Explore the relations between social economic factors and construction equipment repair transactions by using machine learning techniques

Li Jen Shao(N11096837)

IFN703/4 Assessment 3

Executive Summary

In this research we would like to predict Hastings Deering's customers monthly number of transactions. We also added external data from Government database to see the possibility of helping the prediction and getting useful insight. We added Producer Price Index, Number of Building Approvals, Value of Building Jobs, Number of Covid Cases because we expect to see these four external variables can have impact on the monthly number of transactions by each customer. We built a Random Forest model to predict monthly number of transactions, classification report, confusion matrix and permutation feature importance are the method to evaluate the prediction result. In terms of the result, we don't see any significant sign that prove the external variables can affect the monthly number of transactions.

Introduction

The Covid-19 pandemic was a major public health issue during 2020, 2021 and 2022, the virus had spread across the world massively and took tremendous lives, some specialists said the virus changed people's life and behaviour forever. Most of the people experienced lockdown during 2020 and 2021, people were forced to stay home to help stopping the spread of the virus. Thus, some business behaviour has been permanently changed since the pandemic, there were lockdown happened in major cities across the world, the Covid-19 recession had caused the stop of production activities, the disruption of worldwide supply chain, the virus gave such an impact on the base structure of global social economy.

Hastings Deering is a construction equipment supplier in Queensland Australia, it is one of the official suppliers of Caterpillar the US construction machinery and equipment company. In this research we would like to discuss the relations between social economic factors and the construction equipment repair transactions which provided by Hastings Deering, to see if there are any change of customer behaviour or insights we can reveal. The dataset contains customer repair and parts replacement transactions based in Toowoomba region Queensland from 2020 to 2021, which is the main period when the pandemic swept across Australia.

In this research, we would like to explore the relations from a social economic view in related to Toowoomba's housing market, we will use Producer Price Index (Brisbane), number of building approvals (Toowoomba), value of building jobs (Toowoomba) and number of Covid cases in Toowoomba as the external features and also as the indicators of Toowoomba housing market to add into the original dataset, this can give us some insights of the relations between Toowoomba housing market and construction equipment repair transactions. Potentially, we would like to see number of customer transactions increase as the housing market indicators (the four external features) indicates the market is flourish and the related activities increase also. The reason is when there are more building activities in Toowoomba, these Hastings Deering's customers would make more repair activities and order more replacements due to the usage of their construction equipment increase.

Producer Price Index (Brisbane)

The Producer Price Index (PPI) and Consumer Price Index (CPI) can be seen as general indicators of inflation and price, according to the definition from Australian Bureau of Statistics (ABS) [1], PPI measures the price change of products (goods and services) as they leave the place of production or as they enter the production process. This price change is measured from the perspective of the industries that produce goods and services. Whereas other measures, such as the Consumer Price Index (CPI), measure price change from the consumers perspective. There are many PPIs, one of the PPIs is the input to the house construction industry, this can reflect the trend of cost on building a house within Australia, as the cost increase the house price would also increase. From figure 1 we understand that during 2021 the cost of building houses and house price increased tremendously.

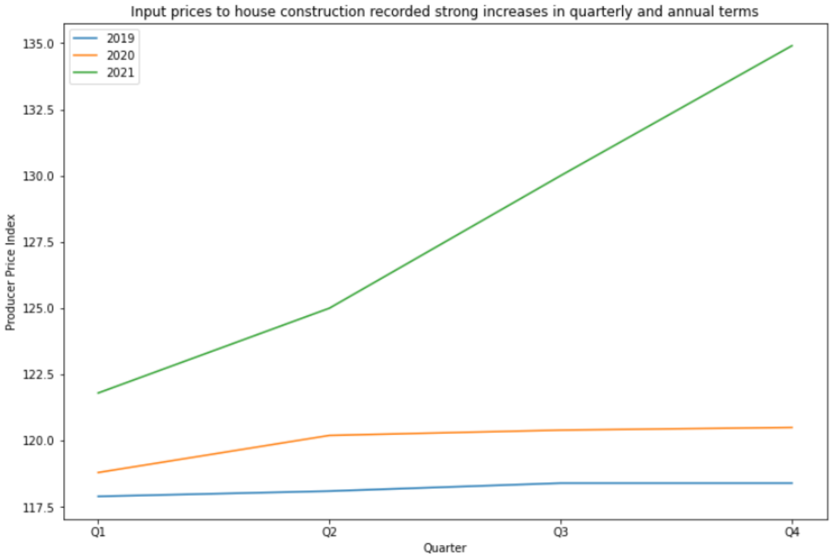


Figure 1: Quarterly PPI in Australia shows tremendously increase during 2021, 12% increased annually since January 2020. Data from ABS [1].

Monthly Number of building approvals and monthly value of building jobs in Toowoomba

These two features are more intuitive, they indicate the building activities around Toowoomba region. According to Queensland government site [2], a building development approval (or building permit) is needed before construction can start on most types of domestic building work. These two features should be highly correlated, if number of building approvals increase, the value of it should follows the trend. From figure 2 we understand that from late 2020 the monthly number of building approvals increased tremendously comparing to 2019. Figure 3 shows during 2021, the monthly value of building jobs is generally higher than 2020 and 2021.

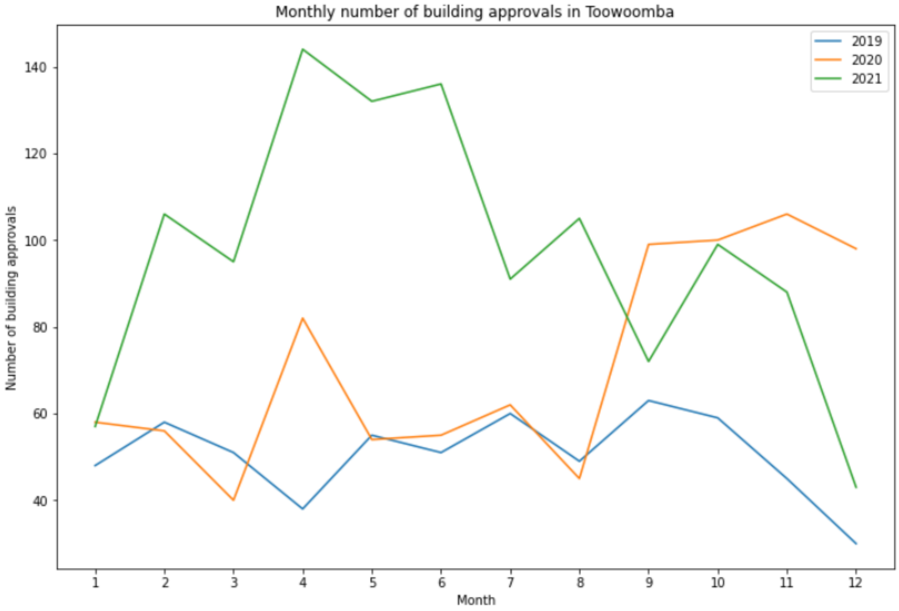


Figure 2: Monthly number of building approvals in Toowoomba region shows tremendously increase from late 2020 to 2021. Data from ABS [3].

