Computer Science & IT, School of Science

COSC2669 Case Studies in Data Science

WIL Final Report

Group 38 (THE 6 DIMENSIONS)

Oct 2020

Seoungyeon Back: s3805769@student.rmit.edu.au

Vijay Lakshmanan Iyer: s3797863@student.rmit.edu.au

Xian Jing Wong: s3772149@student.rmit.edu.au

Yufei Wang：s3246253@student.rmit.edu.au

Table of contents

1.0 Introduction

1.1 Project Background

The COVID-19 pandemic, also known as the coronavirus pandemic, is a worldwide ongoing pandemic of coronavirus disease 2019 (COVID-19) [1]. The disease was first identified in December 2019 in Wuhan, China and has been spreading worldwide for over 10 months . The outbreak was declared a Public Health Emergency of International Concern in January 2020 and was recognised as a pandemic in March 2020 [2]. As of 9 October 2020, 36.5 million cases had been confirmed worldwide, and more than 1.06 million deaths had been attributed to COVID-19 [3].

The spread of COVID-19 and the efforts to quarantine the disease has caused unprecedented impacts on countries around the world, both socially and economically. The pandemic caused the largest global recession in history, with more than a third of the global population at the time being placed on lockdown [4]. Global stock markets was significantly impacted and displayed the worst performance since 1987 [5]. The pandemic caused a worldwide unemployment rate surge and income loss for over 400 million full-time jobs [6]. Major global consequences due to the pandemic include the supply-side manufacturing issues as well as the decreased business in the services sector such as retail, restaurant, tourism and entertaining industry [4].

This contagious disease has brought tremendous impacts to Australia with more than 23000 Australian infected and over 420 killed [7]. Even though the pandemic in Australia is fading, it has caused unprecedented economic damage across all industries with high unemployment rate and elevated debts [7]. One of the industries that is being potentially impacted is the real estate industry. There are a number of factors influencing property price including population, economy, interest rate as well as government regulations [8]. Of all the states and territories in Australia, Victoria has experienced two waves of COVID-19 attacks, causing the state to stay in a number of stages of lockdown for more than 6 months.

1.2 Problem statement

This project aims to gain further understanding of the impact of COVID-19 on the real estate business in Victoria, Australia. The business consideration behind this is that property price is related to a range of social and economical factors which could be directly or indirectly linked to the COVID-19 pandemic. For example, the overall national economy performance, currency rate, national unemployment rate, interest rate and so forth. How does the pandemic influencing Victorian property price and how will the property price change in the future upon the severity of this pandemic, is the core issue to be addressed in this project.

The use-case this project addresses is:

▪ As a real estate agent or a property investor, I would like to know how the property price has been influenced by the COVID-19 pandemic so far in 2020 and what are the predictions of property price in the future

2.0 Methodology

2.1 Dataset features and metrics

A range of features were considered for this investigation including housing sale relevant data, COVID-19 test cases & Victorian lockdown status as well as other social and economical factors such as unemployment rate and Australian dollar currency rate. Table 1 below summaries the attributes considered as well as their metrics.

Table 1. Variables investigated and their metrics

Attribute	Metrics
Housing sale data including sale price and sale date	Data retrieved from RealEstate website from December 2018 – July 2020 [9]
Property related data including property location - suburb, region, no. of bedroom in property, property type,	Data retrieved from RealEstate website from December 2018 – July 2020 [9]
Other property related data including rank of suburb, distance of suburb to CBD	Data from March 2020 to July 2020 retrieved from Domain [10] and myboot.com [11]
COVID-19 related data including daily test numbers, confirmed cases by surburb and lockdown status	Data from March 2020 to July 2020 retrieved from informgram.com [12] and covid19data.com.au [13].
Victoria unemployment rate	Data in Australian labour market information website from March 2020 to July 2020 [14]
AUD currency rate	Data in Australian reserve bank website from March 2020 to July 2020 [15]

2.2 Data preparation

Data cleaning/preparation was performed using the Python packages pandas [16] and NumPy [17]. Prior to the modelling process, feature selection was performed using RandomForestRegressor to eliminate the unnecessary variables. Feature importance table was extracted from the RandomForestRegressor from the Scikit-learn package [18]. Although the table showed that several features displayed 0.00% importance, those less important features were dummy variables that transformed from the original features (house type, suburb name, race and lockdown stage). Therefore, all features were remained. The data has been normalized, by using the Preprocessing. MinMaxScaler from Scikit-learn package, therefore different features can be assessed uniformly in the following modelling process.

2.3 Exploratory analysis

Exploratory data analysis was performed by visualising the change of property price over the time of 2018 to 2020. The python plotting package Seaborn [19] and R studio plotting package ggplot [20] were used for data visualisation. The relationships between pairs of attributes were investigated using the correlation heat map (Figure 1). The links between the features and the property price were also displayed by using the “group by class” function during the visualisation process.

