Surprise Housing – House Price Prediction

Submitted by:

Satya Jyothi. T

Flip Robo

Internship Batch – 34

ACKNOWLEDGMENT

I would like to express my deep and sincere gratitude towards Flip Robo Technologies for providing me the internship opportunity and a great chance for learning and professional development.

A big special thanks to my SME Ms. Khushboo Garg for providing necessary help in solving the problems and for providing clarifications on time throughout.

My sincere thanks to “Data Trained” who are the reason behind my internship at Flip Robo.

Last but not least my all-well-wishers including parents/ spouse/ friends who have been my backbone in every step of my life.

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.
- We are required to model the price of houses with the available independent variables. This model 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

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. For this company wants to know:

- Which variables are important to predict the price of a variable?
- How do these variables describe the price of the house?

Motivation for the Problem Undertaken

This project was given by Flip Robo Technologies as a part of the internship program. This opportunity gives the exposure to real world data and using my skillset in solving a real time problem has been the primary motivation.

In this document, the focus will be on

- How to analyse the dataset of Surprise House – House Price Prediction
- What are the variables that impact House Price Prediction
- Overall data analysis on the problem

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

The main objective of doing this project is to build a machine learning model to predict the house prices with the help of other supporting features.

As the target is to predict house prices which is continuous data, it is a regression problem. Hence in this project, all regression related machine learning algorithms are used to build the models. Multiple regression metrics (r2 score, MAE, MSE and RMSE) have been checked to compare various regression models.

Hyper parameter tuning is performed for the best model with varying multiple key parameters of the algorithm.

Best model is chosen based on the least difference between model r2 score and cross validation score and saved as final model.

Model is built using train dataset and then used test dataset to predict house prices

Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). There are 2 data sets that are given. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. The dimension of data is 1168 rows and 81 columns.

2) Test file contains all the independent variables, but not the target variable. The dimension of data is 292 rows and 80 columns.

```
df.columns.to_series().groupby(df.dtypes).groups
```

```
{int64: ['Id', 'MSSubClass', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'BsmtFinSF1', 'BsmtFinSF2',
'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'H
alfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchS
F', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'MoSold', 'YrSold', 'SalePrice'], float64: ['LotFrontag
e', 'MasVnrArea', 'GarageYrBlt'], object: ['MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig',
'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Ext
erior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtF
inType2', 'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual', 'Functional', 'FireplaceQu', 'GarageType', 'Garage
Finish', 'GarageQual', 'GarageCond', 'PavedDrive', 'PoolQC', 'Fence', 'MiscFeature', 'SaleType', 'SaleCondition']}
```

Data Pre-processing Done

Data integrity is checked by checking for duplicate values, white spaces and missing values. There are no duplicate values, white spaces found in the dataset.

There are missing values present in 18 columns out of 81 columns. Columns 'PoolQC', 'MiscFeature', 'Alley', 'Fence', 'FireplaceQu' have more than 45% of missing values hence dropped these columns from the dataset.

For remaining columns with missing values, used 'mode' to replace missing values for categorical variables and 'median' to replace missing values for numerical variables.

There are 3 data types in the data set. a) Object b) int64 c) float64

Features with ‘int64’ datatype contains both numerical and categorical data.

Segregated all the columns based on datatypes for better EDA.

Statistical summary is checked for both numeric and categorical data to draw the insights of descriptive statistics (count, mean, min, max, std deviation and IQR values) in a simple manner.

Data Inputs- Logic- Output Relationships

There are many factors that have an impact on house prices. Some of the insights such as

How the location of the lot decides the selling price of the property.

How the amenities like Garages, Swimming pool, parking lot, fence type, insulation type and quality increase the selling prices.

How the number of rooms directly increases the cost and size of the property.

What is the type of building and the year built mostly comes for the sales?

The year the building build versus the cost of the property.

The year the modified versus the cost of the property.

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

Dropped column “Id” from the dataset as it will not have any significance or contribute to model learning.

