

A Presentation on Housing Projects price predictions



Submitted To: Khusboo Garg
(SME of Internship Batch no-28)

Submitted By: Abhishek Behera
Internship Batch no-28

Content

- Introduction
- Analytical Problem Framing
 - I. Exploratory Data Analysis (EDA)
 - II. Visualizations
- Data Pre-Processing on train and test datasets
- Model/s Development and Evaluation
- Performing hyper parameter tuning, saving the best model and predicting the label
- Conclusion and future work discussion



INTRODUCTION

- House Price prediction, is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improve Real Estate efficiency. 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. There are three factors that influence the price of a house which include physical conditions, concept and location.
- Houses are one of the necessary need 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 modeling, Market mix modeling, 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

- As we are provided with two sets of data, one is for training and other for testing. Here we need to build a machine learning model using train dataset and then by using that model we will make predictions for test data set.
- Both the datasets are in csv format, train dataset has 1168 rows and 81 columns whereas test dataset has 292 rows and 80 columns. Here in the test dataset we do not have the target label and need to predict the same.
- As we have to predict house sale prices in this problem which is a continuous data, I will be using different regression machine learning models.

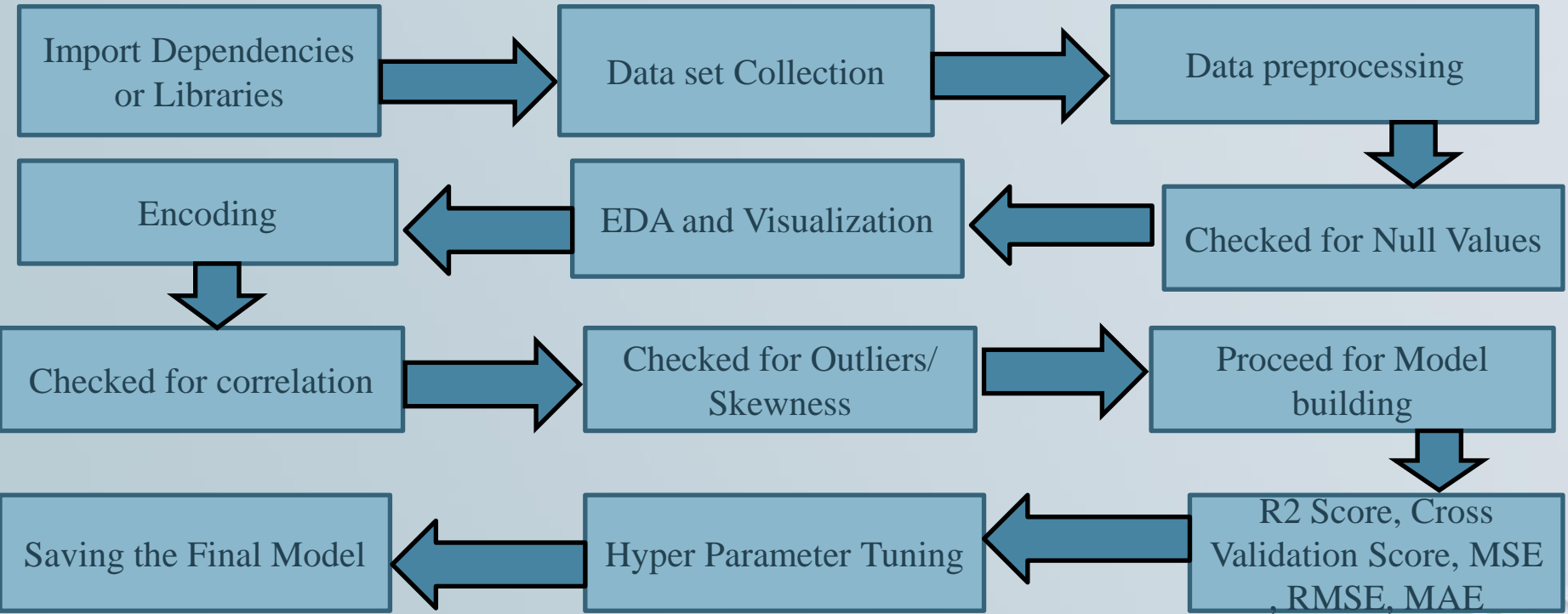


Hardware & Software Specification

Requirements	Tools Used
Hardware	<p>RAM: 8 GB</p> <p>CPU : Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz</p> <p>GPU : Intel(R) HD Graphics 5500and NVIDIA GeForce 940M</p>
Software	<p>Programming language : Python</p> <p>Distribution : Anaconda Navigator</p> <p>Browser based language shell : Jupyter Notebook</p> 
Libraries/Packages	<div> <p>Pandas</p>  </div> <div> <p>NumPy</p>  </div> <div> <p>Matplotlib</p>  </div> <div> <p>seaborn</p>  </div> <div> <p>scikit-learn</p>  </div>

