RICHAR LUIS MARTINEZ CASTRILLON

Number Ryerson - 500930115

CKME 136 – CAPSTONE

PARCEL SALE PRICE PREDICTION WITH MULTINOMIAL REGRESSION, DECISION TREE, NAIVES BAYES AND RANDOM FOREST.

ABSTRACT.

To estimate the parcel sale price of goods in the market is obtained based on the supply and demand of goods, the environment of the economy, etc. However, there are some exogenous and endogenous determinants or variables that affect the price. In this study, the parcel sale price will be estimated by a set of variables including information about land, values, sales, abatements, and building characteristics (i.e.residential) by parcel this are considered as Hedonic Prices. The public database was obtained from the Office of Property Assessments in Allegheny County, City of Pittsburgh, Western Pa Data Center. https://catalog.data.gov/dataset

The object of the study is to create a model to predict the parcel sale price based on a set of characteristics related to the parcel.

Based on previous research in this field, it has been clear that there are several techniques used to predict the parcel sale price. One remarkable case are Supervised Machine Learning algorithms such as Multinomial Regression, Decision Tree, Naives Bayes, and Random Forest, after comparing the performance, the most accurate will be selected.

1. INTRODUCTION

One of the important aspects when making a decision to sell or buy a property is the estimated price, also the price of the land is relevant, in the sense that the land is considered a heterogeneous good, meaning that there are own characteristics of the territory such as “Propertycity”, “Propertyaddress”, “Usedesc”, “Lotarea”, among others, there are also other characteristics related to the building like: “Extfinish_Desc”, “Roofdesc”, “Basementdesc”, “Condition”, “Totalrooms”, “Bedrooms”, and others.

All these factors can affect the sale or purchase price of the territory, and the use techniques and tools for prediction will be adequate only when the obtained data is based on historical transactions (benchmark).

For this purpose there are companies, real state groups, and web pages specialized in the prediction of property values, one of the more reputable ones is Zillow Group, as referenced by Nguyen (2018).

In the Real State market, as in other one, the objective would be for the estimated price to be as close as possible to the real market price or equilibrium price, in which buyers get a fair price and the sellers maximize their profits.

The prices of the territories will be obtained from Allegheny County, PA. USA, for the year 2019. Many studies in the past, have applied prediction models and the results are not always accurate, due to there is no 100% reliable methodology stablished, in that sense, there are problems related to ambiguity , since there is a margin Probability of obtaining statistically non-significant data, however, the result obtained after an study like this, allows the merchants to have a baseline price, that will be useful when making decisions.

As part of Scope of this project, two machine learning algorithms will be created: the first is the estimation of a multinomial regression model, based on the theory of hedonic prices, initially used in the United States of America and that was formally proposed by Rosen (1974), and that has been widely applied to the Real Estate market. The second method being applied is a regression tree, at the end, their performance will be compared to determine which model is more useful and accurate to make the forecast.

For both approaches, it will be necessary to make a selection process of attributes that contribute or significantly impact the sale price of the territory, in the literature it is common to find dimensionality reduction techniques, when there is a large volumes of data to be analyze. For this particular study, the methodology selected will be based on information Gain (Fselector).

At the end of the project, the results will be evaluated and suggestions for future research will be proposed.

1.1. LITERATURE REVIEW

Nguyen (2018), based his study on the price prediction in 5 different counties of the USA, using internal characteristics, such as number of bathrooms, bedroom, etc. also external characteristics such as public schools, with data from 1,457 houses obtained from Zillow, Trulia and Redfine webpage, the models that he applied were Linear Regression based on the concept of Hedonic Prices and machine learning algorithms such as Random Forest (RF) and Support Vector Regression (SVR).

The author found in his study, four significant attributes: number of bathrooms, assessment, listed price and comparable houses' sold price and managed to reduce the overestimation and underestimation of the house ratio from 3:2 to 1:1 resulting in a success of the research.

Regarding the models applied he compared the obtained scores with Zillow's prediction score baseline, and determined that the SVR model is more accurate in predicting the sale price than the baseline score in Hunt County (TX). The RF model found values very close to the baseline in Cowlitz (WA) and Montgomery (IL) and becoming the same in Upson County (GA).

Sonia del Rey Simeón (2017), also supported her work, in the application of the model of the function of hedonic prices, widely used by researchers in the literature, in her study she took data from 935 homes in Seville as a sample, and the main result of the analysis showed that the price of housing depends significantly on the number of square meters of construction and the area of the location.

Luis M. Vilches-Blázquez (2017), in his work assumed the territory as an asset or factor of production that has a value in the market (price), the author took characteristics of the territory in two municipalities: Ibagué and Ortega in the department of Tolima, Colombia, and based on this, he applied machine learning techniques such as algorithms, grouped into three categories: Classification (decision tree and Naïve Bayes), Regression (linear and linear) and Clustering (K-Means), however, the results obtained were considered as exploratory in nature and cannot be taken into consideration when making inferences about the reality of the territory.

Ashray Kakadiya, Khushal Shingala, Shiv Raj Sharma (2018), carried out a house price prediction work in Ames, Iowa for the period between 2006 to 2010, where they used different regression techniques Lasso, Ridge, SVM regression, and Random Forest regression and Classification, Naive Bayes, logistic regression, SVM classification, and Random Forest classification. Finally, in their work they found that living area square feet, material of the roof, and neighborhood were the best predictors of the price of housing.

Similarly, Gerald Muriuki (Oct 11, 2017), used data that containedprices and features of residential houses sold from 2006 to 2010 in Ames, Iowa. In order to estimate a linear regression model capable of predicting the sale price of a house.

George Lever D. Performed a conceptual work and possible applications of the Hedonic price model, in his development George made a theoretical and mathematical explanation of the model, and argued that this type of model is applicable to real estate, and other types of goods that are affected by externalities or characteristics of goods that are not their own.

Vincenza Chiarazzoa, Leonardo Caggiania, Mario Marinellia and Michele Ottomanellia, apply an Artificial Neural Network (ANN) model, in the city of Taranto (Italy), the most relevant factors to determine the price of a property are: quality of the environment and some characteristics of the property.

As detailed in the literature review, the application of hedonic prices has been widely used in the prediction of prices of goods that from their own characteristics are also affected by exogenous or external variables.

Additionally, the authors of those studies suggest the application of Machine Learning Algorithms techniques in order to improve and obtain more accurate predictions every time.,

1.2. ANALYSIS APPROACH

The work development will approach from two methodologies to predict the sale price of Land in Allengency county, USA

The Scope of the work will be based on the use of two methodologies to predict the sale price of land in Allengency County, USA.

1.2.1. Hedonic Prices Model

Hedonic Prices Model was proposed by Rosen in 1974, and is widely used in the estimation of prices and demand of real estate. This type of model is characterized by estimating a price function based on characteristics. These characteristics of implicit attributes are categorized or grouped, as an example territory characteristics, building, environmental factors, local developments, public services, among others.The mathematical expression is:

Ecuacion

$P = f(Xi; v)$

i: 1,2,3……n.

Where: i represents each of the categories or group of characteristics.

P: Prices(Category Attibute).

f: Function Arguments

X: Attribute group for each categories.

v: Coefficient of each attribute in the function.

log(prob=category(i)/prob=category(n))= B0 + B1x1 + B2x2 + .. + Bnxn + ei

Where:

P: Price to estimate(Category)

B= Coefficient of independent attributes.

x= Attributes.

ei= Random error.

i:1,2,3,…………n

1.2.2. Decision Tree.

Decision tree algorithm are used for classification tree or regression tree purposes, but the difference between them is the type of dependent variable to be predicted. however in classification, the dependent variable must be categorical, regardless, it under this argument that it will be used the land sale price attribute to predict in this project, which in this case is numerical and discrete attribute, it will be converted to categorical to build the model.

1.2.3. Random Forest

The Random Forest algorithm assembly, can be used in data mining, works by creating multiple numbers of trees and then combines the results obtained in each tree, is based on the concept of information gain, in each tree built, classifying each node where it is You get more information gain.

Random Forests does not select all data points and variables in each of the trees. It randomly displays data points and variables in each of the trees it creates and then combines the output at the end. Eliminates the bias that a decision tree model could introduce into the model.

