Housing Price Prediction

Using Machine Learning

FLIP ROBO

Project By -

Amar Kumar

Project Description

- The most essential need of every human around the world are houses.

- A US-based housing company named “Surprise Housing” has decided to enter the Australian market. The company is looking at prospective properties to buy houses to enter the market.

- The purpose is to build a Machine Learning model in order to predict the actual value of the prospective properties and help the company in deciding whether to invest in them or not.

- The requirement is to model the price of houses with the available independent variables in the test data set.

Business Goal

- The objective of the company is to know

1. Which aspects of the property are most important to predict the price of variable?
2. How these aspects are co-related with the price of the house?

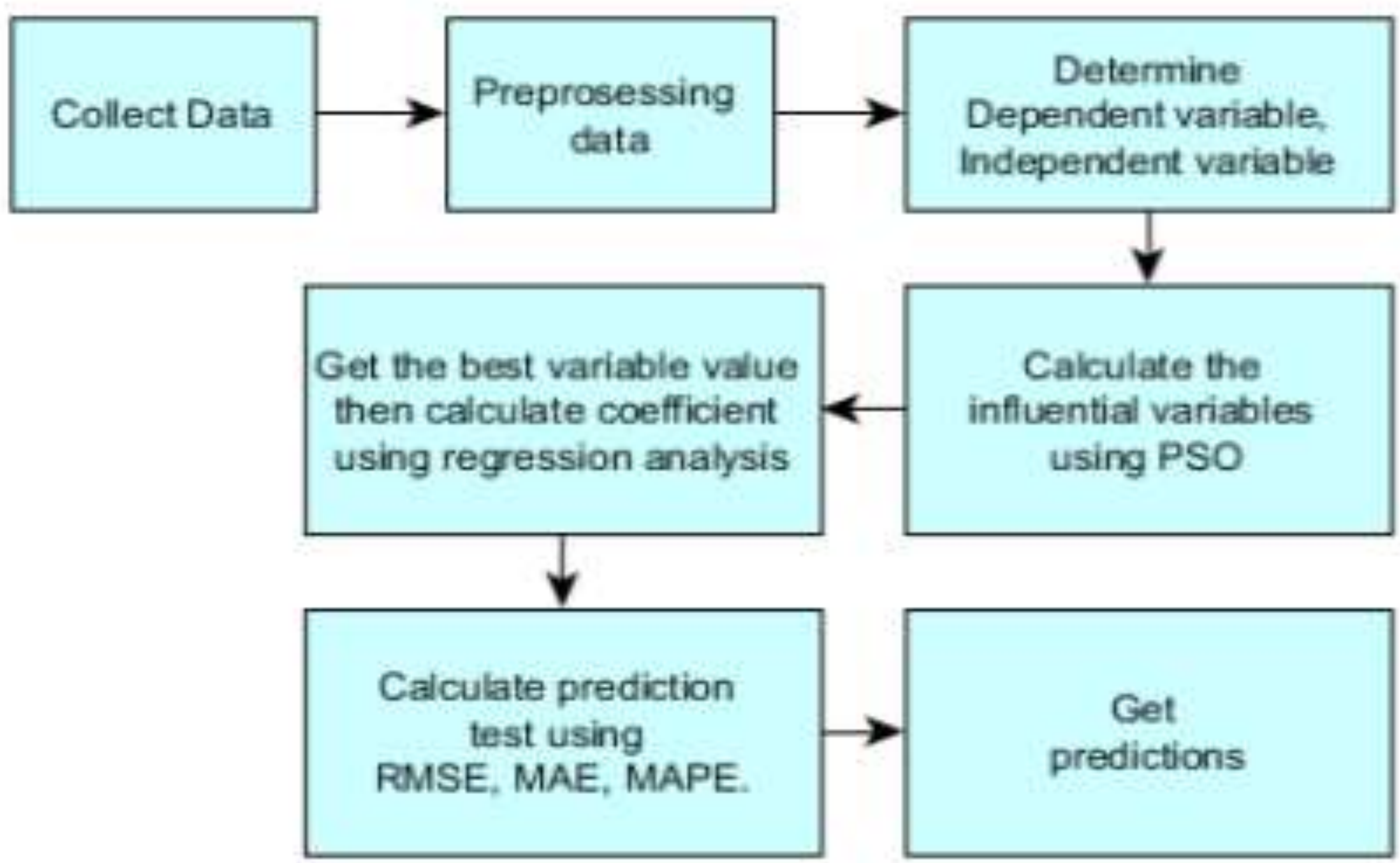
- The model built will then be used by the company management to understand which variables make a huge impact in the prices of the houses and how prices vary with each variable.

