Department of Computer Science and Engineering

Institute of Engineering

JK Lakshmipat University, Jaipur, India

Shubham Bajaj

shubhambajaj@jklu.edu.in

Abstract

This case study mainly focuses on predicting the factors which affect house price.  
 Using stepwise regression and applying it on suitable dataset, we can set some features on which buyers will rely on suggested sale price. This case study can help sellers to improve their pricing methods and also the features on which customers are attracted to.

House Price Prediction

A case study on factors affecting house prices

Case Study Synopsis

This Case Study is based on house price prediction. People in this modern world are always in search for houses. That’s why the dealers or the sellers need to price their houses accordingly so that they get the best price of the property. In this case study we are predicting the factors which have most significance on the house prices. We got a dataset from kaggle in which several features of different houses are given and a sales price is also given which will be used as a predictor variable. After looking though data we can see there are lot of categorical variables and missing values. So, first they needed to be dealt with. We converted categorical variables into dummy variables and filled the missing values. We also removed two columns i.e. Id and SalePrice as Id is not a predictor variable and SalePrice is our target variable. After that we got the data on which we have to do modelling.

We applied stepwise regression (mainly forward selection) on our pre-processed data which will give the columns which affect sale price. We got some results after modelling which we used for evaluation and answering our case questions. After evaluation, we can suggest dealers and sellers to focus on some qualities extracted from our case study and price their properties accordingly.

Learning Objectives

The primary objective of this case study is to demonstrate multiple regression model to establish relationship between the sale prices of houses and lot area or overall qualities as predictor variables. Some other learning objectives include:

1. Develop regression model using transformation of response variable.
2. Validate the model using statistical measures such as mean.
3. Demonstrate the connection between target variables and a collection of predictors.

Study Questions

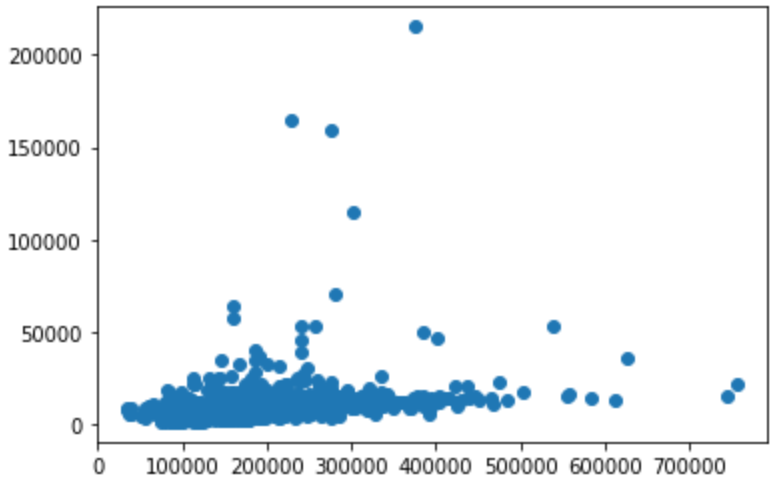
1. Identify statistically significant predictors that influence the sales price of houses.
2. People really like having pool in their home. Is there any evidence that sales price is affected by pool quality?
3. Which factor is affecting most the sales price and by what accuracy?

Suggested Answers of Case Questions

1. **Identify statistically significant predictors that influence the sales price of houses.**

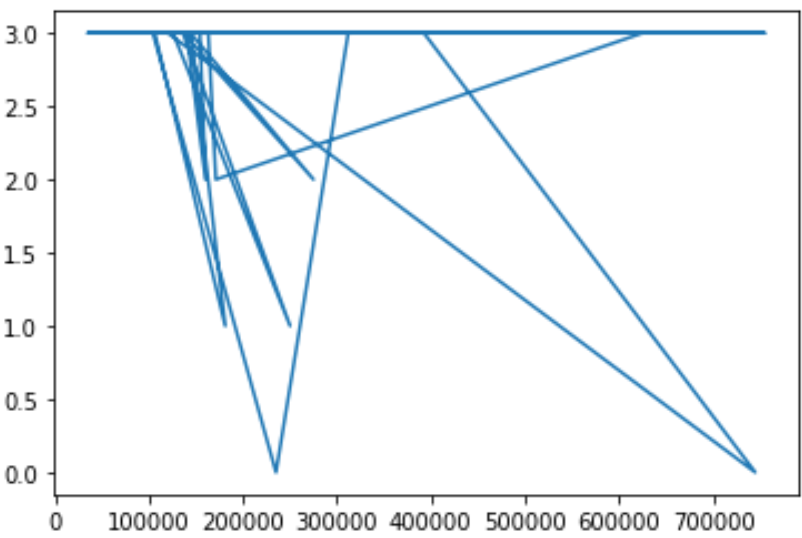
* We have applied model on all of the columns we had in dataset. So, there were total 79 predictors, we made model on. The four most statistically significant predictors **are Street, OverallQual, LotArea and LotShape.** As we can see there is significant change in the cross validation score (i.e. mean and standard error) after adding them to model.

I have tried mapping LotArea vs SalePrice. We can see how they are related in it.



