

**Housing Price Prediction**

Submitted by:

Akshaykumar Torangatti

**Acknowledgement**

In successfully completing this project, I would like to thank all those who are related to this project.

I have worked on Customer Retention analysing key factors which are important aspects that affects customer shopping decisions. I have found some insights which helps e-retailers customer activation and retention.

Primarily, I would thank God for being able to complete this project with success. Then I will thank FlipRobo Technologies for providing me this opportunity, my SME Srishti Maan, under whose guidance I learned about this project. The suggestions and directions have helped in the completion of this project.

Finally, I would like to thank my parents and friends who have helped me directly or indirectly throughout my journey.

**INTRODUCTION**

**Business Problem Framing**

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. Investment is a business activity that most people are interested in this globalization era. The result of this census indicates that the younger generation will need a house or buy a house in the future. Based on preliminary research conducted, there are two standards of house price which are valid in buying and selling transaction of a house that is house price based on the developer. There are several approaches that can be used to determine the price of the house, one of them is the prediction analysis.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

**Conceptual Background of the Domain Problem**

The real estate industry has been one of the leading research projects focusing on modern economics, for its significant implications on relevant industries and fields such as construction, investment, and public welfare. In general, purchasing and investing in any real estate project will involve various transactions between different parties. Thus, it could be a vital decision for both households and enterprises. How to construct a realistic model to precisely predict the price of real estate has been a challenging topic with great potential for further research. It is generally believed by academia that correctly predicting the special price for a specific real estate is impractical since it involves plenty of factors exerting influence on the eventual cost.

The company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not.

**Review of Literature**

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy.

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price.

We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

**Motivation for the Problem Undertaken**

To understand real world problems where Machine Learning and Data Analysis can be applied to help organizations in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data. One of such domains is Real Estate.

Houses are one of the necessary needs of each person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

**Analytical Problem Framing**

**Mathematical/ Analytical Modelling of the Problem**

In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and visualized it using heatmap. Then we have used Z-Score to plot outliers and remove them.

Graphical user interface, application

Description automatically generated

**Data Sources and their formats**

The sample data is provided to Flip Robo from their client database. It is hereby given to me for this exercise. The data provided is in the form csv file, I’ll be converting it into DataFrame to perform basic operations on rows/columns like selecting, deleting, adding, and renaming. To improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

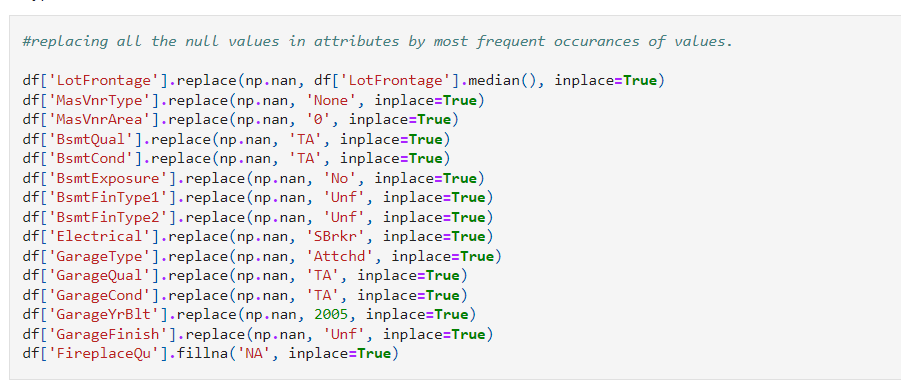
There are total of 1168 observations and 81 features including the target feature Sale Price in train dataset, and 292 observations and 80 features including target feature Sale Price in test dataset.

Graphical user interface, application, Word

Description automatically generated

**Data Pre-processing Done**

I have handled the missing values in both train and test data set. Based on the Data description I have imputed the missing data. Most of the features having nan values which were described as absence of feature in data description, so I have replaced them with ‘not available’ for each feature having nan value.



We have checked the correlation between top 10 the independent attributes and target variables. We have to convert these values using one hot encoding or Label Encoding techniques.