- The company management can then accordingly
manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

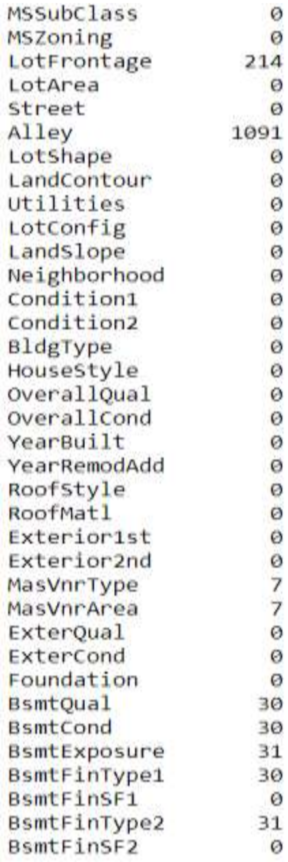
Research & Solution

- The information consists of two data sets i.e. train.csv and test.csv. Training of data will be done on train.csv and the predictions will be made on test.csv. Both data sets are provided in the form of a csv file.
- Data contains 1460 entries each having 81 variables in the train data set but contains 292 entries for the test data set having the same variables as present in the train data set.
- Data contains null values that needs to be treated and extensive EDA has to be performed to gain relationships of important variables and it’s price.
- Data contains both numerical as well as categorical variables hence will be handled accordingly.
- Applying regularization and determining the optimal values of hyper parameters, a machine learning model will be built to pin down the important features which affect the house price positively or negatively.

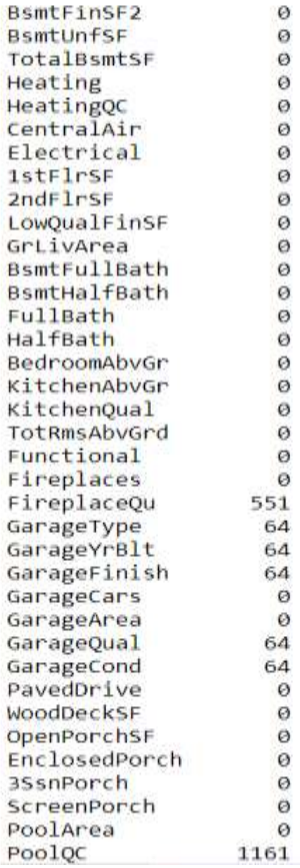
Proposed Model Architecture


Collect Data
Preprosessing data
Determine Dependent variable, Independent variable
Calculate the influential variables using PSO
Get the best variable value then calculate coefficient using regression analysis
Calculate prediction test using RMSE, MAE, MAPE.
Get predictions

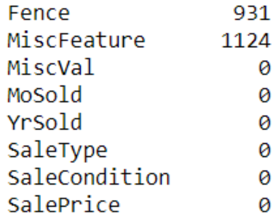
Missing Values



```
MSSubClass          0
MSZoning            0
LotFrontage       214
LotArea             0
Street              0
Alley            1091
LotShape            0
LandContour         0
Utilities           0
LotConfig           0
LandSlope           0
Neighborhood        0
Condition1          0
Condition2          0
BldgType            0
HouseStyle          0
OverallQual         0
OverallCond         0
YearBuilt           0
YearRemodAdd        0
RoofStyle           0
RoofMatl            0
Exterior1st         0
Exterior2nd         0
MasVnrType          7
MasVnrArea          7
ExterQual           0
ExterCond           0
Foundation          0
BsmtQual           30
BsmtCond           30
BsmtExposure       31
BsmtFinType1       30
BsmtFinSF1          0
BsmtFinType2       31
BsmtFinSF2          0
```