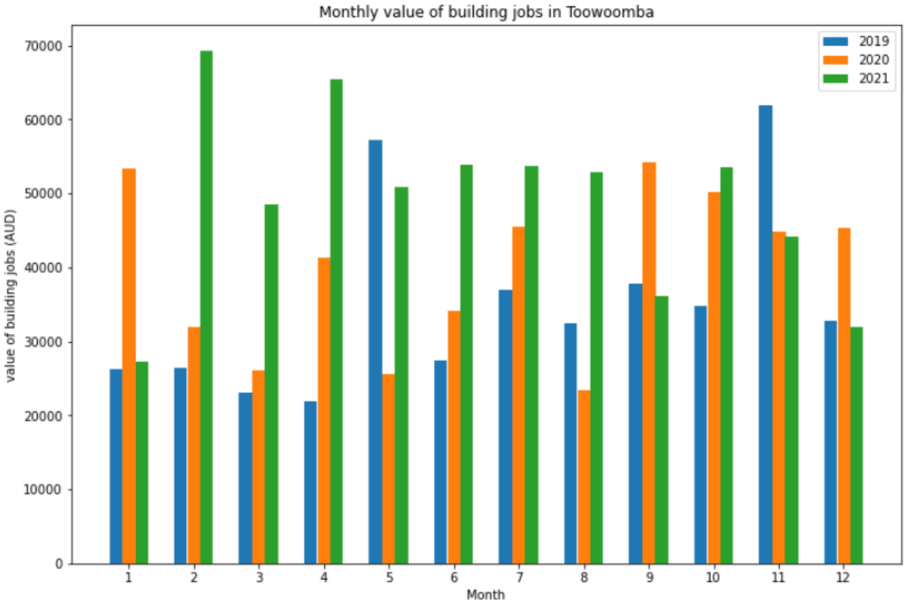


Figure 3: Monthly value of building jobs in Toowoomba region shows tremendously increase in 2021. Data from ABS [3].

Monthly number of Covid cases in Toowoomba

This feature indicates how hard the Covid-19 pandemic hit Toowoomba region and having effect on their business. Potentially, if the Covid cases increase, we would like to see the customer transactions decrease to show Covid can affect customers behaviour. Generally, before December 2021, Toowoomba region did not suffer from Covid-19, the only big outbreak is December 2021.

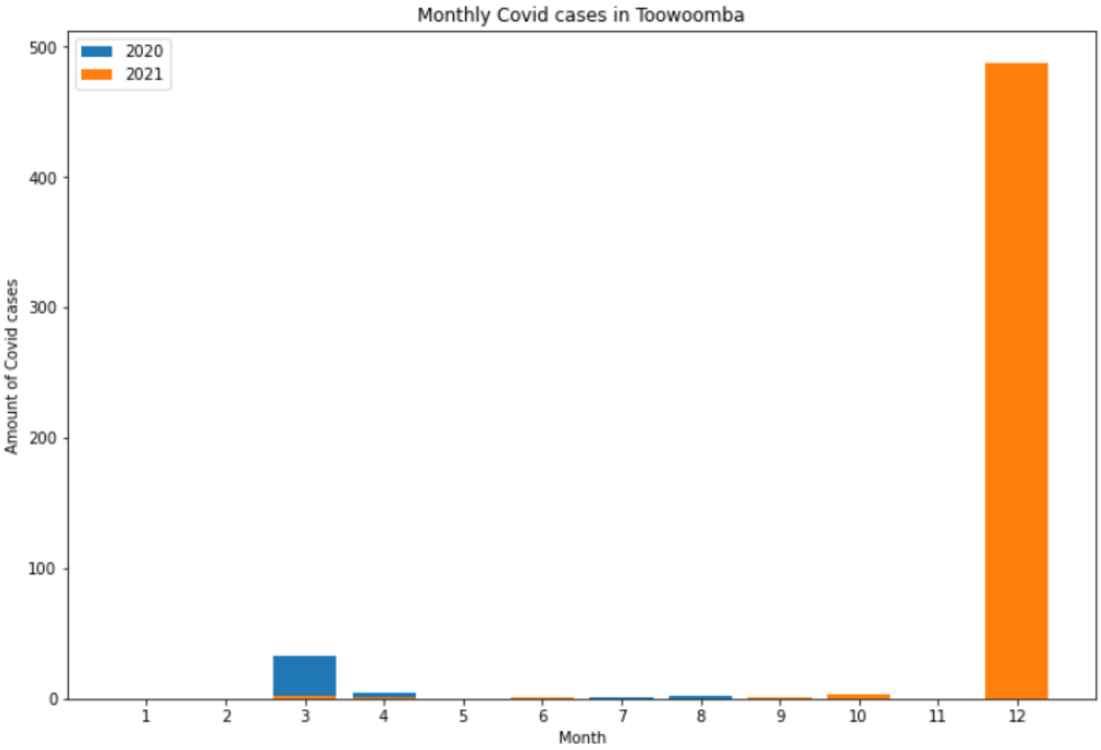


Figure 4: Monthly Covid cases in Toowoomba shows Covid-19 didn't spread to Toowoomba until December 2021. Data from Queensland Government Open data portal [4].

These four features can give us some imagination about the background, the housing market of Toowoomba. Combining four plots we now understand that comparing to 2019(the last year without Covid-19), the housing market flourished during late 2020 and 2021, the rise of house price and number of house constructions were recorded, the value of building jobs also increased during 2021. In terms of Covid cases, there was a small outbreak in March 2020, the major outbreak happened in December 2021. These insights can help us make sure the directions of this research, as we saw the Toowoomba housing market flourished during 2020 and 2021, we assume the monthly customer transactions would also increase due to the increase of usage of construction equipment.

