρκή δεδομένα για την ομοιόμορφη κάλυψη που απαιτείται από αυτές τις μεθόδους. Καθώς οι τεχνικές παρεμβολής είναι εγγενώς τοπικές, ο θόρυβος στις παρατηρήσεις δεν υπολογίζεται κατά μέσο όρο σε ένα ευρύτερο πεδίο, που σημαίνει ότι οι προσεγγίσεις αυτές τείνουν να είναι λιγότερο ανεκτικές στην αβεβαιότητα των δεδομένων.

Από τη σκοπιά του επαγγελματία, το GPR είναι ένα μη γραμμικό, μη παραμετρικό εργαλείο παλινδρόμησης, χρήσιμο για την παρεμβολή μεταξύ σημείων δεδομένων που είναι διασκορπισμένα σε ένα χώρο εισόδου υψηλής διάστασης. Βασίζεται στη θεωρία πιθανοτήτων του Bayes και έχει πολύ στενές σχέσεις με άλλες τεχνικές παλινδρόμησης, όπως η παλινδρόμηση πυρήνων (KRR) και η γραμμική παλινδρόμηση με συναρτήσεις ακτινικής βάσης. Η μη παραμετρική παλινδρόμηση δεν υποθέτει μια αναπαράσταση ή μια κλειστή λειτουργική μορφή, ούτε προσπαθεί να εξηγήσει τη διαδικασία που βρίσκεται πίσω από τα δεδομένα χρησιμοποιώντας θεωρητικές εκτιμήσεις. Αντίθετα, βασίζεται σε ένα μεγάλο όγκο δεδομένων για να προσαρμόσει μια ευέλικτη συνάρτηση με την οποία μπορούν να γίνουν προβλέψεις- αυτό είναι που αποκαλούμε "μηχανική μάθηση".Η ΓΠΡ παρέχει μια λύση στο πρόβλημα της μοντελοποίησης, έτσι ώστε η τοπικότητα της παρεμβολής να μπορεί να ελεγχθεί ρητά και ποσοτικά, κωδικοποιώντας την στην αρχική υπόθεση της ομαλότητας της υποκείμενης συνάρτησης. Για να εισαγάγουμε την GPR, θεωρούμε μια ομαλή, κανονική συνάρτηση, y(x), η οποία λαμβάνει ένα διαστατικό διάνυσμα ως είσοδο και το απεικονίζει σε μια μόνο κλιμακωτή τιμή:

$$y: R^d \rightarrow R \qquad (14)$$

Δεν γνωρίζουμε τη συναρτησιακή μορφή του y, αλλά έχουμε κάνει N ανεξάρτητες παρατηρήσεις, yn, της τιμής του στις θέσεις xn,με αποτέλεσμα ένα σύνολο δεδομένων

$$D = \{xn; yn\}^N \; n = 1 \qquad (15)$$

Μπορούμε να θεωρήσουμε ότι οι παρατηρήσεις,yn είναι δείγματα του y(x) στη δεδομένη θέση, τα οποία μπορεί να περιέχουν θόρυβο παρατήρησης. Ο στόχος είναι τώρα να χρησιμοποιήσουμε αυτές τις τιμές δεδομένων για να δημιουργήσουμε έναν εκτιμητή που μπορεί να προβλέψει τη συνεχή συνάρτηση y (x) σε απόκεντρες τοποθεσίες x και επίσης να ποσοτικοποιήσουμε την αβεβαιότητα ("αναμενόμενο σφάλμα") αυτής της πρόβλεψης.Υπάρχουν δύο ισοδύναμες προσεγγίσεις για την εξαγωγή του πλαισίου GPR: η θεώρηση του χώρου βαρών και η θεώρηση του χώρου συναρτήσεων, η καθεμία από τις οποίες αναδεικνύει κάπως διαφορετικές πτυχές της διαδικασίας προσαρμογής.

Κεφάλαιο 4-Η Γλώσσα Προγραμματισμού Python

Για να υλοποιήσουμε τα μοντέλα μηχανικής μάθησης χρησιμοποιούμε τη γλώσσα προγραμματισμού Python και το γραφικό περιβάλλον της Google, Google Colaboratory. Η Python είναι μια σύγχρονη και η πιο διαδεδομένη γλώσσα παγκοσμίως στις μέρες μας. Είναι, πλέον, γνωστό πως χρησιμοποιείται ευρέως από σπουδαστές, ερευνητές και εργαζόμενους σε τεχνολογικές εταιρείες για την ανάπτυξη λογισμικού και εφαρμογών, αλλά και την πρακτική εφαρμογή μεθόδων Τεχνητής Νοημοσύνης και Ανάλυσης Δεδομένων. Η Python αποτελεί, μια εύκολη γλώσσα προγραμματισμού όσον αφορά την εκμάθηση της, παρέχοντας μέσα από το Διαδίκτυο και τους χρήστες πληθώρα οδηγών μάθησης και βοήθειας. Το Google Colaboratory, εν συντομία Google Colab αποτελεί ένα γραφικό αλληλεπιδρούμενο περιβάλλον (GUI), το οποίο κατασκευάστηκε από την Google με σκοπό την μεγαλύτερη ευκολία των χρηστών στην σύνταξη κώδικα, την όσο πιο 66 γίνεται απλουστευμένη χρήση της Python με άμεση πρόσβαση στις διάφορες βιβλιοθήκες της χωρίς να χρειάζεται η εγκατάσταση τους και την παροχή υπολογιστικής μνήμης (RAM) και κάρτας γραφικών (GPU) με σκοπό την βελτιστοποίηση της απόδοσης του προγράμματος. Το Google Colab λειτουργεί ως σημειωματάριο (notebook) της Python σε παρόμοια μορφή με την οποία λειτουργεί το Jupyter Notebook.

4.1 Python

Για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python και συγκεκριμένα Python 3.6.

Η Python είναι διερμηνευόμενη (interpreted), γενικού σκοπού (general-purpose) και υψηλού επιπέδου, γλώσσα προγραμματισμού. Ανήκει στις γλώσσες προστακτικού προγραμματισμού (Imperativeprogramming) και υποστηρίζει τόσο το διαδικαστικό (proceduralprogramming) όσο και το αντικειμενοστραφές (object-orientedprogramming) προγραμματιστικό υπόδειγμα (programmingparadigm). Είναι δυναμική γλώσσα προγραμματισμού (dynamicallytyped) και υποστηρίζει συλλογή απορριμμάτων (garbagecollection ή GC)[49], [50].

Δημιουργήθηκε από τον Ολλανδό Γκίντο βαν Ρόσσουμ (GuidovanRossum) στο ερευνητικό κέντρο CentrumWiskunde&Informatica (CWI) το 1989 και κυκλοφόρησε για πρώτη φορά το 1991.

Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Το συντακτικό της, επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ότι θα ήταν δυνατόν σε γλώσσες όπως η C++ ή η Java. Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες και για την ταχύτητα εκμάθησης της. Μειονεκτεί στο ότι επειδή είναι διερμηνευόμενη είναι πιο αργή από τις μεταγλωττιζόμενες (compiled) γλώσσες, όπως η C και η C++. Για αυτόν τον λόγο δεν είναι κατάλληλη για γραφή λειτουργικών συστημάτων.

Οι διερμηνευτές της Python είναι διαθέσιμοι για εγκατάσταση σε πολλά λειτουργικά συστήματα, επιτρέποντας στην Python την εκτέλεση κώδικα σε ευρεία γκάμα συστημάτων. Χρησιμοποιώντας εργαλεία τρίτων, όπως το Py2exe ή το Pyinstaller, ο κώδικας της Python μπορεί να πακεταριστεί σε

αυτόνομα εκτελέσιμα προγράμματα για μερικά από τα πιο δημοφιλή λειτουργικά συστήματα, επιτρέποντας τη διανομή του βασισμένου σε Python λογισμικού για χρήση σε αυτά τα περιβάλλοντα χωρίς να απαιτείται εγκατάσταση του διερμηνευτή της Python.

Η Python αναπτύσσεται ως ανοιχτό λογισμικό (opensource) και η διαχείρισή της γίνεται από τον μη κερδοσκοπικό οργανισμό Python Software Foundation. Ο κώδικας διανέμεται με την άδεια Python Software FoundationLicense η οποία είναι συμβατή με την GPL. Το όνομα της γλώσσας προέρχεται από την ομάδα των Άγγλων κωμικών ΜόντυΠάιθον.

Για να εκτελεστούν διαδραστικά (interactively) μεμονωμένες εντολές ή προγράμματα είναι απαραίτητη η εγκατάσταση του διερμηνευτή της Python, ο οποίος είναι ελεύθερα διαθέσιμος για «κατέβασμα» (download) από τον επίσημο ιστότοπό της (www.python.org). Για Microsoft Windows υπάρχουν εκδόσεις των 32 ή 64 bits. Στα λειτουργικά συστήματα Linux και Mac OS X συνηθίζεται να είναι προεγκατεστημένη, πιθανότατα όμως να είναι μια παλαιότερη έκδοσή της.

Για τη συγγραφή προγραμμάτων είναι απαραίτητος ένας κειμενογράφος ή ακόμα καλύτερα ένα ολοκληρωμένο περιβάλλον ανάπτυξης (Integrated Development Environment - IDE), το οποίο είναι ένα ειδικό λογισμικό για την ανάπτυξη εφαρμογών. Η Python έρχεται μαζί με ένα εύχρηστο και απλό περιβάλλον ανάπτυξης με την ονομασία IDLE. Τα αρχικά του προέρχονται από τις λέξεις Interactive DeveLopmentEnvironment και είναι γραμμένο σε Python από τον Γκίντο βαν Ρόσσουμ. Χρησιμοποιεί τη βιβλιοθήκη γραφικών Tkinter, οπότε μπορεί να εκτελεσθεί σε περιβάλλον Linux, Windows και Mac OS X. Για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας ως ολοκληρωμένο περιβάλλον ανάπτυξης το PyCharm, με δυνατότητες επεξεργασίας, συγγραφής αποσφαλμάτωσης (debugging), κ.λπ.

4.2 Η Συμβολή της γλώσσας Python

NumPy

Το πακέτο NumPy εξασφαλίζει την αποτελεσματική αποθήκευση και επεξεργασία αριθμητικών πινάκων.Η αποτελεσματική αποθήκευση και επεξεργασία αριθμητικών πινάκων είναι απολύτως θεμελιώδης για την τη διαδικασία της επιστήμης των δεδομένων. Η Python διαθέτει εξειδικευμένα εργαλεία για το χειρισμό τέτοιων αριθμητικών πινάκων: το πακέτο NumPy και το πακέτο Pandas. Η NumPy, η συντομογραφία του Numerical Python, είναι ένα από τα σημαντικότερα θεμελιώδη πακέτα για τους αριθμητικούς υπολογισμούς στην Python. Τα περισσότερα υπολογιστικά πακέτα που παρέχουν επιστημονικές λειτουργίες χρησιμοποιούν τα αντικείμενα συστοιχιών του NumPy για την ανταλλαγή δεδομένων.

Η NumPy είναι μια βιβλιοθήκη ανοικτού κώδικα διαθέσιμη στην Python, η οποία βοηθάει στην μαθηματικό, επιστημονικό, μηχανικό και επιστημονικό προγραμματισμό δεδομένων. Είναι μια πολύ χρήσιμη βιβλιοθήκη για την εκτέλεση μαθηματικών και στατιστικών πράξεων στην Python. Λειτουργεί τέλεια για πολυδιάστατους πίνακες και πολλαπλασιασμό πινάκων και είναι εύκολο να να ενσωματωθεί με τη C/C++ και τη Fortran. Η NumPy παρέχει αναρίθμητα χαρακτηριστικά που μειώνουν τα περίπλοκα καθήκοντα των αναλυτών δεδομένων, των επιστημόνων δεδομένων και των ερευνητών μια

κ.λπ.

Η NumPy παρέχει τόσο την ευελιξία της Python όσο και την ταχύτητα ενός καλά βελτιστοποιημένου μεταγλωττισμένου κώδικα C. Το εύχρηστο συντακτικό του τη καθιστά ιδιαίτερα προσιτή και παραγωγική για τους προγραμματιστές. Έχει κατασκευαστεί για να λειτουργεί με πίνακες N-διαστάσεων,γραμμική άλγεβρα, τυχαίους αριθμούς, μετασχηματισμό Fourier κ.λπ. Αφού αποκτηθεί η βασική άνεση με το περιβάλλον της Python, αξίζει να εξερευνηθεί η βιβλιοθήκη NumPy. Ενώ η NumPy παρέχει μια υπολογιστική βάση για γενική επεξεργασία αριθμητικών δεδομένων, πολλοί αναγνώστες θα θελήσουν να χρησιμοποιήσουν το pandas ως βάση για τα περισσότερα είδη στατιστικής ή ανάλυσης, ειδικά για δεδομένα σε πίνακες. Το Pandas παρέχει επίσης κάποια πιο εξειδικευμένα πεδία λειτουργικότητας, όπως ο χειρισμός χρονοσειρών, κάτι που δεν υπάρχει στο NumPy.Η NumPy δημιουργήθηκε το 2005 από τον TravisOliphant και είναι ένα εργαλείο ανοιχτού κώδικα που μπορεί να χρησιμοποιηθεί ελεύθερα.

Η NumPy είναι γρήγορη και αποδοτική στη μνήμη, που σημαίνει ότι μπορεί να χειριστεί τεράστιο όγκο δεδομένων καθιστώντας την πιο προσιτή από οποιαδήποτε άλλη βιβλιοθήκη. Στην πραγματικότητα, το TensorFlow και το Scikit learn χρησιμοποιούν πίνακες NumPy για τον υπολογισμό τον πολλαπλασιασμό πινάκων στο backend.[22]

Κύρια πλεονεκτήματα της εργασίας με το NumPy για την ανάλυση δεδομένων είναι πως:

- Η NumPy είναι ιδιαίτερα χρήσιμη για τη δημιουργία αντικειμένων δεδομένων με N διαστάσεις.
- Το πλαισιο του αποδίδει γρήγορα και ομαλά όταν εργάζεται σε ομοιογενή σύνολα δεδομένων.

- Όταν χρησιμοποιούνται για αριθμητικούς υπολογισμούς, οι πίνακες NumPy χρησιμοποιούν λιγότερη μνήμη από τις λίστες της Python. Επιτρέπει επίσης στους χρήστες να καθορίζουν τους τύπους δεδομένων στα περιεχόμενα, γεγονός που μπορεί να βελτιστοποιήσει τον κώδικα.
- Η NumPy μπορεί να αποθηκεύει αποτελεσματικά δεδομένα και λειτουργίες δεδομένων, ειδικά καθώς οι πίνακες αυξάνονται σε μέγεθος.
- Δεν είναι δύσκολο να εκτελεστούν μαθηματικές πράξεις στα δεδομένα που είναι αποθηκευμένα στη NumPy.
- Το NumPy επιτρέπει στους χρήστες να αυξήσουν την ταχύτητα της ροής εργασιών τους.
- Είναι σε θέση να διασυνδεθεί με άλλα πακέτα Python. Δεδομένου ότι το NumPy υπάρχει εδώ και σχετικά μεγάλο χρονικό διάστημα, σχεδόν όλα τα πακέτα μηχανικής μάθησης και ανάλυσης δεδομένων για την Python χρησιμοποιούν το NumPy με κάποια ιδιότητα.

Το Scikit-learn αξιοποιεί αυτό το πλούσιο περιβάλλον για να παρέχει υπερσύγχρονες υλοποιήσεις πολλών γνωστούς αλγορίθμων μηχανικής μάθησης, διατηρώντας παράλληλα μια εύχρηστη διεπαφή στενά ενσωματωμένη με τη γλώσσα Python. Αυτό απαντά στην αυξανόμενη ανάγκη για στατιστική ανάλυση δεδομένων με μη ειδικούς στις βιομηχανίες λογισμικού και διαδικτύου, καθώς και σε τομείς εκτός της επιστήμης των υπολογιστών, όπως η βιολογία ή η φυσική. Η **scikit-learn (sklearn)**[23], αποτελεί μια ανοιχτού κώδικα βιβλιοθήκη της Python η οποία υλοποιήθηκε με στόχο την ανάπτυξη αλγορίθμων μηχανικής μάθησης.Η βιβλιοθήκη sklearn περιέχει πολλά αποτελεσματικά εργαλεία για μηχανική μάθηση και στατιστική μοντελοποίηση, όπως ταξινόμηση, παλινδρόμηση, ομαδοποίηση και μείωση διαστάσεων.

Το Scikit-learn διαφέρει από άλλες εργαλειοθήκες μηχανικής μάθησης στην Python για διάφορους λόγους:

i) διανέμεται υπό την άδεια BSD

ii) ενσωματώνει μεταγλωττισμένο κώδικα για αποτελεσματικότητα, σε αντίθεση με το MDP (Zito et al., 2008) και το pybrain (Schaul et al., 2010),

iii) εξαρτάται μόνο από το Numpy και Scipy για να διευκολύνει την εύκολη διανομή,

iv) εστιάζει στον προστακτικό προγραμματισμό

Ενώ το πακέτο είναι γραμμένο κυρίως σε Python, ενσωματώνει τις βιβλιοθήκες C++ LibSVM (Chang and Lin, 2001) και LibLinear (Fan et al., 2008) που παρέχουν υλοποιήσεις αναφοράς των SVM και των γενικευμένων γραμμικών μοντέλων με συμβατές άδειες χρήσης.

Η sklearn διαθέτει μερικές από τις ακόλουθες δυνατότητες:

1) Supervised learning algorithms
2) Cross-validation
3) Unsupervised learning algorithms
4) Various toy datasets

5) Feature extraction

Και πολλά άλλα.

