Predict House Accomodations Prices

M'GHARI TARIQ
FAKIRI ISMAIL

In this data mining project, you will use data science techniques such as machine learning to predict the price of real estate in a particular area. This project finds application in real estate industries for predict house prices based on previously available data such as the location and size of the house and facilities near the house

Cycle d'ingenieur Systeme d'information et BIG DATA
Ecole nationale des sciences Appliquées Berrechid
01/02/2021

Abstract

This year in "Data mining" course, we will do a prject supervised by Pr.HRIMECH HAMID. In this project we will predict house accomodation using kaggle dataset that you can find in the link below https://www.kaggle.com/c/house-prices-advanced-regression-techniques.The dataset is a history of prices in a particular area.And we will be using R language as our statistical language.

1 Dataset Explanation

This dataset contains multiple variables that we will use to make our prediction model for SalePrice which is the dependant variable (Y). Each line in the dataset reffers to a house and it's sale price. The data are both categorical/numeric. We will be using encoding algorithms for our categorical data. For the NA, we will be using different approaches such us eliminating the NA, or replacing them with the mean, Max or Min of a certain column. We can see in figure1 a scatter plot of the SalePrice of the dataset.

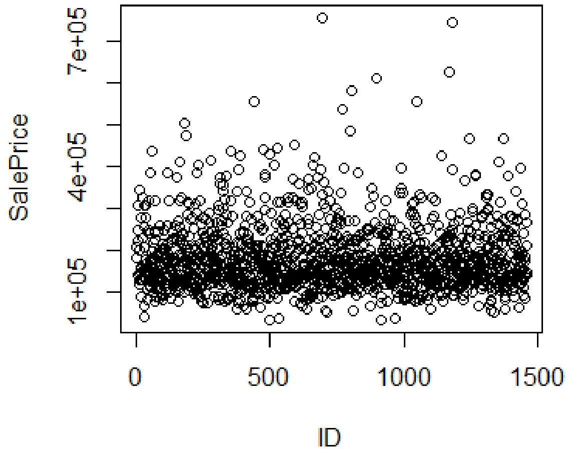


Figure 1: Scatter plot of the prices of the data!

A look into our data In this part we will visualize the content of our data to have an idea about what we can do to this data. On the following figure we have plotted the house prices with each variable, and we took the ones that they significantly influence the houses Price

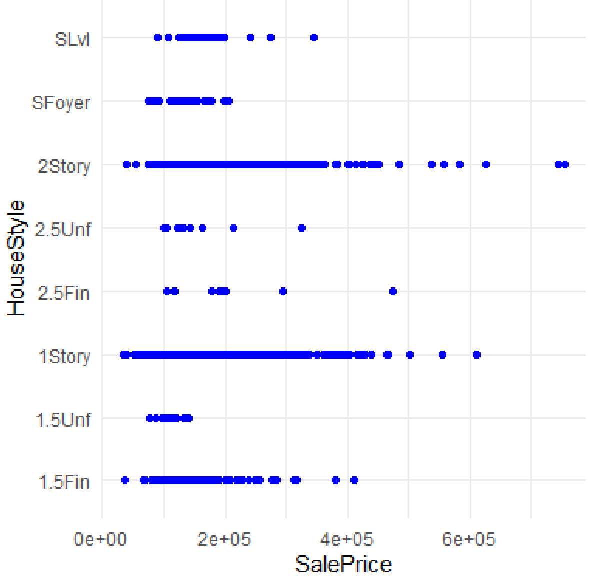


Figure 2: Scatter plot of the prices of the data!

The figure 2 shows us the price changes between different home types, we can conclude that the houses with "1.5Fin One and one-half story: 2nd level finished" has generally the lowest prices comparing to other types of homes.