The aim of this research is…

Literature review

As we mentioned above about the house price and inflation, we cannot help ourselves to ask, how is the house price in Toowoomba during 2020 and 2021? How to define inflation is good or bad? PRD is a professional real estate agency across Queensland, according to their report of Toowoomba market update 2022 made by PRD Toowoomba [5], it says in Q4 2021 Toowoomba recorded annual (Q4 2020 to Q4 2021) median price growth of 14.2% for houses and -5.9% for units. Between Q4 2020 – Q4 2021 total sales in both markets increased, by 24.7% for houses and by 59.6% for units. Owner occupiers and down-sizers can benefit from real returns on capital investment, as median price growth is alongside higher sales numbers. Ready-to-sell houses are in low supply, creating an opportunity for developers. This means the Toowoomba housing market had massive trading and developing activities during late 2020 and 2021, the demand and price for houses increased tremendously. For our construction contractors, this means new jobs and opportunities to use their equipment. When a trading happened, usually people would ask contractors to tidy up their new property, developing new property is more intuitive, the more properties they develop, the more usage of their machinery is needed, and more repair plans and parts would be ordered.

In terms of price and inflation, why is inflation important here? A good inflation can stimulate the housing market and lead to more trading and property development simply because the activities are profitable. Naturally we can expect more jobs and equipment usage for construction contractors if we have volumes in housing market. In contrast, a bad inflation can be vice versa, according to a report of construction inflation alert 2021 made by Associated General Contractors of America [6], it mentioned about how harmful can a bad inflation be. It says the extreme cost increases and supply-chain disruptions are affecting construction, the PPI of input to construction in US soared more than 24% annually since May 2020 to 2021. Given that materials often represent half or more of the cost of a contract, such an increase could easily wipe out the profit from a project and potentially put the contractor out of business. In another words, this means the extremely increase of material price can be seriously harmful to property developers, leads to the decrease of number of building projects simply because the activity is not profitable anymore. It can also lead to fewer jobs and opportunities to use their equipment for our construction contractors if the building projects becomes fewer. The less property they develop, the less usage of their machinery is needed, and less repair plans and parts would be ordered.

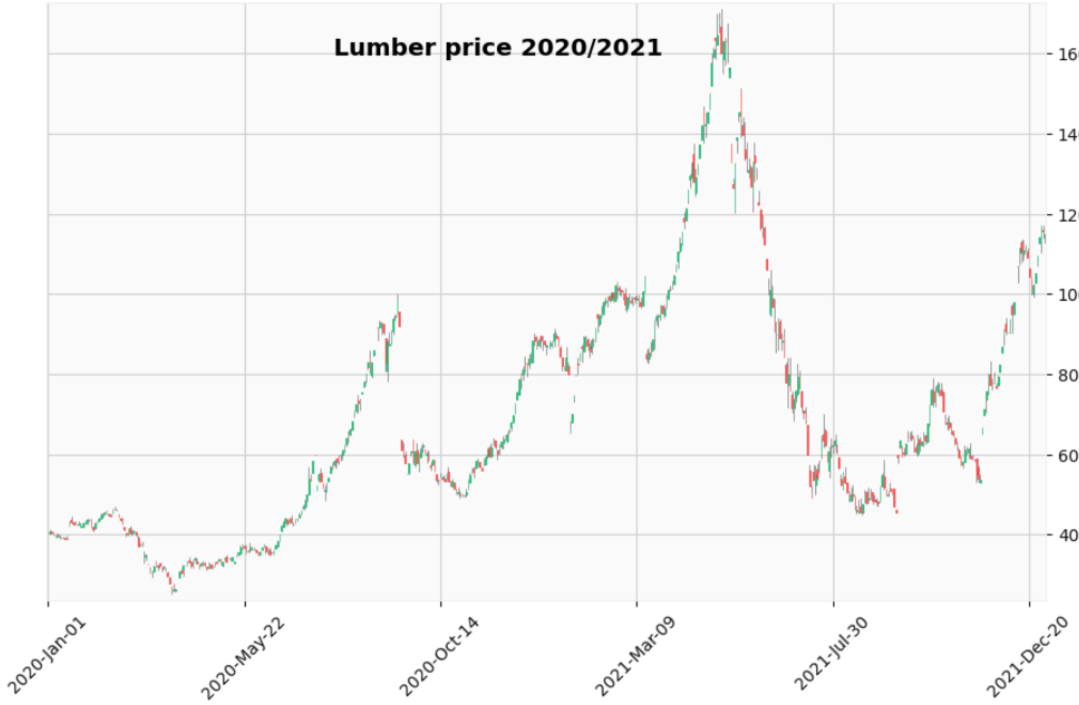


Figure 5: Price of lumber during 2020 and 2021. Lumber as one of the important materials to house construction, the price significantly rose after outbreak of the pandemic. Data from investing.com.au [7].

Generally, from consumer side of view, we expect our construction contractors make more transactions from late 2020 and 2021 due to the rise of activities in Toowoomba's housing market. From producer side of view, we concern about our construction contractors would make less transactions due to the significant rise of price for construction materials, that can lead to less building projects and less job opportunities. Note that comparing to over 24% rise of PPI in US, in figure 1 we recorded 12% rise of PPI in Australia, which indicates that inflation in Australia were not as harmful as US. Further, from figure 2 we can understand that property developers are willing to apply building approvals more than usual, this indicates that the developers think the housing market in Toowoomba is profitable and willing to make more building projects. Therefore, the initial assumption about a flourished Toowoomba housing market can cause more transactions by Hastings Deering's customers are still valid.

Approach

The Hastings Deering's dataset contains 35144 transactions from their Toowoomba warehouse, as we mentioned before these transactions are made by 874 construction contractors. We apply monthly scale on the dataset, a dictionary for the variables we will use has been provided:

- Customer_ID: Customer unique number generated by the system, a unique identifier.
- Customer_Market_Segment_ID: Market Segmentation code, categorical variable.
- Source_of_Supplier_ID: Describes type and make of parts, e.g., Caterpillar, etc.,

categorical variable.

