CDS 590:

Consultancy Project and Practicum

Consultancy Project Mid-Term Report:

Time Series Prediction for Predicting Real Private Consumption

Prepared by:

Muhammad Azzubair bin Azeman P-COM0019/19

Submitted for:

Dr. Zarul Fitri Zaaba

Date of Submission:

20 November 2020

School of Computer Science

Academic Session 2019 / 2020

CHAPTER 1

INTRODUCTION & RELATED WORKS

1.1 Background of Practicum Company

DataMicron Systems Sdn Bhd is a technology company which offers consultant services for business intelligence and big data-related solutions. As many companies nowadays have their own databases, they face challenges and difficulties in gaining insights from their big and complex data using traditional techniques. Therefore, this is where DataMicron comes in whereby their managing director said in his interview with *The Star* newspaper publication:

"*We extract data from various databases provided to us by our clients and merge them in the data warehouse, where from there, we do analysis of the data to provide our clients with business intelligence and predictive analysis, which in turn, would help them in their decision-making process*" (Hooi, 2014).

On 2004, DataMicron company was granted by Government of Malaysia through Malaysia Digital Economy Corporation (MDEC) with Malaysia Status Services (MSC) status which enables their company to enhance their product and service developments on multimedia technologies. As a result, the company has extended their scope of services to more than five countries as in 2014. The success of this company was reflected by their Microsoft Asia Pacific Keystone Award on 2005, and SME Corp Innovation Award (ICT) on 2013 (Hooi, 2014).

DataMicron provides innovative solutions for Big Data, and Business Intelligence for many local and international organisations. As data value is significantly increasing, DataMicron offers three types of data-related services which are Training, Consultancy, and Support. In terms of training, DataMicron together with other industry partners agreed to develop future talents by conducting one-year placement under their company for Bachelor students of Universiti Teknologi Malaysia under 2u2i mode programme (MDEC, 2019).

1.2 Background of Domain

Level of economic advancement of one country to another differs by the main macroeconomics indicator which is the Gross Domestic Product (GDP). As such, world countries are categorised into three categories which are Developed Economies, Economies in Transition, and Developing Economies (United Nations, 2020). In determining country classification, World Gross Product (derived from GDP) is included as one of the indicators.

Gross Domestic Product (GDP) is defined as “the market value of all final goods and services produced in an economy annually” (Hashim et. al, 2018). It is measured based on three main approaches which are the Production, Expenditure, and Income approaches (Department of Statistics Malaysia (DoSM), 2020). In terms of Production approach, it reflects on economic activities of individuals towards GDP as an overall; while for Expenditure approach, it determines the values of services and products consumed by consumers. As for income approach, it includes all income sources and amounts gained in economy. Therefore, in order to determine the economic values of each approach, macroeconomic indicators (also known as econometrics) are used as input to calculate the values of each approach.

Expenditure approach plays a crucial role in overall GDP as it contribute the most to the overall GDP since 2013 until 2018 (Asada et.al, 2019). This approach is dependent on five main macroeconomic indicators (econometrics) namely as Real Private Consumption (RPC), Real Government Consumption (RGC), Gross Fixed Capital Formation (GFCF), and Net Export (NE) (DoSM, 2020). Out of the four variables, RPC is the major contributor to Malaysia’s expenditure since 2018 (Ministry of Finance (MoF), 2019). Since RPC is the most significant attribute towards contribution to Malaysia’s GDP, it is very crucial for RPC to be predicted as it were the reference for Malaysia’s MoF in making decisions for future financial planning.

In brief, RPC is defined as the amount of goods and services consumed by households to fulfill their basic needs and wants (DoSM, 2020). It reflects on the expenditure of people in Malaysia as an overall. Most commonly, RPC is predicted using statistical techniques such as vector autoregression and ARIMA models (Razak, Khamis & Abdullah, 2017).

RPC prediction using machine learning has been a major topic discussed in many literatures in the last two decades (Taieb, 2014). Several machine learning models such as Neural Network, Support Vector Machine, and K-Nearest Neighbour were proposed and discussed. However, machine learning techniques were foreign among Malaysians until it was first recommended publicly by the Minister of International Trade and Industry (MITI) (Bernama, 2018). As a result, machine mearning models and techniques are gradually being learned by Malaysians in many online courses recently (Fadzil, Latif, & Munira, 2015).

1.2 Problem Statement

Much studies have been done on comparing between performances of machine learning models with statistical models (Makridakis, Spiliotis, & Assimakopoulos, 2018) in general and especially in economics (Yu, 1999; Dematos et. al, 1996; Kumar, 2018). However, there is no specific study have been done in comparing model performances between machine learning models with statistical models for time series prediction of real private consumption in Malaysia. Addressing the performances between statistical with machine learning were have practical benefits for researchers in economics and contribute in understanding of both approaches for time series prediction of real private consumption specifically and macroeconomics generally.

1.3 Research Question

This research proposal makes an attempt to predict time series of real private consumption. This were consequently results in determining whether machine learning models or statistical models is better for RPC prediction. This project proposes machine learning approaches for RPC prediction. This brings to the following research questions:

- Which machine learning model is the most suitable for RPC prediction?
- Which approach is better for time series prediction of real private consumption?
- What is the RPC prediction of the best machine learning model for second quarter of 2020?

Throughout this project, the answers of research question answers were retrieved from literature reviews and discussion sections of this project report.

1.4 Objectives

Therefore, the aim of this project is to propose machine learning techniques as a new approach to improve econometric forecasting in Malaysia. This project will be focusing on forecasting one of the econometric indicators which is RPC published quarterly by DoSM. Typically, to achieve the aim of this project, there are two objectives listed below which are:

1. To investigate the suitable machine learning technique for RPC prediction model.
2. To develop RPC prediction model using the selected machine learning approach.
3. To evaluate the RPC prediction model.

1.5 Benefits of the project

This problem is actually a consultation project between DataMicron with Malaysia's Ministry of Finance (MoF). Therefore, this project will benefit DataMicron in providing proposed solution for their client's problem. In particular, this project will propose new methodologies of RPC prediction using machine learning approaches and were develop reliable models for suggesting RPC predictions as a reference for MoF in making effective decisions. In return, MoF will use these insights to optimise their financial planning and reduce financial loss.

1.6 Related Works

In this section, literatures were referred to explore the details of established and proposed methods in RPC predictions using machine learning techniques. Published researches demonstrating on analysis behind time series of economic predictions were reviewed.

1.6.1 Review on Domain

Since 2018, RPC contributed the most to Malaysia's GDP followed by I and RGC (MoF, 2019). However, in terms of annual RPC growth, RPC has shown a fluctuating trend between 6.0% – 8.0% (BNM, 2019; BNM, 2018). This trend is contributed mostly by the growth of employments and wages, thus this shows that affecting this sector resulted much to the overall annual RPC. Other than that, RPC is also being contributed by other variables such as imports of consumption goods, narrow money, and loans disbursed by banks (BNM, 2016).

1.6.2 Review on Data Science & Analytics techniques

Statistical techniques such as ARIMA, and VAR were reported and compared for Malaysia's economics forecasting. (Razak, Khamis & Abdullah, 2017). Both of these models can be used for univariate time series (UTS) forecasting. This means both models can be used to forecast a variable for several periods ahead in future. The findings of this study found out that VAR is more accurate than ARIMA due VAR's less mean absolute percentage error (MAPE) compared to ARIMA (Razak, Khamis & Abdullah, 2017). In addition, they also highlighted that VAR outweighed ARIMA by having multivariate time series (MTS) forecasting which enables for more dynamic forecasting using multiple variables to forecast stock market index.

