



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 is 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 : Muhammad Azzubair bin Azeman

Date : 27 January 2021.....

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 his assistance and dedicated involvement 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 are 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 binti Hanafi Chia, Ruzbihan Hadi bin Ahmad Bakhtiar, Muhamad Haris bin Idris, Faiz bin Amirul Fuad and to all my supportive friends, thank you for helping me 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

Since the 20th century, prediction of real private consumption (RPC) using statistical techniques has been prevalent and widely applied by economists. Recently, machine learning techniques has emerged and much more economic research has been done to compare the prediction performance between machine learning and statistical techniques. The Ministry of Finance is currently having a similar situation in determining whether to retain the current statistical techniques or machine learning techniques for implementing RPC prediction. Thus, this project is carried out to evaluate the performance of machine learning techniques and statistical methods for time series of RPC prediction. This project also aims to evaluate the performance of selected machine techniques for time series RPC prediction. The result of this evaluation shows that the statistical method still produces better prediction performance in comparison with machine learning techniques. Based on this finding, the best machine learning algorithm that is Random Forest. Hence, more research is needed to further optimise the machine learning model for better prediction performance.

ABSTRAK

Sejak abad ke-20, ramalan penggunaan peribadi sebenar (PPS) menggunakan teknik statistik sudah menjadi kebiasaan dan banyak digunakan secara meluas oleh para ahli ekonomi. Baru-baru ini, teknik pembelajaran mesin telah muncul dan lebih banyak kajian ekonomi lin telah dilakukan untuk membandingkan prestasi ramalan antara pembelajaran mesin dengan teknik statistik. Kementerian Kewangan kini mengalami situasi yang sama dalam menentukan sama ada akan mengekalkan teknik statistik semasa atau teknik pembelajaran mesin untuk ramalan PPS. Oleh itu, projek ini dijalankan untuk menilai prestasi teknik pembelajaran mesin dan kaedah statistik untuk ramalan siri masa PPS. Projek ini juga bertujuan untuk menilai prestasi teknik mesin terpilih untuk ramalan siri masa PPS. Hasil penilaian ini menunjukkan bahawa kaedah statistik masih menghasilkan prestasi ramalan yang lebih baik berbanding dengan teknik pembelajaran mesin. Berdasarkan penemuan ini, algoritma pembelajaran mesin terbaik adalah *Random Forest*. Lebih banyak penyelidikan diperlukan untuk mengoptimumkan lagi model pembelajaran mesin untuk prestasi ramalan yang lebih baik.

TABLE OF CONTENTS

| | |
|---|--------------|
| DECLARATION..... | i |
| ACKNOWLEDGEMENTS..... | ii |
| ABSTRACT..... | iii |
| ABSTRAK..... | iv |
| LIST OF TABLES | vii |
| LIST OF FIGURES | viii |
| LIST OF ABBREVIATIONS AND SYMBOLS | ix |
| CHAPTER 1 INTRODUCTION & RELATED WORKS | 1 |
| 1.1 Background of Practicum Company..... | 1 |
| 1.2 Background of Domain..... | 2 |
| 1.3 Problem Statement..... | 3 |
| 1.4 Research Question | 3 |
| 1.5 Objectives | 4 |
| 1.6 Benefits of the project..... | 4 |
| 1.7 Related Works | 4 |
| 1.7.1 Problem Domain | 4 |
| 1.7.2 Data Science & Analytics techniques | 5 |
| 1.7.3 Data Science & Analytics tools..... | 7 |
| CHAPTER 2 RESEARCH METHODOLOGY | 9 |
| 2.1 Activities plan and Gantt Chart..... | 9 |
| 2.2 Data Science Project Lifecycle | 10 |
| 2.3 Problem Analysis | 11 |
| 2.3.1 Initial Suggestions..... | 11 |
| 2.3.2 Specific case to be addressed..... | 11 |
| 2.3.3 Exploratory Data Analysis | 12 |
| 2.3.4 Implementation of Machine Learning Techniques for RPC Prediction..... | 14 |
| 2.3.4.1 Data Cleansing..... | 15 |
| 2.3.4.2 Data Preparation..... | 16 |
| 2.3.4.3 Model Development..... | 18 |

| | |
|---|-----------|
| 2.3.4.4 Model Optimisation | 19 |
| 2.4 Proposed Solution to MoF | 20 |
| 2.5 Justification of Data Science Techniques and Tools | 21 |
| 2.6 Chapter Summary | 22 |
| CHAPTER 3 RESULTS AND DISCUSSIONS | 23 |
| 3.1 Model Evaluation..... | 23 |
| 3.2 Data Prediction..... | 26 |
| 3.3 What have been achieved and not achieved | 27 |
| 3.4 Challenges and Solutions..... | 27 |
| 3.5 Practicum Experience Applicability from Class..... | 28 |
| 3.6 Observations during Practicum related to Professional and Operational Issues..... | 29 |
| 3.7 Chapter Summary | 30 |
| CHAPTER 4 CONCLUSION AND LESSON LEARNED | 31 |
| 4.1 Conclusion | 31 |
| 4.2 Lesson Learned | 31 |
| 4.2.1 Have a basic knowledge of client's domain..... | 31 |
| 4.2.2 Master communication skills to engage with surrounding people and clients | 32 |
| 4.2.3 Possess storytelling skills..... | 32 |
| 4.3 Future Works..... | 32 |
| 4.4 Useful Data Science Pipeline and Theories | 33 |
| 4.5 Suggestion for Improvement of Practicum and its Preparations | 34 |
| 4.6 Integration of Practicum Experience with Student Coursework in DSA program | 35 |
| 4.7 Chapter Summary | 35 |
| REFERENCES | 36 |
| APPENDIX 1 | 40 |
| APPENDIX 2..... | 44 |
| APPENDIX 3 | 55 |

LIST OF TABLES

| | |
|---|----|
| Table 1 Description of Real Private Consumption and its indicators | 12 |
| Table 2 Model prediction errors (RMSE) during Model Development..... | 18 |
| Table 3 Summary of Literature Reviews | 21 |
| Table 4 Averaged RMSE of machine learning models with respect to windowing techniques.... | 25 |
| Table 5 Averaged RMSE of statistical models..... | 25 |

LIST OF FIGURES

| | |
|---|----|
| Figure 1 Gantt Chart of Project Consultancy and Practicum..... | 9 |
| Figure 2 Lifecycle of Data Science Projects..... | 10 |
| Figure 3 Overview of Real Private Consumption trends and its indicators..... | 13 |
| Figure 4 Flow Chart of RPC Prediction..... | 15 |
| Figure 5 Output of each process in Data Cleansing phase..... | 15 |
| Figure 6 Output of RPC indicators before (left) and after (right) Windowing process | 16 |
| Figure 7 Generation of horizontal windows via column transpose | 17 |
| Figure 8 Illustration of sliding window (left) and expanding window (right) method..... | 17 |
| Figure 9 Recommended configuration of RPC prediction by Random Forest (Sliding Window) 19 | |
| Figure 10 Generated heatmap using Expanding Window..... | 20 |
| Figure 11 Generated heatmap using Sliding Window | 20 |
| Figure 12 Generated heatmap using Non-Random Percentage Split..... | 20 |
| Figure 13 Comparison of RMSE before and after model optimisation..... | 23 |
| Figure 14 Comparison of RMSE between machine learning with statistical methods..... | 25 |
| Figure 15 RPC prediction using Random Forest with sliding window | 26 |
| Figure 16 Waterfall Model of DataMicron project management methodology..... | 30 |
| Figure 17 DataMicron's Flow of Project Management | 30 |

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 is 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 is 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 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 are 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 are foreign to Malaysians until these techniques are 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

Numerous studies have been done to compare 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 study have been done to compare model performances between machine learning algorithms and statistical techniques for time series prediction of real private consumption in Malaysia. Addressing the performances between statistical techniques and machine learning algorithms 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 is aimed at predicting time series of real private consumption using machine learning techniques. Thus, this led to determination of whether machine learning algorithms or statistical techniques is better for RPC prediction. This project proposes machine learning approaches for RPC prediction. That leads into the following research questions:

- Which machine learning model is the most suitable for RPC prediction?
- Which machine learning algorithm is better for prediction of real private consumption?
- What is the best machine learning model for RPC prediction?

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

1.5 Objectives

This project aims to propose machine learning algorithms as a new approach to improve time series prediction of real private consumption in Malaysia. Thus, this project will focus on predicting one of the econometric indicators which is real private consumption (RPC) based on data published by Department of Statistics Malaysia (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 techniques.
3. To evaluate the performance of RPC prediction models.

1.6 Benefits of the project

This is a consultation project between DataMicron Sdn Bhd 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 for RPC prediction using machine learning approaches for MoF. Thus, the MoF will use that information to optimise its financial planning and reduce its financial loss.

1.7 Related Works

This chapter contains the review of the finance domain particularly macroeconomic variables related to real private consumption. Then, established statistical techniques for RPC prediction, and the proposed machine learning algorithms and tools are presented.

1.7.1 Problem Domain

Since 2018, RPC has been the largest contributor 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 mainly due to the growth of employments and wages. This indicates that the impact of this sector has been significant throughout the overall annual RPC. Since RPC is the most significant attribute towards contribution to Malaysia's GDP, it is very crucial for RPC to be predicted as it reflects the most of Malaysia's GDP for MoF to make decisions on economy improvement.

1.7.2 Data Science & Analytics techniques

Statistical techniques such as ARIMA, and VAR have been reported and compared for Malaysia's economics forecasting (Razak, Khamis & Abdullah, 2017). These two models can be used for univariate time series forecasting. Both models can be used to predict economic indicators such as Currency in Circulation, Exchange Rate, External Trade, and Reserve Money for several periods ahead in future. The findings of this study have 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, 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. Hence, machine learning techniques have started to be used for economics prediction since 20th century. The earliest attempt is 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 predicted using machine learning models than linear models such as ARIMA. Similar observation is obtained by Dematos et. al (1996) in which they found Recurrent Neural Network (RNN) and Feedforward Neural Network (FNN) outperformed ARIMA in predicting Japanese yen/U.S. dollar exchange rate. In summary, nonlinear trend of economic indicators is more accurate to be predicted using machine learning.

A comparative study that was done by Kumar et. al (2018) compares machine learning performances between selected machine learning techniques for predicting stock market trend. Machine learning techniques used in this study are support vector regression (SVR), random forest (RF), k-nearest neighbour (KNN), naive bayes (NB), and SoftMax (SM). Interestingly, when large dataset (4500 entries) is input to the models, RF outperformed others by having the highest accuracy and f-measure followed by SVR. This indicates that RF and SVR are suitable models for predicting nonlinear trends of economic growths. In addition, RF is good for prediction due to its robustness against outliers (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 performances between machine learning algorithms with statistical techniques 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 the two models had the least RMSE percentage points of -0.56 and -0.43 respectively. In contrast, MARS and 4QMA are suggested to be avoided as they had significantly poor prediction performance compared to others.

Recently, Maehashi and Shintani (2020) also performed comparison study on several machine learning models in predicting macroeconomic variables. Interestingly, this 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, 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, 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 concluded that ensemble learning methods outperformed 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, their predictions would become better, more accurate, and robust against errors.

1.7.3 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 many available libraries for dealing with time series data and predictions. One of the popular libraries for time series forecasting is the statsmodels library. According to McKinney et al. (2011), this library provides many statistical models such as ordinary least squares, VAR, and ARIMA. Another library for time series prediction 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 pound sterling would rise against dollar and euro by ± 0.02 by end of 2019.

Another approach of time series prediction 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) explained about the sliding window in detail 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 produced as the output. In Python, this windowing feature is available in several libraries such as tslearn, cesium-ml, ts-fresh, and seglearn. Burns and Whyne (2018) compared these libraries features for time series prediction. The outcome of the comparison shows that all libraries provide prediction feature except tslearn. Besides this, seglearn library happens to be the only library with the most features such as sliding window, and compatible with machine learning models.

Another important library for time series prediction using machine learning models is the scikit learn library. This library contains most of the common machine learning models for data scientists. According to Hackeling (2017), scikit learn library provided linear models such as linear and multiple linear regressions, nonlinear models such as bagging, AdaBoost, RF, 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 prediction.

1.8 Chapter Summary

Background of practicum company has been described in section, followed by domain background. By having these information, problem statement of this project has been clearly defined, together with the research questions. To answer the questions, the objectives of this project has been clearly stated. Data science analytics techniques and tool have also discussed in this chapter. Next chapter will describe the Methodology of this project in details.

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 carried 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. 98 contact hours with the practicum company are recorded in a logbook (refer Appendix 1) and weekly meetings with supervisor are recorded to update about project progress.