DATA ANALYSIS & MODEL BUILDING FLOW

CHART



DATA PRE-PROCESSING

- Importing the necessary dependencies and libraries.
- Reading the CSV file and converted into data frame.
- Checking the data dimensions for the original dataset.
- Looking for null values and accordingly fill the missing data.
- Checking the summary of the dataset.
- Checking unique values.
- Checking all the categorical columns in the dataset.
- Visualizing each features using matplotlib and seaborn.
- Performing encoding using the ordinal encoder on categorical features.
- Checking for co-relation/multi-collinearity in a heat map.
- Checking for Outliers/Skewness using boxen plot and distribution plot.
- Perform Scaling using Standard Scaler method.
- Checking for the final dimension of dataset to confirm the input details.
- Creating train test split and the best random state found in the range 1-1000.



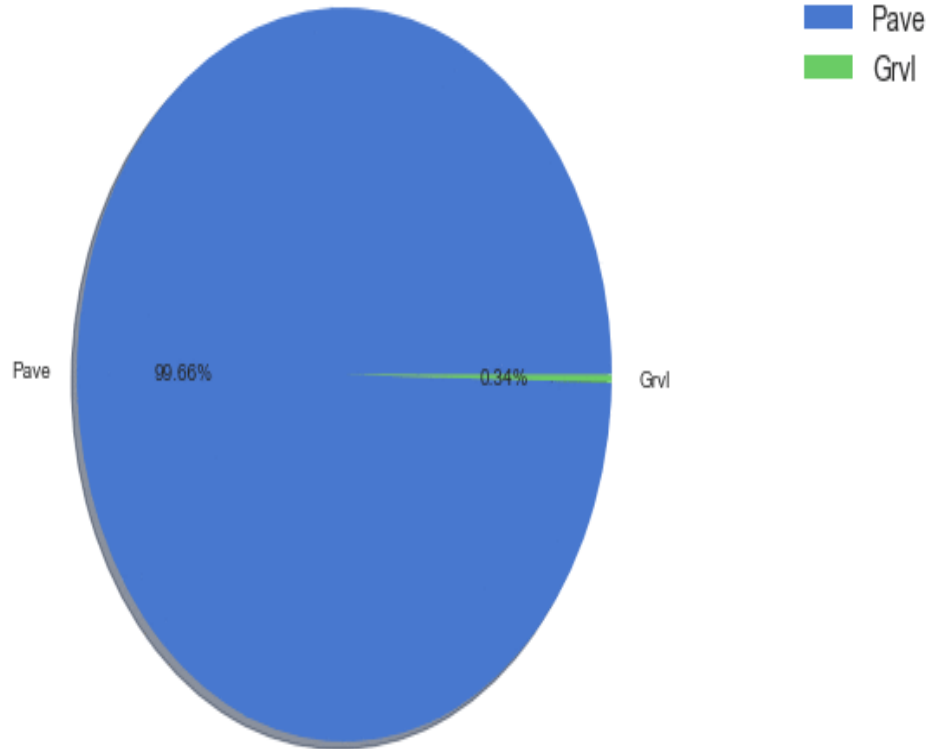
EXPLORATORY DATA ANALYSIS (EDA)

<u>1. Univariate Analysis</u>	<u>2. Multivariate Analysis</u>	<u>3. Correlation of Dataset</u>	<u>4. Correlation with Target variable</u>	<u>5. Conclusion</u>
Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable.	Multivariate analysis is a set of statistical techniques used for analysis of data that contain more than one variable.	Correlation is used to test relationships between quantitative variables or categorical variables.	Correlation with the target variable to know how the data is related.	Summary with the conclusion of all the analysis.



PIE PLOT

- A Pie Chart is a circular statistical plot that can display only one series of data.
- The area of the chart is the total percentage of the given data.
- The area of slices of the pie represents the percentage of the parts of the data.

