Housing-Project

Submitted by:

Akash Rawat

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

sklearn.preprocessing

sklearn.ensemble

sklearn.model_selection

StandardScaler

OrdinalEncoder

Pandas

Seaborn

Numpy

Matplotlib

train_test_split

Random Forest Regressor, Extra Trees Regressor, Decision Tree Regressor

Grid Search CV

classification_report, confusion_matrix, roc_curve, accuracy_score

r2_score, mean_squared_error

INTRODUCTION

- ## Business Problem Framing

  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.

  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.

- ## Conceptual Background of the Domain Problem

  The project will require knowledge and practice in building Graphs plots and analysing them to get the relationship between dataset, Knowledge of Different Learning Models to build and predict the required output. Basic Data science concepts to increase the quality of the dataset and Python Knowledge (Coding Language) which will be used to solve the complete Micro Credit Defaulter project. Understanding of calculating F2 score, accuracy, skewness, and basic mathematics/statistical approaches will help to build an accurate model for this project.

- ## Review of Literature

  Market price is what a willing, ready and bank-qualified buyer will pay for a property and what the seller will accept for it. The transaction that takes place determines the market price, which will then influence the market value of future sales. Price is

determined by local supply and demand, the property's condition and what other similar properties have sold for without adding in the value component.
Market value is an opinion of what a property would sell for in a competitive market based on the features and benefits of that property (the value), the overall real estate market, supply and demand, and what other similar properties have sold for in the same condition.
The major difference between market value and market price is that the market value, in the eyes of the seller, might be much more than what a buyer will pay for the property or it's true market price. Value can create demand, which can influence price. But, without the demand function, value alone cannot influence price. As supply increases and demand decreases, price goes down, and value is not influential. As supply decreases and demand increases, the price will rise, and value will influence price. Market value and market price can be equal in a balanced market.

- Motivation for the Problem Undertaken

  I wanted to solve the real-life problem using the Technical skills gathered during the course of being a Data Analyst and improving the skill set.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Decision Tree –

  It is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs and utility.

  Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

Random Forest –

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

- Hardware and Software Requirements and Tools Used

Software: Jupyter Notebook - To code and build the project in python
Libraries:

a. numpy - To perform basic math operations
b. pandas - To perform basic File operations
c. Matplotlib - To plot Different Graphs/ Plots
d. Seaborn - Advance library to enhance the quality of graphs/plots
e. warnings - To ignore the unwanted warnings raised while interpreting the code
f. sklearn - To build the Prediction models
g. imblearn - To balance our dataset distribution

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

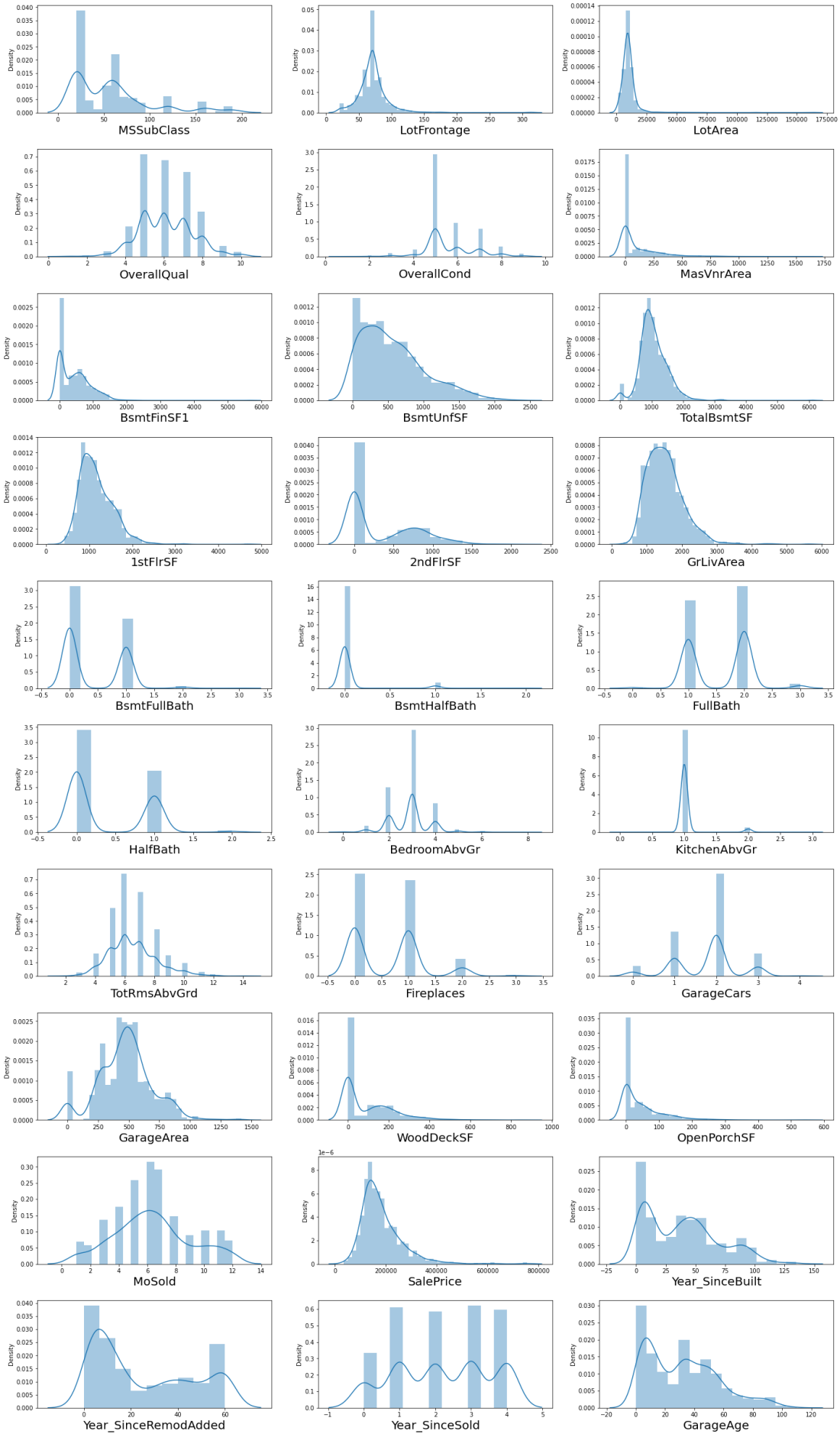
We used different approaches from checking the dataset quality to building the model. We checked the null values and repeated rows in the dataset. For checking the Outliers, we used Box Plot and to remove