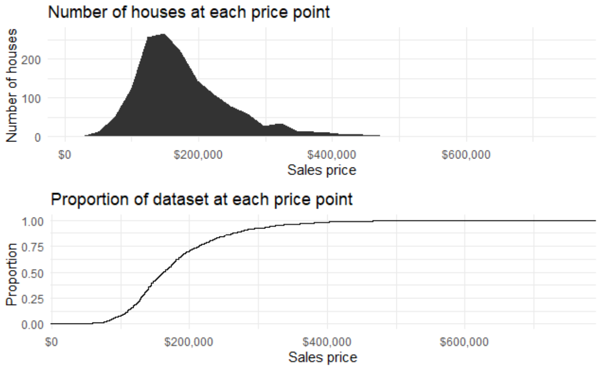


Figure 3: Houses per sale Price!

Here we discover that the most common house price is between USD 100k and USD 300K with a peak near USD 150k and a meaningful proportion of outliers above USD 300K

2 Data Preparation

Our dataset contains multiple NA's in the next figure we will show you a graphic that will demonstrate the NA values on the first 10 columns, we will be using VIM for this graphics

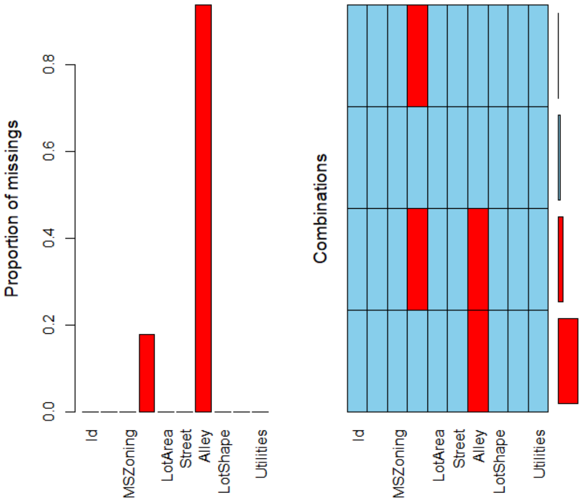
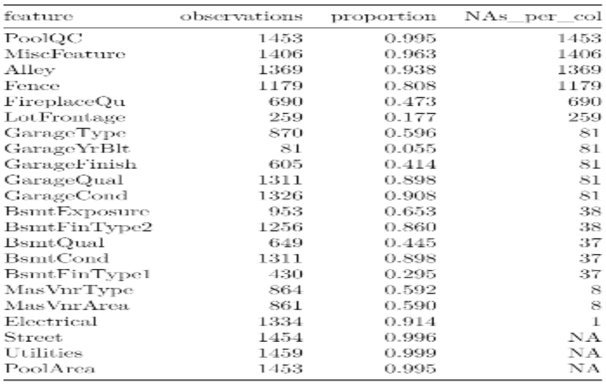


Figure 4: Missing data!

As we can see in Figure 3, the first graphic shows us the proportion of NA's in these columns, and for the second graphic it shows a combination of each variable with another variable, where they have NA's in common

The following figure will show us the number of NA's per each column so that we would know if we can eliminate one of the variables that has a lot of NA's



feature	observations	proportion	NAs_per_col
PoolQC	1453	0.995	1453
MiscFeature	1406	0.963	1406
Alley	1369	0.938	1369
Fence	1179	0.808	1179
FireplaceQu	690	0.473	690
LotFrontage	259	0.177	259
GarageType	870	0.596	81
GarageYrBlt	81	0.055	81
GarageFinish	605	0.414	81
GarageQual	1311	0.898	81
GarageCond	1326	0.908	81
BsmtExposure	953	0.653	38
BsmtFinType2	1256	0.860	38
BsmtQual	649	0.445	37
BsmtCond	1311	0.898	37
BsmtFinType1	430	0.295	37
MasVnrType	864	0.592	8
MasVnrArea	861	0.590	8
Electrical	1334	0.914	1
Street	1454	0.996	NA
Utilities	1459	0.999	NA
PoolArea	1453	0.995	NA

Figure 5: Sum of NA'S for each column in the dataset!

