

**School of Computer Sciences**

CDS590 – Consultancy Project & Practicum

Final Report

Predicting Real Private Consumption using Time Series:  
A Machine Learning Approach

MUHAMMAD AZZUBAIR BIN AZEMAN  
P-COM 0019/19

Supervisor : Mohd Azam Osman

Mentor : Patrick Lam Kar Jun

SEM 1 2020 / 2021

DECLARATION

“I declare that the following is my own work and does not contain any

***unacknowledged*** work from any other sources. This project was undertaken to fulfil

the requirements of the Consultancy Project & Practicum for the Master of Science

(Data Science and Analytics) program at Universiti Sains Malaysia”.

Signature : ……………………………

Name : ……………………………

Date : ……………………….……

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to Allah S.W.T., whom with His will, I have been given the opportunity to complete this practicum project entitled Predicting Real Private Consumption using Time Series Data: A Machine Learning Approach.

I would like to thank my supervisor, Mr Mohd Azam Osman, for his encouragement, guidance, and patience. Without ~~her~~ his assistance and dedicated involvement ~~in every step~~ ~~throughout the process~~ in this project, this practicum report would never accomplish.

I would also like to thank all my lecturers and teaching assistants who have given directly or indirectly contributed in this project. Their assistances were very helpful and meaningful to me, and despite all those handy helps, they also gave me moral support and endless guidance.

Not to forget, my project colleagues, Nur Farahin, Abdul Hadi, Haris Idris, Faiz and to all my supportive friends, thank you for helping me a lot throughout this project and for your endless support, I have owned them.

Finally, I want to thank you my beloved family, Azeman, and Kartinah for constantly supporting, encouraging, and praying for my success.

# ABSTRACT

Prediction comparison of real private consumption (RPC) between machine learning ~~with~~ and statistical techniques is still questionable among economists. It is well-established that statistical techniques dominated the time series prediction of macroeconomics since 20th century. This project aims at predicting time series of real private consumption using machine learning technique. ~~Specifically, this project investigates the~~ Best machine learning models are investigated for macroeconomics, develop, and evaluate them for RPC prediction together with the currently applied statistical techniques. In this context, RPC is defined as the quarterly aggregated amount of goods and services consumed by households to fulfil their basic needs and wants. To make the comparison between machine learning with statistical techniques in predicting future values of RPC, the best machine learning models found from literature review were developed, optimised, evaluated, and used for predicting RPC. The results showed the majority of the current statistical techniques surpassed the optimised machine learning models by having five times RMSE lower than the prior models. These results suggest that much more optimisation steps must be done on the developed machine learning models. On this basis, the best feature selection algorithms, and configuration of model hyperparameters should be taken into account when modelling machine learning models for them to surpass the current statistical techniques used for prediction of RPC future values.

1. **General statement on RPC prediction using statical approach. Machine Trend in prediction. (2) Problem faced by ..MoF in order to predict.. (3) proposed solution (4) aims of your project. (5) which ML models give best performance. (6) Result of comparison are presented in this report. (7) based on the result, ML model will be suggested to MoF for ,…**

# ABSTRAK

Perbandingan ramalan perbelanjaan penggunaan akhir swasta (PPAS) antara teknik pembelajaran mesin dengan teknik statistik masih dipersoalkan dalam kalangan ahli ekonomi. Sudah terbukti bahawa teknik statistik mendominasi ramalan siri ekonomi makro sejak abad ke-20. Projek ini bertujuan untuk membuat percubaan untuk meramalkan siri masa perbelanjaan penggunaan akhir swasta. Secara khusus, projek ini menyelidik model pembelajaran mesin yang terbaik bagi ekonomi makro, membina, dan menilai mereka untuk ramalan PPAS bersama dengan teknik statistik yang sedang digunakan. Dalam konteks ini, PPAS didefinisikan sebagai jumlah keseluruhan barang dan perkhidmatan yang digunakan oleh isi rumah untuk memenuhi keperluan dan kehendak asas mereka. Untuk membuat perbandingan antara teknik pembelajaran mesin dengan teknik statistik dalam meramalkan nilai PPAS di masa depan, model pembelajaran mesin terbaik yang dijumpai dari kajian lepas dikembangkan, dioptimumkan, dinilai, dan digunakan untuk meramal PPAS. Hasilnya menunjukkan sebahagian besar teknik statistik ketika ini kekal lebih baik daripada model pembelajaran mesin yang dioptimumkan dengan RMSE dengan kadar kesilapan lima kali lebih rendah daripada model-model statistik. Hasil ini menunjukkan bahawa langkah pengoptimuman yang lebih banyak harus dilakukan pada model pembelajaran mesin yang telah dibina. Atas dasar ini, algoritma pemilihan pemboleh ubah yang terbaik, dan konfigurasi hiperparameter model harus dipertimbangkan ketika memodelkan model pembelajaran mesin supaya mereka mejadi lebih baik dan lebih tepat daripada teknikstatistik yang masih lagi digunakan sehingga kini untuk meramal nilai masa depan PPAS.

**TABLE OF CONTENTS**

[DECLARATION i](#_Toc62203570)

[ACKNOWLEDGEMENTS ii](#_Toc62203571)

[ABSTRACT iii](#_Toc62203572)

[ABSTRAK iv](#_Toc62203573)

[LIST OF TABLES vii](#_Toc62203574)

[LIST OF FIGURES viii](#_Toc62203575)

[LIST OF ABBREVIATIONS AND SYMBOLS ix](#_Toc62203576)

[**CHAPTER 1 INTRODUCTION & RELATED WORKS** 1](#_Toc62203577)

[1.1 Background of Practicum Company 1](#_Toc62203578)

[1.2 Background of Domain 2](#_Toc62203579)

[1.3 Problem Statement 3](#_Toc62203580)

[1.4 Research Question 3](#_Toc62203581)

[1.5 Objectives 4](#_Toc62203582)

[1.6 Benefits of the project 4](#_Toc62203583)

[1.7 Related Works 4](#_Toc62203584)

[1.7.1 Review on Domain 4](#_Toc62203585)

[1.7.2 Review on Data Science & Analytics techniques 5](#_Toc62203586)

[1.7.3 Review on Data Science & Analytics tools 7](#_Toc62203587)

[**CHAPTER 2 RESEARCH METHODOLOGY** 8](#_Toc62203588)

[2.1 Activities plan and Gantt Chart 8](#_Toc62203589)

[2.2 Data Science Project Lifecycle 9](#_Toc62203590)

[2.3 Problem Analysis 10](#_Toc62203591)

[2.3.1 Initial Questions 10](#_Toc62203592)

[2.3.2 Specific use case to be addressed 10](#_Toc62203593)

[2.3.3 Exploratory Data Analysis 11](#_Toc62203594)

[2.3.4 Final Analysis 13](#_Toc62203595)

[2.4 Proposed Solution 15](#_Toc62203596)

[2.5 Justification on Selected Data Science Techniques and Tools 16](#_Toc62203597)

[2.6 Contribution 16](#_Toc62203598)

[**CHAPTER 3 RESULTS AND DISCUSSIONS** 17](#_Toc62203599)

[3.1 Investigation on Machine Learning techniques 17](#_Toc62203600)

[3.2 Development of selected Machine Learning models 18](#_Toc62203601)

[3.2.1 Data Cleansing 18](#_Toc62203602)

[3.2.2 Data Preparation 19](#_Toc62203603)

[3.2.3 Model Development 20](#_Toc62203604)

[3.2.4 Model Optimisation 21](#_Toc62203605)

[3.3 Model Evaluation 23](#_Toc62203606)

[3.4 What have been achieved and not achieved 27](#_Toc62203607)

[3.4 Challenges and Solutions 27](#_Toc62203608)

[3.5 Practicum Experience Applicability from Class 28](#_Toc62203609)

[3.6 Observations during Practicum related to Professional and Operational Issues 29](#_Toc62203610)

[**CHAPTER 4 CONCLUSION AND LESSON LEARNED** 31](#_Toc62203611)

[4.1 Main Conclusion 31](#_Toc62203612)

[4.2 Lesson Learned 31](#_Toc62203613)

[4.2.1 Have a basic knowledge of client’s domain 31](#_Toc62203614)

[4.2.2 Master communication skills to engage with surrounding people and clients 32](#_Toc62203615)

[4.2.3 Possess storytelling skills 32](#_Toc62203616)

[4.3 Future Works 32](#_Toc62203617)

[**REFERENCES** 33](#_Toc62203618)

# LIST OF TABLES

[Table 1 Description of Real Private Consumption and its indicators 11](#_Toc62209620)

[Table 2 Summary of Literature Reviews 17](#_Toc62209621)

[Table 3 Model prediction errors (RMSE) during Model Development 20](#_Toc62209622)

[Table 4 Averaged RMSE of machine learning models with respect to windowing techniques 25](#_Toc62209623)

[Table 5 Averaged RMSE of statistical models 25](#_Toc62209624)

# LIST OF FIGURES

[Figure 1 Gantt Chart of Project Consultancy and Practicum 8](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209124)