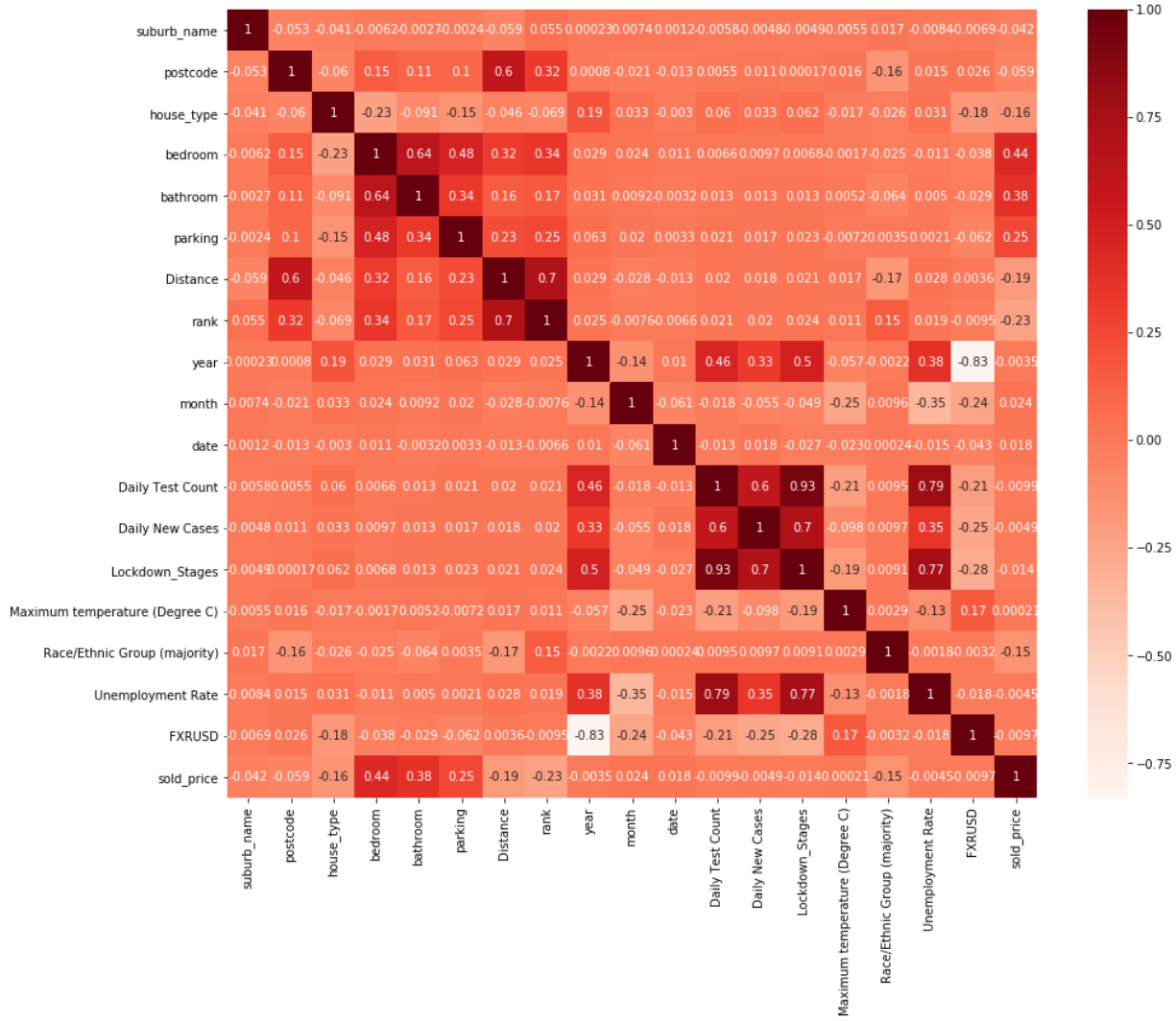

Figure 1 Correlation heat map of features considered

2.4 Modelling

Following data preprocessing, modelling process commenced by using a 60:40 training-testing splitting on the entire dataset. Comparison table of regression accuracy among different regressor (Ridge, lasso, ElasticNet, SupportVectorRegressor, RandomForestRegressor, extratreesregressor, baggingregressor, huberregressor, bayridge, xgb, decisiontreeregressor, KneighborsRegressor and GradientBoostingRegressor). RandomForestRegressor was selected as it could achieve the highest accuracy of prediction (Figure 2). Hyperparameter tuning on RandomForestRegressor was carried out.

