ΤΕΛΙΚΗ ΑΝΑΦΟΡΑ

Μηνάς Αχλαδιανάκης
Ιωάννης Καμινάρης
Ιωάννης Τρανταλίδης

UNIVERSITY
OF CRETE

26 Μαΐου 2021

1 ΕΙΣΑΓΩΓΗ

Αντικείμενο της εργασίας, είναι η εκτίμηση της τιμής πώλησης ακινήτων,μέσω μοντέλων μηχανικής μάθησης για παλινδρόμηση. Μια ακριβής πρόβλεψη για την τιμή του ακινήτου είναι σημαντική για τους μελλοντικούς ιδιοκτήτες επενδυτές, εκτιμητές και φορολογητές. Η παραδοσιακή πρόβλεψη της τιμής κατοικίας γίνεται με βάση τη σύγκριση κόστους και τιμής πώλησης, που όμως δεν έχει την απαραίτητη πιστοποίηση. Επομένως, ένα μοντέλο πρόβλεψης της τιμής βοηθά στη συμπλήρωση πληροφοριών και στη βελτίωση της αποτελεσματικότητας της αγοράς ακινήτων. Χρησιμοποιώντας δεδομένα από 1460 ακίνητα της πόλης Ames, της πολιτείας Αϊόβα των Ηνωμένων Πολιτειών της Αμερικής, στοχεύουμε στην ανάπτυξη αλγορίθμων για την εκτίμηση της τιμής των κατοικιών, βασιζόμενοι σε 79 σταθερά χαρακτηριστικά, δηλαδή χαρακτηριστικά που δεν αλλάζουν εύκολα όπως τοποθεσία, τετραγωνικά, αριθμός υπνοδωματίων κ.τ.λ. Ο κώδικας που φτιάξαμε είναι διαθέσομος στο Github.

2 ΜΕΘΟΔΟΛΟΓΙΑ

Τα δεδομένα έχουν ληφθεί από έναν διαγωνισμό για προχωρημένες τεχνικές παλινδρόμησης του

Kaggle. [1] Στο Kaggle και γενικότερα στη βιβλιογραφία, υπάρχουν μοντέλα με αποτελέσματα ως προς το Root Mean Squared Error,της τάξεως του 0.1,ανάμεσα στο λογάριθμο της τιμής πώλησης και της προβλεπόμενης τιμής των ακινήτων. Εμείς αναπαράγουμε αυτά τα αποτελέσματα,αναλύοντας πλήρως τη διαδικάσια που ακολουθήσαμε ,καθώς και τα μοντέλα που χρησιμοποιήσαμε.Επίσης,προγραμματίσαμε πλήρως τον κώδικα για τα μοντέλα μηχανικής μάθησης που χρησιμοποιήσαμε. Ο κώδικας ειναι διαθέσιμος στο Github. [2]

Η προσέγγιση που χρησιμοποιήσαμε έχει ως βάση την τεχνική της μηχανικής μάθησης, η οποία ονομάζεται gradient boosting.Η πρόβλεψη αυτής της τεχνικής, βασίζεται σε ένα σύνολο(ensemble) αδύναμων μοντέλων πρόβλεψης. Συγκεκριμένα χρησιμοποιήσαμε 2 παραλλαγές αυτης της τεχνικής,βασιζόμενοι σε δέντρα αποφάσεων:

1).Gradient boosted trees με συνάρτηση κόστους Mean squared Error.

[1] ΒΙΒΛΙΟΓΡΑΦΙΑ
[2] ΒΙΒΛΙΟΓΡΑΦΙΑ

2).XGBoost για δέντρα αποφάσεων, με συνάρτηση κόστους Mean squared Error.Αυτός ο αλγόριθμος, είναι παρόμοιος με τον προηγούμενο,αλλά προσθέτει στη συνάρτηση κόστους, έναν όρο εξομάλυνσης $L2$.

Παρόλλο που έχουμε προγραμματίσει εξ ολοκλήρου αυτά τα μοντέλα,για λόγους ταχύτητας θα χρησιμοποιήσουμε τις παρακάτω βιβλιοθήκες της python:

1. Gradient boosted trees: Sklearn [3]
2. XGBoost: XGBoost [3]

Τέλος, αναλύσαμε τη σημάντικότητα των χαρακτηριστικών που δίνουν τα μοντέλα ,για την επιρροή τους στην τιμή πώλησης και προτείναμε μελλοντικούς τρόπους προσέγγισης του προβλήματος,για την βελτίωση των αποτελεσμάτων. Τα δεδομένα είναι από 1460 ακίνητα της πόλης Ames. Το σύνολο δεδομένων περιλαμβάνει 79 μεταβλητές (23 ονομαστικές, 23 διατάξιμες, 14 διακριτές, 19 συνεχείς) που περιγράφουν σχεδόν όλες τις πτυχές των ιδιοτήτων που μπορεί να ενδιαφέρουν αγοραστές και επενδυτές, κατά την αξιολόγηση ενός ακινήτου.

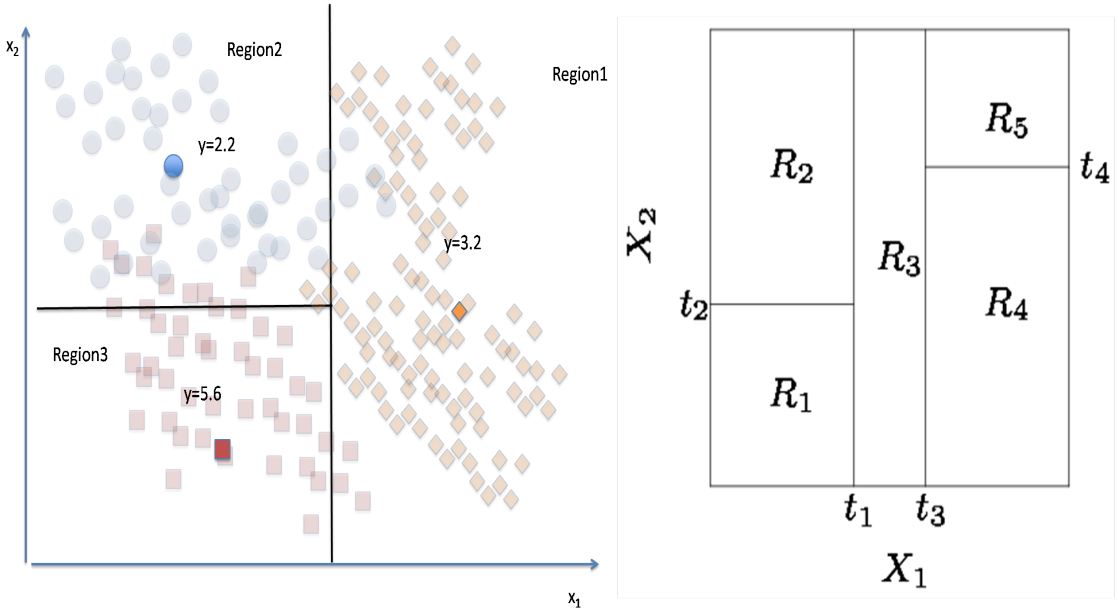
[3]ΒΙΒΛΙΟΓΡΑΦΙΑ
[3]ΒΙΒΛΙΟΓΡΑΦΙΑ

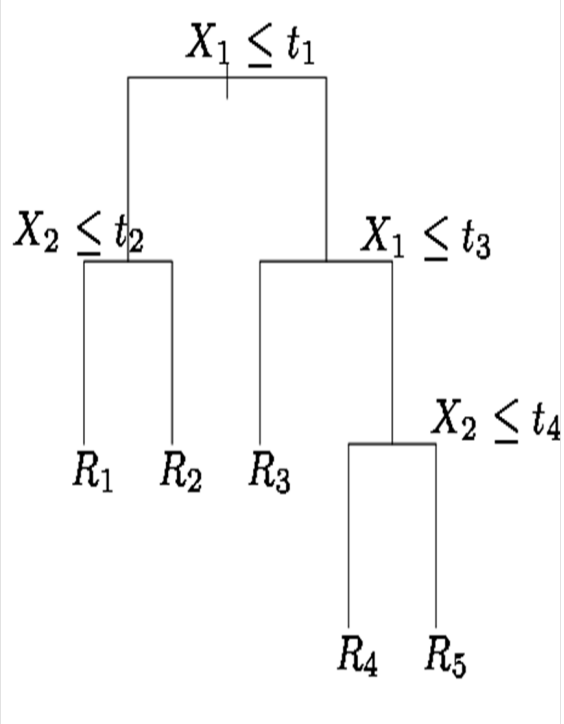
3 ΑΛΓΟΡΙΘΜΟΙ-ΛΟΓΙΣΜΙΚΟ

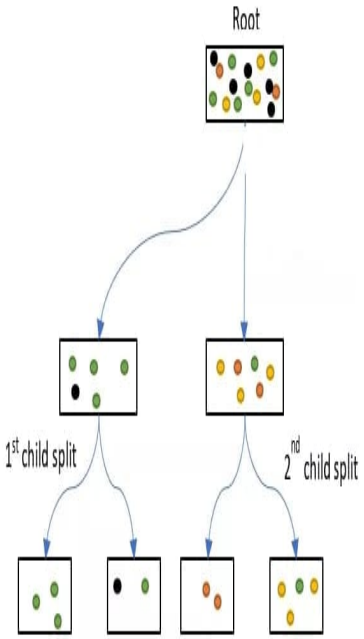
3.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ ΓΙΑ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΣΥΝΑΡΤΗΣΗ ΚΟΣΤΟΥΣ MSE

ΒΑΣΙΚΗ ΙΔΕΑ

- Χώριζουμε τα δεδομένα σε **περιοχές-regions** $R_i ... R_k$
- Για κάθε δείγμα $X = (X_1, \dots X_n)$ που πέφτει σε μια περιοχή,η προβλεψή μας είναι **η μέση τιμή των** $\boldsymbol{y^{(i)}}$, για τα δείγματα εκπαίδευσης $X^{(i)}$ που πέφτουν σε αυτή την περιοχή.