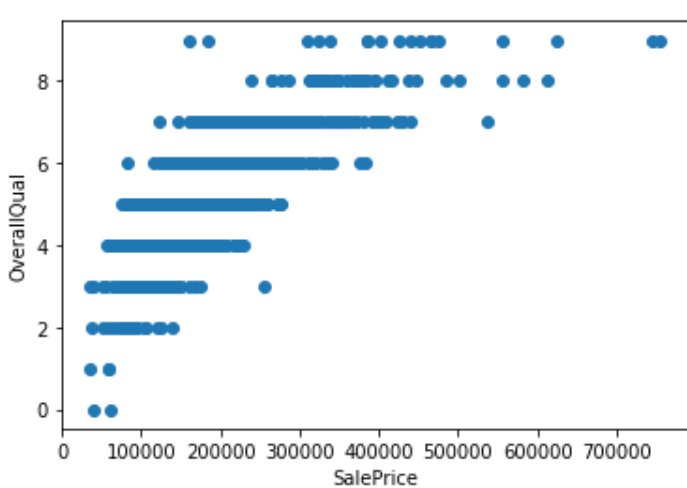
1. **People really like having pool in their home. Is there any evidence that sales price is affected by pool quality?**

* In this dataset, if we see after our result the accuracy reduced from 81% to 79% after adding ‘PoolQC’. So, it means that pool quality has a negative impact on the sale price of house. Though change of 3% is not that significant, but it affects the sale price a little. Here is a line chart showing how sale price and pool quality are related.



1. **Which factor is affecting most the sales price and by what accuracy?**

* Looking through the model and accuracy, the most statistically significant predictor which is playing a key role in affecting the sales price is “**OverallQual**” which tells us about the **overall material and finish quality**. Accuracy increased from 22% to around 67% after the addition of “OverallQual”. So overall material and finish quality affects most the sale price of house.



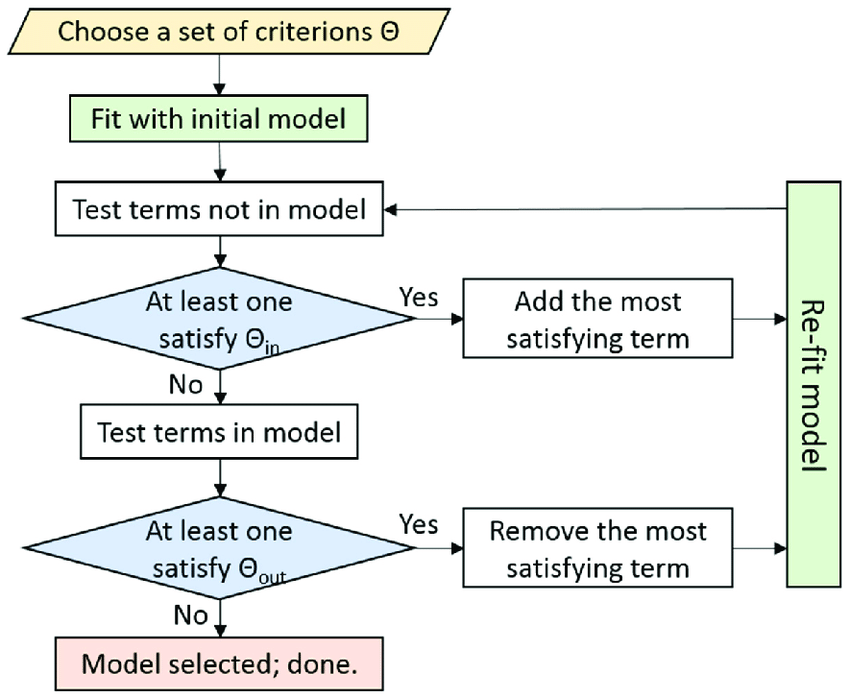
**Appendix A: Description of Utilized technique: [Stepwise Regression]**

In statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre specified criterion. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R2, Akaike information criterion, Bayesian information criterion, Mallows's Cp, PRESS, or false discovery rate.[1]

Equation: Y = b0 + b1 x1 + b2 x2 +……+ bnxn

It can be applied by three methods:

1. Forward selection: Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, we add the one variable that gives the single best improvement to the model. This is the one which is used in our case study.
2. Backward elimination: Backward elimination is reverse process. All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.
3. Bidirectional elimination: A combination of the above, testing at each step for variables to be included or excluded.



Advantages:

* The ability to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options.
* It’s faster than other automatic model-selection methods.
* Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables.