- Major_Minor_Class: Parts grouping, e.g., Undercarriage, Idler, etc., categorical variable.
- Sales_Channel: Describes how the sale comes into the business, e.g. Over the Counter, Workshop, etc., categorical variable.
- Month: The month that made the transaction, categorical variable.
- Number_building_approvals(private_sector): Monthly number of building approvals (private sector) in Toowoomba region, Building Approvals by Statistical Area (SA2) and above, source from Australian bureau of statistics, measurement is number of dwelling units, continuous variable.
- Value_building_jobs(private_sector_AUD_thousands): Monthly value of building jobs (private sector) in Toowoomba region, source from Australian bureau of statistics, measurement is thousand AUD, continuous variable.
- Producer_price_index(brisbane): Quarterly PPI of input to the house construction industry in Brisbane, source from Australian bureau of statistics, continuous variable.
- Monthly_Covid_cases_Toowoomba: Monthly Covid-19 cases in Toowoomba region, source from QLD government open data portal, continuous variable.
- Quantile_monthly_transactions: An aggregated column, this is also the target column. Percentage of monthly number of transactions by the customer in quantiles, e.g., Q1 means the customer's monthly number of transactions are in first 25% of the monthly total number of transactions, ordinal categorical variable.

First 6 variables are originally from the dataset which provided by Hastings Deering, they are all categorical variables. Number_building_approvals, Value_building_jobs, Producer_price_index, Monthly_Covid_cases_Toowoomba are the external variables we added into the dataset, Producer Price Index are quarterly data based on Brisbane, the other 3 external variables are monthly data based on Toowoomba. Quantile_monthly_transactions is an aggregated variable made from the monthly counts of transactions by each customer, it has been processed to percentage of the counts for the month total number of transactions by all customers, then categorised as Q1, Q2, Q3 and Q4, depending on what quantile it is in the month, also transform to quantiles means we will have an extreme balanced ordinal categorical target variable.

We will use a Random Forest Classifier to classify the Quantile_monthly_transactions. Random Forest Classifier is an ensemble method machine learning algorithm, it is a combination of multiple decision trees, it will generate large number of decision trees and use the vote system to vote for the most popular class, the generalization error of a forest will converge as we add more trees into it, also the error of a forest depends on the strength of individual trees in the forest and the correlation between them. According to the inventor Leo Breiman[8], Random Forest has some advantages such as better accuracy, relatively robust to outliers and noise, faster process time, gives useful internal estimates of error, strength, correlation and variable importance and it supports parallel computing. We expect the dataset would have a lot of noise because we don’t see any significant pattern in terms of our target variable.

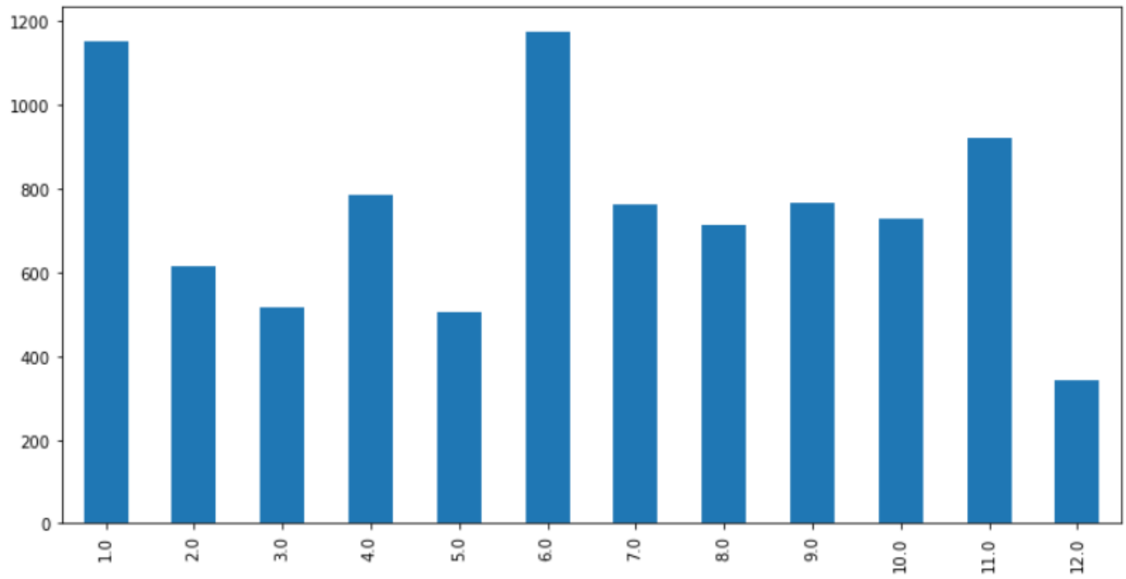


Figure 6: Counts for the monthly transactions in 2020, total number transactions for the year is 17429.

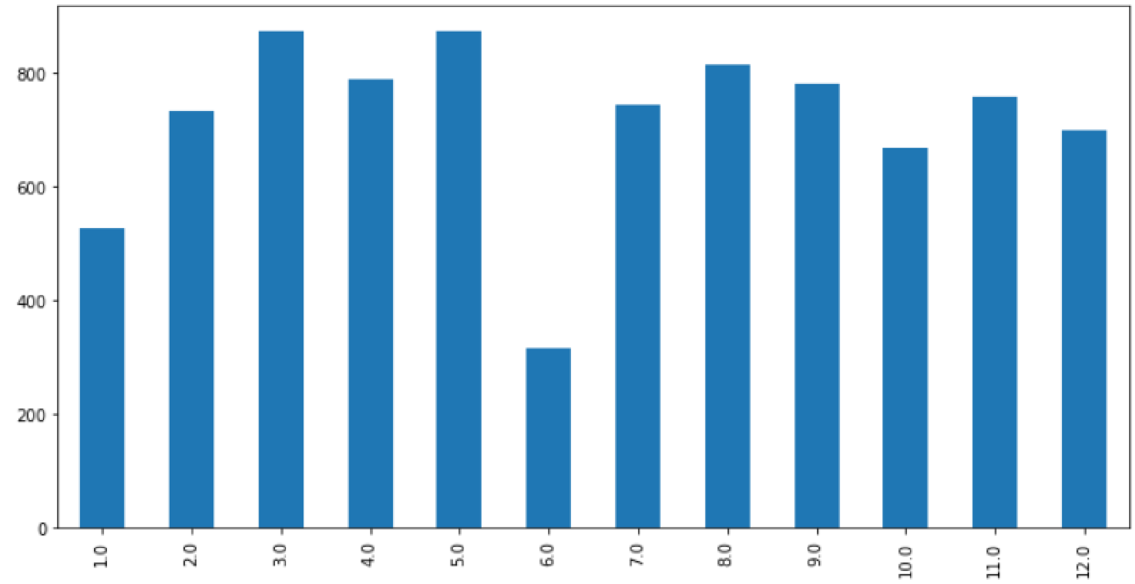


Figure 7: Counts for the monthly transactions in 2021, total number transactions for the year is 17715.