Observations: Number of observations of the most frequent feature value
Proportion: Proportion of total observations
Observations: Number of missing values in a feature. Note that 'NA' here means that there are no missing features in a given column
We took a 50% rate of NA's, so every variable that has more than 1400/2 =700 NA, we will exclude it before building our regression model. In this case, the variables to exclude are: Alley,FireplaceQu,PoolQC,Fence,MiscFeature

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.
A correlation matrix consists of rows and columns that show the variables. Each cell in a table contains the correlation coefficient.

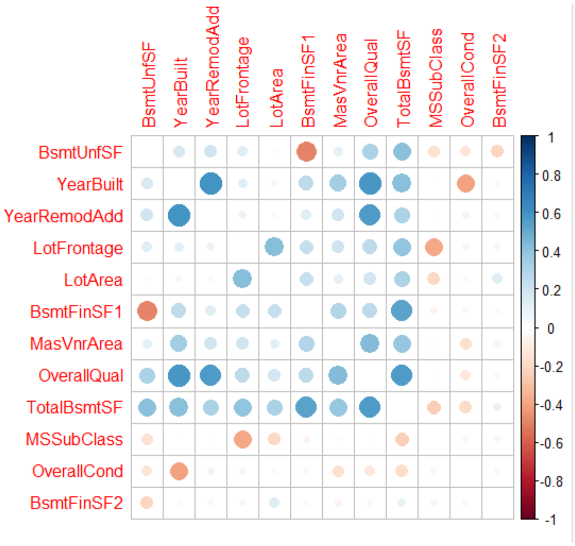


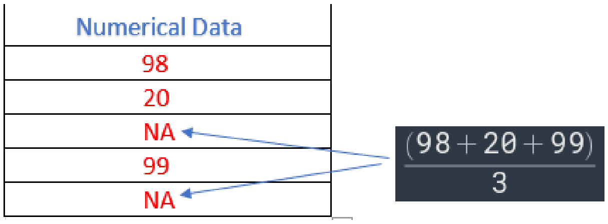
Figure 6: Correlation Matrix!

3 Building Models

I.Preparing data for our Models

a.Replacing NA Values

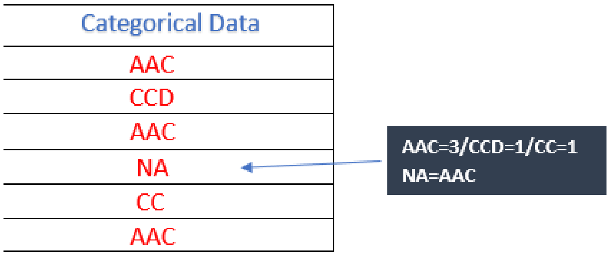
For this first model we will perform certain replacements regarding NA values. The method we will be using is to replace Numerical variables with the mean of the columns. And categorical variables we will perform frequently used value in the column the table below will explain.



Numerical Data
98
20
NA
99
NA

$$\frac{(98+20+99)}{3}$$

Figure 7: Numerical data NA's replacement!



Categorical Data
AAC
CCD
AAC
NA
CC
AAC

AAC=3/CCD=1/CC=1
NA=AAC

Figure 8: Categorical data NA's replacement!

After replacing the NA values into our data, we need to encode categorical values.

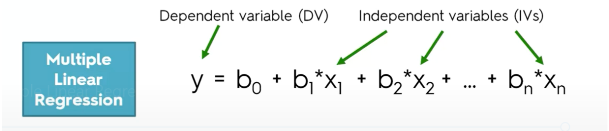
b.Encoding categorical values

Numerization of categorical variables by taking the values 0 or 1 to indicate the absence or presence of each category. If the categorical variable has k categories we would need to create k binary variables (technically speaking, k-1 would suffice). In the following example, the categorical variable "Trend" with three values transformed to three separate binary numerical variables. The main drawback with this method is when the categorical variable with many values (e.g., city) which can tremendously increase the dimension of data.

b. Model Building

i. Linear Regression Model

linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.



$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \ldots + b_n{}^*x_n$$

Figure 9: Multiple Linear Regression!