Dropped column “Utilities” also from the dataset as this variable contains only one type of unique data throughout hence there will not be any relation between this variable and target variable “SalesPrice”.

Hardware and Software Requirements and Tools Used

Software Used -

Jupyter Notebook

Python 3.9.13

NumPy

Pandas

Matplotlib

Seaborn

scikit-learn

Hardware Used -

Processor — Intel i7 processor 8th Generation

RAM — 32 GB

GPU — 4GB NVIDIA Graphics card

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

There are many features in dataset. Used various statistical and analytical techniques to solve the problem.

- Performed univariate, bivariate and multivariate graphical analysis to draw the key insights for the relation between input and output variables
- Used Pearson's coefficients of correlation to identify the strong, negative and very weak correlations of features with target variable. Considered -0.15 to +0.15 range as very weak correlations.
- Used SelectKBest algorithm with F-regression test to know the top features that have strong relation with target variable. Considered the threshold score of 50 as it is 10% of top 10^{th} feature score with which identified the weak features.
- Dropped various columns that have very weak correlations and have lowest SelectKBest feature scores
- Multi-collinearity between features is checked using VIF (Variance Inflation Factor) and dropped the features with VIF>10 in multiple steps. VIF>10 indicates strong multi-collinearity of the features. VIF can also be addressed with PCA (Principal Component Analysis) without dropping the features. In this project, PCA is not used as the final model need to be used to predict the house prices with test dataset which contains set of fixed features which may be changed if PCA is applied. Hence dropped the columns with VIF > 10 by checking in multiple steps.
- Outliers are checked for the X numerical data. In z-score method, removed outliers if z-score > 3 and < -3. In IQR method, removed outliers beyond Inter Quartile Range (Q3-Q1). Calculated %data loss with both methods. z-score method resulted in data loss with less than 10% (6.8%) which is acceptable where as IQR method resulted

in the data loss of 19.3% hence finally used z-score method for outliers removal.

- Skewness is checked for X numerical data. Applied multiple transformation techniques (Power Transformer, Quantile Transformer) to check the skewness reduction. Considered -0.5 to +0.5 range as fairly symmetrical. Power Tranformer has reduced the skewness within this range, hence considered the data for further processing from Power Transformer.
- Scaled the X data using Standard Scaler. Scaled data is used for model building.

Testing of Identified Approaches (Algorithms)

The different regression algorithms used in this project to build ML model are as below:

Linear Regression

KNeighborsRegressor

DecisionTreeRegressor

SGDRegressor

SVR

Lasso

Ridge

ElasticNet

RandomForestRegressor

ExtraTreesRegressor

GradientBoostingRegressor

AdaBoostRegressor

XGBRegressor

Run and Evaluate selected models

- Started with Linear Regression to identify the best random state and no. of folds to be used in the cross validation.
- Once both are identified, used the same best random state and no. folds for all algorithms. Scoring parameter used as default. i.e. r2 score
- Multiple regression metrics (r2 score, MAE, MSE, RMSE) were checked for all algorithms.

Linear Regression Results:

```
Model : Linear Regression
R2 Score :  0.8651597418842483
Mean Absolute Error(MAE) :  17799.682312775694
Mean Squared Error(MSE) :  538696584.3679109
Root Mean Squared Error(RMSE) :  23209.83809439245
```

```
The CV r2 score is: 85.1640517196786
The model r2 score is: 86.51597418842482
Difference is: 0.013519224687462361
```

Results for other algorithms:

	Model	Model R2 Score	Cross Validation R2 Score	Difference in R2 Score	MAE	MSE	RMSE
9	ExtraTreesRegressor()	85.051538	85.075834	0.000243	16994.802844	5.972019e+08	24437.714091
11	AdaBoostRegressor()	80.544974	80.483049	0.000619	21503.411184	7.772423e+08	27879.066471
2	KNeighborsRegressor()	80.221440	79.836669	0.003848	20479.301835	7.901678e+08	28109.923233
7	ElasticNet()	84.066605	83.681661	0.003849	18575.331044	6.365507e+08	25229.955886
4	SVR()	-4.773593	-5.242437	0.004688	46508.014892	4.185781e+09	64697.610585
...	...	...	...	...	...	...	...
5	Lasso()	86.516825	85.165001	0.013518	17798.793278	5.386626e+08	23209.106236
0	LinearRegression()	86.515974	85.164052	0.013519	17799.682313	5.386966e+08	23209.838094
8	RandomForestRegressor()	87.494459	85.849240	0.016452	15907.636927	4.996054e+08	22351.854905
12	XGBRegressor()	87.880527	85.679750	0.022008	17025.305763	4.841817e+08	22004.129535
1	DecisionTreeRegressor()	62.366094	71.891860	0.095258	26019.036697	1.503502e+09	38775.014814

13 rows × 7 columns

Key Metrics for success in solving problem under consideration

- All the regression metrics parameters (r2 score, MAE, MSE and RMSE) were checked for all algorithms.
- Based on the comparison between model r2 score and cross validation scores for multiple models, considered "ExtraTreesRegressor()" is the best model as the difference in r2 score is least among all the models
- In this project, best model is chosen based on scoring parameter: r2_score which is default scoring parameter
- However, best model can change if scoring parameter is changed to MAE or MSE or RMSE
- Hyper parameter tuning is performed for "ExtraTreesRegressor()“

ExtraTreeRegressor (Hyper Parameter Tuned) Results:

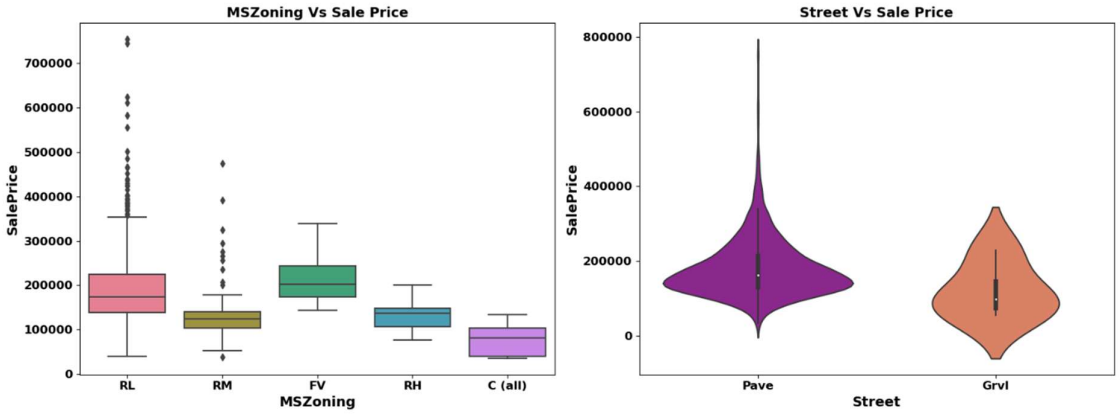
```
Model : ExtraTreesRegressor
R2 Score :  0.8669615273635047
Mean Absolute Error(MAE) :  16452.560665137615
Mean Squared Error(MSE) :  531498321.05247027
Root Mean Squared Error(RMSE) :  23054.24735384936
```

Observations for ExtraTreesRegressor:

- Hyper Parameter Tuned model: The difference between Cross Validation Score and model score is 0.01718
- Model with default parameters: The difference between Cross Validation Score and model score is 0.000243
- As the delta between model score and cross validation score is least in the model with default parameters, considered the model with default parameters as the best model for saving