**Annexure B: Crisp Model Steps**

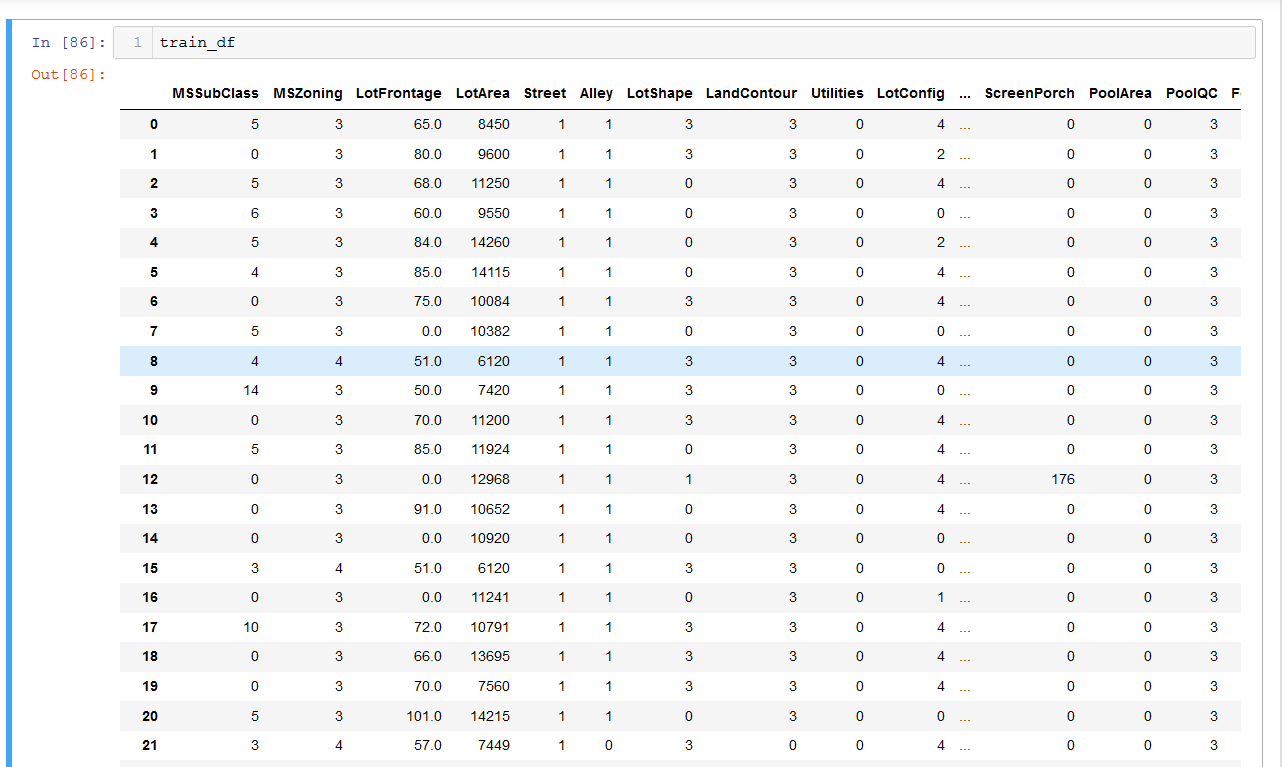
B.1 Business Understanding

House Price Prediction: People move from one place to another for better employment, culture, lifestyle, opportunities. They need home after moving to some new place. And the first thing anyone see for is the price of the house. And price of a certain house depends upon a lot of factors like locality, area, utilities, proximity to public transport, quality and material of house, garage, garden, etc. Our main aim in this case study is to find out the factors which have more effect on price compared to others based on the dataset.

B.2 Data Understanding

“House Prices: Advanced Regression Techniques”

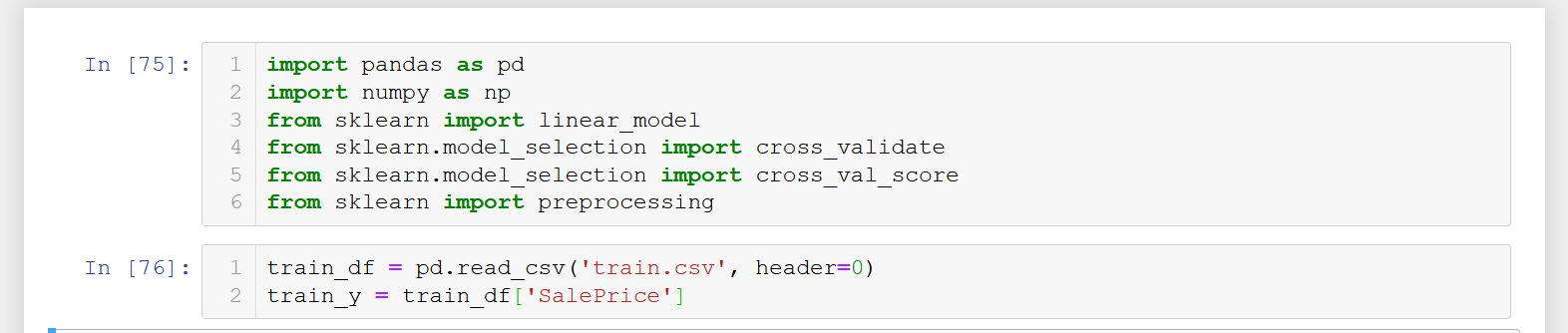
It contains two csv file, I worked with one. Dataset contains some features of the various houses. It has 1459 data rows and 81 columns (information and factors). We have taken 79 predictor variables and one target variable i.e. SalePrice.Dataset contains different kind of data types i.e. float, integer, character, etc.



B.3 Data Preparation

1. Data Gathering

We have collected data from kaggle which is suitable for our case study.

