PROJECT HOUSING

Submitted by:

KIRTARATH SINGH SAINI

ACKNOWLEDGMENT

All the references, research papers, data sources, professionals and other resources that helped and guided in completion of the project.

- Google
- Youtube

INTRODUCTION

- ## Problem Statement:

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 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.

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 variable?
- How do these variables describe the price of the house

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  We were 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.

- Data Sources and their formats

  we were provided with the CSV files that contained the data which was divided into two parts

  TRAIN.CSV

  TEST.CSV

- Data Preprocessing Done

  The files were read on the jupyter using codes and then the EDA (EXPLORATORY DATA ANALYSIS (EDA) was done

  1. Viualization was done.
  2. Null values were checked and treated.
  3. Duplicate values were Checked.
  4. Unique values were Checked.
  5. Encoding was done to deal with the categorical column.
  6. Outliers were checked
  7. Skewness was checked and removed

- Hardware and Software Requirements and Tools Used

  Essential libraries used were Pandas , numpy, matplotlib ,seaborn ,warning sklearn (preprocessing ,label encoder , linear

regression , knn regression , random forest regressor, gridsearchCV , crossvalidation.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- Testing of Identified Approaches (Algorithms)

  Algorithms used were

  1. Knn regresson
  2. Random forest regressor
  3. Linear regression
  4. Support vector regression

- Run and Evaluate selected models

  Various models were used such as-

  Knn regresson
  Test score=-0.7948038426154977
  Train score= 0.8315528576257694

  Random forest regressor
  Test score= 0.9804907483553653
  Train score= -18.492445466720998

  Linear regression
  Test score= -2229085.1471461407

Train score= 0.8361452528603943

Support vector regression
Test score= -0.7948038426154977
Train score= -0.04822528865901643

Among all of them the best working model or the best fit model was the Random forest regressor it gave the train accuracy of 0.9804907483553653 and the test accuracy of -18.492445466720998

- Cross validation

Cross val score was considered as 3 because at cross fold 3 the cv score is 0.8572790977988548 and accuracy score for training is 0.9804907483553653 and accuracy score for testing is -18.492445466720998

- Grid searchCV was used
  The predicted cross val score was 84.7391886947602

  The R2 score was -1214.495956927553

- Visualizations
  1.Group plot were used
  2.Histograms were used
  3.Countplot were used

- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

The best fit model was the Random Forest regressor.

There are various factors that affect the pricing of the house.