Pandas

Το Pandas είναι μια βιβλιοθήκη επεξεργασίας και ανάλυσης δεδομένων ανοικτού κώδικα για την Python. Παρέχει εύχρηστες δομές δεδομένων, όπως Series και DataFrame, καθώς και μια ποικιλία συναρτήσεων για τον χειρισμό και την ανάλυση δομημένων δεδομένων. Το Pandas είναι ιδιαίτερα κατάλληλο για το χειρισμό δεδομένων σε πίνακες και χρονοσειρές.Διαθέτει λειτουργίες για την ανάλυση, τον καθαρισμό, την εξερεύνηση και τον χειρισμό δεδομένων. Το όνομα "Pandas" έχει αναφορά τόσο στο "Panel Data", όσο και στο "Python Data Analysis" και δημιουργήθηκε από τον Wes McKinney το 2008.Το Pandas είναι ένα πακέτο Python ανοικτού κώδικα με άδεια BSD που βασίζεται στο NumPy. Χρησιμοποιείται γενικά για εργασίες μηχανικής μάθησης, καθώς και για την ανάλυση δεδομένων και την επιστήμη των δεδομένων. Το Pandas προσφέρει φιλικές, εύχρηστες δομές δεδομένων και εργαλεία ανάλυσης για την εργασία με χρονοσειρές και αριθμητικά δεδομένα.[24]

Το Pandas θεωρείται ένα από τα καλύτερα πακέτα επεξεργασίας δεδομένων. Λειτουργεί επίσης καλά με διάφορες άλλες μονάδες Python για την επιστήμη δεδομένων. Συνδυάζοντας τη λειτουργικότητα του Matplotlib και του NumPy, το Pandas προσφέρει στους χρήστες ένα ισχυρό εργαλείο για την εκτέλεση αναλύσεων και οπτικοποίησης δεδομένων.

Ο παρακάτω κατάλογος επισημαίνει ορισμένες από τις πιο χρήσιμες λειτουργίες που προσφέρει το Pandas για την ανάλυση δεδομένων:

- Το Pandas είναι γνωστό για την εξαιρετική του ικανότητα να αναπαριστά και να οργανώνει δεδομένα.
- Η βιβλιοθήκη Pandas δημιουργήθηκε για να μπορεί να εργάζεται με μεγάλα σύνολα δεδομένων πιο γρήγορα και πιο αποτελεσματικά από οποιαδήποτε άλλη βιβλιοθήκη. Υπερέχει στην ανάλυση τεράστιων ποσοτήτων δεδομένων.
- Τα δεδομένα μπορούν να εισαχθούν στην Pandas από μια ποικιλία μορφών αρχείων, όπως SQL, Excel και JSON, μεταξύ άλλων.
- Όταν ένας χρήστης του Pandas γράφει μια ή δύο γραμμές κώδικα, είναι δυνατόν να εκτελέσει εργασίες που θα απαιτούσαν περισσότερες από δέκα ή δεκαπέντε γραμμές κώδικα χρησιμοποιώντας Java ή C++. Αυτή η αποτελεσματικότητα βοηθά τους αρχάριους να εργαστούν με το Pandas.
- Το Pandas θεωρείται μια ισχυρή βιβλιοθήκη που διαθέτει μια σειρά από χαρακτηριστικά και εντολές που διευκολύνουν την ανάλυση δεδομένων.
- Επειδή η Python είναι μια από τις πιο δημοφιλείς γλώσσες προγραμματισμού στον κόσμο, η εκμάθηση κώδικα σε Pandas για Python είναι μια ευέλικτη και εμπορεύσιμη δεξιότητα που μπορεί να κερδίσει την προσοχή των εργοδοτών.
- Οι χρήστες μπορούν να επεξεργαστούν και να προσαρμόσουν το Pandas επιλέγοντας από τον εκτεταμένο κατάλογο χαρακτηριστικών του.

Τα βασικά χαρακτηριστικά των πάντα περιλαμβάνουν:

- DataFrame: Μια δισδιάστατη, μεταβλητή σε μέγεθος και δυνητικά ετερογενής δομή δεδομένων σε μορφή πίνακα με επισημασμένους άξονες (γραμμές και στήλες). Είναι παρόμοια με ένα λογιστικό φύλλο ή έναν πίνακα SQL.
- Series: Ένας μονοδιάστατος επισημασμένος πίνακας που μπορεί να περιέχει οποιονδήποτε τύπο δεδομένων.
- Data Cleaning: Το Pandas παρέχει εργαλεία για τον καθαρισμό και την προεπεξεργασία δεδομένων, τον χειρισμό ελλιπών τιμών και την αναδιαμόρφωση συνόλων δεδομένων.
- Data Analysis: Υποστηρίζει διάφορες λειτουργίες ανάλυσης δεδομένων, όπως φιλτράρισμα, ομαδοποίηση, συνάθροιση και στατιστική ανάλυση.
- Ενσωμάτωση με άλλες βιβλιοθήκες: Το Pandas ενσωματώνεται καλά με άλλες δημοφιλείς βιβλιοθήκες της Python, όπως οι NumPy, Matplotlib και scikit-learn.

```python
import pandas as pd
```

Μετά την εισαγωγή, μπορούμε να δημιουργήσουμε και να χειριστούμε DataFrames, να εκτελέσουμε ανάλυση δεδομένων και να απεικονίσουμε τα αποτελέσματά μας χρησιμοποιώντας την εκτεταμένη λειτουργικότητα του pandas.

-import pandas: Αυτό το μέρος του κώδικα δίνει εντολή στην Python να εισάγει τη βιβλιοθήκη pandas. Η λέξη-κλειδί import χρησιμοποιείται για την εισαγωγή εξωτερικών ενοτήτων ή βιβλιοθηκών στο σενάριο ή το περιβάλλον της Python.

-as pd: Αυτό το μέρος του κώδικα εκχωρεί ένα ψευδώνυμο στην εισαγόμενη βιβλιοθήκη pandas, διευκολύνοντας την αναφορά της στον κώδικά σας. Σε αυτή την περίπτωση, το ψευδώνυμο "pd" είναι μια συνήθης σύμβαση που χρησιμοποιείται στην κοινότητα της επιστήμης δεδομένων της Python.

Το *Shap (SHapley Additive exPlanations)* είναι μια δημοφιλής βιβλιοθήκη στη μηχανική μάθηση για την ερμηνεία της εξόδου των μοντέλων μηχανικής μάθησης. Βασίζεται στις τιμές Shapley, μια έννοια από τη θεωρία συνεργατικών παιγνίων. Οι τιμές Shapley παρέχουν έναν τρόπο για τη δίκαιη κατανομή μιας αξίας μεταξύ μιας ομάδας συνεισφερόντων. Στο πλαίσιο της μηχανικής μάθησης, οι τιμές Shapley χρησιμοποιούνται για την απόδοση της πρόβλεψης ενός μοντέλου σε κάθε χαρακτηριστικό, υποδεικνύοντας τη συμβολή κάθε χαρακτηριστικού στην τελική πρόβλεψη.

Η βιβλιοθήκη Shap είναι ιδιαίτερα χρήσιμη για την κατανόηση και επεξήγηση της εξόδου πολύπλοκων μοντέλων, όπως τα μοντέλα συνόλου, τα βαθιά νευρωνικά δίκτυα και άλλα. Παρέχει

ένα ενοποιημένο πλαίσιο για την ανάλυση της σημασίας των χαρακτηριστικών και την ερμηνεία των μοντέλων. Οι τιμές Shap μπορούν να χρησιμοποιηθούν για την εξήγηση μεμονωμένων προβλέψεων ή για την απόκτηση πληροφοριών σχετικά με τη συνολική συμπεριφορά του μοντέλου.

Οι χρήστες μπορούν να απεικονίσουν τις τιμές Shap μέσω διαφόρων γραφημάτων, όπως γραφήματα σύνοψης, γραφήματα δύναμης και γραφήματα εξάρτησης, για να κατανοήσουν πώς τα διάφορα χαρακτηριστικά επηρεάζουν τις προβλέψεις του μοντέλου. Συνολικά, το Shap είναι ένα πολύτιμο εργαλείο για την επεξηγηματικότητα και την ερμηνευσιμότητα του μοντέλου στη μηχανική μάθηση.

Η βιβλιοθήκη Shap στην Python είναι ένα ισχυρό εργαλείο για την επεξήγηση της εξόδου των μοντέλων μηχανικής μάθησης. Παρέχει μια ενοποιημένη προσέγγιση για τη σημασία των χαρακτηριστικών και την ερμηνεία των μοντέλων με τη χρήση των τιμών Shapley. Ακολουθεί μια σύντομη επισκόπηση του τρόπου χρήσης της βιβλιοθήκης shap στην Python:

- Εγκατάσταση
- Βασική χρήση:

Εισαγωγή των απαραίτητων ενοτήτων και εκφόρτωση του εκπαιδευμένου μοντέλου μηχανικής μάθησης.

- Δημιουργία τιμών Shap:

Χρήση της βιβλιοθήκης shap για υπολογισμό των τιμών Shap για ένα σύνολο δειγμάτων εισόδου. Αυτό συνήθως γίνεται με τη χρήση της κλάσης shap.Explainer

- Οπτικοποίηση:

Οπτικοποίηση των τιμών Shap χρησιμοποιώντας διάφορα γραφήματα που παρέχονται από τη βιβλιοθήκη shap. Οι συνήθεις μέθοδοι οπτικοποίησης περιλαμβάνουν συνοπτικά διαγράμματα, διαγράμματα δύναμης και διαγράμματα εξάρτησης.

4.3 Υλοποίηση σε γλώσσα Python

Ξεκινώντας τον κώδικα ενημερώνουμε το colab απο που προέρχονται τα αρχεία που έχουμε για επεξεργασία .

```
from google.colab import drive
drive.mount('/content/drive')
```

Αυτό είναι χρήσιμο όταν έχουμε δεδομένα ή αρχεία αποθηκευμένα στο Google Drive που θέλουμε να χρησιμοποιήσουμε ή να επεξεργαστούμε στο σημειωματάριό Colab. Μετά την τοποθέτηση, μπορεί να πραγματοποιηθεί πλοήγηση στον κατάλογο του Google Drive χρησιμοποιώντας την αναφερόμενη διαδρομή και να αποκτηθεί πρόσβαση στα αρχεία που είναι αποθηκευμένα εκεί. Στον προηγούμενο κώδικα που δώσαμε, το αρχείο Excel φορτώθηκε από μια διαδρομή στο Google Drive και γι' αυτό είναι απαραίτητο να προσαρτήσουμε το Drive.

```
[ ]  !pip install shap
```

Η εντολή !pip install shap χρησιμοποιείται για την εγκατάσταση της βιβλιοθήκης SHAP (SHapley Additive exPlanations) σε ένα περιβάλλον Python, και συγκεκριμένα σε αυτή την περίπτωση, σε ένα σημειωματάριο του Google Colab.

Η SHAP είναι μια δημοφιλής βιβλιοθήκη για την ερμηνεία της εξόδου των μοντέλων μηχανικής μάθησης. Παρέχει ένα ενοποιημένο μέτρο της σημασίας των χαρακτηριστικών και βοηθά στην κατανόηση της επίδρασης κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου.

Η εκτέλεση αυτής της εντολής εγκαθιστά τη βιβλιοθήκη SHAP στο τρέχον περιβάλλον Python, καθιστώντας την διαθέσιμη για χρήση σε επόμενα κελιά κώδικα. Αν αντιμετωπίσετε προβλήματα κατά την εγκατάσταση, ίσως θελήσετε να ελέγξετε τη σύνδεσή σας στο διαδίκτυο ή να βεβαιωθείτε ότι έχετε τα απαραίτητα δικαιώματα για την εγκατάσταση πακέτων στο περιβάλλον Colab.

```
import pandas as pd
import numpy as np
from sklearn.metrics import mean_squared_error
import sklearn
import shap
from sklearn import ensemble
import matplotlib.pylab as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.svm import SVR
from sklearn import linear_model
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import DotProduct, WhiteKernel
from sklearn.metrics import mean_absolute_error
import xgboost
from sklearn.metrics import r2_score
```

Ο παραπάνω κώδικας είναι οι δηλώσεις εγκατάστασης και εισαγωγής για διάφορες βιβλιοθήκες και εργαλεία που χρησιμοποιούνται συνήθως στην ανάλυση δεδομένων και τη μηχανική μάθηση. Ας τον εξετάσουμε βήμα προς βήμα:

1. **Pandas and NumPy:**

   -Εισαγωγή του Pandas για χειρισμό δεδομένων και του NumPy για αριθμητικές πράξεις.

2. **Scikit-Learn (sklearn):**
3. **SHAP (SHapley Additive exPlanations):**
4. **Ensemble Learning (Random Forest, etc.):**
5. **Matplotlib:**
6. **MinMaxScaler:**
7. **Linear Regression:**
8. **Decision Tree Regressor:**
9. **Support Vector Regressor (SVR):**
10. **Linear Model:**
11. **Gaussian Process Regressor:**
12. **Mean Absolute Error and R2 Score:**
13. **XGBoost:**

Συνοπτικά, αυτός ο κώδικας δημιουργεί το περιβάλλον εισάγοντας βασικές βιβλιοθήκες και εργαλεία για την ανάλυση δεδομένων και τη μηχανική μάθηση. Αυτές οι βιβλιοθήκες θα χρησιμοποιηθούν για εργασίες όπως η επεξεργασία δεδομένων, η εκπαίδευση μοντέλων, η αξιολόγηση μοντέλων και η ερμηνεία μοντέλων.

```
path = "/content/drive/MyDrive/diplomatiki/House_pricing_data.xlsx"

df = pd.read_excel(path)

print(df.head())
print(df.keys())

df_drop = df.dropna()

trainX1 = df_drop[['PRICE ZONE',  'AREA', 'LEVELS',
       'FLOOR', 'KITCHENS', 'BATHROOMS', 'BEDROOMS', 'HEATING SYSTEM',
       'ENERGY CLASS', 'YEAR OF MANUFACTURE', 'YEAR OF RECONSTRUCTION',
       'ELEVATOR', 'AIR CONDITIONING', 'TYPE OF FLOORING', 'PETS',
       'SAFETY DOOR', 'FRAMES', 'SIEVES', 'TRANSPARENT', 'PAINTED', 'VIEW',
       'PARKING', 'DISTANCE FROM SUBWAY'
    ]]

trainY1 = df_drop[["PRICE"]]

#convert to np array
trainX, trainY = np.array(trainX1), np.array(trainY1)
print('trainX shape == {}.'.format(trainX.shape))
print('trainY shape == {}.'.format(trainY.shape))

#spliting the data
x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split(trainX, trainY, test_size = 0.2, shuffle=True)
print('x_train shape : ', x_train.shape)
print('y_train shape : ', y_train.shape)

print('x_test shape : ', x_test.shape)
print('y_test shape : ', y_test.shape)
```

Ο παρεχόμενος κώδικας είναι ένας αγωγός προεπεξεργασίας δεδομένων για μια εργασία παλινδρόμησης μηχανικής μάθησης. Ας τον αναλύσουμε βήμα προς βήμα:

1. Φόρτωση του συνόλου δεδομένων
2. Εμφάνιση πληροφοριών συνόλου δεδομένων
3. Χειρισμός ελλιπών τιμών
4. Επιλογή σχετικών στηλών:
5. Κανονικοποίηση των δεδομένων:
6. Εξαγωγή μεταβλητών χαρακτηριστικών και στόχων
7. Μετατροπή σε NumPy Arrays
8. Διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής
9. Εμφάνιση των σχημάτων των συνόλων εκπαίδευσης και δοκιμής

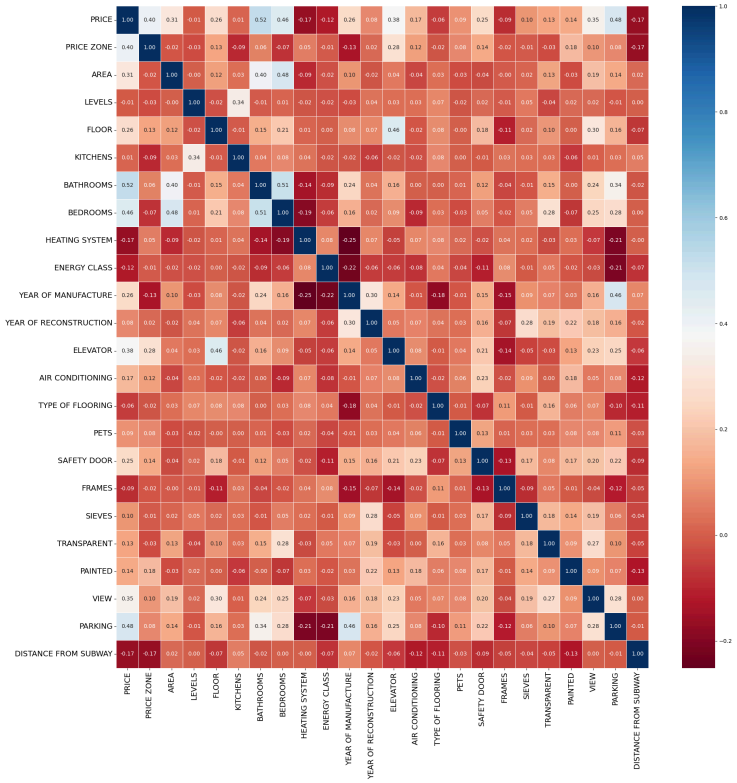
Συνοπτικά, αυτός ο κώδικας προετοιμάζει ένα σύνολο δεδομένων για ένα μοντέλο μηχανικής μάθησης παλινδρόμησης φορτώνοντας, καθαρίζοντας, επιλέγοντας τα σχετικά χαρακτηριστικά, κανονικοποιώντας (αν και δεν εφαρμόζεται) και χωρίζοντας τα δεδομένα σε σύνολα εκπαίδευσης και δοκιμής. Θέτει τις βάσεις για την εκπαίδευση και την αξιολόγηση μοντέλων παλινδρόμησης στα παρεχόμενα δεδομένα τιμολόγησης κατοικιών.

```
import seaborn as sb

cor = train.corr(method ='pearson')

plt.figure()
fig, ax = plt.subplots(figsize=(20,20))
plt.yticks(fontsize=14)
plt.xticks(fontsize=14)
fig = sb.heatmap(cor, cmap='RdBu', linewidths=0.5, annot=True, xticklabels= train.keys(), yticklabels=train.keys(), fmt='.2f')
```

Ο παραπάνω κώδικας παράγει έναν χάρτη θερμότητας για την απεικόνιση του πίνακα συσχέτισης Pearson των χαρακτηριστικών στο πλαίσιο δεδομένων του τρένου χρησιμοποιώντας τις βιβλιοθήκες seaborn (sb) και matplotlib (plt). Ας αναλύσουμε τον κώδικα:

1. Υπολογισμός του πίνακα συσχέτισης Pearson:
2. Δημιουργία σχήματος και υποδιαγράμματος:
3. Ορισμός μεγέθους γραμματοσειράς για ticks:
4. Δημιουργία χάρτη θερμότητας με χρήση του Seaborn:

