

**Project - Housing**

Submitted by:

N THARANI

**ACKNOWLEDGMENT**

Various research papers and websites are used. Research papers like A Model of the Australian Housing Market, Is the Australian housing market in a bubble?, Australian Housing Market: Causes and Effects of Rising Price, Housing markets, economic productivity, and risk: international evidence and policy implications for Australia , etc., are used. Website like researchgate.net, geeksforgeeks, economictimes, stackoverflow.com, etc., are used as references. The data is received from the client which is their own data.

Thanking SRISHTI MAAN, my guide from FLIPROBO TECHNOLOGIES for clearing all my doubts while undergoing the project.

**INTRODUCTION**

* Business Problem

The 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. It is a very large market and there are various companies working in the domain. The company is looking at prospective properties to buy houses to enter the market. We have to build a model which will then be used by the management to understand how exactly the prices vary with the variables. They can 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.

* Conceptual Background of the Domain Problem

Inefficient housing markets can have widespread and lasting impacts on productivity and the wider economy. Poor spatial structures, for example, mean increased travel time and congestion, while lack of affordable housing near employment exacerbates social inequalities and constrains the effective operation of labour markets. Australians traditionally place a high degree of importance on homeownership which is perceived as important for stability of family life and for wealth creation. For a long period of time governments have been encouraging home ownership through direct grants to first-time home buyers. However, the proportion of homeowners without a mortgage has been decreasing over time, which is an indication of increased household indebtedness in the Australian economy.At this time penetrating the Australian housing economy would give us a profit if played safe.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Certain websites and papers that helped me to take insights from are:

1. A Model of the Australian Housing Market - Trent Saunders and Peter Tulip
2. Is the Australian housing market in a bubble? - [Justine Wang](https://www.researchgate.net/profile/Justine-Wang), [Alla Koblyakova](https://www.researchgate.net/profile/Alla-Koblyakova), [Piyush Tiwari](https://www.researchgate.net/profile/Piyush-Tiwari-7), [John Croucher](https://www.researchgate.net/profile/John-Croucher-2)
3. Australian Housing Market: Causes and Effects of Rising Price - Mohammad Mafizur Rahman
4. Housing markets, economic productivity, and risk: international evidence and policy implications for Australia - Nicole Gurran, Peter Phibbs, Judith Yates, Catherine Gilbert, Christine Whitehead, Michelle Norris, Kirk McClure, Mike Berry, Paul Maginn and Robin Goodman

* Motivation for the Problem Undertaken

The 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. 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.

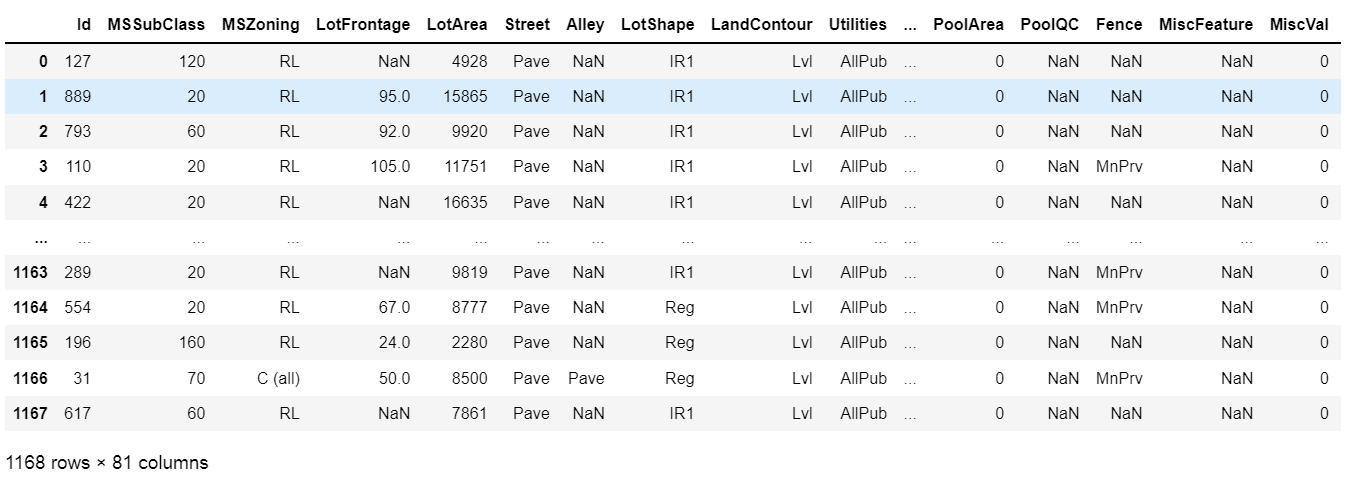
The main motivation of this project is that the company is looking at prospective properties to buy houses to enter the market and to purchase houses at a price below their actual values and flip them at a higher price.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Various statistical, mathematical, analytical algorithms are used. Experimental design, outliers detection using Z-score, transformation, scaling, re-sampling methods, statistical hypothesis tests, estimation statistics, data manipulation, feature selection, feature reduction, balancing the dataset, etc are done for the problem.