[Figure 2 Lifecycle of Data Science Projects 9](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209125)

[Figure 3 Overview of Real Private Consumption trends and its indicators 12](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209126)

[Figure 4 Flow Chart of RPC Forecasting 14](#_Toc62209127)

[Figure 5 Generation of horizontal windows via column transpose 14](#_Toc62209128)

[Figure 6 Illustration of sliding window (left) and expanding window (right) method 15](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209129)

[Figure 7 Output of each process in Data Cleansing phase 18](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209130)

[Figure 8 Output of RPC indicators before (left) and after (right) Windowing process 19](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209131)

[Figure 9 Recommended configuration of RPC prediction by Random Forest (Sliding Window) 21](file:///C:\Users\azzub\OneDrive%20-%20Universiti%20Sains%20Malaysia\Desktop\CDS%20590\Time-Series-Forecasting\Final%20Report\Draft%202%20Report.docx#_Toc62209132)

[Figure 10 Generated heatmap using Expanding Window 22](#_Toc62209133)

[Figure 11 Generated heatmap using Sliding Window 22](#_Toc62209134)

[Figure 12 Generated heatmap using Non-Random Percentage Split 22](#_Toc62209135)

[Figure 13 Comparison of RMSE before and after model optimisation 23](#_Toc62209136)

[Figure 14 Comparison of RMSE between machine learning with statistical methods 25](#_Toc62209137)

[Figure 15 RPC prediction using Random Forest with sliding window 26](#_Toc62209138)

[Figure 16 Waterfall Model of DataMicron project management methodology 30](#_Toc62209139)

[Figure 17 DataMicron’s Flow of Project Management 30](#_Toc62209140)

# LIST OF ABBREVIATIONS AND SYMBOLS

4QMA 4-Quarter Moving Average

ARIMA Autoregressive Integrated Moving Average

BNN Backpropagation Neural Network

BNM Bank Negara Malaysia

CC Loans disbursed for Consumption Credit

CCS Credit Card Turnover Spending

CNN Convolutional Neural Network

COVID-19 Coronavirus Disease 2019

EW Expanding Window

FBM KLCI FTSE Bursa Malaysia Kuala Lumpur Composite Index

FNN Feedforward Neural Network

FSR Forward Stepwise Regression

GB Gradient Boosting

GFCF Gross Fixed Capital Formation

ICG Imports of Consumption Goods

KNN k-Nearest Neighbour

LASSO Least Absolute Shrinkage and Selection Operator

LSTM Long Short Term Memory

MAPE Mean Absolute Percentage Error

MARS Multi-Adaptive Regression Splines

MFVAR Mixed Frequency Vector Autoregression

MIDAS Mixed Data Sampling

MIER Consumer Sentiment Index

MITI Minister of International Trade and Industry

MoF Ministry of Finance

NB Naive Bayes

NDA Non-Disclosure Agreement

NE Net Export

NM Narrow Money

PCE Personal Consumption Expenditure

PCI Private Consumption Index

POC Proof of Concept

QSS Quarterly Service Survey

RBF Radial Basis Function

RF Random Forest

RGC Real Government Consumption

RMSE Root Mean Squared Error

RPC Real Private Consumption

RR Ridge Regression

RT Regression Trees

SM Softmax

SOM Sales of Motorcycle

SOPC Sales of Passenger Cars

SVR Support Vector Regression

SW Sliding Window

UECM Unrestricted Error Correction Model

XGBoost Extreme Gradient Boosting

# 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 definedas “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 contributes 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 (VAR) 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 have been proposed and discussed. However, machine learning techniques for RPC prediction were foreign to Malaysians until these techniques were first publicly recommended by the Minister of International Trade and Industry (MITI) (Bernama, 2018). As a result, machine learning models and techniques are gradually being learned by Malaysians in many online courses recently (Fadzil, Latif, & Munira, 2015).

## 1.3 Problem Statement

Numberous studies have been done to compare ~~on comparing between~~ the 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 studies have been done ~~in comparing~~ to compare model performances between machine learning ~~models~~ ~~with~~ and statistical models for time series prediction of real private consumption in Malaysia. Addressing the performances between statistical model ~~with~~ and machine learning models has practical benefits for researchers and related authority in economics and thus, it will contribute to understanding of both approaches for time series prediction of real private consumption specifically and macroeconomics generally.

## 1.4 Research Question

This research ~~makes an attempt to~~ is aimed at predicting time series of real private consumption using machine learning technique. Thus, this led to determination of ~~usThis were consequently resulting 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~~ That leads into 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 model performance evaluation?

Throughout this project, the answers ~~of~~ to research questions are extracted from literature reviews, ??? and discussion sections in this ~~project~~ report.

## 1.5 Objectives

~~The aim of~~ this project aims to propose machine learning techniques as a new approach to improve econometric forecasting/a prediction ??? in Malaysia. Thus, this project will focus on forecasting/predicting one of the econometric indicators which is real private consumption (RPC) based on data published quarterly by DoSM. The following are the objectives of this project:

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.

## 1.6 Benefits of the project

This is ~~problem~~ is ~~actually~~ a consultation project between DataMicron Sdn Bhd~~with~~ and the Malaysia’s Ministry of Finance (MoF). Therefore, this project will benefit DataMicron Sdn Bhd by providing the proposed efficient solution for MoF. This project will propose new methodologies ~~of~~ for RPC prediction using machine learning approaches for MoF ~~as a reference.~~  ~~and were develop reliable models for suggesting RPC predictions as a reference for MoF in making effective decisions~~. Thus, the MoF will use ~~these insights~~ that information to optimise ~~their~~ its financial planning and reduce its financial loss.

## 1.7 Related Works

In this section, proposed methods in RPC predictions using machine learning techniques are presented. ~~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 reviewe~~ Then, Data science, analytics techniques and tools are discussed in this section.

### 1.7.1 ~~Review on~~ Problem Domain

Since 2018, RPC has been the largest contributor ~~contributed the most~~ to Malaysia’s GDP followed by GFCF 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~~ is mainly due to ~~by~~ the growth of employments and wages. ~~Thus,~~ This ~~shows~~ indicates that ~~affecting~~ the impact of this sector has been significant throughout ~~resulted much to~~ the overall annual RPC. ~~Other than that~~ Besides, RPC ~~is~~ has also been ~~being~~ contributed by other variables such as imports of consumption goods, narrow money, and loans disbursed by banks (BNM, 2016). The importance of RPC prediction ???

### 1.7.2 ~~Review on~~ Data Science & Analytics techniques

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

Machine learning has become more common in recent year. ~~In other parts of the world~~ Hence, machine learning techniques ~~were began~~ have started to be ~~proposed~~ used 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 that BNN produces 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 has been done by Dematos et. al (1996) in which they found that Recurrent Neural Network (RNN) and Feedforward Neural Network (FNN) are outperformed ARIMA in forecasting Japanese yen / U.S. dollar exchange rate. ~~This summarised~~ In summary, the nonlinear trend of economic indicators is more accurate to ~~be forecasted~~ predict using machine learning.

Two years back, a comparative study that was done by Kumar et. al (2018) to compare machine learning performances between several selected machine learning techniques for predicting stock market trend. Machine learning techniques used for this comparison are support vector regression (SVR), random forest (RF), k-nearest neighbour (KNN), naive bayes (NB), and SoftMax (SM). Interestingly, ~~they found out discovered that~~ when large dataset (4500 entries) is input to those models, RF is found that outperformed than 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 be 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 is to compare time series prediction performances between linear models with nonparametric models. Linear models included in the study are 4-Quarter Moving Average (4QMA), Forward Stepwise Regression (FSR), Least Absolute Shrinkage and Selection Operator (LASSO), and Ridge Regression (RR). Meanwhile, nonparametric models implemented in the study are 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 reveals that tree-based ensemble models which are RF and GB are the best models as two models have the least RMSE percentage points of -0.56 and -0.43 respectively. In contrast, MARS and 4QMA are suggested to be avoided due to these two methods are significantly poor prediction performance compared to others.

Recently, Maehashi and Shintani (2020) also performed comparison study ~~between~~ on several machine learning models in predicting macroeconomic variables. Interestingly, ~~their~~ this ~~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 (FNN), Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM) for neural network models are deployed. Meanwhile, for regularised least square method, ~~they included~~ Lasso, Ridge, and Elastic Net regressions are included in this study 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 has concluded that ensemble learning methods outperformed than other models due to their adaptability with nonlinear trends of the macroeconomics variables.

Another important finding 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 time series are included for model training, the predictions will become better, more accurate, and robust against errors.

### 1.7.3 ~~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 done using Python for time series prediction and forecasting because of ~~there are~~ many libraries available ~~that are created~~ for ~~the purpose of~~ dealing with time series data and forecasts/predicting??. One of the popular libraries available in Python for time series forecasting is statsmodels library. According to McKinney et al. (2011), this library provides many statistical models ~~made available~~ for 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 for time series forecasting is by using machine learning approach. This approach recognises time series data as a supervised learning using sliding window for model training and testing. Brownlee (2017) has explained about the sliding window in detail, ~~whereby he described~~ is 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~~ produced 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. The outcome of their comparison shows that all of the libraries has forecasting feature except tslearn. On top of that, ~~they found out that~~ seglearn library is found that the only library ~~that~~ which 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 provides 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.