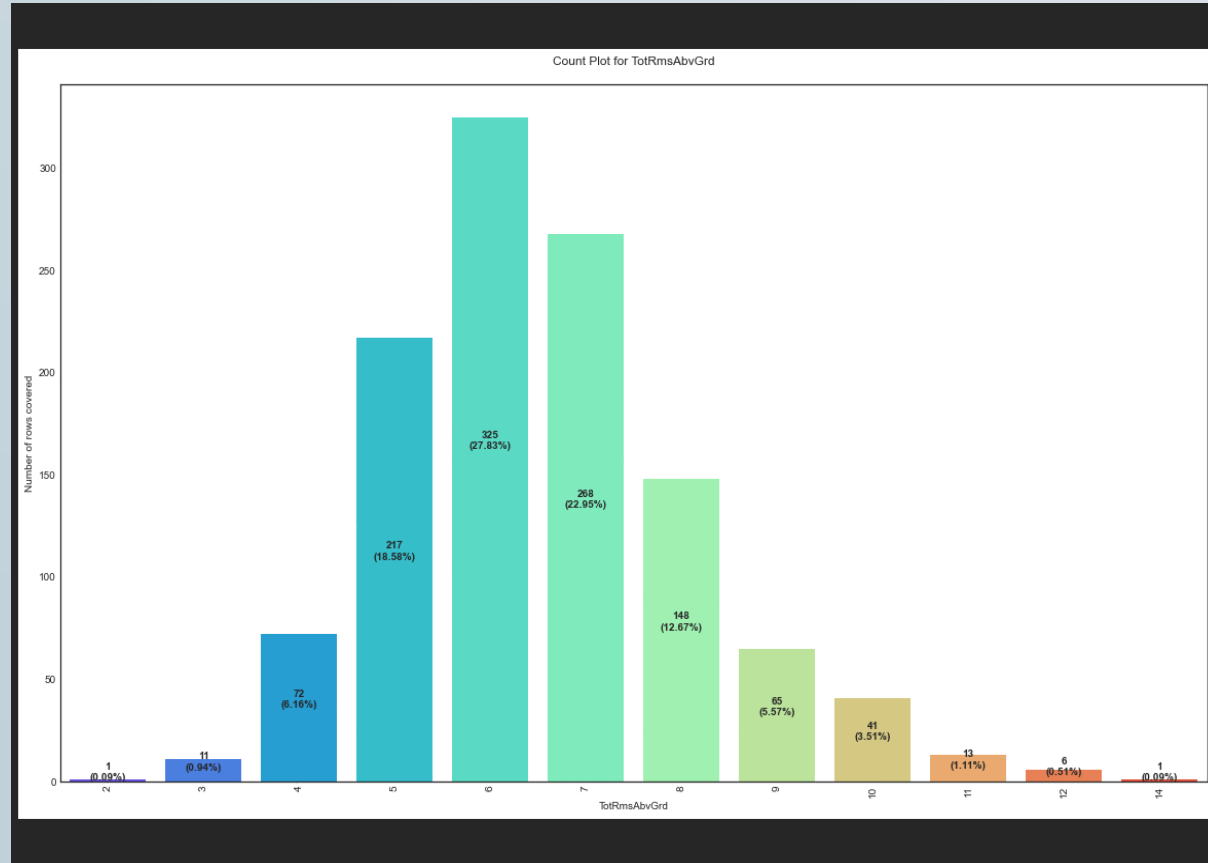


COUNT PLOT

➤ Count plot method is used to show the counts of observations in each categorical bin using bars.

➤ Parameters : This method is accepting the following parameters that are described below: x, y

➤ This parameter takes names of variables in data or vector data, optional inputs for plotting long-form data.