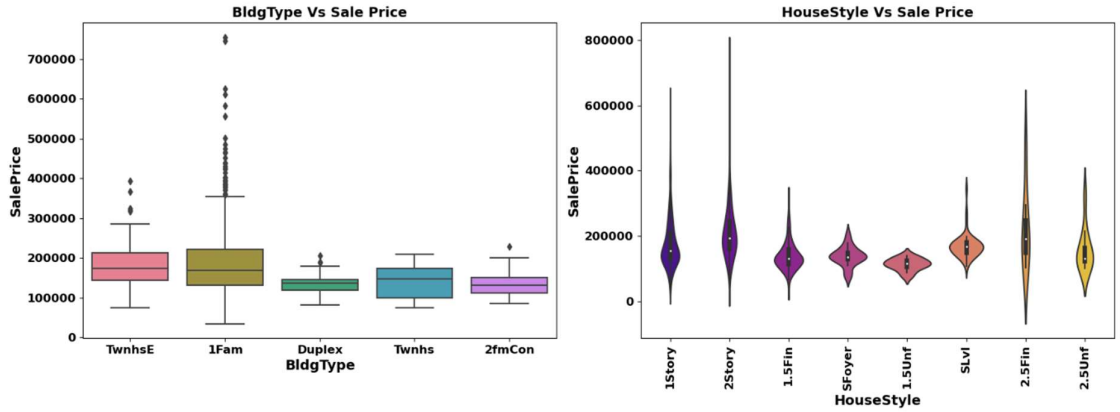
Visualizations

As there are many features, shown some of the visualizations in this document.



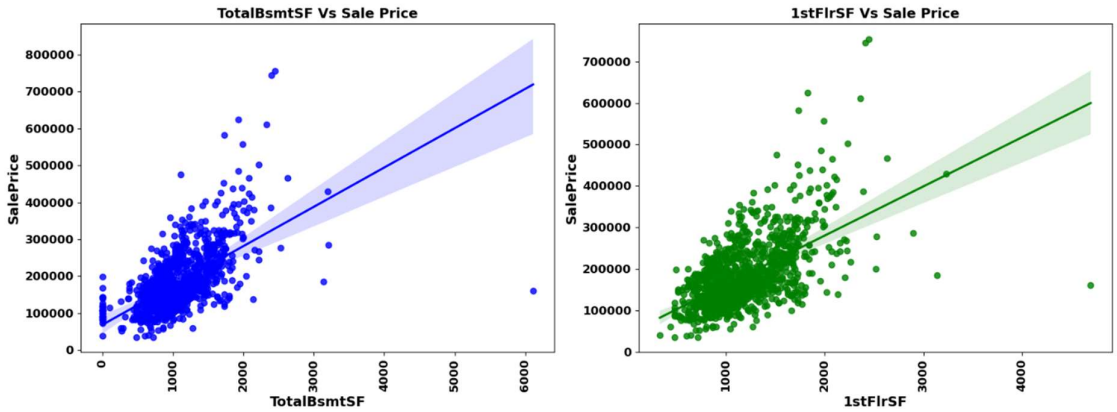
Observations:

- Sale Price of the property is higher in Residential Low-Density zone where as it is lower in Commercial zone
- Sale Price of the property is higher with Paved type of road access to property where as it is lower Gravel type of road access to property



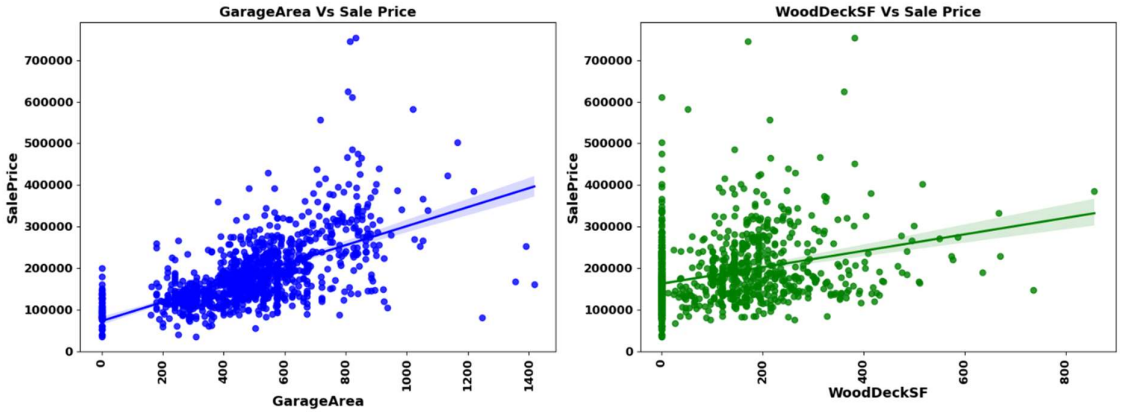
Observations:

- Sale Price of the property is higher with Single-family Detached type of dwelling where as it is lower with Duplex
- Sale Price of the property is higher with Two story style of dwelling where as it is lower with One and one-half story: 2nd level unfinished type of dwelling



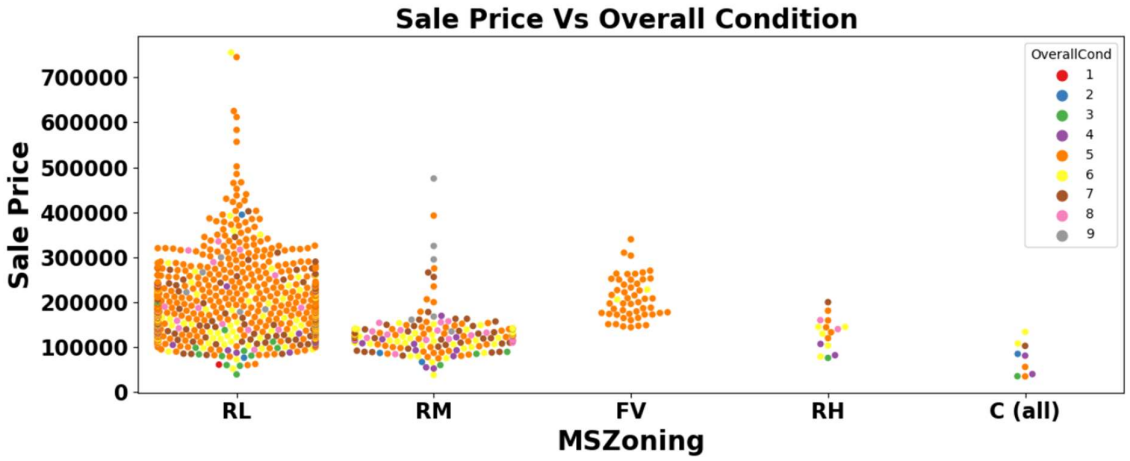
Observations:

- Sale Price of the property increases with increase in Total square feet of basement area
- Sale Price of the property increases with increase in First Floor square feet



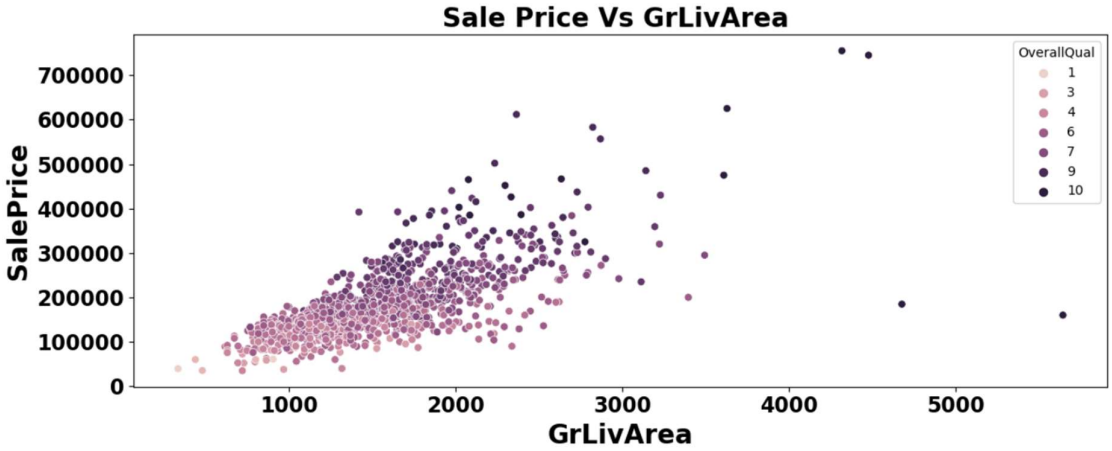
Observations:

- Sale Price of the property increases with Size of garage
- Sale Price of the property increases with increase in Wood deck area



Observations:

- Most of properties for sale have overall condition rating of either 5 or 6.
- Sale price inside RL Zone is much higher than other remaining zones.
- Cheapest properties are available in Commercial zone.



Observations:

Total floor area increases the sale price also get increases corresponding the overall quality of House

Interpretation of the Results

- Multiple ML models were developed using different regression algorithms
- Used r2 score as scoring parameter. Best model was chosen based on the least difference between model r2 score and cross validation score.
- ExtraTreesRegressor was considered as the best model which provided 85% model score as well as cross-validation score.
- Used the best model to predict the house prices using test dataset (test.csv)

CONCLUSION

Key Findings and Conclusions of the Study

	Model	Model R2 Score	Cross Validation R2 Score	Difference in R2 Score	MAE	MSE	RMSE
9	ExtraTreesRegressor()	85.051538	85.075834	0.000243	16994.802844	5.972019e+08	24437.714091
11	AdaBoostRegressor()	80.544974	80.483049	0.000619	21503.411184	7.772423e+08	27879.066471
2	KNeighborsRegressor()	80.221440	79.836669	0.003848	20479.301835	7.901678e+08	28109.923233
7	ElasticNet()	84.066605	83.681661	0.003849	18575.331044	6.365507e+08	25229.955886
4	SVR()	-4.773593	-5.242437	0.004688	46508.014892	4.185781e+09	64697.610585
...	...	...	...	...	...	...	...
5	Lasso()	86.516825	85.165001	0.013518	17798.793278	5.386626e+08	23209.106236
0	LinearRegression()	86.515974	85.164052	0.013519	17799.682313	5.386966e+08	23209.838094
8	RandomForestRegressor()	87.494459	85.849240	0.016452	15907.636927	4.996054e+08	22351.854905
12	XGBRegressor()	87.880527	85.679750	0.022008	17025.305763	4.841817e+08	22004.129535
1	DecisionTreeRegressor()	62.366094	71.891860	0.095258	26019.036697	1.503502e+09	38775.014814

ExtraTreesRegressor is the best model from this study based on r2 score metrics which helps to predict house prices

Learning Outcomes of the Study in respect of Data Science

Price Prediction modeling – This study helps one to understand the business of real estate. How the price is changing across the Properties.

Prediction of Sale Price – This helps to predict the future revenues based on inputs from the past and different types of factors related to real estate & property related aspects. This is best done using predictive data analytics to calculate the future price values of houses. This helps in segregating houses, identifying the ones with high future value, and investing more resources on them.

Deployment of ML models – The Machine learning models can also predict the houses depending upon the needs of the buyers and recommend them, so customers can make final decisions as per the needs.

Limitations of this work and Scope for Future Work

This dataset has 1168 rows which small. Company can get more correct insights if more data is available.

There are some categories which are provided in the data description for all features are not available in train dataset hence if there is any new category other than trained model for any feature, then this model would not be able to identify that and will not work as expected

There are multiple columns with more than 50% of missing values. Imputation of them can decrease effectiveness and dropping them leads to significant loss of data.

There are multiple non-value added features in the dataset with target variable. If not treated them properly can lead to ineffective model building