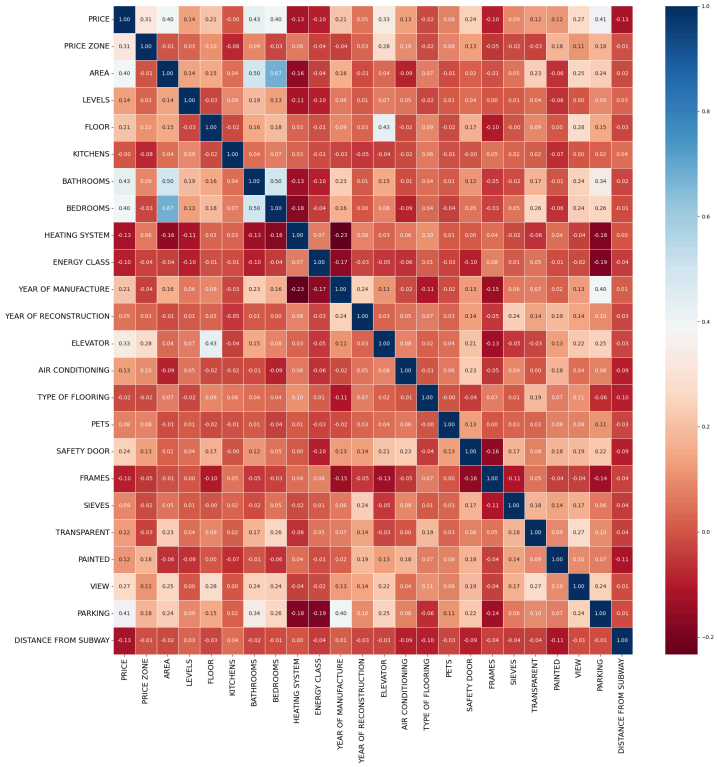
Αυτός ο κώδικας είναι χρήσιμος για την οπτικοποίηση της συσχέτισης μεταξύ διαφορετικών χαρακτηριστικών στο σύνολο δεδομένων. Ο χάρτης θερμότητας παρέχει έναν γρήγορο και διαισθητικό τρόπο για τον εντοπισμό μοτίβων και σχέσεων μεταξύ των χαρακτηριστικών. Οι θετικές συσχετίσεις υποδεικνύονται με ανοιχτότερα χρώματα, οι αρνητικές συσχετίσεις με πιο σκούρα χρώματα και η ένταση του χρώματος αντικατοπτρίζει την ισχύ της συσχέτισης.




```
[ ]  import seaborn as sb

     cor = train.corr(method ='kendall')

     plt.figure()
     fig, ax = plt.subplots(figsize=(20,20))
     plt.yticks(fontsize=14)
     plt.xticks(fontsize=14)
     fig = sb.heatmap(cor, cmap='RdBu', linewidths=0.5, annot=True, xticklabels= train.keys(), yticklabels=train.keys(), fmt='.2f')
```

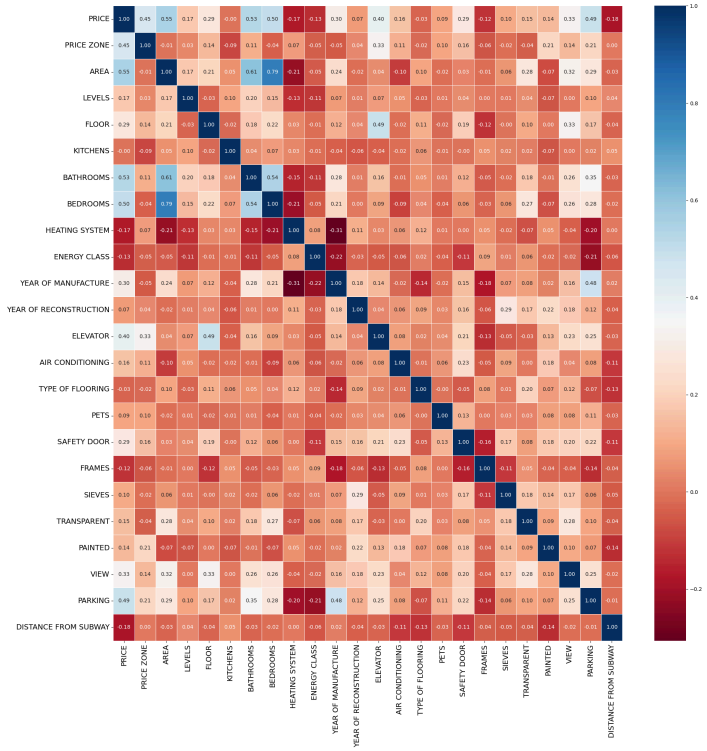
Αυτός ο κώδικας είναι παρόμοιος με τον προηγούμενο, αλλά αντί να χρησιμοποιεί τη συσχέτιση Pearson, υπολογίζει τον συντελεστή συσχέτισης Kendall rank και τον απεικονίζει χρησιμοποιώντας έναν χάρτη θερμότητας.

Είναι ιδιαίτερα χρήσιμος οταν ενδιαφερόμαστε να αξιολογήσουμε την ισχύ και την κατεύθυνση των μονοτονικών σχέσεων μεταξύ μεταβλητών, σε αντίθεση με τις γραμμικές σχέσεις που αξιολογούνται με τη συσχέτιση Pearson. Η συσχέτιση κατάταξης Kendall είναι λιγότερο ευαίσθητη στις ακραίες τιμές και λειτουργεί καλά για ταξινομημένα δεδομένα. Ο θερμικός χάρτης που προκύπτει παρέχει πληροφορίες σχετικά με τις μονοτονικές σχέσεις μεταξύ διαφορετικών χαρακτηριστικών στο σύνολο δεδομένων.




```
[ ]  import seaborn as sb

     cor = train.corr(method ='spearman')

     plt.figure()
     fig, ax = plt.subplots(figsize=(20,20))
     plt.yticks(fontsize=14)
     plt.xticks(fontsize=14)
     fig = sb.heatmap(cor, cmap='RdBu', linewidths=0.5, annot=True, xticklabels= train.keys(), yticklabels=train.keys(), fmt='.2f')
```

Ακόμη ένας κώδικας,παρόμοιος με τους προηγούμενους δύο είναι αυτός που υπολογίζει τον συντελεστή συσχέτισης Spearman και τον απεικονίζει χρησιμοποιώντας έναν ακόμα χάρτη θερμότητας. Είναι αρκετά κοντά στον Kendal όπως θα παρατηρήσουμε τον παρακάτω χάρτη.




```python
# Plain Linear Regression
lr = LinearRegression().fit(x_train, y_train.ravel())

# RMSE - MAE : LR
preds_lr = lr.predict(x_test)

#preds_lr= preds_lr*dif + min
#y_test = y_test*dif + min

rmse_lr = np.sqrt(mean_squared_error(y_test, preds_lr))
mae_lr = mean_absolute_error(y_test, preds_lr)
r2_score_lr = r2_score(y_test, preds_lr)
print("Linear Regression: RMSE on test set: {:.4f}".format(rmse_lr))
print("Linear Regression: MAE on test set: {:.4f}".format(mae_lr))
print("Linear Regression: R^2 on test set: {:.4f}".format(r2_score_lr))
print("")

# Decision Tree
dt = DecisionTreeRegressor(max_depth=2)
dt.fit(x_train, y_train.ravel())

# RMSE - MAE : DT
preds_dt = dt.predict(x_test)

#preds_dt= preds_dt*dif + min

rmse_dt = np.sqrt(mean_squared_error(y_test, preds_dt))
mae_dt = mean_absolute_error(y_test, preds_dt)
r2_score_DT = r2_score(y_test, preds_dt)
print("Decision Tree Regressor: RMSE on test set: {:.4f}".format(rmse_dt))
print("Decision Tree Regressor: MAE on test set: {:.4f}".format(mae_dt))
```

```
print("Linear Regrassion: R^2 on test set: {:.4f}".format(r2_score_lr))
print("")

# Decision Tree
dt = DecisionTreeRegressor(max_depth=2)
dt.fit(x_train, y_train.ravel())

# RMSE - MAE : DT
preds_dt = dt.predict(x_test)

#preds_dt= preds_dt*dif + min

rmse_dt = np.sqrt(mean_squared_error(y_test, preds_dt))
mae_dt = mean_absolute_error(y_test, preds_dt)
r2_score_DT = r2_score(y_test, preds_dt)
print("Decision Tree Regressor: RMSE on test set: {:.4f}".format(rmse_dt))
print("Decision Tree Regressor: MAE on test set: {:.4f}".format(mae_dt))
print("Decision Tree Regressor: R^2 on test set: {:.4f}".format(r2_score_DT))
print("")

# Support Vector Regression
svr = SVR(kernel="linear").fit(x_train, y_train.ravel())

# RMSE - SVR
preds_svr = svr.predict(x_test)

#preds_svr= preds_svr*dif + min

rmse_svr = np.sqrt(mean_squared_error(y_test, preds_svr))
mae_svr = mean_absolute_error(y_test, preds_svr)
r2_score_SVR = r2_score(y_test, preds_svr)
print("Support Vector Regression: RMSE on test set: {:.4f}".format(rmse_svr))
print("Support Vector Regression: MAE on test set: {:.4f}".format(mae_svr))
print("Support Vector Regression: R^2 on test set: {:.4f}".format(r2_score_SVR))
print("")
```

```
# Lasso Regression
lassoReg = linear_model.Lasso(alpha=0.92)
lassoReg.fit(x_train, y_train.ravel())

# RMSE - Lasso
preds_lasso = lassoReg.predict(x_test)

#preds_lasso= preds_lasso*dif + min

rmse_lasso = np.sqrt(mean_squared_error(y_test, preds_lasso))
mae_lasso = mean_absolute_error(y_test, preds_lasso)
r2_score_lasso = r2_score(y_test, preds_lasso)
print("Lasso Regression: RMSE on test set: {:.4f}".format(rmse_lasso))
print("Lasso Regression: MAE on test set: {:.4f}".format(mae_lasso))
print("Lasso Regression: R^2 on test set: {:.4f}".format(r2_score_lasso))
print("")

# Gaussian Process Regressor
kernel = DotProduct() + WhiteKernel()
gpr = GaussianProcessRegressor(kernel=kernel,random_state=0).fit(x_train, y_train.ravel())

# RMSE - Gaussian Process Regressor
preds_gpr = gpr.predict(x_test)

#preds_gpr= preds_gpr*dif + min

rmse_gpr = np.sqrt(mean_squared_error(y_test, preds_gpr))
mae_gpr = mean_absolute_error(y_test, preds_gpr)
r2_score_gpr = r2_score(y_test, preds_gpr)
print("Gaussian Process Regressor: RMSE on test set: {:.4f}".format(rmse_gpr))
print("Gaussian Process Regressor: MAE on test set: {:.4f}".format(mae_gpr))
print("Gaussian Process Regressor: R^2 on test set: {:.4f}".format(r2_score_gpr))
print("")

# Random Forest Regressor
regressor = ensemble.RandomForestRegressor(n_estimators=200)
regressor.fit(x_train, y_train.ravel())
```

```
    # RMSE
    preds_rfg = regressor.predict(x_test)

    #preds_rfg= preds_rfg*dif + min

    #print("preds shape : ", preds.shape)
    #print("y_test shape : ", y_test.shape)

    rmse_rfg = np.sqrt(mean_squared_error(y_test, preds_rfg))
    mae_rfg = mean_absolute_error(y_test, preds_rfg)
    r2_score_rfg = r2_score(y_test, preds_rfg)
    print("Random Forest Regression: RMSE on test set: {:.4f}".format(rmse_rfg))
    print("Random Forest Regression: MAE on test set: {:.4f}".format(mae_rfg))
    print("Random Forest Regression: R^2 on test set: {:.4f}".format(r2_score_rfg))
    print("")

    XGBR = xgboost.XGBRegressor().fit(x_train, y_train)
```

```
# RMSE

preds_xgb = XGBR.predict(x_test)

#preds_xgb= preds_xgb*dif + min

rmse_xgb = np.sqrt(mean_squared_error(y_test, preds_xgb))
mae_xgb = mean_absolute_error(y_test, preds_xgb)
r2_score_xgb = r2_score(y_test, preds_xgb)
print("XGBR : RMSE on test set: {:.4f}".format(rmse_xgb))
print("XGBR : MAE on test set: {:.4f}".format(mae_xgb))
print("XGBR : R^2 on test set: {:.4f}".format(r2_score_xgb))
print("")

from sklearn.neural_network import MLPRegressor
regr_MLP = MLPRegressor(hidden_layer_sizes=1000, random_state=1, max_iter=5000, batch_size= 32, early_stopping= True).fit(x_train, y_train)

preds_regr_MLP = regr_MLP.predict(x_test)

rmse_regr_MLP = np.sqrt(mean_squared_error(y_test, preds_regr_MLP))
mae_regr_MLP = mean_absolute_error(y_test, preds_regr_MLP)
r2_score_MLP = r2_score(y_test, preds_regr_MLP)
print("MLP: RMSE on test set: {:.4f}".format(rmse_regr_MLP))
print("MLP: MAE on test set: {:.4f}".format(mae_regr_MLP))
print("MLP: R^2 on test set: {:.4f}".format(r2_score_MLP))
print("")

import lightgbm as ltb

LGBMR = ltb.LGBMRegressor(n_estimators=200, num_leaves=100)
LGBMR.fit(x_train, y_train.ravel())

preds_LGBMR = LGBMR.predict(x_test)

rmse_LGBMR = np.sqrt(mean_squared_error(y_test, preds_LGBMR))
mae_LGBMR = mean_absolute_error(y_test, preds_LGBMR)
r2_score_LGBMR = r2_score(y_test, preds_LGBMR)
print("LGBMR : RMSE on test set: {:.4f}".format(rmse_LGBMR))
print("LGBMR : MAE on test set: {:.4f}".format(mae_LGBMR))
print("LGBMR : R^2 on test set: {:.4f}".format(r2_score_LGBMR))
```

- Γραμμική παλινδρόμηση (lr): Εκπαιδεύει ένα απλό μοντέλο γραμμικής παλινδρόμησης και αξιολογεί την απόδοσή του στο σύνολο δοκιμών.
- Δέντρο αποφάσεων (dt): Εκπαιδεύει ένα μοντέλο παλινδρόμησης δέντρου αποφάσεων και αξιολογεί την απόδοσή του.
- Υποστήριξη διανυσματικής παλινδρόμησης (svr): Εκπαιδεύει έναν παλινδρομητή διανύσματος υποστήριξης με γραμμικό πυρήνα και αξιολογεί την απόδοσή του.
- Παλινδρόμηση Lasso (lassoReg): Εκπαιδεύει ένα μοντέλο παλινδρόμησης Lasso και αξιολογεί την απόδοσή του.
- Gaussian Process Regressor (gpr): Εκπαιδεύει έναν παλινδρομητή διαδικασίας Gauss και αξιολογεί την απόδοσή του.
- Random Forest Regressor (regressor): Εκπαιδεύει έναν παλινδρομητή τυχαίου δάσους και αξιολογεί την απόδοσή του.
- XGBoost Regressor (XGBR): Εκπαιδεύει έναν παλινδρομητή XGBoost και αξιολογεί την απόδοσή του.
- MLP Regressor (regr_MLP): Εκπαιδεύει έναν παλινδρομητή multi-layer perceptron και αξιολογεί την απόδοσή του.
- LightGBM Regressor (LGBMR): Εκπαιδεύει έναν παλινδρομητή LightGBM και αξιολογεί την απόδοσή του.

Για κάθε μοντέλο, ο κώδικας εκτυπώνει τα αποτελέσματα RMSE, MAE και R^2 στο σύνολο δοκιμών, παρέχοντας πληροφορίες σχετικά με την απόδοση κάθε μοντέλου παλινδρόμησης στο σύνολο δεδομένων σας. Μπορεί να χρειαστούν προσαρμογές και τελειοποίηση των υπερπαραμέτρων ανάλογα με τα ιδιαίτερα χαρακτηριστικά των δεδομένων και του προβλήματος που πρόκειται να επιλυθεί.

```
import matplotlib.pyplot as plt

plt.rcParams["figure.figsize"] = [9.00, 10.50]
plt.rcParams["figure.autolayout"] = True

fig, ax =plt.subplots(1,1)

data=[[rmse_lr,mae_lr,r2_score_lr],
      [rmse_dt,mae_dt, r2_score_DT],
      [rmse_svr,mae_svr, r2_score_SVR],
      [rmse_gpr,mae_gpr, r2_score_gpr],
      [rmse_xgb,mae_xgb, r2_score_xgb],
      [rmse_regr_MLP,mae_regr_MLP, r2_score_MLP],
      [rmse_LGBMR,mae_LGBMR, r2_score_LGBMR],
      [rmse_rfg,mae_rfg, r2_score_rfg]
      ]

column_labels=["rMSE", "MAE", "R^2"]

row=["Linear Regrassion ",
     "Decision Tree",
     "Support Vector Regression",
     "Gaussian Process Regressor",
     "XGBOOST Regressor",
     "Multi-layer Perceptron regressor",
     "LightGBM regressor",
     "Random Forest Regressor"]

df=pd.DataFrame(data,columns=column_labels)

df.update(df.applymap('{:,.4f}'.format))
ax.axis('tight')
ax.axis('off')
the_table = ax.table(cellText=df.values,colLabels=df.columns,rowLabels=row,fontsize=15,loc="center",cellLoc='center')
the_table.auto_set_font_size(False)
the_table.set_fontsize(14)
plt.show()
```

Η ανάγκη εκτέλεσης του κώδικα προκύπτει επειδή τα συγκεκριμένα αριθμητικά αποτελέσματα για μετρικές όπως RMSE, MAE και R^2 εξαρτώνται από το σύνολο δεδομένων με το οποίο εργαζόμαστε. Αυτές οι μετρικές υπολογίζονται με βάση τις προβλέψεις του μοντέλου στο σύνολο δεδομένων σας και τα αποτελέσματα θα διαφέρουν ανάλογα με τα χαρακτηριστικά των δεδομένων σας.