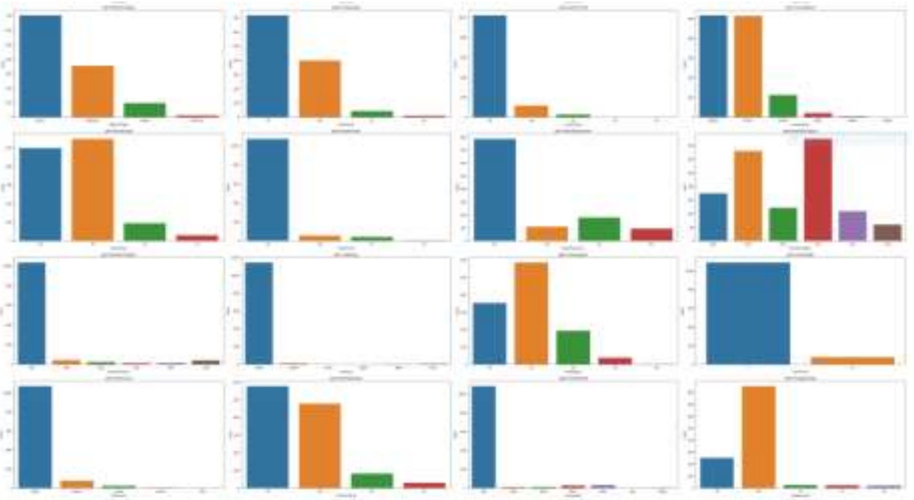
```
BsmtFinSF2          0
BsmtUnfSF           0
TotalBsmtSF         0
Heating             0
HeatingQC           0
CentralAir          0
Electrical          0
1stFlrSF            0
2ndFlrSF            0
LowQualFinSF        0
GrLivArea           0
BsmtFullBath        0
BsmtHalfBath        0
FullBath            0
HalfBath            0
BedroomAbvGr        0
KitchenAbvGr        0
KitchenQual         0
TotRmsAbvGrd        0
Functional          0
Fireplaces          0
FireplaceQu       551
GarageType         64
GarageYrBlt        64
GarageFinish       64
GarageCars          0
GarageArea          0
GarageQual         64
GarageCond         64
PavedDrive          0
WoodDeckSF          0
OpenPorchSF         0
EnclosedPorch       0
3SsnPorch           0
ScreenPorch         0
PoolArea            0
PoolQC           1161
```

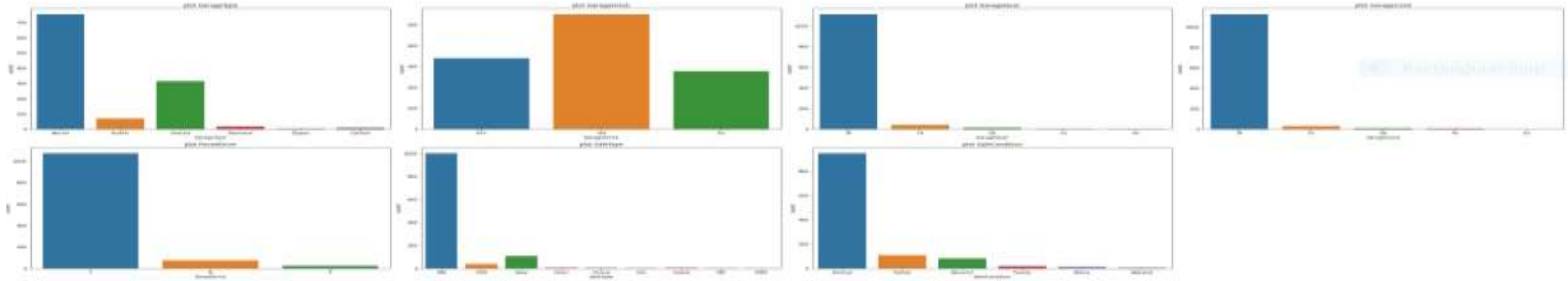


```
Fence             931
MiscFeature      1124
MiscVal             0
MoSold              0
YrSold              0
SaleType            0
SaleCondition       0
SalePrice           0
```