In this section we will create our model following a Multiple Linear Regression Model, this step comes up after the data cleaning and preparation in the next figure we plotted the Multiple Regression Model

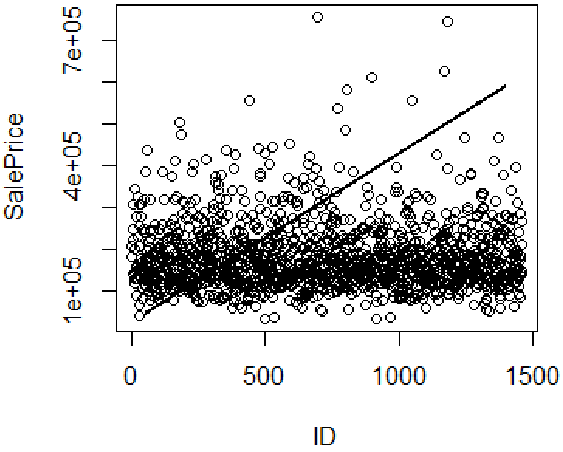


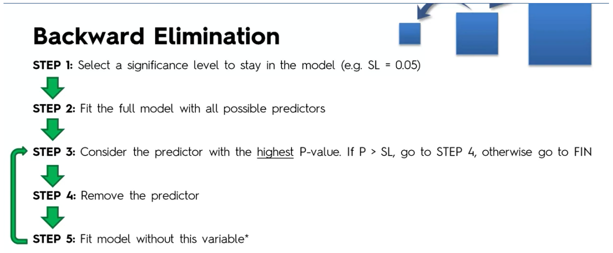
Figure 10: Multiple Linear Regression!

as we can see in the above figure a multiple linear regression will get a bad results because of how the data is scattered.

So we have applied a backward elimination on the data.

Backward Elimination:

Backward elimination is a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output



Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

STEP 2: Fit the full model with all possible predictors

STEP 3: Consider the predictor with the highest P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

Figure 11: Multiple Linear Regression!

Even after applying Backward elimination our model still scores bad results, in the table below you will see the score for this model

R2	RMSE
0.70	76709.92

ii. Principal Component Regression

In statistics, principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA). More specifically, PCR is used for estimating the unknown regression coefficients in a standard linear regression model.

In PCR, instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors. One typically uses only a subset of all the principal components for regression, making PCR a kind of regularized procedure and also a type of shrinkage estimator.

R2	RMSE
0.79	68622.52

iii. Regularized Regression

Regularized regression is a type of regression where the coefficient estimates are constrained to zero. The magnitude (size) of coefficients, as well as the magnitude of the error term, are penalized. Complex models are discouraged, primarily to avoid overfitting.

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. This type is very useful when you have high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

R2	RMSE
0.89	23798.81

iii. M.A.R.S

In statistics, multivariate adaptive regression splines (MARS) is a form of regression analysis introduced by Jerome H. Friedman in 1991. It is a non-parametric regression technique and can be seen as an extension of linear models that automatically models nonlinearities and interactions between variables.

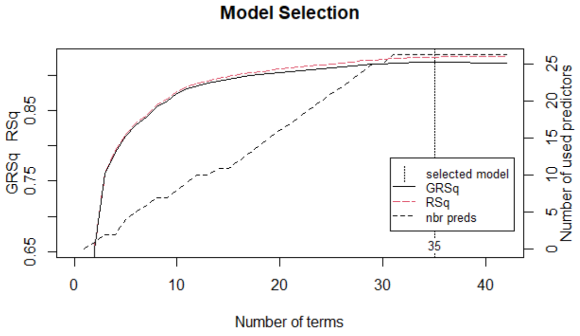


Figure 12: M.A.R.S!

R2	RMSE
0.97	20227.12

4 Conclusion

Model	R2	RMSE
Multiple Linear Regression	0.70	76709.92
PCA	0.79	68622.52
partial least squares	0.76	74849.37
regularized regression	0.89	23798.81
M.A.R.S TUNED	0.97	20227.12

RMSE: Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

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

Figure 13: RMSE!

R2: The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

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

Figure 14: R squared!