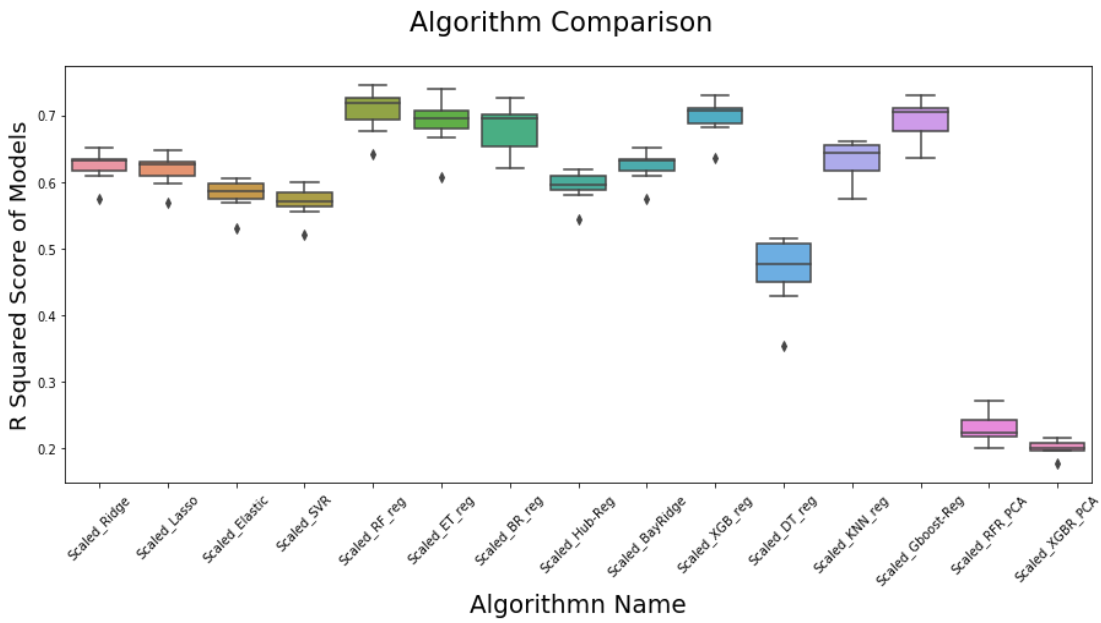

Figure 2 Comparison of regression accuracy among different algorithms of regressor

3.0 Proposed solution (answers/results)

The analysis of house price impact from COVID-19 as well as the prediction of future house pricing upon COVID-19 influences will address the business problems and provide valuable business insights for the real estate agents and property investors. To deliver both the house pricing insight analysis and the prediction of future house price trending, the proposed solution is a website integrated with two functionalities including a dashboard and a prediction page allowing user inputs.

3.1 Dashboard – exploratory data analysis

The dashboard can provide information on how the COVID-19 pandemic is influencing the housing market. Following is an example of the prototype of the dashboard page. Here Plotly graphs are used on the website to make the dashboard interactive and user friendly (Figure 3). Comparisons of house pricing over time and correlations between house pricing and a range of factors were analysed and can be displayed on the dashboard. A few key insights have been drawn from our analysis and are further discussed below.

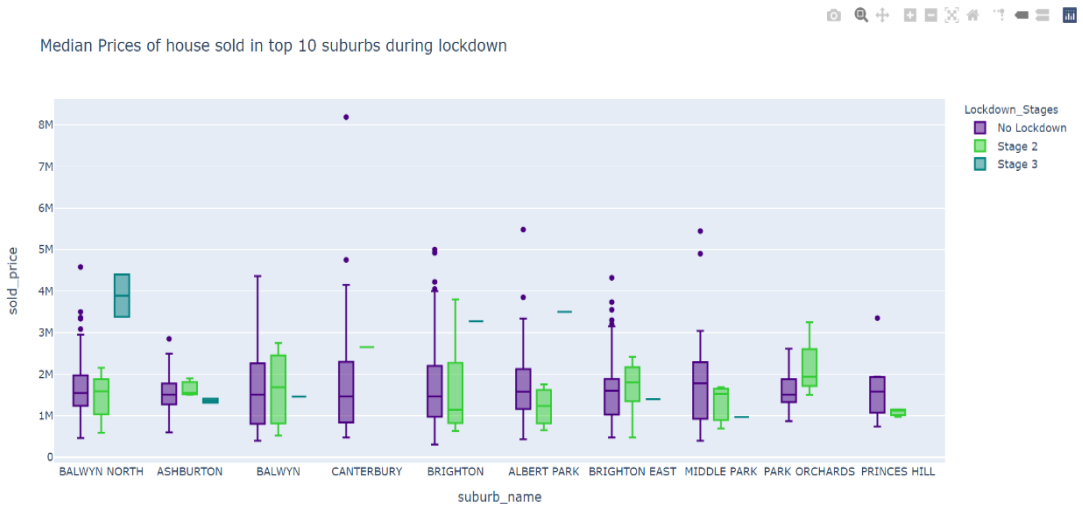

Figure 3 Prototype dashboard page of website

3.1.1 Overall observation

Figure 4a demonstrates that all property prices have risen overall from 2019 in Victoria and have not been affected by the COVID-19 pandemic. This information should bring the agents and investors some comfort that the pandemic has not caused catastrophic damage to the real estate market. A negative effect brought on by COVID-19 was the rise in unemployment. Some real estate agents may be concerned about their business being impacted. However, our analysis (Figure 4b) shows that the real estate industry remains intact during the COVID-19 pandemic. Also, house prices were not significantly affected despite the higher unemployment rate in 2020.