Αλγόριθμος:

- Δεδομένα $\{X^i, y^i\}$ με χαρακτηριστικά $X_1, \ldots X_n$
- Ελέγχουμε όλες τιμές t των $X_i, \forall\ i \in \{1, \ldots, n\}$
- Επιλέγουμε το χαρακτηριστικό X_i και την τιμή t. που δημιουργεί τον καλύτερο διαχωρισμό στα y. (θα δουμε πως)
- Αν $X_i < t$ στέλνουμε το δείγμα αριστερά στο δέντρο,αλλιώς δεξιά.
- Επαναλαμβάνουμε τη διαδικασία στους κόμβους που δημιουργήσαμε.
- Ο τερματικός κόμβος ονομάζεται περιοχή- Region ή φύλλο.

ΚΑΛΥΤΕΡΟΣ ΔΙΑΧΩΡΙΣΜΟΣ

Η μείωση διασποράς ισοδυναμεί με μείωση MSE:

$$Var(y_{i\in R_j}) = \frac{1}{n}\sum_{i\in R_j}(y_i - \bar{y}_{i\in R_j})^2 \quad = \frac{1}{n}\sum_{i\in R_j}(y_i - \hat{y}_{i\in R_j})2 = MSE(R_j)$$

Αρα θέλουμε κάθε διαχωρισμός να ελαχιστοποιεί τη διασπορά των $y^{(i)}$ στους κάτω κόμβους,σε σχέση με τον πάνω, ώστε στις περιοχές να έχουμε το λιγότερο MSE

Για να μετρήσουμε πόσο έχει μειωθεί η διασπορά για τα y των κάτω κόμβων, σε σχέση με του πάνω:

$$variance \quad reduction =$$

$$Var(y_{up}) - \left(\frac{|R_{left}|}{|R|}Var(y_{left}) + \frac{|R_{right}|}{|R|}Var(y_{right})\right)$$

$|R_{left(right)}|$: Αριθμός δειγμάτων που καταλήγουν αριστερά (ή δεξιά) μετά το διαχωρισμό.
$|R|$: Αριθμός δειγμάτων στον κόμβο που κάνουμε το διαχωρισμό.
$|y_{left(right,up)}|$:Τα y των δειγμάτων στον πάνω αριστερά ή δεξιά κόμβο.

ΚΡΙΤΗΡΙΟ ΤΕΡΜΑΤΙΣΜΟΥ/ Overfit

Τα δέντα αποφάσεως κάνουν εύκολα overfit στα δεδομένα εκπαίδευσης,γιαυτο ορίζουμε παραμέτρους τις οποίες θα βελτιστοποιήσουμε με hyper parameter tuning:

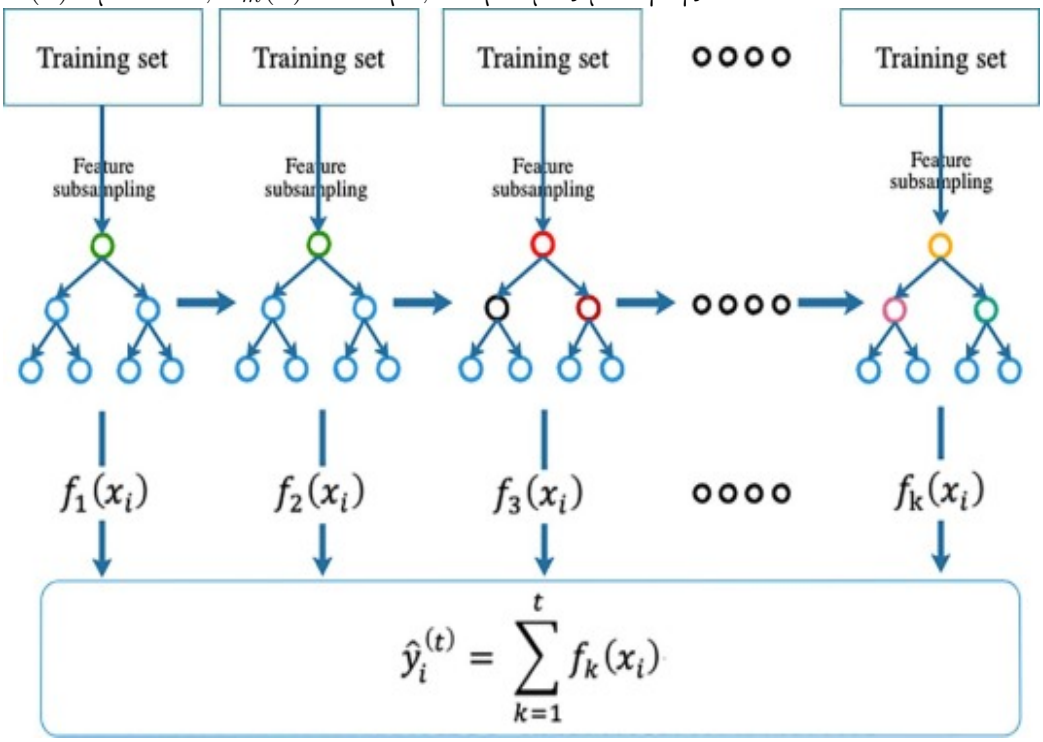
Παράμετροι	**Περιγραφή**
max depth	Μέγιστο επίπεδο που μπορούν να φτάσουν τα δέντρα.
min samples split	Αν ένας κόμβος έχει λιγότερα δείγματα από αυτήν την τιμή,δεν κανούμε επιπλέον διαχωρισμό σε αυτόν.
tol	Αν η μείωση της διασποράς, μετά απο ένα διαχωρισμό γίνει πολύ μικρή π.χ.(10^{-7}) σταματάμε για αυτόν τον κόμβο.

3.2 Gradient Boosting

ΒΑΣΙΚΗ ΙΔΕΑ

Δημιουργία μοντέλου πρόβλεψης, χρησιμοποιώντας ενα σύνολο(ensemble) α-δύναμων μοντέλων(δέντρα).
Το μοντέλο είναι άθορισμα M δέντρων παλινδρόμησης.
$\hat{y} = F(x) = \sum_{m1}^{M} a h_m(x)$

$F(x)$: μοντέλο, $h_m(x)$: δέντρο, a : ρυθμός μάθησης



ΥΛΟΠΟΙΗΣΗ

Η υλοποίηση του αλγορίθμου γίνεται σε βήματα.Σε κάθε βήμα $m = 1 \dots M$ προσθέτουμε ένα δέντρο παλινδρόμησης h_m.

$$F_m = F_{m-1} + ah_m(x)$$

Η συνάρτηση απώλειας που θα χρησιμοπιήσουμε είναι η τετραγωνική:

$$L(y, F) = \sum_{i=1}^{n} (y_i - F)^2/2$$

Αρχικοποίηση

$$F_0 = \underset{\gamma}{\operatorname{argmin}} L(y, \gamma) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \gamma)^2/2$$

$$\frac{\partial L(y, \gamma)}{\partial \gamma} = 0 \Rightarrow F_0 = \sum_{i=1}^{n} y_i/n = \bar{y_n}$$

Δέντρο στο m βήμα

Εστω H :το σύνολο όλων των δέντρων παλινδρόμησης. Θέλουμε να διαλέξουμε ένα $h_m \in H$ για να προσθέσουμε στο m βήμα.

$$h_m = \underset{h \in H}{\operatorname{argmin}} L(y, F_m) = \underset{h \in H,}{\operatorname{argmin}} L(y, F_{m-1}+h) = \underset{h \in H,}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - F_{m-1} - h)^2/2$$

$$\Rightarrow$$

$$h_m(x) = \underset{h \in H,}{\operatorname{argmin}} \sum_{i=1}^{n} (r_{i,m} - h)^2/2$$

$r_{i,m} = y_i - F_{m-1}$, **σφάλματα του μοντέλου στο προηγούμενο βήμα**

Αρα για να ελαχιστοποιήσουμε αυτό το $Loss$,όπως είδαμε στα δέντρα αποφάσεων,το h_m είναι δέντρο παλινδρόμηδης ,που θα εκπαιδευτεί στα $\{X, r_{i,m}\}$ με συνάρτηση κόστους MSE
Ετσι η προβλεψή μας για ενα δείγμα x που ανήκει σε μια περιοχή R_j είναι:

$$h_m(x_{i\in R_j}) = \text{μέση τιμή των } r_{i\in R_j} = \frac{\sum_{i\in R_j} r_{i,m}}{|R_j|}$$

Δηλαδή το h_m προβλέπει τα σφάλματα στο προηγούμενο βήμα,ώστε σε δεδομένα που δεν έχει εκπαιδευτεί, να μπορεί να τα διορθώνει.

