CAPSTONE PROJECT

Melbourne Housing Price Prediction

ABSTRACT

Real estate business is very competitive because the prices of the properties are varying significantly due to several criteria of the property. Because of that estimating price of a certain property is bit challenging and need to be precise. As a solution, a Machine Learning model was developed in this project to determine the estimated value of a certain property using few criteria of the property.

Pasan Dharmasiri
Pasan.dharmasiri@outlook.com

CONTENTS

1) Introduction 2
2) Dataset 2
3) Methodology 3
   a) Exploratory Data Analysis 3
   b) Pre-processing Data 4
   c) Model Training 4
4) Results 5
5) Conclusion 5
6) Discussion 5
7) References 6

1) INTRODUCTION

In present, real state business is very competitive because the prices of the properties are varying significantly due to several criteria of the property. Because of that estimating price of a certain property is bit challenging and need to be precise. If the estimated price is higher and it leads to reduce the sales of the company.

As a solution, a Machine Learning model was developed in this project to determine the estimated value of a certain property using few criteria of the property. Apart from that, an API was integrated with the model to embed this Machine Learning application to another web or any other user application.

2) DATASET

To develop a Machine Learning application that estimates price, it requires a dataset that contains the selling price of each property with the other criteria such as location as specifications. Therefore, a dataset that contains details of the properties in Melbourn, Australia was used in this project.

This is an open dataset which available in Kaggle and the following table explains the columns of the dataset.

Table 1: Data Description

Criteria	Description
Suburb	Suburb of the property
Address	Address of the property
Bedrooms	No. of rooms in the property
Type	Property type (House, unit, Townhouse)
Price	Sold price
Method	Sold method of property (property sold, property sold prior etc.)
Seller	Real Estate Agent
Date	Date sold
Distance	Distance from CBD in Kilometers
Postcode	Postcode of the location
Bedroom 2	No. of rooms in the property (Alternative source)
Bathroom	No. of Bathrooms in the property
Car	Number of car spots
Land size	Land Size in Meters
Building Area	Building Size in Meters
Year Built	Year the house was built
Council Area	Governing council for the area
Latitude	Latitude of the location
Longitude	Longitude of the location
Region name	General Region (West, Northwest, North, Northeast …etc.)
Property count	Number of properties that exist in the suburb.

3) METHODOLOGY

This section describes the Machine Learning process that used to train and evaluate model in this project.

a) Exploratory Data Analysis

As exploratory data analysis some filtering, and conversions must be done in the dataset. These steps can be described as follows.

i) Removing false data such as the data points that have Year Build latest than the Sold data.

ii) Then convert Sold date into days count from current date as well as the Year build into age of the property.

iii) After that remove unwanted columns and by calculating correlation matrix.

iv) There after rows with null data must be removed.

v) Then remove outliers of numerical data and plot box plots to confirm whether there are no outliers.

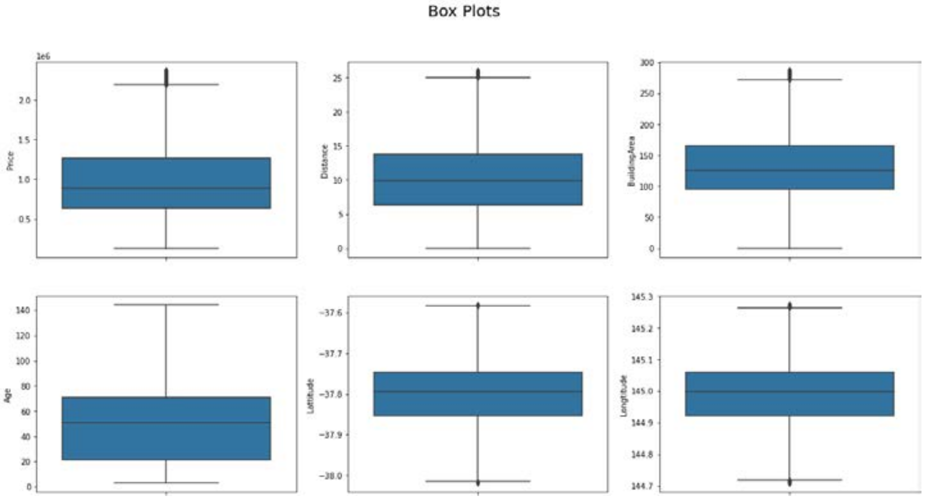

Figure 1: Box plot of numerical data

vi) Plot histogram of categorical data

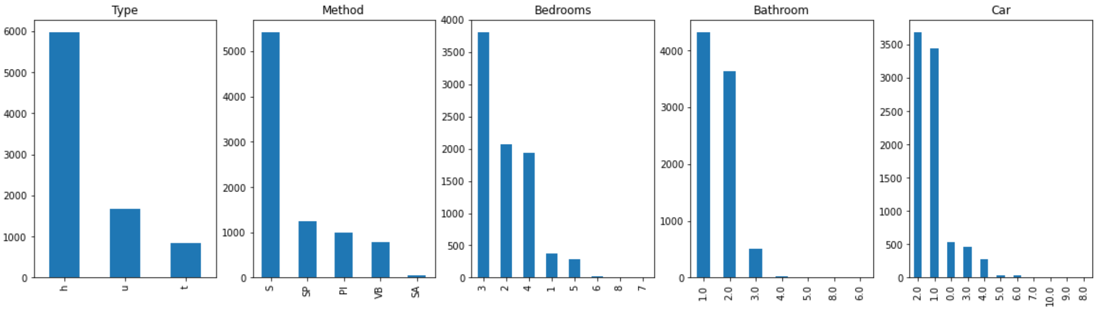

Figure 2: Histogram of categorical data

b) Pre-processing Data

In the pre-processing section label encoder is used to encode categorical data in numerical data. Then Standard scalar is used to Standardize features by removing the mean and scaling to unit variance.

c) Model Training

Main target is to predict price of the property. Therefore, this requires regression model to predict price value. In this project Linear Regression model, Random Forest regression and Decision Tree regression model trained and evaluate which has the best performance on this scenario.