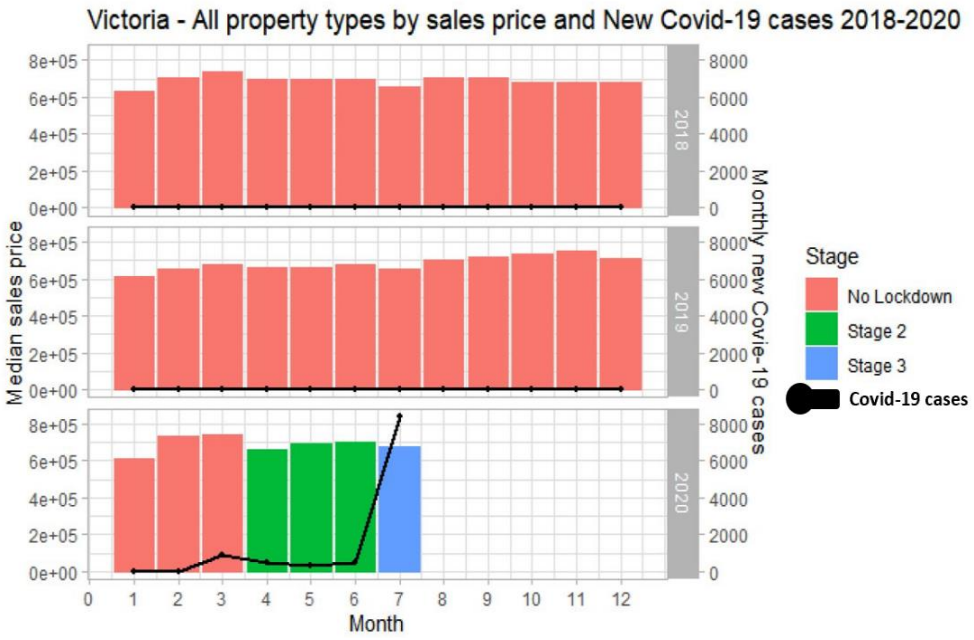

Figure 4a House price for the last 3 years

Figure 4b Unemployment rate and house price for the last 3 years

3.1.2 House pricing by region

Investors would like to know how each region in Victoria is being impacted and where is the safest spot to put their money. They can find out the house pricing trend by region from this graph (Figure 5). Clearly, Eastern metropolitan Victoria is a good place to have already invested given that now the sale price appears to be increasing.

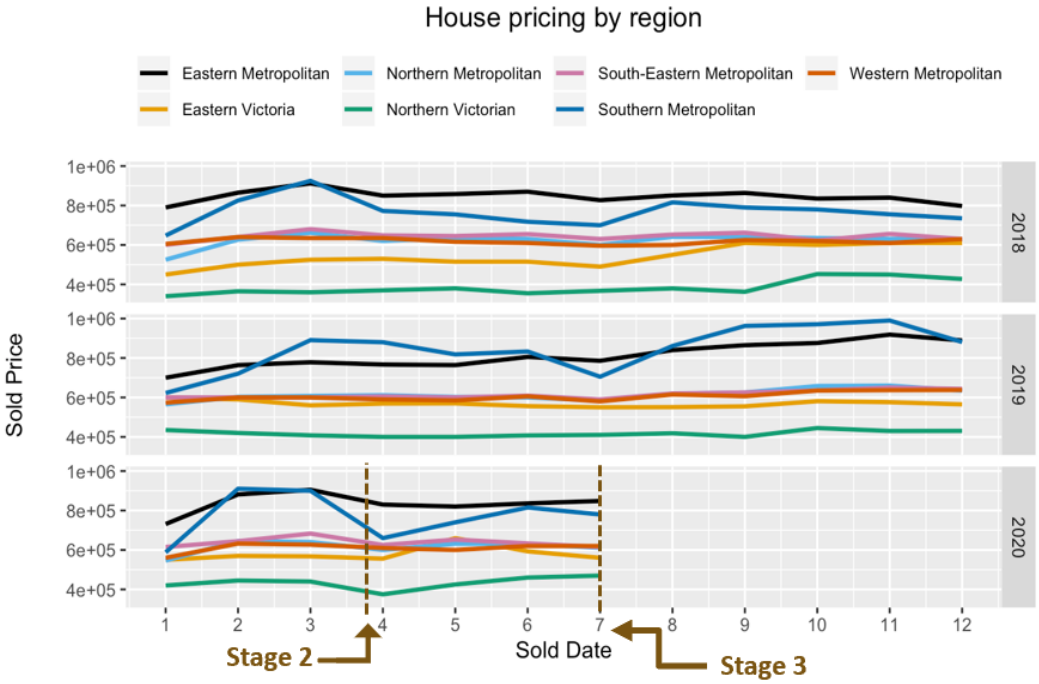

Figure 5 House pricing by region

3.1.3 House pricing by suburb

Real estate agents can provide the information of house price by suburbs with a statistical analysis (Figure 6) and a visualisation (Figure 7). All suburbs are ordered by median house price for the last 3 years with total sales. When investors select the suburb, they can see the trend of the house price of the suburb.

suburb_name <chr>	Median <dbl>	Total_sales <int>
CANTERBURY	1898588.9	55
ALBERT PARK	1752393.8	73
MIDDLE PARK	1746659.1	44
BRIGHTON	1705765.6	192
BALWYN NORTH	1640996.5	195
PARK ORCHARDS	1639657.4	48
BALWYN	1599377.6	145
BRIGHTON EAST	1599106.7	177
ASHBURTON	1517492.2	76
PRINCES HILL	1502300.0	10

