HOSUING PROJECT – SALE PRICE PREDICTION

Submitted by:

NATASHA PODDAR

. INTRODUCTION

- ## Business Problem Framing

  The project was on HOUSE PRICE PREDICTION, it's a very practical issue on how various factors like Garage space, Foundation, Renovation of the house, Exterior Quality and many many more factors contribute in settling the price of a Apartment/Housing model

- ## Conceptual Background of the Domain Problem

  This project is based on House Sale Price prediction, one should have a basic idea on the space, area, locality, quality of the house and other factors which help in determining the price of an apartment. Like a larger space would attract a higher price, a fenced garden bungalow with various amenities would attract a higher price.

  Basic Real estate knowledge and concepts will be useful for better understanding of the project.

- ## Motivation for the Problem Undertaken

  The motivation is to understand that what are the major drivers which help in determining a price of a house. This project had more than 80 independent features which contributed in the sale price.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Firstly missing values were checked and imputed. The columns which had more than 40% data missing were deleted

  Correlation with all independent variables and wrt target were checked

  Outliers were identified but they were removed through zscore as the data had mostly categorical figures

  Skewness was checked and tools were applied to control them and scale the data

  Multi colinearity was checked and worked upon

  Models were applied to train and test the model

- Data Sources and their formats

  Detailed data was procured from Housing Use Case,

  Train data was provided for model training and testing

  Test Data was provided for predicting the results

- Data Preprocessing Done

  What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

  1. Duplicate values check
  2. Unique & Count of all columns were checked
  3. Missing values were imputed
  4. Columns which had more than 40% data missing were removed
  5. Catagorical data was Encoded

6. Skewness removal through Power Transform and scaling of the data
7. VIF Check -for multicollinearity
8. Correlation check
9. Graphical Univariate, Bivariate & Multivariate Analysis
10. Outliers check -ZSCORE

- Data Inputs- Logic- Output Relationships

  If we talk about correlation, some features had a very strong positive correlation wrt target –

  1. Garage Area
  2. Garage Cars
  3. OverAll Quality
  4. Gr Liv Area

  One on the other hand some features had a very strong Negative correlation wrt target –

  1. BSMT QLT
  2. EXTERQLT
  3. KITCHEN QLT

  Where as some features had a neutral correlation wrt target –

  1. MASVNRTYPE
  2. MISC VAL
  3. BSMT HALF BATH
  4. LANDSLOPE
  5. BSMT FIN TYPE 2
  6. LAND CONTOUR

- State the set of assumptions (if any) related to the problem under consideration
  1. VIF SCORES FOR MANY COLUMNS WERE GREATER THAN 10, AFTER DROPPING 10 COLUMNS, WE ASSUMED THAT IF WE

DELETE MORE COLUMNS FROM THE DATA SET WE MAY LOSE IMP FEATURES WHICH WOULD HELP IN DETERMINING THE MODEL ACCURACY

- Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

1. Pandas – For Data Reading and understanding
2. Simple Imputer – (SK LEARN) -for imputing missing values
3. Label Encoder –(SK LEARN) – For Encoding the categorical data into numerical ones
4. Zscore(SCIPY)-For checking & removal of outliers
5. Power Transform ()- Skewness removal
6. Duplicate- To check for duplicate Values
7. CORR-To check Correlation
8. VIF -To check for multicollinearity
9. Numpy- For mathematical operations
10. LINEAR REGRESSION (SKLEARN) – Training & Testing the model
11. KNN REGRESSOR (SKLEARN) – Training & Testing the model
12. DECISION TREE REGRESSOR (SKLEARN) – Training & Testing the model
13. RANDOM FOREST REGRESSOR (SKLEARN) – Training & Testing the model
14. GRADIENT BOOSTING REGRESSOR (SKLEARN) – Training & Testing the model
15. CROSS VAL SCORE – Regularizing the model
16. GRID SEARCH CV- Hyper Tuning the Model for higher accuracy
17. SEABORN- VISUALIZATION LIBRARY – HISTPLOTS, DISTPLOTS, SCATTERPLOTS, COUNTPLOTS, BOXPLOTS and other graphs

18. MATPLOTLIB.PY PLOT -Visualization tool

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Firstly missing values were checked and imputed. The columns which had more than 40% data missing were deleted
  Correlation with all independent variables and wrt target were checked
  Outliers were identified but they were removed through zscore as the data had mostly categorical figures
  Skewness was checked and tools were applied to control them and scale the data
  Multi colinearity was checked and worked upon
  Models were applied to train and test the model

- Testing of Identified Approaches (Algorithms)
  1. LINEAR REGRESSION
  2. KNN
  3. DECISION TREE REGRESSOR
  4. RANDOM FOREST REGRESSOR
  5. GRADIENT BOOSTING REGRESSOR

- Key Metrics for success in solving problem under consideration

  R2 SCORE

- Visualizations

  Seaborn Library was used along with matplotlib Library for visualizations

Histplots, bar plots, count plots, swarmplots, boxplots etc were made and analysed

- Interpretation of the Results

  All the models predicted an accuracy in the range of 70-80 where as gradient Boosting Regressor had the highest accuracy and we had hyper tuned the said model and saved the same

CONCLUSION

- Key Findings and Conclusions of the Study

  Gradient Boosting Regressor had the highest accuracy and the least difference between model score and cv mean score.

- Learning Outcomes of the Study in respect of Data Science

  Gradient Boosting Regressor had the highest accuracy and the least difference between model score and cv mean score. With unique feature we realized the type of data all the columns had, The various visualization tools helped in understanding the different relationships between the variables, imputing methods helped in filling up the missing values. VIF was a very powerful tool to detect multicollinearity and Cross Val score helped in regularizing the model

- Limitations of this work and Scope for Future Work

  VIF Scores can be checked and we can try and eliminate more columns to check if there is an improvement in the model accuracy