# 1.8 Chapter Summary

Several feature detectors has been described in Section 2.3.1. However, based on the literatures, Harris Corner Detector provides better detection of curves and edges. The focus is to obtain most of the points that would represent the features for further process. The selection of the feature detector will be described later in Chapter 4.

# CHAPTER 2

# RESEARCH METHODOLOGY

## 2.1 Activities plan and Gantt Chart

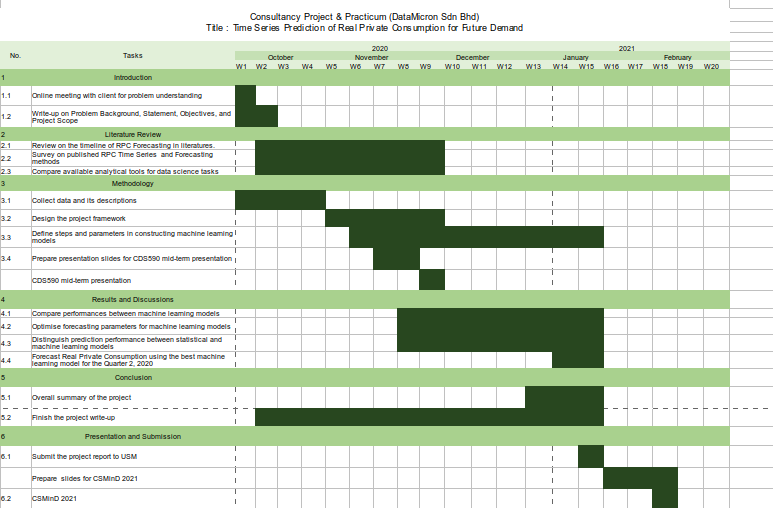
Throughout semester 1 2020 / 2021, the activities for project consultancy and practicum had been conducted based on the plans as illustrated in figure 1. This project is ~~done~~ carrioud individually with the supervision of project supervisor and guided by a mentor from the practicum company. With the limitations of the current coronavirus disease 2019 (COVID-19) situation, all of the process for the project consultancy and practicum had been 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 recorded to update the progress of the project.

Figure Gantt Chart of Project Consultancy and Practicum

## 2.2 Data Science Project Lifecycle

As a data scientist consultant, it is important to implement 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 toMicrosoft (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 going through.

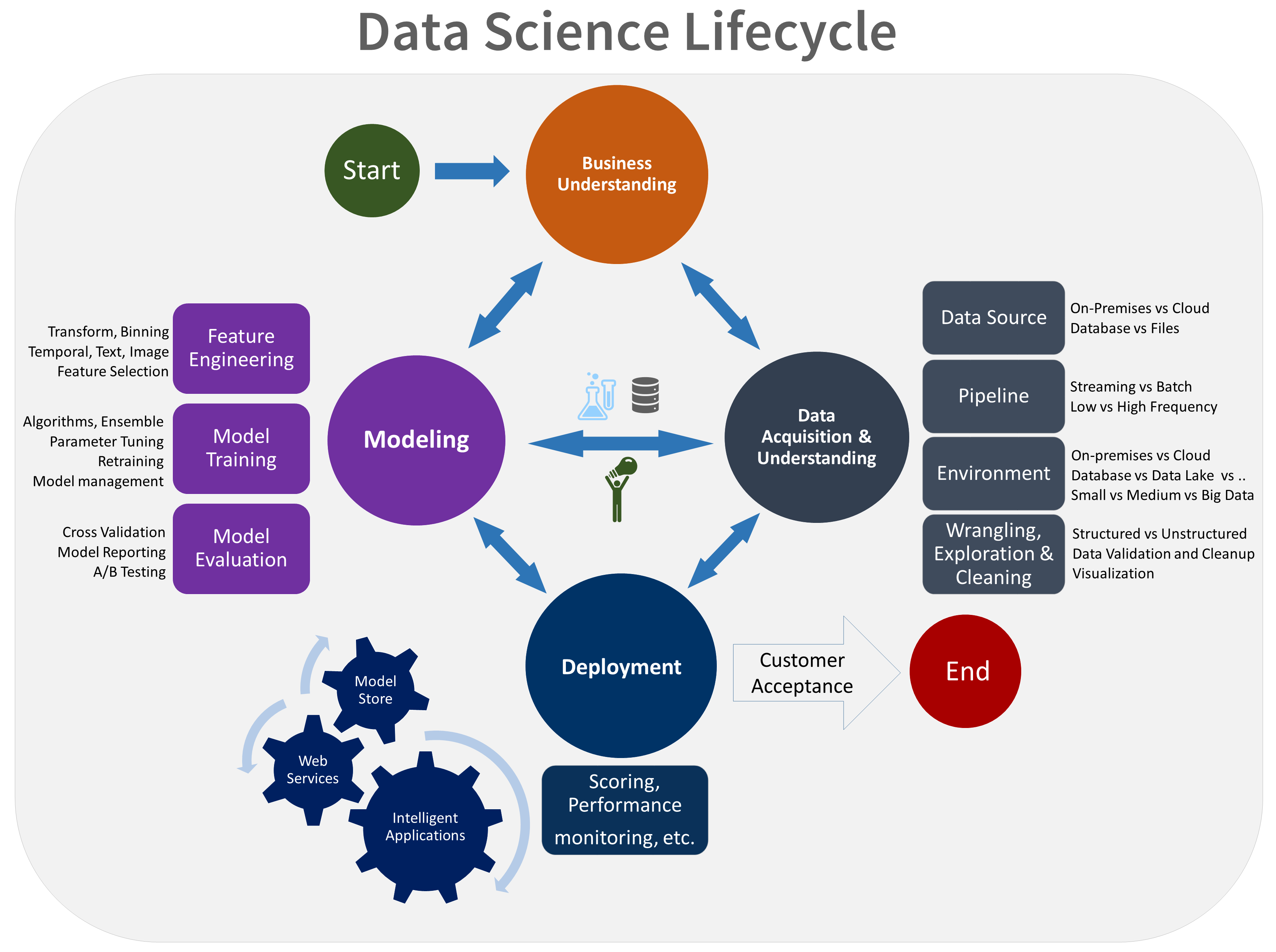


Figure Lifecycle of Data Science Projects

~~In short,~~ Business Understanding stage was conducted during a consultation meeting where MoF explained the background of their RPC problem. The analysis of the problem and the description of their expected solution were uncovered in the form of research questions form. ~~and describe their expected solutions~~. In order to determine the success of the proposed solution, it is measured, and ensured to be within clients’ expectations using success metrics that are specific, measurable, achievable, relevant, and time-bound (Microsoft, 2020). The next stages were conducted individually and ~~lastly~~ finally the proposed solution ~~were~~ was 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 was prepared and asked to the client (see in Appendix ??) !!! Ada sample answer tak?? Kalau ada mohon letak dalam Appendix!!! . The purpose of these questions is mainly to determine the problem framing, identify target and predictor variables, and pinpoint for data source. ~~Before coming into~~ Prior to address the main questions, general questions related to domain background were prepared. The questions are listed in following sub-section. After the responding to the questions, an exploratory data analysis and a final analysis were carried out.

### 2.3.1 Initial Questions

The questions listed below have been prepared to understand the ~~client~~ ~~to know of~~ the domain background of the client, methods currently being used for RPC forecasting, and solution expectations. The answers of these questions have been 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 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 of machine learning models would be expected for RPC forecasting?
* What is the expected model performance metrics to be implemented?

### 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 provided financial support to its citizens through *Prihatin Rakyat* and *Penjana* packages, real private consumption will still 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 was conducted to provide an overview of the variable distributions, insights/information ??, and trends. In terms of Data Science Lifecycle, this step is categorised under Data Acquisition and Understanding. Firstly, the dataset was given in an excel file containing variables of real private consumption indicators as published in BNM (2016). These variables ~~is~~ are collected from data published by DoSM, and BNM from 1995 until 2020. All the numerical variables were cleaned and transformed into quarterly data for them to be aligned with quarterly RPC. Table 1 shows details of variables types and descriptions.