Όταν εκτελέσουμε τον κώδικα, θα εκπαιδευτούν τα μοντέλα παλινδρόμησης στο σύνολο δεδομένων μας, θα κάνει προβλεφθούν στο σύνολο δοκιμών και στη συνέχεια θα υπολογιστούν και θα εμφανιστούν οι μετρικές απόδοσης για κάθε μοντέλο. Έτσι θα μας επιτραπεί να δούμε πόσο καλά αποδίδει κάθε μοντέλο παλινδρόμησης στα συγκεκριμένα δεδομένα.

	rMSE	MAE	R^2
Linear Regrassion	207.0496	160.3613	0.6523
Decision Tree	284.8260	220.6101	0.3421
Support Vector Regression	227.8507	165.9798	0.5790
Gaussian Process Regressor	249.1340	190.2862	0.4966
XGBOOST Regressor	208.3388	152.0668	0.6480
Multi-layer Perceptron regressor	225.4327	179.2784	0.5879
LightGBM regressor	208.6656	153.7848	0.6469
Random Forest Regressor	184.7891	135.8462	0.7231

Ο πίνακας που δημιουργείται στον κώδικα εμφανίζει μετρικές αξιολόγησης (rMSE, MAE και R^2) για διαφορετικά μοντέλα παλινδρόμησης. Ας κατανοήσουμε τι αντιπροσωπεύει κάθε μετρική:

- rMSE (Root Mean Squared Error):

Το rMSE είναι ένα μέτρο του μέσου μεγέθους των σφαλμάτων μεταξύ των προβλεπόμενων και των πραγματικών τιμών.Οι χαμηλότερες τιμές rMSE υποδηλώνουν καλύτερη απόδοση του μοντέλου.Εκφράζεται στις ίδιες μονάδες με τη μεταβλητή-στόχο.

- MAE (Μέσο απόλυτο σφάλμα):

Το MAE μετρά τις μέσες απόλυτες διαφορές μεταξύ προβλεπόμενων και πραγματικών τιμών.Όπως και το rMSE, οι χαμηλότερες τιμές MAE υποδηλώνουν καλύτερη απόδοση του μοντέλου.Εκφράζεται επίσης στις ίδιες μονάδες με τη μεταβλητή-στόχο.

- R^2 (Συντελεστής προσδιορισμού):

Το R^2 αντιπροσωπεύει την αναλογία της διακύμανσης της εξαρτημένης μεταβλητής (στόχος) που είναι προβλέψιμη από τις ανεξάρτητες μεταβλητές (χαρακτηριστικά).Οι τιμές R^2 κυμαίνονται από 0 έως 1, όπου το 1 υποδηλώνει τέλεια προσαρμογή.Υψηλότερες τιμές R^2 υποδηλώνουν καλύτερη απόδοση του μοντέλου.

Ερμηνεία των αποτελεσμάτων:

Γραμμική παλινδρόμηση: Οι χαμηλότερες τιμές rMSE και MAE, οι υψηλότερες τιμές R^2 υποδηλώνουν καλή απόδοση.

Δέντρο αποφάσεων: Αξιολογήστε με βάση το rMSE, το MAE και το R^2. Χαμηλότερες τιμές είναι επιθυμητές.

Υποστήριξη διανυσματικής παλινδρόμησης (SVR): Αξιολογήστε με βάση το rMSE, το MAE και το R^2. Οι χαμηλότερες τιμές είναι καλύτερες.

Gaussian Process Regressor (GPR): Εξετάστε τα rMSE, MAE και R^2. Οι χαμηλότερες τιμές είναι προτιμότερες.

XGBoost Regressor: Αξιολογήστε με βάση τα rMSE, MAE και R^2. Οι χαμηλότερες τιμές είναι καλύτερες.

Ρυθμιστής Perceptron πολλαπλών στρωμάτων (MLP): Αξιολογήστε με βάση τα rMSE, MAE και R^2. Χαμηλότερες τιμές είναι επιθυμητές.

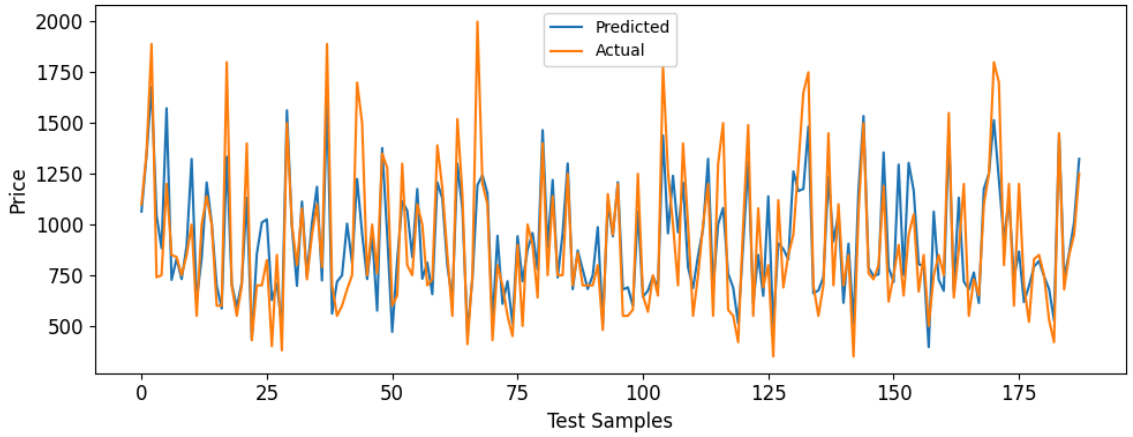
Ρυθμιστής LightGBM: Εξετάστε τα rMSE, MAE και R^2. Οι χαμηλότερες τιμές είναι προτιμότερες.

Ρυθμιστής Random Forest: Αξιολογήστε με βάση τα rMSE, MAE και R^2. Οι χαμηλότερες τιμές είναι καλύτερες.

Συνοπτικά, ο στόχος είναι να επιλεγεί ένα μοντέλο με χαμηλότερες τιμές rMSE και MAE και υψηλότερες τιμές R^2, που υποδηλώνουν καλύτερη ακρίβεια και προβλεπτική απόδοση.

```
from matplotlib import pyplot as plt
from matplotlib.pyplot import figure

figure(figsize=(10, 4))
plt.yticks(fontsize=12)
plt.xticks(fontsize=12)

plt.ylabel('Price', fontsize=12)
plt.xlabel('Test Samples', fontsize=12)

plt.plot(preds_rfg, label = 'Predicted')
plt.plot(y_test, label = 'Actual')
plt.legend()
plt.show()
```

Αυτός ο κώδικας παράγει ένα γραμμικό διάγραμμα όπου ο άξονας x αντιπροσωπεύει διαφορετικά δείγματα δοκιμών, ο άξονας y αντιπροσωπεύει τις τιμές των τιμών και απεικονίζονται δύο γραμμές για τις προβλεπόμενες και τις πραγματικές τιμές. Το υπόμνημα βοηθά στη διάκριση μεταξύ των δύο γραμμών στο διάγραμμα.



Το διάγραμμα απεικονίζει τη σύγκριση μεταξύ των προβλεπόμενων τιμών (preds_rfg) που παράγονται από έναν Random Forest Regressor και των πραγματικών τιμών (y_test) από το σύνολο δοκιμών. Ακολουθεί τι μπορείτε να ερμηνεύσετε από το διάγραμμα:

-Άξονας X ("Δείγματα δοκιμής"): Αντιπροσωπεύει μεμονωμένα δείγματα ή περιπτώσεις από το σύνολο δοκιμών.

-Άξονας Y ("Τιμή"): Αντιπροσωπεύει τις αντίστοιχες τιμές τιμής τόσο για τις προβλεπόμενες όσο και για τις πραγματικές τιμές.

-Γραμμή "Predicted" (Προβλεπόμενη τιμή): Η γραμμή αναπαριστά τις προβλεπόμενες τιμές που παράγονται από τον Random Forest Regressor για κάθε δείγμα δοκιμής.

-Γραμμή "Actual" (Πραγματικές): Η γραμμή αντιπροσωπεύει τις πραγματικές τιμές τιμών από το σύνολο δοκιμών.

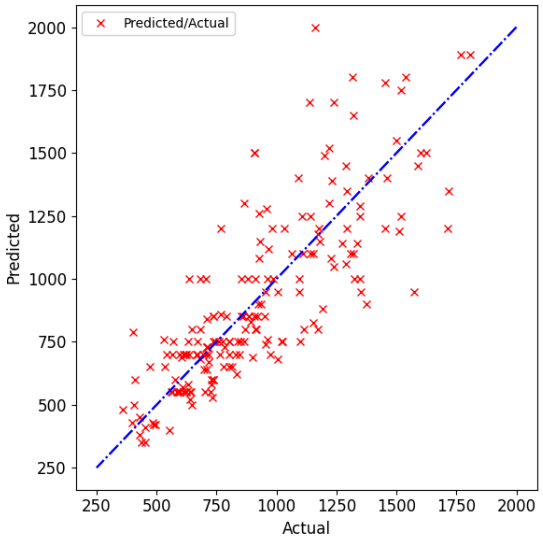
-Υπόμνημα: Το υπόμνημα υποδεικνύει ποια γραμμή αντιστοιχεί στις προβλεπόμενες τιμές ("Predicted") και ποια γραμμή αντιστοιχεί στις πραγματικές τιμές ("Actual").

Συγκρίνοντας τις γραμμές "Predicted" και "Actual",παρατηρούμε πως οι δύο γραμμές ακολουθούν στενά η μία την άλλη, γεγονός που υποδηλώνει ότι οι προβλέψεις του μοντέλου ευθυγραμμίζονται καλά με τις πραγματικές τιμές, υποδεικνύοντας καλή απόδοση.

Συνοπτικά, το διάγραμμα παρέχει μια οπτική επιθεώρηση του πόσο καλά οι προβλέψεις του Random Forest Regressor ταιριάζουν με τις πραγματικές τιμές για κάθε δείγμα δοκιμής. Βοηθά στην κατανόηση της απόδοσης του μοντέλου και στον εντοπισμό πιθανών περιοχών για βελτίωση.

```
# min = trainY1.HeartRate.min()
# max = trainY1.HeartRate.max()

p1 = trainY1.PRICE.max(), trainY1.PRICE.max()
p2 = trainY1.PRICE.min(), trainY1.PRICE.min()

figure(figsize=(6, 6))
plt.yticks(fontsize=12)
plt.xticks(fontsize=12)

plt.ylabel('Predicted', fontsize=12)
plt.xlabel('Actual', fontsize=12)

plt.plot(preds_LGBMR, y_test,'rx')

plt.plot([p1, p2], [p1, p2], 'b-.')
plt.legend(['Predicted/Actual'])
# To show the plot
plt.show()
```

Ο παρεχόμενος κώδικας δημιουργεί ένα διάγραμμα διασποράς για να συγκρίνει τις προβλεπόμενες τιμές (preds_LGBMR) με τις πραγματικές τιμές (y_test) χρησιμοποιώντας τον ρυθμιστή LightGBM. Το διάγραμμα διασποράς βοηθά στην οπτικοποίηση του πόσο καλά ευθυγραμμίζονται οι προβλέψεις με τις πραγματικές τιμές. Τα σημεία κατά μήκος της διαγώνιας γραμμής υποδεικνύουν ακριβείς προβλέψεις, ενώ οι αποκλίσεις από τη γραμμή αντιπροσωπεύουν σφάλματα πρόβλεψης.



Οι τιμές που απέχουν πολύ από τη διαγώνιο γραμμή, υποδηλώνουν σημαντική απόκλιση μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών.Κάτι τέτοιο μπορεί να ευθύνεται σε:

-Σφάλματα πρόβλεψης:

Σημεία πάνω από τη διαγώνια γραμμή υποδεικνύουν ότι το μοντέλο υπερεκτίμησε τις τιμές (οι προβλέψεις είναι υψηλότερες από τις πραγματικές).Σημεία κάτω από τη διαγώνια γραμμή υποδηλώνουν ότι το μοντέλο υποεκτίμησε τις τιμές (οι προβλέψεις είναι χαμηλότερες από τις πραγματικές).

-Ανακρίβεια μοντέλου:

Όσο πιο μακριά βρίσκονται τα σημεία από τη διαγώνια γραμμή, τόσο λιγότερο ακριβείς είναι οι προβλέψεις του μοντέλου.Οι μεγάλες αποκλίσεις μπορεί να υποδηλώνουν ότι το μοντέλο δεν καταγράφει ορισμένα πρότυπα ή χαρακτηριστικά των δεδομένων.

-Ακραίες τιμές ή ανωμαλίες:

Τα ακραία σημεία μακριά από τη γραμμή μπορεί να είναι ακραίες τιμές ή ανωμαλίες στα δεδομένα που το μοντέλο δεν χειρίστηκε καλά.

Αξιολόγηση του μοντέλου:

-Εάν η πλειονότητα των σημείων απέχει πολύ από τη διαγώνια γραμμή, αυτό μπορεί να υποδηλώνει ότι το μοντέλο χρειάζεται βελτίωση και η απόδοσή του μπορεί να μην είναι ικανοποιητική.

Ενδεχόμενο προσαρμογής μοντέλου:

-Θα μπορούσε να είναι μια ένδειξη ότι το μοντέλο χρειάζεται ρύθμιση, πρόσθετα χαρακτηριστικά ή διαφορετικό αλγόριθμο για να βελτιώσει την ακρίβειά του.

Παρατηρούμε πως δεν είναι πολλά τα σημεία που βρίσκονται μακριά απο την διαγώνιο μας , συνεπώς η εγγύτητα των υπολοίπων σημείων στη διαγώνιο γραμμή στο διάγραμμα διασποράς αποτελεί θετικό σημάδι, υποδεικνύοντας ότι οι προβλέψεις του μοντέλου ευθυγραμμίζονται καλά με τις πραγματικές τιμές.

```
# Create object that can calculate shap values
explainer = shap.TreeExplainer(regressor)
# Calculate Shap values
shap_values = explainer.shap_values(x_train)
```

Ο παρεχόμενος κώδικας περιλαμβάνει τη χρήση της βιβλιοθήκης SHAP (SHapley Additive exPlanations) για τον υπολογισμό των τιμών Shapley για ένα εκπαιδευμένο μοντέλο που βασίζεται σε δέντρα (regressor). Ακολουθεί μια ανάλυση του κώδικα:

Επεξήγηση:

-Create Explainer Object:

Αρχικοποιεί ένα αντικείμενο επεξηγητή SHAP ειδικά για μοντέλα που βασίζονται σε δέντρα (TreeExplainer). Αυτό το αντικείμενο χρησιμοποιείται για τον υπολογισμό των τιμών Shapley για το μοντέλο.

-Calculate Shap Values:

shap_values = explainer.shap_values(x_train): Υπολογίζει τις τιμές Shapley για τα δεδομένα εισόδου που παρέχονται (x_train). Οι τιμές Shapley αντιπροσωπεύουν τη συμβολή κάθε χαρακτηριστικού στη διαφορά μεταξύ της εξόδου του μοντέλου για μια συγκεκριμένη περίπτωση και της αναμενόμενης εξόδου του μοντέλου (μέση πρόβλεψη).

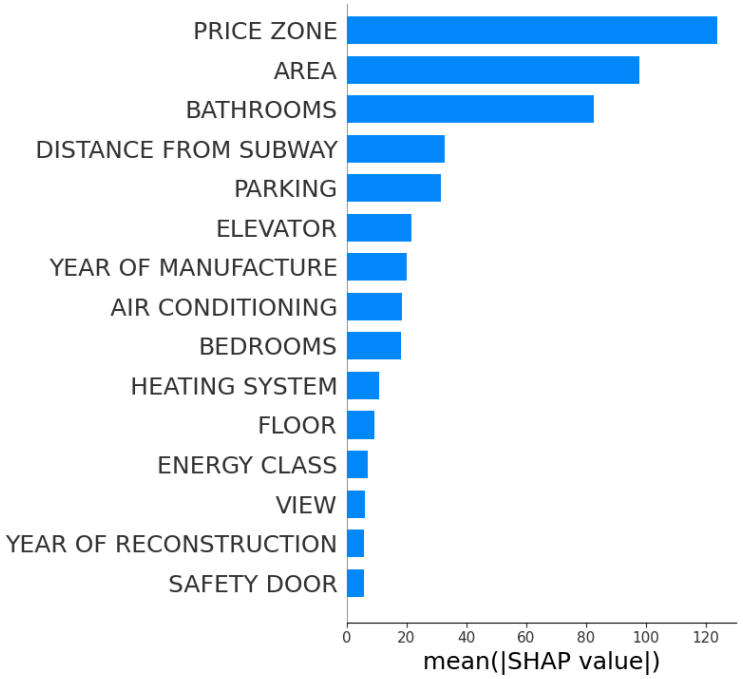
Σημείωση:

Οι τιμές Shapley βοηθούν στην εξήγηση της εξόδου ενός μοντέλου μηχανικής μάθησης αποδίδοντας την πρόβλεψη του μοντέλου σε μεμονωμένα χαρακτηριστικά. Οι θετικές τιμές Shapley υποδηλώνουν θετικό αντίκτυπο στην πρόβλεψη, ενώ οι αρνητικές τιμές υποδηλώνουν αρνητικό αντίκτυπο.Ο παλινδρομητής σε αυτή την περίπτωση θα πρέπει να είναι ένα ήδη εκπαιδευμένο μοντέλο βασισμένο σε δέντρα, όπως ένα δέντρο απόφασης, ένα τυχαίο δάσος ή ένα μοντέλο gradient boosting.Μετά τον υπολογισμό των τιμών Shap, μπορείτε να τις χρησιμοποιήσετε για διάφορους σκοπούς ερμηνευσιμότητας, όπως η κατανόηση της επιρροής των διαφόρων χαρακτηριστικών στις επιμέρους προβλέψεις ή η απόκτηση πληροφοριών σχετικά

με τη διαδικασία λήψης αποφάσεων του μοντέλου. Οι τιμές shap_values θα έχουν το ίδιο σχήμα με τα δεδομένα εισόδου x_train και κάθε γραμμή θα αντιστοιχεί στις τιμές Shapley για μια συγκεκριμένη περίπτωση στα δεδομένα εκπαίδευσης.

```
# SHAP Summary Plot
#shap.summary_plot(shap_values, x_train)
fig = shap.summary_plot(shap_values, x_train, feature_names=trainX1.keys(), max_display=15, plot_type="bar", show=False)
plt.yticks(fontsize=18)
plt.xlabel('mean(|SHAP value|)', fontsize=18)
# plt.savefig(path_to_figs+ modelID +'_shap_'+ModelType+'.eps', format='eps')
# plt.savefig(path_to_figs+ modelID +'_shap_'+ModelType+'.svg', format='svg')
#shap.decision_plot(explainer.expected_value[0], shap_values[0])
```

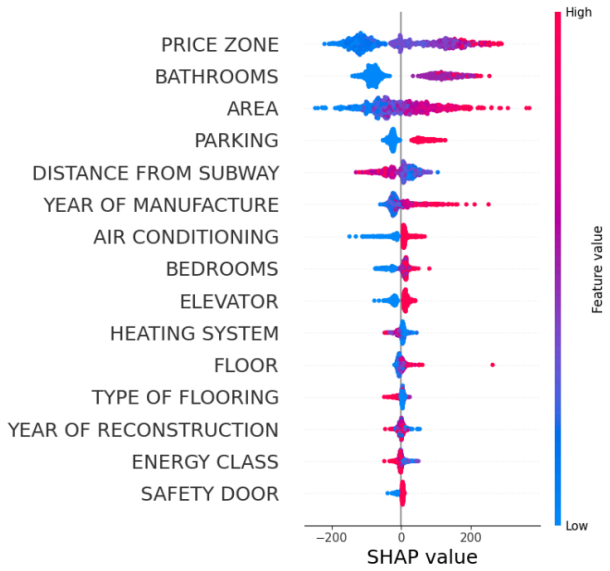
Ο παρεχόμενος κώδικας παράγει ένα συνοπτικό διάγραμμα SHAP (SHapley Additive exPlanations) χρησιμοποιώντας τη βιβλιοθήκη SHAP.