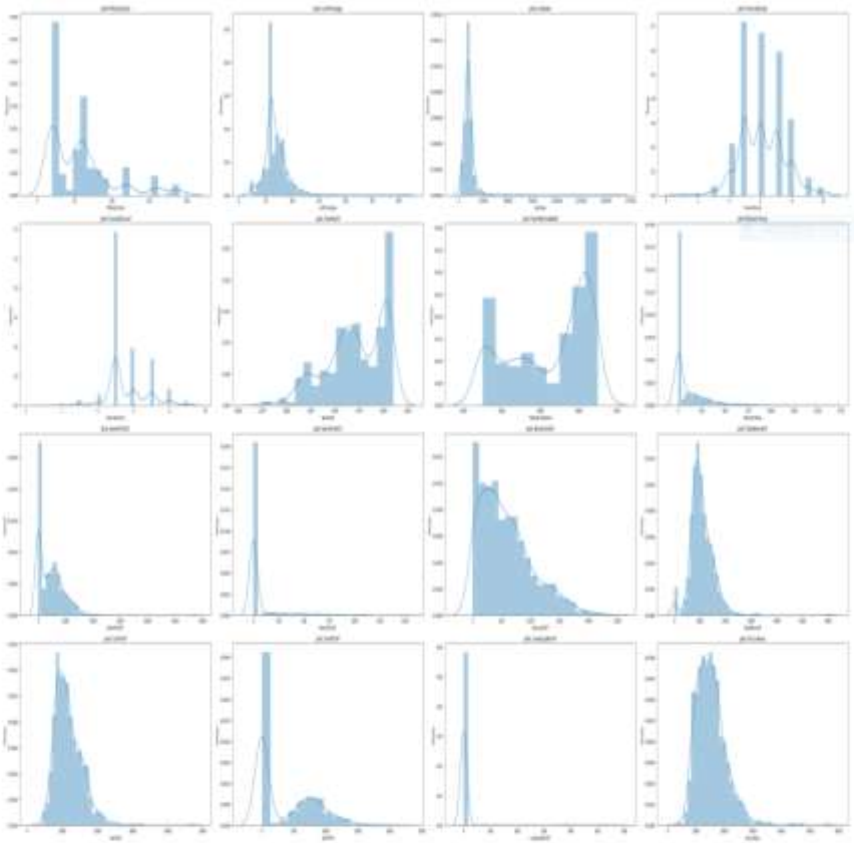
- There are missing values in 18 variables of both train and test data set.

