HOUSING PROJECT

Submitted by:

YOUR NAME

NARASIMHA KARTHIK KOMANDURU

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.

INTRODUCTION

- ## Business Problem Framing

  The problem is related to Identifying the price of the House using sale of houses data set in Australia.Houses are one of the necessary need of each and every person in the globe.It is a very large market and there are various companies working on this domain.Data science comes as a very vital tool to solve problems in this domain to help the companies increase their revenue,profits,improving their marketing strategies and focusing on changing trends in house and sales purchases.

- ## Conceptual Background of the Domain Problem

  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 of such companies.

- ## Review of Literature

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

- ## Motivation for the Problem Undertaken

  The main motivation behind choosing this project is to Develop a model to predict the value of houses using different Machine learning algorithms.The reason behind choosing machine learning domain as it gives most effective solutions in predicting the values of the houses.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The Mathematical and Analytical models used in this are Statistics and different Machine learning algorithm(Random Forest Regressor) to predict the efficiency of the model.The main reason behind using Statistics is to find different statistical features of the data and then perform EDA on the data and the Reason behind Choosing Random Forest regressor is it gives most effective predictive value of the model.

- Data Sources and their formats

  The main source of data set is from Super housing Company from Australian market and the data set contains different parameters that are used in predicting the model.The below screen shot shows the different columns of the data used in analyzing the problem.

```
Index(['Id', 'MSSubClass', 'MSZoning', 'LotFrontage', 'LotArea', 'Street',
       'Alley', 'LotShape', 'LandContour', 'Utilities', 'LotConfig',
       'LandSlope', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType',
       'HouseStyle', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd',
       'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType',
       'MasVnrArea', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual',
       'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinSF1',
       'BsmtFinType2', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', 'Heating',
       'HeatingQC', 'CentralAir', 'Electrical', '1stFlrSF', '2ndFlrSF',
       'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
       'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual',
       'TotRmsAbvGrd', 'Functional', 'Fireplaces', 'FireplaceQu', 'GarageType',
       'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual',
       'GarageCond', 'PavedDrive', 'WoodDeckSF', 'OpenPorchSF',
       'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'PoolQC',
       'Fence', 'MiscFeature', 'MiscVal', 'MoSold', 'YrSold', 'SaleType',
       'SaleCondition', 'SalePrice'],
      dtype='object')
```

- Data Preprocessing Done

  In the Data Preprocessing step we have identified the shape of both train and test data and then identified the null values in both the data sets.After identifying the null values in both the data sets we combined both train and test data to fill the null values.While filling the null values first we identified the categorical columns and filled them with Constant value for columns having a meaning for the null values and with Mode for the categorical columns which do not have any specific meaning for the NA value.Once we filled the NA values for Categorical columns we identified the Numerical columns having null values and filled them with Median and mean based on Statistical data which was obtained using the describe function.Once the NA or missing values are filled identified the Skewness in the data for the numerical columns and applied log transformation to remove the skewness for the Numerical columns.We only removed the skewness of the numerical data as we are not supposed to remove skewness from the Categorical data.Once the skewness is removed we had used the get dummies function to convert all the Categorical columns in to numerical which is also known as encoding.After doing scaling we have scaled our data For more accuracy.

- Data Inputs- Logic- Output Relationships

We have splitted the Target variable from the train and test data set as both the data sets has difference in the columns.As there is a variance it will resulted in a error so we have splitted the Target variable and applied log transformation for the data to remove the skewness so that data will be evenly distributed and later we splitted the train data and test data base on the index and created train test split to predict the output of the model.

- Hardware and Software Requirements and Tools Used

The tools used here are Jupyter notebook for coding purposes and the different libraries used here are Matplotlib for plotting and Seaborn for visualization and sklearn for importing classification reports and for using regression alogrithms.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

The statistical approaches used here are describe function to find the mean,median and mode for the data set and the skew function to remove skewness of the data.The analytical approaches used here are log transformation to remove skweness of the data and using distplot to identifying the distribution of data.

- Testing of Identified Approaches (Algorithms)

The different algorithms used for identifying and testing the data are train test split and random forest regressor for creating train test split and for predicting the model.

- Run and Evaluate selected models

  The only algorithm that we used here is random forest regressor as it gives the best accuracy and most effective metrics below is the code snippet for the random forest regressor.

```
: from sklearn.model_selection import train_test_split
  X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=21)

: from sklearn.ensemble import RandomForestRegressor
  rf=RandomForestRegressor()
  rf.fit(X_train,y_train)

: RandomForestRegressor()

: y_pred=rf.predict(X_test)

: rf.score(X_train,y_train)

: 0.9798602160287279
```