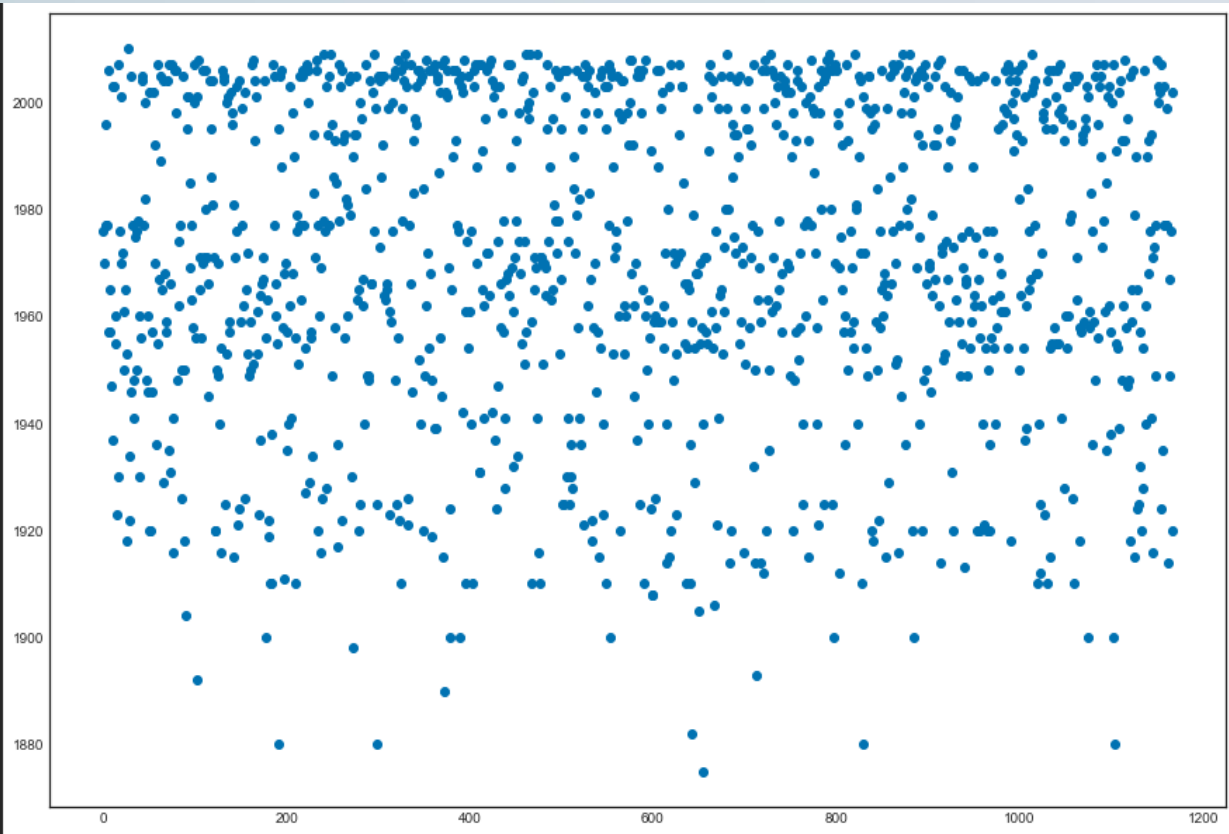


SCATTER PLOT

➤ Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them.

➤ The scatter method in the matplotlib library is used to draw a scatter plot.

➤ Scatter plots are widely used to represent relation among variables and how change in one affects the other.