In other parts of the world, machine learning techniques were began to be proposed for economics forecasting since 20th century. The earliest attempt was done by Yu (1999) whereby she compared model performance between ARIMA and Backpropagation Neural Network (BNN) in forecasting stock index. The outcome of this study is BNN produced lower MAPE and Root Mean Squared Error (RMSE) compared to ARIMA. This reflects that nonlinear trend of stock index is better to be forecasted using machine learning models than linear models such as ARIMA. Similar observation was obtained by Dematos et. al (1996) in which they found that Recurrent Neural Network (RNN) and Feedforward Neural Network (FNN) outperformed ARIMA in forecasting Japanese yen / U.S. dollar exchange rate. This summarised that nonlinear trend of economic indicators is more accurate to be forecasted using machine learning.

Two years back, a comparative study was done by Kumar et. al (2018) in comparing between machine learning performances in predicting stock market trend. Machine learning used were support vector regression (SVR), random forest (RF), k-nearest neighbour (KNN), naive bayes (NB), and softmax (SM). Interestingly, they found out that when large dataset (4500 entries) were input, RF outperformed other models by having the highest accuracy and f-measure followed by SVR. This indicates that RF and SVR are suitable models to be used for predicting nonlinear trends of economic growths. In addition, RF is also good to used for prediction due to its robustness against outliers because of its bagging principle in learning the training set and predicting the test set (Roy & Larocque, 2015). Meanwhile for SVR, it is known to be highly effective and efficient in forecasting values of stock prices (Vo et al., 2016).

Another study done by Chen et al. (2019) also compared between machine learning performances in predicting personal consumption expenditure services (PCE services) which is based on Quarterly Service Survey (QSS). The focus of the study was to compare time series prediction performances between linear models with nonparametric models. Linear models included in the study were 4-Quarter Moving Average (4QMA), Forward Stepwise Regression (FSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Ridge Regression (RR). Meanwhile, nonparametric models implimented in the study were Regression Trees (CART), Random Forests (RF), Gradient Boosting (GB), Multi-Adaptive Regression Splines (MARS), and Support Vector Regression (SVR) with Radial Basis Function (RBF). The outcome of the study found out tree-based ensemble models which are RF and GB are the best models as the had the least RMSE percentage points of -0.56 and -0.43 respectively. In contrast, MARS and 4QMA were suggested to be avoided they had significantly poor prediction performance than others.

Recently, Maehashi and Shintani (2020) also performed comparison study between machine learning models in predicting macroeconomic variables. Interestingly, their comparison study is enriched with factor model (economics model) and multiple machine learning model approaches such as neural networks, regularised least square methods, as well as ensemble learning methods. In particular, they employed Feedforward Neural Network (FFNN), Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM) for neural network models. Meanwhile, for regularised least square method, they included Lasso, Ridge, and Elastic Net regressions which all of them are categorised as linear models. As for ensemble learning methods, Maehashi and Shintani (2020) incorporated bagging, Random Forests, and boosting models in which all of them are the ensemble methods of regression trees. The findings of the study concluded that ensemble learning methods outperformed other models due to their adaptability with nonlinear trends of the macroeconomics variables.

Another important findings highlighted by Maehashi and Shintani (2020) is machine learning models performed excellently when the window size (also known as forecast horizon) is large. This means the more timely data were included for model training, their predictions would become better, more accurate, and robust againts errors.

1.6.3 Review on Data Science & Analytics tools

Analytical tools for time series prediction and forecasting are abundant nowadays. They are available online either as a free software or as an advance premium software. An example of a free software is the Python, a well-integrated, and popular data science tool among data scientists. Multiple studies have been using Python for time series prediction and forecasting because there are many available libraries that are created for the purpose of dealing with time series data and forecasts. One of the popular libraries for time series forecasting is the statsmodels library. According to McKinney et al. (2011), this library provide many statistical models made available to python users such as ordinary least squares, VAR, and ARIMA. Another library for time series forecasting is the Facebook's Prophet as proposed by Usher and Dondio (2020) for short term forecast on pound sterling with respect to euro and dollar currency. They forecasted the pound sterling would rise against dollar and euro by ±0.02 by end of 2019.

Another approach of time series forecasting is by using machine learning approach. This approach recognises time series data as a supervised learning through the use of sliding window for model training and testing. Brownlee (2017) explained about the sliding window in detail whereby he described that time series dataset can be restructured into a supervised learning dataset by using the value of previous data to predict for future data. In short, historical data are taken as input and future data is treated as the output. In python, this windowing feature is available is several libraries such tslearn, cesium-ml, ts-fresh, and seglearn. Burns and Whyne (2018) compared these libraries features for time series forecasting. Their comparison shows that all of the libraries has forecasting feature except tslearn. On top of that, they found out that seglearn library is the only library that has the most features such as multivariate time series, sliding window, and compatible with machine learning models.

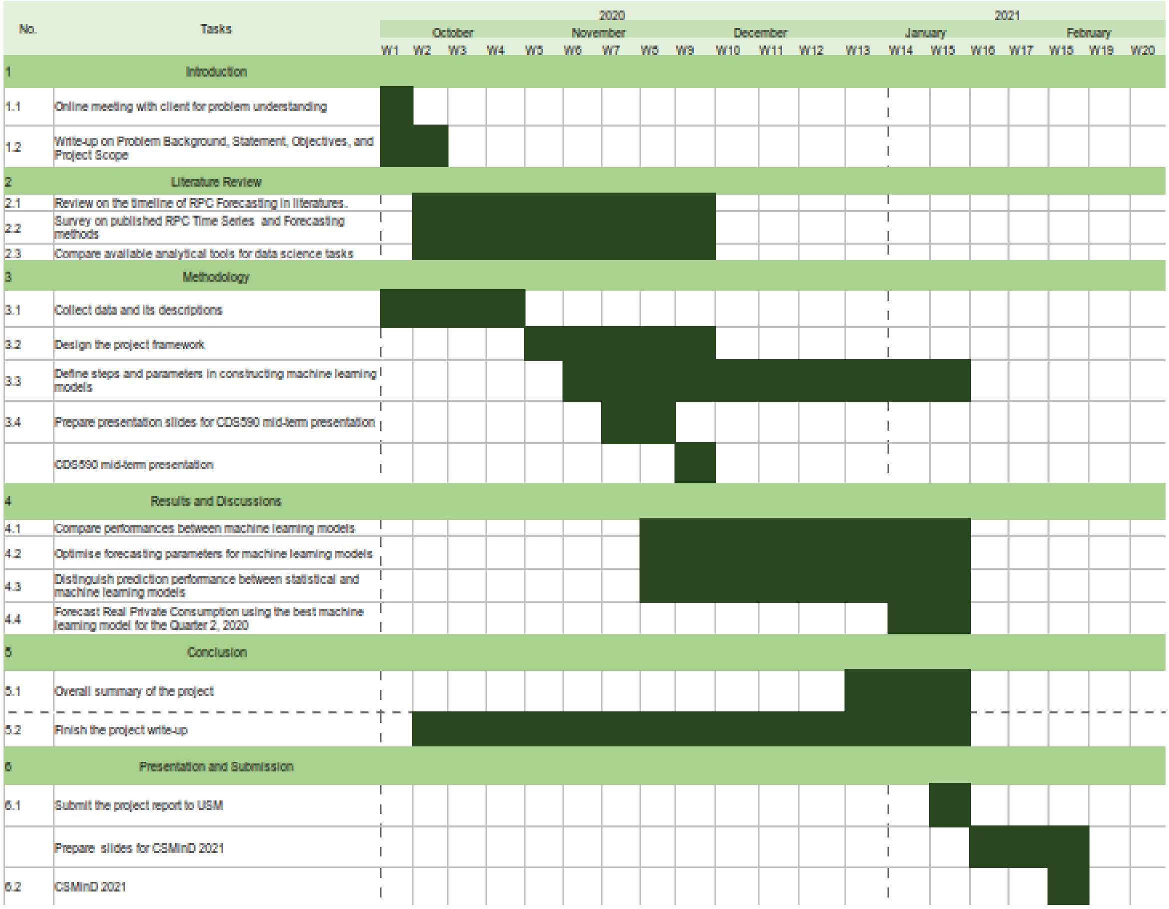
Another important library for tasks using machine learning models is the scikit learn library. This library contains most of the common machine learning models for data scientist. According to Hackeling (2017), scikit learn library provided linear models such as linear and multiple linear regressions, non linear models such as decision trees and random forests, and perceptron derived models such as support vector machines and artificial neural networks. Hence, combination of scikit learn and seglearn libraries are sufficient for time series forecasting.

