

HOUSING PROJECT

Submitted by:

HARSHIL GORADIA

**ACKNOWLEDGMENT**

I have mentioned all the references that helped me and guided you in completion of the project.

1. Stackoverflow.com
2. Kaggle.com

**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. In this project we need to predict the prices of houses using Machine Learning.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

There are quite many independent variables in the dataset, which affects the prediction of the target variable. As the number of independent variables increase features taking part in the prediction also increases. But it gives more efficient and accurate prediction.

* Data Sources and their formats

The source of the data is an US-based housing company named Surprise Housing. The data collect is in CSV file and is of sale of houses in Australia. There are two datasets, one is train.csv with 1168 entries and 80 independent variables with one dependent variable whereas other is test.csv with only 292 entries and 80 independent variables.

* Data Preprocessing Done

For cleaning the data, I divided it into numerical data and object data. Variables with more than 40% nan values were dropped without any processing assuming that replacing 40% of data will reduce the accuracy while predicting. Rest all the nan values were replaced with ‘Unknown’ for categorical values and handled using KNNImputer for numerical values.

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

Assumptions were made that replacing values in a variable with more than 40% nan values will decrease accuracy for prediction.

* 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. Jupyter Notebook - The whole project was done in jupyter notebook which provides python environment.
2. Pandas and Numpy - These libraries which are most important for an predictive analysis project. These libraries provide a function of shaping the data, importing the data and viewing the data.
3. Matplotlib and Seaborn – These libraries were used for visualising the data in the entire project. All the visualization done in the project were done using either one of these libraries.
4. Scipy - Scipy library was only used to import zscore to remove outliers present in the dataset.
5. Sklearn – This library was the most used library as compared to other libraries. This library has been used in pre-processing for imputing numerical null values and for scaling the dataset. Later it was widely used for importing algorithms for machine learning and predicting.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

I followed a usual approach for solving the problems of the project, that goes like importing dataset followed by data cleaning and EDA. After visualizing, I split the data into train and test data test. Later I went on with machine learning using different models and regularised using L1 and L2 algorithms. I chose my best model for predicting the target variable by hyperparameter tuning on RandonForestRegressor and LassoRegressor.

* Run and evaluate selected models

LinearRegression:

Graphical user interface, text, application

Description automatically generated

test accuracy: 0.8831288926214168

Mean absolute error: 0.08646957896167633

Mean squared error: 0.014964898196116315

Root Mean Squared Error: 0.12233110069036539

DecisionTreeRegressor:

Graphical user interface, text

Description automatically generated

test accuracy : 0.6685610183917505

Mean absolute error: 0.15289835540376281

Mean squared error: 0.04243949363742265

Root Mean Squared Error: 0.20600847952796178

RandomForestRegressor:

Graphical user interface, text, application

Description automatically generated

test accuracy : 0.8410025018386513

Mean absolute error: 0.10193340685111044

Mean squared error: 0.02035902137655109

Root Mean Squared Error: 0.14268504258173345

KNeighborsRegressor:

Text

Description automatically generated

test accuracy : 0.6208299371635125

Mean absolute error: 0.1660122201975233

Mean squared error: 0.04855127598801955

Root Mean Squared Error: 0.22034354083571306

LassoCV:

Text

Description automatically generated

Mean Absolute Error : 0.08216570847921147

Mean Squared Error : 0.014063817268392146

Root Mean Squared Error : 0.11859096621746594

test accuracy : 0.8901660488038875

RidgeCV:

Text

Description automatically generated

Mean Absolute Error : 0.0850779148198316

Mean Squared Error : 0.01472106800275973

Root Mean Squared Error : 0.12133040840102588

test accuracy : 0.8850331290777205

* Key Metrics for success in solving problem under consideration

Key metrics used in this project are mean\_squared\_error, mean\_absolute\_error, r2\_score. Deciding the best model based on this metrics is way which is followed by many data scientists. They provide information about errors and accuracy for the model.

* Visualizations

Chart, scatter chart

Description automatically generated Chart

Description automatically generated

Graphical user interface, scatter chart

Description automatically generated

Histogram

Description automatically generated with medium confidence Chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

* Interpretation of the Results

From visualization I found that nearly all the independent variables contribute enough towards predicting the target variable. LassoCV came out to be the best regularization model that gives the best accuracy while predicting the prices of houses.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Findings from the project was that many factors play an important for deciding the prices of houses in Australia. There were many nan values present in the dataset but were handled smoothly. At the end of the project key findings were that LassoCV came out to be the best model for predicting house prices followed by LinearRegression and RandomForestRegressor.

* Learning Outcomes of the Study in respect of Data Science

I learnt many things while working on this project. Visualizations can help a lot for feature selection. Data cleaning helps us get a better accuracy while predicting the target variable and various algorithms were used in this project which gave a variety of models to used to predict and select the best model for predicting the target variable. Working with many features and there were train and test two different csv files so had to work a lot behind getting better model.

* Limitations of this work and Scope for Future Work

Limitations of this project are that it can be used for a small dataset only and not for larger dataset. This prediction can be used only in the location from where it was collected, factors deciding house prices of other countries can be different.