Algorithm 1 Gradient Boosting Για Τετραγωνική Απώλεια

$F_0 = \bar{y_n}$
for $m = 1, \dots M$ **do**:
  $r_{i,m} = y_i - F_{m-1}(x_i)$
  εκπαίδευσε h_m στα $\{X, r_{i,m}\}$
  $h_m(x_{i\in R_j}) = \frac{\sum_{i\in R_j} r_{i,m}}{|R_j|}$
  $F_m = F_{m-1} + ah_m(x)$

ΠΑΡΑΜΕΤΡΟΙ

Οι παράμετροι που θα χρησιμοπουήσουμε στο μοντέλο:

Παράμετροι	**Περιγραφή**
max depth	Μέγιστο επίπεδο που μπορούν να φτάσουν τα δέντρα.
learning rate	Ρυθμός Μάθησης.
number of estimators	Πόσα δέντρα θα χρησιμοποιήσουμε M.
min samples split	Αν ένας κόμβος έχει λιγότερα δείγματα από αυτήν την τιμή,δεν κανούμε επιπλέον διαχωρισμό σε αυτόν.

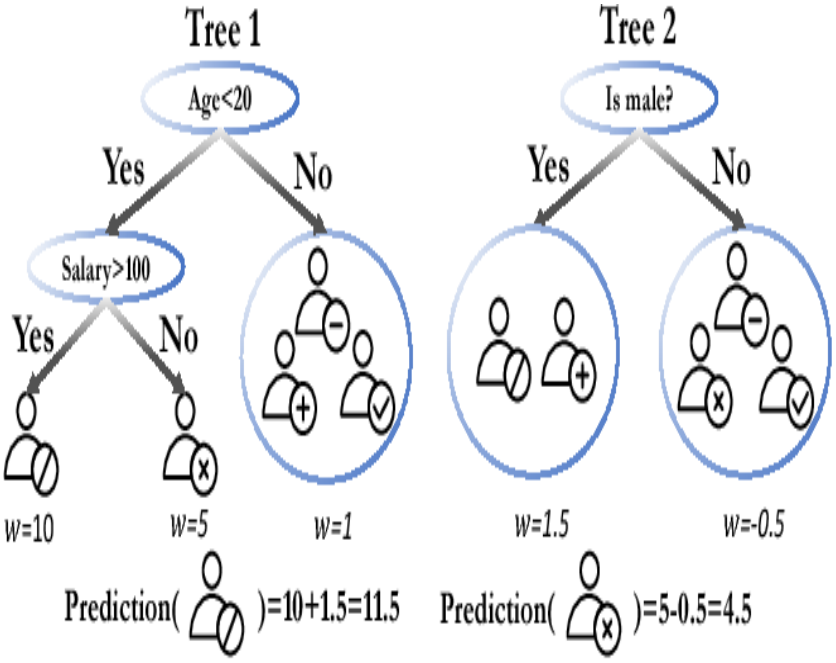
3.3 XGBoost

Το μοντέλο είναι το ίδιο

$$F(x) = \sum_{m=1}^{M} a h_m(x)$$

αλλά προσθέτουμε στη συνάρτηση κόστους έναν όρο εξομάλυνσης $L2$(πέναλτυ).

$$J^{(m)} = \sum_{i=1}^{n} (y_i - F_m(x_i))^2/2 + \frac{1}{2}\lambda \sum_{j=1}^{K} w_j^2, \quad w_j = h_m(x_{i \in R_j})$$



ΥΛΟΠΟΙΗΣΗ

$$
\begin{aligned}
J^{(m)} &= \sum_{i=1}^{n}(y_i - F_m)^2/2 + \frac{1}{2}\lambda\sum_{j=1}^{K} w_j^2 \\
&= \sum_{i=1}^{n}(y_i - F_{m-1} - h_m)^2/2 + \frac{1}{2}\lambda\sum_{j=1}^{K} w_j^2 \\
&= \sum_{j=1}^{k}\sum_{i\in R_j}(y_i - F_{m-1}(x_i) - w_j)^2/2 + \frac{1}{2}\lambda\sum_{j=1}^{K} w_j^2
\end{aligned}
$$

$$
\frac{\partial J^{(m)}}{\partial w_j} = \sum_{i\in R_j}(-r_{i,m} + w_j) + \lambda w_j = 0
$$

$$
w_j^* = \frac{\sum_{i\in R_j} r_{i,m}}{\lambda + |R_j|}
$$

Αντικαθιστούμε το w_j^* στο $J^{(m)}$ και παίρνουμε την τιμή που θέλουμε να μειώνει κάθε διαχωρισμός του δέντρου h_m.

$$
I_{R_j} = J^{(m)}(w_j^*) = \frac{1}{2}\frac{\sum_{i\in R_j} r_{i,m}^2}{|R_j| + \lambda}
$$

Ετσι, κάθε διαχωρισμός του h_m θέλουμε να αυξάνει το:

$$
\begin{aligned}
Criterion &= I_{up} - (I_{left} + I_{right}) \\
&= \frac{1}{2}\frac{\sum_{i\in parent} r_i^2}{|parent| + \lambda} - \frac{1}{2}\left(\frac{\sum_{i\in leftNode} r_i^2}{|leftNode| + \lambda} + \frac{\sum_{i\in rightNode} r_i^2}{|rightNode| + \lambda}\right)
\end{aligned}
$$

Algorithm 2 XGBoost Για Τετραγωνική Απώλεια και εξομάλυνση L2

$F_0 = \bar{y_n}$
for $m = 1, \dots M$ **do**:
$r_{i,m} = y_i - F_{m-1}(x_i)$
εκπαίδευσε h_m στα $\{X, r_{i,m}\}$ με βάση το Criterion
$h_m(x_{i \in R_j}) = \frac{\sum_{i \in R_j} r_{i,m}}{\lambda + |R_j|}$
$F_m = F_{m-1} + a h_m(x)$

ΠΑΡΑΜΕΤΡΟΙ

Οι παράμετροι που θα χρησιμοπουήσουμε στο μοντέλο:

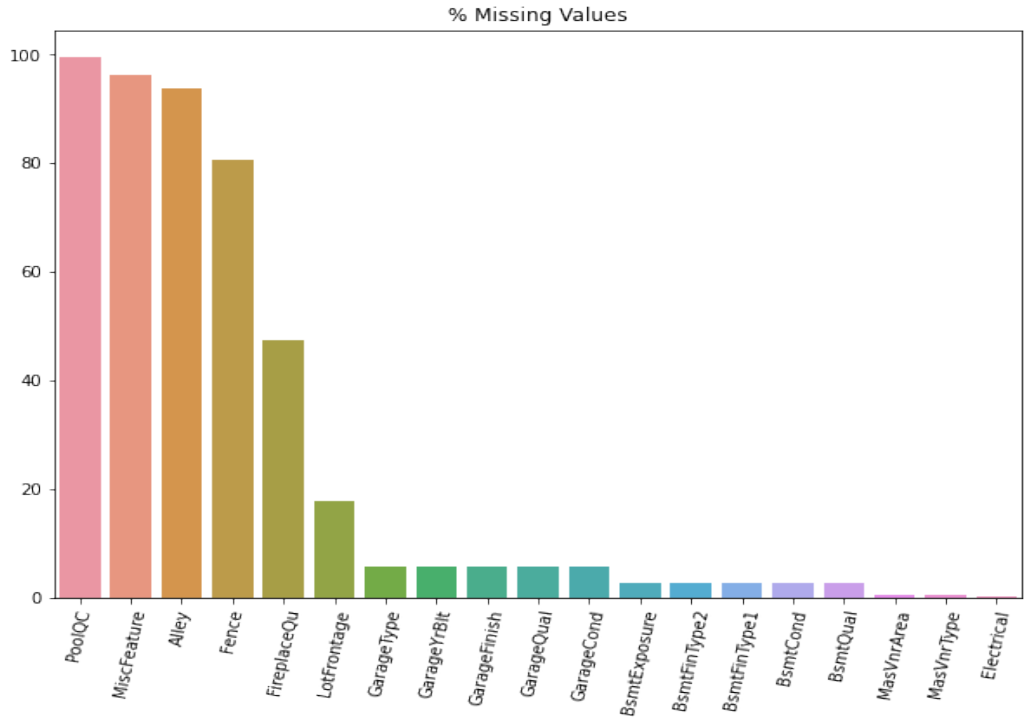
Παράμετροι	**Περιγραφή**
max depth	Μέγιστο επίπεδο που μπορούν να φτάσουν τα δέντρα.
learning rate	Ρυθμός Μάθησης.
number of estimators	Πόσα δέντρα θα χρησιμοποιήσουμε M.
min samples split	Αν ένας κόμβος έχει λιγότερα δείγματα από αυτήν την τιμή,δεν κανούμε επιπλέον διαχωρισμό σε αυτόν.
λ	Εξομάλυνση $L2$.

