Predicting Housing Figures From Local Demographics

Summary

Housing is an important part of the Australian economy. A home is a crucial point of reference—in memory, feeling, and imagination—for inventing the story of ourselves, our life-narrative, for understanding our place in time. Meanwhile, as we have seen in the news, the housing market in Australia has dipped when COVID struck, and has been rising sharply again recently.

In this project, we are curious to see how the relationship between housing data, demographics, and certain economic factors interact and whether we could predict housing data in each postcode/LGA with these variables. We will explore questions such as “Where should one buy a home?”, “Can we predict prices based on socioeconomic data such as income levels, household size, or crime rate, etc.?”.

Description

We will analyse the NSW housing market through the following:

1. **Track the trend** of house sales and rental activities over the period of Q3 2017 to Q1 2021 in terms of the number of sales/bonds and price/rent, aiming to provide insights into following questions: (i) Are more people buying or investing in properties nowadays compared to the pre-COVID period? (ii) How have the housing and rental activities changed since COVID (Q1 2020)? Are there any correlations between both? (iii) Are there any differences across LGAs? What are the most popular and fast growing LGAs in the housing market?
2. **Identify the correlations** between house market activities as measured by the number of sales in the local area and selected demographic and socioeconomic features of the LGAs.
3. **Predict house prices** using the selected socioeconomic data, socioeconomic data from previous periods, and house prices from previous periods. In
4. **Find similar LGAs**: Addition to overall predictions, we will try to identify clusters of similar LGAs, e.g. rural vs metropolitan, and predict within these clusters to have higher predictive power for features in those clusters.

Datasets

House sales and Rent Data (NSW Family & Community Services)

1. These data include house and rent prices in different quartiles per postcode
2. These data are in xlsx format. Data cleaning would involve: (i) Selecting sheet name and Headers (ii) Taking care of NA values, removing unnecessary columns (iii) Clustering columns in a way that makes sense (iv) Combining all datasets

Census Data (TableBuilder service by ABS)

2016 Census data is the latest available and we acknowledge that some information might be outdated by now especially considering the change of circumstances in the past couple years, which might inadvertently impact the accuracy of our prediction.

1. These data are available in both csv and xlsx formats.
2. Some of the census data may include: (i) Personal income of the residents (ii) Household size (iii) Distance to work (iv) Marital Status (v) Hours Worked (vi) SEIFA scores (stat by ABS) (vii) Population and age brackets (by NSW Planning)

Others

1. Map data and shape of NSW (ABS)
2. Crime rate (NSW BOCSAR): Data are provided in xlsx format, data cleaning might involve: (i) Clustering into quarterly data (ii) Filtering which crime might be more relevant

Methods

We expect to use a linear regression model as the base model for predicting housing figures. However, further models will be used to check if there are better-fitting ones, such as XGboost. RFE may be used for feature-engineering. Clustering methods such as kmeans will also be used to see if there are notable similarities within the postcodes in NSW.

Project Plan and Milestones

Milestone 1: A proposal with the required sections is finished and submitted.

Milestone 2: Clean Data (Week 8) - All datasets have been merged, cleaned, and ready to be analysed and ran for a basic machine learning algorithm. The first part includes variable identification, univariate analysis, bi-variate analysis, missing value treatment. Also, normalisation of features has to be considered due to different magnitudes.

Milestone 3: Improve ML Model (Week 11) - A machine learning model has been identified that is significantly better than the base model. Identify suitable metrics to measure model accuracy and to be able to compare different models.

Milestone 4: Analysis Complete (Week 12) - When interpretations and analysis of the results are made. Also, the analysis addresses the initially asked questions. Last, useful graphs are added throughout the notebook to supplement and convey findings.

Milestone 5: Structure and Presentation (Week 13) - When a neat Jupyter notebook is ready for submission and the video presentation is finished to present the results. Furthermore, a brief but meaningful summary of the work is shown in the README file in the GitHub repository and the structure of the repository is checked for completeness and suitable structure.