CHAPTER 2

RESEARCH METHODOLOGY

2.1 Activities plan and Gantt Chart

Throughout semester 1 2020 / 2021, the activities for project consultancy and practicum were conducted based on the plans as illustrated in figure 1. This project is done individually with the supervision of project supervisor and guided by a mentor from the practicum company. With the limitations of the current COVID-19 situation, all of the process for the project consultancy and practicum were conducted online via emails, phone calls, and conference calls. 96 contact hours with the practicum company were recorded in a logbook and weekly meetings with supervisor were also be recorded to update about project progress.



No.	Tasks	2020												2021							
		October				November				December				January				February			
		W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
1	Introduction																				
1.1	Online meeting with client for problem understanding	■																			
1.2	Write-up on Problem Background, Statement, Objectives, and Project Scope	■	■																		
2	Literature Review																				
2.1	Review on the timeline of RPC Forecasting in literatures.		■	■	■	■	■	■	■	■											
2.2	Survey on published RPC Time Series and Forecasting methods		■	■	■	■	■	■	■	■											
2.3	Compare available analytical tools for data science tasks		■	■	■	■	■	■	■	■											
3	Methodology																				
3.1	Collect data and its descriptions	■	■	■	■																
3.2	Design the project framework					■	■	■	■	■											
3.3	Define steps and parameters in constructing machine learning models						■	■	■	■	■	■	■	■	■	■					
3.4	Prepare presentation slides for CDS590 mid-term presentation							■	■												
	CDS590 mid-term presentation									■											
4	Results and Discussions																				
4.1	Compare performances between machine learning models								■	■	■	■	■	■	■	■					
4.2	Optimise forecasting parameters for machine learning models								■	■	■	■	■	■	■	■					
4.3	Distinguish prediction performance between statistical and machine learning models								■	■	■	■	■	■	■	■					
4.4	Forecast Real Private Consumption using the best machine learning model for the Quarter 2, 2020														■	■					
5	Conclusion																				
5.1	Overall summary of the project													■	■	■					
5.2	Finish the project write-up		■	■	■	■	■	■	■	■	■	■	■	■	■	■					
6	Presentation and Submission																				
6.1	Submit the project report to USM															■					
	Prepare slides for CSMinD 2021																■	■	■		
6.2	CSMinD 2021																		■		

Figure 1 Gantt Chart of Project Consultancy and Practicum

2.2 Data Science Project Lifecycle

As a data scientist consultant, it is important to impliment the fundamentals of data science project lifecycle in daily life. The reason is to structure the process of data science projects so that these projects provide beneficial insights for clients effectively and the outcomes of these project are deliverable on time. According to Microsoft (2020), there are five main stages of data science project lifecycle which are Business Understanding, Data Acquisition and Understanding, Modelling, Deployment, and Customer Acceptance. Figure 2 illustrates the Data Science Lifecycle stages. Throughout practicum, all of these stages were went through.

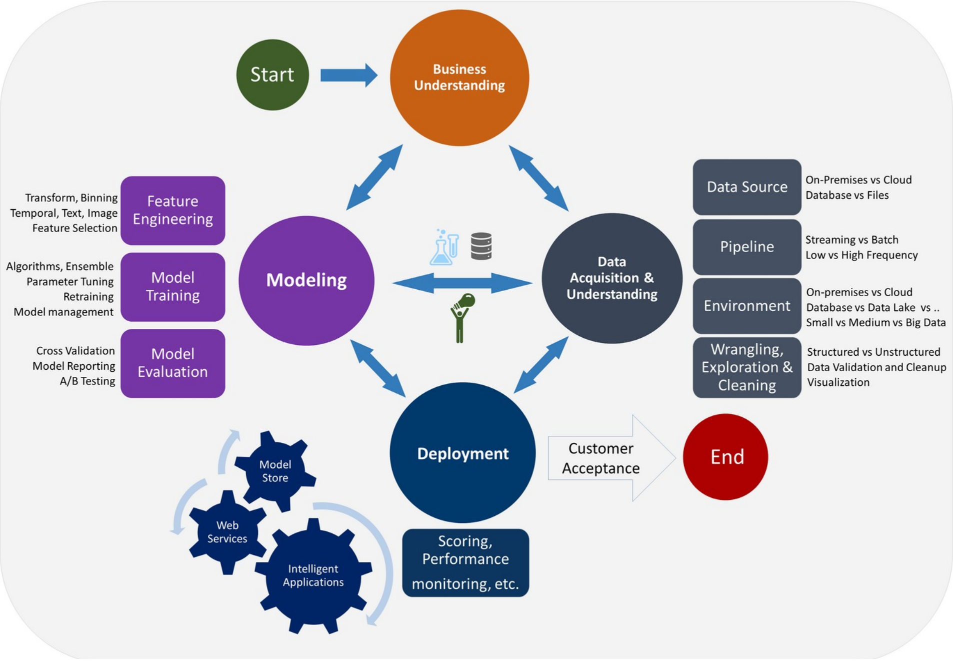


Figure 2 Lifecycle of Data Science Projects

In short, Business Understanding stage were conducted in a consultation meeting whereby MoF were explain the background of their RPC problem, analyse the problem together in the form of research questions, and describe their expected solutions. In order to determine the success of the proposed solution, it were measured, and ensured to be within clients' expectations using success metrics that are specific, measurable, achievable, relevant, and time-bound (Microsoft, 2020). Next stages were conducted individually and lastly the proposed solution were presented to MoF during the Customer Acceptance stage.

2.3 Problem Analysis

Analysis of client's problem were conducted during Business Understanding stage. At this stage, a list of formulated questions were prepared and asked to the client. The purpose of these questions are mainly to determine the problem framing, identify target and predictor variables, and pinpoint for data source. Before coming into the main questions, general questions related to domain background were prepared. The listed questions are as follow. After the questions are answered, exploratory data analysis and final analysis were done.

2.3.1 Initial Questions

Questions listed below were prepared to the client for getting to know of the domain background, methods currently being used for RPC forecasting, and solution expectations. The answers of these questions were used as a reference throughout this consultancy project.

- What is Real Private Consumption?
- What are the variables considered in forecasting RPC?
- How are the variables collected before going into data analytics?
- Why is it important to forecast Real Private Consumption?
- Among all of the mentioned variables, which variable would be the target variable?
- What are the current methods of forecasting RPC?
- How good are those methods in modelling and forecasting RPC?
- What are the limitations of current methods in forecasting RPC?
- What type machine learning models would be expected for RPC forecasting?
- What is the expected model performance metrics to be implimented?
- What is the threshold of acceptance for the RPC forecast values?