Το συνοπτικό διάγραμμα SHAP παρέχει πληροφορίες σχετικά με τον αντίκτυπο κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου. Τα χαρακτηριστικά με υψηλότερες απόλυτες μέσες τιμές Shapley έχουν σημαντικότερη επίδραση στην παραγωγή του μοντέλου.Παρατηρούμε πως στην δική μας περίπτωση οι πέντε τιμές που επηρεάζουν την τιμή ενοικίασης είναι αδιαμφισβήτητα η τιμή ζώνης,η επιφάνεια και τα μπάνια μα επίσης και συγκοινωνιακά χαρακτηριστικά , όπως η απόσταση από μετρό και η παρεχόμενη θέση στάθμευσης .

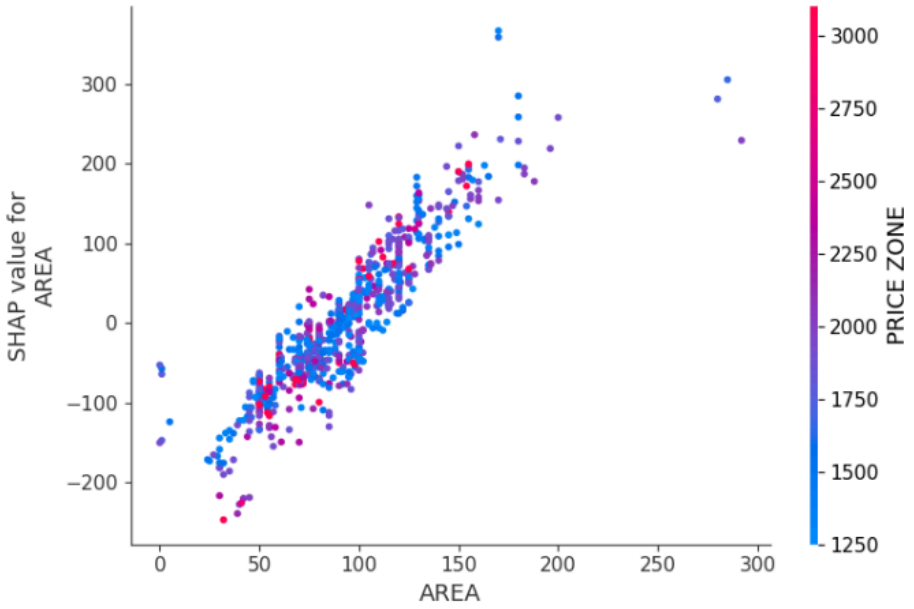
```
[ ] import matplotlib.pyplot as plt
    f = plt.figure()
    fig = shap.summary_plot(shap_values, x_train,feature_names=trainX1.keys(), max_display=15, show=False)
    plt.yticks(fontsize=18)
    #plt.xticks(fontsize=18)
    plt.xlabel('SHAP value', fontsize=18)
    # plt.savefig('_impact.eps', format='eps')
    # plt.savefig('_impact.svg', format='svg')
```

Αυτό το απόσπασμα κώδικα υποδηλώνει ότι δημιουργούμε ένα συνοπτικό διάγραμμα SHAP, ρυθμίζετε την εμφάνισή του και προαιρετικά το αποθηκεύουμε ως αρχείο εικόνας. Το συνοπτικό διάγραμμα SHAP απεικονίζει τον αντίκτυπο κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου, βοηθώντας στην ερμηνεία της συμπεριφοράς του μοντέλου και της σημασίας του χαρακτηριστικού.




```
fig = shap.dependence_plot("AREA",shap_values, x_train, feature_names=trainX1.keys(), show=False, interaction_index='PRICE ZONE')
fig = shap.dependence_plot("PRICE ZONE",shap_values, x_train, feature_names=trainX1.keys(), show=False, interaction_index='AREA')
fig = shap.dependence_plot("PARKING",shap_values, x_train, feature_names=trainX1.keys(), show=False, interaction_index='DISTANCE FROM SUBWAY')
fig = shap.dependence_plot("BATHROOMS",shap_values, x_train, feature_names=trainX1.keys(), show=False, interaction_index='PRICE ZONE')
fig = shap.dependence_plot("DISTANCE FROM SUBWAY",shap_values, x_train, feature_names=trainX1.keys(), show=False, interaction_index='PRICE ZONE')
```

Ο παρεχόμενος κώδικας παράγει πολλαπλά διαγράμματα εξάρτησης SHAP, εξετάζοντας τη σχέση μεταξύ συγκεκριμένων χαρακτηριστικών και της εξόδου του μοντέλου. Κάθε κλήση shap.dependence_plot επικεντρώνεται σε ένα διαφορετικό χαρακτηριστικό και εξετάζει την αλληλεπίδρασή του με ένα άλλο καθορισμένο χαρακτηριστικό.Κάθε ένα από αυτά τα διαγράμματα παρέχει πληροφορίες για το πώς τα καθορισμένα χαρακτηριστικά επηρεάζουν τις προβλέψεις του μοντέλου και πώς οι αλληλεπιδράσεις τους επηρεάζουν την έξοδο. Ο δείκτης αλληλεπίδρασης βοηθά στην οπτικοποίηση της εξάρτησης λαμβάνοντας υπόψη τη συνδυασμένη επίδραση δύο χαρακτηριστικών



Κεφάλαιο 5ο-Σχεδιασμός και Υλοποίηση

5.1 Πειραματική διαχείριση

Στα πλαίσια του πρακτικού μέρους της διπλωματικής μας εργασίας και των αντίστοιχων μοντέλων-παραδειγμάτων χρησιμοποιήσαμε δεδομένα από την πολύ γνωστή πλατφόρμα εύρεσης ακινήτων SPITOGATOS. Συγκεκριμένα, χρησιμοποιήσαμε το σύνολο δεδομένων , το οποίο βρίσκεται στην διεύθυνση https://www.spitogatos.gr/enoikiaseis-katoikies και τα δεδομένα που περιέχει έχουν να κάνουν με ποικίλες πληροφορίες για 1000 σπίτια στην πρωτεύουσα της Ελλάδας, Αθήνα αλλά και στην ευρύτερη περιοχή της. Επιπρόσθετα, οι πληροφορίες αυτές κατανέμονται δηλαδή εκφράζονται από τις 26 στήλες - μεταβλητές που έχει το πλαίσιο δεδομένων. Αυτές οι μεταβλητές προσδιορίζουν συγκεκριμένα χαρακτηριστικά των σπιτιών και είναι οι εξής:

Price zone: η τιμή ζώνης της ευρύτερης περιοχής

Price: η τιμή του σπιτιού

Price per sq.m: η τιμή του σπιτιού ανά τ.μ

Area: τα τετραγωνικά μέτρα του σπιτιού

Levels: τα επίπεδα του σπιτιού

Floor: ο όροφος του σπιτιού

Heating system: το σύστημα θέρμανσης του κάθε σπιτιού

Air conditioning : η ύπαρξη τοπικού αεροψύκτη

Energy class: η ενεργειακή κλάση του κάθε σπιτιού

Elevator: η παρουσία ανελκυστήρα

Distance from Subway: η απόσταση από μετρό από το κάθε σπίτι

Type of flooring: ο τύπος δαπέδου κάθε σπιτιού

Pets : κατοικία ευπρόσδεκτα

Safety door : πόρτα ασφαλείας του σπιτιού

Frames : κουφώματα του σπιτιού

Sieves : σίτες του σπιτιού

Transparent : διαμπερότητα του σπιτιού

Painted : προσφάτως βαμμένοι τοίχοι του σπιτιού

View : θέα του σπιτιού

Kitchens: ο αριθμός κουζινών του σπιτιού

Bedrooms: ο αριθμός των υπνοδωματίων του σπιτιού

Bathrooms: ο αριθμός των μπάνιων του κάθε σπιτιού

Parking: ο αριθμός των διαθέσιμων χώρων στάθμευσης

Year of manufacture: η χρονιά που χτίστηκε το σπίτι

Year of reconstruction: η χρονιά που ανακαινίστηκε το σπίτι

Το πρόβλημα με το οποίο ασχολούμαστε στην εργασία πάνω στα δεδομένα που παρουσιάστηκαν προηγουμένως είναι η δημιουργία μοντέλων πρόβλεψης Μηχανικής Μάθησης με σκοπό την πρόβλεψη της τιμής ενοικίασης των σπιτιών από το σύνολο δεδομένων. Η μεταβλητή πρόβλεψης που χρησιμοποιείται είναι το χαρακτηριστικό Price, ενώ όλες οι υπόλοιπες μεταβλητές χρησιμοποιούνται στην δημιουργία, την εκπαίδευση και τον έλεγχο των μοντέλων που κατασκευάζονται

5.2 Περιγραφή δεδομένων

Στο κομμάτι αυτό της προεπεξεργασίας χρειάστηκε να εφαρμοστούν διαφορετικές μέθοδοι ανά χαρακτηριστικό, καθώς το κάθε ένα είχε διαφορετική μορφή τιμών. Αρχικά, θα δοθεί έμφαση στην δημιουργία αντιπροσωπευτικής αναπαράστασης του κάθε χαρακτηριστικού σε αριθμητικές τιμές. . Όπως έχει προαναφερθεί, το σύνολο δεδομένων έχει συλλεχθεί από το διαδίκτυο και συνεπώς οι τιμές των χαρακτηριστικών δεν είναι σε κατάλληλη μορφή για να τροφοδοτηθούν στα μοντέλα μηχανικής μάθησης. Αφότου έχει ολοκληρωθεί το στάδιο της απομάκρυνσης χαρακτηριστικών που δεν θα χρειαστούν στην εκπαίδευση των μοντέλων, πρέπει τώρα όλα τα εναπομείναντα χαρακτηριστικά να έρθουν στην κατάλληλη μορφή. Αυτό σημαίνει πως όλα τα δεδομένα πρέπει να είναι αριθμοί που να μην περιέχουν σύμβολα, χαρακτήρες και επιπρόσθετα οι αριθμοί αυτή να μεταφέρουν όσο πιο πιστά γίνεται την φυσική σημασία του κάθε χαρακτηριστικού. Πρέπει δηλαδή, για το κάθε χαρακτηριστικό, να δημιουργηθεί μια αντιπροσωπευτική αναπαράσταση. Ακόμα, παράλληλα με αυτή την διαδικασία θα εφαρμόζεται και συμπλήρωση των εκλιπόντων τιμών όπου αυτές εμφανίζονται, με σεβασμό πάντα στην φυσιολογία του εκάστοτε χαρακτηριστικού.

Παρακάτω θα παρατηρήσουμε κάποιες στατιστικές τιμές για τα δεδομένα που τελικα χρησιμοποιήσαμε.

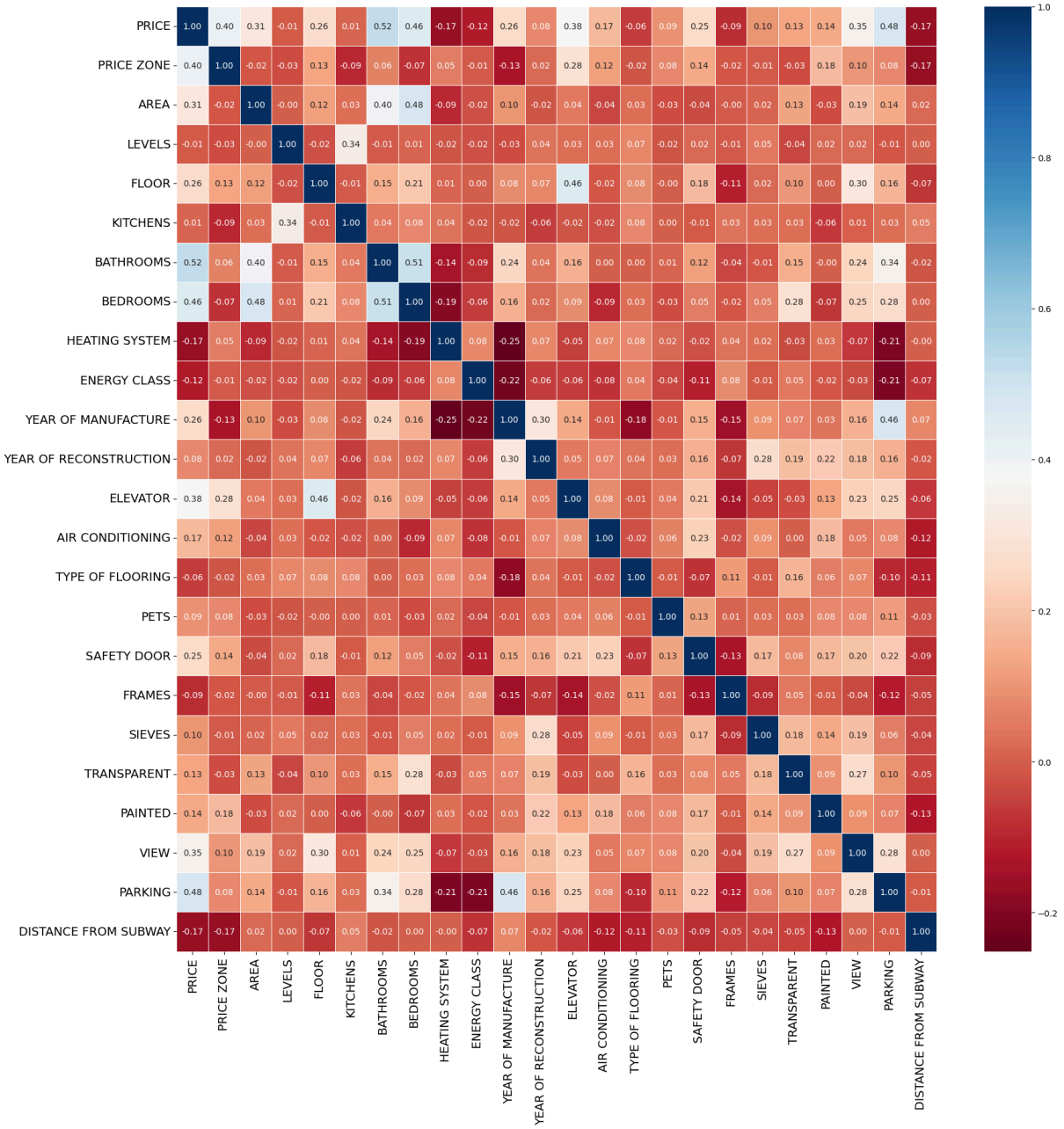
	PRICE ZONE	PRICE	PRICE PER sq.m	AREA	LEVELS	FLOOR	KITCHENS
count	938	938	938	938	938	938	938
mean	1852.985075	899.7270789	10.10998934	93.01705757	1.156183369	1.997334755	1.004264392
std	613.3055475	365.0994165	3.835413219	35.10559961	5.869659442	1.580545863	0.1032189621
min	1050	250	0.13	0	0	-1	-1
25%	1450	640	7.42	70	1	1	1
50%	1800	800	8.97	90	1	2	1
75%	2000	1115	12.095	115	1	3	1
max	4250	2000	27.8	292	180	9	2

BATHROOMS	BEDROOMS	HEATING SYSTEM	ENERGY CLASS	YEAR OF MANUFAC	YEAR OF RECONST	ELEVATOR	AIR CONDITIONING
938	938	938	938	938	938	938	938
1.47761194	2.103411514	1.301705757	6.198294243	1986.803838	2001.637527	0.6140724947	0.6737739872
0.6517604849	0.7981897095	1.657947861	2.885417144	16.1285493	19.12060132	0.4870732876	0.4690812111
1	0	0	0	1901	1901	0	0
1	2	0	4	1975	1985	0	0
1	2	1	6	1982.5	2005	1	1
2	3	2	9	2000	2021	1	1
5	5	10	33	2023	2023	1	1

TYPE OF FLOORIN	PETS	SAFETY DOOR	FRAMES	SIEVES	TRANSPARENT	PAINTED	VIEW	PARKING	DISTANCE FROM S
938	938	938	938	938	938	938	938	938	938
1.144989339	0.3081023454	0.7846481876	0.0618336887	0.3038379531	0.5959488273	0.5948827292	0.5586353945	0.315565032	1879.733475
1.606510716	0.4619554075	0.4164429004	0.2929510653	0.4601588786	0.4909692519	0.4911766165	0.7136981516	0.4649884387	963.7219036
0	0	0	0	0	0	0	0	0	10
0	0	1	0	0	0	0	0	0	1200
0	0	1	0	0	1	1	0	0	1800
2	1	1	0	1	1	1	1	1	2500
5	1	2	3	1	1	1	4	1	6000

Ισχύει πως η μέση τιμή των δεδομένων μας , μιλά για ένα ακίνητο επιφάνειας Ε=93,02 τ.μ με την τιμή ενοικίασης να ανέρχεται στα 899,72 € και με τις παρακάτω παροχές:

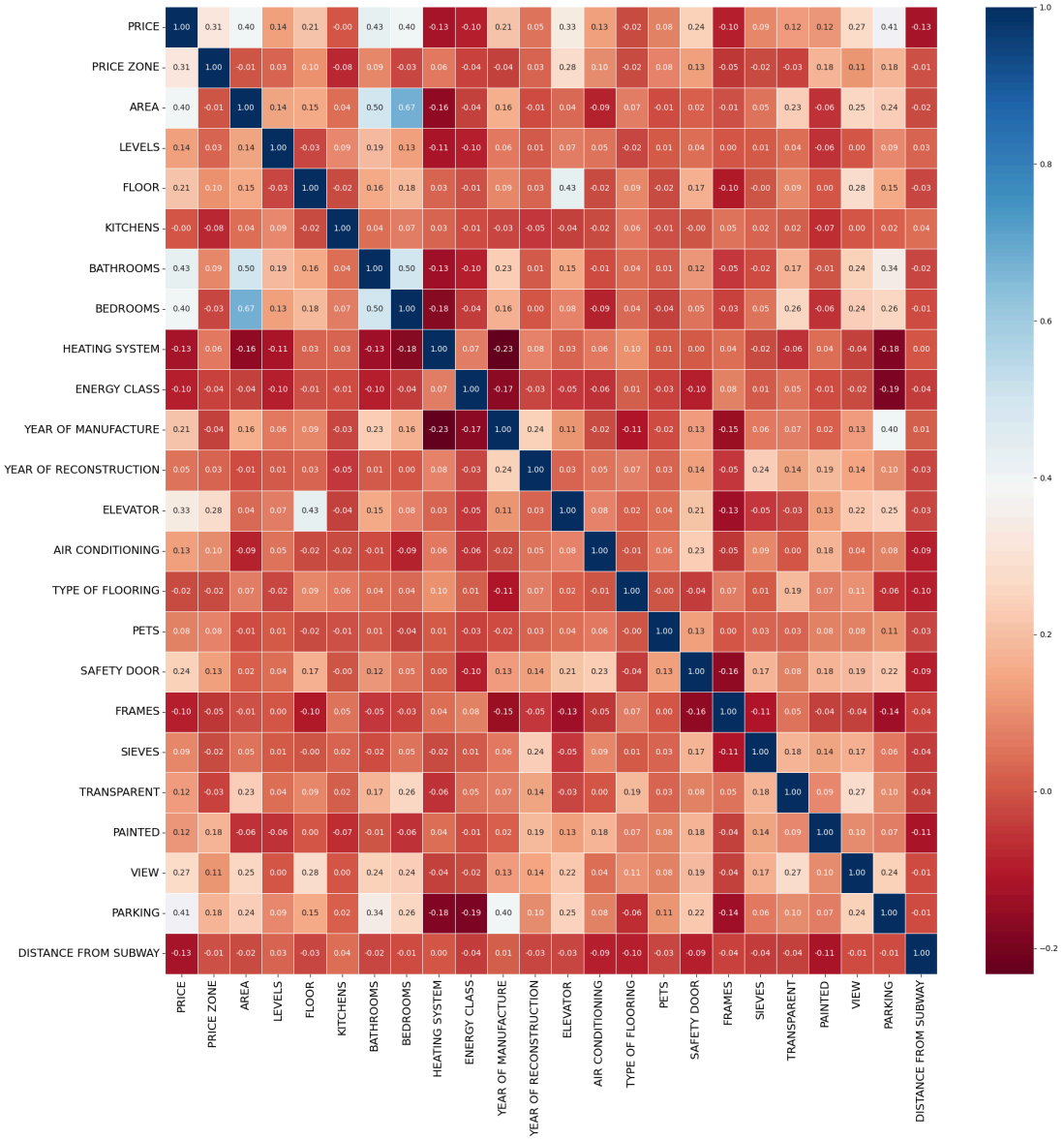
- 1 κουζίνα ,
- παραπάνω από 1 λουτρό,
- παραπάνω από 2 υπνοδωμάτια,
- θέρμανση πετρελαίου
- έτος κατασκευής 1986 και έπειτα
- χωρίς ανελκυστήρα
- χωρίς κλιματισμό
- χωρίς κατοικίδια
- χωρίς θέση στάθμευσης
- με απόσταση απο το μετρό γύρω στα 1879 μ.



Πίνακας 2: Συντελεστής pearson

Παρατηρούμε πως οι παροχές που επηρεάζουν θετικά την τιμή ενοικίασης σύμφωνα με τον *Pearson* είναι :

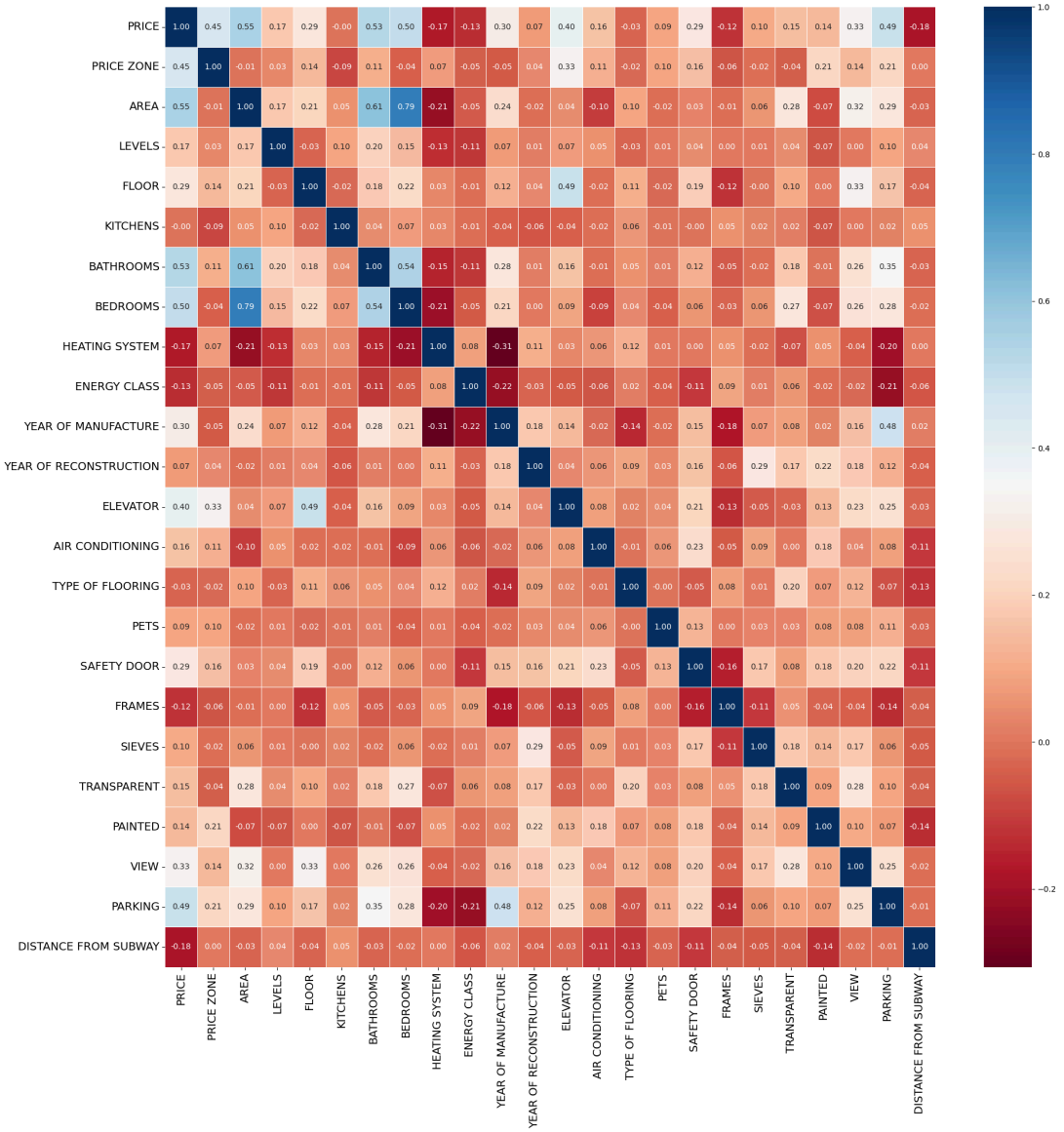
- ☐ μπάνια
- ☐ τιμή ζώνης
- ☐ υπνοδωμάτια
- ☐ ανελκυστήρας
- ☐ θέα
- ☐ θέση στάθμευσης



Πίνακας 4: Συντελεστής kendall

Εδώ βλέπουμε εναν κώδικα παρόμοιος με τον προηγούμενο, αλλά αντί να χρησιμοποιεί τη συσχέτιση Pearson, υπολογίζει τον συντελεστή συσχέτισης Kendall .Συνεπώς ισχύει, πως οι παροχές που επηρεάζουν θετικά την τιμή ενοικίασης σύμφωνα με τον *Kendall* είναι επιπλέον :

- ☐ επιφάνεια



Πίνακας 5: Συντελεστής spearman

Τέλος, ένας ακόμα κώδικας παρόμοιος αλλα με μικρές διαφορές από τους δύο προηγούμενους είναι ο spearman,ο οποίος όπως παρατηρούμε έχει τις περισσότερες τιμές που επηρεάζουν θετικά το μοντέλο μας.

Κανένας από τους παραπάνω κώδικες να σημειωθεί ότι δεν έχει πολύ αρνητικές τιμές ,συνεπώς εκτός της απόστασης απο το μετρό όλα τα υπόλοιπα δεν επηρεάζουν αρνητικά.

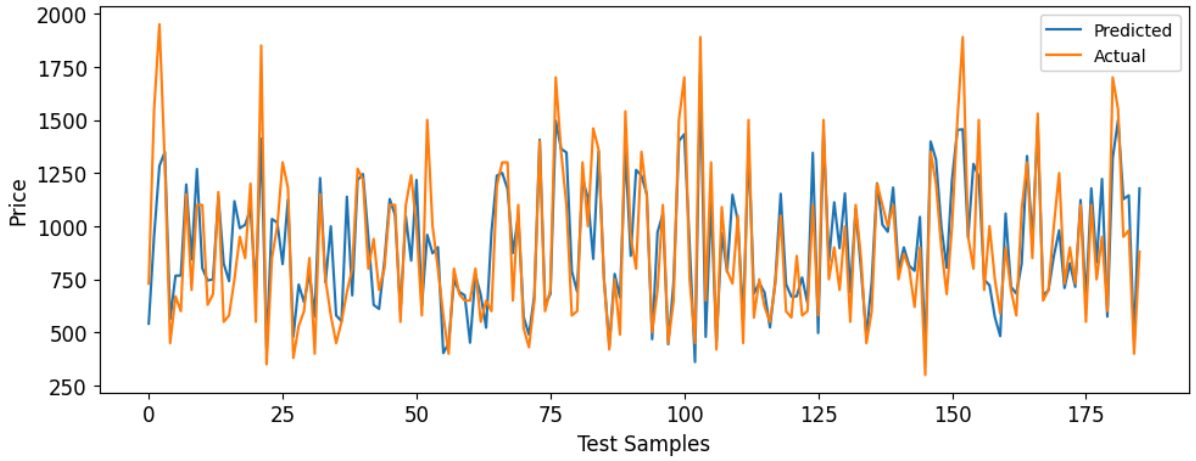
5.3 Αξιολόγηση αποτελεσμάτων

Συνοπτικά, ο στόχος είναι να επιλεγεί ένα μοντέλο το οποίο όπως είδαμε και στο κεφάλαιο 4.3, θα συνδυάζει όσο το δυνατόν καλύτερα τις εξής ιδιότητες:

- Χαμηλότερη τιμή rMSE
- Χαμηλότερη τιμή MAE
- Υψηλότερη τιμή R^2

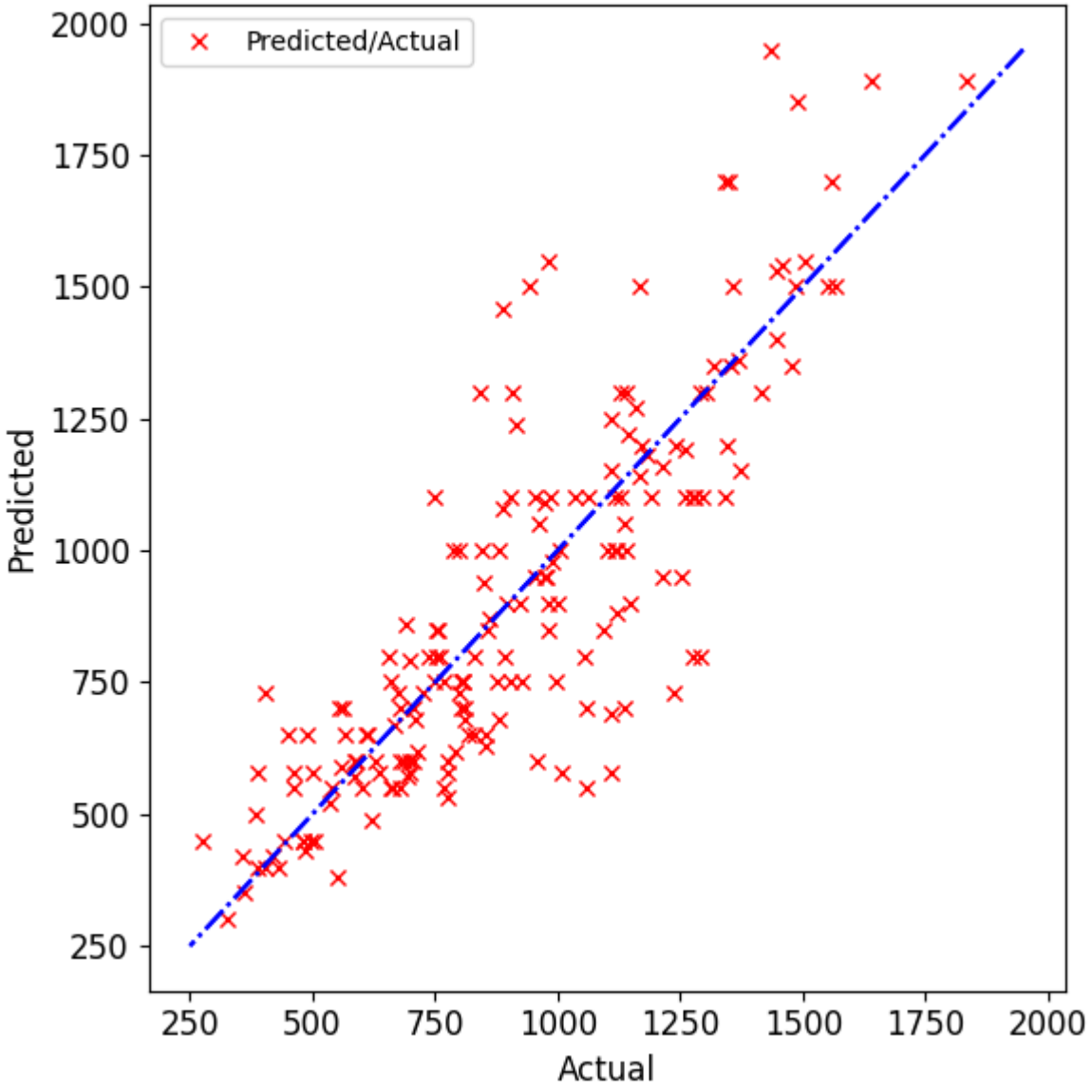
Παρατηρούμε λοιπόν, πως στην προκειμένη περίπτωση , τα αριθμητικά αποτελέσματα που λαμβάνουμε βρίσκουν στην ιδανικότερη θέση **το Random Forest Regressor (RFR) ,**συνεπώς αυτό είναι το μοντέλο που αποδίδει καλύτερα στο σύνολο δεδομένων μας

	rMSE	MAE	R^2
Linear Regrassion	209.6780	155.9988	0.6450
Decision Tree	259.9789	202.9718	0.4542
Support Vector Regression	225.2458	165.0180	0.5903
Gaussian Process Regressor	268.4540	207.6750	0.4181
XGBOOST Regressor	202.8234	143.7528	0.6678
Multi-layer Perceptron regressor	223.4886	172.3025	0.5967
LightGBM regressor	192.6328	138.8266	0.7004
Random Forest Regressor	185.9699	133.3793	0.7207



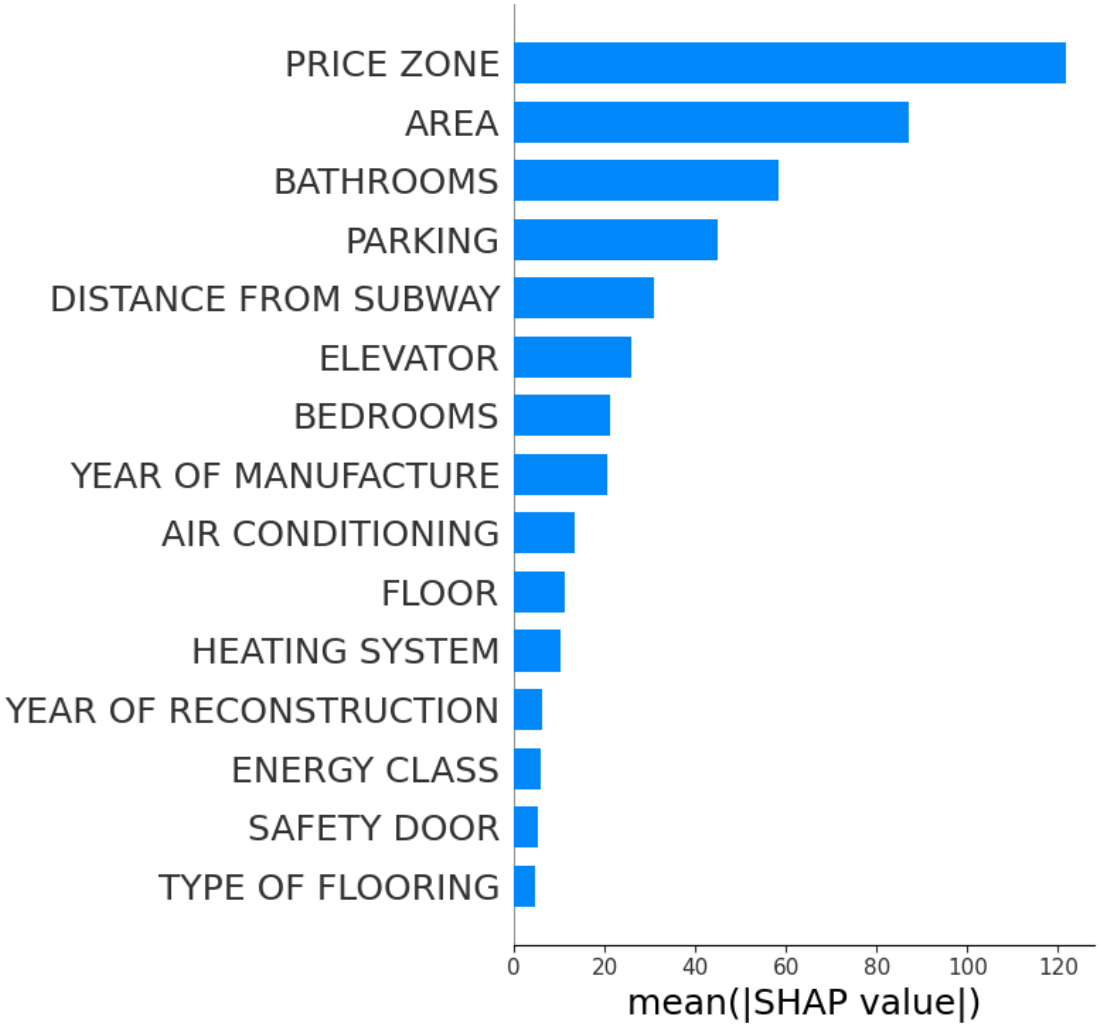
Συγκρίνοντας τις γραμμές "Predicted" και "Actual", παρατηρούμε πως οι δύο γραμμές ακολουθούν στενά η μία την άλλη, γεγονός που υποδηλώνει ότι οι προβλέψεις του μοντέλου ευθυγραμμίζονται καλά με τις πραγματικές τιμές, υποδεικνύοντας καλή απόδοση.