  Our training data gives 97% accuracy

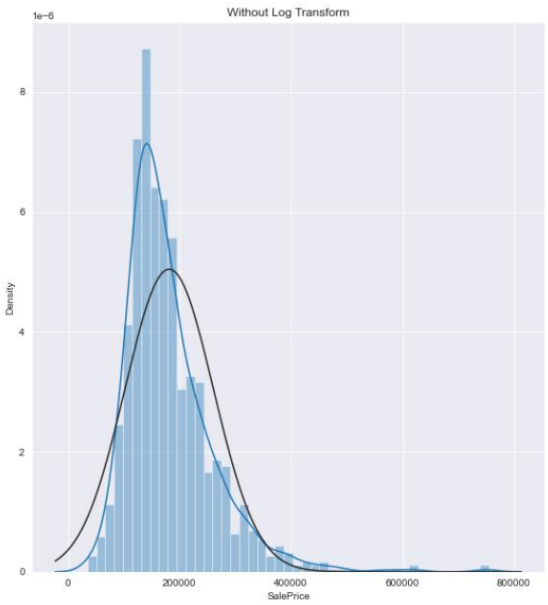
```
: rf.score(X_test,y_test)

: 0.8342374482118036
```

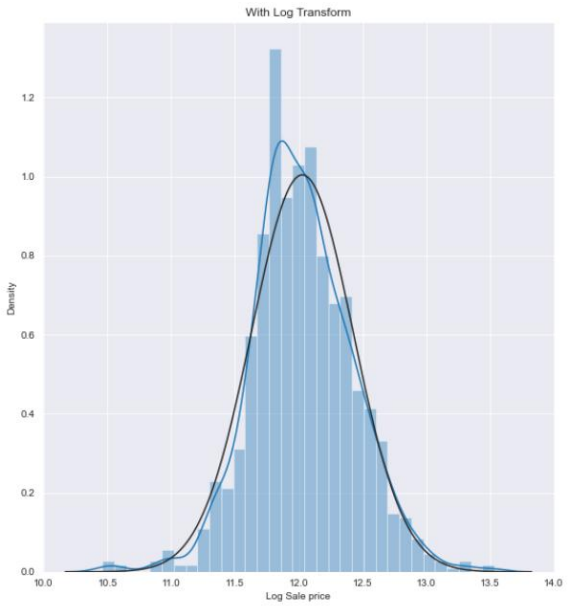
- Key Metrics for success in solving problem under consideration

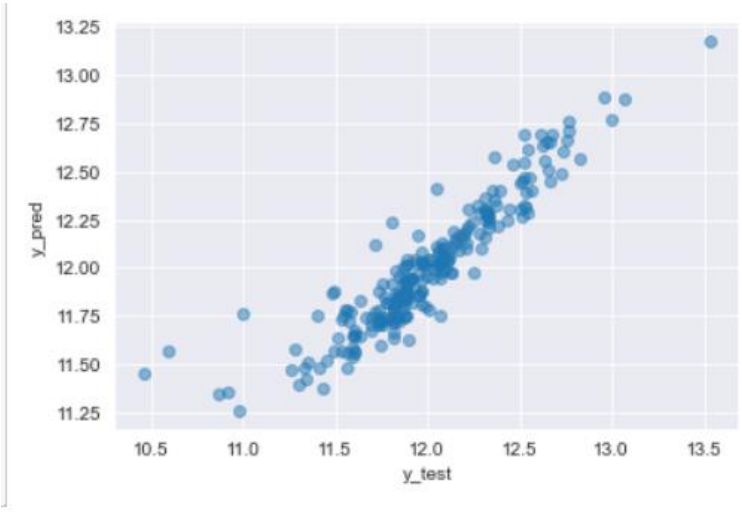
  The key metrics used in resolving our problem are MAE,MSE and RMSE as the problem is regression problem and these will be the key metrics to solve the problem.

- Visualizations

The visulazation techniques used here are scatter plot to find the distirbution of model once done and then dist plot to find the distribution of the target variable after applying log transformation.Below are the screen shots for the target variable and the distribution of result.







our test data set gives more fine results as it is increasing order.

- Interpretation of the Results

  After completing the visulaization and preprocessing and modelling of our data we predicted the final percentage which gives the predicting value using MSE,RMSE And MAE.Once calculated these metrics we plotted and scatter plot for our model after applying randomized search and found that the relation ship is linear for the predict value with the trained data.

CONCLUSION

- Key Findings and Conclusions of the Study

  The key findings of the problem is that there are a lot of values that need to be interpreted and need to visualize and import more time to analyze the problem as the train and test data has different columns and there is a need to different regression techinques and statistical approches to predict the model

- Learning Outcomes of the Study in respect of Data Science

  The learning from this project is that we need to put lot of effort in doing EDA and then cleaning the data.The main challenge I had here is identifying the key attributes and I over comed it by reviewing youtube and going through the concepts like removing skewness and transformation techniques.The main use of visualization technique is it helps us to infer the data more clearly .

- Limitations of this work and Scope for Future Work

  The limitation of the project is we used only one regression algorithm which is random forest regressor and we can use other regression algorithms to predict the models in future.For time being used only one regression algorithm.