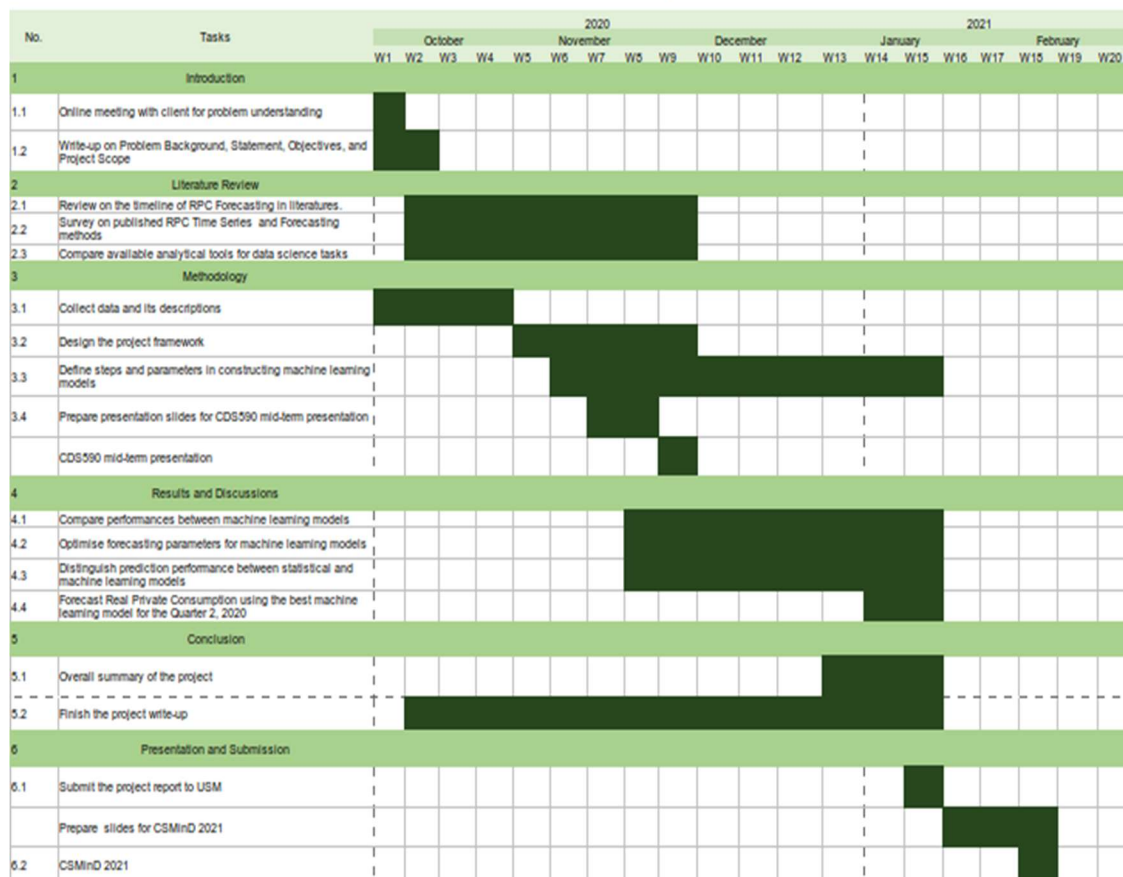


Figure 1 Gantt Chart of Project Consultancy and Practicum

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 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 are going through.

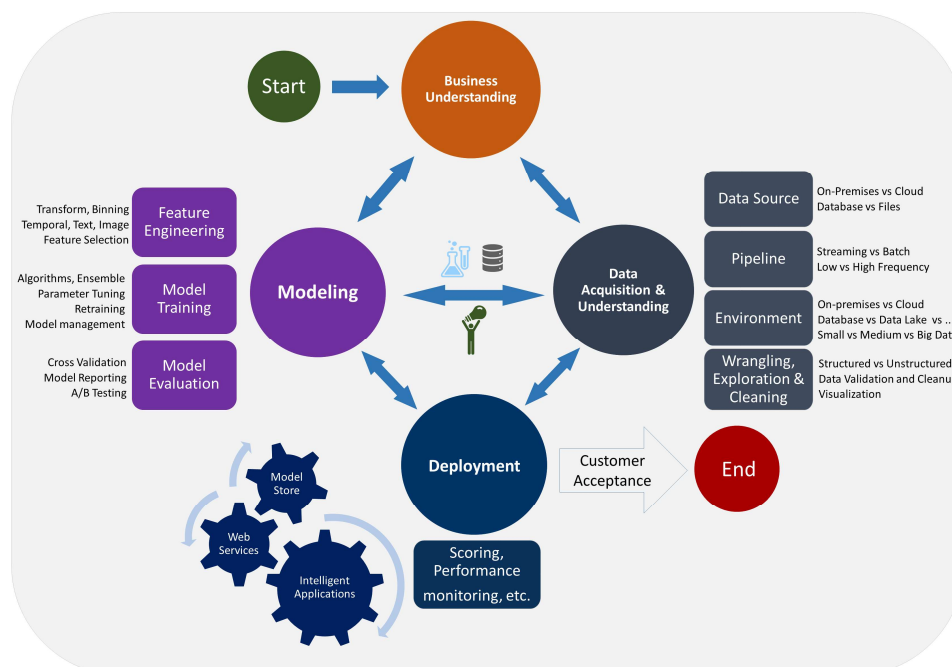


Figure 2 Lifecycle of Data Science Projects

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. 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 finally the proposed solution was presented to MoF during the Customer Acceptance stage.

2.3 Problem Analysis

Analysis of client's problem are conducted during Business Understanding stage. At this stage, a presentation slide by MoF describing on their problem has been given to the DataMicron for reference (refer Appendix 2). The purpose of the slide is mainly to describe their current statistical techniques and metrics used for RPC prediction and their predictions since 2018 until 2020. Based on this slide, there were several approaches suggested by MoF to improve RPC prediction. The listed suggestion is as follow. After went through the suggestion thoroughly with mentor, approach using machine learning algorithms were agreed to be used for this project.

2.3.1 Initial Suggestions

Suggestions listed below are prepared by the client to ease DataMicron in choosing the best approach. As a Data Science consultant company, approaches that would require domain expert is omitted, instead, approach related with DataMicron's expertise is agreed for the project.

- How to improve prediction of Real Private Consumption?
 1. Replacing current indicators with others which are more accurate to predict the movement of private consumption (RMSE comparison as a benchmark)
 2. Adding more error performance criterion (MAE, MAPE, MSE)
 3. Increase the number of indicators used (cost: overfitting the model)
 4. Implementing Machine Learning and Deep Learning approach
 5. Build Private Consumption Index
 6. Track higher frequency data in unstructured or semi structured form in order to get more regular forecast update

2.3.2 Specific 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 (refer Appendix 3) from MoF, an exploratory data analysis has been conducted to provide an overview of the variable distribution, information, and trends. In terms of Data Science Lifecycle, this step is categorised under Data Acquisition and Understanding. Firstly, the dataset is given in an excel file containing variables of real private consumption indicators as published in BNM (2016). These variables are collected from data published by DoSM, and BNM from 1995 until 2020. All numerical variables are 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 (CC) | Amount of RM millions lent by banks for loans in Consumption Credits |
| 5. | Loans disbursed to Wholesale & Retail Trade, Restaurant, & Hotels (LOWT) | Amount of RM billions lent 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 |

After describing each attribute, Python is used to be visualise the trends of each attributes.

Figure 3 shows the overall visualisation of RPC and its indicators.

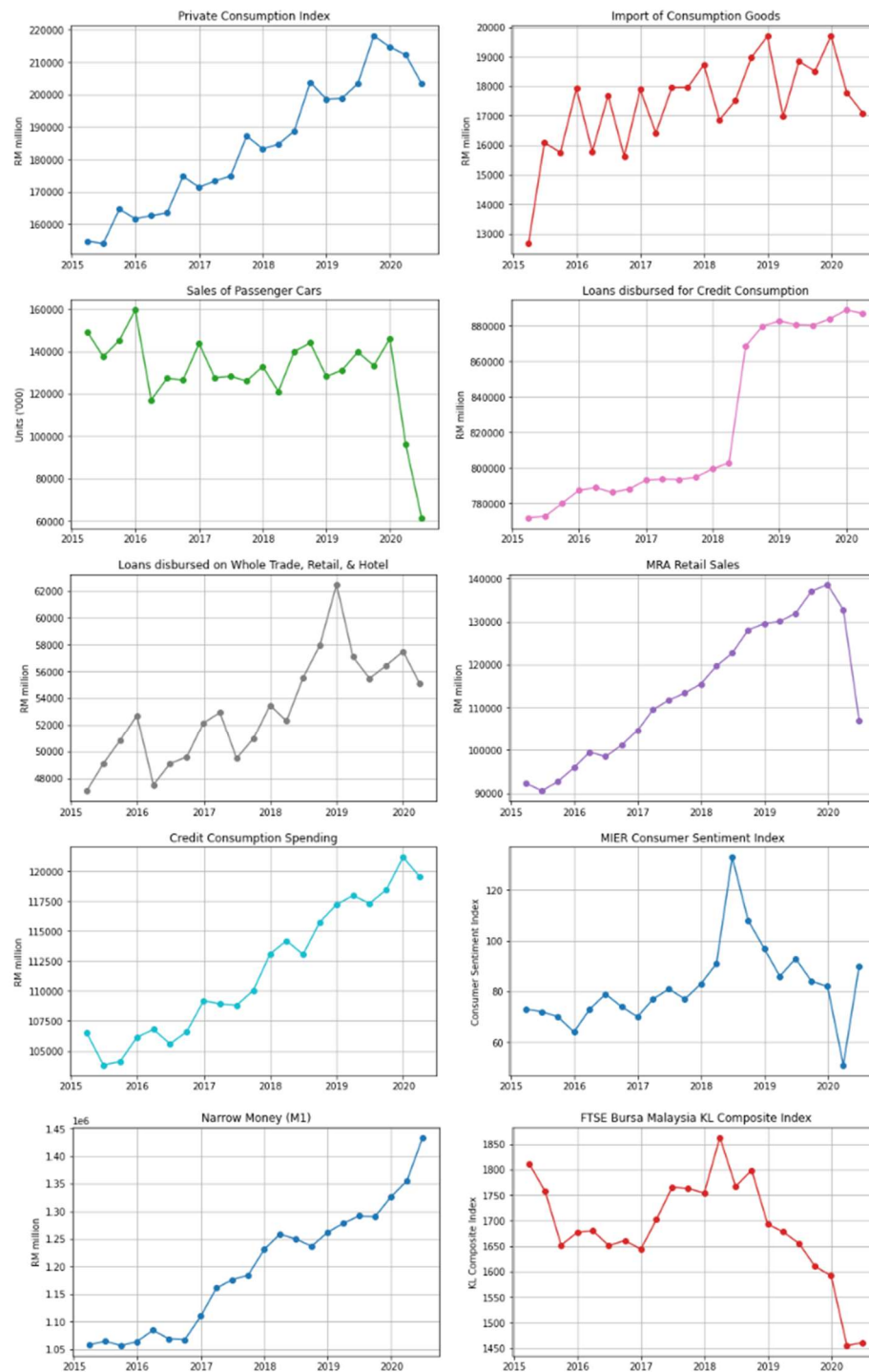


Figure 3 Overview of Real Private Consumption trends and its indicators

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 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 until the end of the second quarter of 2018. Beyond this quarter, pessimists started to outperform optimists due to the global challenges affecting economic growth in Malaysia (Rasid, 2019). In contrast, FBM displays an overall downward trend with some exceptions in 2018. This attribute is heavily influenced 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 Implementation of Machine Learning Techniques for RPC Prediction

RPC prediction using machine learning techniques has been 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 has been implemented as stated by Rasel et al. (2015). Flow chart of the proposed methodology is shown in Figure 4. The process flow of the final analysis and discussions are organised based on this flow chart.

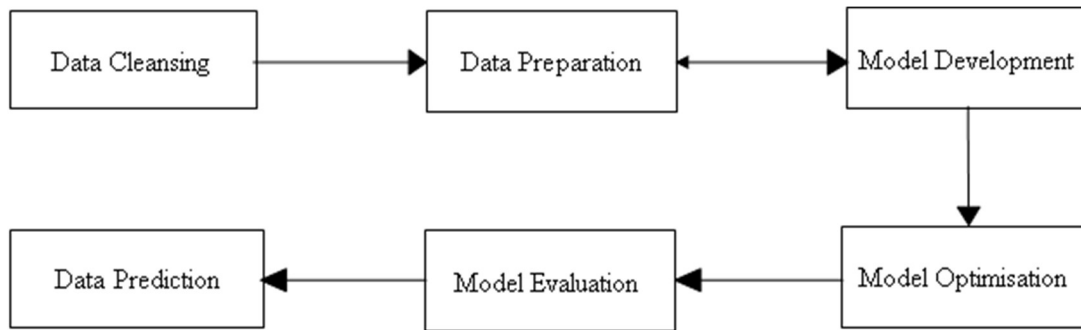


Figure 4 Flow Chart of RPC Prediction

2.3.4.1 Data Cleansing

Upon receiving dataset (excel file) from Ministry of Finance's Fiscal and Economics Division, the data is explored and cleaned using python's pandas library. Firstly, the timestamp of RPC indicators is not aligned with RPC which are collected monthly and quarterly respectively. Therefore, the indicators have been aggregated by sum and average depending on attribute type accordingly. Next, columns with zero values throughout the dataset have been discarded which reduced the overall attributes from 18 into 16 attributes. Thirdly, it is noticed during data exploration that the timestamp at which each attribute is begun to be collected are different. Therefore, dataset timestamp has been standardized by filtering it to be starting from date at which all attributes are collected simultaneously which is 1 January 2015. Figure 5 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 | 2036.700000 | 75641.0 |
| 1995Q2 | 1771.000000 | 2756.300000 | 2223.900000 | 81561.0 |
| 1995Q3 | 1775.800000 | 3022.800000 | 2618.600000 | 86385.0 |
| 1995Q4 | 1762.200000 | 3283.100000 | 2501.600000 | 71463.0 |
| 1996Q1 | 1882.600000 | 2090.800000 | 2330.400000 | 84606.0 |
| ... | ... | ... | ... | ... |
| 2019Q2 | 19154.941736 | 18837.769720 | 2501.836000 | 124764.0 |
| 2019Q3 | 19038.573194 | 18504.265237 | 2552.025000 | 142484.0 |
| 2019Q4 | 19217.321673 | 19704.177128 | 2602.998000 | 146849.0 |

