**Predicting Real Estate Price**

Lincoln Brown and James Mulvihill

Bellevue University

Predictive Analytics 630

Andrew Hua

**Predicting Real Estate Prices**

What contributes to house price? What are good predictors of housing market bubbles? How does a home appraiser estimate how much my home is worth? These are questions that data science applied to real estate can shed light on.

**The Data**

In 2017, Melbourne, Australia was voted for the seventh year in a row the world’s most liveable city. [[1]](#footnote-0) While it hasn’t done as well since then (mainly because of COVID), we will be looking at data from the housing market as it was in 2017. The data we will be using comes from Kaggle.[[2]](#footnote-1) The dataset has over 13,000 rows and columns that include the suburb the house is in, number of rooms, size, year built and the all important target variable - price.

**The Models**

It is likely we will need to experiment with multiple models before selecting the one that best fits our data. For starters, we might try a linear regression model. We also might try a random forest.

**Why are these Models Used?**

**Plan of Evaluation**

We will likely use RMSE to evaluate the model towards the end.

**What We hope to learn**

The goal would be to find a good model for predicting house price given the features of the dataset.

**Risks and Ethical Consequences**

One risk is that bad data could skew our results. Inside the dataset, there are a few houses that have huge square meterage. This is unlikely to be the case, especially if the price is around average. The data will need to be carefully cleaned to make sure rows like this do not skew our results. Another risk would be that the data does not contain important factors that would influence house price. Most notably, the quality of the house is not stated in the dataset. While year built might be correlated with this, we don’t have a clear way of knowing if the house is in disrepair.

The data is collected from an open source repository on kaggle, so there are no ethical concerns with how the data was obtained. If there are any other ethical concerns that emerge we’ll seek guidance at that time.

**Plan for the Unexpected**

It’s possible that we might need to use a different dataset. This could happen if there is confusion about what some of the columns in the Melbourne dataset signify. For example, one column is titled “Rooms”, which might just be bedrooms, but, coming from Australia, might mean something else. As a backup, we will most likely use a dataset from kaggle or a similar source that looks at real estate data from Ames, Iowa.

1. *The Guardian.* Melbourne 'world's most liveable city' for seventh year running. Retrieved 3/21/2024. <https://www.theguardian.com/australia-news/2017/aug/16/melbourne-worlds-most-liveable-city-for-seventh-year-running> [↑](#footnote-ref-0)
2. kaggle. Melbourne Housing Snapshot. Retrieved 3/21/2024. <https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot> [↑](#footnote-ref-1)