the outliers we used IQR method. Then we moved to next step of checking data distribution and skewness. To scale the data, we used MinMax Scaler method and to remove the skewness we first checked the log and square root method, but skewness of the dataset was not getting removed from it, so we performed the Power transform to remove skewness. We started building different models and checked their R2 score and selected the best suited model to perform Hyper tuning on. We got Random Forest Algo with the best result and after performing Hyper tuning we finalized the model.

- Testing of Identified Approaches (Algorithms)

1) Decision Tree
2) Random Forest

3) Visualizations


MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
MasVnrArea
BsmtFinSF1
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
MoSold
SalePrice
Year_SinceBuilt
Year_SinceRemodAdded
Year_SinceSold
GarageAge


MSZoning
Street
LotShape
LandContour
Utilities
LotConfig
LandSlope
Neighborhood
Condition1
Condition2
BldgType
HouseStyle
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Heating
HeatingQC
CentralAir
Electrical
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
SaleType
SaleCondition

Interpretation of the Results

Random Forest Regressor

R2_score: 91.30783395015679
mean_squared_error: 616302251.4064877
mean_absolute_error: 16891.917606837607
root_mean_squared_error: 24825.435573348714

Cross validation score : 84.74032771295785

R2_Score - Cross Validation Score : 6.567506237198941

Extra Tree Regressor

R2_score: 91.5494040671515
mean_squared_error: 599174160.8796107
mean_absolute_error: 17234.899116809116
root_mean_squared_error: 24478.034252766513

Cross validation score : 84.23214440054994

R2_Score - Cross Validation Score : 7.317259666601558

Decision Tree Regressor

R2_score: 72.69646613973157
mean_squared_error: 1935907493.3618233
mean_absolute_error: 28379.31623931624
root_mean_squared_error: 43998.948775644894

Cross validation score : 68.07667347237657

R2_Score - Cross Validation Score : 4.619792667355

CONCLUSION

- Key Findings and Conclusions of the Study

We found that to predict the House price using Data Science the best way after performing Data Cleaning is to use Random Forest Algorithm it provides 88% accuracy which is better than other Regression algorithms.

- Learning Outcomes of the Study in respect of Data Science

In data science, there are various steps involved during Data analysis and cleaning. With the help of various Visualization tools like plots, Graphs we were able to perform the actions and observe different things. Like for finding the outliers we used Box Plot visualization, for finding the skewness and normalization we used Count Plot visualization, for finding skewness we visualized the skewness using Heat Map for the clear picture of how the variables are co-related to each other in the dataset. We used different metrics to check which model best fits the prediction for the dataset.

- Limitations of this work and Scope for Future Work

Data was unbalanced if data was balanced more accurate and clearer picture of the output -> result is dependent on the data

Neural network classifier which are still unexplored & can be taken for future consideration