From figure 6 and 7 we can understand that 2021 had only 286 more total transactions than 2020. From monthly scale we see there are more months that the monthly transactions are more than 700 in 2021, but we have 2 months in 2020 that the monthly transactions are more than 1000. We don’t see any strong sign that can fit the previous assumption about transactions in 2021 should be more than 2020, but averagely, monthly transactions seem increased a bit from the middle of 2020. Additionally, we don’t see significant decrease in December 2021 due to

Covid-19 outbreak in Toowoomba.

Pre-Processing

For internal variables provided by Hastings Deering, we will use all of them except customer_ID because it is a unique identifier variable, itself does not contain any information. For external variables, we use variable selection to decide what variables we shoud fit into the model, a simple method to do it is a spearman correlation metrix. From below we see Number of building approvals are highly correlated with value of building jobs, we will get rid of value of building jobs because Number of building approvals are more intuitive to interprete with our assumption. For monthly number of transactions and Quantile of monthly transactions, here is just a sanity check that Q1 and Q4 should be highly correlated to original monthly number of transactions, because Quantile of monthly transactions are made by monthly number of transactions. Eventually, we will use Quantile of monthly transactions as our target variable, thus a classification problem to classify the four quantiles.

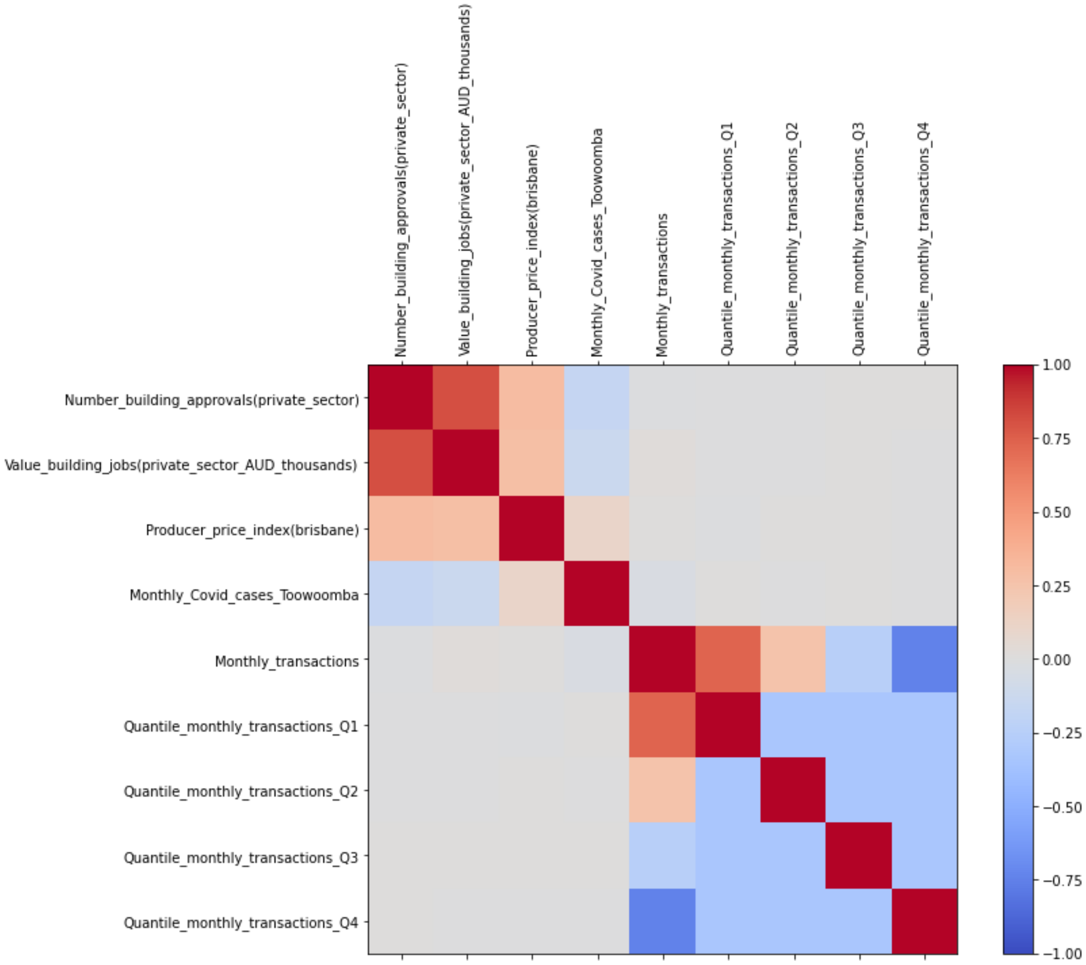
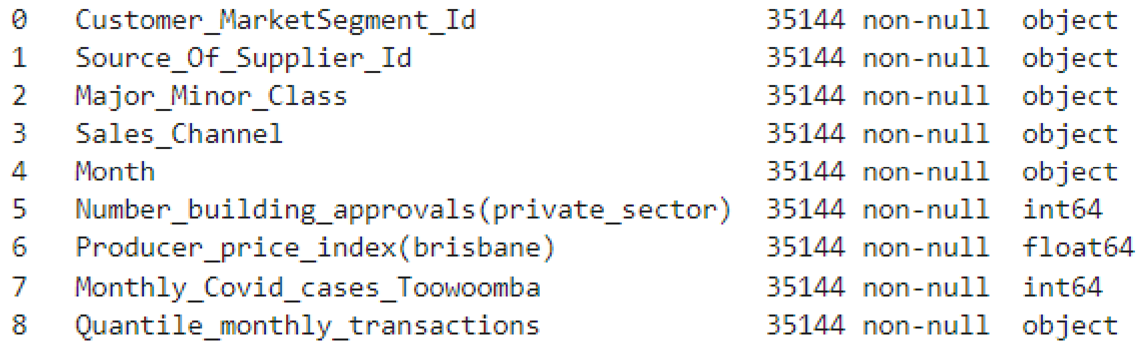


Figure 8: The spearman correlation metrix says we should make decision between Number_building_approvals and Value_building_jobs.