Συνοπτικά, το διάγραμμα παρέχει μια οπτική επιθεώρηση του πόσο καλά οι προβλέψεις του Random Forest Regressor ταιριάζουν με τις πραγματικές τιμές για κάθε δείγμα δοκιμής. Βοηθά στην κατανόηση της απόδοσης του μοντέλου και στον εντοπισμό πιθανών περιοχών για βελτίωση.

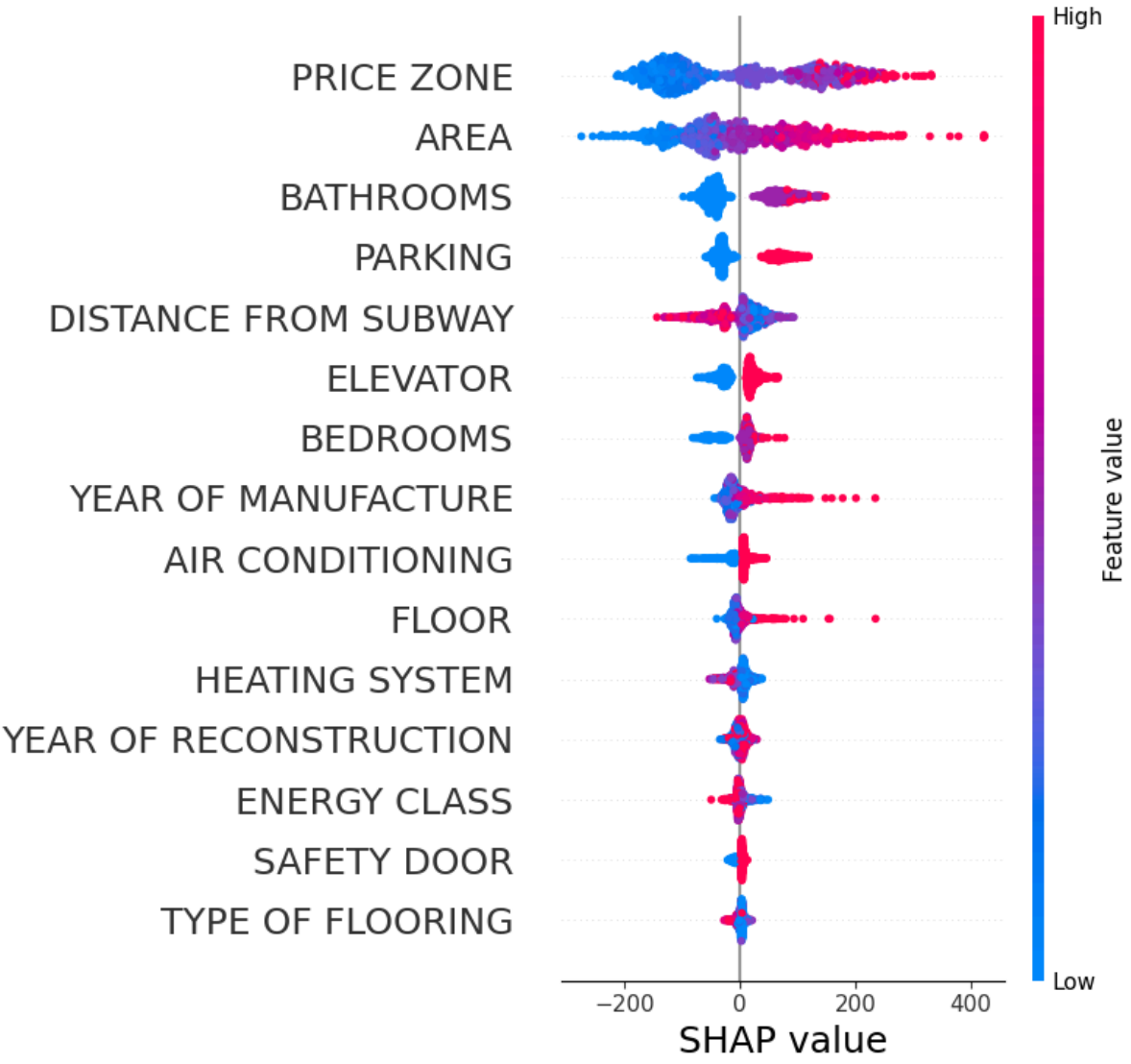


Παρατηρούμε πως δεν είναι πολλά τα σημεία που βρίσκονται μακριά από την διαγώνιο μας , συνεπώς η εγγύτητα των υπολοίπων σημείων στη διαγώνιο γραμμή στο διάγραμμα διασποράς αποτελεί θετικό σημάδι, υποδεικνύοντας ότι οι προβλέψεις του μοντέλου ευθυγραμμίζονται καλά με τις πραγματικές τιμές.

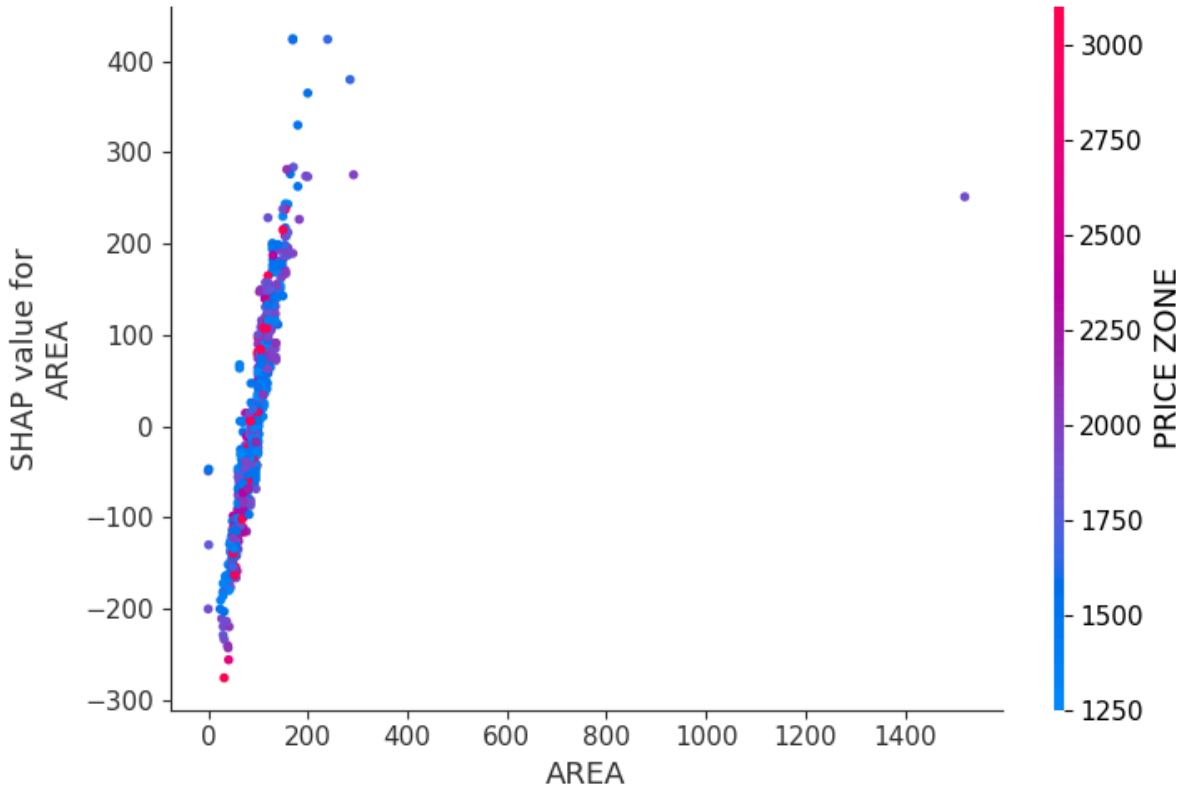
5.4 Explainable AI



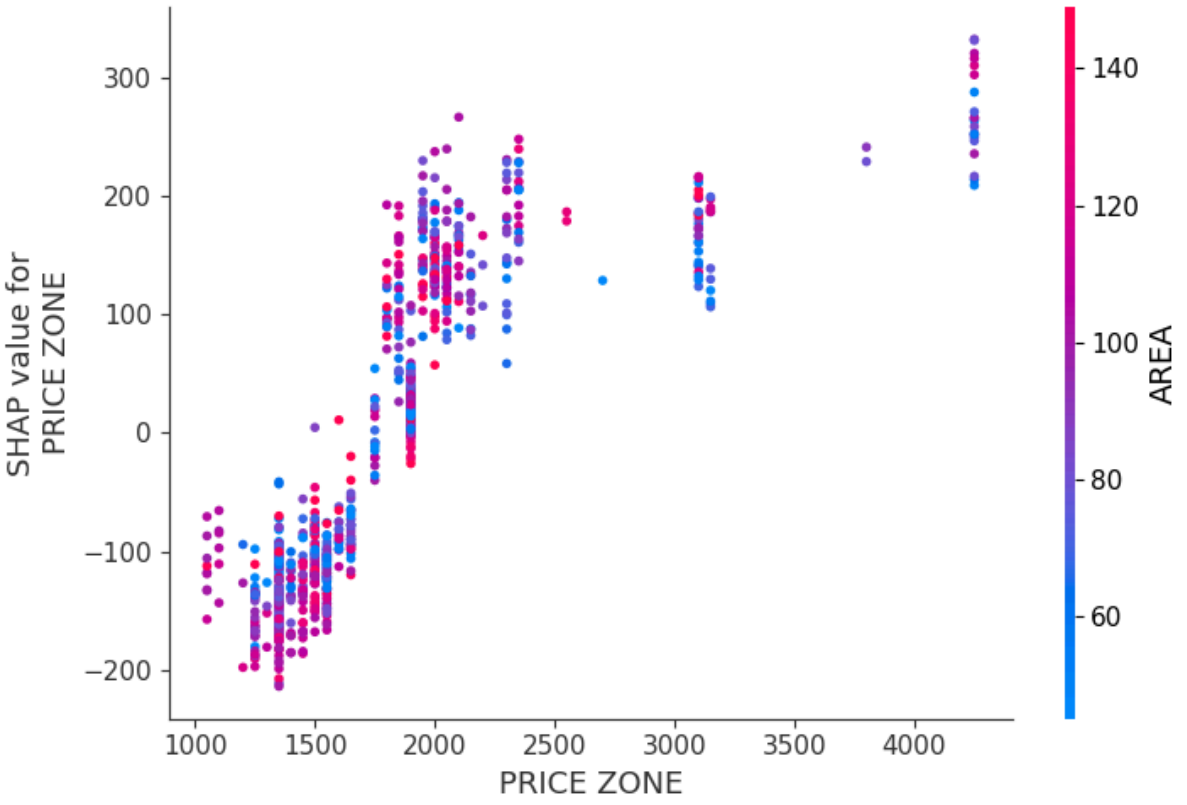
Η σημασία του χαρακτηριστικού SHAP μετράται ως η μέση απόλυτη τιμή των τιμών Shapley όπως γνωρίζουμε από το Κεφάλαιο 4.3 .Συνεπώς αυτό που παρατηρούμε στα δικά μας αποτελέσματα είναι πως η τιμή ζώνης αποτελεί το σημαντικότερο χαρακτηριστικό με διαφορά από τα επόμενα κάτι που θεωρούμε αυτονόητο.Ακολουθεί βεβαίως η επιφάνεια του ακινήτου, η οποία σύμφωνα με την λογική δικαιωματικά βρίσκεται στην δεύτερη θέση σημαντικότητας.Συνεχίζουμε με τον αριθμό των λουτρών , που συνδέονται με την εξυπηρέτηση των μελών του νοικοκυριού μα και την υγιεινή αυτών .Οι επόμενες δύο κατηγορίες είναι και κατά την άποψή μου οι πιο σημαντικές, αφού στην Ελλάδα του 2024 η συγκοινωνιακή πολιτική κυριαρχεί.Παρατηρούμε πως στην θέση 4 βρίσκεται η θέση στάθμευσης και στην θέση 5 η απόσταση από σταθμό μετρο , γεγονός που δείχνει πως πολίτης του σήμερα έχει ανάγκη να διευκολύνει την καθημερινότητα του όσον αφορά την μεταφορά του από τόπο σε τόπο.Συνεπώς, είναι γεγονός πως το να μπορεί ο πολίτης να μετακινηθεί από και προς την οικεία του με ασφάλεια αυτού και της περιουσίας ,του γρήγορα και αποτελεσματικά είναι κάτι που τον επηρεάζει ώστε να επιλέξει την τοποθεσία της διαμονής του.

