CDS 590

Mid-Term Presentation

Predicting Real Private Consumption Using Time Series Data: A Machine Learning Approach

Supervisor : Mohd Azam Osman

Student Name : Muhammad Azzubair bin Azeman

Mentor : Patrick Lam Kar Jun

Practicum Company : DataMicron Systems Sdn Bhd

Introduction: Background of Company

- **DataMicron Systems Sdn Bhd** is a **consultant services** company which offers **big data solutions** and consultation services in **data analytics**.
- As **many companies** nowadays have their own databases, they **face difficulties** in **gaining insights** from raw data using relational database queries.
- This is where **DataMicron** comes as a solution provider in **implementing data warehouses** for data storage, equipped with **visualization tools** for data analytics.

Office Locations

Kuala Lumpur, Malaysia
DataMicron System Sdn.Bhd.

Suite 9-11
Wisma UOA II, No.21 Jalan Pinang
50450 Kuala Lumpur
MALAYSIA

Phone: (+603) 2163 3168
Fax: (+603) 2162 2168
Email: info@datamicron.com

United States
DataMicron

199383 Stevens Creek Blvd
Cupertino, CA 95014-2358
UNITED STATES

Phone: 408-359-8816
Fax: 408-725-8888
Email: info@datamicron.com

Singapore
DataMicron Pte. Ltd.

20 Maxvell Road,
#09-17, Maxwell House
SINGAPORE 069113

Phone: (65) 6408 9659
Fax: (65) 6234 4416
Email: info@datamicron.com

Introduction: Background of Domain

1. **What is Real Private Consumption (RPC)?**
    - The amount of goods and services consumed by households to fulfill their basic needs and wants (DoSM, 2020)

2. **Why is it important to predict RPC?**
    - RPC is the major contributor (58.7% in 2019) to Malaysia's GDP (Asada et al., 2019).

3. **What is the relation of RPC with GDP?**
    - RPC is one of the indicator of GDP :

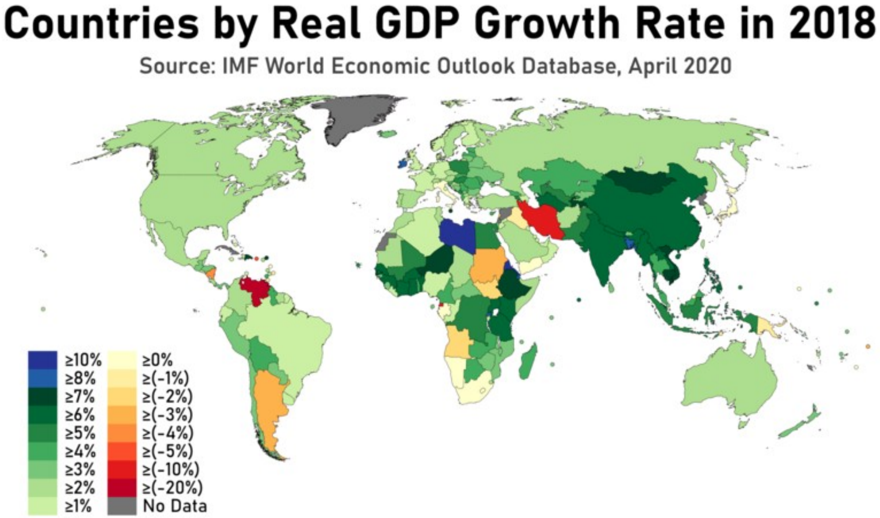
$$GDP = C + G + I + NX$$

- Where :
    - GDP = Gross Domestic Product
    - C = Private Consumption
    - G = Government Consumption
    - I = Investment
    - NX = Net Export (Import - Export)

Introduction: Background of Domain

4. **What is Gross Domestic Product?**

- “Market value of all final goods and services produced in economy annually” (Hashim et al., 2018)

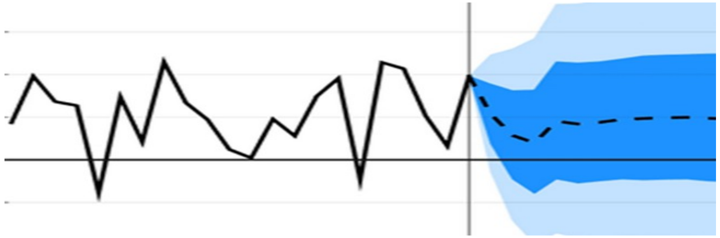


5. **Why it is important to be concern on GDP?**

- GDP determines economic advancement of a country. It distinguishes economic status of one country to another.