Figure 6 Suburbs ordered by median house price

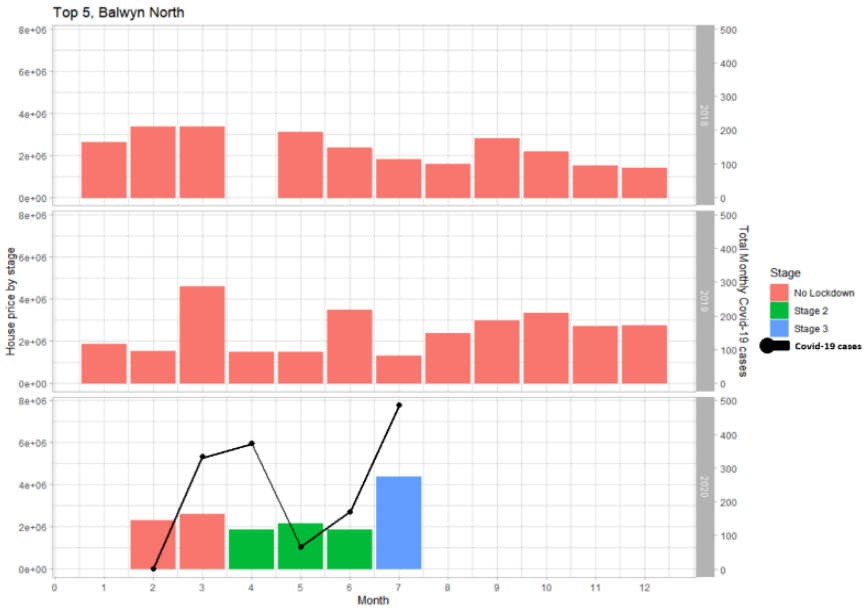

Figure 7 Balwyn North visualisation

3.1.4 House pricing by house type

Overall, the sales price of all house types in Victoria has not been affected by the COVID-19 pandemic but our analysis (Figure 8) shows that there is a different aspect of the property price regarding the house type. From Figure 9, it is found that the sale numbers of ‘Studios’ and ‘Flats’ dramatically dropped while ‘House’ price increased

over the time of COVID-19 (Figure 10).

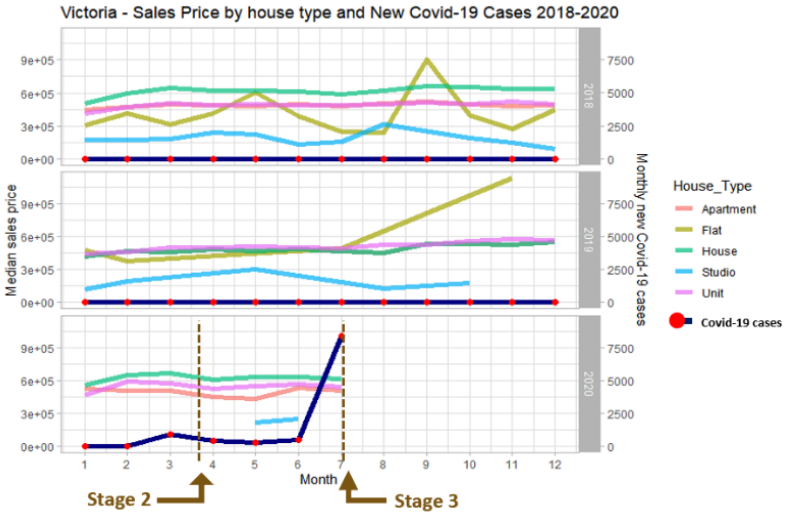

Figure 8 Sales price by house type

Figure 9 Studio sales Price

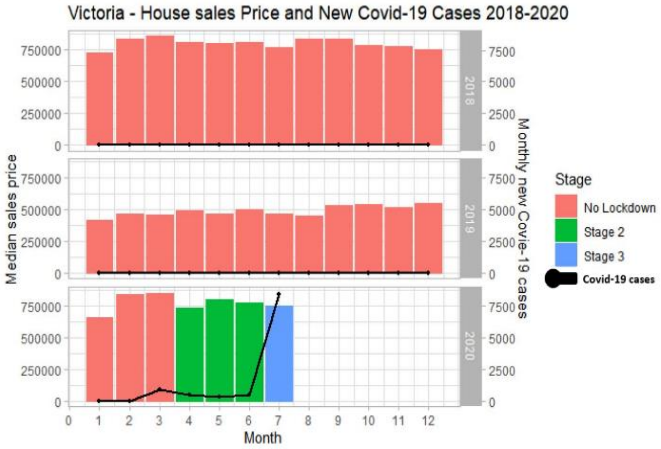

Figure 10 House sales Price

3.2 Predictive webpage – predictive modelling

Another major functionality, the website provides is the prediction model as in a form allowing user inputs [21]. The input fields are all relating to the variables used for training the model for prediction. Then, on the button click “submit”, the underlying model would then give the estimated house price. This feature is useful for the customers if they are planning to purchase an apartment or an individual house in the future, then this would help them in making necessary plans to arrange money for purchasing the same. Also, this feature might help the real estate companies in helping the house owners, by giving them an insight on for what price their house or apartment can be put for sale and also, for putting up an ad for sale in the company’s website.

Following is the screenshot of the form page, which requires the user input.