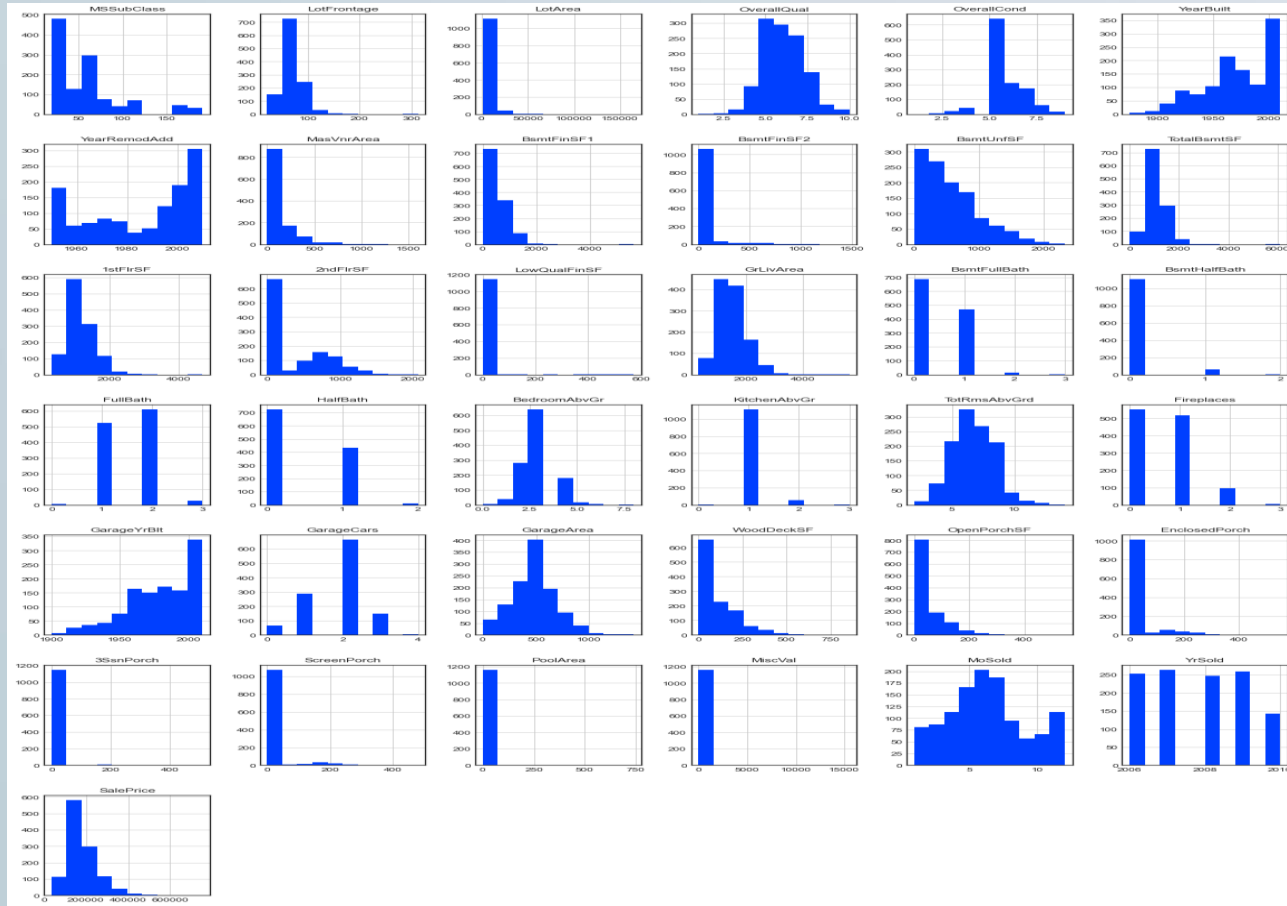


HISTOGRAM

➤ A histogram is basically used to represent data provided in the form of some groups.

➤ It is an accurate method for the graphical representation of numerical data distribution.

➤ It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

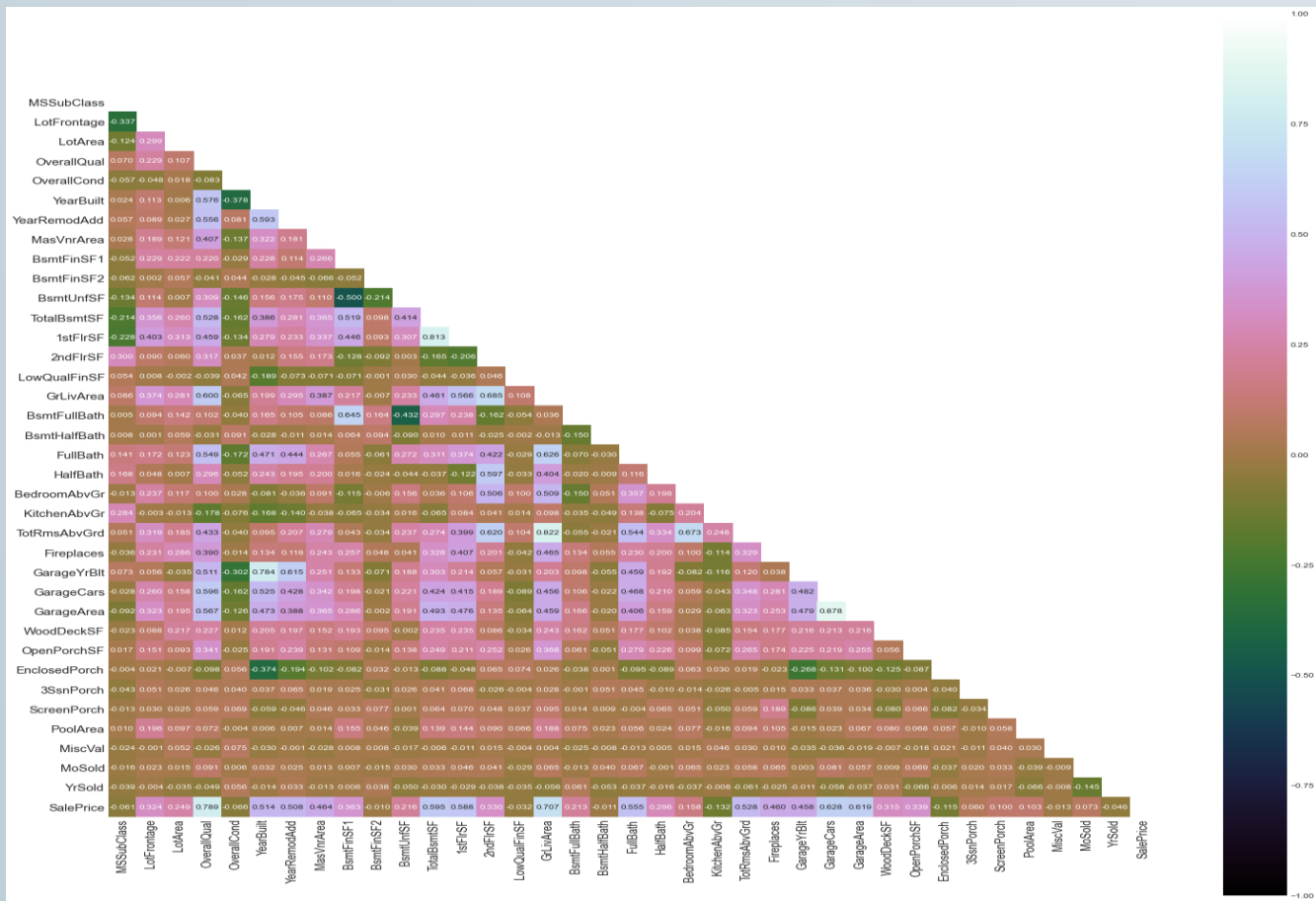


HEATMAP

➤ A heat map contains values representing various shades of the same color for each value to be plotted.