2.3.2 Specific use case to be addressed

According to the World Bank Group (2020), Malaysia's annual private consumption is projected to be declining from 1.2% in 2019 into -4.9% in 2020 due to the recent COVID-19 pandemic. Although Malaysia government had already provide financial support to its citizens through *Prihatin Rakyat* and *Penjana* packages, real private consumption willstill be affected due to social restrictions which reduced the household demands of purchasing wants carefreely.

2.3.3 Exploratory Data Analysis

Upon receiving dataset from MoF, an exploratory data analysis were done to overview the variable distributions, insights, and trends. In terms of Data Science Lifecycle, this step is categorised under Data Acquisition and Understanding. Firstly, the dataset were given in an excel file containing variables of real private consumption indicators as published in BNM (2016). These variables is collected from data published by DoSM, and BNM from 1995 until 2020. All of the numerical variables were cleaned, and transformed into quarterly data for them to be aligned with quarterly RPC. Table 1 shows the variables types and descriptions in detail.

Table 1 Description of Real Private Consumption and its indicators

No.	Variable	Descriptions
1.	Private Consumption Index (PCI)	Measures consumer spending on goods and services in RM millions
2.	Imports of Consumption Goods (ICG)	Import of any tangible commodity produced and purchased by consumers in RM million amount
3.	Sales of Passenger Cars (SOPC)	Amount of sold cars manufactured by local Malaysian brands in ‘000 units
4.	Loans disbursed for Consumption Credit (LCC)	Amount of RM millions lended by banks for loans in Consumption Credits
5.	Loans disbursed to Wholesale & Retail Trade, Restaurant, & Hotels (LOWT)	Amount of RM billions lended by banks for loans in consumers’ Consumption Credits
6.	Sales of Motorcycle (SOM)	Amount of sold motorcycles manufactured by local brands in ‘000 units
7.	Credit Card Turnover Spending (CCS)	Total amount of credits spent in RM millions amount
8.	MIER Consumer Sentiment Index (MIER)	Measure consumer confidence on Malaysia’s economy status
9.	Narrow Money (NM)	Aggregate amount of monetary assets available in Malaysia in RM millions
10.	FBM KLCI	Capitalised-weighted stock market index comprised of 30 largest companies on Bursa Malaysia

After describing each attributes, python were used to be visualise the trends of each attributes. Figure 3 below shows the overall visualisation of RPC and its indicators.

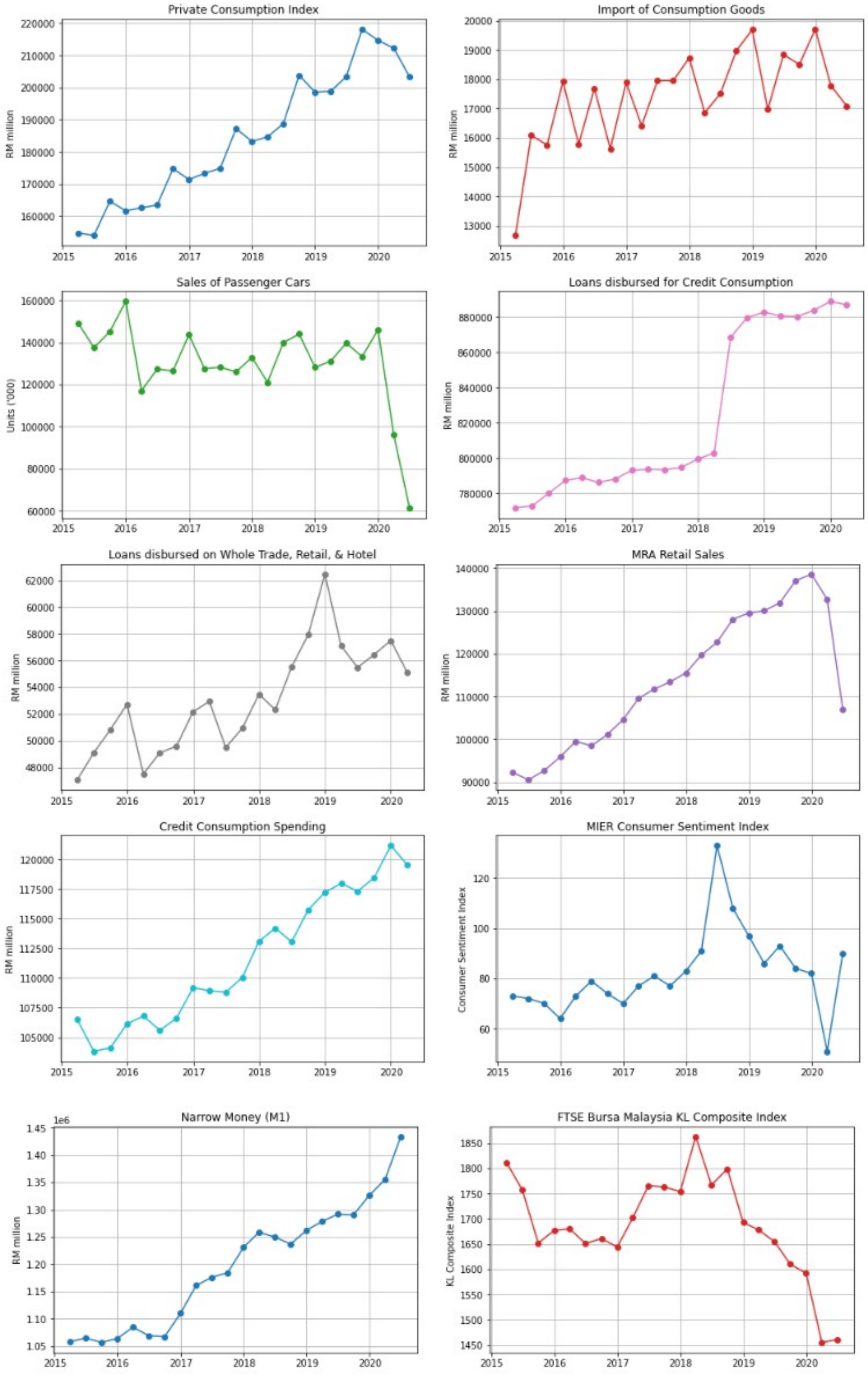


Figure 3 Overview of Real Private Consumption trends and its indicators

Referring to Figure 3, prior to the COVID-19 pandemic, PCI shows an increasing trend with seasonal patterns from 2015 until 2019. Meanwhile, ICG, LOWT, LCC, MRS, CCS, and NM also show similar trends. This shows that most of RPC indicators shows an improvement of RPC growth throghout the years. In addition, MIER also shows an increasing trend but only until second quarter of 2018, beyond than that, MIER declined. Because of Malaysia's General Election held during the second quarter of 2018, consumer sentiments went skyrocketed until the end of the second quarter of 2018. Beyond that quarter, pessimists began outnumbered optimists due to the global challenges affecting Malaysia's economic growth (Rasid, 2019). In contrast, FBM displays an overall decreasing trend with some exceptions in 2018. This attribute is heavily affected by Malaysia's political issues in which investors did not want to take risk in investment while Malaysia is having political turbulence (Afandi & Khoo, 2020).