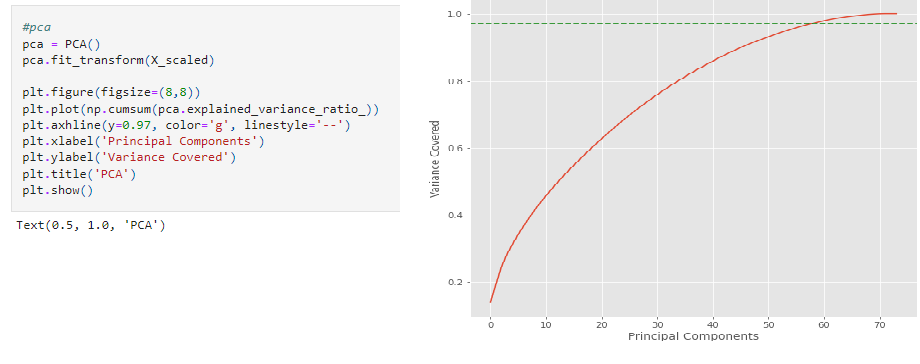
Chart

Description automatically generated with medium confidence

**Data Inputs- Logic- Output Relationships**

I have found out that with continuous numerical variable there is a Linear Relationship with the Sale Price. And for categorical variable, I have used Boxplot for each categorical feature that shows the relation with the median sale price for all the sub categories in each categorical variable. For continuous numerical variables I have used scatter plot to show the relationship between continuous numerical variable and target variable.

We observed multicollinearity in between columns so we assumed that we will be using Principal Component Analysis (PCA).



**Hardware and Software Requirements and Tools Used**

Following are the recommended hardware requirement to build and run machine learning model.

1. 7th generation (Intel Core i7 processor)
2. 8GB RAM / 16 GB RAM (recommended)

We have used following software and tools for the machine learning model.

ANACONDA

Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS.



The figure shows the important libraries I have imported to execute the project.

I have used built in Data science libraries like pandas, NumPy, Visualization libraries like matplotlib and seaborn. Jupyter Notebook, a shareable notebook that combines live code, visualizations, and text.

Machine learning libraries like scikit-learn for data pre-processing, model selection, model evaluation, SciPy for standardizing& normalizing the data,

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

The dataset provided has train and test data separately which have null values and there were outliers present in the dataset, unless removal of missing values outlier treatment there is possibility of our machine learning model overfitting the data or increase the variability in the data. We have implemented replacing null values with most occurred value from attribute and Z-Score method is implemented to reduce the outliers.

The attributes which are having less correlation with target variable have been dropped and removed based on the inference learned from heatmaps and bar plot.

To feed the dataset to model, the independent and dependent variables are to be split and the independent attributes are standardized using ‘StandardScaler’ library.

Now the data obtained is clean and is having least multicollinearity, independent and balanced data. ‘train\_test\_split’ function in model selection is used for splitting data arrays into two subsets: for training data and for testing data. We train the model using the training set and then apply the model to the test set.

We have feature engineered some of the attributes to get most of the outcome from attribute.

Text

Description automatically generated

**Testing of Identified Approaches (Algorithms)**

**Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, Word

Description automatically generated**

The algorithms we used for the training and testing are as follows:

1. Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analysing the relationship between one or more existing independent variables.

1. Random Forest Classifier

The random forest classifier is a versatile classification tool that makes an aggregated prediction using a group of decision trees trained using the bootstrap method with extra randomness while growing trees by searching for the best features among a randomly selected feature subset.

1. Decision Tree Classifier

A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class.

1. XGBoost Classifier

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.

1. KNN Classifier

K Nearest Neighbors(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition

1. SVR

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs.