3.4 Feature importance of models

Και με τα 2 μοντέλα μπορούμε να υπολογίσουμε το πόσο σημαντικό ήταν κάθε χαρακτηριστικό στην πρόβλεψη της τιμής πώλησης. Σε κάθε ένα απο τα M δέντρα , ορίζουμε ως σημαντικότητα ενός διαχωρισμού,την αύξηση του κριτηρίου με το οποίο επιλέγεται ο καλύτερος διαχωρισμός για αυτό το δέντρο(π.χ. για δέντρα με συνάρτηση κόστους MSE είναι το variance reduction).Ο διαχωριμός γίνεται με βάση ένα χαρακτηριστικό, αρα έχουμε τη σημαντικότητα αυτού του χαρακτηριστικού.Σε ένα δέντρο μπορεί να υπάρχουν 2 και παραπάνω διαφορετικοί διαχωρισμοί με βάση το ίδιο χαρακτηριστικό.Η σημαντικότητα του χαρακτηριστιτκου σε κάθε δέντρο είναι το αθροισμα της αύξησης του κριτηρίου ,όλων των διαχωρισμών για το συγκεκριμένο χαρακτηριστικό στο δέντρο.Η τελική εκτίμηση για τη σημαντικότητα ενός χαρακτηριστικού, ειναι η μέση τιμή της σημαντικότητας του σε κάθε ενα απο τα M δέντρα.Τέλος,κανονικοποιούμε τη σημάντικότητα των χαρακτηριστιτκών με τη μοναδιαία νόρμα.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

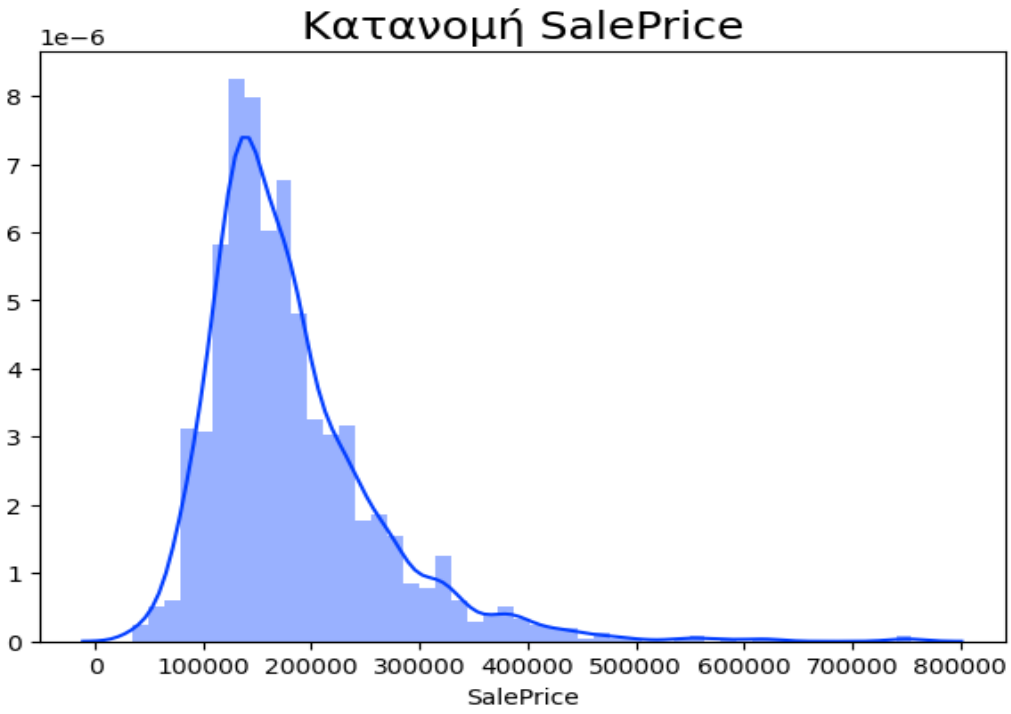
4.1 Missing Values



Σύμφωνα με την περιγραφή των δεδομένων,οι τιμές που λείπουν,για τα περισότερρα χαρακτηριστικά, δηλώνουν την ανυπαρξία του χαρακτηριστικού στο εκάστοτε δείγμα.Αρα σε όλες εκτός απο μία στήλη αντικαταστήσαμε τις τιμές που λείπουν με 0 για τις ποσοτικές και None για τις ποιοτικές.
Για το Lot Frontage,δηλαδή την έκταση του δρόμου (σε πόδια) που συνδέεται με την κατοικία,οι τιμές που λείπουν αποτελούν το 17% Θεωρήσαμε οτι οι κατοικίες μιας γειτονίας έχουν παρόμοια έκταση δρόμου και αντικαταστήσαμε τις τιμές που λείπουν, με τη διάμεσο των κατοικιών της γειτονίας, που ανήκει η εκάστοτε κατοικία.

4.2 ΣΥΣΧΕΤΙΣΗ

Θα περιγράψουμε τη γραμμική συσχέτιση των χαρακτηριστικών με την τιμή πώλησης.Αρχικά ας δούμε την κατανομή της τιμής πώλησης:



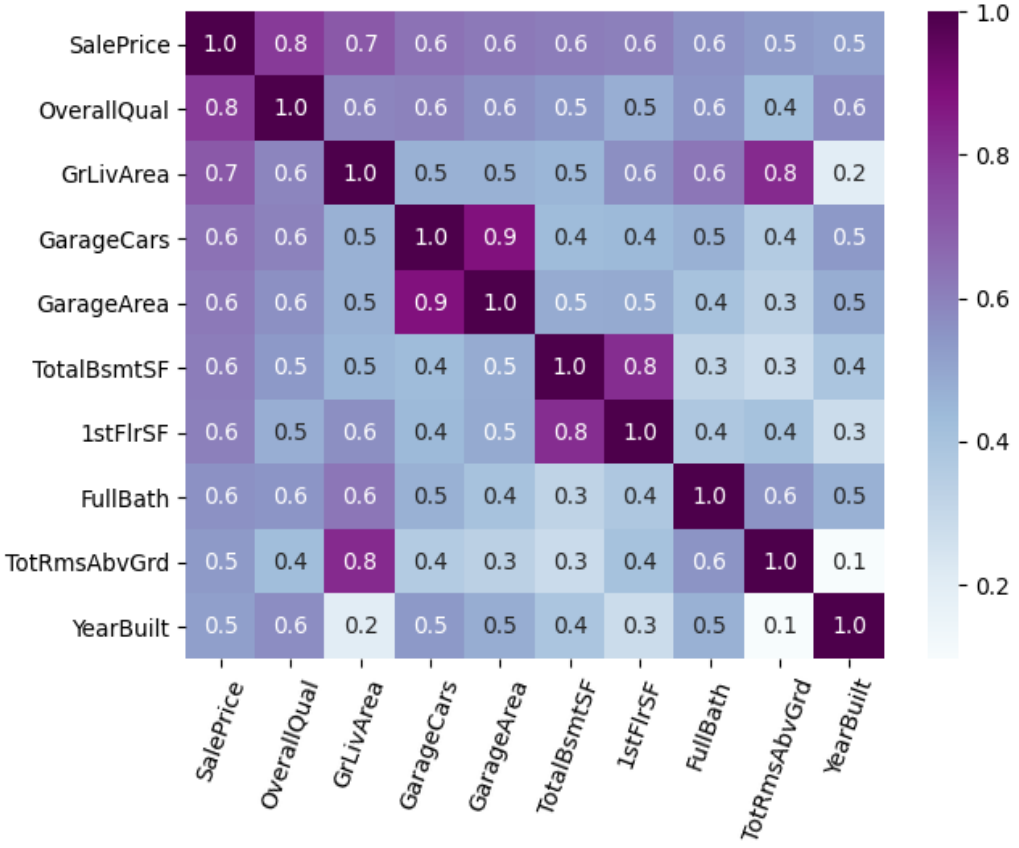
Οπως φαίνεται στον παρακάτω πίνακα,οι μεταβλητές που έχουν θετική γραμμική συσχέτιση με την τιμή πώλησης του ακινήτου είναι αυτές της ποιότητας του ακινήτου, της περιοχής που βρίσκεται και της χωρητικότητας του αμαξοστασίου:

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	ΣΥΣΧΕΤΙΣΗ με SalePrice
Sale Price	1.000000
Overall Quality	0.790982
Greater Living Area	0.708624
Garage Cars	0.640409
Garage Area	0.623431
Total Basemnet Square Feet	0.613581
First Floor Square Feet	0.605852
Full Bathrooms	0.560664
Total Rooms Above Ground	0.533723
Year Built	0.522897