* Data Sources and their formats

The data is received from the client with necessary data description. The data is in excel format which can be imported using pandas from local library easily.

* Data Preprocessing Done

Based on the data description, features which does not help to determine the housing prices are removed. Id which is not going to help is removed. Utiities is a feature which has the same data for the whole dataset which will not help in determining the result. Check for null values, string data type to encode. Dropped Alley, MiscFeature, PoolQC, FenceCheck since the NaN values in these columns cannot be treated. The dataset doesn't have any duplicates. Check for outliers and check the curve for normalization. Correlation and multi collinearity are checked and treated.

* Data Inputs- Logic- Output Relationships

There are various features used to detect the defaulter. All the features which are helping to detect the defaulter is visualised using matplotlib and seaborn. The relationship between the features are determined. After data cleaning, the relationship between input and output variables is fairly linear.

* State the set of assumptions (if any) related to the problem under consideration

In housing/real estate industry, pricing a property is very difficult for which we are using the past data. Sometimes we come with a model and it does not work because of any economic crisis or war, etc., So, we have to be prepared for any issues that can be resolved at any situations.

* Hardware and Software Requirements and Tools Used

Any laptop and computer can be used as hardware. Processor used is Intel(R) Core(TM) i7-4510U CPU. System type is 64-bit operating system, x-64 based processor. RAM of the systerm is 8.00 GB. Microsoft Windows 8.1 version 6.3 is the OS used. Python 2.7.10 is used with the interface Jupyter notebook with many installed libraries.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

There are statistical and analytical approaches in problem-solving. Data cleaning, data mining, wrangling comes under statistical approach whereas data manipulation, feature selection, feature reduction, creating ML models comes under analytical approach.

**Statistical:**

* + - ***Experimental Design*** - Methods to design systematic experiments to compare the effect of independent variables on an outcome
  + Univariate - measure relies on only one variable - Statistical summary, information on variables, etc
  + Bivariate - measure relies on two variables - Correlation, multicollinearity
    - ***Outliers detection using Z-Score (z = (x-μ)/σ)*** - where x is the observed value, μ is the population mean, and σ is the population standard deviation
    - ***Transformation*** - Transformed the data using ordinal encoder for the data to be analysed
    - ***Scaling*** - Scaled the data using Standard Scaler to get unbiased result.
    - ***Re-sampling Methods*** - Train Test Split is used to systematically split a dataset into subsets for the purposes of training and evaluating this predictive model.
    - ***Statistical Hypothesis Tests*** - Cross validation quantifies the likelihood of observing the result given an assumption or expectation about the result whether the model is overfitting/underfitting or fitting good.
    - ***Estimation Statistics*** - GridSearchCV is used to quantify the best parameter from the listed to fit in the model and give better result. It uses data analysis framework which has a combination of effect sizes, confidence intervals, precision planning, and meta-analysis to plan experiments, analyze data and interpret results.

**Analytical:**

It concerns the design and development of algorithms.