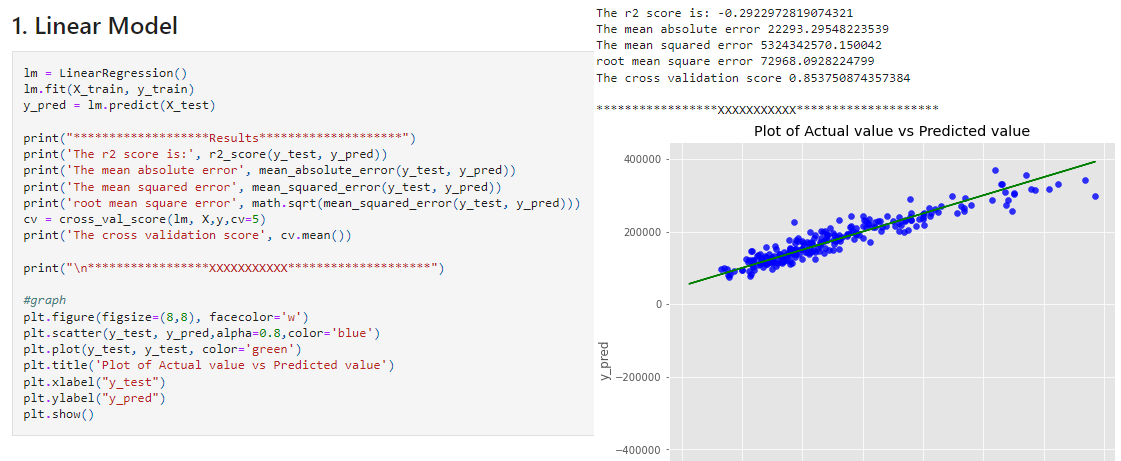
1. Ada Boost Regressor

An AdaBoost regressor. An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

1. Gradient Boosting Regressor

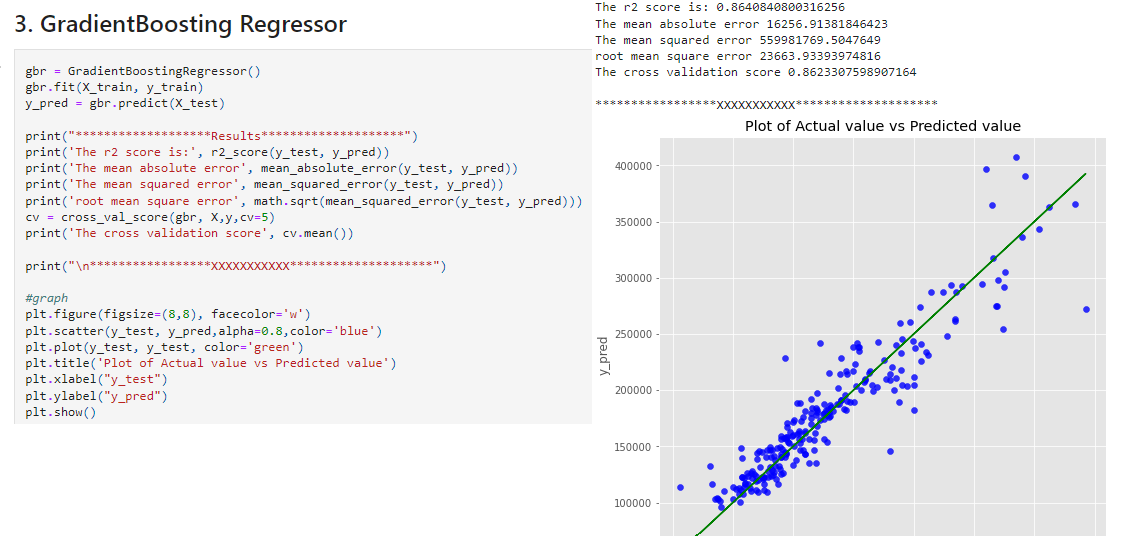
Gradient Boosting for regression. GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.