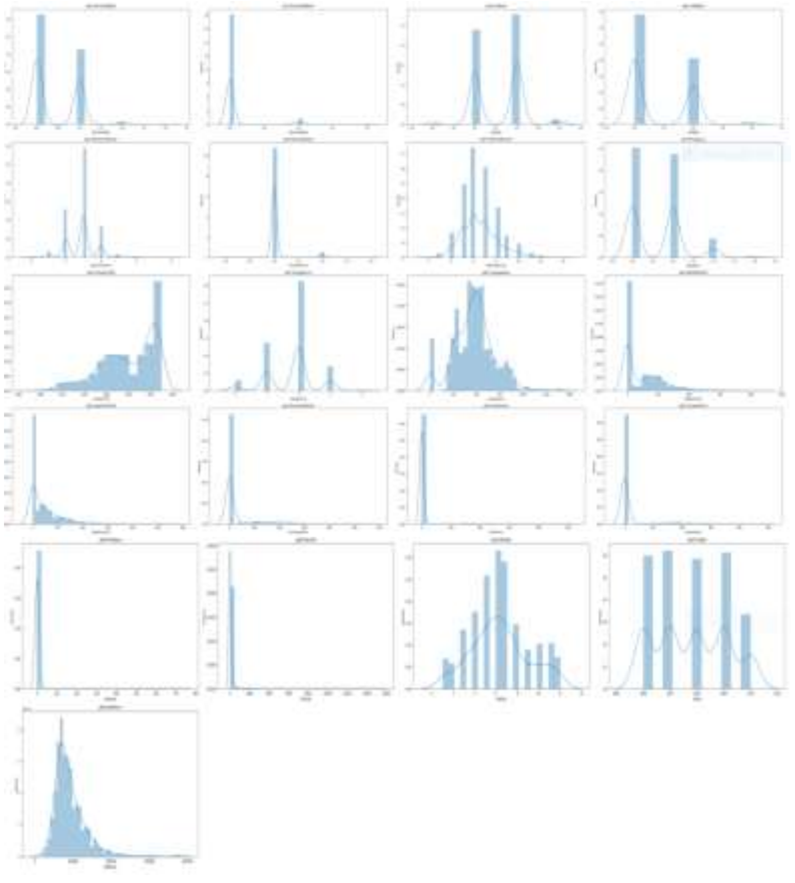
Dashboard of Categorical Data





Dashboard of Continuous Values Data





Observations Of Data Visualization

Categorical

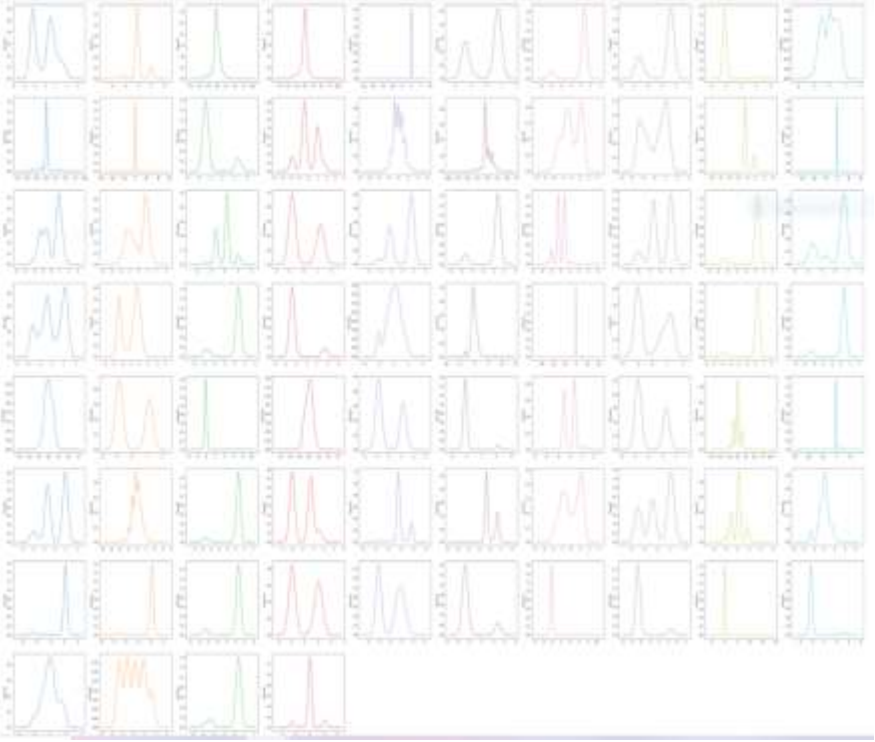
- All records consists of the "All public Utilities (E,G,W,& S)" in the 'Utilities' variable. Since the data in 'Utilities' column does not make any difference to the data set, the variable is deleted.
- It is clearly observed from the graphs that in maximum variables, the data belongs to one particular category appearing in blue the most.

Continuous

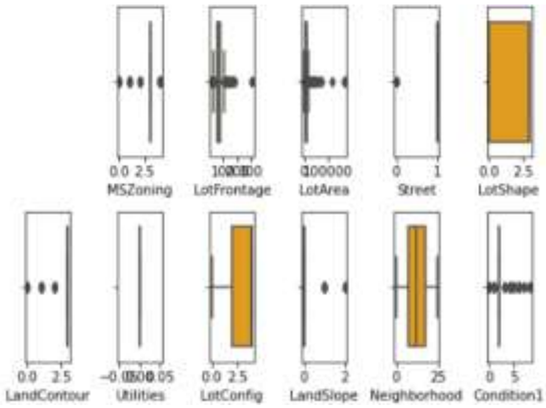
- The graphs in the previous slide clearly shows the skewness present in data.
- It is also observed that values in most of the variables is “0” for maximum records in the data set.