Change Dataset Timeline

2015 - 2020

| | FS | ICG | TSP | SOM |
|--------|--------------|--------------|-------------|----------|
| Date | | | | |
| 2015Q1 | 12328.555125 | 12670.250538 | 1690.895000 | 109248.0 |
| 2015Q2 | 12275.627835 | 16086.134717 | 1701.273000 | 93418.0 |
| 2015Q3 | 12560.931648 | 15755.989512 | 2439.373000 | 86952.0 |
| 2015Q4 | 13161.240861 | 17918.012761 | 2790.716000 | 91184.0 |
| 2016Q1 | 13713.463269 | 15783.503816 | 1919.767000 | 106224.0 |
| 2016Q2 | 13513.860603 | 17687.022792 | 1802.066000 | 95922.0 |
| 2016Q3 | 13501.733745 | 15623.794846 | 2240.072000 | 95237.0 |
| 2019Q1 | 18652.156040 | 16967.177973 | 2278.060000 | 132716.0 |
| 2019Q2 | 19154.941736 | 18837.769720 | 2501.836000 | 124764.0 |
| 2019Q3 | 19038.573194 | 18504.265237 | 2552.025000 | 142484.0 |

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 |

Figure 5 Output of each process in Data Cleansing phase

2.3.4.2 Data Preparation

After cleansing the dataset, it is noted that initially there are no value for RPC indicators for second quarter of 2020 since they are future values that are not published yet. Therefore, in order to fill up the values of the RPC indicators, windowing method integrated with tree-based ensemble learning models has been used to predict the future values. In particular, the windows have been 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 future values of RPC indicators have been predicted using bagging, RF, and boosting models simultaneously. Figure 6 illustrates the windowing process before and future value predictions using Random Forest.

| Date | NM | FBM | MCF | MSW | EMP | MIER | PCI | 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 | 2019-03-31 | 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 | 2019-06-30 | 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 | 2019-09-30 | 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 | 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 | 2020-03-31 | 1355344.0 | 1455.0 | 1539.0 | 22728.0 | 45895.0 | 51.0 | 212257.0 |
| 2020-06-30 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | 2020-06-30 | 1287626.0 | 1516.0 | 1652.0 | 21880.0 | 45438.0 | 75.0 | NaN |

Figure 6 Output of RPC indicators before (left) and after (right) Windowing process

Furthermore, two types of data splitting methods are used for Model Development stage which are percentage split and time-based cross validation. For percentage split's initial run, the time series dataset has been splitted into 2 sets by 70:30 for training and testing sets. This splitting has been 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 have been 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 is set) and sliding throughout iterations. The output of data splitting methods has been recorded in Model Development stage. Figure 7 illustrates the concept of transposing a column into horizontal windows and Figure 8 displays the concept of sliding window in time-based cross validation.

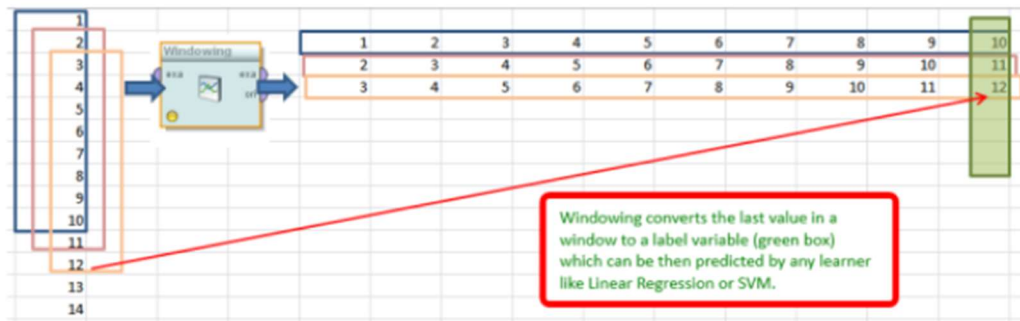


Figure 7 Generation of horizontal windows via column transpose

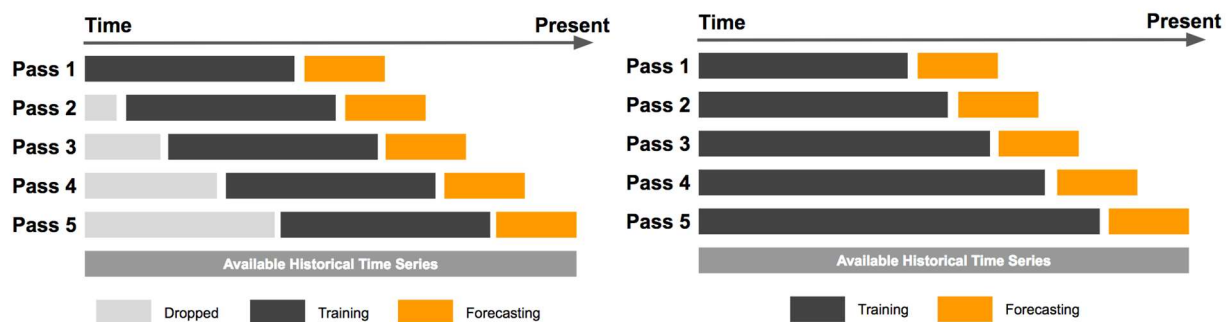


Figure 8 Illustration of sliding window (left) and expanding window (right) method

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 have been evaluated using RMSE evaluation metrics similar to the current evaluation metrics currently being used by MoF. Model training and testing are repetitively conducted together with model optimization to achieve the best model performance. During **model optimization** stage, two components are tested to produce the best model performance which are feature selection, data split ratios. The best attributes to be used in model training are determined using feature selection which is available in Python's scikit learn library while the best window size is identified using loop function. During model evaluation stage, model performances are evaluated Root Mean Squared Error (RMSE) which is similarly used by MoF for comparison. Lastly, future data of RPC for the second quarter of 2020 is predicted using the best evaluated model. The predicted value will be presented to MoF during Customer Acceptance of Data Science Project Lifecycle.

2.3.4.3 Model Development

At this stage, models are trained using multiple data preparation methods. Based on Table 2, it is 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 is 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 machine learning models including Random Forest and Bagging for macroeconomics variables. This shows 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 are optimised during Model Optimisation stage.

Table 2 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 |

2.3.4.4 Model Optimisation

After the models are successfully developed during Model Development stage, their prediction accuracy is 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 is 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 is iterated from 60:40 until 90:10 ratio.

For expanding window, all models' prediction error is optimised by selecting the best attributes via feature selection during model training. While for sliding window, all models are optimised by selecting the best attributes via feature selection together with the best window size. Similar process is done for percentage split for both random and non-random split in which all models are 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.

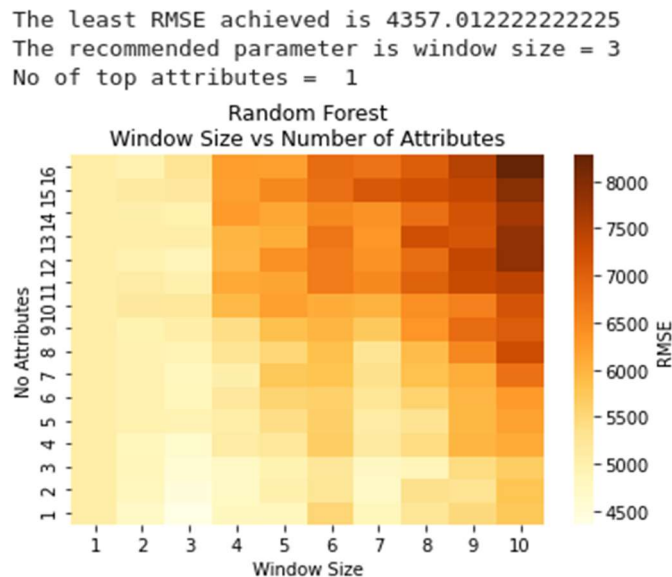


Figure 9 Recommended configuration of RPC prediction by Random Forest (Sliding Window)

Referring to Figure 10-12, it is 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 are applied (window size = 1 until 3). Lastly, Figure 12 displays that non-random percentage split performed better when 75:25 data split ratio is used together with a maximum of 5 significant attributes.

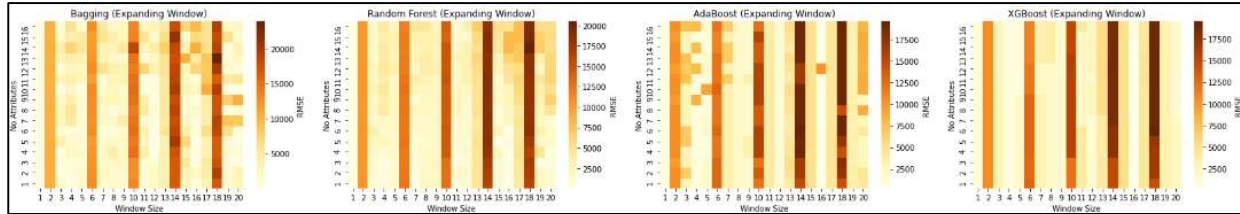


Figure 10 Generated heatmap using Expanding Window

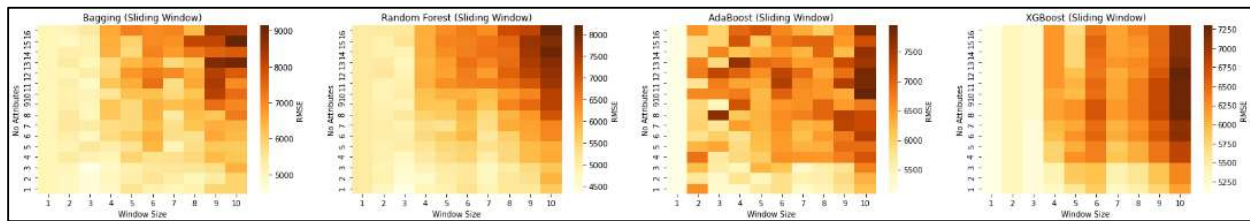


Figure 11 Generated heatmap using Sliding Window

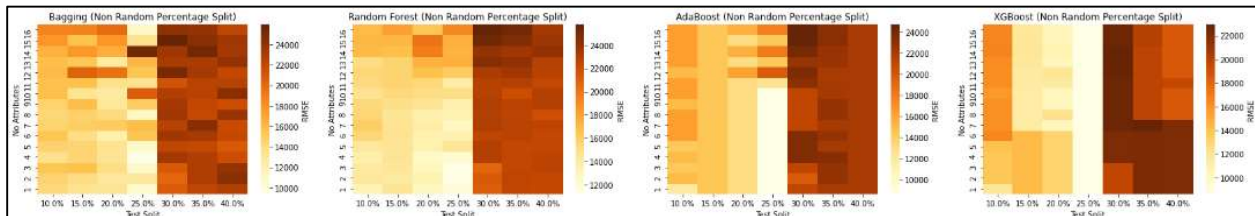


Figure 12 Generated heatmap using Non-Random Percentage Split

2.4 Proposed Solution to MoF

After evaluation stage of Machine learning techniques has been done, 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 then be evaluated by the client whether it is acceptable or not. In addition, the predicted value also will be compared with the actual RPC value after it is officially published by DoSM.

2.5 Justification of Data Science Techniques and Tools

Based on a comparative study done by Kumar et al. (2018), Chen et al. (2019), and Maehashi and Shintani (2020), found out that tree-based ensembled learning such as Bagging, Random Forest, and Boosting performed are the best for models predicting macroeconomic indicators. Table 3 summarises the literature reviews findings. Throughout literature reviews, it is found out that ensemble learning models based on regression trees are the best models to be used for RPC prediction. The reason is because these models 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 are selected to be implemented for RPC prediction. Hence, it is justified that Bagging, Boosting, and RF are selected due to their high accuracy for economics predictions.

Table 3 Summary of Literature Reviews

| Author | Algorithm | | | Findings |
|------------------------------------|--|--|---|---|
| Kumar et. al (2018) | Parametric | Nonparametric | | <ul style="list-style-type: none">• When dataset is large, RF outperformed others• When dataset is small, NB performed the best |
| | <ul style="list-style-type: none">• NB• KNN• Softmax | <ul style="list-style-type: none">• RF• SVR | | |
| Chen et al. (2019) | Parametric | Non parametric | | <ul style="list-style-type: none">• Tree-based ensemble models such as RF, and GB are the most accurate models.• Due to underfitting of MARS and 4QMA models, they had poor prediction performance |
| | <ul style="list-style-type: none">• 4QMA• LASSO• Ridge | <ul style="list-style-type: none">• CART• RF• XGBoost• SVR• MARS | | |
| Maehashi and Shintani (2020) | Linear | Ensemble | Neural Network | <ul style="list-style-type: none">• Tree-based ensemble models are the majority of the best models.• Large window size is recommended for better time series predictions. |
| | <ul style="list-style-type: none">• LASSO• Ridge• EN | <ul style="list-style-type: none">• Bagging• RF• AdaBoost | <ul style="list-style-type: none">• FFNN• CNN• LSTM | |

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, only the seglearn library is compatible to be used with machine learning models from the scikit learn library, with windowing feature. Basic libraries such as “Pandas”, “Matplotlib”, and “Numpy”, are also included in this project for data cleansing, data preparation, and visualisations.

2.6 Chapter Summary

In this chapter, 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. Next chapter will present result and discussion on performance evaluation of machine learning techniques and statistical model for RPC prediction.