When training Linear Regression model, there is no hyper parameters to tune. Therefore, this model can be trained directly.

In Random Forest regression model, there were 3 main hyperparameters (n estimators, max depth and min samples leaf) to tune. Those parameters tuned manually using few loops and plot performance.

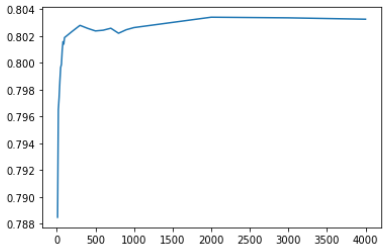

Figure 3:N Estimators performance

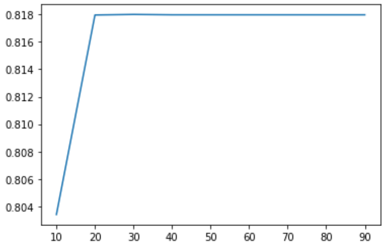

Figure 4:Max Depth performance

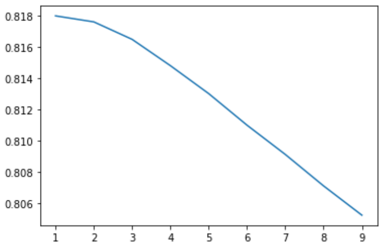

Figure 5: Min Sample leaf performance

For the Decision Tree regression model, there are several hyperparameters had to be tuned and for that process grid search cross validation was used with the following hyper-parameters.

Min samples split: [2,3,4,5,6,7]
Max depth: [12,14,16]
Min samples leaf: [1,2,3,4,5,6,7,8,9]
Max leaf nodes: [100,200,300]

After hyperparameters was tuned following was returned as the best parameters from the grid search cross validation function.

Max depth: 14
Max leaf nodes: 200
Min samples leaf: 8
Min samples split: 2

4) RESULTS

The trained models that mentioned above were evaluate using main 3 types of values. Mean Square Error, Root Mean Square Error and R2 Score. While training those models a model function was developed calcite these values at end of training using test dataset. The calculate values for each different models are in below table.

Table 2: Model Performance Comparison

Model Name	Mean Square Error	Root Mean Square Error	R2 Score
Logistic Regression	8.055428e+10	283820.864989	0.594920
Random Forest Regressor	3.639213e+10	190767.210631	0.816997
Decision Tree Regressor	5.229548e+10	228682.056369	0.737024

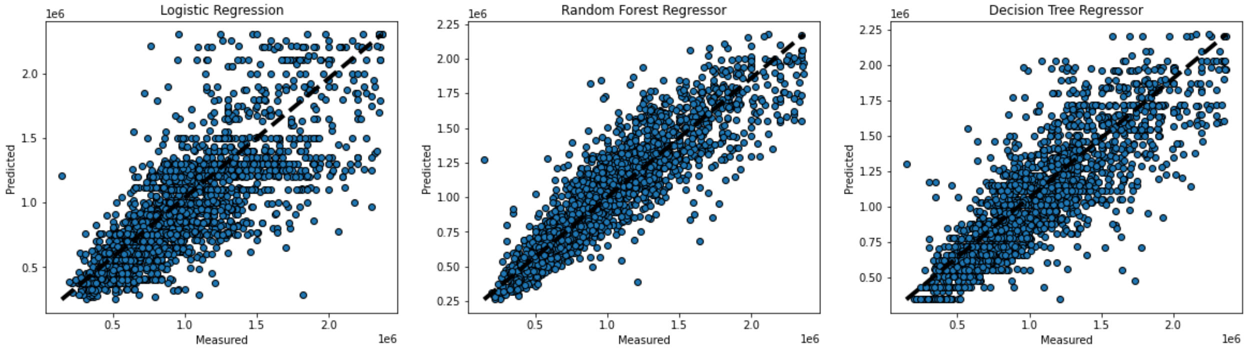

Figure 6: Model Prediction cross validation

5) CONCLUSION

When comparing the performance of the above three models, Random Forest Regressor has the highest score and the minimum error. By referring the cross validation figure this model performed with low bias and low variance when comparing to the other models. According that figure Random Forest Regressor model can be used in real life application because price prediction can be allowed within a range.

6) DISCUSSION

The main drawback of the trained model is the slightly lower score. This may reduce by balancing the dataset. As another suggestion Deep Leaning models can be used to get high accuracy.

7) REFERENCES

[1]"Melbourne Housing Market", *Kaggle.com*, 2021. [Online]. Available: https://www.kaggle.com/anthonypino/melbourne-housing-market.