1.2.4. Naives Bayes (Classifier)

The Naive Bayes Algorithm, is a probabilistic classifier based on Baye's theorem, mathematically:

$P(B \mid A) = P(B) * P(A \mid B) / P(A)$

The algorithm makes a strong assumption about the data having features independent of each other while in reality, they may be dependent in some way. it assumes that the presence of one feature in a class is completely unrelated to the presence of all other features. If this assumption of independence holds, Naive Bayes performs extremely well and often better than other models.

The study aims to predict the sale price using a set of features as independent variables.

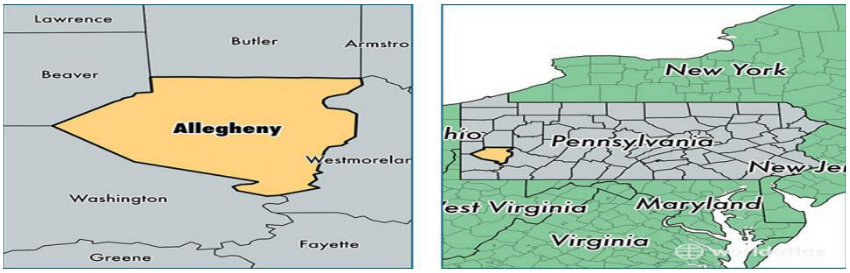
Aproximation.

PREDICT SALEPRICE		
PRE-PROCESSING	**DATA PREPARATION**	**BUILD MODELS**
Review and data type. **Univariate Analysis** Histograms , plot and Barplots. Treatment outliers and missing values. Definition target attribute. **Bivariate Analisys.** Correlation. Cero Variance. Chi-squered test Data cleaning	**Feature Selection** Identification of key attributes Gain(Using FSelector). Select Attribute with high information gain.	Split data in training 70% and test set 30%. Build Decision tree. Then use Run K-folds Cross Validation model. Build a model to multinomial Logistic regression. Build Random Forest Model. 3 models tuning parameters. Build Naives Bayes Clasifiers Get confusion matrix all of them.

2.DATA EXPLORATION

2.1 Data Description

The map of Allegheny County, shows our sample of location from where the database is taken, object of analysis:



Source:Google.com

For the development of this work, a database was obtained from the Office of Property Assessments in Allegheny County, City of Pittsburgh, Western Pa Data Center. The website to access the data is located at *https://catalog.data.gov/dataset*.

This database contains information on the characteristics of property such as: land, values, sales, Abatements, Propertycity, Propertyaddress, Usedesc, Lotarea, Taxes, among others, and some characteristics of the buildings such as: Extfinish_Desc, Roofdesc, Basementdesc, Condition, Totalrooms, Bedrooms etc.

The sample contains a total of 86 attributes and 579,186 or observations, which are distributed as a factor and numerical attributes.

of the Data Dictionary is shown as part of the Appendix 1 of this project, taken from the source mentioned above.

A first approximation to the description of the data set was the summary, it is clearly observed according to the definition of data dictionary attribute that there is a group of variables that have the same meaning and represent a single value. To simplify the analysis, one of each pair of them was eliminated. Below, there is a list of the referred variables:

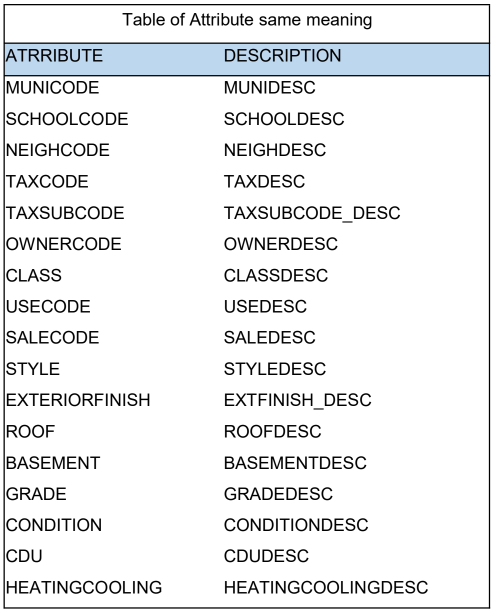


Table of Attribute same meaning

ATRRIBUTE	DESCRIPTION
MUNICODE	MUNIDESC
SCHOOLCODE	SCHOOLDESC
NEIGHCODE	NEIGHDESC
TAXCODE	TAXDESC
TAXSUBCODE	TAXSUBCODE_DESC
OWNERCODE	OWNERDESC
CLASS	CLASSDESC
USECODE	USEDESC
SALECODE	SALEDESC
STYLE	STYLEDESC
EXTERIORFINISH	EXTFINISH_DESC
ROOF	ROOFDESC
BASEMENT	BASEMENTDESC
GRADE	GRADEDESC
CONDITION	CONDITIONDESC
CDU	CDUDESC
HEATINGCOOLING	HEATINGCOOLINGDESC

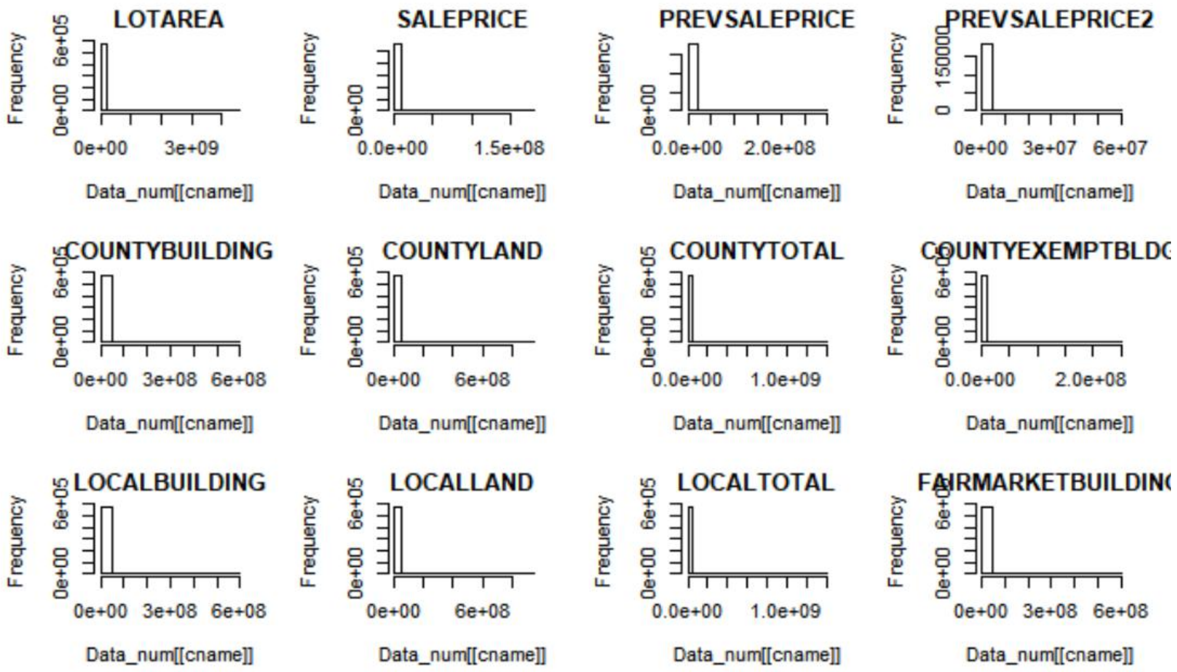
The dataset contains attributes with more than 95% of missing values, for the purpose of this study, they were discarded: Propertyfraction, Propertystate, Propertyunit, Taxsubcode, Taxsubcodedesc, Farmsteaflag, Cleangreen, Abatemanflag, Altid, Asofdate, Taxyear.