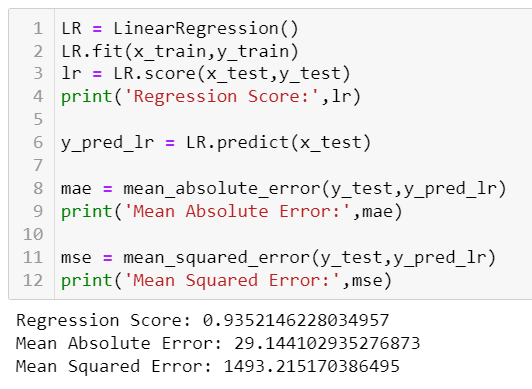
* ***Data Manipulation*** - Check for skewness and if there is no more skewed data, not necessary to treat it.
* ***Feature selection*** - Check the relationship between independent and dependent variables. Select the features which gives meaning to the problem based on the subject knowledge.
* ***Feature reduction*** - Reduce the features which does not give proper solution to the problem using Principal Component Analysis. But, the data is scraped by us. So, all the variables are in need for the analysis.
* ***Balancing the dataset*** - Balance the dataset and normalize the data to get better result.
* Testing of Identified Approaches (Algorithms)

Various evaluation metrics can be used for this regression type of model. Some of the popular algorithms are the following:

* Linear Regression
* k-Nearest Neighbors
* Decision Trees
* Support Vector Machine
* Random Forest Regression
* Run and Evaluate selected models

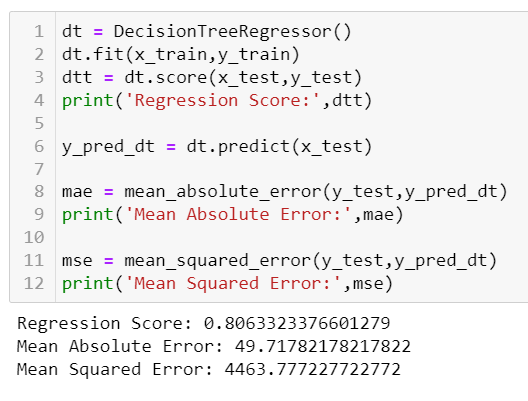
Regression Score - Percentage of correct predictions for dataset