Upon the COVID-19 pandemic emergence in Malaysia, all RPC indicators show a significance decrease in the first quarter of 2020 in comparison with fourth quarter of 2019 excluding the narrow money attribute. This reflects that external virus has inflicted severely on Malaysia's RPC especially when Malaysia enforced the Movement Control Order starting on March 2020. In second quarter of 2020, some of the RPC indicators such as the MRS, NM, and FBM show a rebound trend whereby the values are slightly improving during Malaysia's Recovery Movement Control Order (RMCO). However, still most of the remaining variables such as ICG, SOPC, and MRS are having declining trends which resulted in the overall downward trend of RPC (PCI) in the second quarter of 2020. In contrast, NM shows an increasing trend throughout the years which means money supply for Malaysia is not affected by the pandemic. This indicates that more money are being supplied in the economy over time.

2.3.4 Final Analysis

Forecasting RPC using machine learning techniques were conducted according the proposed method by Kumar et al. (2018) with some modifications. This method consists mainly of six steps: data cleansing, data preparation, model training, model optimisation, model evaluation, and data forecasting. In particular, windowing implimentation were conducted as stated by Rasel et al. (2015). Flow chart for the proposed methodology is illustrated in Figure 4. The flow of the final analysis and discussions were arranged based on this flow chart.

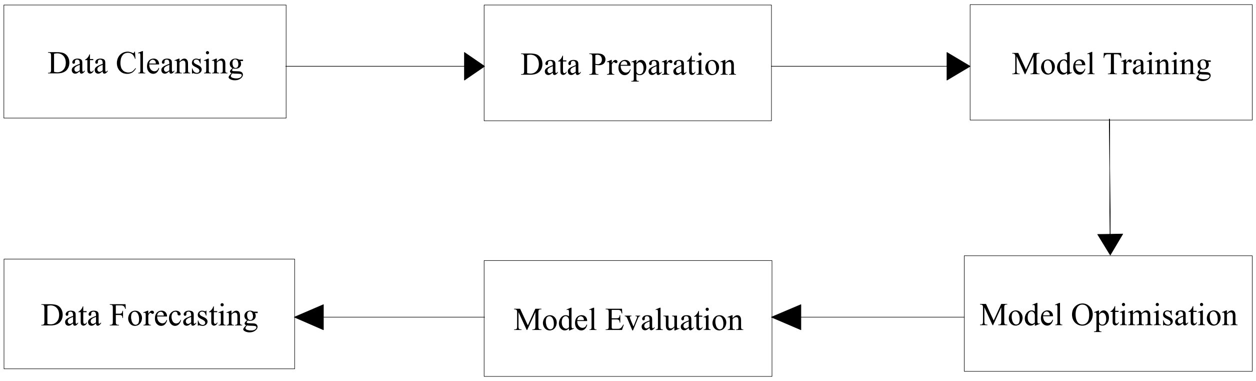
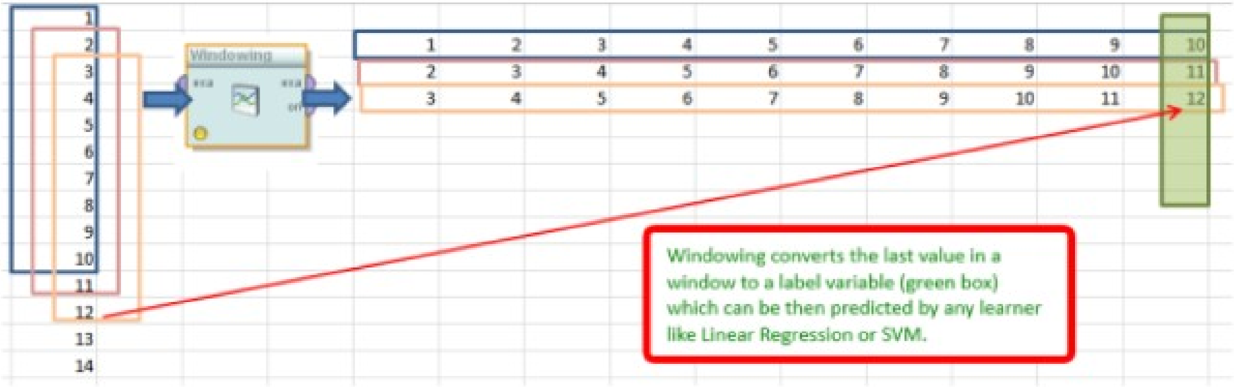


Figure 4 Flow Chart of RPC Forecasting

The first step to be done after obtaining the dataset were **data cleansing** by filtering the timeline of dataset until the year at which all attributes are begin to be collected simultaneously. After data cleansing were be **data preparation** stage. At this stage, initially there are no value for RPC indicators for second quarter of 2020 since they are future values that were not published yet. In order to fill up those attributes, they were predicted using windowing method integrated with bagging, RF, and boosting respectively. According to Rasel et al. (2015), these windows were generated by transposing the column of RPC indicators into horizontal windows in which the last row were become the target value to be predicted. This sliding window were move horizontally from the beginning of the time series until the end through a time-based cross-validation. Figure 5 illustrates the concept of transposing a column into horizontal windows and Figure 6 displays the concept of sliding window during time-based cross validation.



Windowing converts the last value in a window to a label variable (green box) which can be then predicted by any learner like Linear Regression or SVM.

Figure 5 Generation of horizontal windows via column transpose

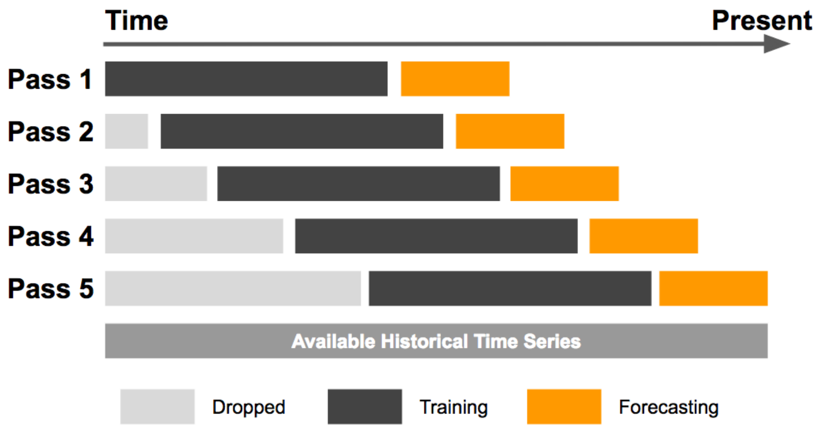


Figure 6 Illustration of sliding window concept for model training, and prediction

After the data has been prepared, the windowed data were input into bagging, RF, and boosting models for **model training** stage. Throughout the training, the models were evaluated using RMSE evaluation metrics similar to the current evaluation metrics currently being used by MoF. Model training and testing were repetitively conducted together with model optimization to achieve the best model performance. During **model optimization** stage, two components were tested to come out with the best model performance which are feature selection, and window size. The best attributes to be used in model training were determined using feature selection available in python's scikit learn library while the best window size were identified using loop function. Before forecasting, model performances were evaluated in terms of their evaluation metrics during the **model evaluation** stage. Lastly, future data of RPC for second quarter of 2020 were forecasted using the best model evaluated. This forecasted value will be presented to client during Customer Acceptance stage of Data Science Project Lifecycle.

2.4 Proposed Solution

After the models passed the evaluation stage, the tree-based ensemble models will be proposed to the MoF as a new method for future RPC prediction. Along with these, the predicted value of Malaysia's RPC of second quarter of 2020 will also be presented. This predicted value will be evaluated by the client whether it is acceptable or not. Plus, the predicted value also will be compared with the actual RPC value after it is officially published by DoSM.