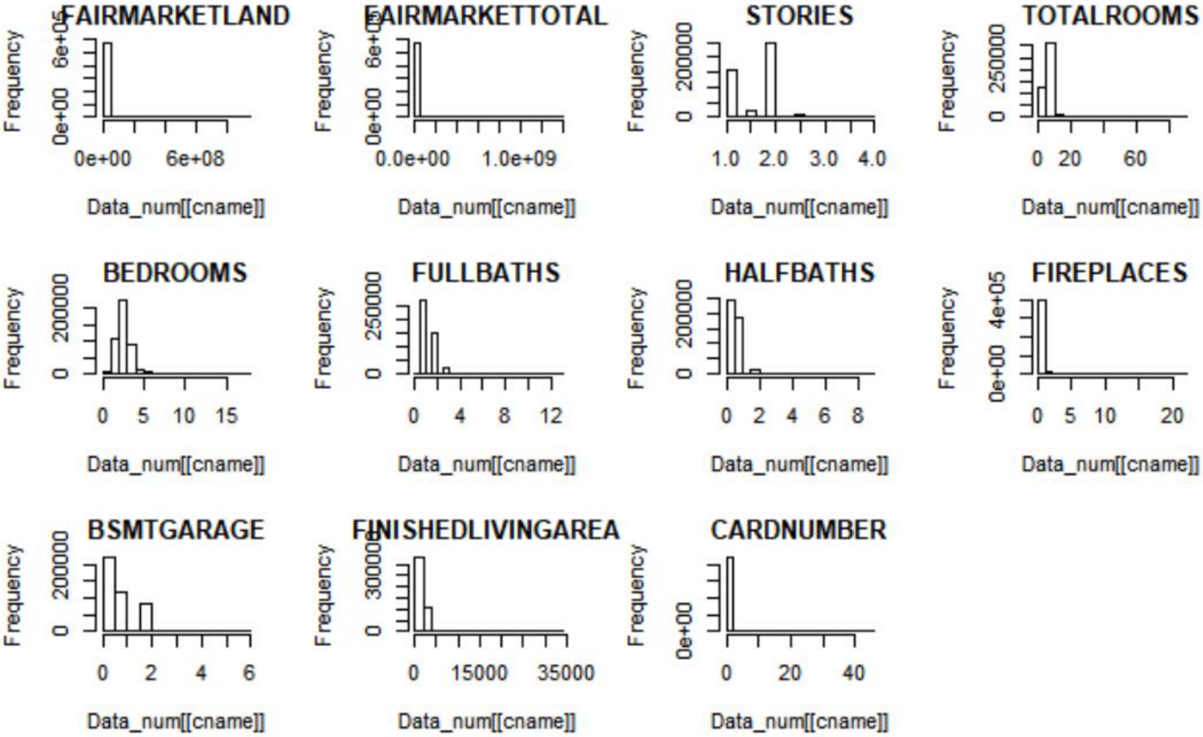
Descriptive attributes that are not providing real insights were also discarded :Legal1, Legal2, Legal3, DeepPage, Deebook, Changenoticeaddress1, Changenoticeaddress2, Changenoticeaddress3, Changenoticeaddress4, Homesteadflag (It contains a single factor), Propertyhousenum, Propertyzip, Countyexemptbldg And Dates.

Once the first analysis was carried out, a reduction in dimensionality was obtained, moving to 40 variables with 579,186 observations. Simultaneously, the Propertyaddress, Propertycity, Munidesc, Usedesc attributes were converted to their correct type of factor-to-character data.

In a graphical analysis of the statistical behavior of all variables, the data set was divided into numerical and non-numerical attributes.

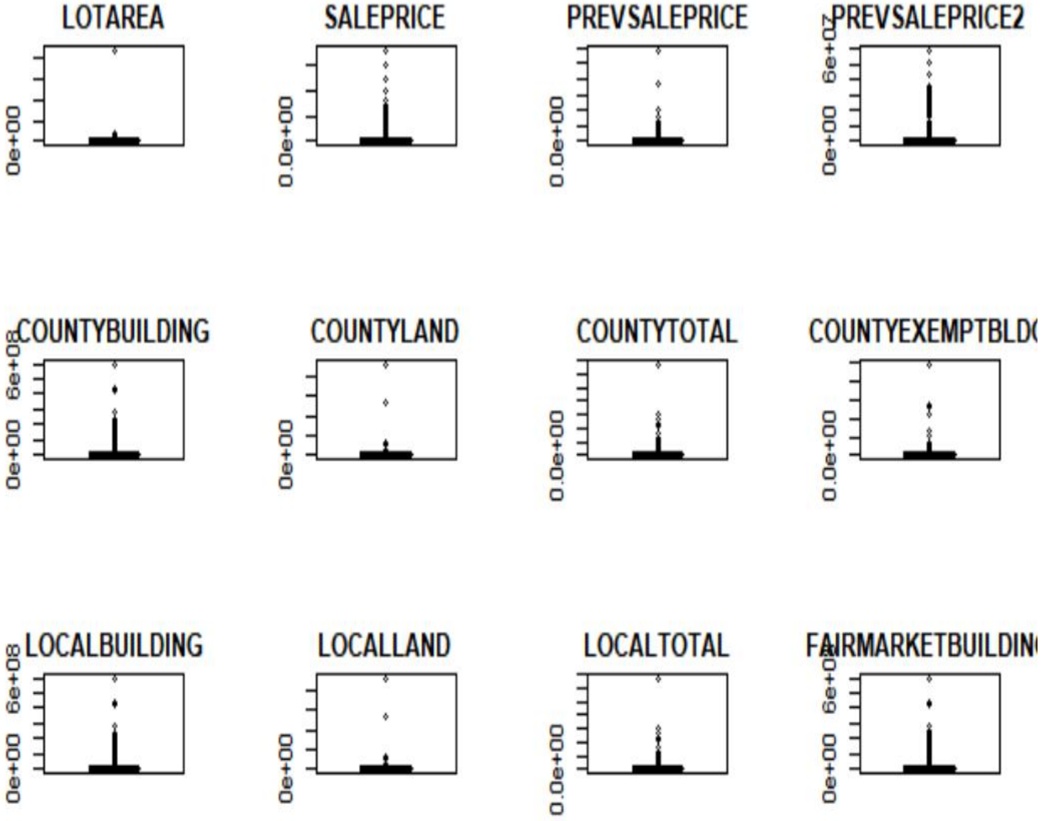
The histograms of the numerical attributes show below, are highly skewed to the left in all the cases, concentrating their values near the average of each variable with a small dispersion of data in the sample, therefore a normality test is not considered necessary since it is observed a statistical bias in the distribution.