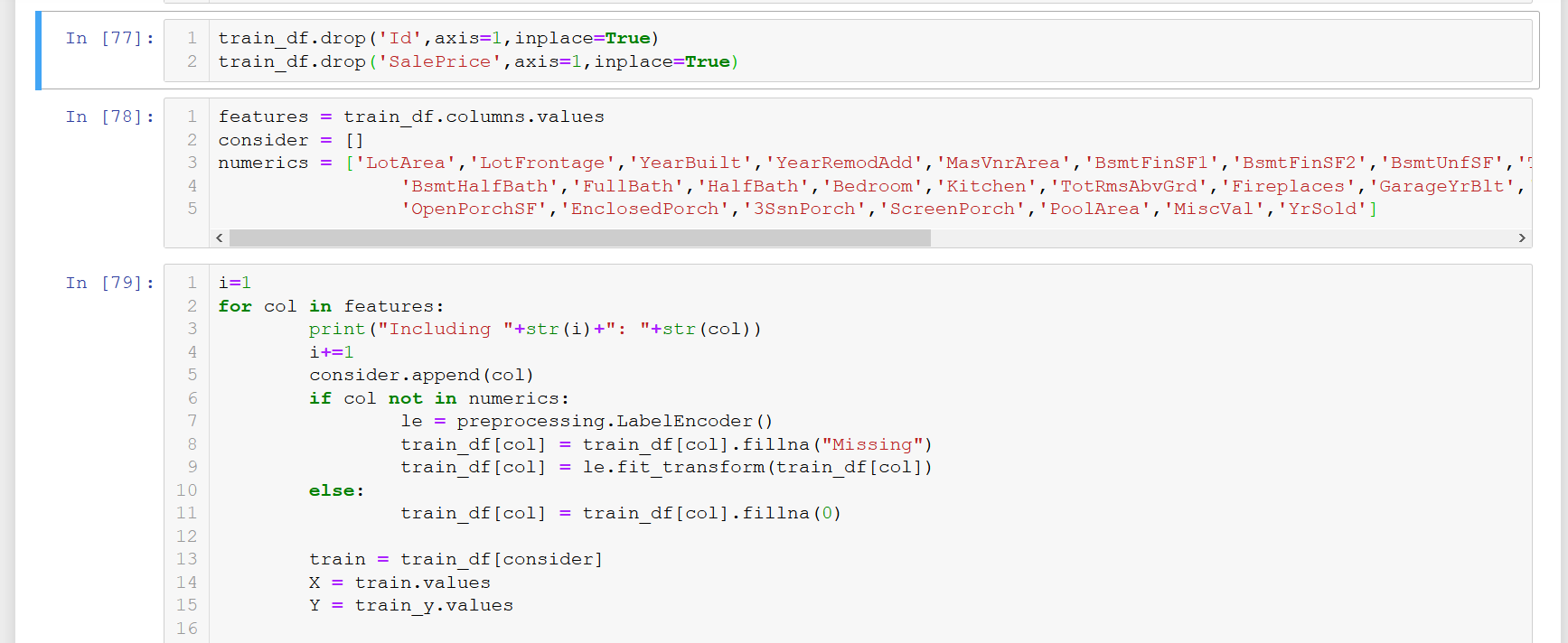
2. Accessing the data

3. Data Preprocessing

a) Dropped two fields i.e. Id and SalePrice. Because Id is not a predictor variable and SalePrice is a target variable.

b) Encoded fields containing datatypes other than integer and float.

c) Filled the missing values.

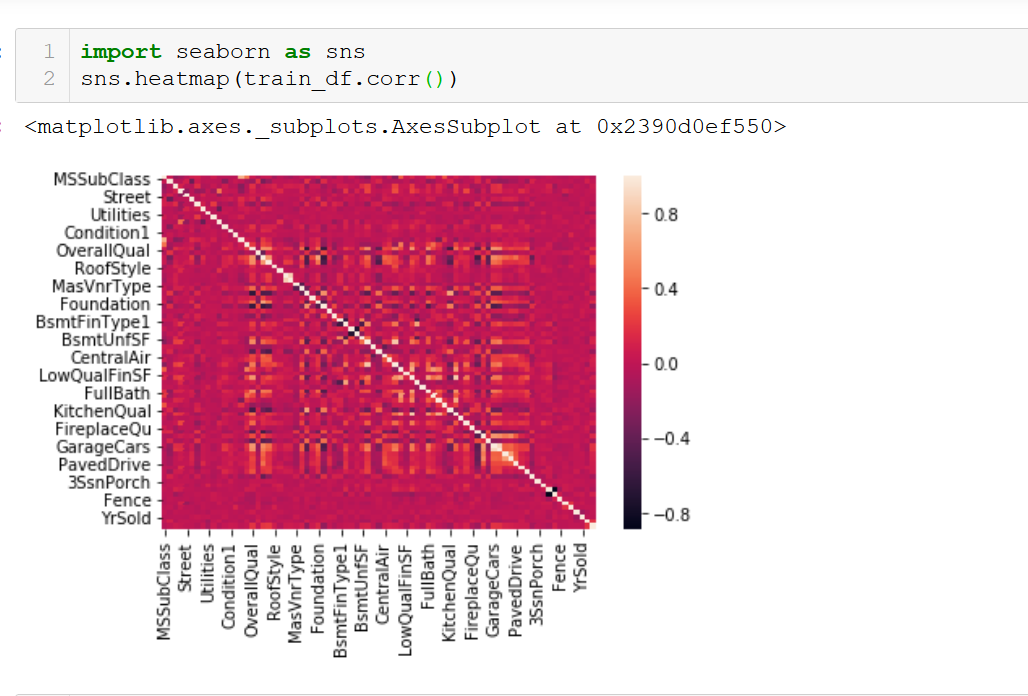


Metadata after encoding:

|  |  |  |
| --- | --- | --- |
| Features | Data Types | Description |
| MSSubClass | int64 | building class |
| MSZoning | int32 | general zoning classification |
| LotFrontage | float64 | Linear feet of street connected to property |
| LotArea | int64 | Lot size in square feet |
| Street | int32 | Type of road access |
| Alley | int32 | Type of alley access |
| LotShape | int32 | General shape of property |
| LandContour | int32 | Flatness of the property |
| Utilities | int32 | Type of utilities available |
| LotConfig | int32 | Lot configuration |
| LandSlope | int32 | Slope of property |
| Neighborhood | int32 | Physical locations within Ames city limits |
| Condition1 | int32 | Proximity to main road or railroad |
| Condition2 | int32 | Proximity to main road or railroad (if a second is present) |
| BldgType | int32 | Type of dwelling |
| HouseStyle | int32 | Style of dwelling |
| OverallQual | int64 | Overall material and finish quality |
| OverallCond | Int64 | Overall condition rating |
| YearBuilt | Int64 | Original construction date |
| YearRemodAdd | Int64 | Remodel date |
| RoofStyle | Int32 | Type of roof |
| RoofMatl | Int32 | Roof material |
| Exterior1st | Int32 | Exterior covering on house |
| Exterior2nd | Int32 | Exterior covering on house (if more than one material) |
| MasVnrType | Int32 | Masonry veneer type |
| MasVnrArea | Float64 | Masonry veneer area in square feet |
| ExterQual | Int32 | Exterior material quality |
| ExterCond | Int32 | Present condition of the material on the exterior |
| Foundation | Int32 | Type of foundation |
| BsmtQual | Int32 | Height of the basement |
| BsmtCond | Int32 | General condition of the basement |
| BsmtExposure | Int32 | Walkout or garden level basement walls |
| BsmtFinType1 | Int32 | Quality of basement finished area |
| BsmtFinSF1 | Int64 | Type 1 finished square feet |
| BsmtFinType2 | Int32 | Quality of second finished area (if present) |
| BsmtFinSF2 | Int64 | Type 2 finished square feet |
| BsmtUnfSF | Int64 | Unfinished square feet of basement area |
| TotalBsmtSF | Int64 | Total square feet of basement area |
| Heating | Int32 | Type of heating |
| HeatingQC | Int32 | Heating quality and condition |
| CentralAir | Int32 | Central air conditioning |
| Electrical | Int32 | Electrical system |
| 1stFlrSF | Int64 | First Floor square feet |
| 2ndFlrSF | Int64 | Second floor square feet |
| LowQualFinSF | Int64 | Low quality finished square feet (all floors) |
| GrLivArea | Int64 | Above grade (ground) living area square feet |
| BsmtFullBath | Int64 | Basement full bathrooms |
| BsmtHalfBath | Int64 | Basement half bathrooms |
| FullBath | Int64 | Full bathrooms above grade |
| HalfBath | Int64 | Half baths above grade |
| BedroomAbvGr | Int64 | Number of bedrooms above basement level |
| KitchenAbvGr | Int64 | Number of kitchens |
| KitchenQual | Int32 | Kitchen quality |
| TotRmsAbvGrd | Int64 | Total rooms above grade (does not include bathrooms) |
| Functional | Int32 | Home functionality rating |
| Fireplaces | Int64 | Number of fireplaces |
| FireplaceQu | Int32 | Fireplace quality |
| GarageType | Int32 | Garage location |
| GarageYrBlt | Int64 | Year garage was built |
| GarageFinish | Int32 | Interior finish of the garage |
| GarageCars | Int64 | Size of garage in car capacity |
| GarageArea | Int64 | Size of garage in square feet |
| GarageQual | Int32 | Garage quality |
| GarageCond | Int32 | Garage condition |
| PavedDrive | Int32 | Paved driveway |
| WoodDeckSF | Int64 | Wood deck area in square feet |
| OpenPorchSF | Int64 | Open porch area in square feet |
| EnclosedPorch | Int64 | Enclosed porch area in square feet |
| 3SsnPorch | Int64 | Three season porch area in square feet |
| ScreenPorch | Int64 | Screen porch area in square feet |
| PoolArea | Int64 | Pool area in square feet |
| PoolQC | Int32 | Pool quality |
| Fence | Int32 | Fence quality |
| MiscFeature | Int32 | Miscellaneous feature not covered in other categories |
| MiscVal | Int64 | $Value of miscellaneous feature |
| MoSold | Int64 | Month Sold |
| YrSold | Int64 | Year Sold |
| SaleType | Int32 | Type of sale |
| SaleCondition | Int32 | Condition of sale |
| SalePrice | Int32 | Property’s sale price in dollars. This is the target variable that you're trying to predict. |