6. **How RPC is currently predicted by Government?**

- Mixed frequency vector auto regression (MFVAR)
- Mixed data sampling (MIDAS)
- Unrestricted error correction model (UECM)



Introduction: Problem Statement

- **Much studies have been done** on **comparing** between performances of **machine learning** models with **statistical** models especially in **economics** (Yu, 1999; Dematos et. al, 1996; Kumar, 2018).

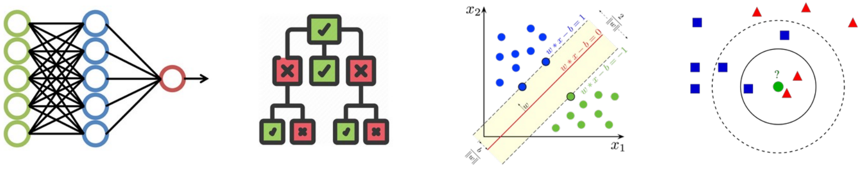
- However, **there is no specific study have been done in comparing model performances between machine learning models with statistical models for time series prediction of real private consumption in Malaysia**.

- The study of machine learning performance will contribute to the **understanding of ML approach for time series prediction of real private consumption** specifically and **macroeconomics** generally.

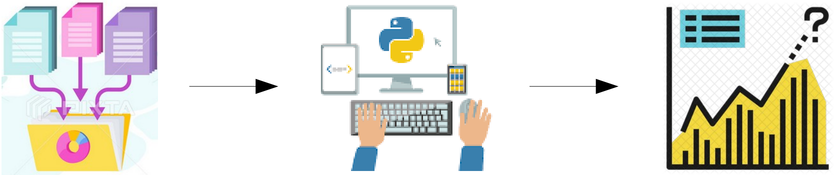
Introduction: Research Question

Throughout this project, the research questions are as following :

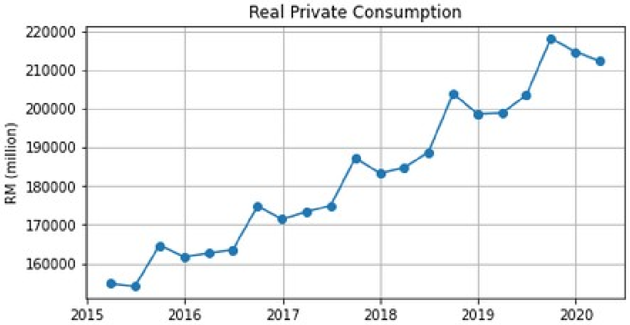
1) Which **machine learning model** is the **most suitable** for RPC prediction?



2) What are the **important steps** in developing machine learning models to **predict** RPC?



3) Between **statistical** and **machine learning** approaches, **which** of them is **better** in **model performance evaluation**?



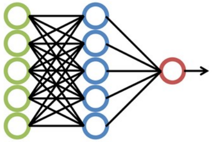
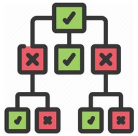
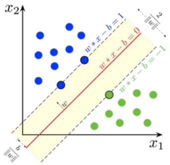
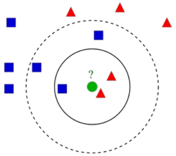
RPC (Q1 2020)
→ RM 212257 mil.

RPC (Q2 2020)
→ ?