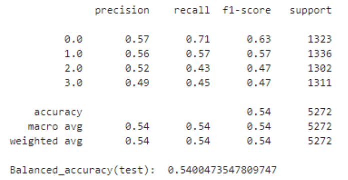


```
0   Customer_MarketSegment_Id                     35144 non-null  object
1   Source_Of_Supplier_Id                         35144 non-null  object
2   Major_Minor_Class                             35144 non-null  object
3   Sales_Channel                                 35144 non-null  object
4   Month                                         35144 non-null  object
5   Number_building_approvals(private_sector)     35144 non-null  int64
6   Producer_price_index(brisbane)                35144 non-null  float64
7   Monthly_Covid_cases_Toowoomba                 35144 non-null  int64
8   Quantile_monthly_transactions                 35144 non-null  object
```

Figure 9: A variable summary, these are the final variables we will fit the model, note that they are monthly data.

Next, we One-Hot-Encoding the categorical variables and Ordinal Encoding the target variables, this will give us a machine learning acceptable form of data, then we standardise the data because the three continuous variables have different scale, standardise the data can help reduce the noise and improve accuracy. Finally, we build a Random Forest classifier to classify the target variable – Quantile_monthly_transactions by using the other 8 variables. Random Forest classifier has several hyper-parameters, the most important three are number of trees(n_estimators), the depth of trees(max_depth) and class weight, the adjustment of hyper-parameters can affect the performance of a model. To optimise the model performance, we will use RandomizedSearchCV to find out the optimal hyper-parameters until the model converge. Eventually, we end up with n_estimators = 38, max_depth = 28 and class_weight="balanced". We will use Accuracy as the measurement to measure model performance.

Results



```
              precision    recall  f1-score   support

         0.0       0.57      0.71      0.63      1323
         1.0       0.56      0.57      0.57      1336
         2.0       0.52      0.43      0.47      1302
         3.0       0.49      0.45      0.47      1311

    accuracy                           0.54      5272
   macro avg       0.54      0.54      0.54      5272
weighted avg       0.54      0.54      0.54      5272

Balanced_accuracy(test):  0.5400473547809747
```

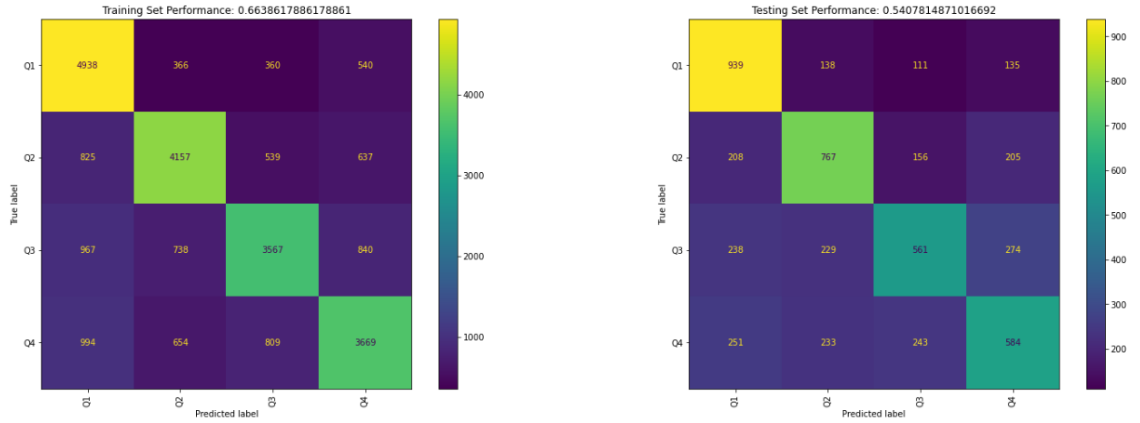


Figure 10: A classification report and confusion matrix for the model.

Figure 10 is the classification report and confusion matrix for the model, we end up with training accuracy of about 0.66 and test accuracy of about 0.54. Generally, training accuracy is a bit higher than testing accuracy, which means the model is not overfitting but a little bit bias. Precision is the percentage of truly positive out of all the positive predicted. Recall is the percentage of predicted positive out of the total positive. F1-score is the harmonic mean between precision and recall, it seems Q1 is easier to be classified by the model.