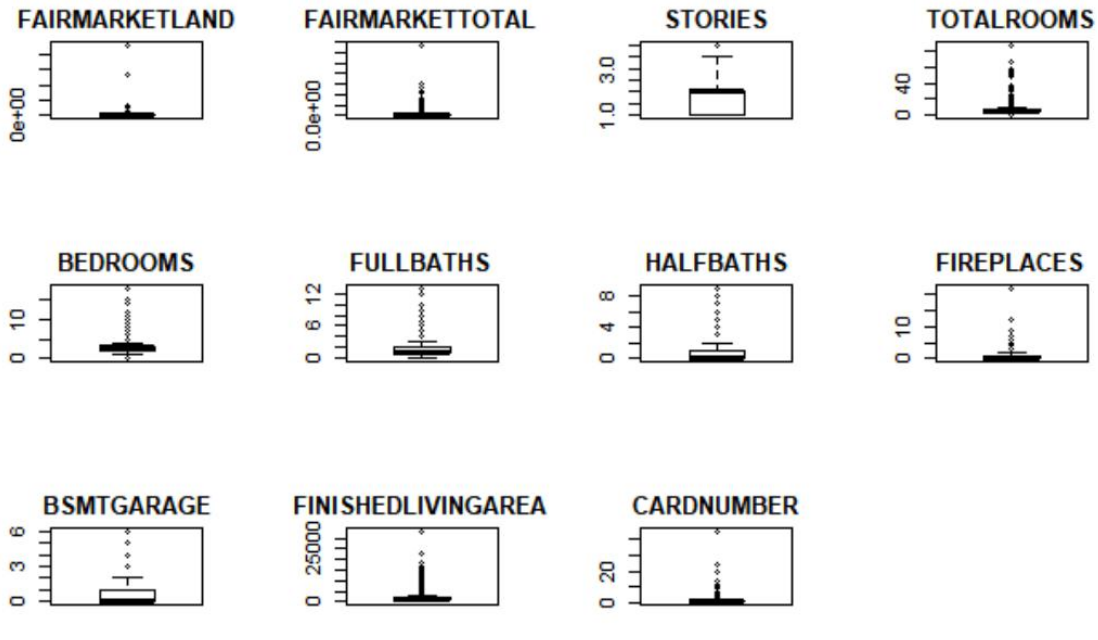




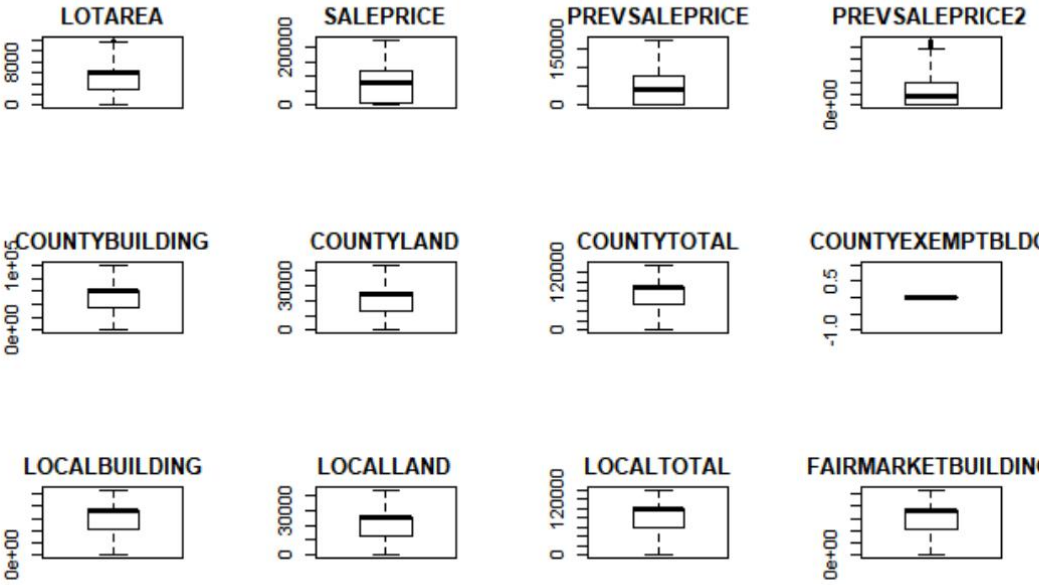
The visualization of the Boxplot, shows the presence of Outliers in all the numerical attributes, the presence of Outliers in the sample can cause bias in the prediction of the variable to be predicted, reflecting a reduction in model accuracy, under this assumption, the Outliers were charged with the median as a measure of central tendency, avoiding the loss of information, see next .

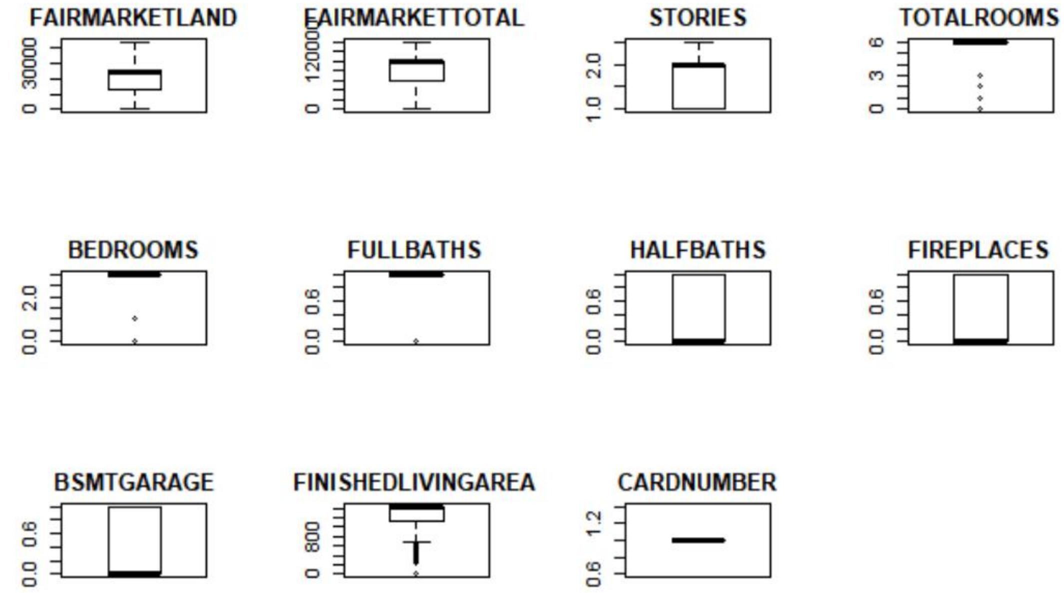
Graphic representation with Outliers





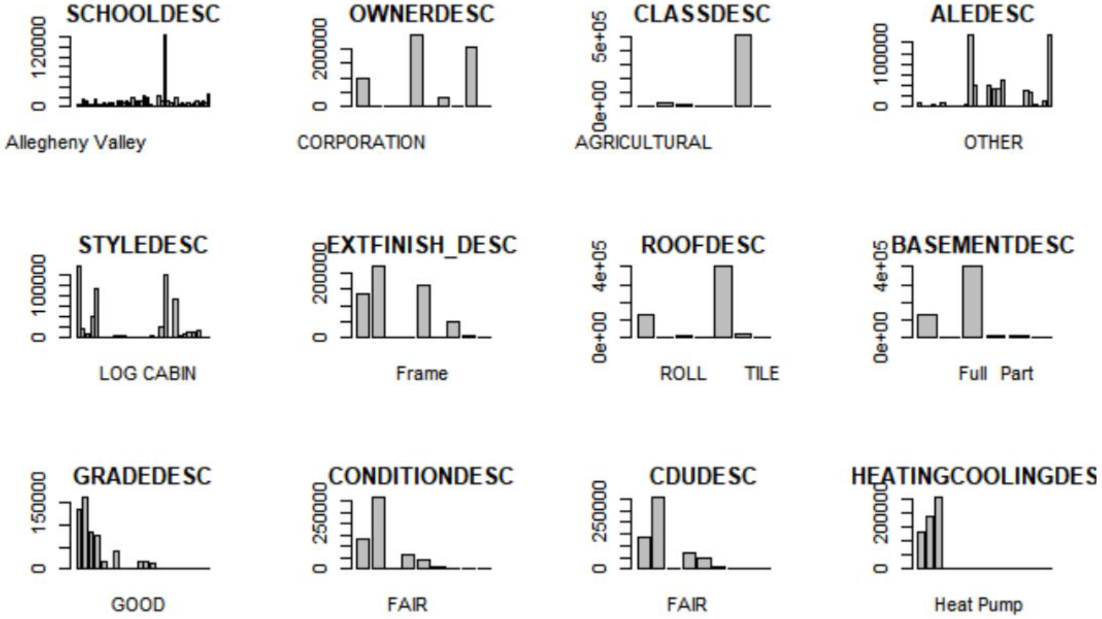
Graphic representation with imputed Outliers





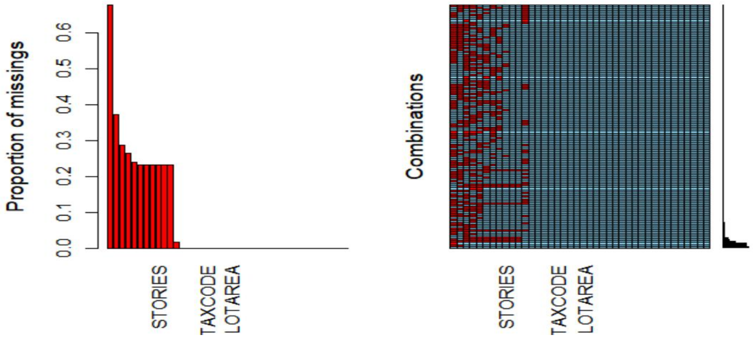
The variables of type factor, were analyzed with Barplots, the results, data dispersed in the variables Schooldesc, Ownerdesc, Saledesc, Styledesc, Ext_finishdesc and high degree of concentration in variables such as Roofdesc, Basemendesc, and Classdesc towards one of each attribute factors and finally bias to the left in the variables Gradedesc, Conditiondesc and Cdudesc.

Graphic representation

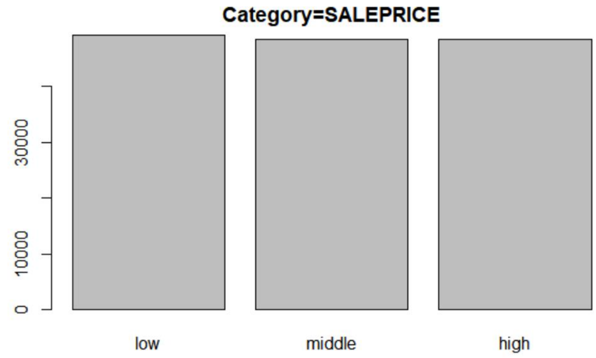


The data under study could verify that there were missing values in the data frame, a graphical representation showed an average of 23% of missing value in all attributes, which were omitted from the sample to simplify the analysis.