CHAPTER 3

RESULTS AND DISCUSSIONS

3.1 Model Evaluation

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

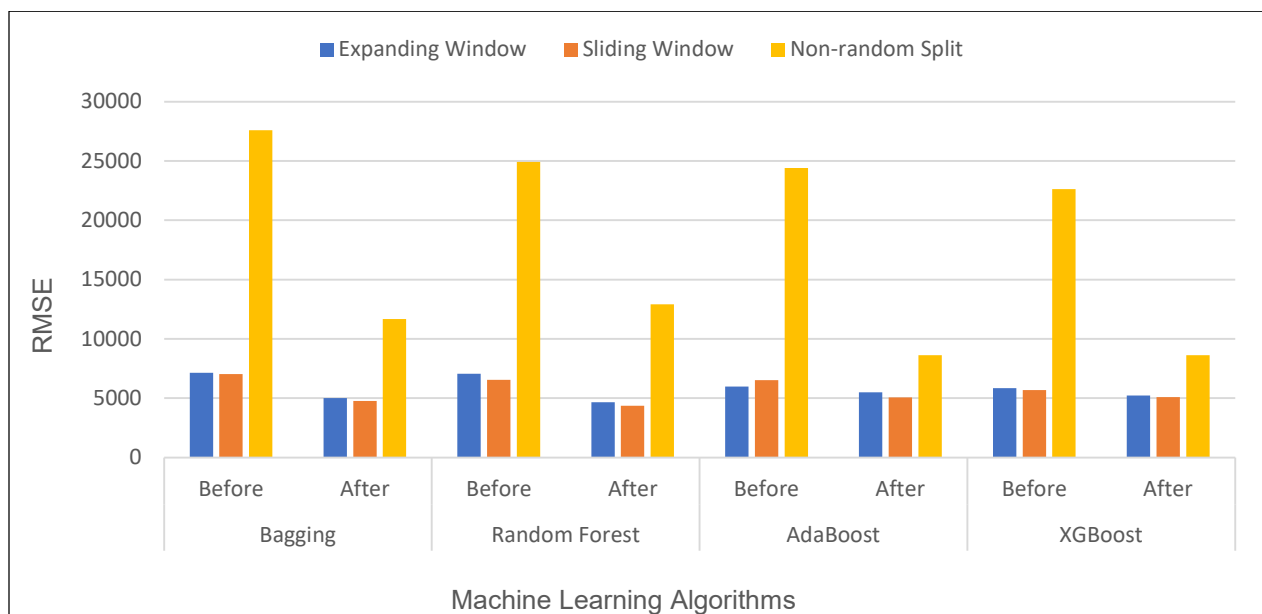


Figure 13 Comparison of RMSE before and after model optimisation

Based on Figure 13, it is observed that all models' prediction error is 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 is 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 is observed before and after model optimisation whereby majority of the models (3 out of 4) produced the least RMSE when sliding window is 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 is identified that RF with sliding window produced the least RMSE after model optimisation. This contradicts with the models' performances during model development whereby XGB is 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 is trained, thus producing a better RF model.

Finally, based on the results (see Figure 13), 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 is 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 are considered for comparison with statistical model as illustrated in Figure 14 and their averaged values is tabulated in Table 4.

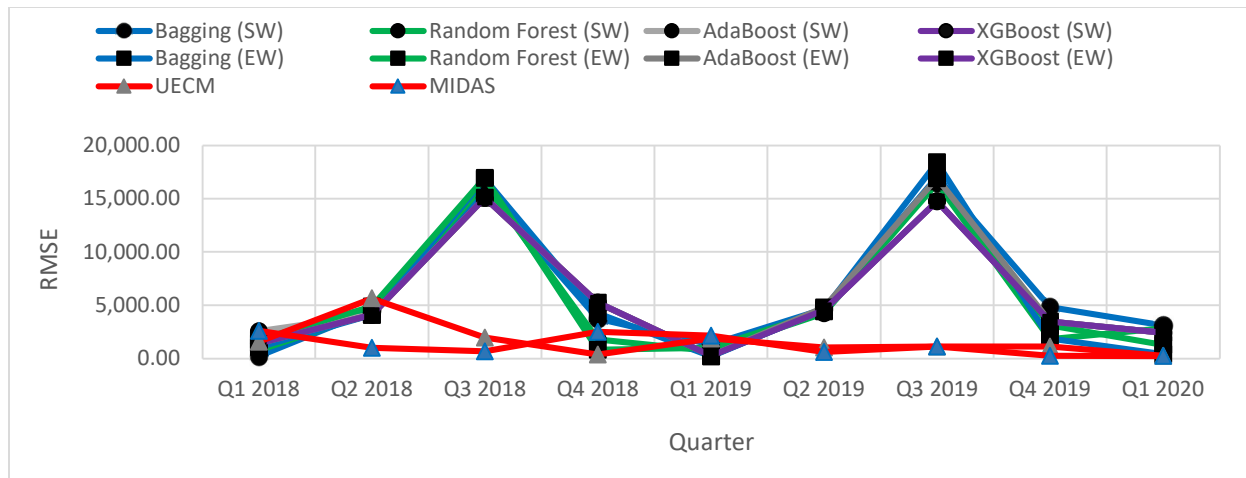


Figure 14 Comparison of RMSE between machine learning with statistical methods

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 14, Table 4 and 5, it is 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.

3.2 Data Prediction

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 is used for data prediction. Figure 15 below displays the prediction results along with the actual RPC plots from 2015 until 2020.

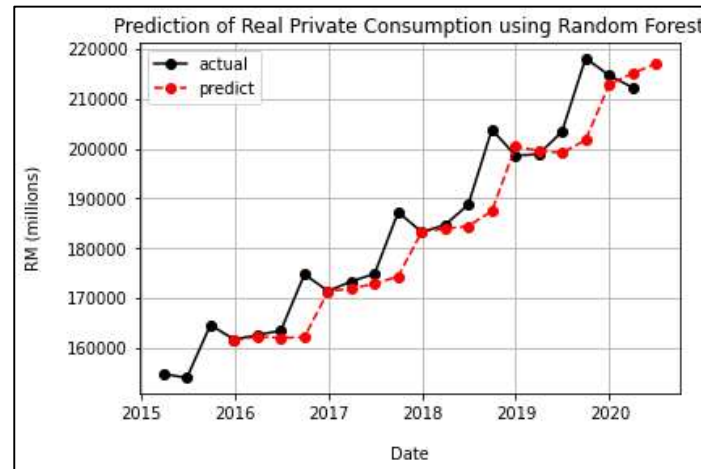


Figure 15 RPC prediction using Random Forest with sliding window

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; Tyrallis, 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 is 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.

3.3 What have been achieved and not achieved

Throughout the practicum, all of the objectives are successfully achieved in which the best machine learning algorithms for real private consumption are identified, developed, and evaluated. However, it is 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 have been faced during the implementation this project. At first, the first challenge is reviewing the best algorithms for RPC prediction. Since machine learning algorithms are very new in macroeconomics prediction, not many publications are 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 is to keep searching related publications with the right keyword.

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. Furthermore, much time has been spent on model optimisation which would be faster to be run on gaming computers. Since much time is 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. it is 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 are 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 is 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 are 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 is 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.

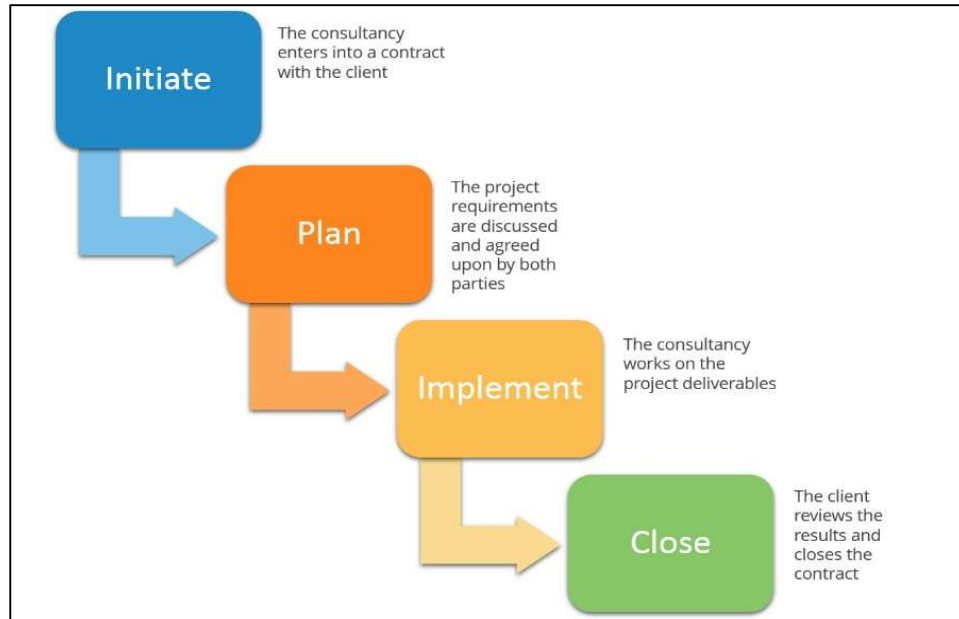


Figure 16 Waterfall Model of DataMicron project management methodology

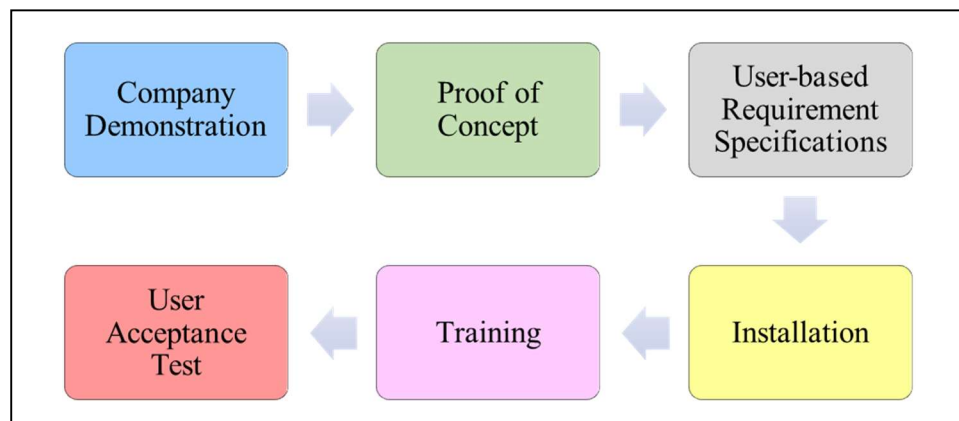


Figure 17 DataMicron's Flow of Project Management

3.7 Chapter Summary

In this chapter, results and discussion on performance evaluations of machine learning algorithms for this project have been presented. Model evaluation, and RPC prediction using the best machine learning technique has been explained in this chapter. Finally, the overall project reflections have been discussed at the end of this chapter.

CHAPTER 4

CONCLUSION AND LESSON LEARNED

4.1 Conclusion

Overall, this project has 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 are 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 are successfully developed and optimised by determining the best number of significant attributes, and number of window sizes to be input into the models. Comparison between machine learning models with statistical models shows that statistical models remained the best than machine learning models. Among machine learning models, Random Forest with sliding window has been found to be the best after it has been optimised. In order to further improve this model, it has been suggested to apply gridsearch during model optimisation stage to further improve its prediction accuracy in hopes of surpassing prediction accuracy of the current statistical models.

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 am 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 am 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 informations, and predictions, storytelling is the utmost important skill a data science must have to tell the storyline of the project outcome. Although the storytelling is done by my mentor, I observed that he possesses an excellent storytelling skill when he delivered our POC outcome to the client. If I am 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 is 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.

4.4 Useful Data Science Pipeline and Theories

Throughout this project, the data science pipeline implemented in this project's flow chart as illustrated in Figure 4 have been a huge help in structuring a data science project. Step-by-step process from data collection until data forecasting in Figure 4 has been organised very well resulting in a structured method of doing time series prediction of real private consumption.

Firstly, during data collection, since the dataset has been retrieved from Ministry of Finance which took from publications of DoSM and Bank Negara (in Microsoft excel spreadsheet format), thus, this step has been efficiently done by downloading from the client's email account and verify the data content by researching the actual values of each attributes from DoSM and BNM publications. In addition, since the dataset is already in spreadsheet format, therefore it is considered as a structured data. This type of data is easily managed for cleansing and management.

Next, data cleansing is a very crucial step in data science pipeline as the time series prediction requires a very high quality of data to produce accurate predictions. Even in Data Science there is a term "*Garbage in, garbage out*" which means machine learning predictions is only meaningful when a quality data is used. Otherwise, machine learning predictions are meaningless and cannot be interpreted. This is the reason for the removal of non-meaningful attributes, and invalid values in the time series dataset.

In addition to the data cleansing, the data has also been prepared for it to be trained and tested using windowing method and non-random percentage split respectively. This is because the dataset is a time series which requires the data to be specifically trained and tested in sequence. Using the prepared data, the machine learning models have been developed, optimised, and evaluated with RMSE during model development, optimisation, and evaluation stages. The final step of model evaluation has been comparing between machine learning models with the statistical models. Then, future value of RPC has been predicted using the best machine learning model.

Finally, the approach machine learning implemented in this project is a supervised learning in which the time series prediction of RPC requires several attributes for the machine learning algorithms to learn and predict for the respective RPC based on the given training data.

4.5 Suggestion for Improvement of Practicum and its Preparations

In future, I would suggest for data science students to explore good opportunity to enrol with data-driven companies for practicum placement. It is important for them to review on the practicum placements and match their project with the university's requirement for the practicum. If all requirements match, students may proceed with preparing the company's domain background and seek for the practicum coordinator's approval for having practicum there. If there are problems arise, do contact frequently with practicum coordinator as they would try as much as possible to assist students in obtaining the best practicum placement.