***Linear Regression:***



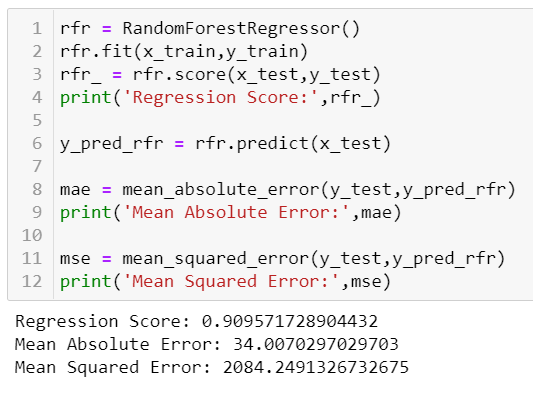
Regression Score - 0.93

***Decision Tree:***



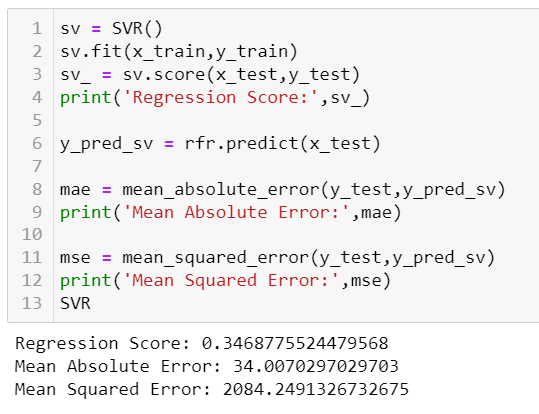
Regression Score - 0.80

***Random Forest Regressor:***



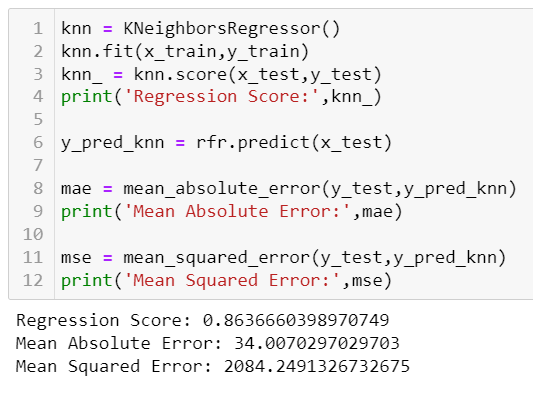
Regression Score - 0.90