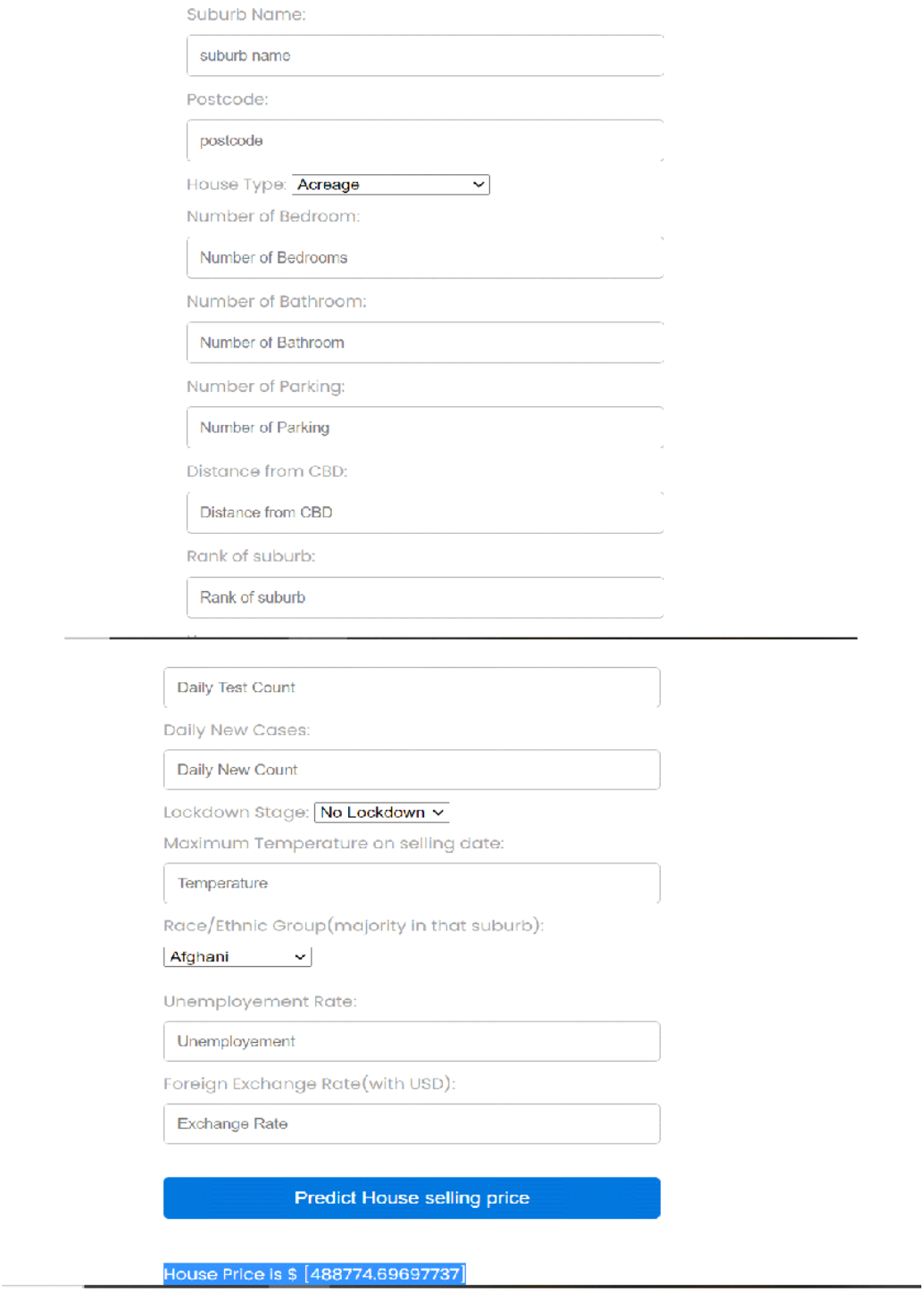

Figure 11 Predictive model webpage inputs

4.0 Significance of results

The business value provided by the entire project is to have a better understanding of how does the COVID-19 pandemic have impacted the Victorian real estate industry by analysing the property prices in various suburbs in correlation to the COVID-19 tests, confirmed cases as well as state lockdown status and displaying as in a form of dashboard. Furthermore, a predictive model was developed to give both the customer and a real estate company, an insight on what would be the expected in terms of house price trending and estimation on a given date in the future (pandemic recovery stage).

The outcome of this study can be used by the real estate agents and the customers to gain insights of the market price trending based on the different parameters like the unemployment rate, stages of lockdown. In turn, this allows them to fix the weekly / monthly rent for the properties associated with their company. For example, the real estate companies can use these results to propose the reasonable rent for the customers without jeopardising the business and incurring loss to the landlord owner and also, ensuring the customers who are affected by the pandemic are able to find home within their budget. Moreover, property investors can use the predictive model outcomes to have a better idea of the market in various suburbs, so to facilitate their investment decision making. Customers, who might have lost jobs, can easily study various analysis and visualisation to find out the right suburbs where the price of the house is reasonable.

5.0 Project management

5.1 Project Management Plan

The team planned 4 stages to progress the entire project using a lean methodology.