2.5 Justification on Selected Data Science Techniques and Tools

Based on a comparative study done by Maehashi and Shintani (2020), they found out that tree-based ensembled learning such as Bagging, Random Forest, and Boosting performed the best for predicting macroeconomic indicators. Therefore, since the target variable of this project is the Real Private Consumption which is also an economic indicator of GDP, thus tree-based ensemble learning were selected as they were the most accurate in predicting macroeconomic indicators (Maehashi and Shintani, 2020). Hence, it is justified that Bagging, Boosting, and Random Forest were selected due to their high accuracy for economics predictions.

Despite of having abundant statistical tools online for RPC prediction, python language is still the best among data scientist due to its dynamic and flexible usage. According to Burns and Whyne (2018), python has multiple open source libraries for time series prediction using windowing techniques such as tslearn, cesium-ml, ts-fresh, and seglearn. However, they also found out that only the seglearn library is compatible to be used with machine learning models from the scikit learn library, plus with windowing feature. Other than seglearn and scikit learn libraries, other basic libraries such as “Pandas”, “Matplotlib”, and “Numpy”, were included in this project for data cleansing, data preparation, and visualisations. Hence, it is justified that python were selected in this project due its dynamic usage throughout this project.

2.6 Contribution

In this project, we show that the key to modeling RPC relies on the process as illustrated in Figure 4 which consists of data cleansing until data forecasting. Throughout this flow chart process, we analyze, predict, evaluate and explain on machine learning models in RPC predictions on the time series data. The principal contributions of this project are:

- We investigated the most suitable machine learning methods for RPC predictions by reviewing literatures related to RPC and economics modelling.
- We developed RPC prediction models using the tree-based ensemble models.
- We evaluated the RPC prediction models among machine learning models and with the statistical methods currently being used by client.

CHAPTER 3

RESULTS AND DISCUSSIONS

3.1 Investigation on Machine Learning techniques

Machine learning performance comparison for macroeconomics indicator were numerously discussed in related works (Chapter 1). Literature reviews were done by discussing on the machine learning model comparisons discussed by researchers. Table 2 summarises the literature reviews findings. Throughout the literature reviews, it was found out that ensemble learning models based on regression trees were the best models to be used for RPC prediction. The reason is because these models are able to learn and predict better on the non-linearity of RPC time series trend compared to other models. Hence, bagging, RF, and boosting algorithms were selected to be implimented in this project for RPC prediction.

Table 2 Summary of Literature Reviews

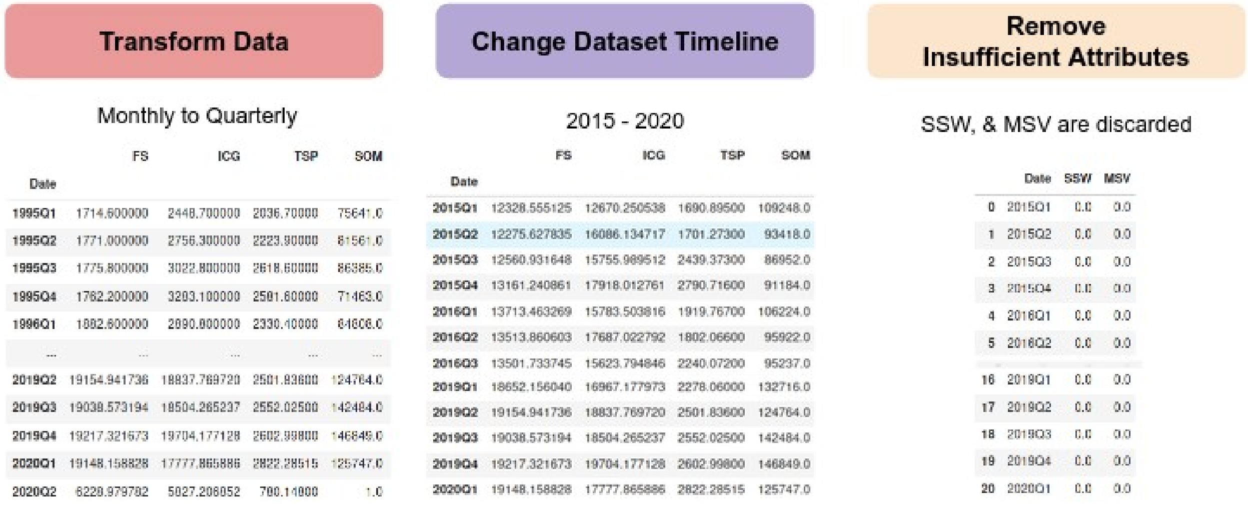
<table>
<tr><th>Author</th><th colspan="4">Algorithm</th><th>Findings</th></tr>
<tr><td rowspan="2">Kumar et. al (2018)</td><td colspan="2">Parametric</td><td colspan="2">Nonparametric</td><td rowspan="2">• When dataset is large, RF outperformed others
• When dataset is small, NB performed the best</td></tr>
<tr><td colspan="2">• NB
• KNN
• Softmax</td><td colspan="2">• RF
• SVR</td></tr>
<tr><td rowspan="2">Chen et al. (2019)</td><td colspan="2">Parametric</td><td colspan="2">Non parametric</td><td rowspan="2">• Tree-based ensemble models such as RF, and GB were the most accurate models.
• Due to underfitting of MARS and 4QMA models, they had poor prediction performance</td></tr>
<tr><td colspan="2">• 4QMA
• LASSO
• Ridge</td><td colspan="2">• CART
• RF
• XGBoost
• SVR
• MARS</td></tr>
<tr><td rowspan="2">Maehashi and Shintani (2020)</td><td>Linear</td><td colspan="2">Ensemble</td><td>Neural Network</td><td rowspan="2">• Tree-based ensemble models were the majority of the best models.
• Large window size is recommended for better time series predictions.</td></tr>
<tr><td>• LASSO
• Ridge
• EN</td><td colspan="2">• Bagging
• RF
• AdaBoost</td><td>• FFNN
• CNN
• LSTM</td></tr>
</table>

3.2 Development of selected Machine Learning models

Machine learning model development were constructed as illustrated in flowchart in Figure 4. It is consist of of six phases which are data cleaning, data preparation, model training, model optimisation, model evaluation, and finally data forecasting. These phases are iterated until the best model evaluation is achieved which also would result the best prediction models.

3.2.1 Data Cleansing

Upon receiving dataset (excel file) from Ministry of Finance's Fiscal and Economics Division, the data was explored and cleaned using python's pandas library. Firstly, the timestamp of RPC indicators were not aligned with RPC which are collected monthly and quarterly respectively. Therefore, the indicators were aggregated by sum and average depending on attribute type accordingly. Next, columns with zero values throughout the dataset were discarded which reduced the overall attributes from 18 into 16 attributes. Thirdly, it was noticed during data exploration that the timestamp at which each attribute were begin to be collected are different. Therefore, dataset timestamp were standardized by filtering it to be starting from date at which all attributes were collected simulataneously which is 1 January 2015. Figure 7 shows the output of each data cleansing process done in python's pandas library.