Graphic representation.



For the application of machine learning algorithms, the sale price is selected as the response variable to predict in the model, this discrete type variable is converted into a three-level category: low, medium and high, using Quantail as a tool.Graphic representation.



In the bivariate analyzes, first applying Pearson correlation test to the numerical attributes, Fairmarketbuilding, Fairmarkettotal, Localland, Fairmarketland, Localtotal, Localbuilding, Countytotal obtained a high correlation> 0.8, in addition, in the zero variance analysis the variables Countyexemptbldg, Cardnumber, Countybuilding, Bedrooms, Fullbaths, Finishedlivingarea contain a variance close to or equal to 0, assuming they could be eliminated.

Secondly, in the factor-type variables the chi-square test, applied to determine the independence of the variables, the value P = 0.015 between Taxcode and Roofdesc and the value P = 2.2 exp -16 between Roofdesc and Basementdesc, obtained are lower that <0.05 rejecting the hypothesis of null independence of attributes.

3.FEATURE SELECTION

3.1 Information Gain

When working with machine learning tools, the selection of attributes is relevant to find the ones with greater predictive power in the model, it is the selection based on the entropy of the attributes, considering those with the highest information gain.

As a result, 9 attributes were selected out of 27: Lotarea, Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalroomshal, Fbaths , Fireplaces, Bsmtgarage in addition to the variable to predict Category (Saleprice).

4.BUILD THE MODELS AND INTERPRETATION

During the process of creation of the model, a set of the 10 variables with the higher information gain were used.

4.1 Build Decision Tree Model

For the development of the decision tree model, in the first instance the formula to be used to predict the model was defined, there was chosen Category = Saleprice, three levels: low, middle and high, as the response or dependent variable and Lotarea , Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalroomshal, Fbaths, Fireplaces, Bsmtgarage, as independent variables.

The Decision tree model, for this study acts as a classifier, so that in order to obtain the projection, the data set T_Data was split, into 70% into training set with a total of 102,183 observations and 30% test set with 43,794 observations.

The model was initially run in the training set and after the test set, then the prediction of the model in the test set was calculated, a good tool to do this is the confusion matrix that allows the researcher to revise the precision obtained and also the classification error.

For the model, it was 58.23%, which represents the observations that were correctly assigned to the level of its category, represented by the main diagonal of the confusion matrix divided by the total of the sample and the error = 1- accuracy = 1-58.23 = 41.77%, which represent the observations that were not correctly assigned to their level.

On the other hand, K-folds cross validation (10) is applied in the training set, to determine if there is a variation or improvement in the accuracy of the model, the average error obtained equal to 44.90%, with a 55.1% accuracy.

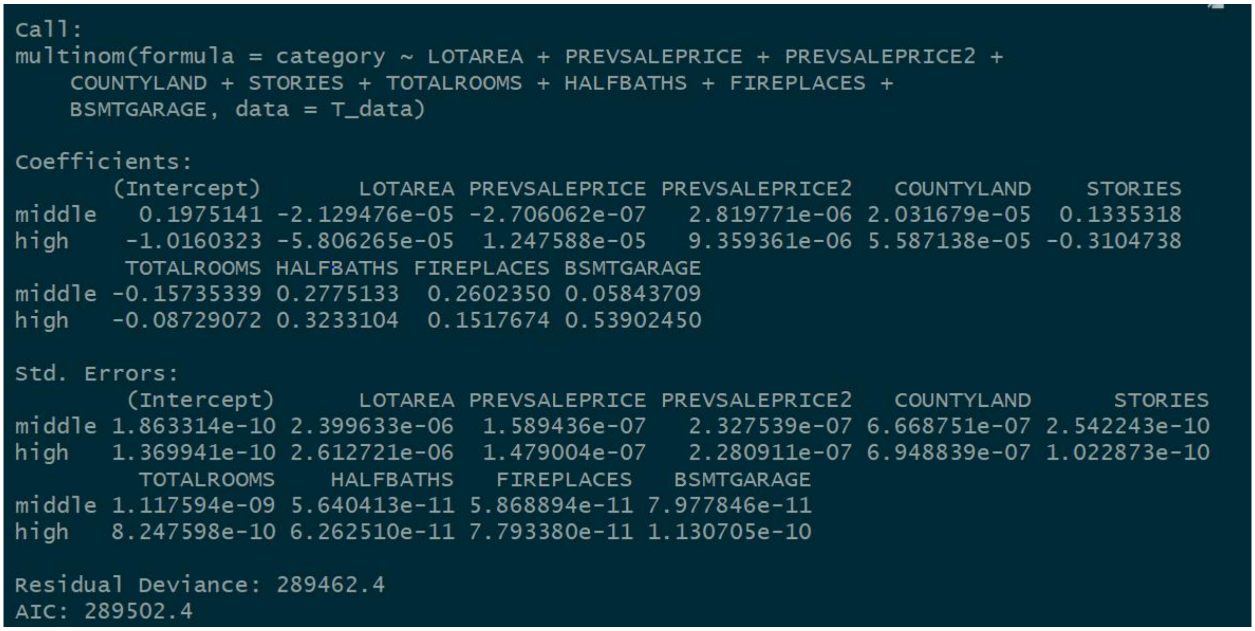
4.2 Multinomial Logistic Regression

The objective of the logistic regression model is to predict the probability to assign an observation to a espesific category based on one of the categories, there are three levels in the study, low is taken as base and two levels middle and high are predicted in the Regression.

To proceed with the development of the model it is necessary to convert our category = saleprice variable to a factor.

In the model the “low” category was chosen as the base group, ref = 1 and the database T_data is used. Next, the formula of the multinomial regression model to predict is established, as a dependent variable the category = saleprice and independent variables were chosen: Lotarea, Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalroomshal, Fbaths, Fireplaces, Bsmtgarage.

The results obtained in the regression are divided into two parts, the first corresponds to the coefficients and the second to the errors



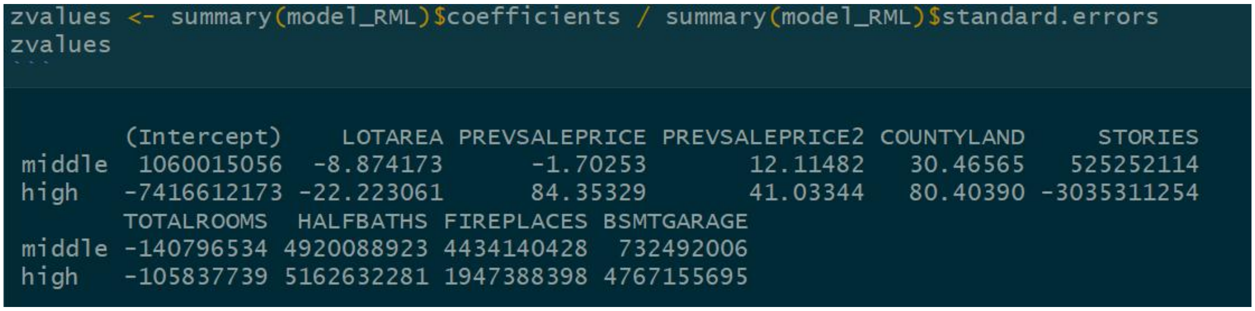
```
Call:
multinom(formula = category ~ LOTAREA + PREVSALEPRICE + PREVSALEPRICE2 +
    COUNTYLAND + STORIES + TOTALROOMS + HALFBATHS + FIREPLACES +
    BSMTGARAGE, data = T_data)

Coefficients:
       (Intercept)       LOTAREA PREVSALEPRICE PREVSALEPRICE2   COUNTYLAND    STORIES
middle   0.1975141 -2.129476e-05 -2.706062e-07   2.819771e-06 2.031679e-05  0.1335318
high    -1.0160323 -5.806265e-05  1.247588e-05   9.359361e-06 5.587138e-05 -0.3104738
        TOTALROOMS HALFBATHS FIREPLACES BSMTGARAGE
middle -0.15735339 0.2775133  0.2602350 0.05843709
high   -0.08729072 0.3233104  0.1517674 0.53902450

Std. Errors:
        (Intercept)      LOTAREA PREVSALEPRICE PREVSALEPRICE2   COUNTYLAND      STORIES
middle 1.863314e-10 2.399633e-06  1.589436e-07   2.327539e-07 6.668751e-07 2.542243e-10
high   1.369941e-10 2.612721e-06  1.479004e-07   2.280911e-07 6.948839e-07 1.022873e-10
         TOTALROOMS    HALFBATHS   FIREPLACES   BSMTGARAGE
middle 1.117594e-09 5.640413e-11 5.868894e-11 7.977846e-11
high   8.247598e-10 6.262510e-11 7.793380e-11 1.130705e-10

Residual Deviance: 289462.4
AIC: 289502.4
```