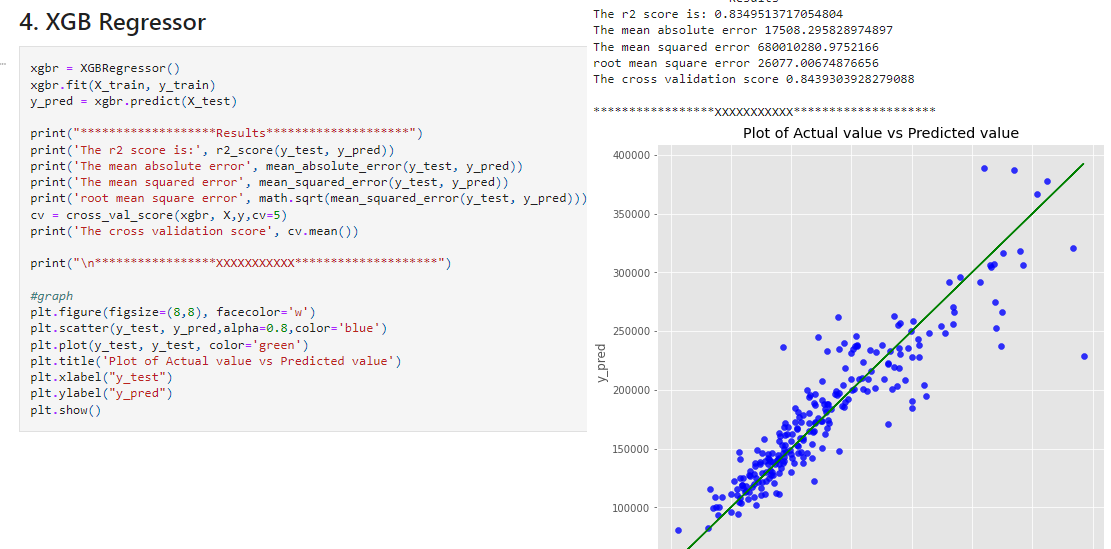
**Run and Evaluate selected models**



Chart, scatter chart

Description automatically generated

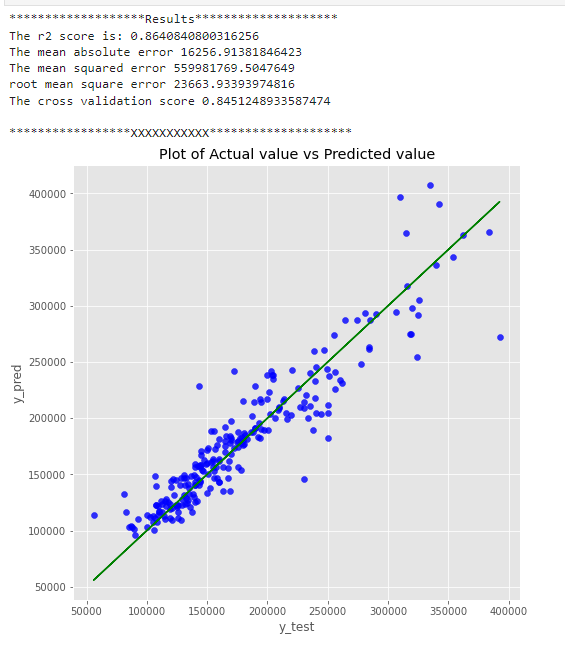




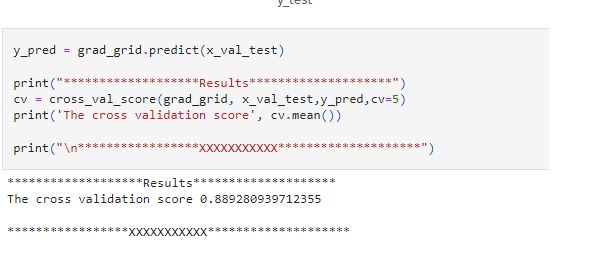
We can see too big values in MAE and MSE, that is due to individual parameters we got after the encoding of the Attributes. As per the models we trained, we got several CV score, and graphs for the values predicted. Gradient boosting algorithm has least difference between CV score to R2score value. Hence choosing Gradient boosting to be BEST Model.