Transform Data

Monthly to Quarterly

	FS	ICG	TSP	SOM
Date				
1995Q1	1714.600000	2448.700000	2036.70000	75641.0
1995Q2	1771.000000	2756.300000	2223.90000	81561.0
1995Q3	1775.800000	3022.800000	2618.60000	86385.0
1995Q4	1762.200000	3203.100000	2581.60000	71463.0
1996Q1	1882.600000	2890.800000	2330.10000	84806.0
...	...	...	...	...
2019Q2	19154.941736	18837.769720	2501.83600	124764.0
2019Q3	19038.573194	18504.265237	2552.02500	142484.0
2019Q4	19217.321673	19704.177128	2602.99800	146849.0
2020Q1	19148.158828	17777.865886	2822.28515	125747.0
2020Q2	6228.979782	5027.206052	780.14000	1.0

Change Dataset Timeline

2015 - 2020

	FS	ICG	TSP	SOM
Date				
2015Q1	12328.555125	12670.250538	1690.89500	109248.0
2015Q2	12275.627835	16086.134717	1701.27300	93418.0
2015Q3	12560.931648	15755.989512	2439.37300	86952.0
2015Q4	13161.240861	17918.012761	2790.71600	91184.0
2016Q1	13713.463269	15783.503816	1919.76700	106224.0
2016Q2	13513.860603	17687.022792	1802.06600	95922.0
2016Q3	13501.733745	15623.794846	2240.07200	95237.0
2019Q1	18652.156040	16967.177973	2278.06000	132716.0
2019Q2	19154.941736	18837.769720	2501.83600	124764.0
2019Q3	19038.573194	18504.265237	2552.02500	142484.0
2019Q4	19217.321673	19704.177128	2602.99800	146849.0
2020Q1	19148.158828	17777.865886	2822.28515	125747.0

Remove Insufficient Attributes

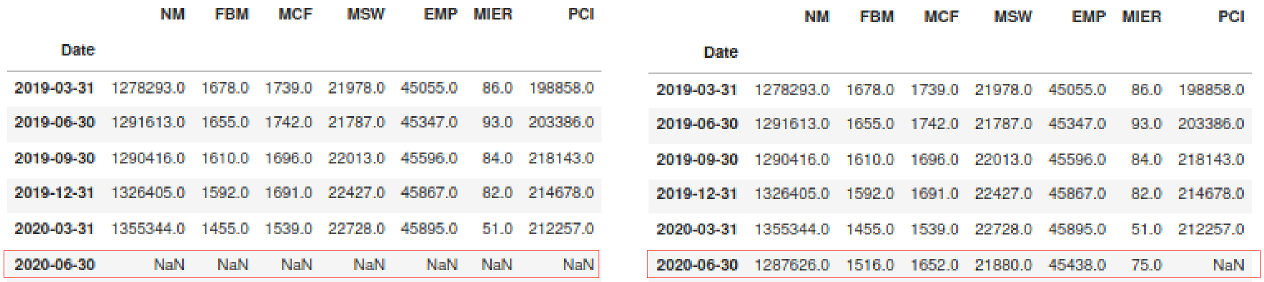
SSW, & MSV are discarded

	Date	SSW	MSV
0	2015Q1	0.0	0.0
1	2015Q2	0.0	0.0
2	2015Q3	0.0	0.0
3	2015Q4	0.0	0.0
4	2016Q1	0.0	0.0
5	2016Q2	0.0	0.0
16	2019Q1	0.0	0.0
17	2019Q2	0.0	0.0
18	2019Q3	0.0	0.0
19	2019Q4	0.0	0.0
20	2020Q1	0.0	0.0

Figure 7 Output of each process in Data Cleansing phase

3.2.2 Data Preparation

After cleansing the dataset, it was noted that initially there are no value for RPC indicators for second quarter of 2020 since they are future values that were not published yet. Therefore, in order to fill up the values of the RPC indicators, windowing method integrated with tree-based ensembled learning models were used to predict the future values. In particular, the windows were evaluated using RMSE with the historical data and the windowing process is iteratively done by changing the window size until the predicted values achieved the least RMSE in comparison with the actual value. Using the best window size, all of the future values of RPC indicators were predicted using bagging, RF, and boosting models simultaneously. Figure 8 illustrates the windowing process before and and future value predictions using Random Forest.



Date	NM	FBM	MCF	MSW	EMP	MIER	PCI
2019-03-31	1278293.0	1678.0	1739.0	21978.0	45055.0	86.0	198858.0
2019-06-30	1291613.0	1655.0	1742.0	21787.0	45347.0	93.0	203386.0
2019-09-30	1290416.0	1610.0	1696.0	22013.0	45596.0	84.0	218143.0
2019-12-31	1326405.0	1592.0	1691.0	22427.0	45867.0	82.0	214678.0
2020-03-31	1355344.0	1455.0	1539.0	22728.0	45895.0	51.0	212257.0
2020-06-30	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Date	NM	FBM	MCF	MSW	EMP	MIER	PCI
2019-03-31	1278293.0	1678.0	1739.0	21978.0	45055.0	86.0	198858.0
2019-06-30	1291613.0	1655.0	1742.0	21787.0	45347.0	93.0	203386.0
2019-09-30	1290416.0	1610.0	1696.0	22013.0	45596.0	84.0	218143.0
2019-12-31	1326405.0	1592.0	1691.0	22427.0	45867.0	82.0	214678.0
2020-03-31	1355344.0	1455.0	1539.0	22728.0	45895.0	51.0	212257.0
2020-06-30	1287626.0	1516.0	1652.0	21880.0	45438.0	75.0	NaN

Figure 8 Output of RPC indicators before (left) and after (right) Windowing process

3.2.3 Model Training

Currently, model training is still an ongoing process. As of now, basic tree-based ensemble learning models were developed using default parameters. The dataset was splitted into three sets which were training, validating, and testing sets. Model training uses training, and validating sets for the model to learn on the underlying pattern of RPC indicators. Meanwhile, the testing set were used later during the model evaluation phase. Table 3 tabulates the current results of the prediction models with RMSE used as the evaluation metrics and Figure 9 plots the overall output all models' training phase.