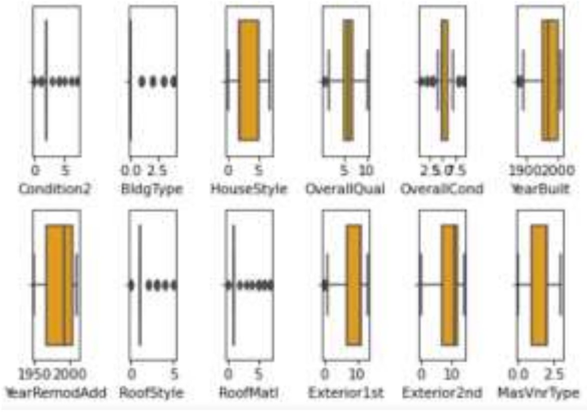
Skewness Before Transformation

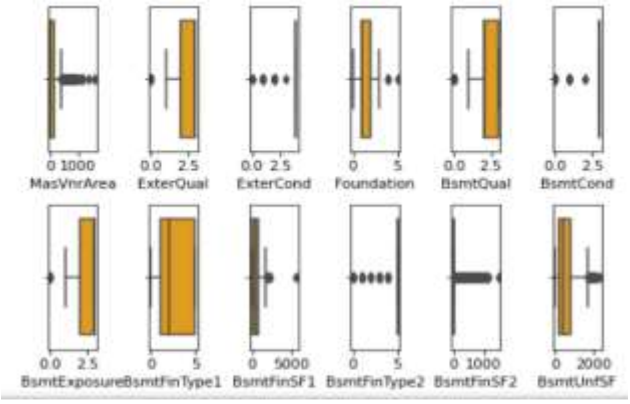
Skewness After Transformation

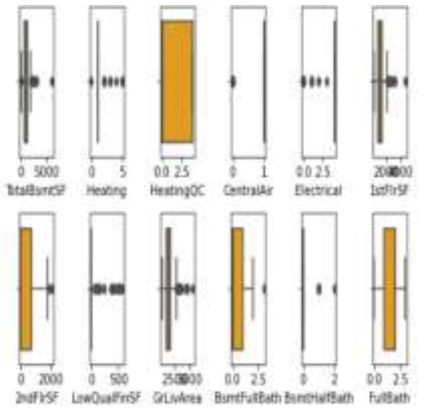


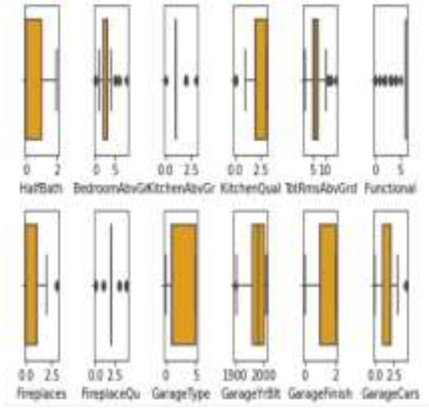
Outliers Detection

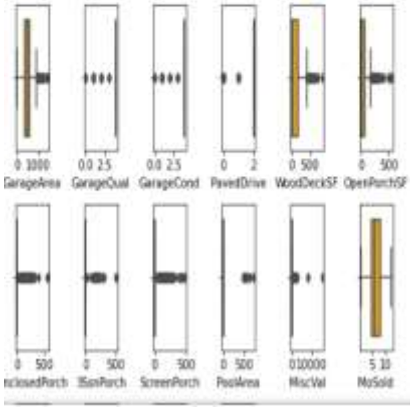


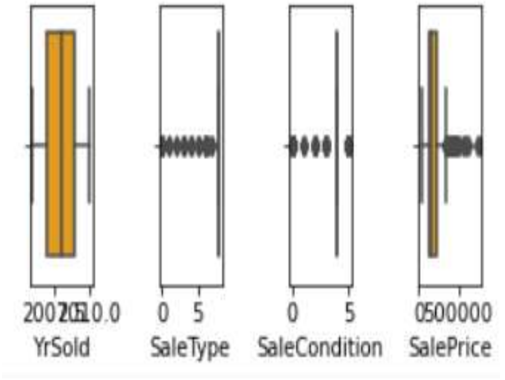












Observation

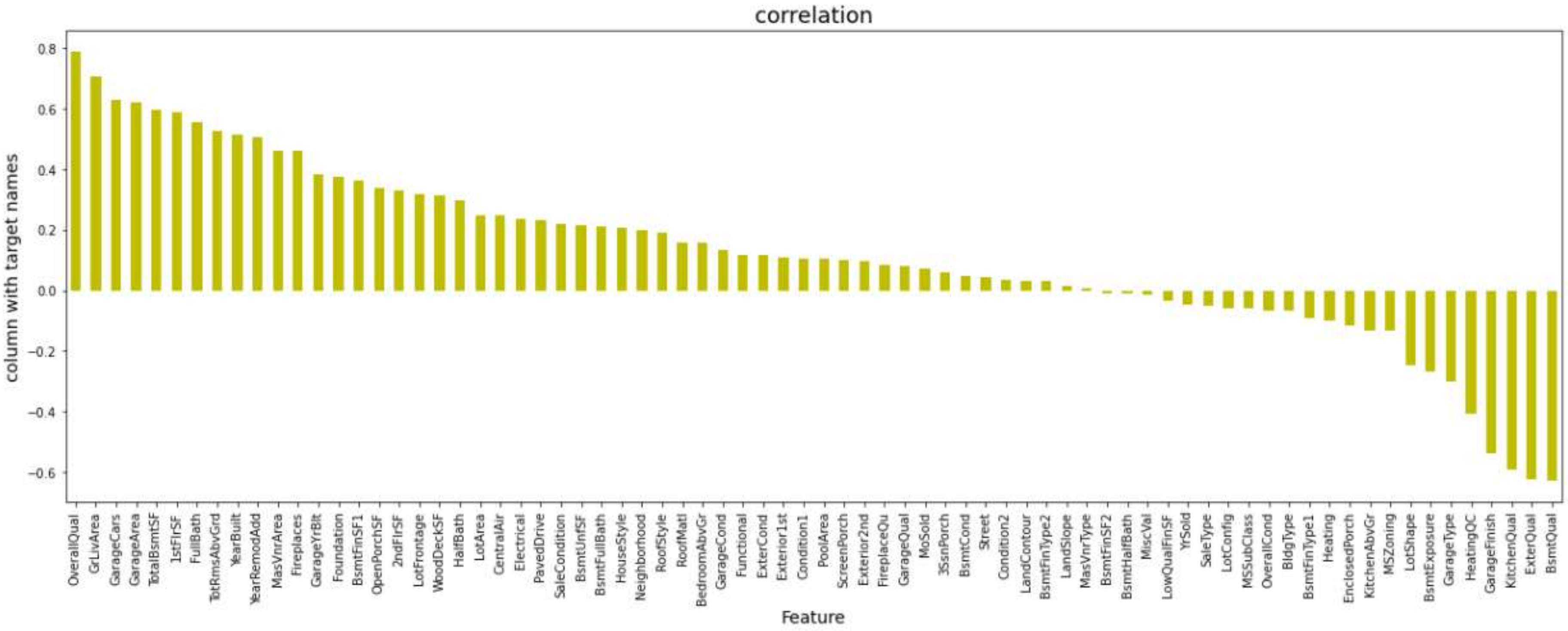
Skewness

- Skewness in some features are resolved but still can be seen in few variables even after applying transformation methods.
- In few variables, even though the skewness is high, the variables cannot be cleared from the data set because it has good correlation with the target variable and few variables have categorical data hence we don't consider having skewness or outliers.

Outliers

- Too many outliers are observed in data. Removal of outliers gives an information loss of 58.73% which is huge data loss, hence data removal is not preferred. Alternatively, data transformation using “Power Transform” is applied on data to normalize it.

Correlation of Variables with the Target Variable



Most positively correlated feature is “Overall Qual” and most negatively correlated feature is “BsmtQual”