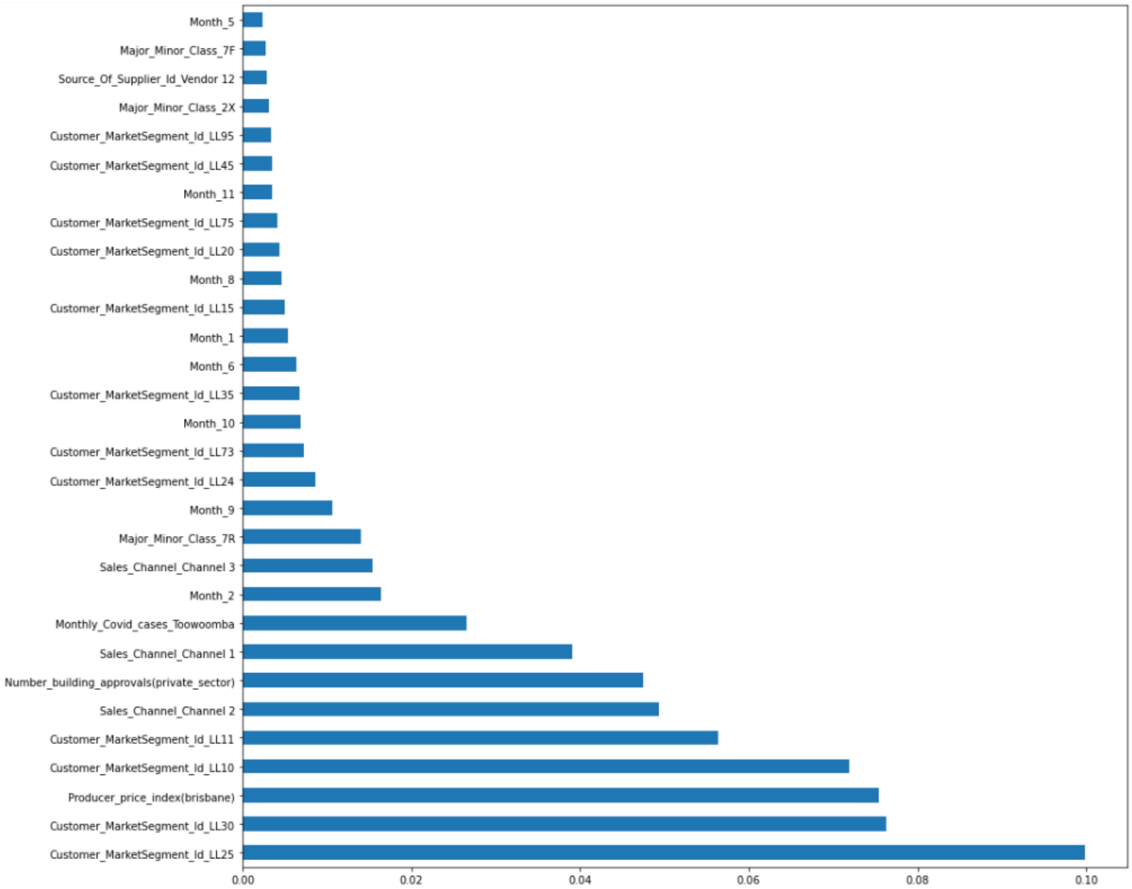
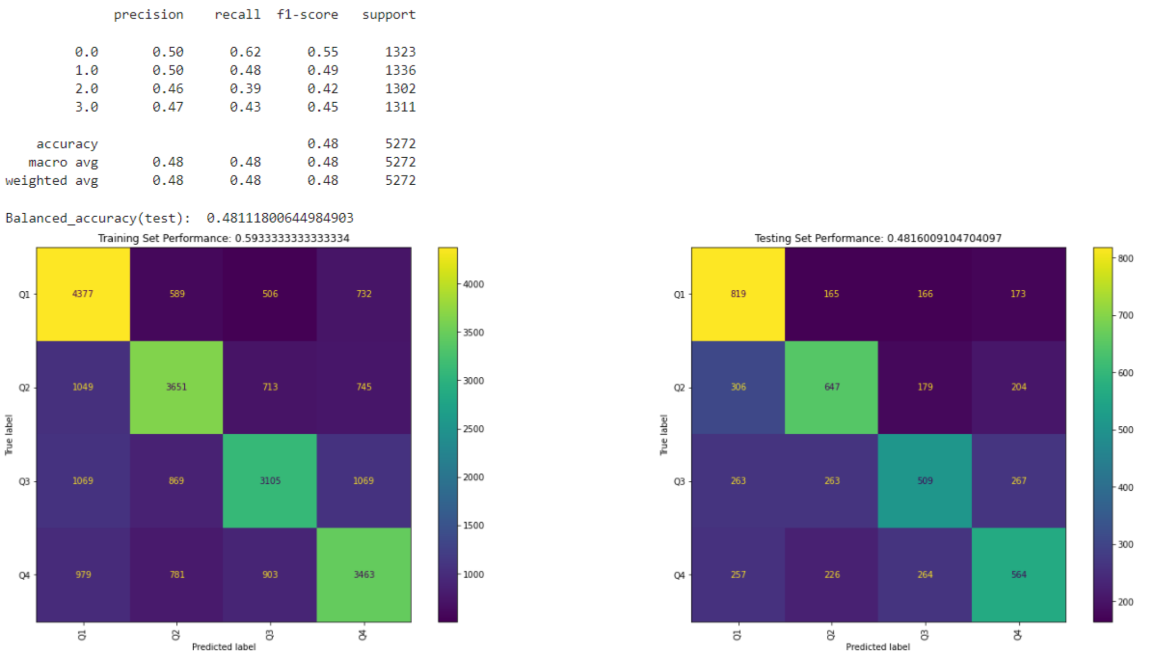


Figure 11: Permutation feature importance can tell us which variable was contributing to the prediction.

Permutation feature importance is a model inspection technique, it is defined as the decrease of the model score when a single feature value is not available, which means each time we make a feature unavailable, and see how it affect the model performance. From figure 11 we see Producer price index and number of building approvals seems can contribute to the prediction.

Discussion

Firstly, the model accuracy is not satisfactory, we only have 0.54 for testing set, means the prediction is not that reliable, ideally, we expect the accuracy can be over 0.6 before hyper-parameter optimisation. For model evaluation, by looking at figure 11 it seems our external variables can contribute and have impact on prediction. However, from previous content we don’t see any sign that can reflect to the assumption about the transactions, can external variables really affect the target variable? If we focus on the x-axis of the permutation importance we may find out, the most important feature we have is Customer_MarketSegment_id_LL25, which has the importance score of about 0.1, note that the feature importance values do not sum up to one, because they are not normalized. Now we understand that each variable in the model contribute very little to the prediction, we are actually ranking feature importance in a bunch of variables that contribute very little to the prediction, there are no variable that can dominantly affect the monthly number of transactions, even the most important variable. To verify this result, we simply train the model again without any external variables to see the model accuracy with original dataset, we use the same hyper-parameters.



```
              precision    recall  f1-score   support

         0.0       0.50      0.62      0.55      1323
         1.0       0.50      0.48      0.49      1336
         2.0       0.46      0.39      0.42      1302
         3.0       0.47      0.43      0.45      1311

    accuracy                           0.48      5272
   macro avg       0.48      0.48      0.48      5272
weighted avg       0.48      0.48      0.48      5272

Balanced_accuracy(test):  0.48111800644984903
```

Figure 12: A classification report and confusion matrix for the model without any external variables.

	Training accuracy	Testing Accuracy
With external variables	0.66	0.54
Without external variables	0.59	0.48

Table 1: A table for accuracy of both models.

From above we see without any external variables, we end up with training accuracy of about 0.59 and testing accuracy of about 0.48, which means with 3 external variables we can only increase the model accuracy by 0.06 for both training set and testing set. That is saying there is no significant evidence to determine our external variables can have impact on the target variable.