Furthermore, in preparing for the practicum, it would be better to approach companies having industrial lecturers who are expert in both academic research and real-world applications of data science. By approaching them to be the mentor, it would be much beneficial for data science practicum student to gain valuable experience in doing data science projects and at the same time doing research of the project domain in solving the company's problem related to data. My experience as an intern to DataMicron has opened my eyes on the importance of academic people in industries as they are highly respected by the company's chief executive officer and being the advisor to the company for every project. Since these industrial lecturers have many years of experience in dealing with data related projects, they are always the person to be referred by data analysts, and data scientists. Hence, it would be much better to approach companies having industrial lecturers as they can help on suggesting the best methods for solving clients' problem.

My last suggestion for practicum improvement is to deal with companies to only focus the task for data science projects only. This is because my experience as an intern has brought myself for various data science tasks for various domains other than time series prediction such as developing current prototype of DataMicron's Eagleye in doing natural language processing tasks for Crime and Investigative domain and developing image processing engine for extracting information from invoices, forms, and receipts. All of these experiences have indeed elevated my experience to a whole another level, however, the side effect is my research quality on time series prediction has been dropped as I have other tasks to complete as an intern to the company. Hence, I would suggest future data science student to only focus on the research for better research quality.

4.6 Integration of Practicum Experience with Student Coursework in DSA program

Based on my experience during practicum at DataMicron, skills for time series prediction using machine learning techniques, extracting entities and relations from texts, and image processing of forms, invoices, and receipts would be highly demanded. Hence, I would suggest for the DSA program to integrate students' coursework with these skills.

Time series prediction using machine learning techniques can be integrated in Predictive Business Analytics (CDS 512) subject in which students can use the sample data in Kaggle to predict future values of a time series data. Currently, statistical techniques are taught in this subject for predicting future values. In order to enrich students' knowledge and equip them with the demanding skills, windowing techniques incorporated with machine learning and deep learning algorithms are encouraged to be taught as they have the potential to be applied in many domains.

Secondly, extracting entities and relations from texts are advanced skill and not taught in any DSA program. I would suggest this skill to be taught in Text and Speech Analytics (CDS 522) subject. This is because extracting entities and relations in texts would automate the tedious job of extracting information from texts currently being done manually. This skill is very demanded by document-based companies such as crime and investigation, law and judiciary, and accountancy domains. Hence, integrating this skill would be very effective for DSA students to be employed.

Finally, image processing is an advance skill currently being taught in Multimodal Information Retrieval (CDS 521). Currently, development of a deep learning model to classify images is taught. It would be fruitful for this subject to extend the image processing to object detection in which specific regions in images are classified. Again, this would be helpful for document-based agencies in extracting information from forms, and invoices. Hence, DSA students would have high chances to be employed by artificial intelligence companies.

4.7 Chapter Summary

In this chapter, the main conclusion for this project has been described. Overall reflections of this project have been presented in this chapter. In a nutshell, this chapter attempts to summarise the whole project and experiences throughout practicum.

REFERENCES

- Afandi, A. and Khoo, R. (2020). Ringgit are 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.
- 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*. Ishington: 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

APPENDIX 1



UHM UNIVERSITI
SAINS
MALAYSIA



DATAMICRON

Project Consultancy and Practicum

DataMicron Systems Sdn Bhd

Practicum Logbook submitted in partial fulfilment of the requirement for

Masters of Science (Data Science and Analytics)

For

Dr. Nasuha Lee Abdullah

Mohd Azam Osman

By

Muhammad Azzubair bin Azeman

P-COM0019/19

Session

Semester 2020 / 2021

| DATE | WORK DONE | CONTACT HOURS | REMARKS |
|------------|--|---------------|--|
| 12/10/2020 | Meeting with Client together with mentor. | 2 hours | Client presented their problem statement, current methodology, and solution expectations. In addition, they shared their dataset. Problem Statement: - Inaccurate PCI Forecasting Solution Expectation: - More Accurate PCI Forecasting using Machine Learning |
| 12/10/2020 | Internal Meeting with Mentor and colleagues | 2 hours | Mentor discussed strategies to meet up clients' expectations. Have to find additional variables to enrich given dataset and more accurate PCI forecasting. |
| 13/10/2020 | Researched on additional variables referring on published papers | 7.75 hours | Found 2 additional variables based on published papers: |
| 14/10/2020 | Presented to Mentor on researched additional variables. | 1 hour | Mentor approved on the additional attributes and decided to proceed with data science tasks: - Additional Data Collection - Data Modelling - Future PCI Forecasting |
| 14/10/2020 | Collected Additional Variables | 3 hours | Searched additional variables in official data sources such as: - Department of Statistics Malaysia (DOSM) - Bank Negara Malaysia (BNM) |

| | | | |
|------------|--------------------|------------|--|
| 14/10/2020 | Data Preprocessing | 0,25 hours | Filter Dataset <ul style="list-style-type: none"> - Discarded invalid values (zeros) - Omitted invalid attributes Transform Dataset <ul style="list-style-type: none"> - Aggregated values from daily to quarterly values |
| 15/10/2020 | Data Modelling | 36 hours | Modelling Dataset <ul style="list-style-type: none"> - Researched Modelling python codes for Bagging, Random Forest, AdaBoost, and XGBoost. - Evaluated the Models |
| 19/10/2020 | Model Optimisation | 36 hours | Compared Model Performance <ul style="list-style-type: none"> - Compared before and after Feature Selection - Compared before and After tuning window size - Compared before and after combining Feature Selection and Tuning Horizon Size combination |
| 26/10/2020 | Model Evaluation | 24 hours | Evaluated Model Performance <ul style="list-style-type: none"> - RMSE during Model Development and Optimisation were evaluated. - Discussed the effect of window size and feature selection with model performance (RMSE) - Compared prediction errors between ML models with statistical models. |

| | | | |
|------------|---|----------|--|
| 2/11/2020 | Model Forecasting | 24 hours | <p>Forecasted 1 period in future</p> <ul style="list-style-type: none"> - Using the top three performing models, new future values were generated and input into the models to predict for future PCI. - Models were tuned to achieve better performing forecasting accuracy by comparing with historical PCI values. - The best forecasting hyperparameters were set to forecast future PCI values |
| 30/11/2020 | Modeling and Forecasting Presentation to mentor | 2 hours | Mentor approves on the methodology and outcomes. The forecasting values will be evaluated by comparing with published PCI values that will be released by November 2020. |

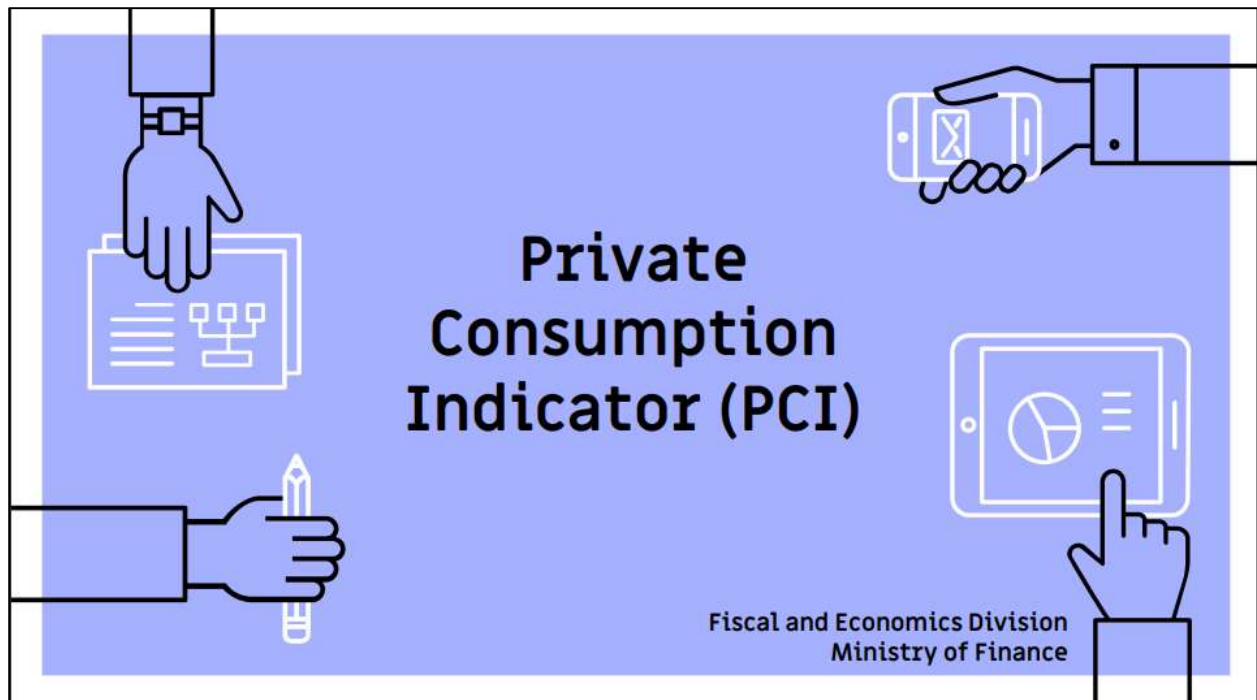
| | |
|-------------------|---|
| Remarks by Mentor | <p>Azzahar always come in on time, follow schedule and adhere to designated lunch time. He works best for the team and always ready to help other staffs in assisting their tasks. He can adapt easily to multitasking jobs given by team leader and come out with good outcomes that benefit the team.</p> |
|-------------------|---|

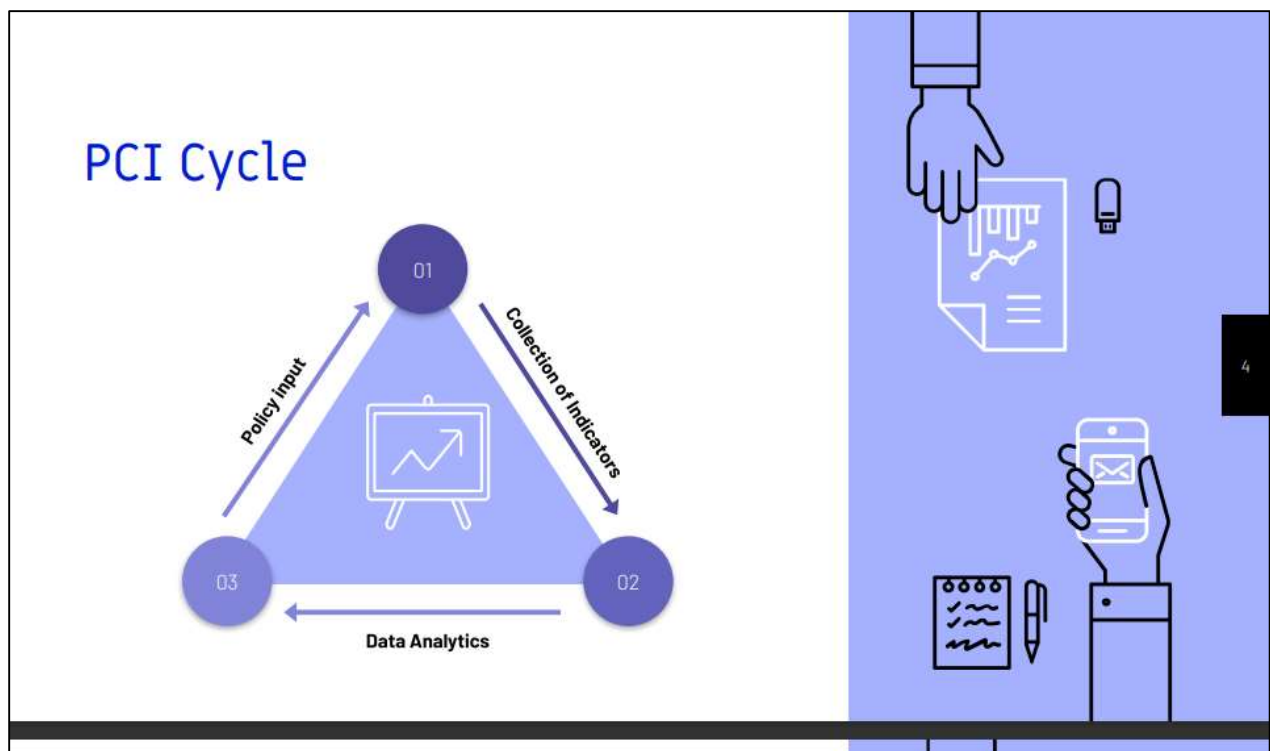
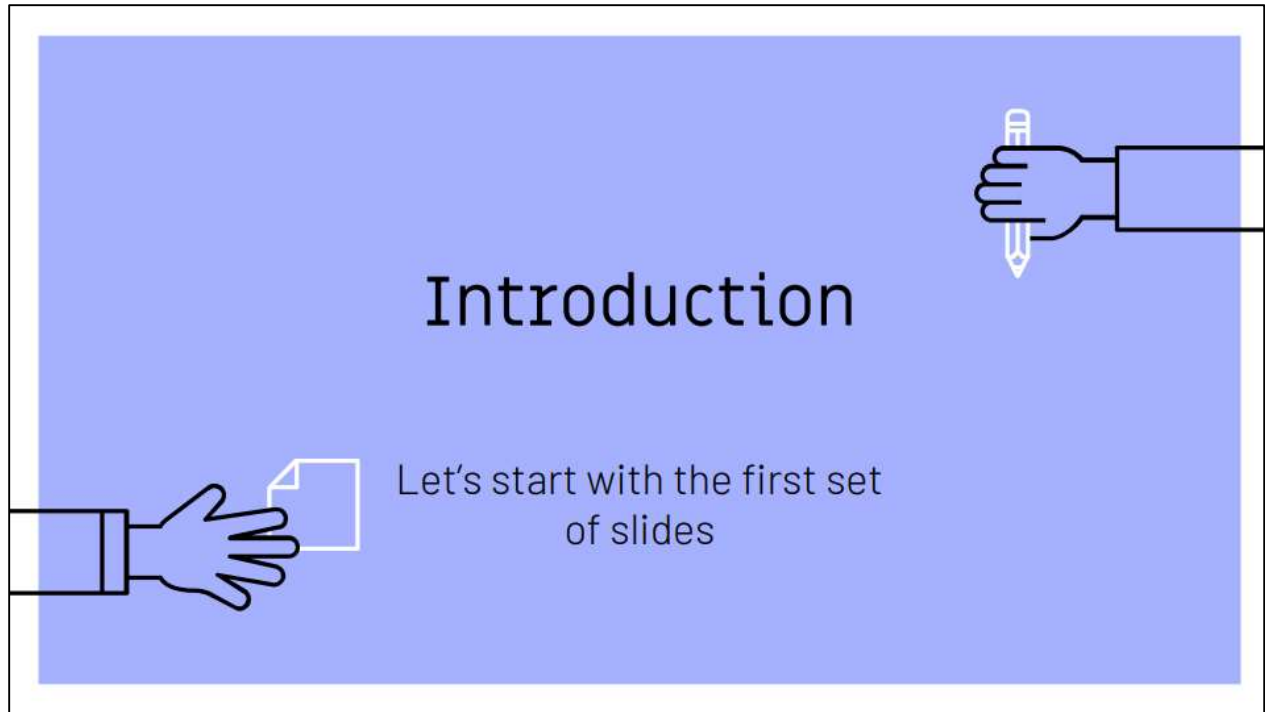
Mentor's Signature:


 DATAMICRON SYSTEMS SDN BHD
 (Reg No: 874340-H)
 Suite 09-11, 8th Floor, Wisma UDA I
 21, Jalan Pintas
 50450 Kuala Lumpur

Date: 12/1/2021

APPENDIX 2





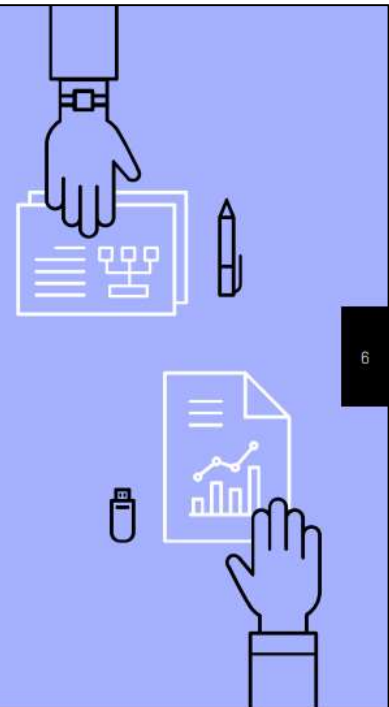
Collection of Indicators

- ▶ Excel with monthly data/ higher frequency (structured data)



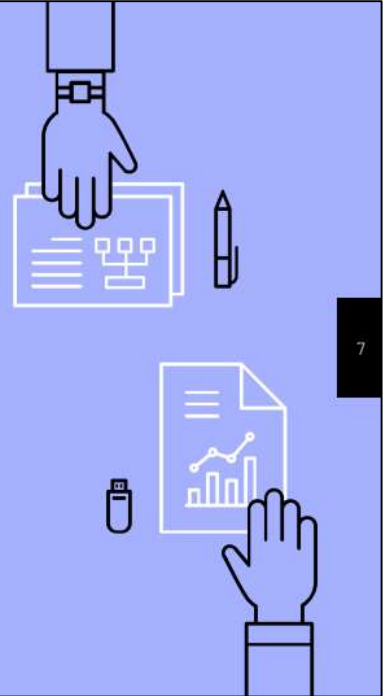
Data Analytics

- ▶ Transforming data into valuable asset
- ▶ Using various econometrics and machine learning as well as deep learning (in future) approach



Policy Input

- Visualisation (table/ interactive charts & dashboard/ apps in future)
- Policy brief



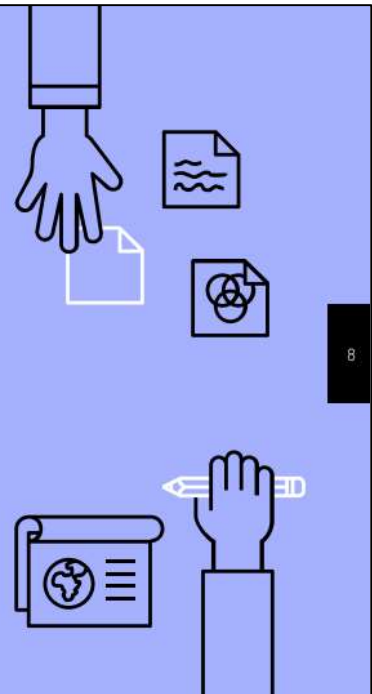
Introduction

Share of private consumption to Malaysian economy

Private consumption constitute the largest contributor to the Malaysian economy, more than 55%.

Private consumption growth

Ranging from 6.5% to 8.9% between 2018Q1 and 2020Q1.



Share of private consumption to Malaysian economy

| | 2018Q1 | 2018Q2 | 2018Q3 | 2018Q4 | 2019Q1 | 2019Q2 | 2019Q3 | 2019Q4 | 2020Q1 |
|---------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| GDP | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Private Consumption | 56.5 | 56.6 | 59.0 | 55.6 | 58.2 | 58.2 | 60.5 | 58.0 | 61.7 |
| Public Consumption | 11.1 | 11.8 | 11.6 | 15.0 | 11.3 | 11.3 | 11.3 | 14.7 | 11.8 |
| Private Investment | 18.0 | 20.2 | 17.9 | 13.3 | 17.3 | 19.5 | 17.2 | 13.4 | 16.8 |
| Public Investment | 7.2 | 5.9 | 6.9 | 9.3 | 6.0 | 5.2 | 5.6 | 8.3 | 5.2 |
| Export | 68.5 | 66.9 | 67.3 | 66.7 | 65.5 | 64.2 | 63.1 | 62.2 | 60.5 |
| Import | 60.5 | 61.5 | 61.1 | 59.7 | 56.9 | 57.3 | 56.5 | 56.2 | 55.1 |
| Inventories | -0.9 | 0.1 | -1.6 | -0.3 | -1.5 | -1.2 | -1.2 | -0.4 | -0.9 |

Private consumption growth

| | 2018Q1 | 2018Q2 | 2018Q3 | 2018Q4 | 2019Q1 | 2019Q2 | 2019Q3 | 2019Q4 | 2020Q1 |
|---------------------|--------|--------|--------|--------|--------|---------|--------|--------|--------|
| GDP | 5.2 | 4.7 | 4.4 | 4.8 | 4.5 | 4.8 | 4.4 | 3.6 | 0.7 |
| Private Consumption | 6.5 | 7.9 | 8.9 | 8.4 | 7.7 | 7.8 | 7.0 | 8.1 | 6.7 |
| Public Consumption | 0.2 | 2.9 | 5.0 | 3.9 | 6.3 | 0.3 | 1.0 | 1.2 | 5.0 |
| Private Investment | 1.1 | 5.6 | 5.0 | 5.9 | 0.6 | 1.5 | 0.4 | 4.3 | -2.3 |
| Public Investment | -1.2 | -10.0 | -2.7 | -6.0 | -13.7 | -7.8 | -14.6 | -8.0 | -11.3 |
| Export | 2.3 | 2.0 | 0.5 | 2.9 | 0.1 | 0.5 | -2.1 | -3.4 | -7.1 |
| Import | -2.0 | 3.7 | 2.3 | 2.0 | -1.6 | -2.3 | -3.5 | -2.4 | -2.5 |
| Inventories | -242.8 | -116.4 | 84.8 | -132.2 | 75.3 | -2009.0 | -20.5 | 31.7 | -35.5 |

9

Model usage

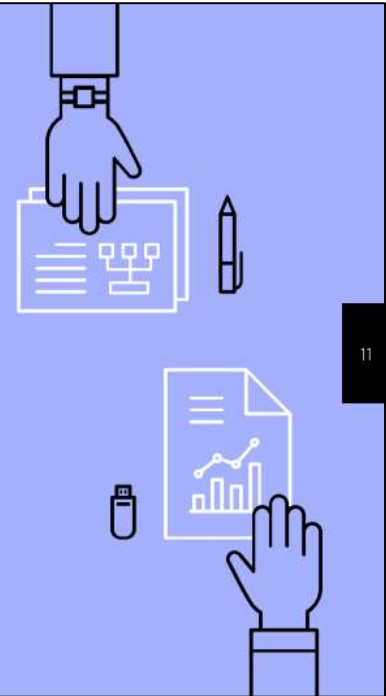
Continue with the second
set of slides

Model Framework

$$PCON = f(FS, ICG, SOPC, LOWT)$$

Where

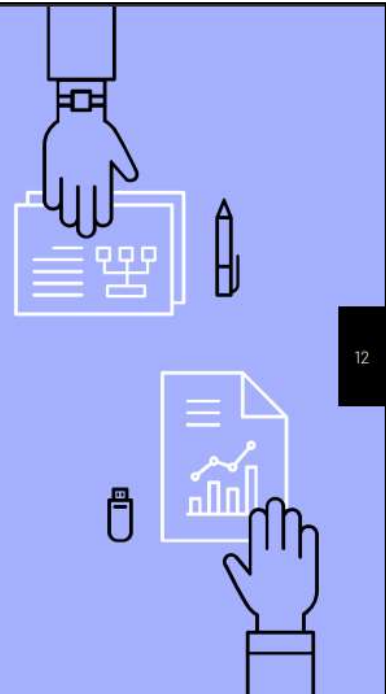
- ▶ PCON is real private consumption
- ▶ FS is food sales
- ▶ ICG is import of consumption goods
- ▶ SOPC is sales of passenger vehicles
- ▶ LOWT is loan to wholesale and distributive trade



11

Data collection

- ▶ Department of Statistics Malaysia (DOSM)
- ▶ Malaysia Automotive Association (MAA)
- ▶ PCON is in quarterly frequency
- ▶ Others are in monthly frequency



12

Unrestricted Error Correction Model (UECM)

Quarterly frequency

- ▶ Stationarity (ADF and PP tests)
- ▶ Cointegration tests (Bounds test)
- ▶ Long run estimates (equilibrium adjustment)
- ▶ Short run estimates



13

Mixed Data Sampling Regression (MIDAS)

Quarterly - monthly frequency

- ▶ Dealing with mixed-frequency data
- ▶ Able to solve the problem of parameter proliferation
- ▶ While preserving some timing information

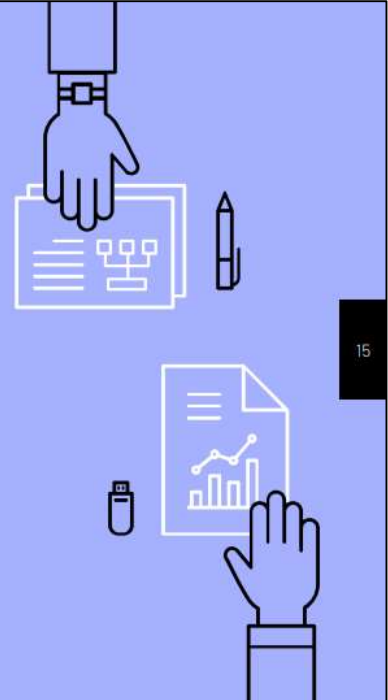


14

Mixed Frequency Vector Autoregressions (MFVAR)

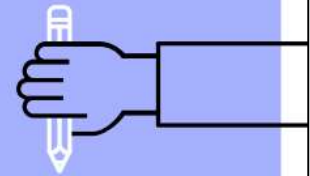
Quarterly – monthly frequency

- ▶ System approach that jointly explains indicators and predictant without imposing a-priori restrictions on the dynamics
- ▶ Can be an advantage when few variables are modelled and the dynamics is limited
- ▶ VAR provides a good approximation to the data generating process (DGP)