With the coefficients obtained, the formula can be written as follows:

Log(*Pr(middle=2)Pr(low=1)****)=** 0.1975 - 2.1 exp-5***(****Lotarea) - 2.7 exp-7 (Prevsaleprice) + 2.8 exp-6 (Prevsaleprice2) + 2.exp-5(Countyland) + 0.13(Stories) - 0.15(Totalroomshal) + 0.27(Fbaths) + 0.26(Fireplaces) + 0.058B(smtgarage)*

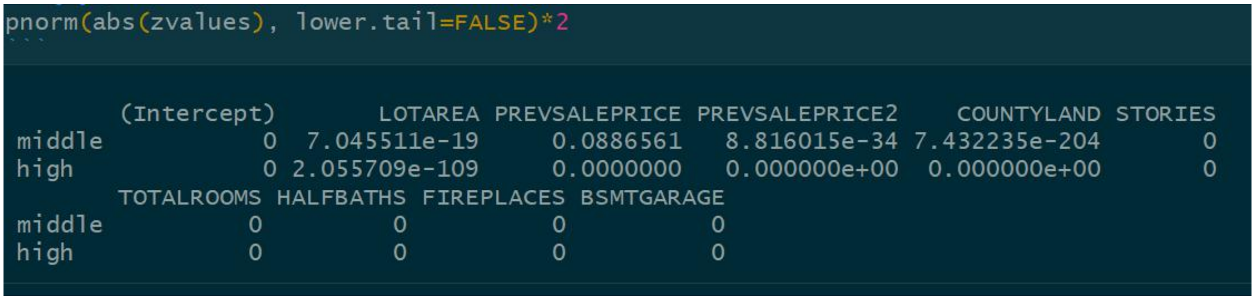
Log(*Pr(high=3)Pr(low=1)****)=** -1.01 - 5.8 exp-5***(****Lotarea) + 1.24 exp-5 (Prevsaleprice) + 9.3 exp-6 (Prevsaleprice2) + 5.58 exp-5(Countyland) -0.31(Stories) -0.08(Totalroomshal) + 0.32(Fbaths) + 0.15(Fireplaces) + 0.53 (Bsmtgarage)*

The z-values and the p-values. To get the z-values divide the coefficients by the standard errors, these are obtained to evaluate the statistical significance of the coefficients in the model, the following is the result obtained:



```
zvalues <- summary(model_RML)$coefficients / summary(model_RML)$standard.errors
zvalues
```

```
        (Intercept)    LOTAREA PREVSALEPRICE PREVSALEPRICE2 COUNTYLAND     STORIES
 middle  1060015056  -8.874173      -1.70253       12.11482   30.46565   525252114
 high   -7416612173 -22.223061      84.35329       41.03344   80.40390 -3035311254
        TOTALROOMS  HALFBATHS FIREPLACES BSMTGARAGE
 middle -140796534 4920088923 4434140428  732492006
 high   -105837739 5162632281 1947388398 4767155695
```

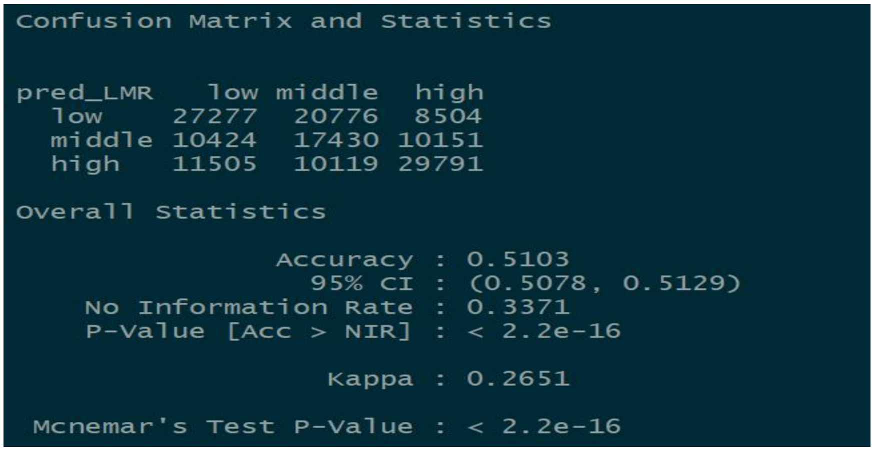


```
pnorm(abs(zvalues), lower.tail=FALSE)*2
```

```
        (Intercept)       LOTAREA PREVSALEPRICE PREVSALEPRICE2    COUNTYLAND STORIES
 middle           0  7.045511e-19     0.0886561   8.816015e-34 7.432235e-204       0
 high             0 2.055709e-109     0.0000000   0.000000e+00  0.000000e+00       0
        TOTALROOMS HALFBATHS FIREPLACES BSMTGARAGE
 middle          0         0          0          0
 high            0         0          0          0
```

It can be seen that all the P - values obtained from each of the variables were found to be <0.05, so it can be affirmed that the coefficients are statistically significant, and these variables can be used to predict the probability of the model. Note that variables such as: countyland stories, presaleprice2, Halfbat, Fireplaces and Bsmtgarage, have a positive impact on the probability of correctly assigning an observation.

The confusion matrix shows an accuracy of 51.035, detailed in the following output:

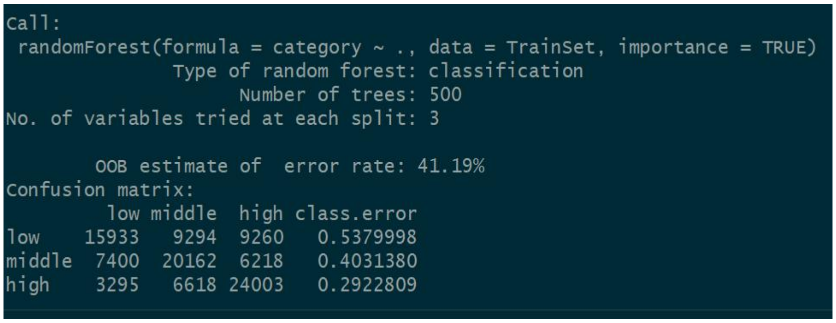


```
Confusion Matrix and Statistics


pred_LMR    low middle  high
  low     27277  20776  8504
  middle  10424  17430 10151
  high    11505  10119 29791

Overall Statistics

               Accuracy : 0.5103
                 95% CI : (0.5078, 0.5129)
    No Information Rate : 0.3371
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2651

 Mcnemar's Test P-Value : < 2.2e-16
```

4.3 Random Forest

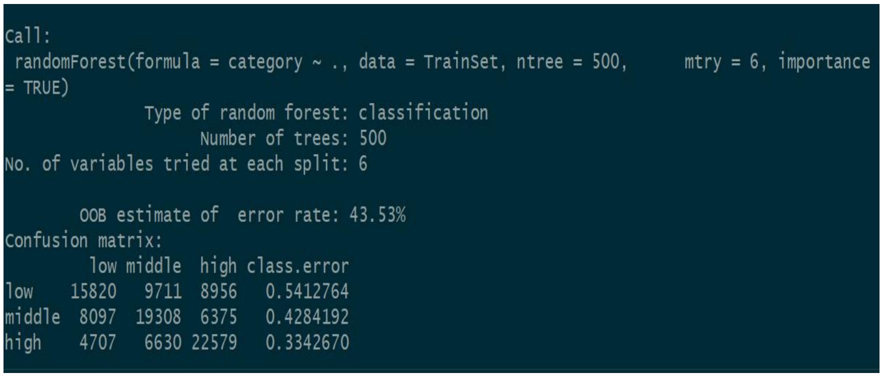
To build the Random forest model, the T_data was divided into training and test set in a ratio of 70:30. The dependent variable is category = saleprice, as following independent variables were used: Lotarea, Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalroomshal, Fbaths, Fireplaces, Bsmtgarage.