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 (CC) | 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. | 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 |

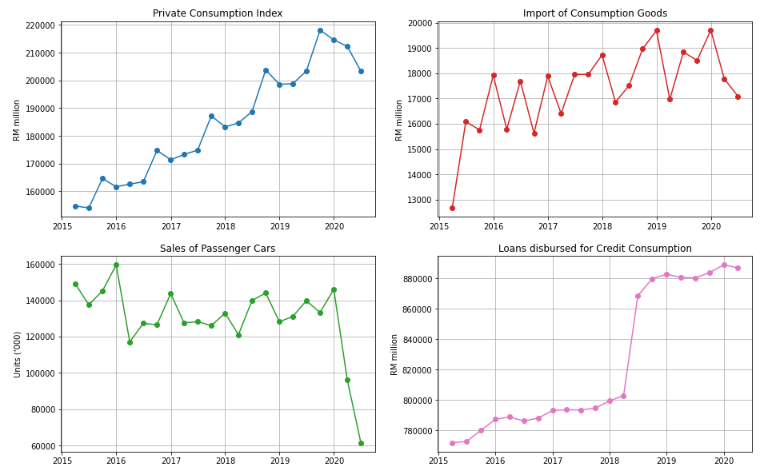
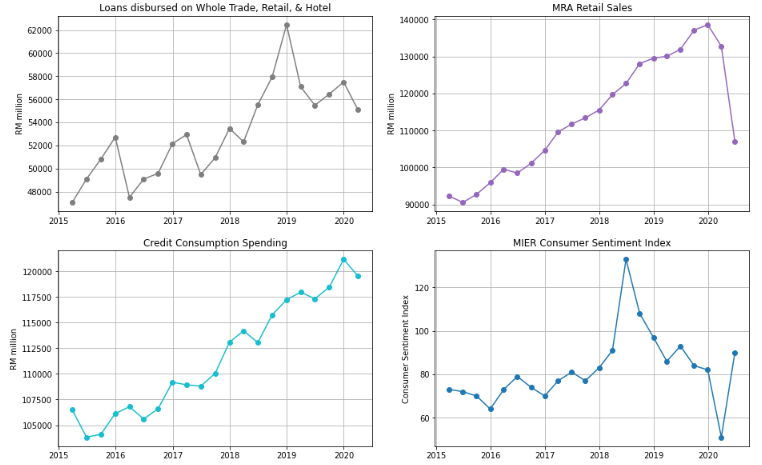
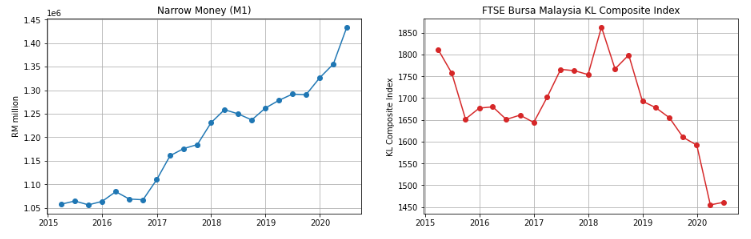


Figure Overview of Real Private Consumption trends and its indicators

After describing each attribute, Python were used to ~~be~~ visualise the trends of each attribute. Figure 3 ~~below~~ shows the overall visualisation of RPC 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 show an improvement of RPC growth throughout the years. In addition, MIER also shows an increasing trend but only until the second quarter of 2018, beyond than that, MIER has declined. It is due to Malaysia’s General Election held during the second quarter of 2018; consumer sentiments exploded ~~went skyrocketed~~ until the end of the second quarter of 2018. Beyond ~~that~~ this quarter, pessimists started to outperform ~~began outnumbered~~ optimists due to the global challenges affecting ~~Malaysia’s~~ economic growth in Malaysia (Rasid, 2019). In contrast, FBM displays an overall ~~decreasing~~ downward trend with some exceptions in 2018. This attribute is heavily affected by Malaysia’s political issues in which causes 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 significant decrease in the first quarter of 2020 in comparison with the fourth quarter of 2019 excluding the narrow money attribute. This reflects the fact that external virus has inflicted severely on Malaysia’s RPC especially, when Malaysia enforced the Movement Control Order starting in March 2020. In the 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. On the other hand, 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~~ Implementation

Forecasting/Prediction RPC using machine learning techniques is implemented according to the method proposed 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. Windowing implementation is ~~conducted~~ implemented as stated by Rasel et al. (2015). Flow chart ~~for~~ of the proposed methodology is shown in Figure 4. The process flow of the final analysis and discussions are arranged based on this flow chart.

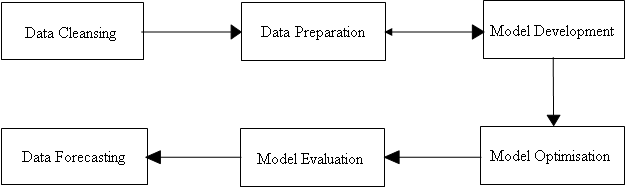
****

Figure Flow Chart of RPC Forecasting

The first step to be done after obtaining the dataset ~~were~~ **is data cleansing** by filtering the timeline of dataset until the year at which all attributes are begun to be collected simultaneously. After data cleansing ~~were~~ is the **data preparation** stage. At this stage, initially there ~~are~~ is no value for RPC indicators for the second quarter of 2020, as these are future values that ~~were~~ have not yet been published. In order to fill up those attributes, they ~~were~~ predicted predicted using windowing method integrated with bagging, RF, and boosting respectively. According to Rasel et al. (2015), these windows are generated by transposing the column of RPC indicators into horizontal windows in which the last row are become the target value to be predicted. This sliding window is then moved 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.

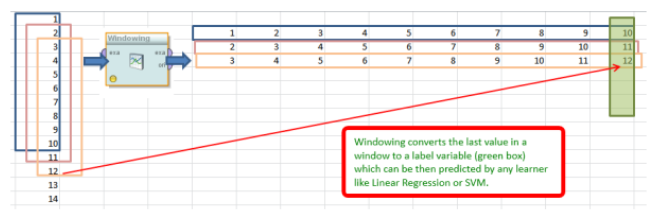


Figure Generation of horizontal windows via column transpose

Figure 5

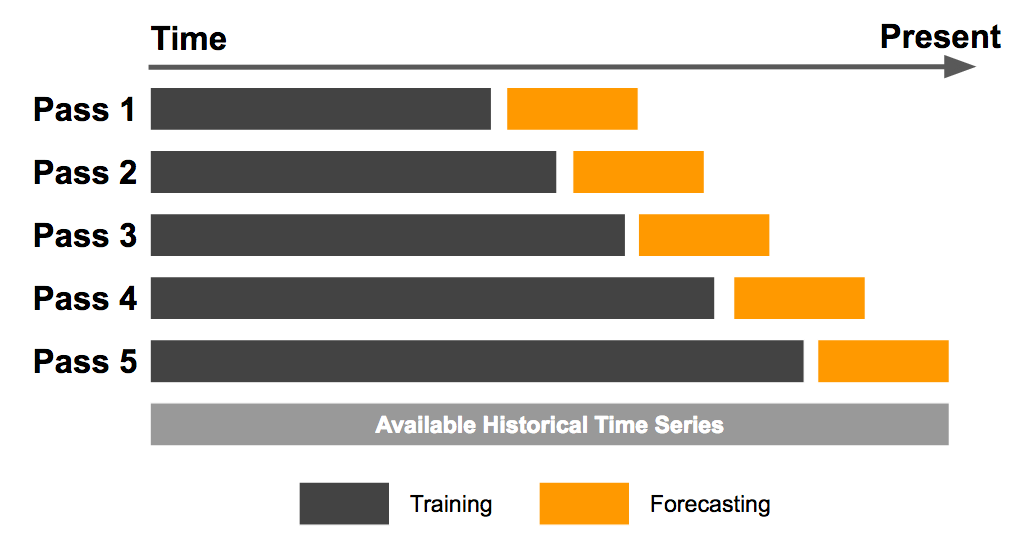
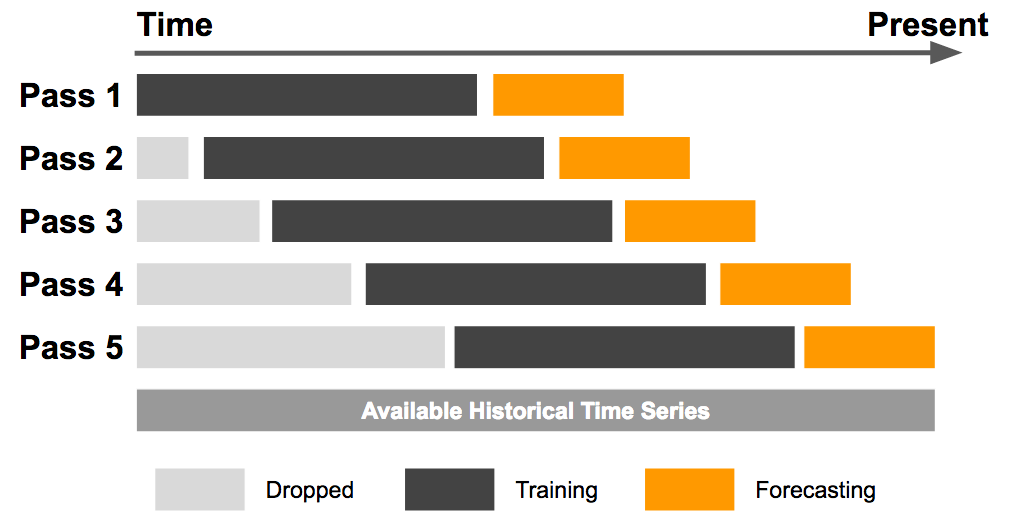