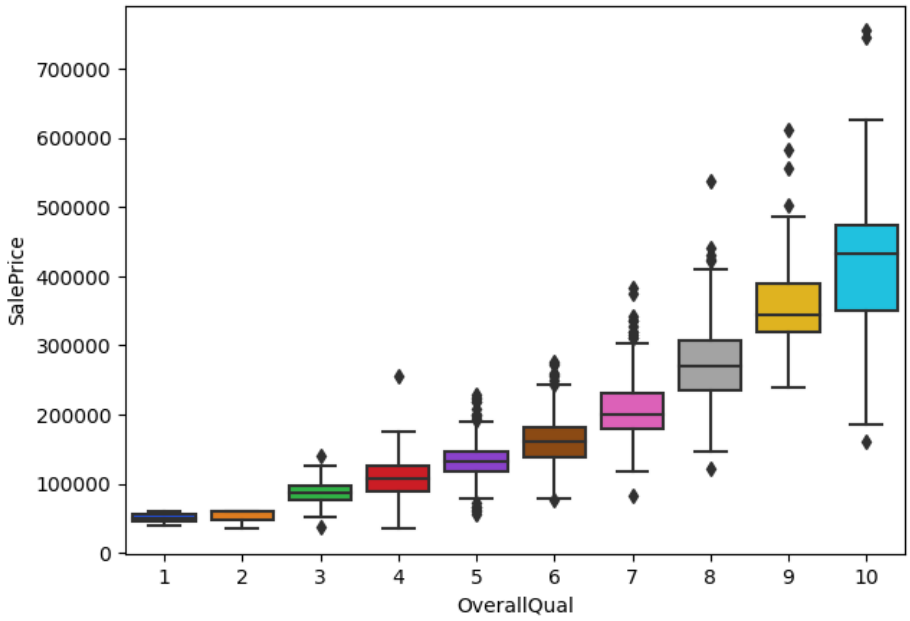
Ενώ, οι μεταβλητές που έχουν αρνητική γραμμική συσχέτιση με την τιμή πώλησης είναι αυτές της θέσης της κουζίνας, της κλειστής βεράντας (αν υπάρχει) και της γενικότερης κατάστασης του ακινήτου, όπως φαίνεται στον παρακάτω πίνακα:

ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ	ΣΥΣΧΕΤΙΣΗ με SalePrice
Kitchen Above Ground	-0.135907
Enclosed Porch	-0.128578
MSSubClass	-0.084284
Overall Condition	-0.077856
Year Sold	-0.028923
Low Quality FinancialSF	-0.025606
Miscaleneous Values	-0.021190
Basement and Half Bathroom	-0.016844
BsmtFinSF2	-0.011378

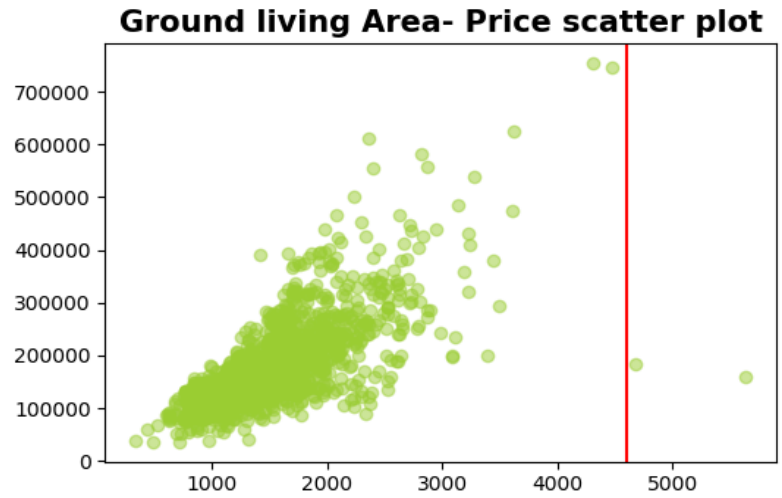
Η συσχέτιση των 10 χαρακτηριστικών με τη μεγαλύτερη γραμμική συσχέτιση με την τιμή πώλησης:

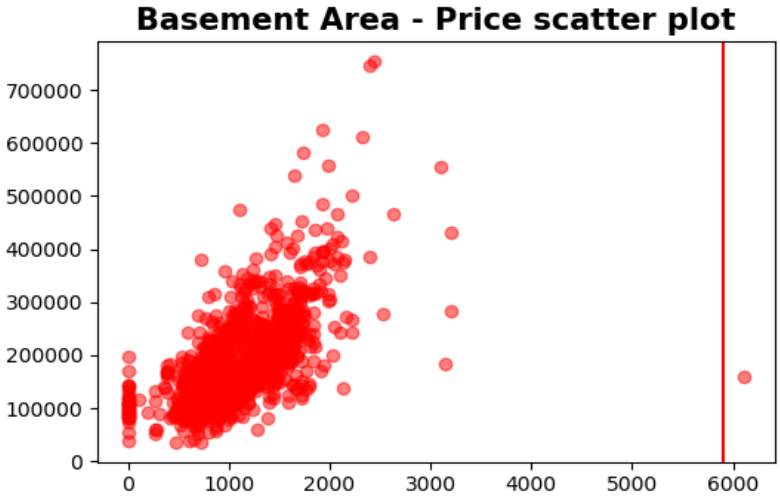


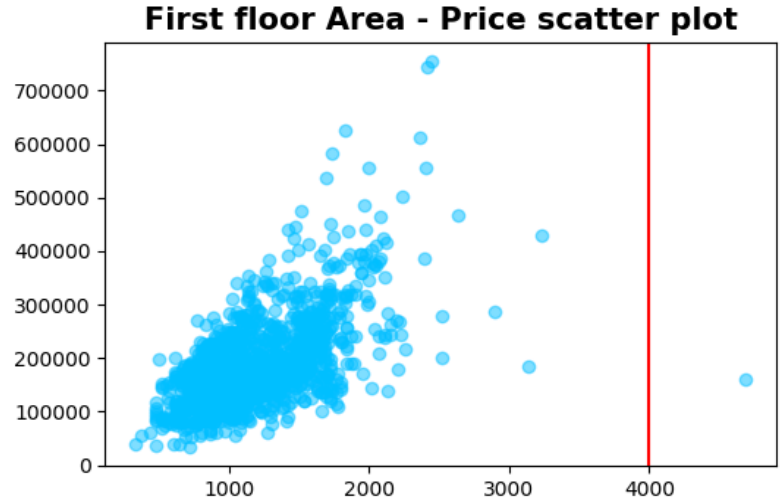
Στο παρακάτω box plot βλέπουμε πως η τιμή πώλησης και ποιότητα ακινήτου έχουν ισχυρή (εκθετική) συσχέτιση, ενώ το πλήθος των ακραίων τιμών (sale price) φαίνεται να είναι μεγαλύτερο όσο αυξάνεται η ποιότητα του σπιτιού.

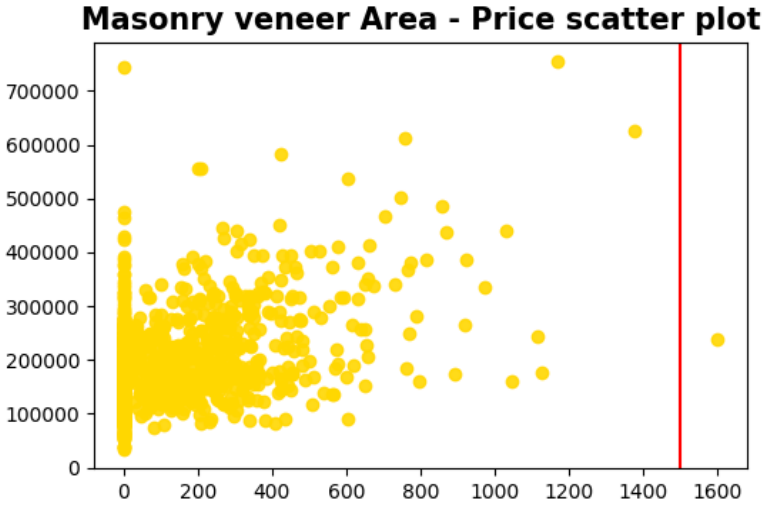


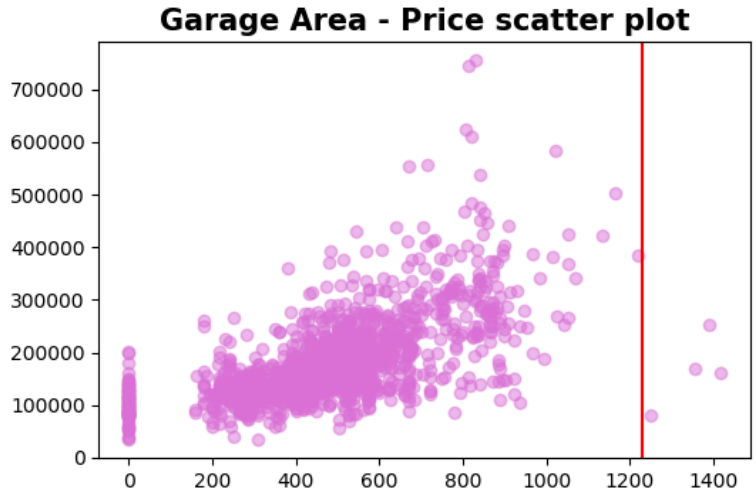
Παρακάτω,είναι τα γραφήματα των χαρακτηριστικών,με την τιμή πώλησης.Η κόκκινη γραμμή υποδεικνύει ότι πέρα από αυτή, είναι οι ακραίες τιμές τις οποίες αφαιρέσαμε.