***Support Vector Regression:***



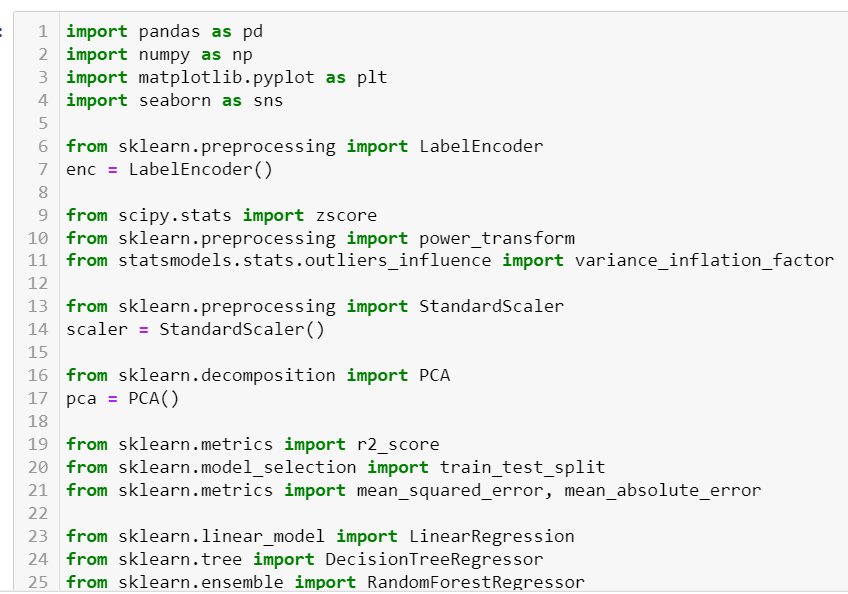
Regression Score - 0.34

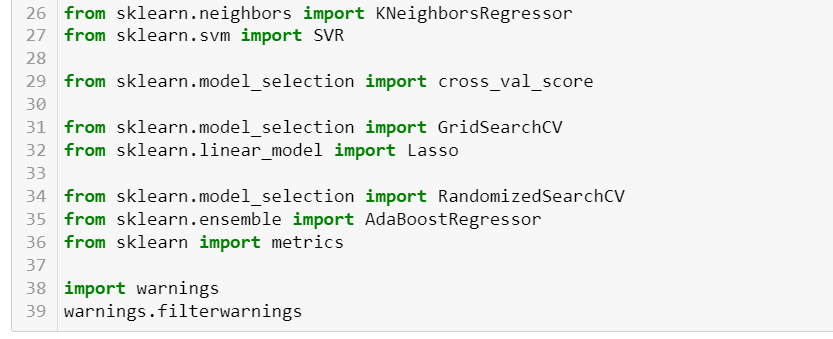
***KNeighbors Regressor:***



Regression Score - 0.86

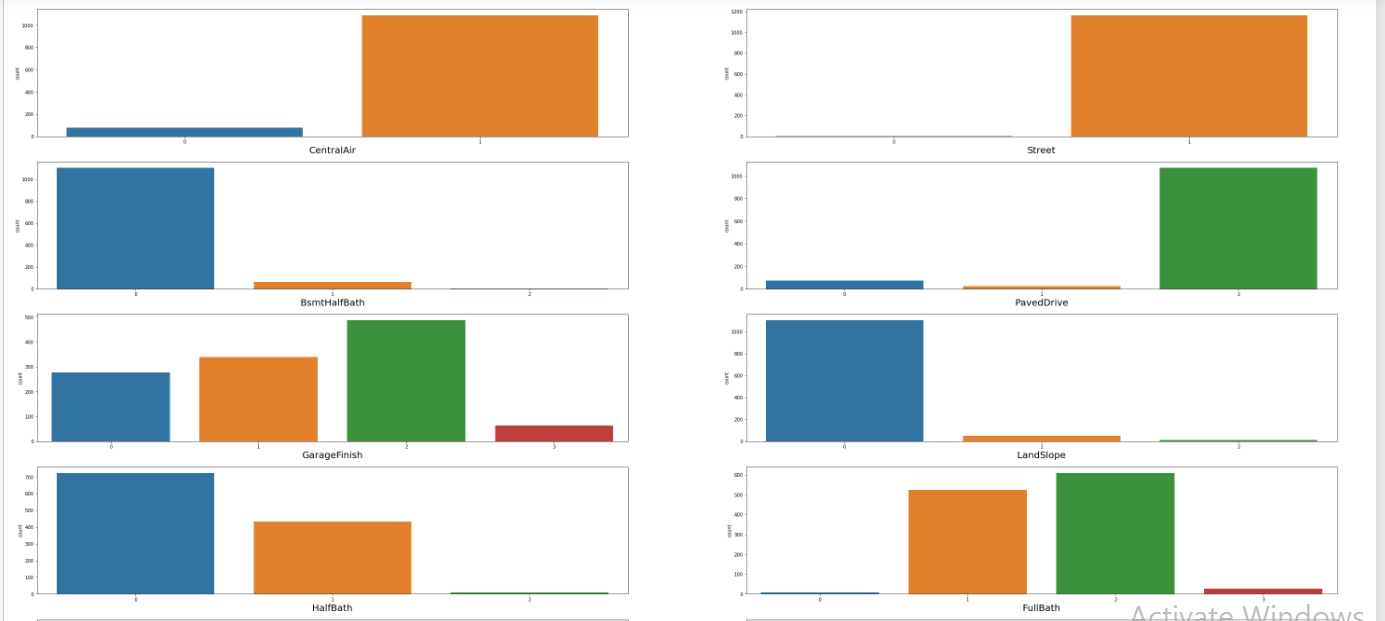
* Key Metrics for success in solving problem under consideration