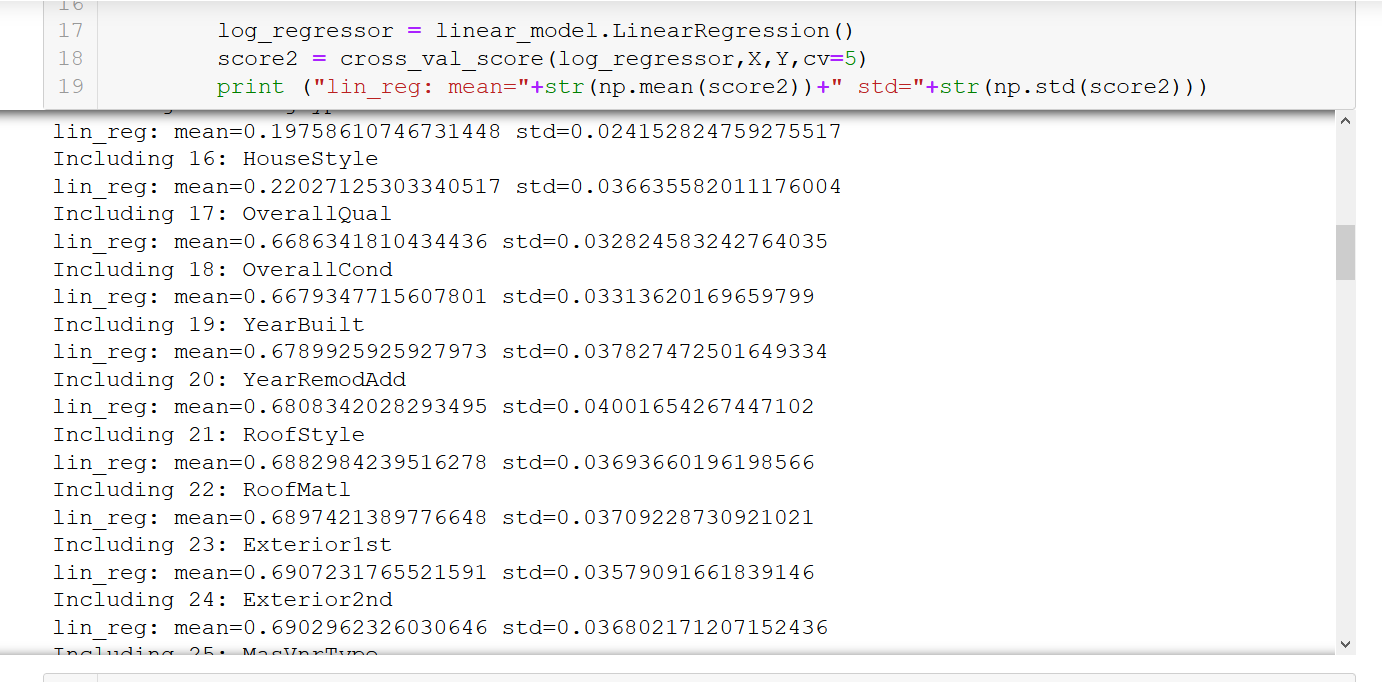
1. Storing Data

After that data was stored in X and y to further work on them. And seen the correlation between columns.



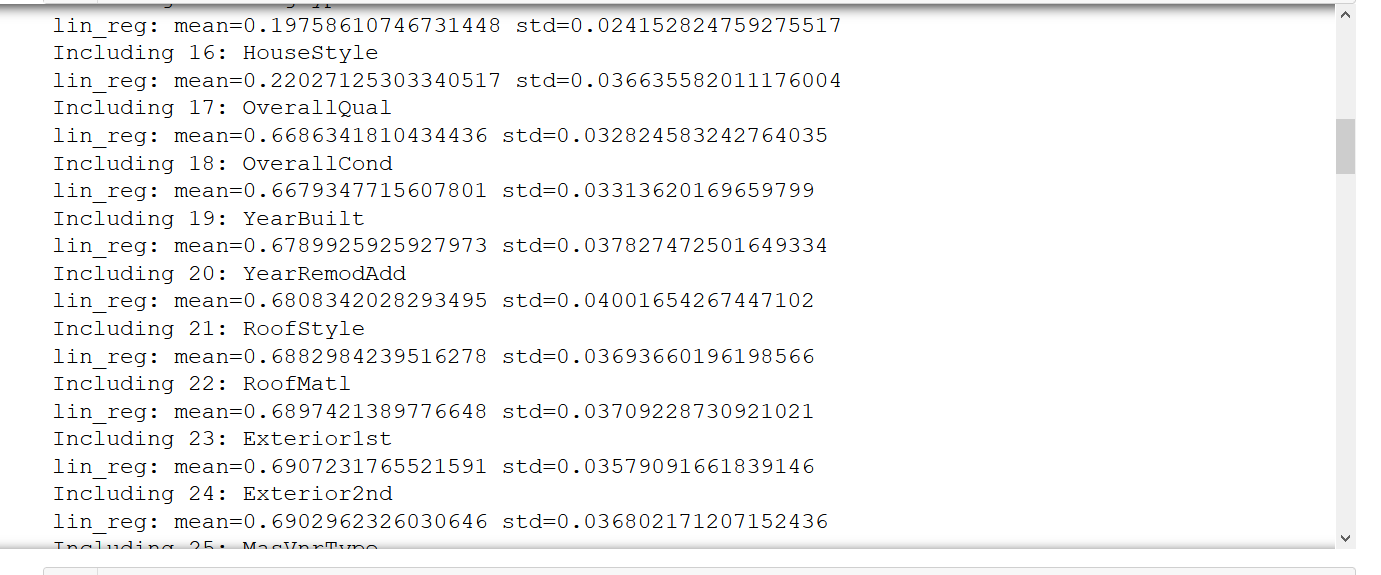
B.4 Modelling

After doing pre-processing, I applied cross\_val\_score function which has default value RMSE (Root Mean Square Error) and calculated the RMSE and stored it in score2. And after that took the mean and standard deviation.



B.5 Evaluation

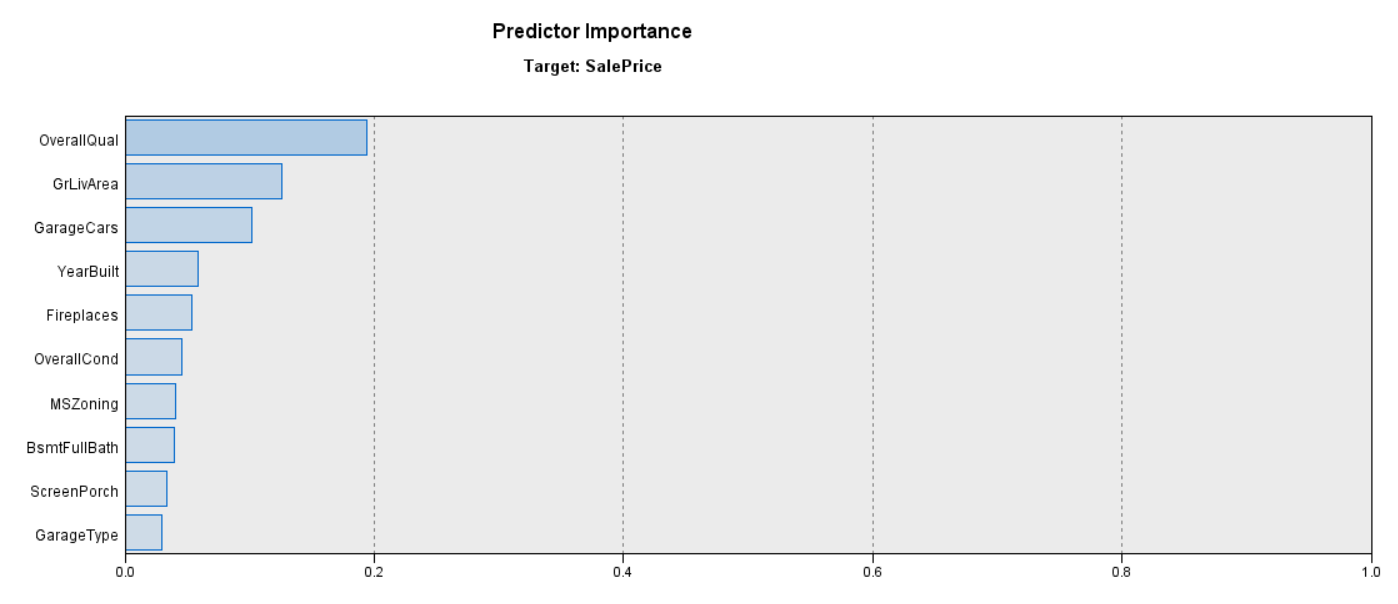
Evaluated the result after running the code and saw some significant increase after adding certain columns.