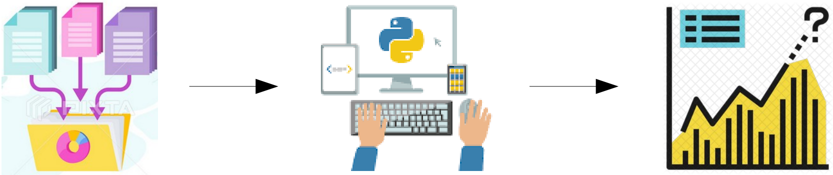
Introduction: Objectives

Following the research questions, the objectives are as following :

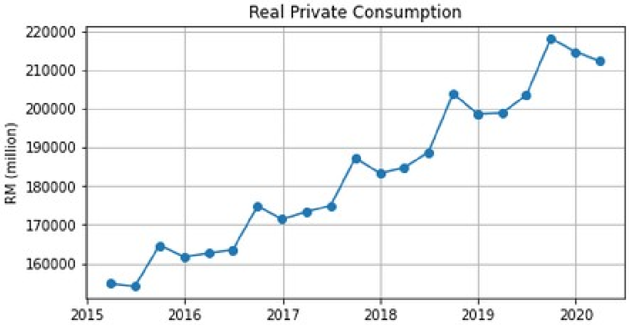
1) To **investigate** the **suitable machine learning technique** for RPC prediction

2) To **develop RPC prediction model** using the **selected machine learning** approach.



3) To **evaluate** prediction performance of the **RPC prediction models**.



RPC (Q1 2020)
→ RM 212257 mil.

RPC (Q2 2020)
→ ?

Introduction: Benefits of the Project

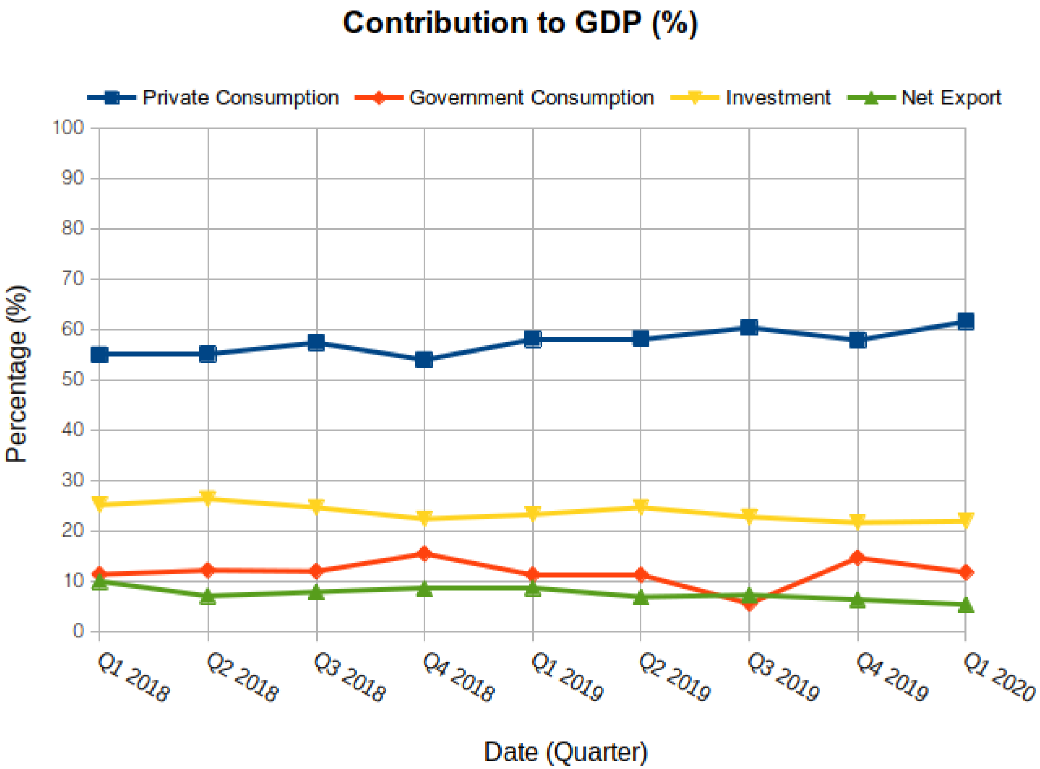
- This project is a **consultation project** between **DataMicron** with **Fiscal and Economics Division** of Malaysia's **Ministry of Finance** (MoF).

- Therefore, this project will benefit **Data Micron** in **providing proposed solution** for their client's problem.