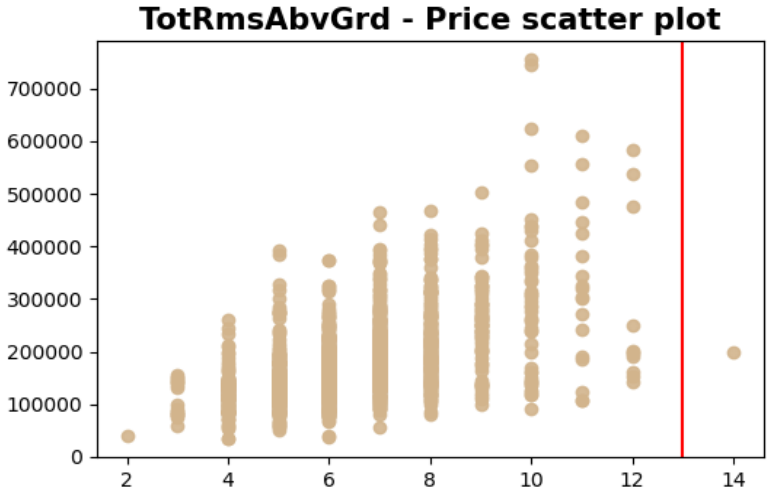

Ground living Area- Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000
5000


Basement Area - Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
6000


First floor Area - Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
1000
2000
3000
4000

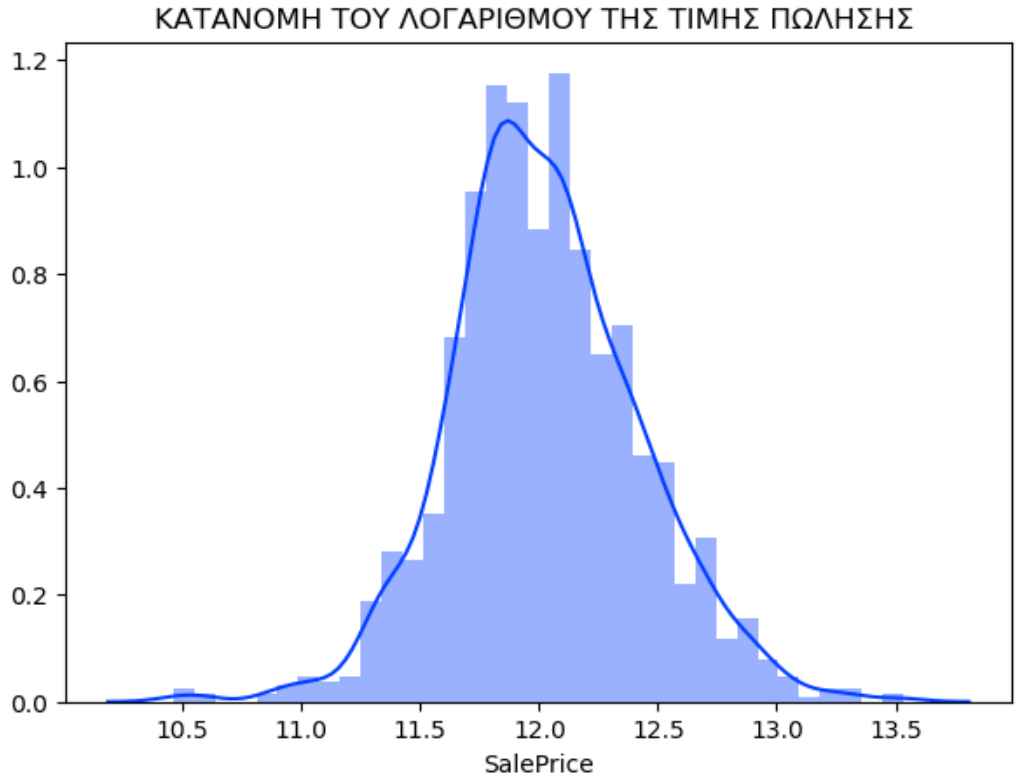

Masonry veneer Area - Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400
1600


Garage Area - Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
0
200
400
600
800
1000
1200
1400


TotRmsAbvGrd - Price scatter plot
700000
600000
500000
400000
300000
200000
100000
0
2
4
6
8
10
12
14

4.3 ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Αρχικά,μετασχηματίσαμε την τιμή πώλησης με το λογάριθμο της. Ετσι τα σφάλματα στην πρόβλεψη ακριβών και φτηνών ακινήτων,θα έχουν ίδια επιρροή.



Τα υπόλοιπα χαρακτηριστικά χωρίζονται σε ποιοτικά και ποσοτικά. Τα ποιοτικά, χωρίζονται σε 2 κατηγορίες: Τα διατάξομα (ordinal) και τα μη-διατάξιμα.

ΔΙΑΤΑΞΙΜΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

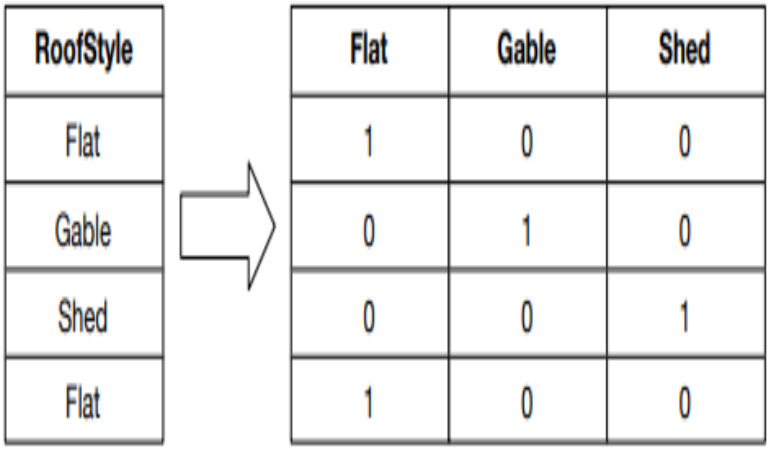
Τα διατάξιμα ποιοτικά χαρακτηριστικά, είναι αυτά στα οποία οι κατηγορίες του χαρακτηριστικού είναι ταξινομημένες.

Σύμφωνα με την περιγραφή των δεδομένων , κάναμε τον ακόλουθο μετασχηματισμό στις κατηγορίες των διατάξιμων χαρακτηριστικών:

ΚΑΤΗΓΟΡΙΑ	ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ
Lvl	3
Bnk	2
HLS	1
Low	0
Ex	4
Gd	3
TA	2
Fa	1
Po	1
None	0
Y	1
N	0
Reg	3
IR1	2
IR2	1
IR3	0
GLQ	6
ALQ	5
BLQ	4
Rec	3
LwQ	2
Unf	1
Av	3
Mn	2
No	1
Sev	2
Mod	1
Gtl	0

ΜΗ-ΔΙΑΤΑΞΙΜΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Για το μετασχηματισμό των μη-διατάξιμων ποιοτικών χαρακτηριστικών ,χρησιμοποιήσαμε One-hot encoding.Ενα παράδειγμα:



RoofStyle
Flat
Gable
Shed
Flat

Flat	Gable	Shed
1	0	0
0	1	0
0	0	1
1	0	0

ΠΟΣΟΤΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Επειδή το μοντέλο μας βασίζεται σε δέντρα αποφάσεων,δεν απαιτείται κανονικοποίηση των ποσοτικών χαρακτηριστικών.