Figure Illustration of sliding window (left) and expanding window (right) method

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

## 2.4 Proposed Solution???

After the ML models have 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 are selected as ~~they~~ these methods are given the most accurate in predicting macroeconomic indicators (Maehashi and Shintani, 2020). Hence, it is justified that Bagging, Boosting, and Random Forest are 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~~ are also 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

~~this project~~ In this chapter, ~~we show that~~ research methodology for this project has been described. Problem analysis and the prediction process of RPC on time series data have been presented in this chapter. Finally, the proposed solution and justification for selected data science techniques and tools have also been explained. ~~the key to modeling RPC relies on the process as illustrated in Figure 4 has been described. 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~~. In a nutshell this chapter attempts to fulfil the first objective and the second objective of this project. The principal contributions of this chapter are:   
a

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

# 

# 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 implemented in this project for RPC prediction.

Table 2 Summary of Literature Reviews

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Author | Algorithm | | | | Findings |
| Kumar et. al (2018) | Parametric | | Nonparametric | | * When dataset is large, RF outperformed others * When dataset is small, NB performed the best |
| * NB * KNN * Softmax | | * RF * SVR | |
| Chen et al. (2019) | Parametric | | Non parametric | | * 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 |
| * 4QMA * LASSO * Ridge | | * CART * RF * XGBoost * SVR * MARS | |
| Maehashi and Shintani (2020) | Linear | Ensemble | | Neural Network | * Tree-based ensemble models were the majority of the best models. * Large window size is recommended for better time series predictions. |
| * LASSO * Ridge * EN | * Bagging * RF * AdaBoost | | * FFNN * CNN * LSTM |

## 3.2 Development of selected Machine Learning models

Machine learning model development were constructed as illustrated in flowchart in Figure 4. It is consisting 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

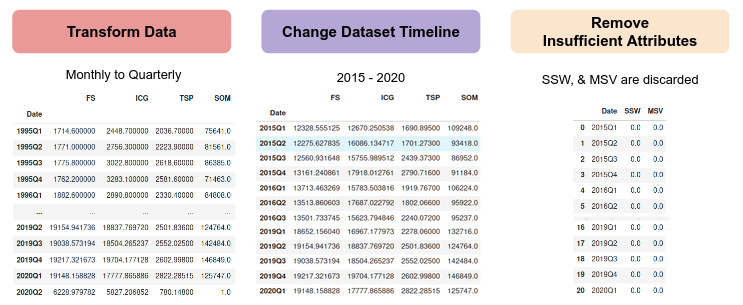
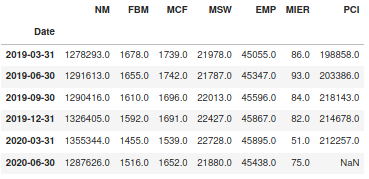
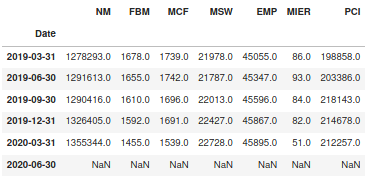
Upon receiving dataset (excel file) from Ministry of Finance’s Fiscal and Economics Division, the data was explored and cleaned using python’s pandas library. Firstly, the timestamp of RPC indicators was not aligned with RPC which are collected monthly and quarterly respectively. Therefore, the indicators 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 was begun to be collected are different. Therefore, dataset timestamp was standardized by filtering it to be starting from date at which all attributes were collected simultaneously which is 1 January 2015. Figure 7 shows the output of each data cleansing process done in python’s pandas library.

Figure 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 future value predictions using Random Forest.

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

****

Furthermore, two types of data splitting methods were used for Model Development stage which are percentage split and time-based cross validation. For percentage split’s initial run, the time series dataset was splitted into 2 sets by 70:30 for training and testing sets. This splitting was applied with non-random splitting due to ordinal property of the time series. Meanwhile for time-based cross validation’s initial run, sliding window and expanding window methods were used. As illustrated in Figure 6, the window size of training dataset for expanding window will keep on expanding until the final iteration, while the window size of training dataset for sliding window is constant (for initial run, window size = 5 was set) and sliding throughout iterations. The output of all data splitting methods was recorded in Model Development stage.

### 3.2.3 Model Development

At this stage, models with initial were trained using multiple data preparation methods. Based on Table 3, it was observed that data splitting method using sliding window are the best as it produces the least prediction error for majority (3 out of 4) of the models. This is because sliding window method only learns the recent trend of the time series and predicts future values based on the trend. This finding is supported by Vasconcelos (2017) in which he stated that fixed rolling window (also known as sliding window) produces lower prediction error than expanding window. Vasconcelos (2017) justifies by proving the null hypothesis saying that both models are similar is rejected, concluding sliding window is better than expanding window.

In terms of model comparison, it was observed that boosting models (AdaBoost and XGBoost) are better at prediction by having lower RMSE than bagging and random forest. This is because boosting models learn the time series trends in sequential manner, whereby both model applies the concept of penalties for each error made by previous models. As a result, boosting models learned and predicted better than bagging and random forest. This outcome is also similar to Weng et. al (2018) in which they also found out that boosting model outperformed other 6 models including Random Forest and Bagging for macroeconomics variables. This shows that boosting model is the best model regardless of data splitting methods used.

To be specific, when both data splitting methods and model algorithms are taken into account in choosing the best technique for RPC prediction, it is determined that boosting model (typically XGBoost) with sliding window is the best technique for RPC prediction. In order to improve the models’ prediction, these models were optimised during Model Optimisation stage.

Table 3 Model prediction errors (RMSE) during Model Development

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Splitting Methods | Bagging | Random Forest | AdaBoost | XGBoost |
| Expanding Window (EW) | 7138.13 | 7057.10 | 5982.15 | 5837.29 |
| Sliding Window (SW) | 7036.21 | 6554.78 | 6530.31 | 5693.98 |
| Non-Random Percentage Split (NRS) | 27585.97 | 24905.06 | 24387.61 | 22622.79 |

### 3.2.4 Model Optimisation

After the models were successfully developed during Model Development stage, their prediction accuracy was further improved by optimising their window size, percentage split ratio, and feature selection. In particular, the purpose of this stage is to determine the best configurations of data preparation methods to improve prediction accuracy of the developed models. Feature selection was done by iterating the number of significant attributes from 1 until all of the attributes are included. Meanwhile for window size, they are iterated from 1 until window size = 10, and percentage split was iterated from 60:40 until 90:10 ratio.

For expanding window, all models’ prediction error was optimised by selecting the best attributes via feature selection during model training. While for sliding window, all models were optimised by selecting the best attributes via feature selection together with the best window size. Similar process was done for percentage split for both random and non-random split in which all models were optimised by selecting the best attributes via feature selection together with the test size. Figure 9 shows the recommended configurations for RPC prediction using Random Forest (sliding window) associated with the heatmap generated. Other heatmaps generated for other algorithms and data preparation methods are in figure 10-12 below.

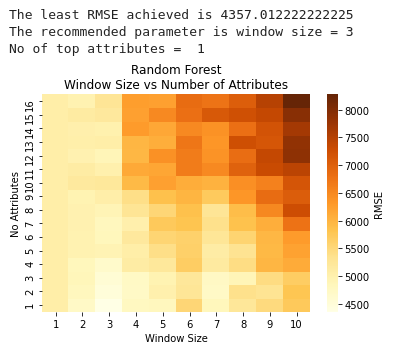


Figure Recommended configuration of RPC prediction by Random Forest (Sliding Window)

Referring to Figure 10-12, it was observed in general, that there are patterns in prediction errors depending on the data split method used. Figure 10 shows that expanding window shows a significantly higher error for each 4 incremental window sizes. Meanwhile in Figure 11, sliding window displays lower prediction error when small window sizes were applied (window size = 1 until 3). Lastly, Figure 12 displays that non-random percentage split performed better when 75:25 data split ratio was used together with a maximum of 5 significant attributes.