- In return, **MoF** will **use these insights** to discuss with **MoF**'s top management in **considering new approach** for time series **prediction** of **real private consumption**.

Related Works: Review on Domain

- According **Ministry of Finance (2019)**, contribution to GDP is **dominantly** contributed by **Real Private Consumption** since 2018.
- Following RPC are **Investment**, **Real Government Consumption**, and finally **Net Export**.
- Moreover, **RPC's contribution** to GDP is **quarterly increasing** since 2018 (MoF, 2019).
- This shows that Malaysia's GDP is **significantly dependent** on **RPC**.



Source: Ministry of Finance (2020)

Related Works: Review on DSA Techniques

OFF TO THE RACES: A COMPARISON OF MACHINE LEARNING AND ALTERNATIVE DATA FOR PREDICTING ECONOMIC INDICATORS

Table 1 Model Performance Comparison for Quality Service Survey on RPC

	Statistics			Machine Learning				
Algorithm	4QMA	LASSO	Ridge	CART	RF	XGBoost	SVR	MARS
Normalised RMSE	0.23	0.04	0.07	0.11	0.05	0.05	0.10	0.13

Findings : **Tree-based Ensemble models** resulted on the **best predictions**. Such models are **random forests** and **gradient boosting**. The reason is because they can **learn nonlinear patterns** of economic indicators. (Chen et al., 2018)

Related Works: Review on DSA Techniques

Macroeconomic forecasting using factor models and machine learning: an application to Japan☆

Table 2 Model Performance Comparison for Macroeconomic Forecasting

	$h = 1$	$h = 2$	$h = 3$	$h = 6$	$h = 12$	$h = 18$	$h = 24$	$h = 30$	$h = 36$
(1) Best method under specification A to C									
IIP	B-lasso	C-RNN	B-RNN	C-RNN	C-RNN	B-lasso	C-CNN	C-lasso	B-EN
UTIL	C-EN	C-EN	C-EN	C-EN	C-EN	C-EN	C-EN	C-EN	C-EN
UR	A-FAAR	A-FAAR	A-FAAR	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost
WAGE	A-FAAR	A-FAAR	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost
CONS	C-lasso	C-lasso	C-lasso	C-EN	B-boost	B-bagging	B-bagging	B-bagging	B-boost
WPI	A-FAAR	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost	B-boost
CPI	B-EN	B-boost	B-RF	B-RF	B-boost	B-RF	B-boost	B-boost	B-boost

Findings : **Ensemble learning** based on **regression trees** (**bagging**, **random forests**, and **boosting**) is the **best method** due to their **adaptability with nonlinear trends** of macroeconomic indicators. (Maehashi and Shintani, 2020)

Related Works: Review on DSA Tools

Seglearn: A Python Package for Learning Sequences and Time Series

Findings : **seglearn** library has the **most feature** for **time series prediction**.

: This library also can **incorporate** classification, regression, clustering, and **forecasting tasks**.

: Most importantly, seglearn has **sliding window segmentation**, and adaptable to **sklearn** models (Burns and Whyne, 2018).

:

Table 3 Comparison of features between libraries

	tslearn	cesium-ml	ts-fresh	seglearn
Active development (2018)	✓	✓	✓	✓
Documentation	✓	✓	✓	✓
Unit Tests	✓	✓	✓	✓
Multivariate time series	✓	✓	✓	✓
Context data	X	✓	X	✓
Time series target	X	X	X	✓
Sliding window segmentation	X	X	X	✓
Temporal folds	X	X	X	✓
sklearn compatible model selection	X	X	X	✓
Feature representation learning	X	✓	✓	✓
Number of implemented features	N/A	58	64	20
Deep learning	X	X	X	✓
Classification	✓	✓	✓	✓
Clustering	✓	✓	✓	✓
Regression	✓	✓	✓	✓
Forecasting	X	✓	✓	✓

Research Methodology: Contribution

Table 4 Contribution Table on achieving the objectives

Objectives	Method used	Contribution
• To **investigate** the **suitable machine learning technique** for RPC prediction.	**Literature Review**	Most **suitable ML models**
• To **develop** **RPC prediction model** using the **selected machine learning** approach.	**Develop prediction models** using the most suitable models	**RPC prediction models**
• To **evaluate** prediction performance of the **RPC prediction models**.	**Compare performance** between: • Machine Learning models • ML vs Statistical models	**Accuracy** Result