➤ Usually the darker shades of the chart represent higher values than the lighter shade.

➤ For a very different value a completely different color can also be used.

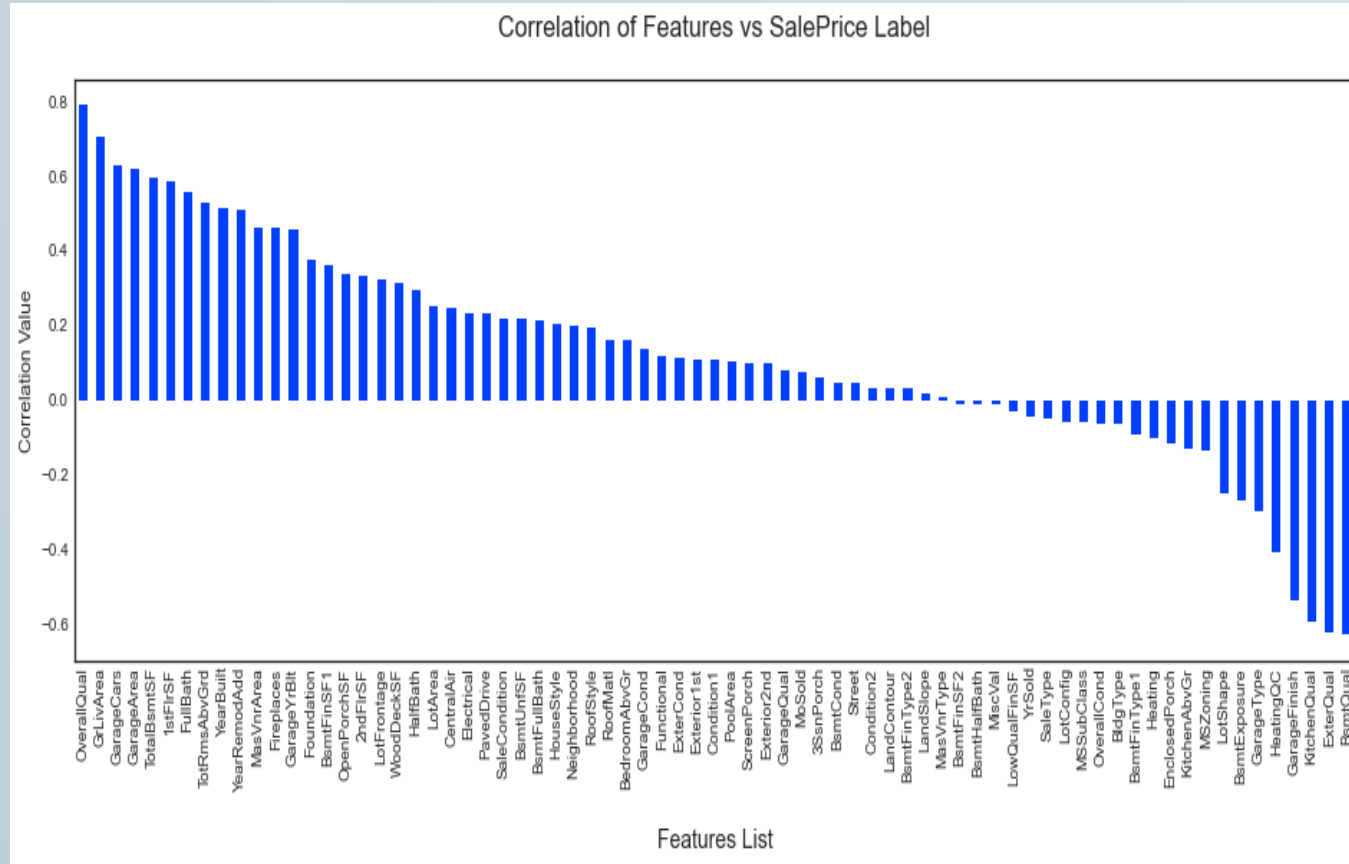


BAR GRAPH

➤ Bar graphs are used to compare things between different groups or to track changes over time.

➤ Here we are comparing the correlation values between the feature columns and the target label column which is Sale Price in our scenario.