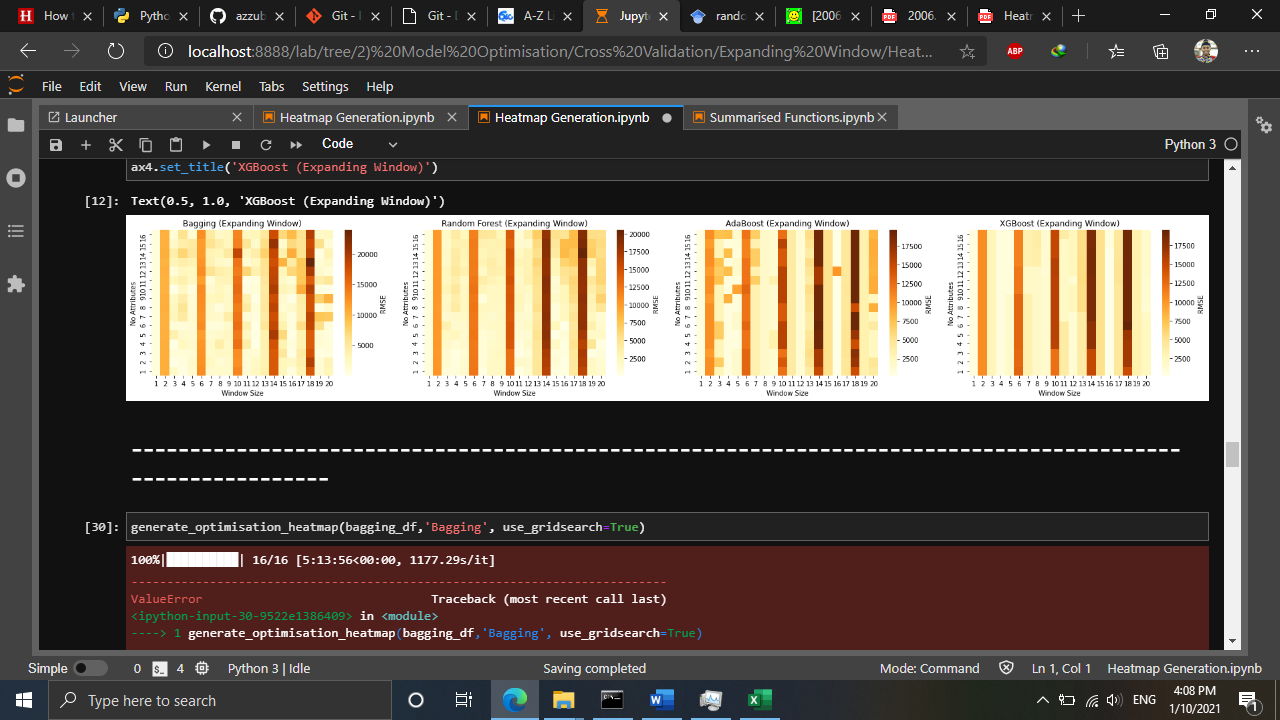


Figure Generated heatmap using Expanding Window

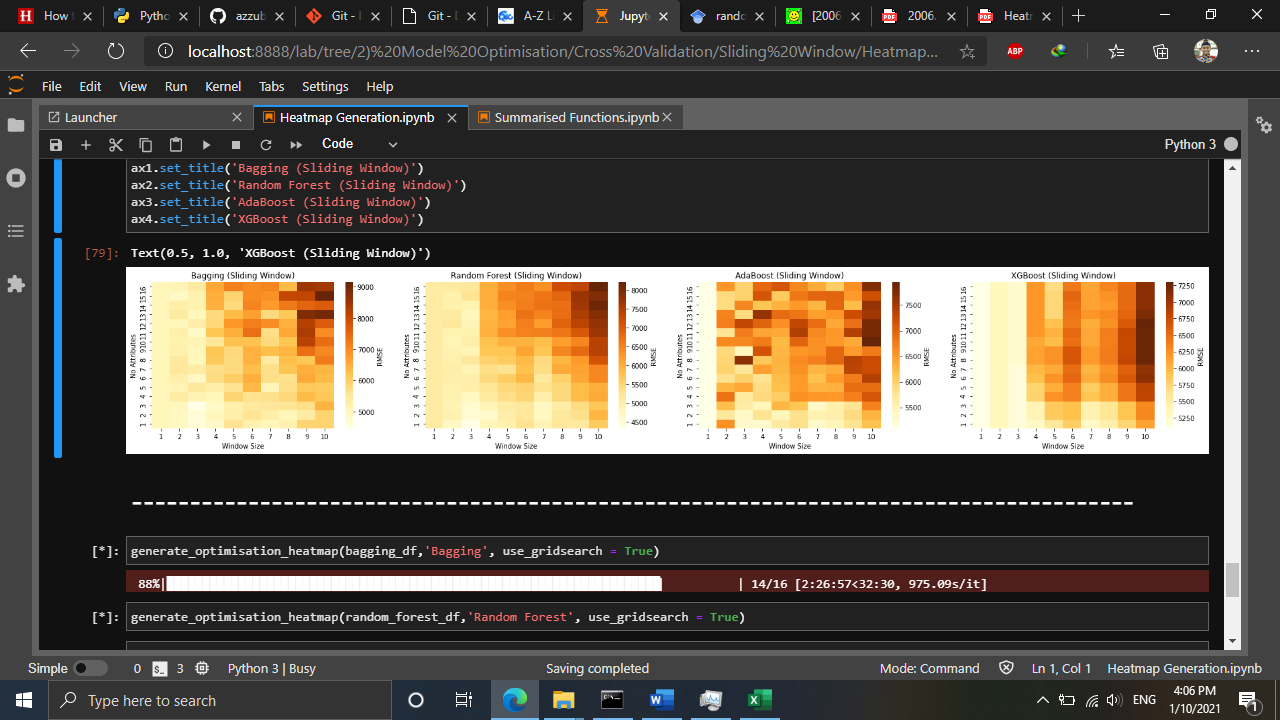


Figure Generated heatmap using Sliding Window

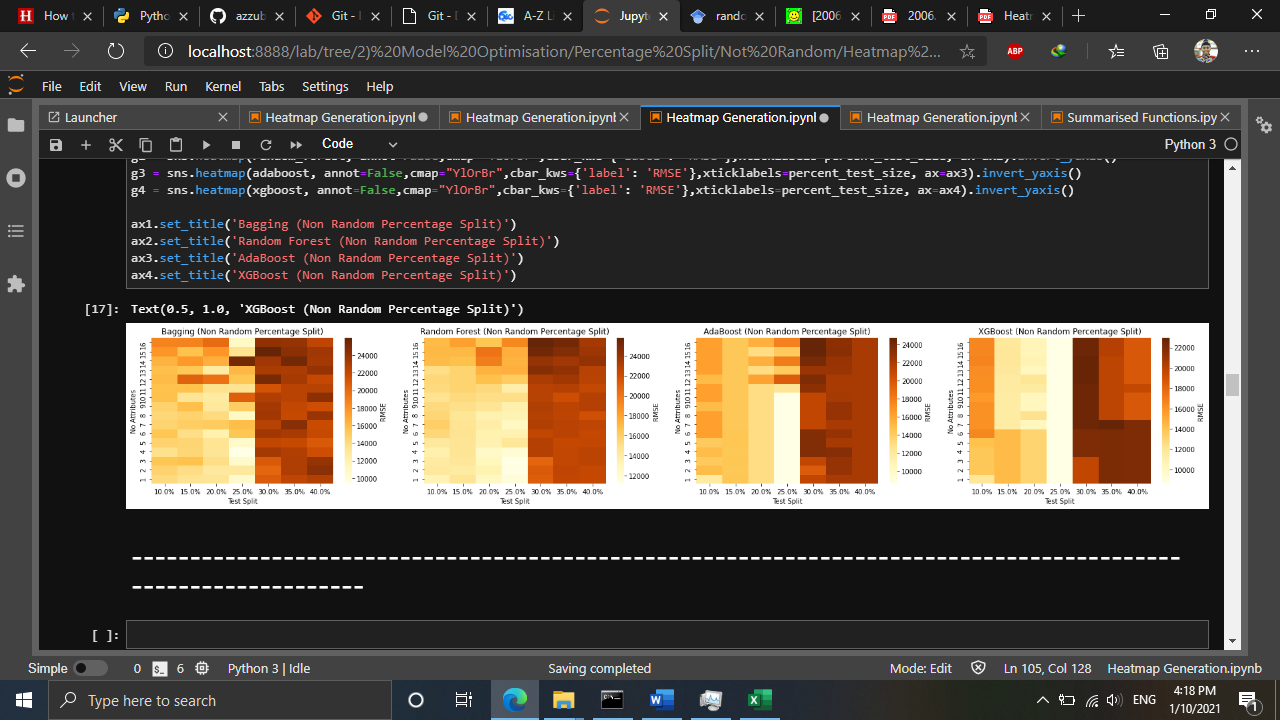


Figure Generated heatmap using Non-Random Percentage Split

## 3.3 Model Evaluation

After discussing Model Optimisation, the best configurations were evaluated during Model Evaluation stage. By taking the best configurations for each algorithm and data splitting methods, the results were illustrated in bar charts in Figure 13.

Figure Comparison of RMSE before and after model optimisation

Based on Figure 13, it was observed that all models’ prediction error was reduced after model optimisation indicating that model optimisation is very effective in improving models’ prediction accuracy. The key to this improved prediction accuracy is feature selection, and data splitting size. Here, the feature selection which is done by sklearn’s (a python library for machine learning tasks) SelectKBest algorithm ranks each attribute by scoring them based on their trend similarity with the target attribute (RPC) (Shaikh, 2018). Thus, attributes with higher similarity pattern have higher rank score. In contrast, when more attributes are included for model training, attributes with low rank scores are considered as noise that increases the models’ prediction error.