Research Methodology: Problem Analysis

- Currently, **MoF** is finding a **methodology** to optimise **RPC predictions** in order to predict future values with **higher accuracy**.
- As of now, their best model is **MIDAS** followed by **UECM** and lastly **MFVAR**.
- This is indicated by their prediction error (RMSE) with **MIDAS** having the **least** while **MFVAR** having the **most error**.
- Could **Machine Learning** outperform these?

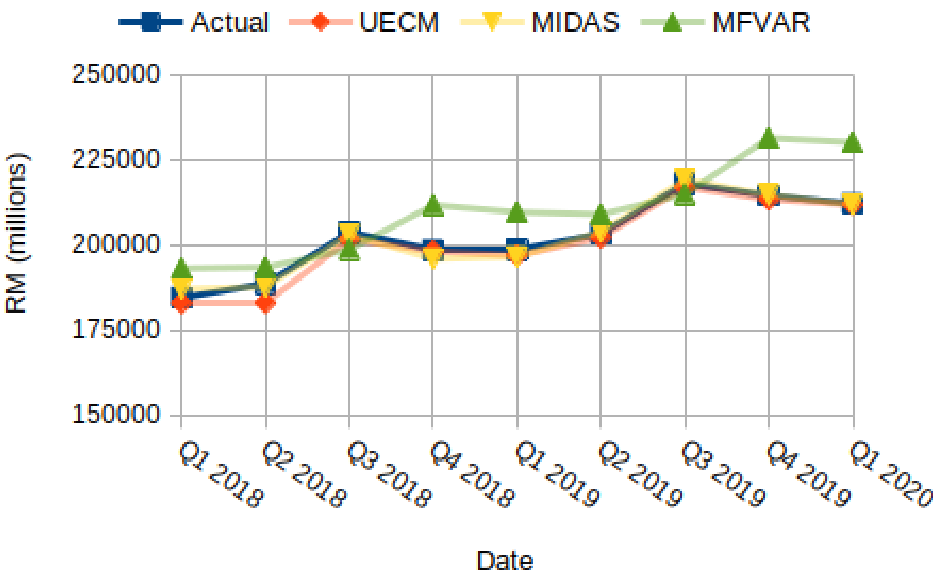
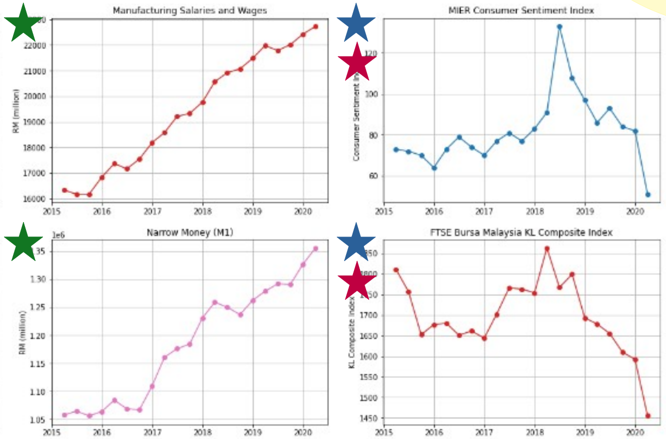