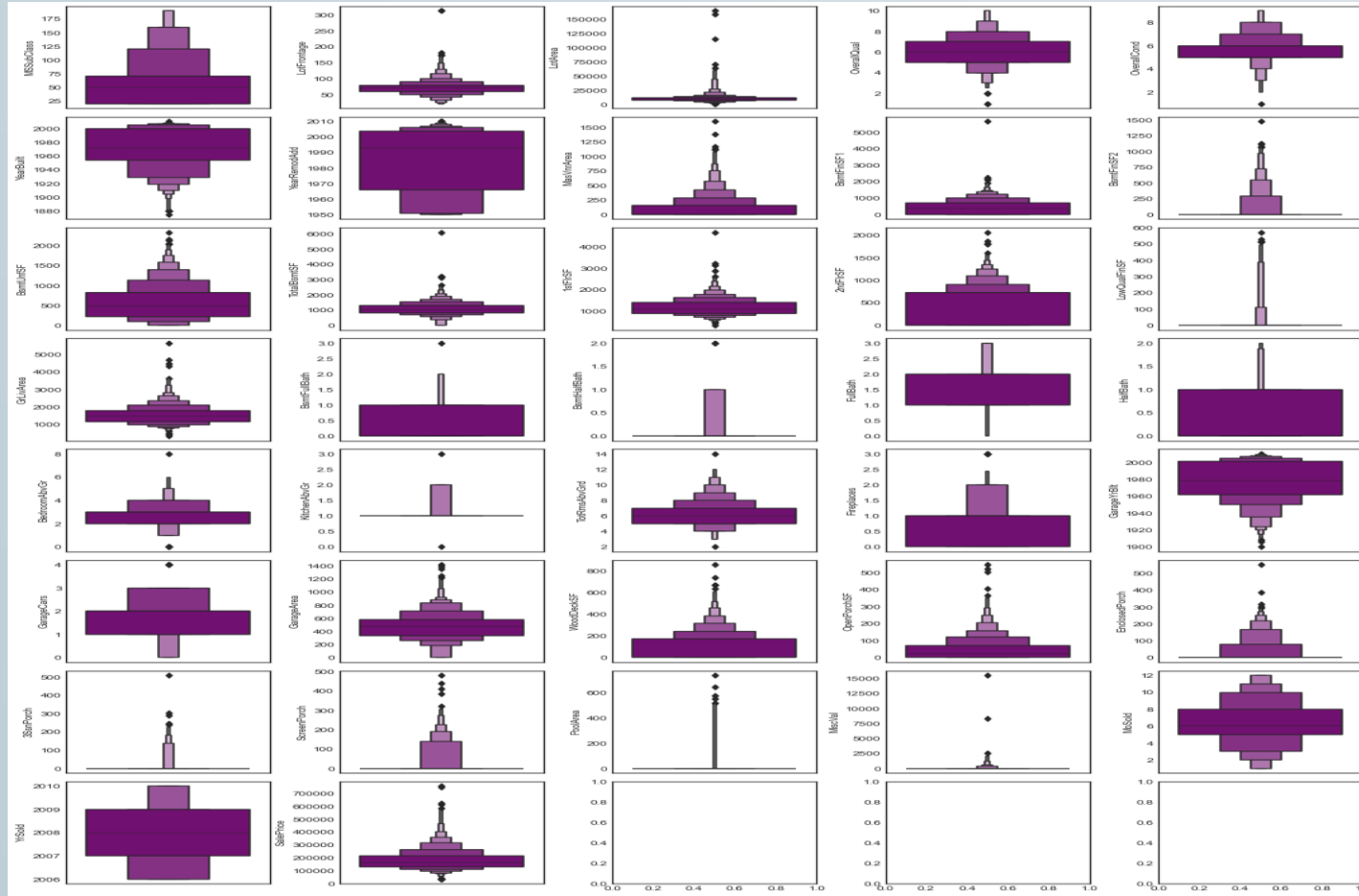
➤ It gives us an insight on positive and negative correlated column details.



BOXEN PLOT

➤ A Boxen Plot is also known as Whisker plot. It is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum.

➤ We have used it to identify the outlier details for all the numeric data type column values.