Secondly, data split size plays a major role in models’ prediction accuracy. Currently, there are no recommended data split size published in literatures as different models have different accuracy towards different data split size. If the data split is 50:50 (Training:Testing), the model will be underfitted. Otherwise, if the data split is 90:10, the models will be overfitted (Koehrsen, 2018). In order to improve prediction accuracy, it is important to avoid model underfitting and overfitting. Similar concern is also considered for windowing methods in which window sizes must also be optimised to prevent model underfitting and overfitting. Hence, this signifies that the best data split size is crucial to reduce model prediction errors.

Similar observation was stated by Seyedzadeh et al. (2019) and Li et al. (2018) in which both concluded that model optimisation significantly improved their forecasting result in modelling building energy consumption, and wind speed respectively. Therefore, this justifies the importance of feature selection, and data split for enhancing model prediction accuracy.

Moreover, in terms of data splitting methods, similar result was observed before and after model optimisation whereby majority of the models (3 out of 4) produced the least RMSE when sliding window was applied. This strengthens the findings of sliding window as in Model Development stage and finalises the best data splitting method for RPC prediction.

Furthermore, in terms of model comparison, it was identified that RF with sliding window produced the least RMSE after model optimisation. This contradicts with the models’ performances during model development whereby XGB was the best. This is because after considering the best window sizes and number of significant attributes to be input into the RF, training set with lesser noise was trained, thus producing a better RF model.

Finally, based on the results from Figure 10, it is determined that the best data splitting method and machine learning algorithm for RPC prediction is RF with sliding window. Other than that, it was also identified that non-random percentage split is not an effective method for data splitting of time series due to its high RMSE for all models. Therefore, all models and data splitting methods except non-random percentage split were considered for comparison with statistical as illustrated in Figure 45 and their averaged values is tabulated in Table 5.

Figure Comparison of RMSE between machine learning with statistical methods

Table 4 Averaged RMSE of machine learning models with respect to windowing techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Splitting Method | Bagging | Random Forest | AdaBoost | XGBoost |
| Expanding Window | 5,748.12 | 5,644.75 | 5,962.00 | 5,692.29 |
| Sliding Window | 6,140.59 | 5,472.99 | 5,819.56 | 5,692.35 |

Table 5 Averaged RMSE of statistical models

|  |  |  |  |
| --- | --- | --- | --- |
|  | UECM | MIDAS | MFVAR |
| RMSE | 1,663.40 | 1,236.63 | 9,577.90 |

Based on Figure 15, Table 4 and 5, it was determined that MIDAS produced the least RMSE compared to other statistical methods, and machine learning techniques. This shows that MIDAS is indeed an extremely powerful statistical technique for macroeconomics forecasting especially for RPC predictions. In fact, MIDAS is the most widely used statistical model for economics and finance for time series forecasting (Ghysels et al. 2016; Ferrara, 2012; Ghysels et al., 2004). This shows that MIDAS is the most reliable model for forecasting time series of RPC. The second-best model is UECM, another statistical model that performs best with economics data particularly for prediction (Sa’ad et al., 2018; Pesaran et al., 2001). The third-best model is RF with sliding window from the previous Model Optimisation stage. This comparison shows that the optimised random forest model is not yet at par with the current statistical methods, suggesting the machine learning for further optimisation to improve their prediction accuracy.

**3.3.1 Data Forecasting**

Regardless of machine learning models’ higher prediction error than statistical methods, the best machine learning model with the best data splitting method which is Random Forest with sliding window was used for data forecasting. Figure 12 below displays the prediction results along with the actual RPC plots from 2015 until 2020.

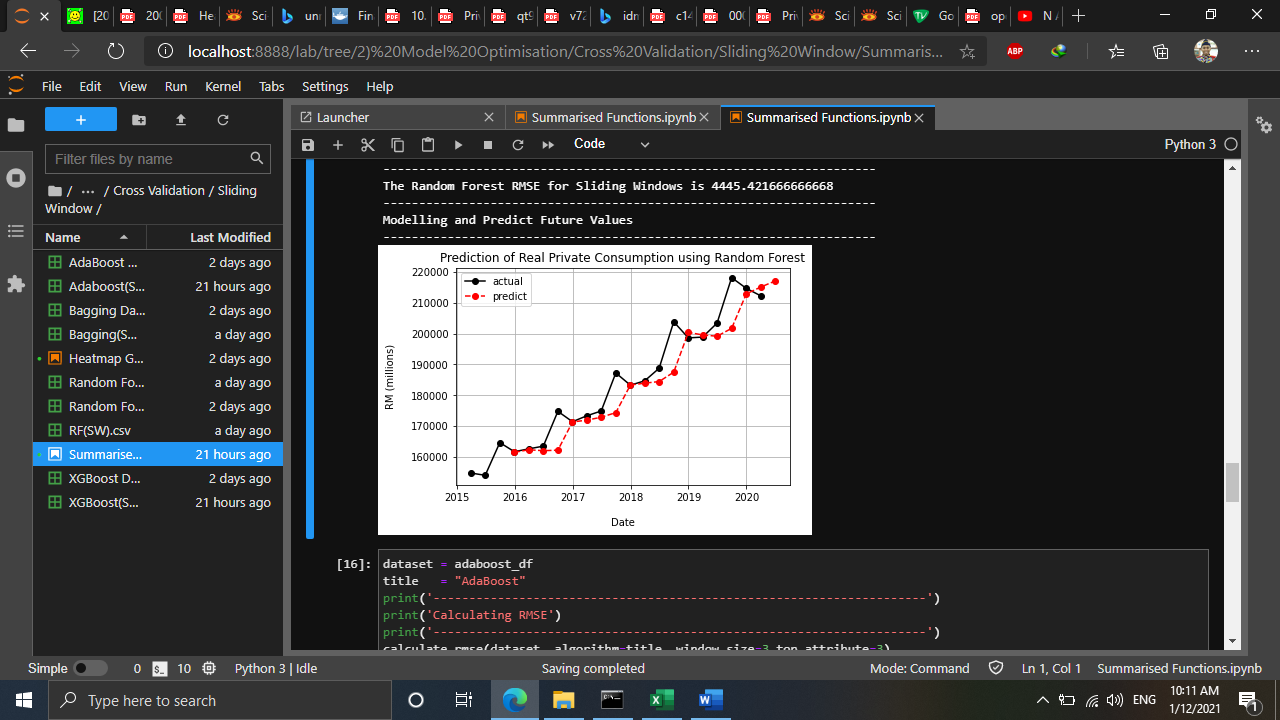


Figure RPC prediction using Random Forest with sliding window

Based on Figure 15, random forest shows a good prediction results throughout the time series except for the fourth quarter of each year. This describes that Random Forest is considerably good for Real Private Consumption predictions. Despite that Random Forest is still not yet better compared to MIDAS and MFVAR, Random Forest shows a potential to surpass statistical methods if it is further optimised. This is because Gonzalez et al. (2019) has demonstrated that Random Forest able to surpass other statistical models (Maehashi & Shintani, 2020; Chen et al., 2019; Tyralis, 2017). Thus, Gridsearch will be considered in future works.

In particular, random forest forecasted that the future value for RPC in the second quarter of 2020 is RM 217055 million. When this value is compared with the actual value (RM 16576 million), the predicted value was completely deviated away, and this is due to outlier factor which is the COVID-19 pandemic. Hence, to overcome this, more data is required particularly data related to COVID-19 so that the model learn the significance of COVID-19 to RPC.

## 3.4 What have been achieved and not achieved

Throughout the practicum, all of the objectives were successfully achieved in which the best machine learning algorithms for real private consumption were identified, developed, and evaluated. However, it was expected that the tree-based ensemble models to have a better prediction, instead, the developed models needed to be optimised much more for them to become better than statistical models. As of now, optimisation using grid search is already in progress but due to time constraint and bug fixing, it is still not achievable to be done and reported yet. In future, we will finish optimising using grid search and add this to the current model optimisation.

## 3.4 Challenges and Solutions

Various challenges faced during this project. At first, the first challenge was reviewing the best algorithms for RPC forecasting. Since machine learning algorithms are very new in macroeconomics forecasting, not many publications were found in comparing between machine learning with statistical models. Only a few publications studies in this comparison and most of them are recently published ranging from 2019-2020. The solution to this challenge was to keep searching related publications with the right keyword. Also, rephrasing the keyword helps a lot.

The second challenge is python coding during model development and optimisation. Despite that python is a dynamic programming language for Data Science tasks, as a beginner in python, it requires a lot of time to develop the models and fix the coding bugs. Plus, much time has been spent on model optimisation which would be faster to be run on gaming computers. Since much time was expected to be spent on these stages, the solution is to start the project as early as possible to utilise the remaining time effectively despite of lack of hardware resources.

The final challenge is to discuss the model evaluation. At first, it was hypothesised that machine learning models would surpass current statistical models as they are one of the elements of artificial intelligence, however, the outcome of this project revealed otherwise, strengthening its position as the most reliable methods for RPC predictions. The solution to this is to further optimise the machine learning models using gridsearch, and to try other feature selection algorithms other than SelectKBest, such as SelectFromModel, and mutual\_info\_regression.