Let us tune the hyper parameters and check if we can increase the accuracy of the model.





Output:



Graphical user interface, text, application, email

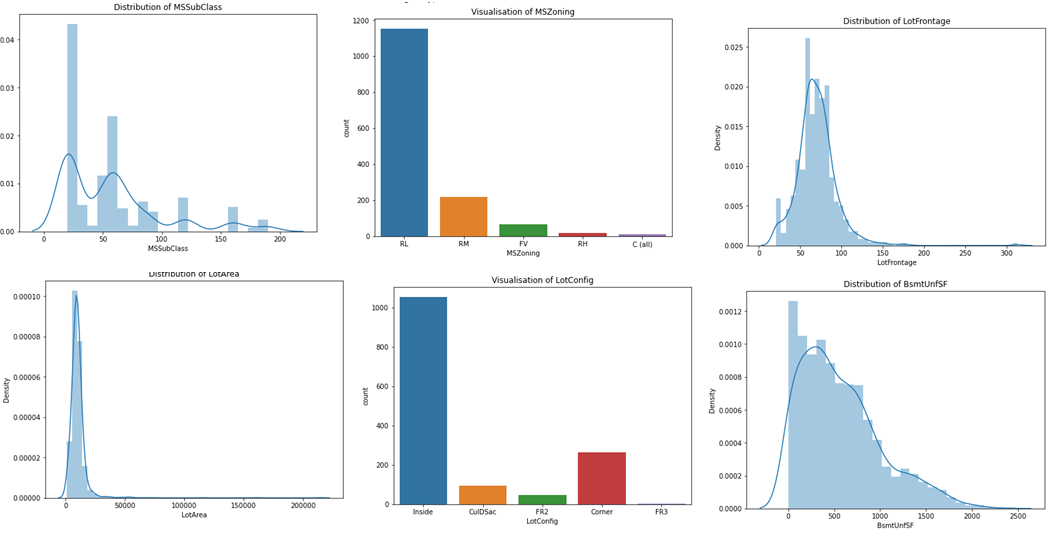
Description automatically generated

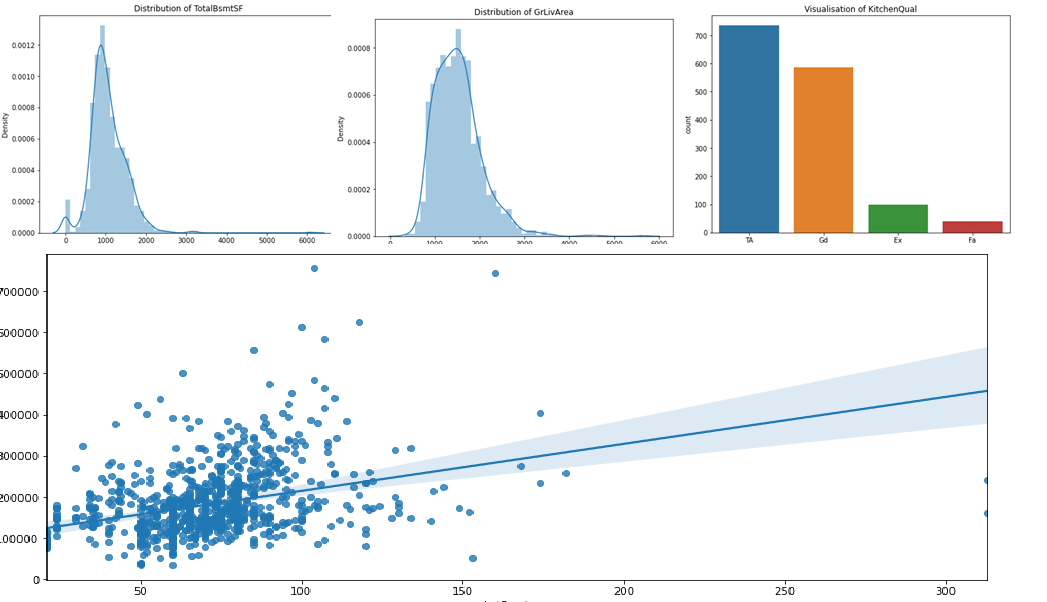
**Key Metrics for success in solving problem under consideration**

An key/evaluation metric quantifies the performance of a predictive model This typically. Following are the metrics I have used to evaluate the model performance.