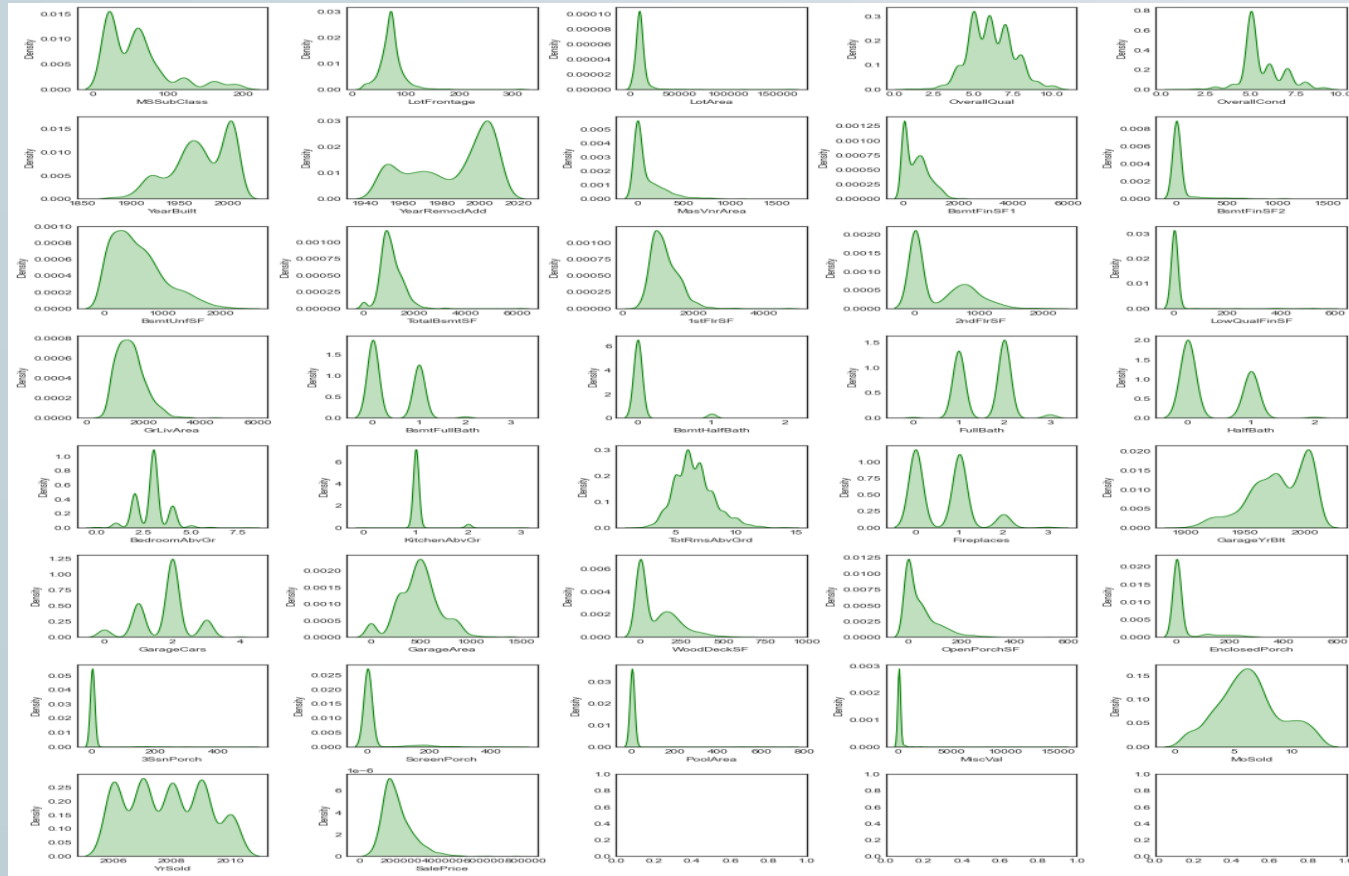


DISTRIBUTION PLOT

➤ Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution.

➤ Here we have used it to analyze the skewness information for numeric data type column values.

➤ The acceptable form usually is a normal distribution resembling a bell shape curve.



MODEL'S DEVELOPMENT

The algorithms used on training and test data are as follows:

1. Linear Regression Model
2. Ridge Regularization Regression Model
3. Lasso Regularization Regression Model
4. Support Vector Regression Model
5. Decision Tree Regression Model
6. Random Forest Regression Model
7. K Nearest Neighbors Regression Model
8. Gradient Boosting Regression Model
9. Ada Boost Regression Model
10. Extra Trees Regression Model



EVALUATION AND HYPER PARAMETER TUNING

The key metrics used here were:

- ☐ R2 score
- ☐ Cross Validation Score
- ☐ MAE
- ☐ MSE
- ☐ RMSE

➤ We tried to find out the best parameters list to increase our accuracy scores by using Hyper parameter Tuning.

➤ In order to achieve a higher score we used the Grid Search CV method with 5 folds.



CONCLUSION AND SCOPE FOR FUTURE WORK

- During this project I have faced a problem of low amount of data for training the machine learning models upon.
- Many columns are with same entries in more than 80% of rows which lead to reduction in our model performance.
- One more issue present is there are large number of missing values in this data set, so we have to fill those missing values in correct manner manually.
- We can still improve our model accuracy with some feature engineering and by doing some extensive hyper parameter tuning on it.





*Thank
you...*