## 3.5 Practicum Experience Applicability from Class

Practicum has sure taught me a lot on the practical side of Data Science. It has shown me that Data Science courses taught in class is just the foundation of what has been practiced in real world. Subjects taught in class such as Principles and Practices of Data Science (CDS 501), Machine Learning (CDS 503), and Predictive Business Analytics (CDS 512) indeed helped me a lot in providing me with the basis of Time Series Forecasting using Machine Learning techniques. Having those subjects as the foundation, practicum has brought me to a whole another level in which those foundations are essential to be familiarised and mastered.

Principles and Practices of Data Science has taught me the basics of data science general processes and some statistical knowledges. Without having these in mind, flowchart of this project and overall practicum would be very unorganised as there are many possible ways to deal with data. Also, this subject revealed to me that data science could not get away from statistics. Things that statistics could not do can be done using machine learning. Even some of the machine learning algorithms require some statistical knowledge such as Naïve Bayes, and Linear Regression. Thus, Principles and Practices of Data Science indeed helps.

Furthermore, as I am working on time series data, Predictive Business Analytics helped me a lot on working with time series. This subject gave the idea of using windowing techniques for time series using machine learning algorithms. Despite this subject taught me the basics of time series forecasting using both statistical (ARIMA) and windowing techniques, it is already enough, and it is up to me to explore these techniques even more. Algorithms used in this project were also taught in class during Machine Learning. This subject provided me on the theoretical concepts of algorithms and helped me on justifying their significance to be used in this project.

Also, we also applied knowledges from other subjects for tasks not related to this project such Text and Speech Analytics, Consumer Behaviour and Social Media Analytics for Research and Development purposes. Overall, this practicum indeed made me to apply the knowledges we learned from class. If in class we apply the techniques and algorithms using example dataset, during practicum, we applied it similarly but with real dataset published by DoSM and guided by both mentor and supervisor. Hence, we would like to express our gratitude for these experiences.

## 3.6 Observations during Practicum related to Professional and Operational Issues

During practicum, it was observed that data privacy is a crucial concern when working with big companies such as banks, private companies and even government sectors. Local companies such as DataMicron is trusted by government sectors due to its local workers who are local Malaysian citizens. In comparing with other international companies that also operating on IT consultation, they are less trusted due to the data privacy issue between Malaysia and other country. Also, when we were having a consultation session with the client, they even asked about the company’s staff citizenship. Again, this shows data privacy is a serious matter in government sectors. In addition, as a Universiti Sains Malaysia (USM) student who works on one of DataMicron’s proof of concept project, all of us who are in the same team working on this project are needed to sign the Non-Disclosure Agreement (NDA) to prevent data leakage to irresponsible individuals.

In terms of operational issues, throughout practicum, it was observed that DataMicron has a similar project flow as other IT consultant companies that implements the waterfall model (Figure 16). This model is much similar to the consultation phases learned during Research, Consultancy, and Professional Skills. During the first consultation phase, the company demonstrated their background and products to their clients. Such products are data warehouse, and business intelligence tool (DataMicron InstaBI), data management and cleansing tool (EZ Data) and data science studio (DataMicron Foresight). When the client is interested in these products to solve their current problems, the second consultation phase will be held in which they will require DataMicron to held the Proof of Concept session to prove that DataMicron is capable of solving their problem using a sample of their dataset. If the client agrees to continue towards implementation, software installation and training will be conducted. After the client is satisfied with the consultancy, they will sign the User Acceptance Test and the consultancy closes. Figure 17 illustrates the operation flow of DataMicron for consultancy services.

Diagram

Description automatically generated

Figure Waterfall Model of DataMicron project management methodology

Figure DataMicron’s Flow of Project Management

# CHAPTER 4 CONCLUSION AND LESSON LEARNED

## 4.1 Main Conclusion

Overall, this project demonstrated the applicability of time series forecasting using machine learning techniques. Firstly, during determination of the best machine learning models for the first objective, the literature review section revealed that tree-based ensemble models were the best machine learning for forecasting macroeconomics attributes such as real private consumption. Secondly, during model development for the second objective, all of the tree-based ensemble models were successfully developed and optimised by determining the best number of significant attributes, and number of window sizes to be input into the models. Lastly, during model evaluation for the last objective, Figure 14 shows that the models were successfully optimised by having lower RMSE, and the best model is Random Forest with sliding window. However, Figure 15 displayed that the current machine learning models were not able to surpass the current statistical models yet as much more optimisation is needed. Despite that, machine learning models have the potential to surpass statistical model if more optimisation is applied.

## 4.2 Lesson Learned

Throughout the practicum, I experienced only a glimpse of real data science projects. Having almost 100 contact hours of practicum proved that I am a very new to the real practical data science in the real world. Data science taught me that there are much more lessons to be learned and I will definitely embrace and practice those lessons each day. Here, I would like to emphasize on three main lessons learned throughout my practicum at DataMicron company.

### 4.2.1 Have a basic knowledge of client’s domain

Being involved in economics domain really shocked me as I never have basic knowledge on economics. It took me almost one month to understand the basics of economics to really understand each attribute given by client. By really understand these attributes, I can add Feature Engineering section to flowchart to further process my dataset before model development stage.

### 4.2.2 Master communication skills to engage with surrounding people and clients

Communicating with clients is the most important element in consultancy process. Small miscommunications may lead to terrible implications afterwards. During my practicum, I was lucky to have a very understanding, friend-like, and supportive mentor. Since he also has background in data science, he knows the basic data science processes, algorithms, and how to deal with clients, and managements. I was inspired by him by the way he communicates with people around him regardless of their position in the company. When the management asked about our team’s progress on the POC project, he always has the best answer which is not too technical and easily understood by the management. I am gradually learning his communication skills and will master the communication skills like my mentor in future.

### 4.2.3 Possess storytelling skills

Data cleansing, analysis, and prediction are all technical skills a data scientist must master and apply behind the scenes. However, when it comes to delivering data insights, and predictions, storytelling is the utmost important skill a data science must have to tell the storyline of the project outcome. Although the storytelling was done by my mentor, I observed that he possesses an excellent storytelling skill when he delivered our POC outcome to the client. If I were to put myself in his shoes, I would surely become nervous and run out of idea in front of Ministry of Finance crowds. This shows that as a data scientist, we must train ourselves to tell the outcome of data science projects in form of storytelling so that clients can easily understand our approach. In brief, storytelling is everything for a data scientist in front of his clients.

## 4.3 Future Works

As discussed in chapter 3, predictions using machine learning models is still not yet capable of surpassing statistical models. Therefore, in future, feature selection algorithms will be included as part of the model optimisation stage. Now, feature selection using SelectKBest was applied, and in future, all of the feature selection algorithms will be considered. Furthermore, in future, another model optimisation element will be added which is the gridsearch, an algorithm which selects the best model parameters depending on the listed range of parameters. By having gridsearch, models are expected to be further optimised and have better prediction results.

# 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=dUtRR1JYWjk

2TEJha1BrZml0REY4UT09

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.

Ferrara, L., & Marsilli, C. (2013). Financial variables as leading indicators of GDP growth: Evidence from a MIDAS approach during the Great Recession. *Applied Economics Letters*, *20*(3), pp. 233-237.

Ghysels, E. (2016). Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, *193*(2), 294-314.

Gonzalez-Vidal, A., Jimenez, F., & Gomez-Skarmeta, A. F. (2019). A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. *Energy and Buildings*, *196*, 71-82.

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

Koehrsen, W. (2018). Overfitting vs Underfitting: A Complete Example. *towards data science*. Retrieved on Jan. 2, 2020 from https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

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

Li, C., Xiao, Z., Xia, X., Zou, W., & Zhang, C. (2018). A hybrid model based on synchronous optimisation for multi-step short-term wind speed forecasting. *Applied Energy*, *215*, pp. 131-144.

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

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of applied econometrics*, *16*(3), pp. 289-326.

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

Sa'ad, S., Dahoro, D., & Ahmed, M. N. (2019). Accounting for non‐economic factors in demand for transportation fuels: a comparative study of South Korea and Indonesia. *OPEC Energy Review*, *43*(1), pp. 50-66.

Seyedzadeh, S., Rahimian, F. P., Rastogi, P., & Glesk, I. (2019). Tuning machine learning models for prediction of building energy loads. *Sustainable Cities and Society*, *47*, 101484.

Shaikh, R. (2018), Feature Selection Techniques in Machine Learning with Python. *towards data science.* Retrieved on Jan. 1, 2021 from https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

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

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms*, *10*(4), 114.

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

Vasconcelos, G. (2017). Formal ways to compare forecasting models: Rolling windows. *R-Bloggers*. Retrieved on January 1, 2021 from: https://www.r-bloggers.com/2017/11/formal-ways-to-compare-forecasting-models-rolling-windows/

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

Weng et al., (2018). Macroeconomic indicators alone can predict the monthly closing price of major U.S. indices: Insights from artificial *intelligence*, time-series analysis and hybrid models. *Applied Soft Computing*, 71, pp. 685-697.

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