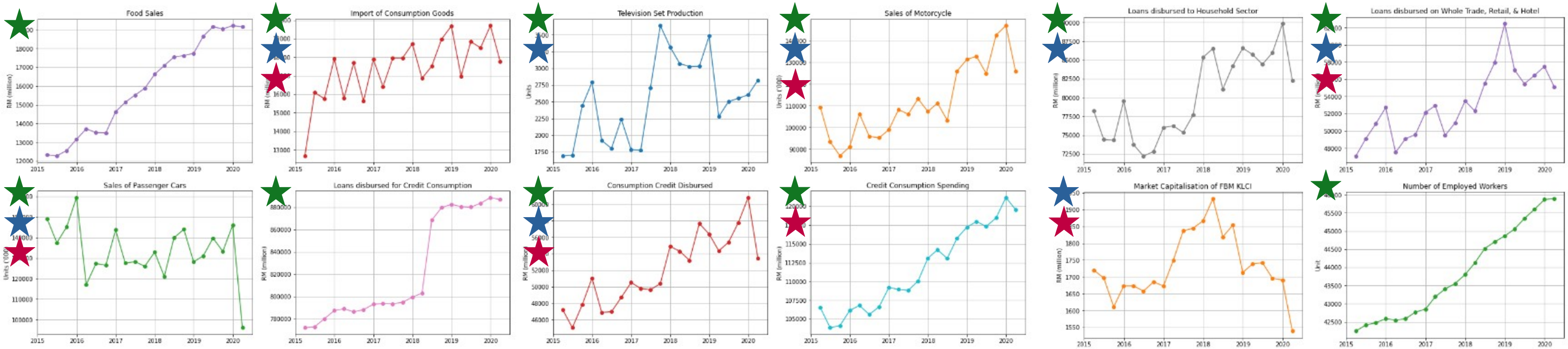
Figure 3 Current Prediction Methods of RPC

	2018Q1	2018Q2	2018Q3	2018Q4	2019Q1	2019Q2	2019Q3	2019Q4	2020Q1
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

Research Methodology: Exploratory Data Analysis

- **Prior** to the COVID-19 pandemic (**Q1 2020**), 13 out of 16 **RPC indicators** show an **increasing trend**.
- Overall, 11 **of the 16 indicators** show **nonlinear trends** except for MSW, NM, FS, EMP, and CC.
- During **Q1 2020**, 9 indicators **significantly declined**





Research Methodology: Final Analysis

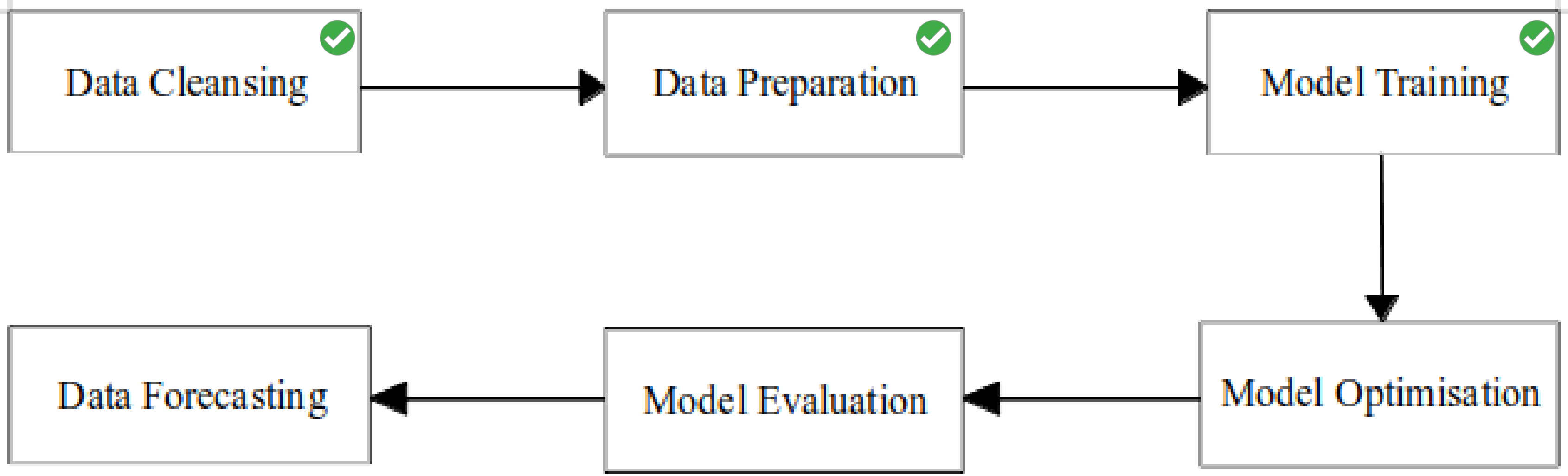


Figure 4 Flow chart of model development for RPC prediction models

Results and Discussion

Objective 1: To **investigate** the **suitable machine learning technique** for RPC prediction