The first model is created with the default parameters with 500 trees and the number of variables treated in each division is 3, here an error of 41.19% was obtained.



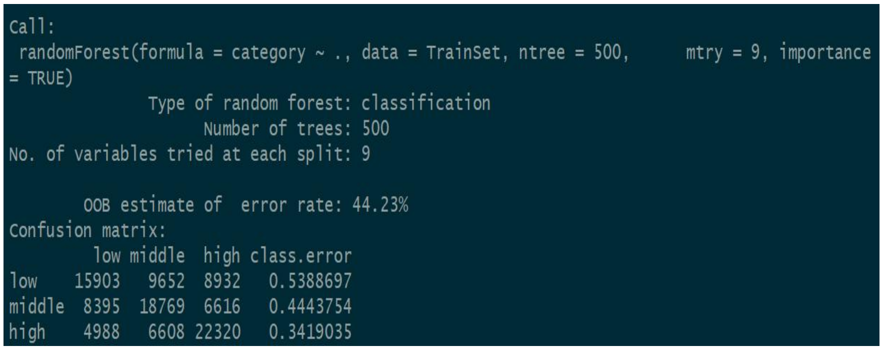
```
Call:
 randomForest(formula = category ~ ., data = TrainSet, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 41.19%
Confusion matrix:
         low middle  high class.error
low    15933   9294  9260   0.5379998
middle  7400  20162  6218   0.4031380
high    3295   6618 24003   0.2922809
```

Then the number of parameters mtry = 6 attributes for the following model was changed. It is observed that the error increases from 41.19% to 43.53%.



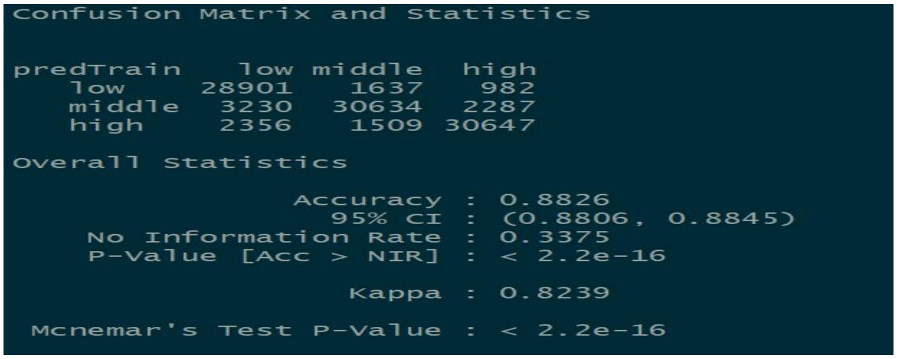
```
Call:
 randomForest(formula = category ~ ., data = TrainSet, ntree = 500,      mtry = 6, importance 
= TRUE) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 43.53%
Confusion matrix:
         low middle  high class.error
low    15820   9711  8956   0.5412764
middle  8097  19308  6375   0.4284192
high    4707   6630 22579   0.3342670
```

Finally it is changed to Random Forest model with 9 attributes, where the error goes to 44.23%..



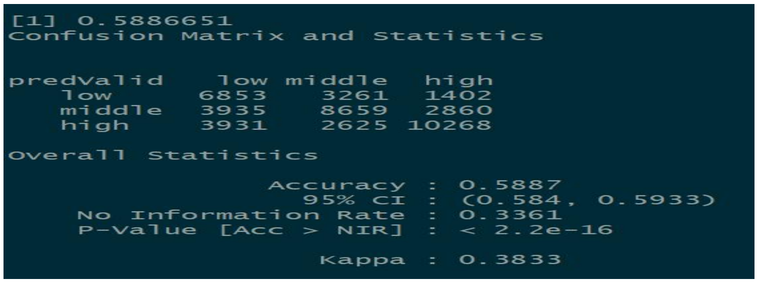
```
Call:
 randomForest(formula = category ~ ., data = TrainSet, ntree = 500,      mtry = 9, importance 
= TRUE) 
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 9

        OOB estimate of  error rate: 44.23%
Confusion matrix:
         low middle  high class.error
low    15903   9652  8932   0.5388697
middle  8395  18769  6616   0.4443754
high    4988   6608 22320   0.3419035
```

After of compare all the possible scenarios, the one with the smallest error rate was selected, in this case it was model1.



```
Confusion Matrix and Statistics

predTrain   low middle  high
   low    28901   1637   982
   middle  3230  30634  2287
   high    2356   1509 30647

Overall Statistics
                                          
               Accuracy : 0.8826          
                 95% CI : (0.8806, 0.8845)
    No Information Rate : 0.3375          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.8239          
                                          
 Mcnemar's Test P-Value : < 2.2e-16       
```

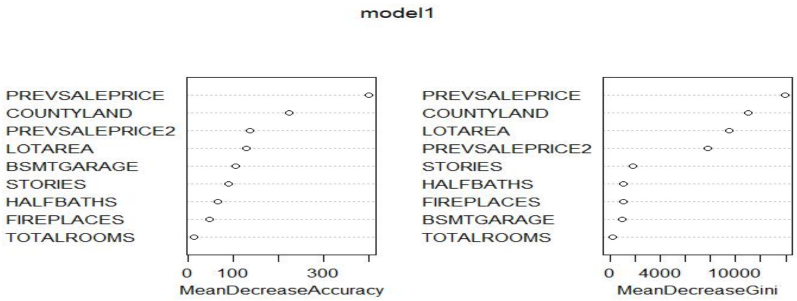
Predicting on Validation set model1 and Checking classification accuracy, you get:



```
[1] 0.5886651
Confusion Matrix and Statistics

predvalid   low middle  high
   low     6853   3261  1402
   middle  3935   8659  2860
   high    3931   2625 10268

Overall Statistics

               Accuracy : 0.5887
                 95% CI : (0.584, 0.5933)
    No Information Rate : 0.3361
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3833
```

The accuracy obtained in the training set was 88.26%, meaning that the number of misclassified observations was 11.74%. However, predicting in test sample it was only 58.87% of accuracy, of the observations correctly classified in the category = saleprice.

Graphically, MeanDecreaseAcurrancy and MeanDecreaseGini show how the accuracy falls on average for each of the variables.

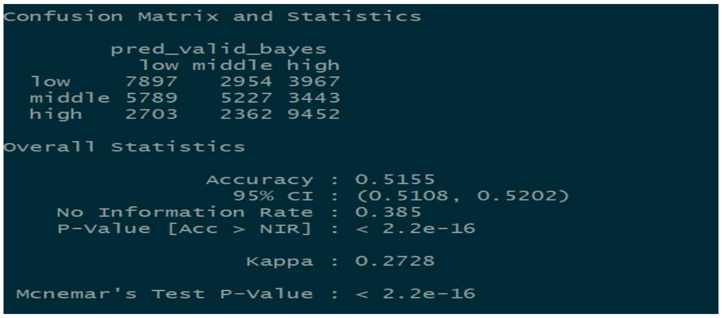


According to the results obtained, it is evident that the formulated model1 presents an accuracy of 58.87%, this model can be considered as the best classifier, its error rate is around 41.13%.

4.4 Naive Bayes Classifier

To build the Naives Bayes Classifier, the T_data was divided into training and test set in a ratio of 70:30. The dependent variable is category = saleprice and as independent variables, Lotarea, Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalroomshal, Fbaths, Fireplaces, Bsmtgarage.