Table 3 Model prediction results using default parameters

	Bagging	Random Forest	Boosting
RMSE	16264	18359	15594

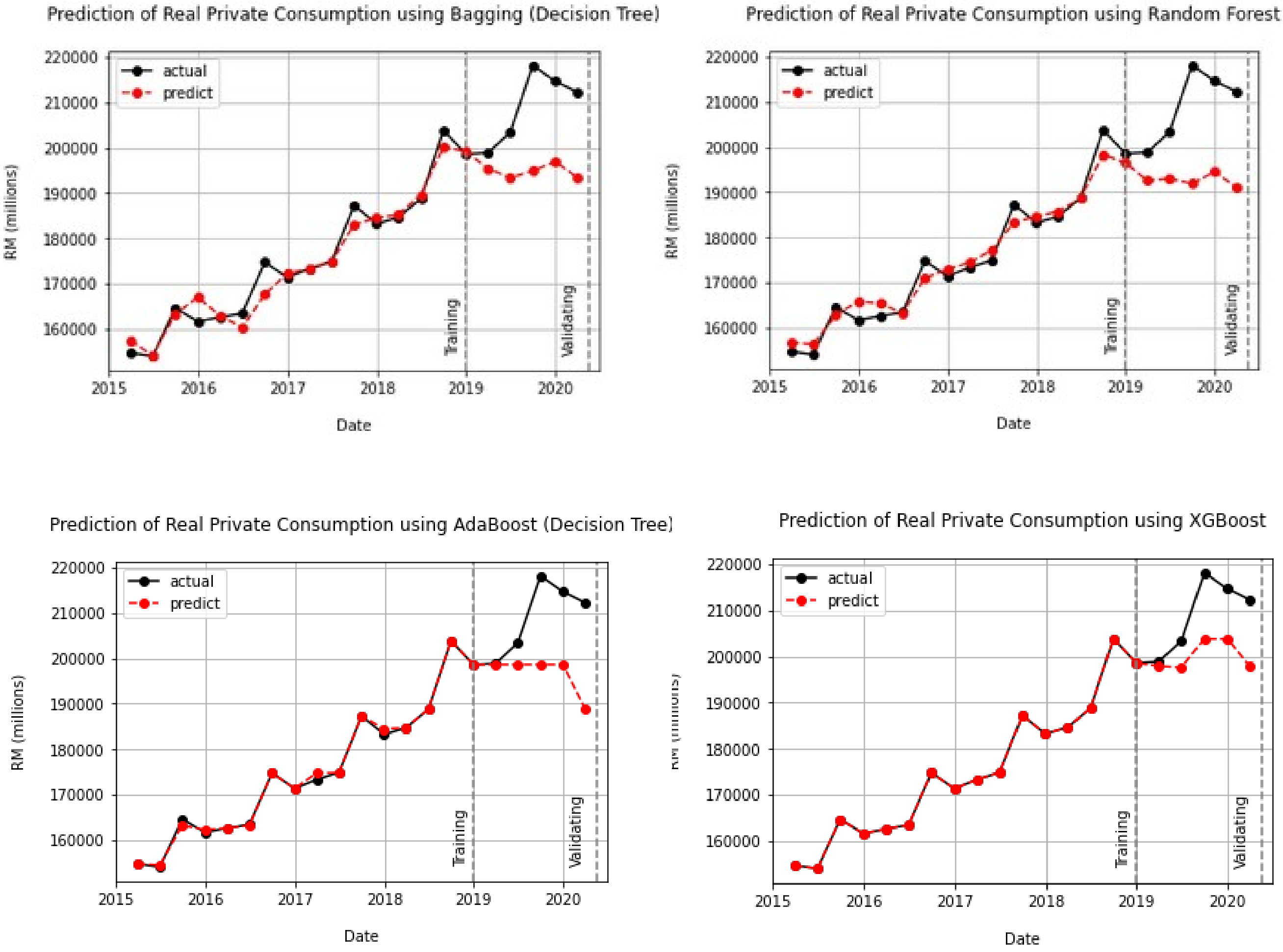


Figure 9 Outputs of all model predictions using default parameter

References

Afandi, A. and Khoo, R. (2020). Ringgit were not face extreme volatility thanks to managed float. *Bernama*. Retrieved on Oct 18, 2020 from https://www.bernama.com/en/general/news_covid-19.php?id=1816684

Asada, H., Kiang, T. K., Espinoza, R., and Vandeweyer, M. (2019). *OECD Economic Surveys 2019: Malaysia*. Retrieved on Oct. 8. 2020, from http://www.oecd.org/economy/surveys/Malaysia-2019-OECD-economic-survey-overview.pdf

Bank Negara Malaysia (2020). *Annual Report 2019*. Retrieved on Oct. 15, 2020 from https://www.bnm.gov.my/ar2019/files/ar2019_en_full.pdf

Bank Negara Malaysia (2019). *Annual Report 2018*. Retrieved on Oct. 15, 2020 from https://www.bnm.gov.my/files/publication/ar/en/2018/ar2018_book.pdf

Bank Negara Malaysia (2018). *Annual Report 2017*. Retrieved on Oct. 15, 2020 from https://www.bnm.gov.my/files/publication/ar/en/2017/ar2017_book.pdf

Bank Negara Malaysia (2017). *Annual Report 2016*. Retrieved on Oct. 15, 2020 from https://www.bnm.gov.my/files/publication/ar/en/2016/ar2016_book.pdf

Bernama (2018). Azmin: Statistical Community needs to Embrace Digital Revolution. *The Edge Markets*. Retrieved on Oct. 9, 2020, from https://www.theedgemarkets.com/article/azmin-statistical-community-needs-embrace-digital-revolution

Brownlee, J. (2017). *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. 1st ed. Machine Learning Mastery.

Burns, D. M., and Whyne, C. M. (2018). Seglearn: A python package for learning sequences and time series. *Journal of Machine Learning Research*, *19*(1), pp. 3238-3244

Chen, J. C., Dunn, A., Hood, K., Driessen, A., & Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Dematos, G., Boyd, M.S., Kermanshahi, B. (1996). Feedforward versus recurrent neural networks for forecasting monthly japanese yen exchange rates. *Financial Engineering and the Japanese Markets, 3*, pp. 59–75

Department of Statistics Malaysia (2020). *National Accounts FAQ.* Retrieved on Oct. 8, 2020 from https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=dUtRR1JYWjk2TEJha1BrZml0REY4UT09

Fadzil, M., Latif, L. A., and Munira, T. A. (2015). MOOCsin Malaysia : A preliminary case study. *MOOCs and Educational Challenges around Asia and Europe, 1*(6), pp. 65-86

Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn*. United Kingdom: Packt Publishing Ltd

Hashim, E., Ramli, N. R., Romli, N., Jalil, N. A., Bakri, S. M., and Ron, N. W. (2018). Determinants of Real GDP in Malaysia. *The Journal of Social Sciences Research*, No. 3, pp. 97-103

Kumar, I., Dogra, K., Utreja, C., and Yadav, P. (2018). A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Prediction. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1003-1007

Maehashi, K., & Shintani, M. (2020). Macroeconomic forecasting using factor models and machine learning: an application to Japan. *Journal of the Japanese and International Economies*, *58*, 101104.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889

McKinney, W., Perktold, J., and Seabold, S. (2011). Time Series Analysis in Python with statsmodels. *Proceedings of the 10th Python in Science Conference,* pp 107-113

Microsoft (2020). *The Business Understanding Stage of the Team Data Science Process Lifecycle.* Retrieved on Oct 17, 2020, from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-business-understanding

Rasic, A. H. (2019). Consumer Sentiment, Business Condition Indexes Down in Q4. *New Straits Times.* Retrieved on Oct. 17, 2020, from https://www.nst.com.my/business/2019/01/456084/consumer-sentiment-business-condition-indexes-down-q4

Rasel, R. I., Sultana, N., & Meesad, P. (2015). An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators. *International Journal of Computational Intelligence Studies*, *4*(2), pp. 134-150

Razak, N. A. A., Khamis, A., and Abdullah, M. A. A. (2017). ARIMA and VAR Modeling to Forecast Malaysian Economic Growth. *Journal of Science and Technology: Special Issue on the Application of Science and Mathematics, 9*(3), pp. 16-24

Roy, M., and Larocque, D. (2012).Robustness of Random Forests for Regression. *Journal of Nonparametric Statistics*, 24(4), pp. 993-1006

Taieb, S. B. (2014). Machine learning Srategies For Multi-Step-Ahead Time Series Forecasting. *Université Libre de Bruxelles, Belgium*, pp.75-86

United Nations, (2020). *World Economic Situation and Prospects 2020*. New York: United Nations Publication.

Usher, J., and Dondio, P. (2020). BREXIT Election: Forecasting a Conservative Party Victory through the Pound using ARIMA and Facebook's Prophet. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pp. 123-128

Vo, V., Luo, J., and Vo, B. (2016). Time Series Trend Analysis Based on K-Means and Support Vector Machine. *Computing and Informatics*, *35*, pp. 111-127

World Bank Group (2020). *Malaysia Economic Monitor (June): Surviving the Storm.* Washington: World Bank Publications

Yu, S. (1999), Forecasting and Arbitrage of the Nikkei Stock Index Futures: An Application of Backpropagation Networks. *Asia-Pacific Financial Markets, 6*, pp. 341–354