- **Tree-based ensemble models** were the **most suitable models**
- Those models are:
  - **Bagging**
  - **Random Forest**
  - **AdaBoost**
  - **XGBoost**
- Reason:
  - **Capable** of predicting **nonlinear trends**

Table 4 Summary of Selected Literature Reviews

<table>
<tr><th rowspan="2">Author</th><th colspan="3">Algorithm</th><th rowspan="2">Findings</th></tr>
<tr><th>Parametric</th><th colspan="2">Non parametric</th></tr>
<tr><td>Chen et al. (2019)</td><td>• 4QMA
• LASSO
• Ridge</td><td colspan="2">• CART
• RF
• XGBoost
• SVR
• MARS</td><td>• Tree-based ensemble models such as RF, and XGBoost were the most accurate models.
• Due to underfitting of MARS and 4QMA models, they had poor prediction performance</td></tr>
<tr><td rowspan="2">Maehashi and Shintani (2020)</td><td>Linear</td><td>Ensemble</td><td>Neural Network</td><td rowspan="2">• Tree-based ensemble models were the majority of the best models.
• Large window size is recommended for better time series predictions.</td></tr>
<tr><td>• LASSO
• Ridge
• EN</td><td>• Bagging
• RF
• AdaBoost</td><td>• FFNN
• CNN
• LSTM</td></tr>
</table>

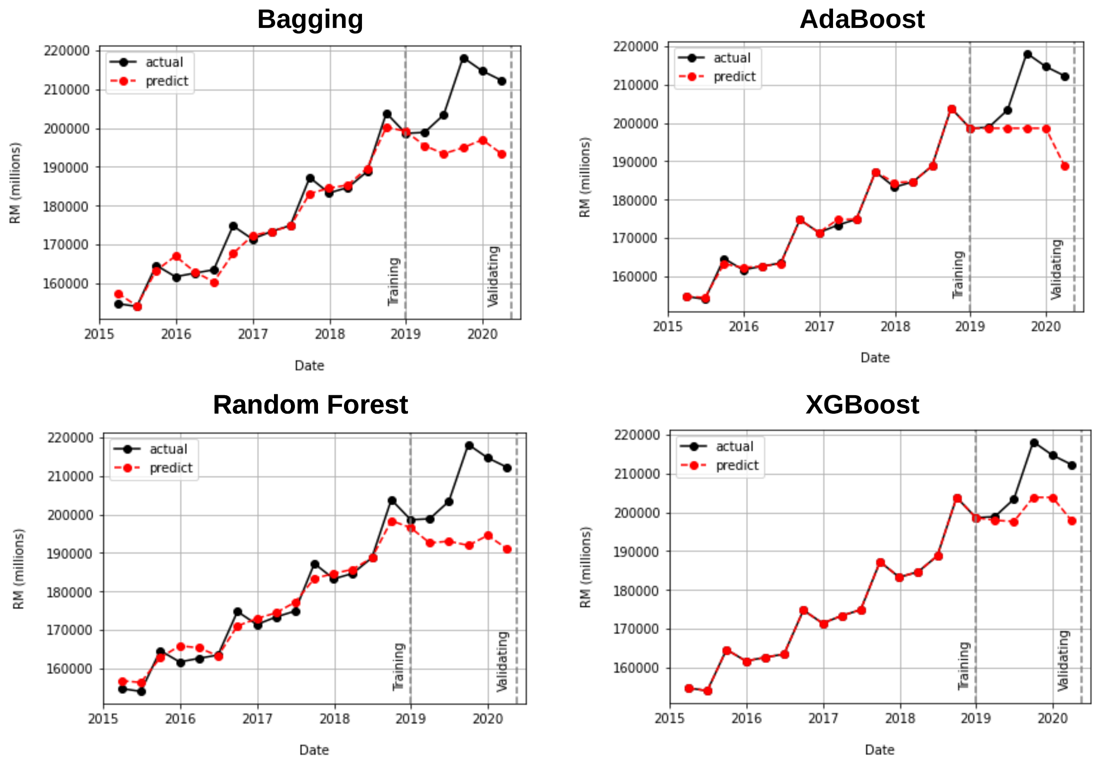
Results and Discussion

Objective 2: To **develop** **RPC prediction model** using **selected machine learning** approach.