Techniques Used

Label Encoder - The technique is used in ML to convert values in object data type into integer/ float format. In this case, all categorical data are converted from object data to float data.

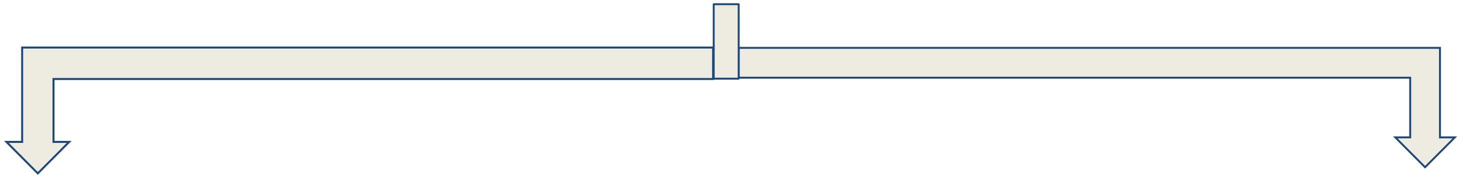
Power Transform - Statistics used to transform the skewness in data into a normal distribution curve. This helps to reduce biases and uncertainties in data for better model training.

Standard Scaling - One of the scaling techniques used for standardization of data wherein the values are centered around by removing the mean and scaling data to unit variance. In this case, the value ranges are too high within the dataset is, hence standardization plays a vital role in data normalization.

Simple Imputer - SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. It replaces the NaN values with a specified placeholder i.e. “mean”, “median” or “most frequent”.

Algorithms Used

Linear Regression



Regularization Techniques

- ElasticNet Regression
- Ridge Regression
- Lasso Regression

Ensemble Techniques

- AdaBoostRegressor
- RandomForestRegressor

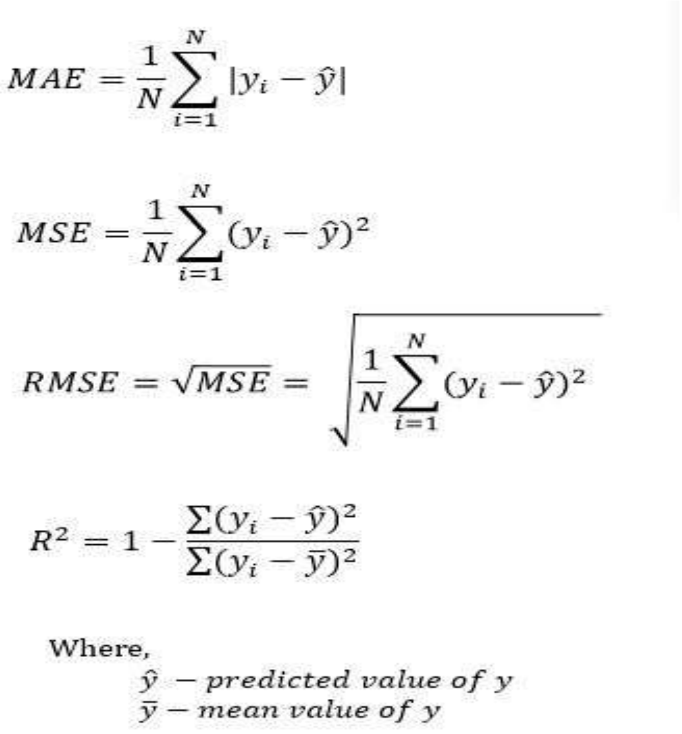
Evaluation Metrics

- R2 Score

- Mean absolute error (MAE)