Forecast accuracy

Continue with the third set of slides



Forecast error measurement

Based on error

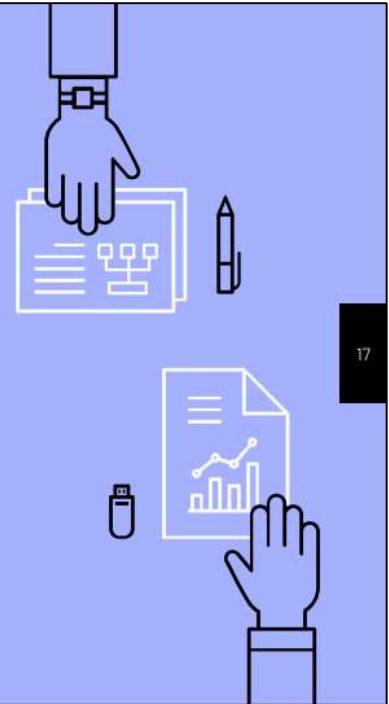
- ▶ Smaller error gives better accuracy

$$e_t = y_t - \hat{y}_t$$

e_t = forecast error

y_t = true value

\hat{y}_t = forecast value



17

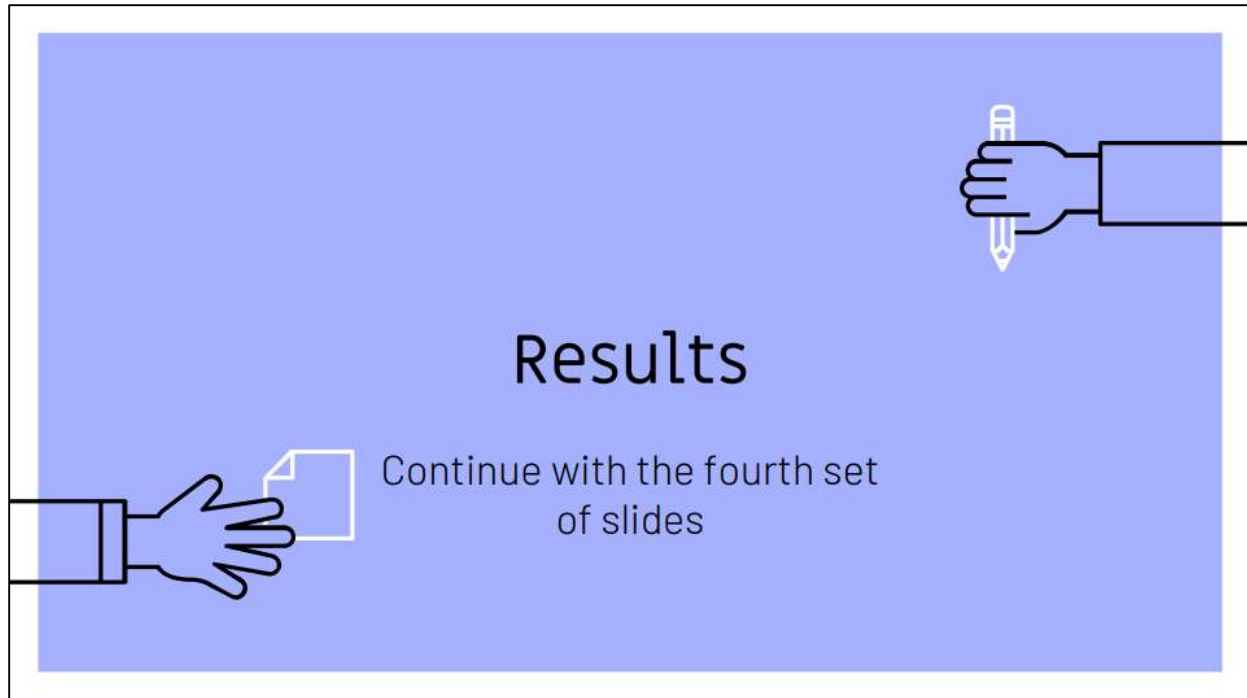
Root Mean Square Error

- ▶ Standard deviation of the residuals
- ▶ Measures how much error there is between two data sets

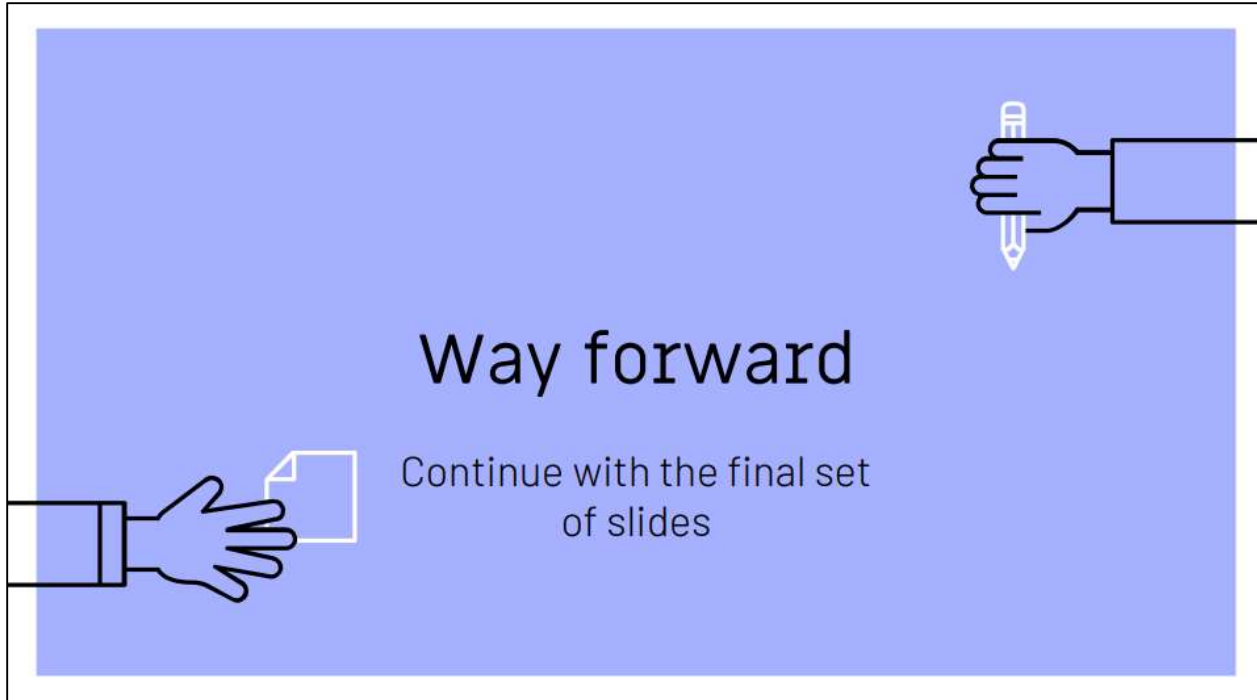
$$RMSE = \left(\left[\sum_{t=1}^n \frac{(F_t - X_t)^2}{n} \right] \right)^{1/2}$$



18



| Able to track up to 8 lags before | | | | | | | | | |
|---|--------|--------|---------|---------|---------|--------|---------|---------|---------|
| | 2018Q1 | 2018Q2 | 2018Q3 | 2018Q4 | 2019Q1 | 2019Q2 | 2019Q3 | 2019Q4 | 2020Q1 |
| ACTUAL | 6.5 | 7.9 | 8.9 | 8.4 | 7.7 | 7.8 | 7.0 | 8.1 | 6.7 |
| UECM | 5.6 | 4.7 | 7.8 | 8.2 | 6.7 | 7.2 | 6.5 | 7.5 | 6.6 |
| MIDAS | 8.0 | 7.4 | 8.5 | 7.0 | 6.5 | 8.1 | 7.6 | 8.2 | 6.6 |
| MFVAR | 11.5 | 10.7 | 6.2 | 15.6 | 13.6 | 10.8 | 5.5 | 16.6 | 15.8 |
| RMSE(UECM) | 1560.1 | 5650.9 | 1952.9 | 382.7 | 1883.3 | 1037.7 | 1127.5 | 1114.1 | 261.4 |
| RMSE(MIDAS) | 2596.6 | 1009.1 | 664.4 | 2495.4 | 2142.8 | 608.2 | 1118.8 | 261.3 | 233.6 |
| RMSE(MFVAR) | 8612.9 | 4829.5 | 4957.7 | 13262.3 | 10993.8 | 5672.4 | 3089.7 | 16843.7 | 17939.1 |
| No error for current quarter since actual values is not released yet | | | | | | | | | |
| | 2018Q2 | 2018Q3 | 2018Q4 | 2019Q1 | 2019Q2 | 2019Q3 | 2019Q4 | 2020Q1 | 2020Q2 |
| ACTUAL | 7.9 | 8.9 | 8.4 | 7.7 | 7.8 | 7.0 | 8.1 | 6.7 | NA |
| UECM | 4.7 | 7.8 | 8.2 | 6.7 | 7.2 | 6.5 | 7.5 | 6.6 | 2.2 |
| MIDAS | 7.4 | 8.5 | 7.0 | 6.5 | 8.1 | 7.6 | 8.2 | 6.6 | -4.2 |
| MFVAR | 10.7 | 6.2 | 15.6 | 13.6 | 10.8 | 5.5 | 16.6 | 15.8 | 11.5 |
| RMSE(UECM) | 5650.9 | 1952.9 | 382.7 | 1883.3 | 1037.7 | 1127.5 | 1114.1 | 261.4 | NA |
| RMSE(MIDAS) | 1009.1 | 664.4 | 2495.4 | 2142.8 | 608.2 | 1118.8 | 261.3 | 233.6 | NA |
| RMSE(MFVAR) | 4829.5 | 4957.7 | 13262.3 | 10993.8 | 5672.4 | 3089.7 | 16843.7 | 17939.1 | NA |

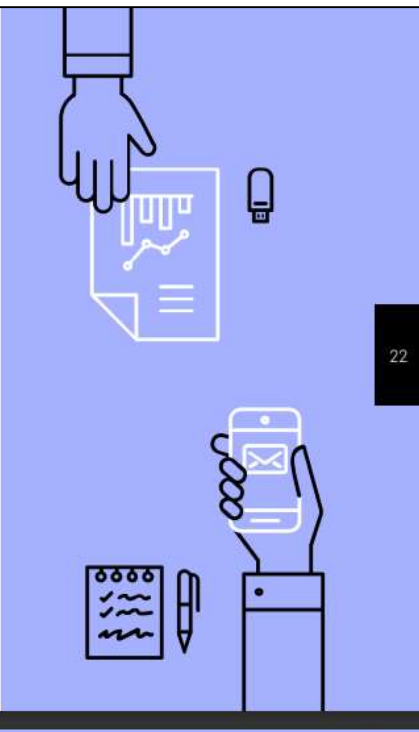


Way forward

Continue with the final set of slides

How to improve the PCI?

- 1) **Replacing current indicators with others** which are more accurate to predict the movement of private consumption (RMSE comparison as a benchmark)
- 2) **Adding more error performance criterion** (MAE, MAPE, Theil Inequality Coefficient, MSE)
- 3) **Increase the number of indicators used** (cost: overfitting the model)
- 4) Implementing **Machine Learning approach** (e.g. Decision Tree, kNN, SVM, Naïve Bayes, K-Means, Random Forest) and **Deep Learning approach** (e.g. Neural Network, Restricted Boltzmann Machine, Long Short-Term Memory)
- 5) **Build Private Consumption Index** (need to identify weightage + engage DOSM, however Dominance Analysis might be good option to start with)
- 6) **Track higher frequency data** in unstructured/ semi structured form in order to get more regular forecast update