References

* <https://en.wikipedia.org/wiki/Stepwise_regression>
* DAVID S. MOORE, ” Introduction to the Practice of Statistics”, GEORGE P. McCABE, BRUCE A. CRAIG, 6th edition, 2009
* <https://scikit-learn.org/stable/index.html>
* Douglas C. Montgomery, “Introduction to Linear Regression”, Elizabeth A. Peck, G. Geoffrey Vining, 5th edition

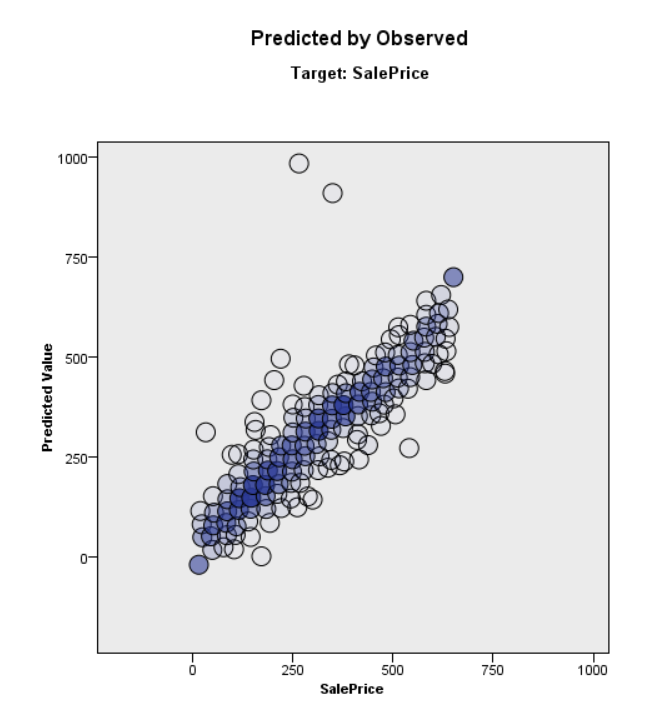
**Annexure C: Modelling**



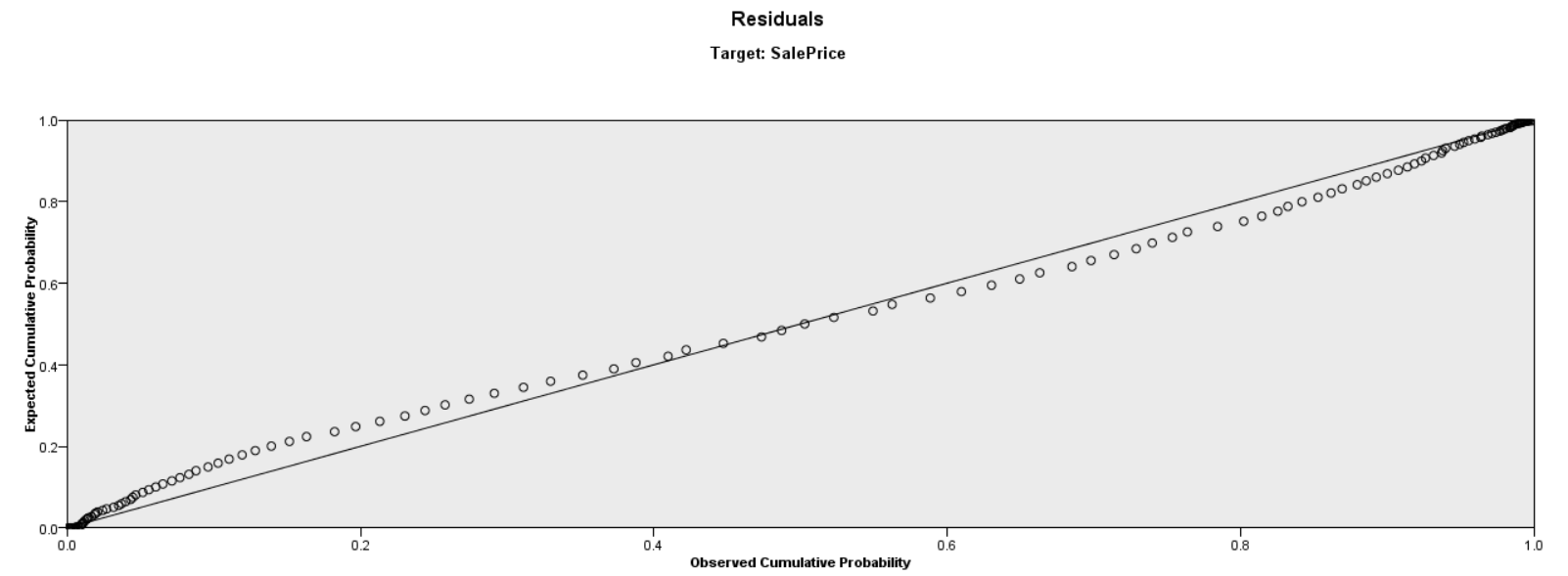
It shows predictor importance when the Criteria taken is F Statistics.