- Stage 1 (Week 2-4):

  Exploration of data sets, define a business problem from the data set that we selected, complete the WIL milestone report

- Stage 2 (Week 5-mid-break):

  Find more data sets to support business problem solving, data pre-processing (E.g. reshaping/ filtering/ cleansing data, identifying outliers, transforming variables

and merging data sets), data exploration, data modelling, identify a clear business problem for industry based on the data exploration

- Stage 3 (Week 7-9):

Data visualisation, API, code checking, finalising oral presentation

- Stage 4 (Week 10-12):

Finalising WIL final report for submission

5.2 Project Management Tools

Using a Microsoft Teams video call, our team had weekly meetings for effective team communication and discussions of what we are working on now and what we are working on in the next step. The business problems, data sets for our project and the progress of each task were discussed and agreed by all team members through video meetings.

Microsoft Teams was the main project management tool for regular chat conversations and sharing files for the WIL project. We also shared links of data sets and resources that might help for our project. Microsoft Teams planner was used to keep track of our work involved in each phase of the project and to outline the due date of each task. Therefore, all team members were able to discuss and approve our tasks (Figure 12). Each task was assigned by an appropriate team member based on our key skills and experience. Hence, all members could contribute equally in the project.

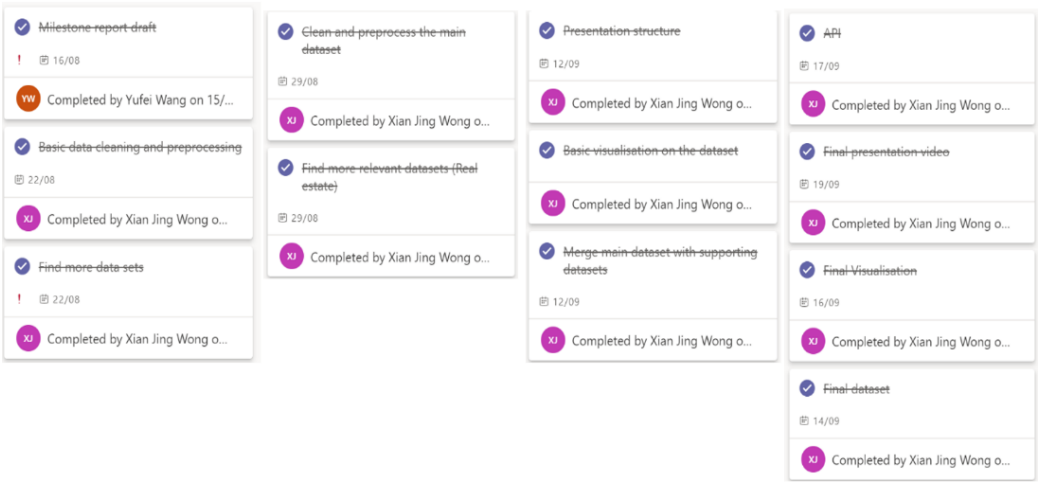

Figure 12 Microsoft Team Planner (week 5-9)

A GitHub repository was used for the code developed by all members during the project.

All data sets that we collected were shared in GitHub and then once a team member started data pre-processing, other members modified and checked the code. All members reviewed every code to approve the task. The progress of data exploration, data visualisation, analysis and data modelling were started and shared in GitHub.

Teamwork of the 6 Dimensions (Group 38) was highly and well-organized project management and good communication through weekly video meetings. A GANTT chart shows our WIL project plan and our members' contribution (Figure 13).

GANTT CHART- The 6 Dimensions

TASK NAME	START DATE	END DATE	START ON DAY*	DURATION* (WORK DAYS)	TEAM MEMBER	PERCENT COMPLETE
First stage						
Find datasets	8/2	8/16	0	15	All	100%
WIL milestone 1 -report	8/2	8/16	0	15	All	100%
Decide data sets	8/9	8/23	7	15	All	100%
Define a business goal	8/9	8/23	1/7	15	All	100%
Second stage						
Find more data sets to support business problem solving	8/17	8/29	15	13	All	100%
Data pre-processing	8/22	9/13	20	1/23	Jack, Vijay	100%
Data exploration	9/13	9/20	42	1/8	Jack, Julia	100%
Data Modelling	9/13	9/20	42	8	Jack	100%
Find solutions of business problems based on the data exploration	9/6	9/20	35	15	All	100%
Third Sample Project						
Data visualisation	9/13	9/20	2/11	1/8	Julia, Emily	100%
API	9/13	9/20	2/11	1/8	Vijay	100%
Code checking and Bug fix from stage 2	9/13	9/20	2/11	1/8	All	100%
Make presentation slides	9/21	9/23	2/19	1/3	All	100%
Make talking points of a presentation	9/21	9/23	2/19	1/3	Emily	100%
Record Oral presentation	9/23	9/24	2/21	1/2	All	100%
Edit a video	9/24	9/25	2/22	1/2	Jack	100%
Fourth stage						
WIL final report	9/26	10/16	55	21	All	50%