22

APPENDIX 3

| | FS | ICG | TSP | SOM | SOPC | CC | CCD | LOWT | CCS | LDH | NM | FBM | MCF | MSW | SSW | MSV | MIER | EMP |
|---------|---------|---------|---------|-------|-------|--------|---------|----------|---------|---------|--------|---------|---------|---------|-----|-----|------|---------|
| 2014M12 | 2497.36 | 4583.09 | 834.883 | 32596 | 55523 | 256515 | 17255.7 | 20551.93 | 35605 | 27635.4 | 346416 | 1761.25 | 1651.17 | 0 | 0 | 0 | 0 | 13748.4 |
| 2015M01 | 4152.67 | 4384.72 | 725.455 | 38276 | 44696 | 256625 | 15231 | 17119 | 35413.5 | 26178.3 | 346300 | 1781.26 | 1687.84 | 5446.57 | 0 | 0 | 0 | 13991.8 |
| 2015M02 | 4022.58 | 3903.38 | 585.681 | 31611 | 44961 | 257491 | 14411.7 | 12883.57 | 35500.9 | 23331 | 351432 | 1821.21 | 1733.14 | 5371.3 | 0 | 0 | 0 | 14092.9 |
| 2015M03 | 4153.31 | 4382.15 | 379.759 | 39361 | 59480 | 257905 | 17594.6 | 17071.92 | 35588.1 | 28698.4 | 360341 | 1830.78 | 1737.52 | 5515.08 | 0 | 0 | 0 | 14160.5 |
| 2015M04 | 4150.13 | 4859.64 | 380.969 | 29699 | 40902 | 257256 | 14239.9 | 15720.31 | 34583.6 | 25214.3 | 351906 | 1818.27 | 1738.53 | 5372.32 | 0 | 0 | 0 | 14136.2 |
| 2015M05 | 4046.74 | 5405.54 | 669.673 | 31849 | 45874 | 257378 | 14537.6 | 14743.18 | 34524.2 | 23444.7 | 352321 | 1747.52 | 1694.81 | 5326.55 | 0 | 0 | 0 | 14126.7 |
| 2015M06 | 4078.76 | 5820.95 | 650.631 | 31870 | 50695 | 258207 | 16257.1 | 18626.13 | 34722.6 | 25735.1 | 360586 | 1706.64 | 1659.02 | 5465.49 | 0 | 0 | 0 | 14146.7 |
| 2015M07 | 4060.99 | 5272.68 | 699.988 | 29982 | 52636 | 259157 | 16030.2 | 17260 | 34557.2 | 25112.9 | 347181 | 1723.14 | 1680.9 | 5358.13 | 0 | 0 | 0 | 14106.2 |
| 2015M08 | 4142.36 | 4794.47 | 812.24 | 28312 | 47309 | 260466 | 15748.8 | 15878.58 | 34845 | 24206.1 | 351493 | 1612.74 | 1555.89 | 5374.33 | 0 | 0 | 0 | 14214.6 |
| 2015M09 | 4357.59 | 5688.85 | 927.145 | 28658 | 45238 | 260520 | 16027.1 | 17685.87 | 34706.3 | 25000.8 | 357984 | 1621.04 | 1597.35 | 5433.32 | 0 | 0 | 0 | 14151.7 |
| 2015M10 | 4348.2 | 5815.56 | 1074.49 | 32318 | 49063 | 261207 | 16625.2 | 17659.75 | 34744.7 | 26321.1 | 351276 | 1665.71 | 1653.75 | 5476.59 | 0 | 0 | 0 | 14184.7 |
| 2015M11 | 4329.51 | 5826.01 | 945.845 | 27819 | 49324 | 262513 | 16231.3 | 16532.05 | 35343.2 | 25628.8 | 352369 | 1672.16 | 1668.89 | 5477.31 | 0 | 0 | 0 | 14215.7 |
| 2015M12 | 4483.54 | 6276.44 | 770.383 | 31047 | 61102 | 263677 | 18188.4 | 18531.21 | 36043.6 | 27593.1 | 360503 | 1692.51 | 1695.17 | 5878.29 | 0 | 0 | 0 | 14185.7 |
| 2016M01 | 4685.84 | 5839.48 | 642.784 | 36248 | 39942 | 264237 | 16606.5 | 16183.64 | 36370.7 | 25920.2 | 363454 | 1667.8 | 1660.62 | 5857.94 | 0 | 0 | 0 | 14150.5 |
| 2016M02 | 4317.45 | 4601.73 | 637.627 | 30633 | 34054 | 262702 | 14337.6 | 14558.84 | 35409.5 | 22285.9 | 366985 | 1654.75 | 1650.14 | 5765.92 | 0 | 0 | 0 | 14196.9 |
| 2016M03 | 4710.18 | 5342.3 | 639.356 | 39343 | 43088 | 262009 | 15938.2 | 16768.51 | 35017.2 | 25507.1 | 354187 | 1717.58 | 1707.44 | 5742.79 | 0 | 0 | 0 | 14200.7 |
| 2016M04 | 4486.94 | 5602.59 | 556.665 | 33065 | 37783 | 262044 | 15083.5 | 15689.09 | 35078 | 23511.1 | 348628 | 1672.72 | 1672.73 | 5697.5 | 0 | 0 | 0 | 14163.7 |
| 2016M05 | 4561.93 | 6153.28 | 593.337 | 31972 | 38532 | 261882 | 15677 | 16414.53 | 35097.7 | 24057.5 | 356464 | 1626 | 1641.01 | 5643.82 | 0 | 0 | 0 | 14200.2 |
| 2016M06 | 4464.99 | 5931.15 | 652.064 | 30885 | 50978 | 262397 | 16211.5 | 16992.65 | 35418 | 24621.3 | 363911 | 1654.08 | 1660.34 | 5814.75 | 0 | 0 | 0 | 14218.4 |
| 2016M07 | 4320.17 | 4915.25 | 600.656 | 26078 | 37661 | 262293 | 15904.9 | 15326.83 | 35392.2 | 22867.9 | 354251 | 1653.26 | 1673.85 | 5779.19 | 0 | 0 | 0 | 14212.8 |
| 2016M08 | 4628.05 | 5290.79 | 841.347 | 33285 | 46227 | 262799 | 16581.2 | 16818.81 | 35518.8 | 25420.9 | 354935 | 1678.06 | 1694.42 | 5831.44 | 0 | 0 | 0 | 14306.9 |
| 2016M09 | 4553.52 | 5417.76 | 798.069 | 35874 | 42644 | 263070 | 16227.6 | 17440.45 | 35694.4 | 24523.7 | 358243 | 1652.55 | 1686.16 | 5934.5 | 0 | 0 | 0 | 14249.6 |
| 2016M10 | 4712.76 | 5354.08 | 620.557 | 31355 | 42523 | 263636 | 16161 | 16572.72 | 35830.6 | 24385.2 | 360901 | 1672.46 | 1705.05 | 5931.89 | 0 | 0 | 0 | 14253.4 |
| 2016M09 | 4553.52 | 5417.76 | 798.069 | 35874 | 42644 | 263070 | 16227.6 | 17440.45 | 35694.4 | 24523.7 | 358243 | 1652.55 | 1686.16 | 5934.5 | 0 | 0 | 0 | 14249.6 |
| 2016M10 | 4712.76 | 5354.08 | 620.557 | 31355 | 42523 | 263636 | 16161 | 16572.72 | 35830.6 | 24385.2 | 360901 | 1672.46 | 1705.05 | 5931.89 | 0 | 0 | 0 | 14253.4 |
| 2016M11 | 4929.47 | 6143.45 | 644.547 | 33554 | 43560 | 264135 | 16454.4 | 17369.99 | 36208.5 | 25742.1 | 368314 | 1619.12 | 1645.87 | 5946.61 | 0 | 0 | 0 | 14317.2 |
| 2016M12 | 4985.7 | 6385.56 | 517.729 | 34051 | 57593 | 265380 | 17901 | 18220.36 | 37148.7 | 25896.1 | 380861 | 1641.73 | 1667.37 | 6307.76 | 0 | 0 | 0 | 14276.7 |
| 2017M01 | 4925.49 | 5742.65 | 448.593 | 34707 | 40295 | 265504 | 17453.6 | 17600.65 | 36945.6 | 26542.6 | 387128 | 1671.54 | 1705.46 | 6144.17 | 0 | 0 | 0 | 14366.8 |
| 2017M02 | 4898.57 | 4572.11 | 538.256 | 36584 | 38877 | 264183 | 14761.8 | 16292.4 | 36204.7 | 22133.6 | 387020 | 1693.77 | 1739.87 | 6073.74 | 0 | 0 | 0 | 14401.6 |
| 2017M03 | 5306.17 | 6093.61 | 789.142 | 36931 | 48356 | 263998 | 17575.7 | 19051.75 | 35782.5 | 27542.6 | 386806 | 1740.09 | 1801.3 | 6359.87 | 0 | 0 | 0 | 14421.7 |
| 2017M04 | 5147.2 | 5665.12 | 888.3 | 35543 | 37741 | 264143 | 15575.1 | 15673.84 | 36012.7 | 23369.7 | 387032 | 1768.06 | 1840.63 | 6408.91 | 0 | 0 | 0 | 14429.6 |
| 2017M05 | 5393.36 | 6663.86 | 934.332 | 39223 | 45133 | 264501 | 17366.9 | 17004.88 | 36238.3 | 27214.2 | 391248 | 1765.87 | 1832.9 | 6403.92 | 0 | 0 | 0 | 14454.4 |
| 2017M06 | 4973.03 | 5621.18 | 889.437 | 31338 | 45341 | 264857 | 16703.1 | 16805.47 | 36551 | 24763.6 | 397698 | 1763.67 | 1838.18 | 6392.75 | 0 | 0 | 0 | 14519.9 |
| 2017M07 | 5238.11 | 5991.54 | 1091.45 | 37541 | 43524 | 264313 | 16343.1 | 16416.18 | 36452.7 | 24946.8 | 392766 | 1760.03 | 1843.63 | 6412.89 | 0 | 0 | 0 | 14497.4 |
| 2017M08 | 5376.52 | 6236.53 | 1311.62 | 40351 | 46008 | 265182 | 17768.4 | 17807.91 | 36645.5 | 27258.4 | 393776 | 1773.16 | 1847.33 | 6423.4 | 0 | 0 | 0 | 14513.4 |
| 2017M09 | 5260.39 | 5725.25 | 1236.31 | 35316 | 36503 | 265256 | 16267.9 | 16723.17 | 36954.7 | 25476.1 | 397982 | 1755.58 | 1845.49 | 6490.7 | 0 | 0 | 0 | 14544.3 |
| 2017M10 | 5490.07 | 5947.88 | 1243.43 | 35424 | 41670 | 265280 | 17546.4 | 17935.4 | 36868.7 | 28003.2 | 403113 | 1747.92 | 1860.45 | 6452.21 | 0 | 0 | 0 | 14581.7 |
| 2017M11 | 5530.66 | 6552.06 | 1072.98 | 38739 | 43155 | 266563 | 18524.5 | 18507.26 | 37565.2 | 29398.9 | 404897 | 1717.86 | 1834.7 | 6450.7 | 0 | 0 | 0 | 14578.9 |
| 2017M12 | 5602.82 | 6225.14 | 1002.21 | 33153 | 48067 | 267666 | 18824.2 | 17042.45 | 38659.5 | 27970.8 | 422820 | 1796.81 | 1906.84 | 6865.98 | 0 | 0 | 0 | 14640.1 |
| 2018M01 | 5845 | 6328 | 1105 | 38441 | 39967 | 267978 | 18760 | 18263 | 38552 | 30886 | 421396 | 1869 | 1960 | 6772 | 0 | 0 | 0 | 14670.5 |
| 2018M02 | 5340 | 5173 | 865 | 32335 | 36605 | 267607 | 16846 | 15946 | 38020 | 26482 | 420010 | 1856 | 1942 | 6823 | 0 | 0 | 0 | 14721.5 |
| 2018M03 | 5902 | 5355 | 1100 | 40457 | 44488 | 267388 | 18694 | 18103 | 37632 | 29166 | 417233 | 1863 | 1896 | 6971 | 0 | 0 | 0 | 14732.5 |
| 2018M04 | 5904 | 5597 | 1109 | 37018 | 41939 | 289547 | 17889 | 18278 | 37855 | 27731 | 415978 | 1870 | 1894 | 6944 | 0 | 0 | 0 | 14803.1 |
| 2018M05 | 5967 | 6004 | 1055 | 31603 | 40215 | 288682 | 16766 | 17478 | 37368 | 26054 | 417389 | 1741 | 1792 | 6986 | 0 | 0 | 0 | 14852.6 |
| 2018M06 | 5672 | 5911 | 864 | 34664 | 57709 | 290300 | 18546 | 19770 | 37833 | 27347 | 416501 | 1692 | 1771 | 6993 | 0 | 0 | 0 | 14863.2 |
| 2018M07 | 5933 | 6694 | 1095 | 44077 | 61212 | 291432 | 19101 | 19101 | 37981 | 29316 | 411225 | 1784 | 1861 | 6987 | 0 | 0 | 0 | 14882.4 |
| 2018M06 | 5672 | 5911 | 864 | 34664 | 57709 | 290300 | 18546 | 19770 | 37833 | 27347 | 416501 | 1692 | 1771 | 6993 | 0 | 0 | 0 | 14863.2 |
| 2018M07 | 5933 | 6694 | 1095 | 44077 | 61212 | 291432 | 19101 | 19101 | 37981 | 29316 | 411225 | 1784 | 1861 | 6987 | 0 | 0 | 0 | 14882.4 |
| 2018M08 | 5940 | 7127 | 1023 | 43403 | 55772 | 294180 | 20598 | 19969 | 38936 | 28842 | 411016 | 1820 | 1866 | 7011 | 0 | 0 | 0 | 14896.5 |
| 2018M09 | 5749 | 5152 | 911 | 38480 | 27021 | 294146 | 17536 | 18882 | 38846 | 26073 | 414530 | 1793 | 1835 | 7069 | 0 | 0 | 0 | 14926.5 |
| 2018M10 | 6011 | 6452 | 1344 | 49223 | 42364 | 293861 | 18788 | 20952 | 38354 | 28781 | 415920 | 1709 | 1727 | 7043 | 0 | 0 | 0 | 14937.1 |
| 2018M11 | 5795 | 6630 | 1080 | 41958 | 43366 | 294290 | 18051 | 20215 | 38945 | 27609 | 418501 | 1680 | 1712 | 6974 | 0 | 0 | 0 | 14941.3 |
| 2018M12 | 5932 | 6607 | 1059 | 40123 | 42428 | 294548 | 19545 | 21305 | 39920 | 30204 | 427721 | 1691 | 1700 | 7465 | 0 | 0 | 0 | 14986 |
| 2019M01 | 6223 | 6516 | 850 | 48605 | 44264 | 294535 | 19633 | 20629 | 39940 | 31128 | 428231 | 1684 | 1730 | 7362 | 0 | 0 | 0 | 14992.8 |
| 2019M02 | 5769 | 4555 | 741 | 36954 | 36725 | 293161 | 15792 | 15678 | 39132 | 24643 | 422365 | 1708 | 1758 | 7290 | 0 | 0 | 0 | 15026.8 |
| 2019M03 | 6661 | 5896 | 687 | 47157 | 50101 | 292919 | 18931 | 20785 | 38904 | 29965 | 427697 | 1644 | 1730 | 7326 | 0 | 0 | 0 | 15035.2 |
| 2019M04 | 6513 | 6621 | 947 | 43743 | 45302 | 292868 | 18394 | 19987 | 38999 | 29285 | 425324 | 1642 | 1748 | 7274 | 0 | 0 | 0 | 15089.8 |
| 2019M05 | 6456 | 6636 | 881 | 47136 | 55894 | 293683 | 19566 | 19107 | 39061 | 29391 | 434537 | 1651 | 1727 | 7253 | 0 | 0 | 0 | 15122.5 |
| 2019M06 | 6186 | 5581 | 674 | 33885 | 38513 | 293653 | 17424 | 16391 | 39260 | 25768 | 431751 | 1672 | 1753 | 7260 | 0 | 0 | 0 | 15134.6 |
| 2019M07 | 6270 | 6360 | 920 | 50627 | 46189 | 294028 | 19725 | 19109 | 39210 | 29578 | 429595 | 1635 | 1725 | 7328 | 0 | 0 | 0 | 15179.8 |
| 2019M08 | 6508 | 6214 | 859 | 48530 | 46802 | 294699 | 19624 | 18934 | 39489 | 29453 | 426561 | 1612 | 1690 | 7345 | 0 | 0 | 0 | 15185.8 |
| | | | | | | | | | | | | | | | | | | |