Feature to enter at p-values<=0.050

Feature to remove at p-values>0.100

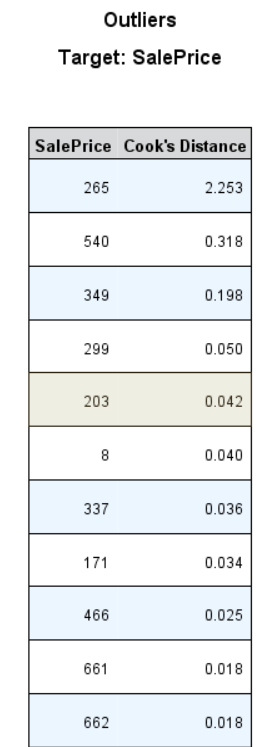


This represents the real SalePrice vs Predicted Value. Almost every predicted value is near the real value.

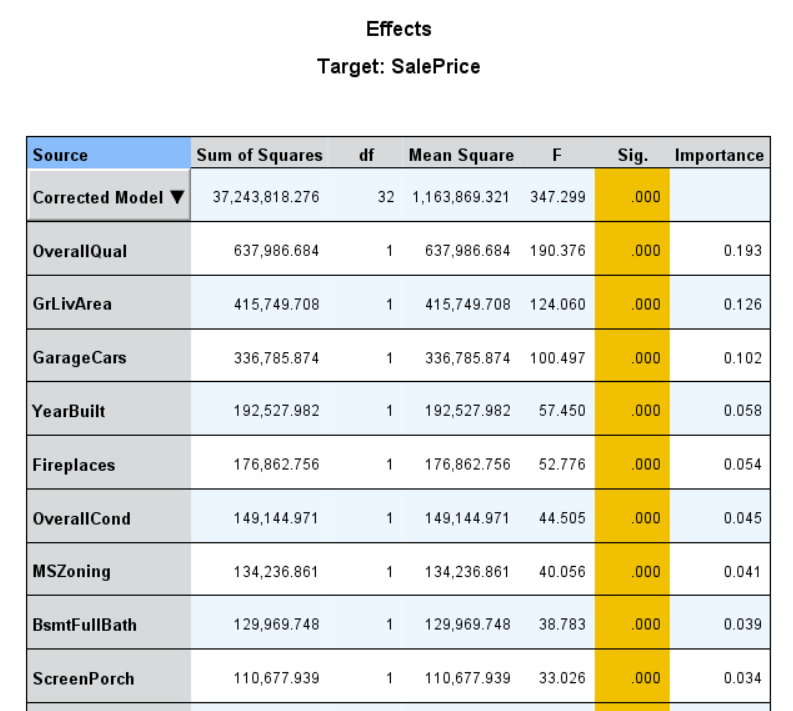


This P-P plot of standardizes residuals compares the distribution of the residuals to a normal distribution. The diagonal line represents the normal distribution. The closer the observed cumulative probabilities of the residuals are to this line, the closer the distribution of the residuals of the normal distribution.

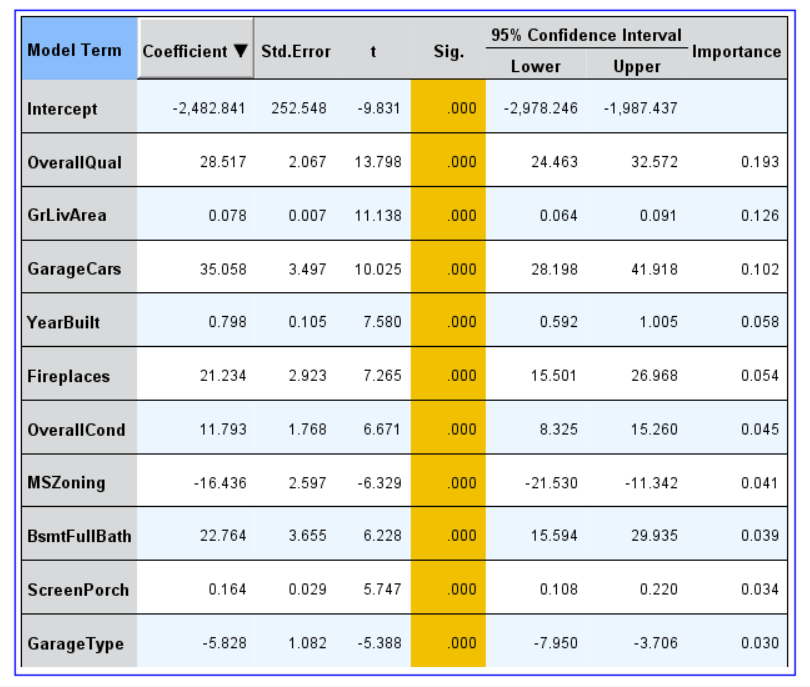
This P-P plot suggests that residuals don’t follow a perfect normal distribution which is an important assumption of the least squares regression.



Observations with high Cook’s distance may significantly change the estimated beta co-efficient values of the predictors depending on whether the observation is included in the sample or not.



This figure shows the F-statistic, sum of squares, mean square value of the top significant features.



Co-efficient which is used to predict or estimate the sale price