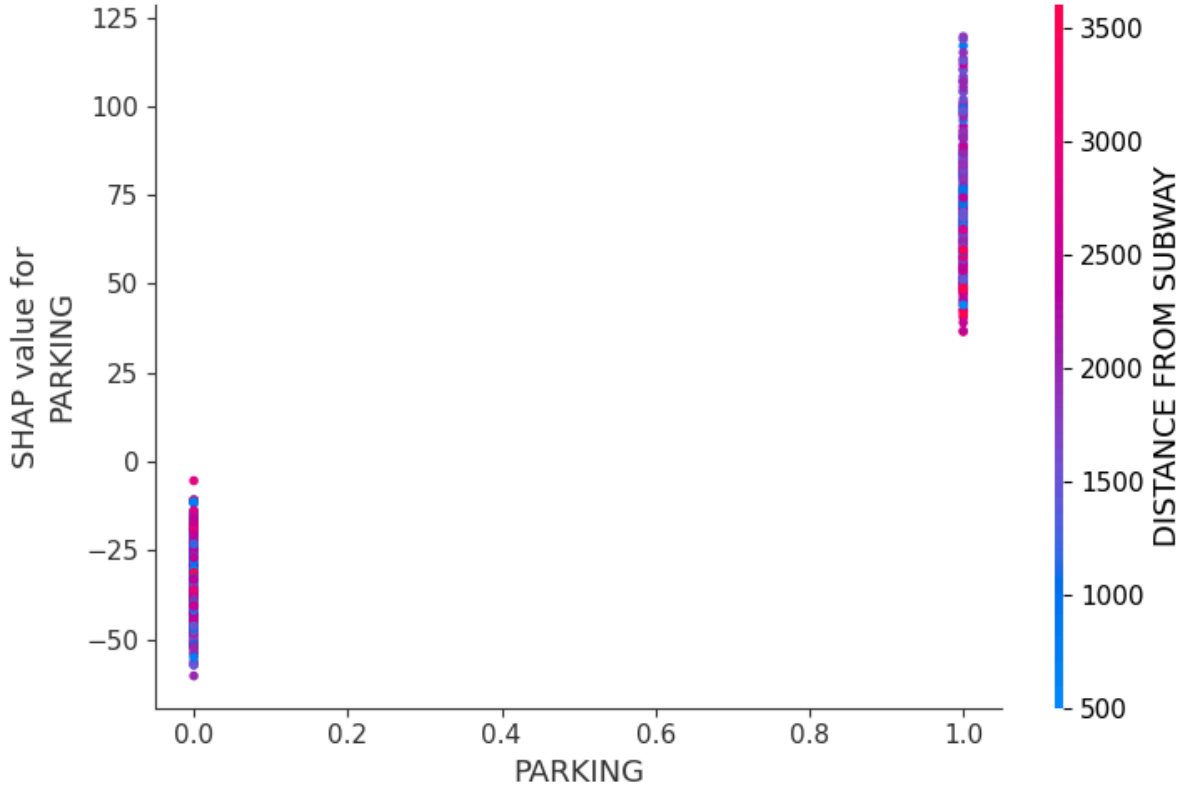


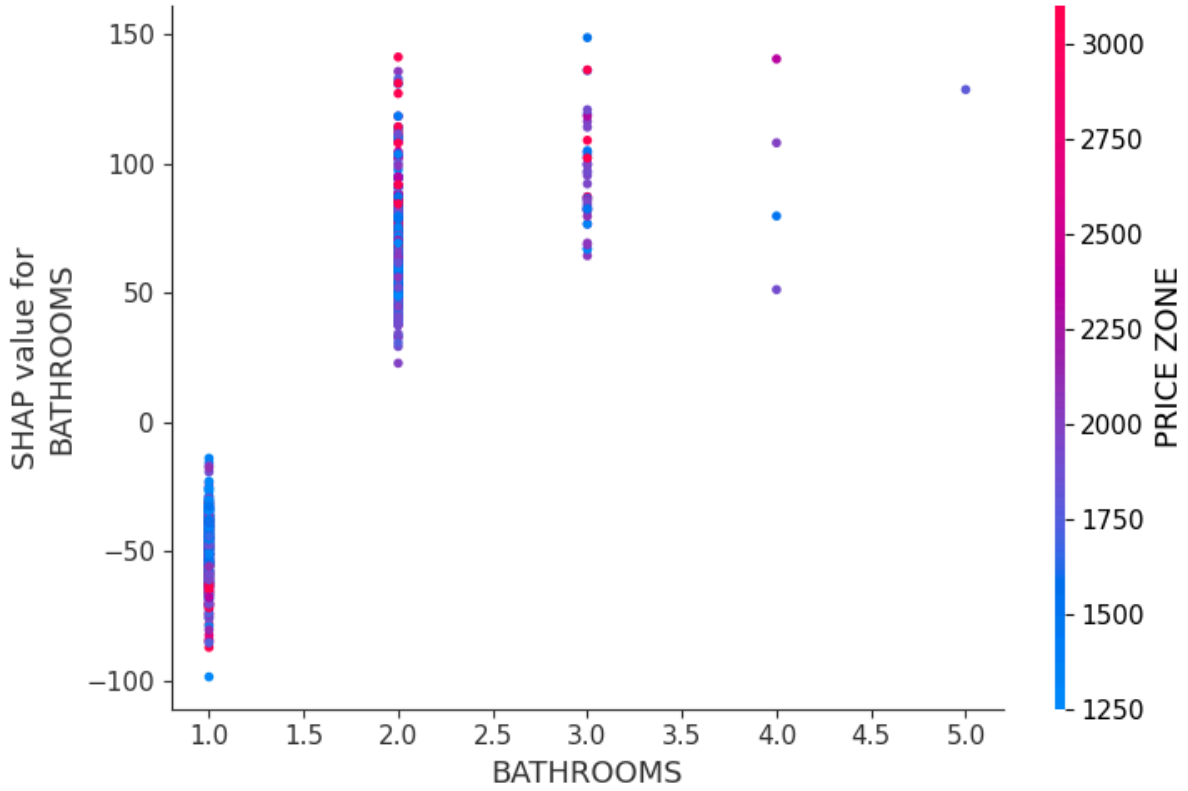
Το συνοπτικό διάγραμμα SHAP απεικονίζει τον αντίκτυπο κάθε χαρακτηριστικού στις προβλέψεις του μοντέλου, βοηθώντας στην ερμηνεία της συμπεριφοράς του μοντέλου και της σημασίας του χαρακτηριστικού. Στο προκείμενο διάγραμμα παρατηρούμε κυρίως, πως το να μην διαθέτει ένα ακίνητο ανελκυστήρα,μπάνιο,θέση στάθμευσης ή κλιματισμό μειώνει κατα πολύ την αξία του χαρακτηριστικού, δηλαδή της τιμής ενοικίασης .



Παρατηρούμε πως όσο αυξάνεται η επιφάνεια αυξάνεται και η αξία του χαρακτηριστικού με τις τιμές να κυμαίνονται κατάλληλα .

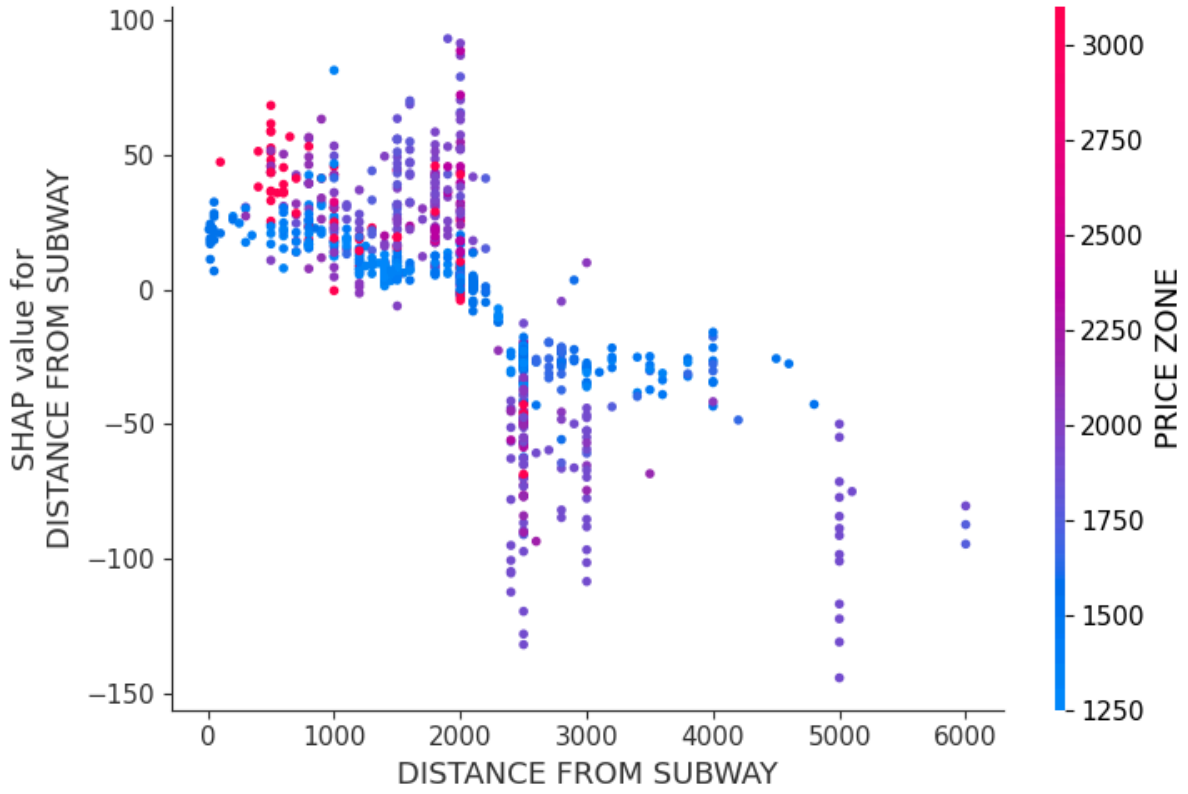




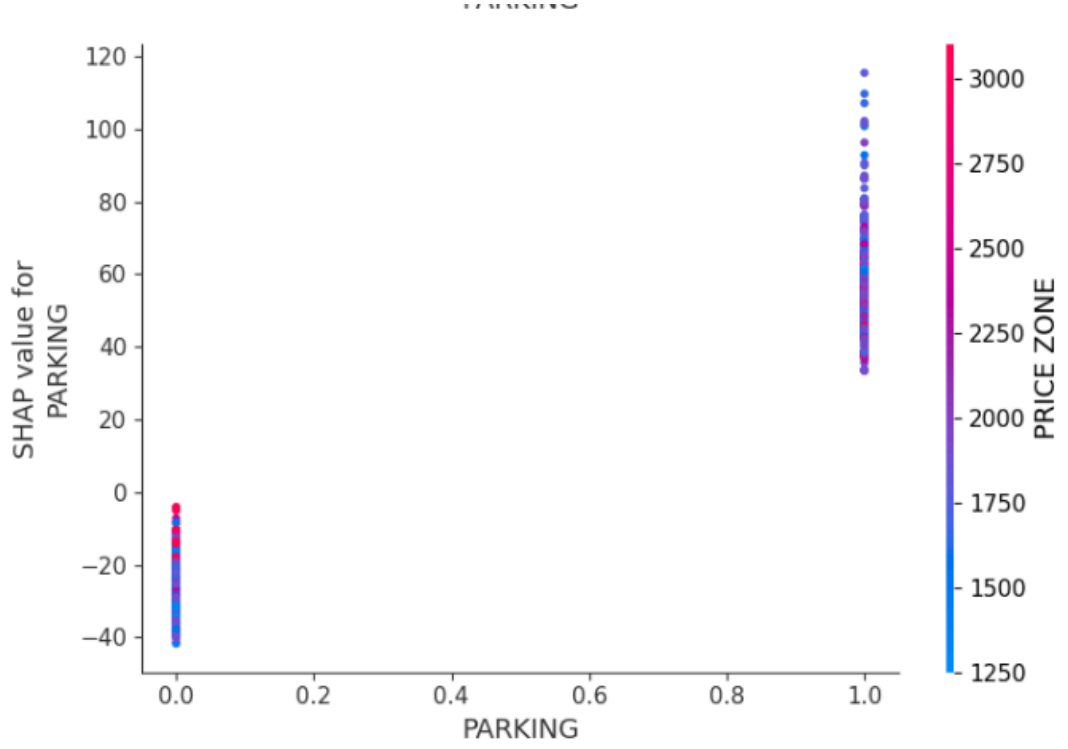


Το διάγραμμα μας δείχνει πως συνηθίζεται η παροχή έως 3 λουτρών.Μεγαλύτερη αξία παρατηρούμε να έχει η παρουσία δύο λουτρών οπου εκεί κατανέμεται καλυτερα το εύρος των

τιμών.Αρκετά παρατηρούμε να είναι και τα ακίνητα που έχουν 1 μπάνιο , με τις τιμές να έχουν και εκεί ποικιλία μα σε αυτή την περίπτωση δεν έχει αρκετή αξία το χαρακτηριστικό .



Παραπάνω βλέπουμε πως όσο αυξάνεται η απόσταση από το μετρό , το χαρακτηριστικό αυτό χάνει την αξία του και σε συνδυασμό έχουμε πτώση της τιμής ενοικίασης.



Εδώ παρατηρούμε πως η ύπαρξη θέση στάθμευσης έχει μεγαλύτερη αξία σαν χαρακτηριστικό καθώς επίσης οι τιμές ενοικίασης ποικίλουν , δεν είναι μόνο υψηλες.

5.5 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της μοντελοποίησης των τιμών των ακινήτων. Ένα πρόβλημα εξέχουσας σημασίας για ένα σεβαστά μεγάλο τμήμα του πληθυσμού, από την στιγμή που η αγορά των ακινήτων επηρεάζει άμεσα την οικονομία μας. Στην συγκεκριμένη μελέτη, επικεντρωθήκαμε στην αγορά των ακινήτων του νομού της Αττικής. Πραγματοποιήθηκε μία προσπάθεια υλοποίησης ενός μοντέλου πρόβλεψης των τιμών ενοικίασης των ακινήτων, ακολουθώντας τις μεθοδολογίες και τη λογική της μηχανικής μάθησης, που προβλέπονται μέσω της βιβλιογραφίας. Αρχικά, συλλέξαμε δεδομένα από την ιστοσελίδα του Σπιτόγατος, τα οποία αφορούσαν ακίνητα προς ενοικίαση, 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 απόκλιση, 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, αρκετοί αλγόριθμοι, για την εκπαίδευση των μοντέλων με είσοδο το σύνολο δεδομένων, από τους οποίους βέλτιστα αποτελέσματα μας έδωσαν οι τεχνικές παλινδρόμησης LGBM,XGBoost η παλινδρόμηση με χρήση της τεχνικής Multi-Layer και τα Τυχαία Δάση. Με βάση αυτά, διακρίνουμε και αναλύουμε τη σπουδαιότητα των χαρακτηριστικών και κατανοούμε τον ρόλο που κατέχει το καθένα από αυτά στην όψη των αποτελεσμάτων. Όλοι οι αλγόριθμοι έκριναν ότι το εμβαδόν των ακινήτων, η τιμή ζώνης , ο αριθμός των λουτρών και η ύπαρξη θέσης στάθμευσης αποτελούν τα κρισιμότερα γνωρίσματα, επηρεάζοντας σε σημαντικό βαθμό την τιμή ενοικίασης. 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.

Όπως παρατηρούμε μέσα από την ενασχόληση με την παραπάνω εργασία η προσπάθεια ακριβής πρόβλεψης της τιμής ενοικίασης ενός καταλύματος αποδεικνύεται να είναι ένα πολυδιάστατο πρόβλημα με πολλές δυσκολίες μα αυτό που αποδεικνύουμε είναι πως οι σύγχρονοι αλγόριθμοι μηχανικής μάθησης μπορούν και πετυχαίνουν σημαντικά κορυφαίες επιδόσεις .Παρόλα αυτά όμως η ανάγκη εύρεσης μεθοδολογιών που θα επιλύουν βέλτιστα το παραπάνω πρόβλημα είναι αναγκαία, καθώς επηρεάζει σημαντικά την καθημερινή ποιότητα ζωής του μέσου ανθρώπου.Πρώτα από όλα , σημαντικό είναι οι χρηματικές απαιτήσεις των ενοικιαστών να συμπίπτουν με τις παροχές .Είναι γνωστή η αύξηση τιμών σχεδόν σε οτιδήποτε χρησιμοποιούμε στην καθημερινότητα μας .Είναι γεγονός πως η παράμετρος της θέσης στάθμευσης είναι άκρως σημαντική , πρόβλημα που παρατηρούμε στην ευρύτερη περιοχή της Αττικής και ως επισκέπτες για σύντομο χρονικό διάστημα σε κάποια τοποθεσία.Σημαντικό θα είναι λοιπόν να επέμβει η πολιτεία ώστε να βρεθεί λύση στο παραπάνω ζήτημα , με επικρατέστερη επιλογή αυτή της δόμησης δημοτικών parking ,έτσι ώστε να μπορεί ο πολίτης ακόμα και αν δεν παρέχεται από τον εκμισθωτή, να έχει τη δυνατότητα να σταθμεύσει κοντά στην οικία του .

Τέλος, αναφορικά με το ερώτημα του ποιος αλγόριθμος μηχανικής μάθησης αποτελεί τον επικρατέστερο, κρίνοντας τις επιδόσεις τους. Η απάντηση είναι Random Forest Regression, Όλα τα μοντέλα φαίνεται να πετυχαίνουν εξαιρετικά ικανοποιητικές επιδόσεις και με μικρή απόκλιση των προβλέψεων από τις πραγματικές τιμές,μα το χαρακτηριστικό του Random Forest με το οποίο εστιάζει με υψηλή ακρίβεια, Ανθεκτικότητα,Ευελιξία σε δύσκολα δεδομένα , ανιχνευόντας ακραίες τιμές και χειρίζοντας ελλιπείς τιμές , είναι αυτό που του επιτρέπει να ξεπεράσει όλα τα υπόλοιπα μοντέλα μηχανικής μάθησης.

Αναφορές

[1]Σερβετάς, Γεώργιος. "Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ." 9 November 2017, https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/15959/Servetas_mppl19049.pdf?sequence=3&isAllowed=y. Accessed 9 March 2024.

[2] R. R. R. Botsman, "Product service systems," in What's Mine, New York: HarperCollins, 2010, pp. 106 - 108.

[3] "Fast facts," Dec 2022. [Online]. Available: https://news.airbnb.com/about-us/.

[4] D. Hill, "How Much Is Your Spare Room Worth?," IEEE Spectrum, vol. 52, no. 9, pp. 32-58, September 2015.

[5] T. Cai, K. Han and H. Wu, "Melbourne airbnb price prediction," 2019.

[6] A. Sihabuddin, "An Extreme Learning Machine Model Approach on Airbnb Base Price Prediction," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, January 2020.

[7] P. Ye, J. Qian, J. Chen, C. Wu, Y. Zhou, S. D. Mars, F. Yang and L. Zhang, "Customized Regression Model for Airbnb Dynamic Pricing," KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 932-940, July 2018. Μεταπτυχιακή Διατριβή Γεώργιος Σερβετάς Μηχανική Μάθηση στην Πρόβλεψη της τιμής ενοικίασης Airbnb στο Άμστερνταμ 59

[8] X. Z. Y. Z. Yuanhang Luo, "Predicting Airbnb Listing Price Across Different Cities," 2019.

[9] R. Pouya, N. Liubov and R. Hoormazd, "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis," International Cross-Domain Conference for Machine Learning and Knowledge Extraction, pp. 173-184, 2021.

[10] L. Lewis, "Predicting airbnb prices with machine learning and deep learning," May 2019.

[11] S. M. a. N. J. T. Mohd, "Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 11, pp. 542-546, September 2019.

[12] B. McNeil, "Price prediction in the sharing economy: A case study," 2020.

[13] R. Deboosere, D. J. Kerrigan, D. Wachsmuth and A. M. Elgeneidy, "Location, location and professionalization: a multilevel hedonic analysis of airbnb listing prices and revenue," Regional Studies, Regional Science,, vol. 6, no. 1, p. :143–156, 2019.

[14] D. Wang and J. L. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com," International Journal of Hospitality Management, vol. 62, pp. 120-131, 2017.

[15] https://csc.gr/machine-learning-michaniki-mathisi-ti-ine/

[16] https://christophm.github.io/interpretable-ml-book/shapley.html

[17] http://artemis.cslab.ece.ntua.gr:8080/jspui/bitstream

[18] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

[19]Sundararajan, Mukund, and Amir Najmi. "The many Shapley values for model explanation." arXiv preprint arXiv:1908.08474 (2019)

[20]Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on Artificial Intelligence and Statistics. PMLR (2020)

[21] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888

[22]NumPy -

[23]https://d1wqtxts1xzle7.cloudfront.net/49599119/Recent_developments_in_multilayer_percep

[24]pandas - Python Data Analysis Library (pydata.org)

[25]G. Hackeling, Mastering Machine Learning with scikit-learn. Packt Publishing Ltd, 2017.

[26]ChatGPT (openai.com)

[27]Cha-5.indd (researchgate.net)

[28]Linear regression - Su - 2012 - WIREs Computational Statistics - Wiley Online Library

[29]Machine Learning Benchmarks and Random Forest Regression (escholarship.org)

[30]Full article: Predicting property prices with machine learning algorithms (tandfonline.com)

[31]House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan | IEEE Conference Publication | IEEE Xplore

[32]JTAER | Free Full-Text | What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values (mdpi.com)

[33]Αγγελίες Ακινήτων - Αγγελίες Σπιτιών | Spitogatos

[34]Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines - ScienceDirect

[35]Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees - ScienceDirect

[36]thesis_Temenos_anastasios.pdf

[37]Remote Sensing | Free Full-Text | Novel Insights in Spatial Epidemiology Utilizing Explainable AI (XAI) and Remote Sensing (mdpi.com)

Παραρτήματα

|A|House_pricing.ipynb - Colaboratory (google.com)