Melbourne Housing Price Prediction – Regression Models

SIT307 8.2D

Introduction

For this task, I worked on predicting housing prices in Melbourne using different regression models. The main idea was to see how well machine learning can estimate property prices based on features like suburb, property type, number of rooms, and land size. I went through the whole process of preparing the data, exploring it with some visualisations, building models, and then checking which one performed the best. In the end, I also set up a small demo application to make the model usable.

Data and Preprocessing

The dataset had around 174 housing records taken from three suburbs. Each property had details such as the number of bedrooms, bathrooms, car spaces, the land size, the suburb it was located in, and the sale date. The target variable was the sold price, which I had to clean because it originally came with dollar signs and commas.

I fixed missing values by filling in medians for numbers like car spaces and land size, and replaced missing agents with "Unknown." I also converted the sale date into a proper datetime format. From there, I created some extra features, such as the year and month of sale, the total number of rooms (bedrooms plus bathrooms), and flags for whether the property was a house or an apartment. These new features helped the models capture more information about each property.

Exploratory Data Analysis

Looking at the dataset, the price distribution was skewed: most houses were below $1 million, but there were some very expensive properties in the millions. The heatmap of correlations showed that bedrooms, bathrooms, and land size all had some relationship with price, though not as strong as suburb and property type. This made sense because location plays a huge role in Melbourne's real estate market.

I also noticed that prices tended to change over time, which showed up clearly when looking at sale dates across months.

Model Development

I tested three different models: linear regression, a decision tree, and a random forest. To evaluate them, I used 5-fold cross validation and looked at three metrics: MAE, RMSE, and R^2.

Linear regression gave reasonable results but missed a lot of the non-linear patterns. The decision tree tended to overfit and wasn't as good on the test set. The random forest came out on top, with the lowest errors and an R^2 score of about 0.78. When I tested it on a hold-out test set, it performed very similarly, which made me more confident it wasn't just memorising the data.

Feature Importance

One of the nice things about the random forest model is that it can show which features matter the most. In my results, suburb and land size were the biggest drivers of price, followed by the

number of rooms and whether it was a house or an apartment. This lined up with common sense: where the house is located and how big it is are usually the first things people look at.

Deployment

To make the model usable, I exported it and built a small Streamlit app. In the app, users can type in details like the suburb, property type, number of bedrooms, bathrooms, car spaces, and land size, and the model will predict the price. This makes the project more practical rather than just theoretical.

Discussion

The project showed that machine learning can give decent price estimates even on a relatively small dataset. The random forest model worked best, and it also gave insights into which features influence prices. At the same time, the dataset was small and covered only three suburbs, so the model wouldn't generalise to all of Melbourne yet. To improve it, more data would be needed, and the model would need to be retrained regularly to reflect changing market conditions.

Conclusion

Overall, I was able to collect, clean, and prepare housing data, try out different models, and evaluate them properly. The random forest turned out to be the strongest model and was deployed in a demo app for users to interact with. While there are limitations, this project shows the value of applying regression models to real-world problems like property price prediction.