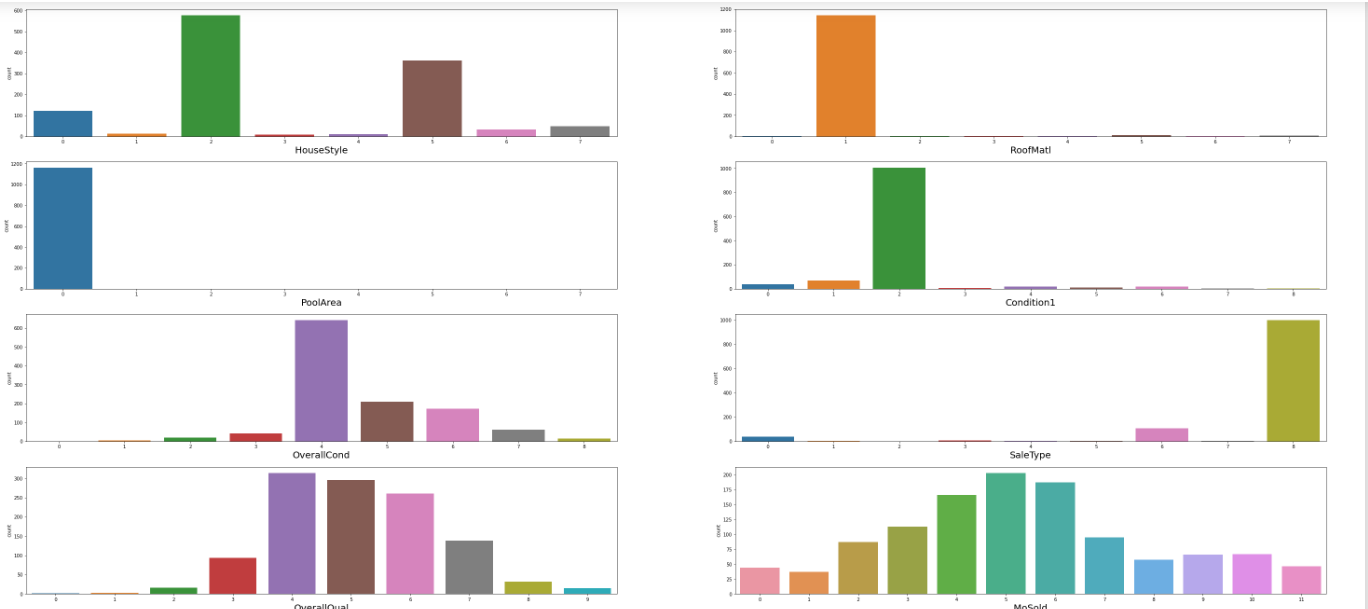




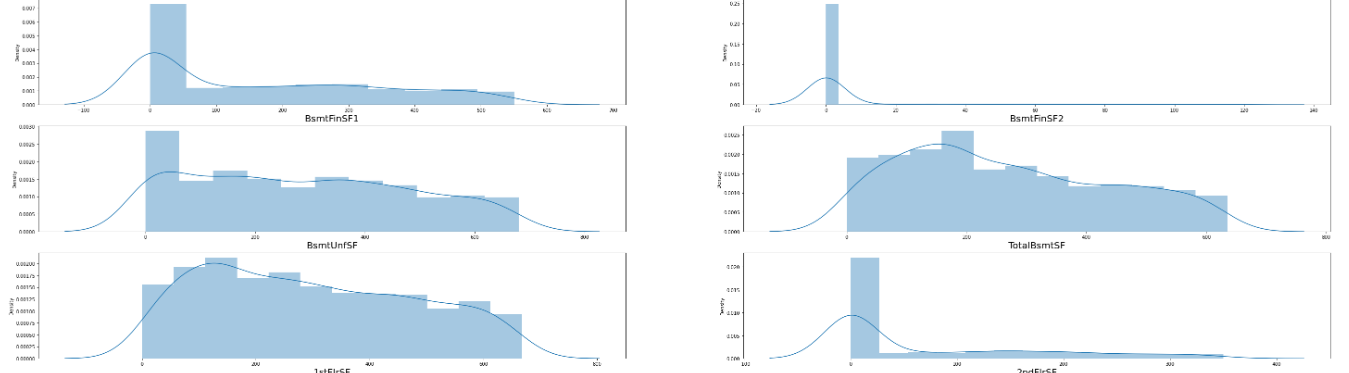
|  |  |
| --- | --- |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| Ordinal Encoder | Encode the data from string to integers |
| Variance\_inflation\_factor | To check the multicollinearity |
| z-score | To check and remove outliers with fixed threshold |
| Standard Scaler | To standardise the data and normalise the curve |
| R2 score, mean\_squared\_error, mean\_absolute\_error | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Linear Regression, DecisionTreeRegressor, RandomForestRegressor, SVR, KNeighborsRegressor | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| RandomisedSearchCV | For hyper parameter tuning |
| AdaBoostRegressor | The weights of instances are adjusted and checks the fitting of model by applying the estimators |
| Metrics | It is used to track the performance by quantitatively assessing |
| **Libraries** | **Usage** |
| Pandas and numpy | Importing dataset, data cleaning, data wrangling and exploratory data analysis |
| Matplotlib and seaborn | Visualisation libraries |
| Variance\_inflation\_factor | To check the multicollinearity |
| z-score | To check and remove outliers with fixed threshold |
| Power transform | Power transformation to remove skewness |
| Standard Scaler | To standardise the data and normalise the curve |
| PCA | Feature reduction |
| SMOTE and Counter | To balance the imbalanced dataset |
| Accuracy score, confusion matrix, classification report | For concluding the results |
| Train\_test\_split | To separate the training and testing dataset |
| Logistic Regression, DecisionTreeClassifier, RandomForestClassifier, SVC, KNeighborsClassifier | All these are machine learning algorithms to find the results |
| Cross\_val\_score | To check the best fitting of the model |
| GridSearchCV | For hyper parameter tuning |
| Plot\_roc\_curve | To check whether the model is good by checking the area under the curve |