- Mean squared error (MSE)

- Root Mean squared error (RMSE)



$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

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

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

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

Where,

$\hat{y}$ $-$ *predicted value of y*

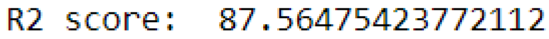
$\bar{y}$ $-$ *mean value of y*

Best Model

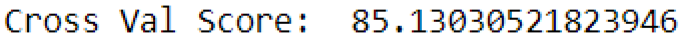
Random Forest Regressor

- Comparing all algorithms, the best model selected is the Random Forest Regressor as the r2 score and cv score is observed to be the best in Random Forest than any other Regression models used. Random Forest is applied using different parameters and hypertune the model performance.

- The Model Performance is as mentioned below.

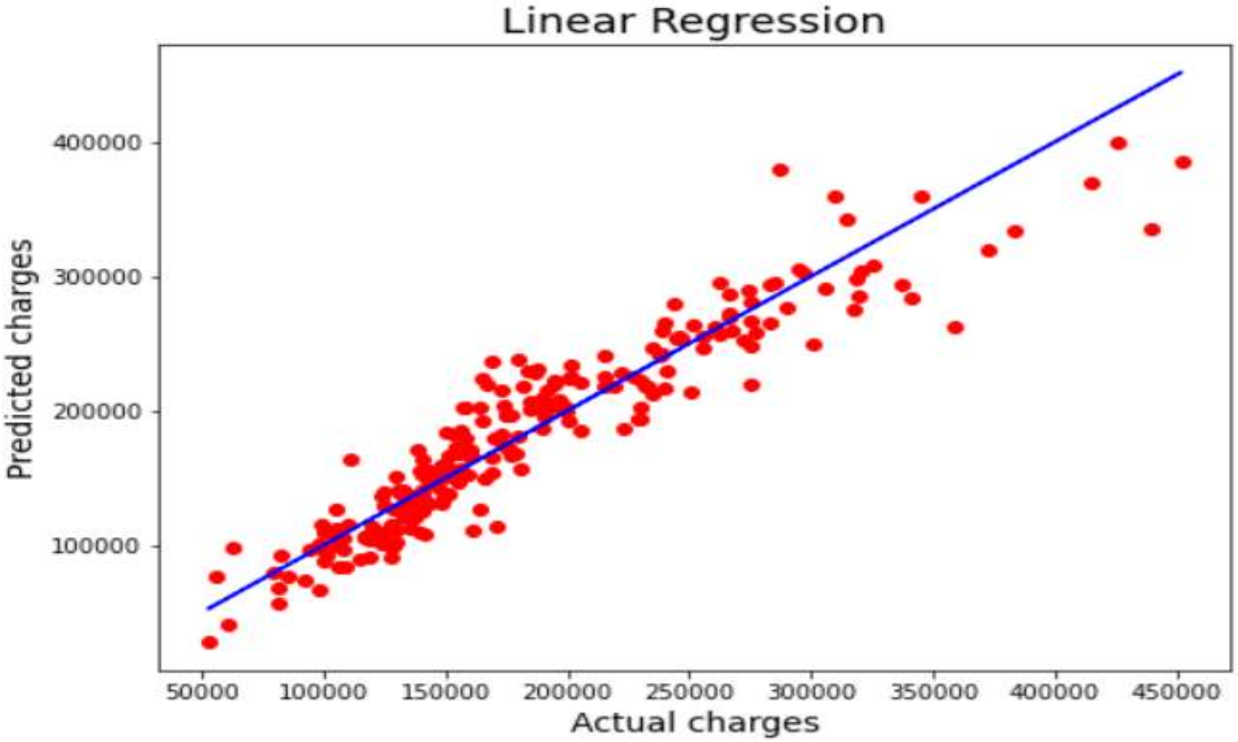


```
R2 score:  87.56475423772112
```



```
Cross Val Score:  85.13030521823946
```

Best Fit Line



The Best Fit line is covering many data points and can be seen near to the actual values which shows quite a good fit of our model.

Thank you!