Initially the model is running in the train dataset and then predicted in the test dataset, the output shows a 51.55% accuracy, as detailed in the following confusion matrix:



```
Confusion Matrix and Statistics

        pred_valid_bayes
          low middle high
  low    7897   2954 3967
  middle 5789   5227 3443
  high   2703   2362 9452

Overall Statistics

               Accuracy : 0.5155
                 95% CI : (0.5108, 0.5202)
    No Information Rate : 0.385
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2728

 Mcnemar's Test P-Value : < 2.2e-16
```

5. CONCLUSION

To observe the results obtained in the prediction of the Category (Saleprice) variable, in Allegheny County, City of Pittsburgh with the application of machine learning algorithms, it is shown in summary in the following table:

Measurement	Multinomial Logistic Regression	Naive Bayes	Decision Tree	Random Forests
Accuracy	51.03	51.55	58.23	58.87
Error	48.97	48.45	41.77	41.13
k-folds cross validation - Accuracy			55.1	
Error			44.9	

With the results presented above, Random forest was the one that obtained the best result, when predicting the saleprice in the validation dataset, for the same algorithm the error rate was 41.13%, , therefore it was possible to reaffirm, the predictive power that the Random forest has, compared to other algorithms.

In the models the significant variables for the analysis were: Lotarea, Prevsaleprice, Prevsaleprice2, Countyland, Stories, Totalrooms, Halfbaths, Fireplaces And Bsmtgarage, applying a single method of dimensionality reduction (Information Gain, within these are classified as characteristics of the building and of the territory and the saleprice variable was categorized into three levels 'low', 'middle'and' high, balanced.

As a final test, a K-folds cross validation was applied to the Naives Bayes method, and the result I did not show a significant improvement in the Accuracy of the model with a 55.1% considered low in the prediction.

6. LIMITATIONS

In the application of the Pearson Chi-Squered test, to evaluate the independence of the factor variables, despite showing the significance value or P-VALUE, the output in R packages displayed a message of “Chi- Squared approach may be incorrect ”, which does not guarantee 100% reliability of the test application. However, in the data processing, the results obtained in a test pair of attributes that apparently have a degree of independence were taken into account.

For dimensionality reduction, only information gain was applied. When applying the Backward Elimination method, an unexpected error was obtained that did not allow to obtain the expected result and thus be able to compare with the result obtained with information gain. What could have generated a different set of relevant variables to predict the models.

The application of K-fold cross validation could not be performed in the Naives Bayes, Random Forests and multinomial logistic regression algorithms, in the sense that R, generates an alert that it cannot be Allocated a vector of size 62 Gb, this can due to the fact that some of the attributes selected to build the model should be converted to another type of data or also because some algorithms are sensitive to the type of data or levels in the attributes, however the test was carried out to the satisfaction of the algorithm Decision tree as an illustration of analysis of how or not to improve the prediction of algorithms.

7. RECOMMENDATIONS

In order to avoid measurement biases, the outliers were imputed with the median, although some attributes after the imputation maintain some outliers, for future work to perform a specific treatment for them, such as, to remove them from the sample, if the Amount is not considerable due to loss of information.

The data set object of the study contained an average of 23% of missing values and NaN, which deserve an imputation treatment with some technique or statistical tool, these missing data may have generated a loss of accuracy of the applied models.

In dimensionality reduction, perform a bivariate analysis on the factor type variables of the sample to determine attribute independence and then proceed to the selection of the relevant factors.

8. REFERENCES

[1] Vilches-Blázquez, L.M. (2016). Identification of the complex techniques applicable to the definition of agricultural agricultural land market prices in Colombia. Version 4. Rural Agricultural Planning Unit (UPRA).

[2] Ashray Kakadiya, Khushal Shingala, Shiv Raj Sharma (2018), House Price Prediction with Regression and Classification.

[3] An Nguyen March 20, (2018),Housing Price Prediction.

[4] Giobertti Raúl, Morantes-Quintana1, Gladys Rincón-Polo y Narciso Andrés Pérez-Santodomingo abril (2018), Modelo De Regresión Lineal Múltiple Para Estimar Concentración De Pm1.

[5] George Lever D. The Hedonic Price Model.

[6] Gerald Muriuki octuber 11 (2017), Predicting House Prices Using Linear Regression.

[7] Vincenza Chiarazzo, Leonardo Caggiani, Mario Marinellia and Michele Ottomanellia July (2014), A Neural Network based Model for Real Estate Price Estimation Considering Environmental Quality of Property Location.

[8] Sonia del Rey Simeón (2018), Un Análisis econométrico del precio de la vivienda en Sevilla en el año 2017.

[9] Link: https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/.

Apendix1. Data dictionary

Field Name	Field Description
PARID	Parcel Identification Number
PROPERTYHOUSENUM	Property Location House Number
PROPERTYFRACTION	Property Location House Number Fraction
PROPERTYADDRESS	Property Location Street Name
PROPERTYCITY	Property Location City Name
PROPERTYSTATE	Property Location State Name
PROPERTYUNIT	Property Location Unit Number
PROPERTYZIP	Property Location Zip code, first 5 digits
MUNICODE	Municipality Code (Tax District)
MUNIDESC	Municipality Name
SCHOOLCODE	School District Code
SCHOOLDESC	School District Name
LEGAL1	Legal Description 1
LEGAL2	Legal Description 2
LEGAL3	Legal Description 3
NEIGHCODE	Neighborhood Number
NEIGHDESC	Neighborhood Name
TAXCODE	Tax Status
TAXDESC	Tax Status Description
TAXSUBCODE	Tax Sub Code (applies to PURTA only)
TAXSUBCODE_DESC	Tax Sub Code Description
OWNERCODE	Owner Type Code 1
OWNERDESC	Owner Description
CLASS	Class

CLASSDESC	Class Description
USECODE	Land Use Code
USEDESC	Land Use Code Description
LOTAREA	Sum of Area of Land
HOMESTEADFLAG	Homestead Flag
CLEANGREEN	Clean and Green Flag
FARMSTEADFLAG	Farmstead Flag
ABATEMENTFLAG	Abatement Flag
RECORDDATE	Record Date
SALEDATE	Sale Date
SALEPRICE	Sale Price
SALECODE	Sale Validity Code
SALEDESC	Sale Validity Code Description
DEEDBOOK	Book Number
DEEDPAGE	Page Number
PREVSALEDATE	Previous Sale Date
PREVSALEPRICE	Previous Sale Price
PREVSALEDATE2	Previous Sale Date 2
PREVSALEPRICE2	Previous Sale Price 2
CHANGENOTICEADDRESS1	Change Notice Full Address 1
CHANGENOTICEADDRESS2	Change Notice Full Address 2
CHANGENOTICEADDRESS3	Change Notice Full Address 3
CHANGENOTICEADDRESS4	Change Notice Full Address 4
COUNTYBUILDING	County Assessed Value for Building
COUNTYLAND	County Assessed Value for Land
COUNTYTOTAL	County Assessed Value Total

COUNTYEXEMPTBLDG	County Exempt Building Amount
LOCALBUILDING	Local Assessed Value for Building
LOCALLAND	Local Assessed Value for Land
LOCALTOTAL	Local Assessed Value Total
FAIRMARKETBUILDING	Fair Market Building Value
FAIRMARKETLAND	Fair Market Land Value
FAIRMARKETTOTAL	Fair Market Total Value
STYLE	Dwelling - Architectural Style
STYLEDESC	Dwelling - Architectural Style Description
STORIES	Dwelling - Number of Stories
YEARBLT	Dwelling - Year Built
EXTERIORFINISH	Dwelling - Exterior Wall code
EXTFINISH_DESC	Dwelling - Exterior Wall Description
ROOF	Dwelling - Roof
ROOFDESC	Dwelling - Roof Description
BASEMENT	Dwelling - Basement
BASEMENTDESC	Dwelling - Basement Description
GRADE	Dwelling - Grade
GRADEDESC	Dwelling - Grade Description
CONDITION	Dwelling - Condition
CONDITIONDESC	Dwelling - Condition Description
CDU	Dwelling - CDU
CDUDESC	Dwelling - CDU Description
TOTALROOMS	Dwelling - Total Rooms
BEDROOMS	Dwelling - Bedrooms
FULLBATHS	Dwelling - Full Baths

HALFBATHS	Dwelling - Half Baths
HEATINGCOOLINGDESC	Description for the type Heating / Cooling system.
HEATINGCOOLING	Dwelling - Heating Cooling
FIREPLACES	Dwelling - Number of Wood burning Fireplaces Stacks
BSMTGARAGE	Dwelling - Integral Basement Garage (Number of Cars)
FINISHEDLIVINGAREA	Dwelling - Total Square Feet of Living Area
CARDNUMBER	Dwelling - Building (card) Number
ALT_ID	Alternate Parcel Identification Number
TAXYEAR	The current certified tax year
ASOFDATE	The run date of this file