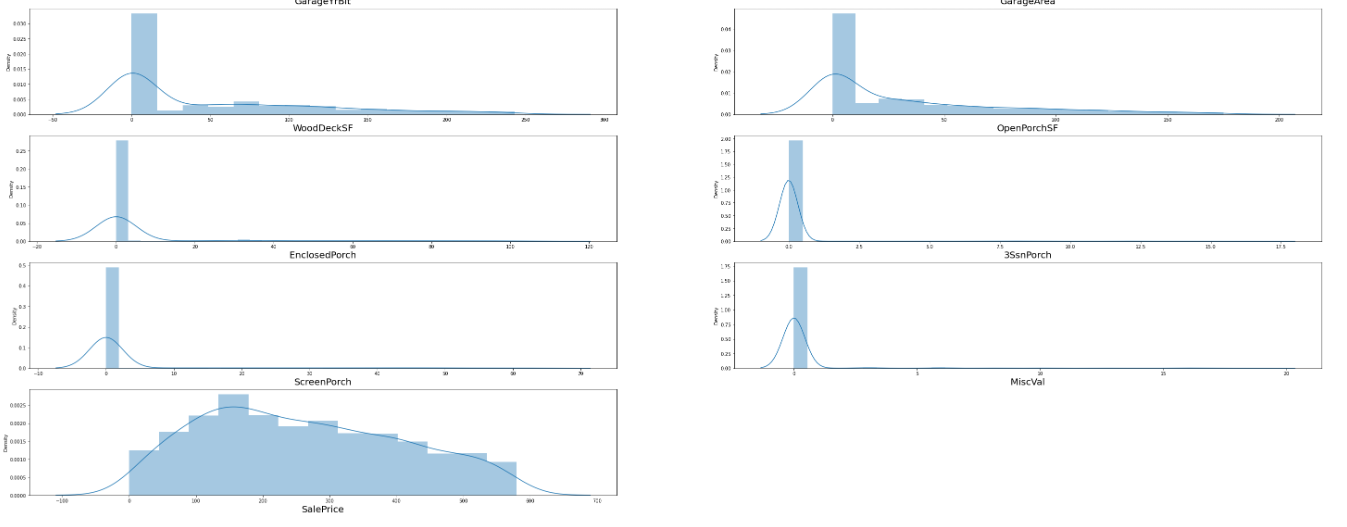
* Visualizations

Matplotlib and seaborn is used for visualisations



Categorical variables are visualized separately to understand the distribution where the target is imbalanced using countplot.





Continuous variables are visualized separately to understand the distribution where they are normally distributed.

To detect outliers boxplot is used which showed many outliers

* Interpretation of the Results
* Categorical variables and continuous variables are visualized separately to understand the distribution where the target is imbalanced and the other variables are normally distributed
* To detect outliers boxplot is used which showed many outliers
* The relationship between dependent and independent variables are visualised where it looks fairly linear
* In data preprocessing, there is no duplicates found. The data has been encoded to avoid biased result.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The relation between the input and output variables looks fairly linear after data cleaning. There is no correlation. Multicollinearity is found in the variables - YearBuilt, RoofMatl, 1stFlrSF, 2ndFlrSF, KitchenAbvGr, PoolArea. From the models, linear regression gives better result but on comparing the cross validation, random forest regressor fits well along with the model score.

Hyper parameter tuning is done to check whether the accuracy can be improved, and the accuracy has increased. So, use the tuned model to predict the flight ticket price.

* Learning Outcomes of the Study in respect of Data Science

Visualising the categorical and continuous variable tells how the variable is distributed and the balancing of the dataset. Checked the correlation. The data already does not have null value and the data is encoded using ordinal encoder.

Decision tree and Random Forest Regressor is good to go with the model. Random forest regression, KNN takes time for training. SVR takes very long time to train the model. Cross-validation took time to complete and checked whether the model is underfitting and overfitting the model.

Random Forest Regressor gave 91% Regression score approximately. After hyper parameter tuning, the model score has increased. So, we can use tuned model for further predictions.

* Limitations of this work and Scope for Future Work

We have got 92 as the regression score after hyper parameter tuning. There is scope to improve the cross validation score with the model score and by changing the parameters in hyper parameter tuning.