I have used R2 score, mean absolute error, mean squared error and root mean squared error.

**Visualizations**

****

****

Observations:

* There are highest number of 1 story old & NEWER ALL STYLES of house in sale.
* Huge number of houses in low residential density are in the sales.
* LotFrontage: There are 259 null values and average of 70 and maximum of 313, we need to check if there are outliers.
* More than 95% of houses have paved type of street. No null values.
* Most of the houses listed are having regular shape of property. It shows that max percent of houses have level of near flat. and each house listed have all public utlilties available comprise of nearly 98%.
* As we know not many houses get corner, we see 80% of inside lot, and rest of all other lots shares for 20%.
* 15-20% of houses listed have neighborhood to North ames which is the highest. And Most of houses listed are suitable for 1 families detached and 1 story buildings.
* 90% of houses listed have an overall quality and material of average and above. We can look for outliers in the attribute. 93% of listed houses have above average overall condition.
* 70% of houses have gable type of roof style and 90% uses composite Shingle which is based on the country wide followed trend.
* Masonry veneer is not used by 75% of dwellings listed hence the value Masonry veneer area is zero.
* 75% of houses listed has average/typical exterior quality materials and condition and 90% of houses foundations are built with poured concrete or cinder blocks method.
* More than 90% houses have above inches of basement height with good quality but with no exposure.
* BsmtFinSF1, BsmtFinSF2, BsmtUnfSF 2ndFlrSF,
* LowQualFinSF,BsmtFullBath,BsmtHalfBath,HalfBath consist of 90% of zero records, null values.
* Most of houses(nearly 90%) are having features like central air conditioning, and excellent Gas forced warm air furnace heating facility.
* Most of dwellings have 1 kitchen with kitchen quality above average.
* On an average 6 or 7 number of rooms in listed houses are with above grade. with 90% typical Home functionality.
* Most of dwellings have good fireplace quality. The house has attached type of garage with good finished interior with average 2 car space in garage.
* WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal have 90% of zero as input.
* The Type of sale proceeded is Conventional Warranty Deed type with Normal Sale condition.
* The price of dwelling is distributed around the mean of 180K with maximum price being 755K.

**Interpretation of the Results**

We have trained several models above for the dataset we had prepared, and we got different results for different algorithm.

Random Forest classifier model gave us 85.5% accuracy. Decision tree classifier model gave us accuracy 63.3 %. XGboost classifier has given us 83.4% accuracy. Gradient Boosting classifier has given us 86.4% accuracy.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

In this project we have tried to show how the house prices vary and what are the factors related to the changing of house prices. The best(minimum) RMSE score was achieved using the best parameters of Gradient Boosting Regressor through GridSearchCV.

**Learning Outcomes of the Study in respect of Data Science**

1. This project has demonstrated the importance of sampling effectively, modelling and predicting data.

2. Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

3. Through different powerful tools of visualization we were able to analyse and interpret different hidden insights about the data

4. The data was improper scaled, so we scaled it to a single scale using sklearns’s package StandardScaler.

5. There were too many features present in the data, so we applied Principal Component Analysis(PCA) and found out the Eigenvalues and on the basis of number of nodes we were able to reduce our features up to 90 columns.

6. The columns were skewed due to presence of outliers which we handled through winsorization technique

7. There were lot of missing values present in different columns which we imputed on the basis of our understanding.

**Limitations of this work and Scope for Future Work**

The data set consist of large number of outliers which hinders the performance of machine learning models. Unless we solve the outlier problems, we are not reaching the best model accuracy. One can focus on collection of real time customer-oriented data which can be useful for EDA. And more inference can be provided based on the analysis.