Our assumption about PPI and number of building approvals to the prediction of monthly number of transactions seems correct, but the result shows there are no relations between external variables and target variable. We can only say that in this dataset we don’t see any sign that can connect two pieces of puzzle together, it can be different if we also have data for 2022, because the inflation in 2022 is more serious than 2020 and 2021, the data of 2022 may show some variation that can correspond to the assumption we have. Also, according to the data we have, Covid outbreak in Toowoomba mainly starts from December 2021, which is the end month of our dataset, a 2022 data maybe can reflect how serious Covid affect the monthly number of transactions. Another thought is due to the lack of experience and knowledge about constructions, we don't know in what situation the contractors would do a repair or order new parts, if this activity is really depending on when the equipment becomes unavailable, then we may not be able to see any patterns in the data because the data is too randomly.

Reflection

This research is a perfect experience regarding practical real world data analysis, the research was vey limited but with lots of possibilities. Permutation feature importance can be misleading by just looking at the plot, we can truly understand which feature is important to the prediction if we also focus on the score the variable has. In this case it is unlucky that no variables can significantly contribute to the prediction. Also, our assumption seems perfect but the behavior in our assumption didn't occur in the dataset, this doesn't mean the assumption is wrong, it could mean we just don't see the pattern in this dataset. This research also reflects a question, what kind of data is good data? Actually, this dataset provided by Hastings Deering is a short version, there were a lot of information lost since Hastings Deering thinks some data are confidential to their company, they could also cut out the information we need in this research. Furthermore, this also reflects the limitation of data analysis, it's not always work, every stage in data analysis is evenly important, just like the old saying says, garbage in garbage out.

References

[1] "Producer Price Index Australia – Australian Bureau of Statistics". https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/producer-price-indexes-australia/latest-release (accessed Jun. 18, 2022)

[2] "Building approvals and inspections – Queensland Government". https://www.business.qld.gov.au/industries/building-property-development/building-construction/approvals-inspections (accessed Jun. 18, 2022)

[3] "Building approvals by SA2 and above - Australian Bureau of Statistics". https://explore.data.abs.gov.au/vis?fs[0]=ABS%20Topics%2C1%7CINDUSTRY%23INDUSTRY%23%7CBuilding%20and%20Construction%23BUILDING_CONSTRUCTION%23&pg=0&fc=ABS%20Topics&df[ds]=ABS_ABS_TOPICS&df[id]=BA_SA2&df[ag]=ABS&df[vs]=2.0.0&pd=2020-01%2C2021-12&dq=..TOT.TOT..317%2B1%2B2%2B3%2B4%2B5%2B6%2B7%2B8%2BAUS.M&ly[cl]=TI

ME_PERIOD&ly[rw]=REGION%2CREGION_TYPE&ly[rs]=MEASURE&lo=4 (accessed Jun. 18, 2022)

[4] “Queensland open data portal”. https://www.data.qld.gov.au/dataset/queensland-covid-19-case-line-list-location-source-of-infection/resource/1dbae506-d73c-4c19-b727-e8654b8be95a (accessed Jun. 18, 2022)

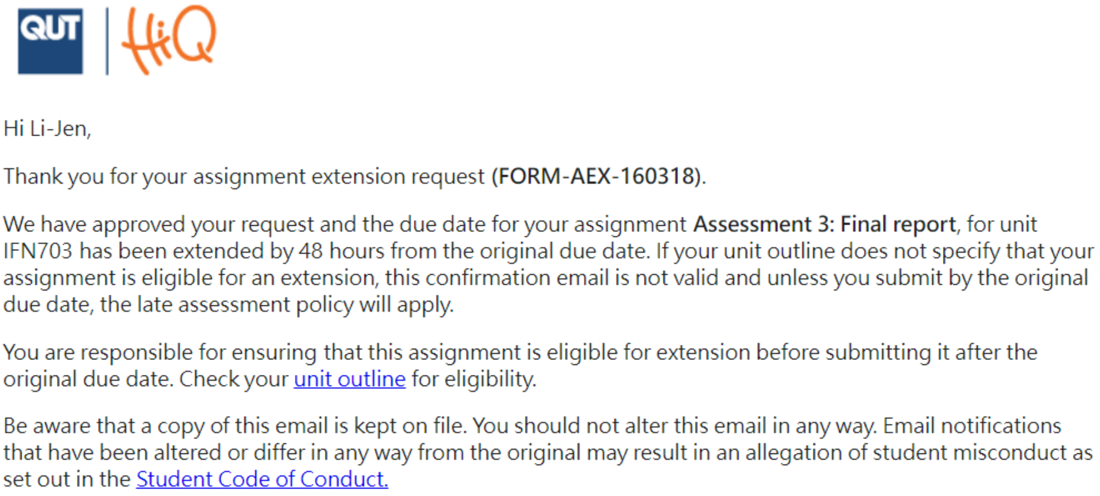
[5] “Toowoomba Property Market Update 1st Half of 2022 – PRD Toowoomba”. https://www.prd.com.au/toowoomba/research-hub/article/toowoomba-property-market-update-1st-half-2022/ (accessed Jun. 18, 2022)

[6] “AGC Construction Inflation Alert - AGC”. https://www.agc.org/learn/construction-data/agc-construction-inflation-alert (accessed Jun. 18, 2022)

[7] “Lumber futures historical data – Investing.com”. https://au.investing.com/commodities/lumber-historical-data (accessed Jun. 18, 2022)

[8] Leo Breiman, Random Forest, 2001

Appendix 1: Supplementary Information



Hi Li-Jen,

Thank you for your assignment extension request (**FORM-AEX-160318**).

We have approved your request and the due date for your assignment **Assessment 3: Final report**, for unit IFN703 has been extended by 48 hours from the original due date. If your unit outline does not specify that your assignment is eligible for an extension, this confirmation email is not valid and unless you submit by the original due date, the late assessment policy will apply.

You are responsible for ensuring that this assignment is eligible for extension before submitting it after the original due date. Check your unit outline for eligibility.

Be aware that a copy of this email is kept on file. You should not alter this email in any way. Email notifications that have been altered or differ in any way from the original may result in an allegation of student misconduct as set out in the Student Code of Conduct.