4.4 XGBoost- ΑΠΟΤΕΛΕΣΜΑΤΑ

Hyper Parameter tuning

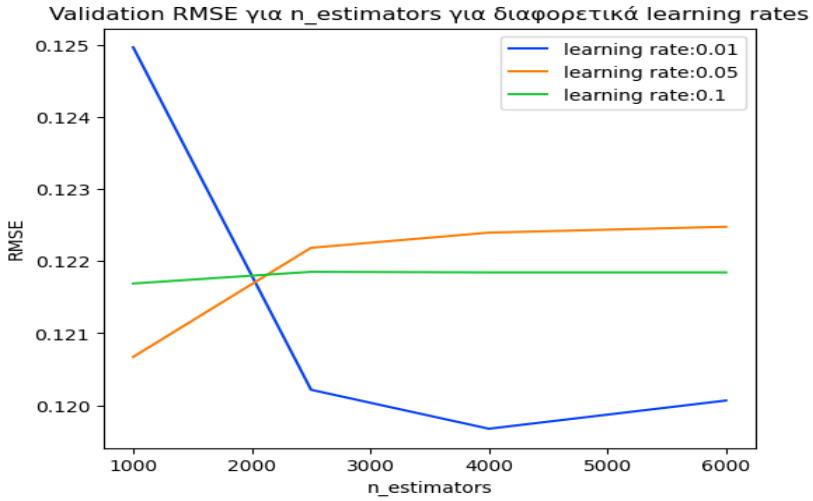
Εξετάσαμε διάφορους συνδυασμούς των παραμέτρων του μοντέλου, και επιλέξαμε αυτόν που δίνει το λιγότερο $RMSE$,με 3-Fold cross validation.Οι παράμετροι και οι τιμές που εξετάσαμε φαίνονται στον ακόλουθα πίνακα:

	Παράμετροι
max depth	$[3, 5, 10]$
learning rate	$[0.01, 0.05, 0.1]$
number of estimators	$[1000, 2500, 4000, 6000]$
λ	$[0.01, 1, 10]$
min child weight	$[1, 5, 10]$

Ο καλύτερος συνδύασμος παραμέτρων φαίνεται στον ακόλουθο πίνακα :

	Παράμετροι
max depth	3
learning rate	0.01
number of estimators	4000
λ	1
min child weight	1

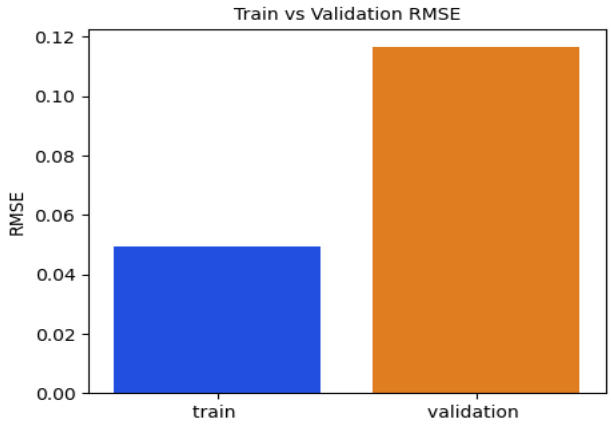
Παρακάτω φαίνονται τα αποτελέσματα για διάφορες τιμές των number of estimators και learning rate, με τις υπόλοιπες παραμέτρους σταθερές , με τιμές αυτές του καλύτερου συνδυασμού που περιγράφει ο παραπάνω πίνακας:



ΑΞΙΟΛΟΓΗΣΗ

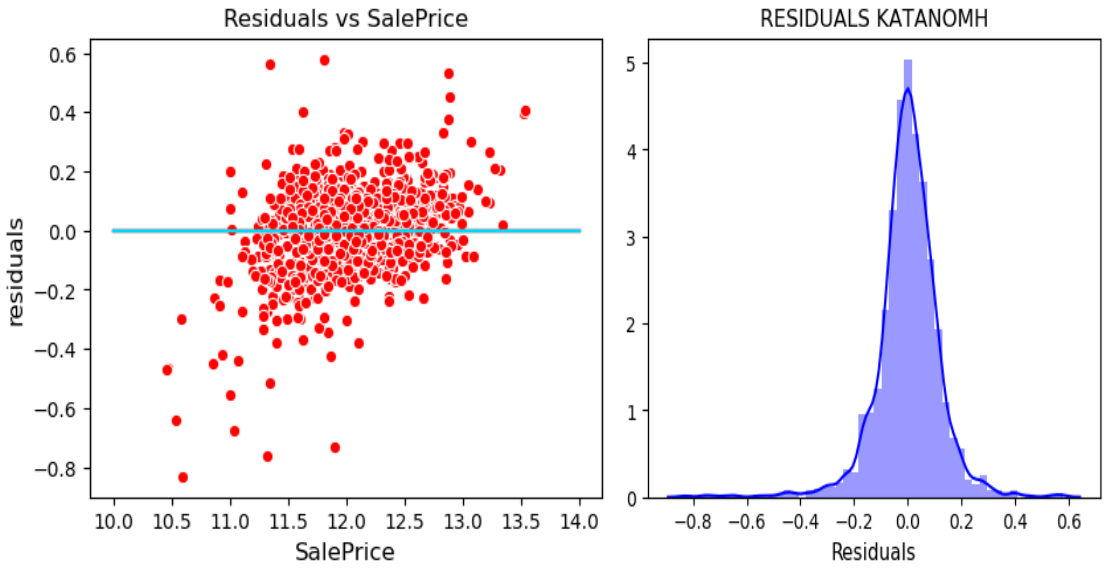
Επειδή η εύρεση των καλύτερων παραμέτρων είναι υπολογιτικά ακριβή,χρησιμοποιήσαμε μόνο 3-Fold cross validation.Γιαυτό η αξιολόγηση του μοντέλου με τι καλύτερες παραμέτρους έγινε με 10-Fold cross validation.Τα αποτελέσματα στα train και validation set ,φαίνονται παρακάτω:

	Train set	Validation set
RMSE	0.049521	0.116555



Το μοντέλο έχει πετύχει τον στόχο του να είναι,το RMSE κοντά στο 0.1.Ομως κάνει overfit στα δεδομένα εκπαίδευσης,άρα υπάρχουν περιθώρια βελτίωσης.

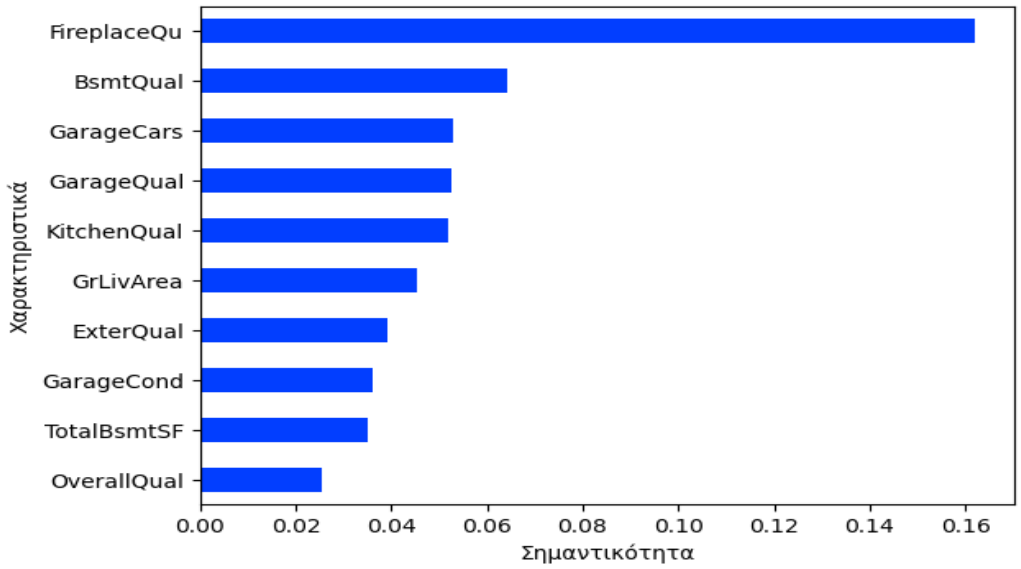
Παρακάτω φαίνεται η κατανομή των $residuals = y - \hat{y}$:



Βλέπουμε οτι για μικρές τιμές του SalePrice , υπάρχουν λίγα σημεία με μεγάλυτερο *error*,από τα άλλα.Αυτό φαίνεται και στην κατανομή των residuals , οπου οι ακραίες αρνητικές τιμές είναι ελαφρώς ανεβασμένες.

Feature importance

Τα 10 πιο σημαντικά χαρακτηριστικά που επηρεάζουν την τιμή πώλησης σύμφωνα με το μοντέλο $XGBoost$:



4.5 Gradient Boosting-ΑΠΟΤΕΛΕΣΜΑΤΑ

Hyper Parameter tuning

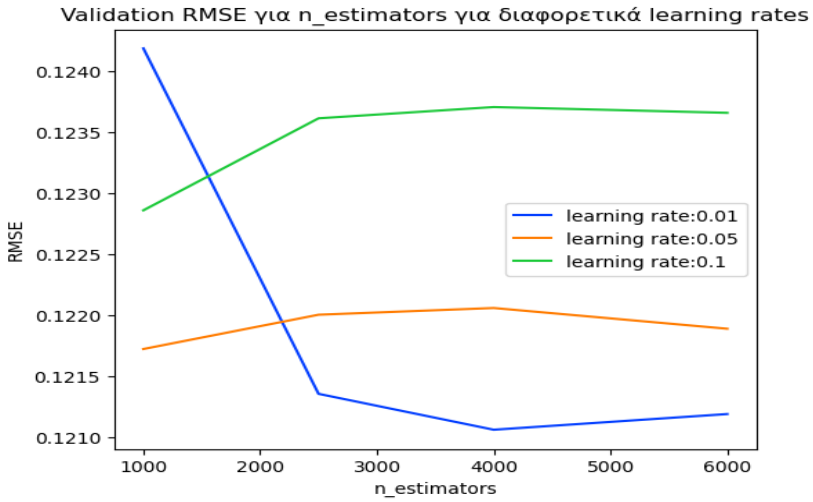
Εξετάσαμε διάφορους συνδυασμούς των παραμέτρων του μοντέλου, και επιλέξαμε αυτόν που δίνει το λιγότερο $RMSE$,με 3-Fold cross validation.Οι παράμετροι και οι τιμές που εξετάσαμε φαίνονται στον ακόλουθα πίνακα:

	Παράμετροι
max depth	$[3, 5, 10]$
learning rate	$[0.01, 0.05, 0.1]$
number of estimators	$[1000, 2500, 4000, 6000]$
min samples split	$[2, 5, 10]$

Ο καλύτερος συνδύασμος παραμέτρων φαίνεται στον ακόλουθο πίνακα :

	Παράμετροι
max depth	3
learning rate	0.01
number of estimators	4000
min samples split	2

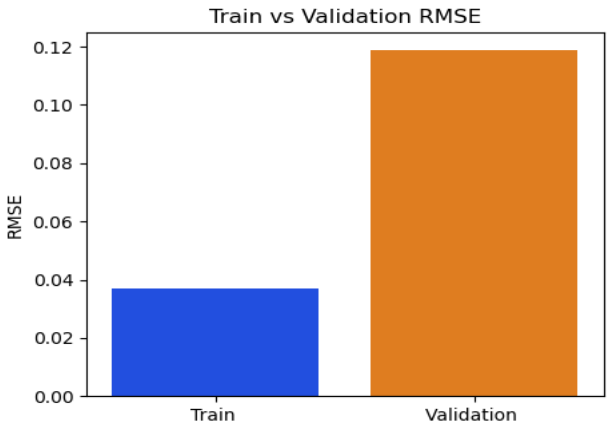
Επίσης παρακάτω φαίνονται τα αποτελέσματα για διάφορες τιμές των number of estimators και learning rate, με τις υπόλοιπες παραμέτρους σταθερές , με τιμές αυτές του καλύτερου συνδυασμού που περιγράφει ο παραπάνω πίνακας:



ΑΞΙΟΛΟΓΗΣΗ

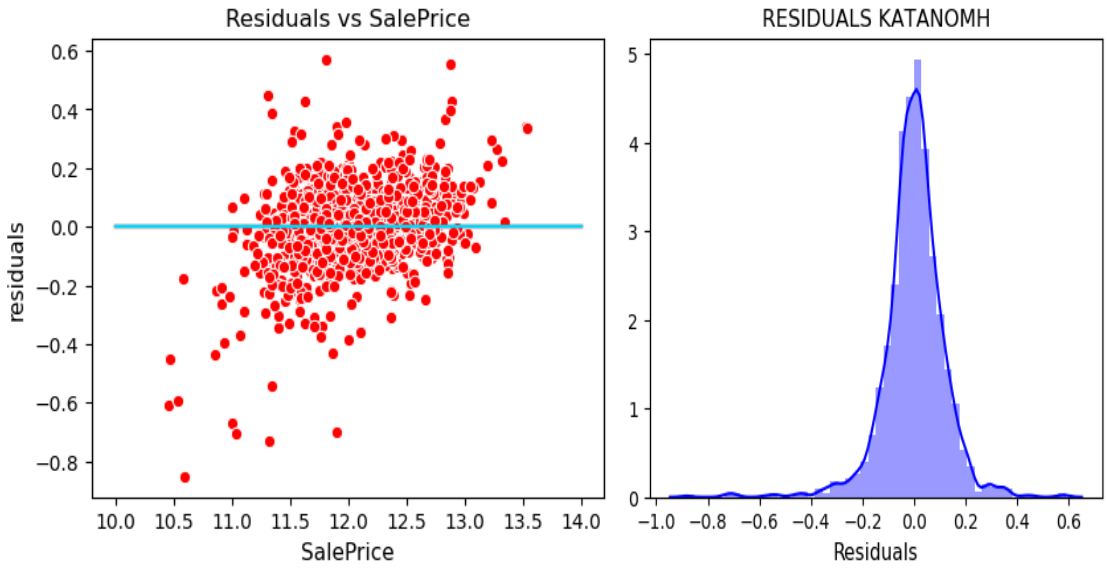
Η αξιολόγηση του μοντέλου με τι καλύτερες παραμέτρους έγινε με 10-Fold cross validation.Τα αποτελέσματα στα train και validation set ,φαίνονται παρακάτω:

	Train set	Validation set
RMSE	0.0437320	0.11798316



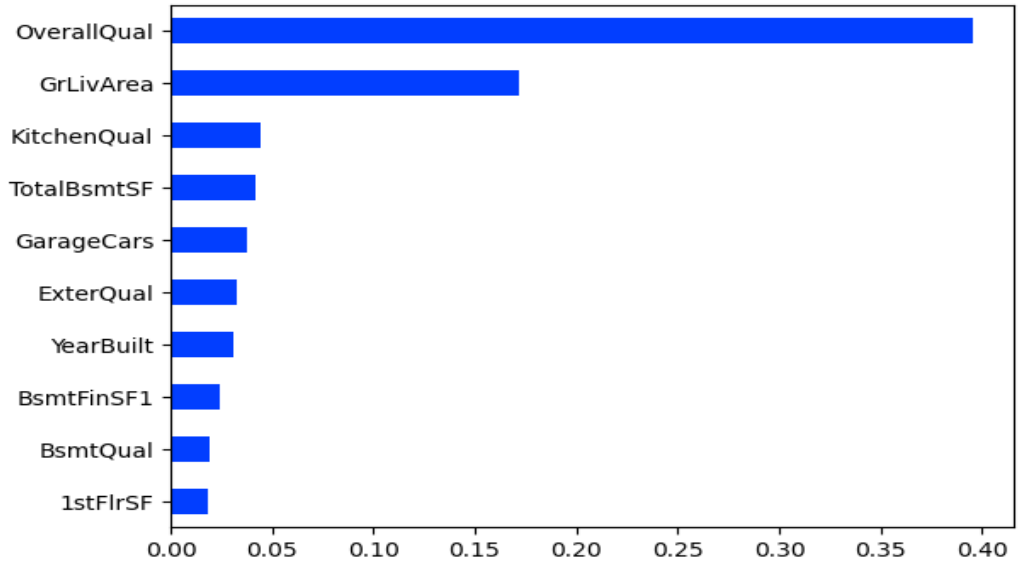
Το σφάλμα είναι μεγαλύτερο σε σχέση με του $XGBoost$,και το $overfit$ έχει αυξηθεί ελάχιστα.Αλλά τα αποτελέσματα είναι κοντά στο στόχο.

Παρακάτω φαίνεται η κατανομή των $residuals = y - \hat{y}$:



Feature Importance

Τα 10 πιο σημαντικά χαρακτηριστικά που επηρεάζουν την τιμή πώλησης σύμφωνα με το μοντέλο:



5 ΣΥΜΠΕΡΑΣΜΑΤΑ

Το αντικείμενο της εγρασίας είναι η πρόβλεψη της τιμής πώλησης ακινήτων μέσω μοντέλων μηχανικής μάθησης και η εύρεση σημαντιτκών χαρακτηριστικών τους.Για να το επιτύχουμε,χρησιμοποιήσαμε την τεχνική Gradient boosting και συγκεκριμένα δύο παραλλαγές της που βασίζονται σε δένρα αποφάσεων: XGBoost και Gradient Boosted trees.Το μοντέλο XGBoost,δίνει καλύτερα αποτελέσματα ως προς το $RMSE$,με λιγότερο $overfit$ στα δεδομένα εκπαίδευσης και είναι κοντά στο στόχο του 0.1 ως προς το $RMSE$ ανάμεσα στο λογάριθμο της τιμής πώλησης των ακινήτων και της προβλεπόμενης τιμής. Το $overfiting$ των μοντέλων,δηλώνει πως μπορεί να υπάρξει βελτίωση στην πρόβλεψη.Αυτό μπορεί να γίνει συλλέγοντας περισσότερα δεδομένα,εξερευνώντας περισσότερα μοντέλα και χρησιμοποιώντας ένα μετα-μοντέλο, βασισμένο σε αυτά.Επιπλέον,τα μοντέλα, μας δίνουν τη σημαντικότητα των χαρακτηριστικών στην τιμή πώλησης.Αρα μπορούμε να χρησιμοποιήσουμε τα πιο σημαντικά χαρακτηριστικά για να κάνουμε προβλέψεις και σε χώρες,οπως η Ελλάδα, που τα ακίνητα διαφέρουν με της Αμερικής,και όλα τα δεδομένα δεν είναι εύκολα διαθέσιμα.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

1. **Δεδομένα**:
   Πηγή: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
   Περισσότερες πληροφορίες : http://jse.amstat.org/v19n3/decock.pdf

2. **Github** :
   https://github.com/TEAM7-UOC-ML/SALEPRICE_PREDICTION

3. **Βιβλιοθήκες** :
   Gradient boosted trees SKLEARN: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
   XGBoost: https://xgboost.readthedocs.io/en/latest/

4. **Σχετικές Εργασίες** :
   https://escholarship.org/uc/item/3ft2m7z5

   https://www.perkinsml.me/ames-housing
   https://www.lexjansen.com/scsug/2017/MK29.pdf