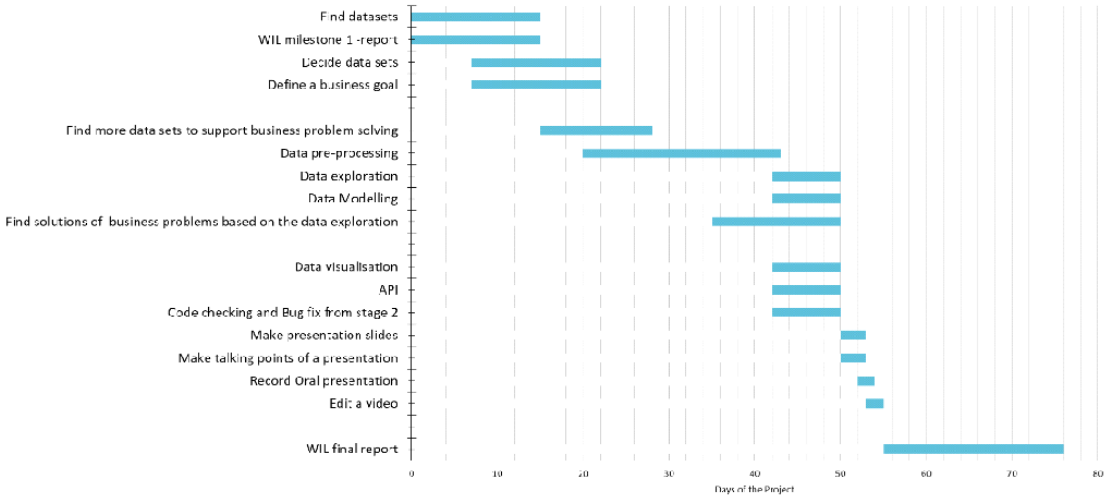

Figure 13 GANTT chart

6.0 References

[1] ‘Naming the coronavirus disease (COVID-19) and the virus that causes it’. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it (accessed Oct. 09, 2020).

[2] ‘Novel Coronavirus – China’, *WHO*. http://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (accessed Oct. 09, 2020).

[3] ‘Coronavirus COVID-19 (2019-nCoV)’. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 (accessed Oct. 10, 2020).

[4] ‘Economic impact of the COVID-19 pandemic’, *Wikipedia*. Sep. 28, 2020, Accessed: Oct. 10, 2020. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Economic_impact_of_the_COVID-19_pandemic&oldid=980858784.

[5] L. J. Palumbo Daniele and D. Brown, ‘Coronavirus: A visual guide to the economic impact’, *BBC News*, Jun. 30, 2020.

[6] V. McKeever, ‘The coronavirus is expected to have cost 400 million jobs in the second quarter, UN labor agency estimates’, *CNBC*, Jun. 30, 2020. https://www.cnbc.com/2020/06/30/coronavirus-expected-to-cost-400-million-jobs-in-the-second-quarter.html (accessed Oct. 11, 2020).

[7] ‘The Costs of COVID: Australia’s Economic Prospects in a Wounded World’. https://www.lowyinstitute.org/publications/costs-covid-australia-economic-prospects-wounded-world (accessed Oct. 11, 2020).

[8] J. Nguyen, ‘4 Key Factors That Drive the Real Estate Market’, *Investopedia*. https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp (accessed Oct. 11, 2020).

[9] ‘Real Estate, Property & Homes for Sale - realestate.com.au’. https://www.realestate.com.au/buy (accessed Oct. 13, 2020).

[10] A. Barnes, ‘Melbourne 307 suburbs ranked for liveability’, *Domain*, Nov. 15, 2019. https://www.domain.com.au/liveable-melbourne/melbournes-most-liveable-suburbs-2019/melbournes-307-suburbs-ranked-for-liveability-2019-898676/ (accessed Oct. 13, 2020).

[11] ‘Australia > Victoria > Search Suburbs by Distance’. http://myboot.com.au/VIC/50/suburblist.aspx (accessed Oct. 13, 2020).

[12] ‘* Average daily tests vs confirmed cases in all jurisdictions - Infogram’. https://infogram.com/1pl653yn0n6klvhqp60evqxmmxbzxvdwzqn (accessed Oct. 13, 2020).

[13] ‘COVID-19 in Victoria: Coronavirus hotspots, postcode lockdowns’, *COVID-19-data-aus*. https://www.covid19data.com.au/victoria (accessed Oct. 13, 2020).

[14] ‘Welcome to the Labour Market Information Portal.’ https://lmip.gov.au/default.Aspx?LMIP/LFR_SAFOUR/LFR_UnemploymentRat

e (accessed Oct. 13, 2020).
[15] scheme=AGLSTERMS A. corporateName=Reserve B. of Australia and scheme=AGLSTERMS A. corporateName=Reserve B. of Australia, 'Historical Data', *Reserve Bank of Australia*. https://www.rba.gov.au/statistics/historical-data.html (accessed Oct. 13, 2020).
[16] 'pandas - Python Data Analysis Library'. https://pandas.pydata.org/ (accessed Oct. 13, 2020).
[17] 'NumPy'. https://numpy.org/ (accessed Oct. 13, 2020).
[18] 'scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation'. https://scikit-learn.org/stable/ (accessed Oct. 13, 2020).
[19] 'seaborn: statistical data visualization — seaborn 0.11.0 documentation'. https://seaborn.pydata.org/ (accessed Oct. 13, 2020).
[20] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016.
[21] 'Turning Machine Learning Models into APIs', DataCamp Community, Oct. 25, 2018. https://www.datacamp.com/community/tutorials/machine-learning-models-api-python (accessed Oct. 18, 2020).