Table 5 Model Performance Comparison

	Bagging	RF	AdaBoost	XGBoost
RMSE	16264	18359	15594	10633

- Currently, **Tree-based ensemble models** with **default parameters** were successfully developed.
- After this, these models will be **optimised** to **reduce** their **RMSE** values as much as possible.



Conclusion

Currently, this project can be concluded as stated in the following table:

Table 4 Conclusion Summary

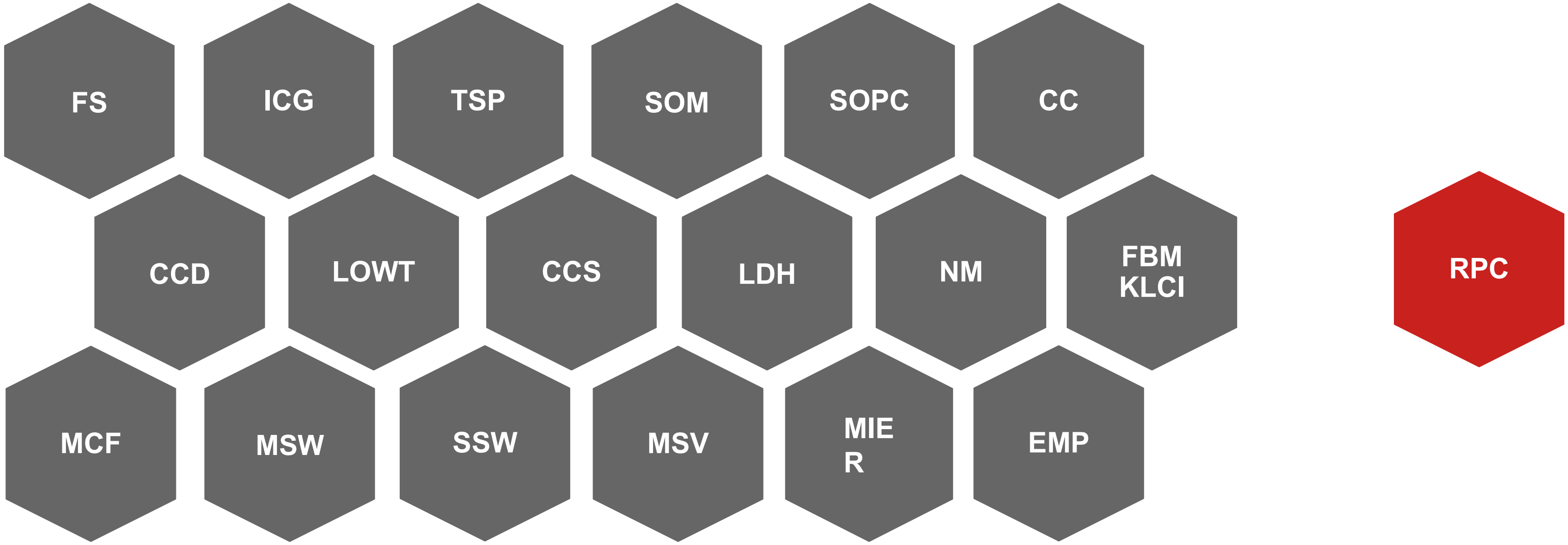
Objectives	Status of Achievement	Findings
• To **investigate** the **suitable machine learning technique** for RPC prediction.	Achieved	• **Tree-based ensemble models** were the **most suitable models**
• To **develop** **RPC prediction model** using the **selected machine learning** approach.	Ongoing	**-**
• To **evaluate** prediction performance of the **RPC prediction models**.	Ongoing	**-**

Thank You

Backup Slides

Related Works: Review on Domain

- In particular, RPC is indicated by 18 indicators as published in MoF (2017).



Source: Ministry of Finance (2017)

Research Methodology: Activities Plan

- From **October** until early of **November**, I spent most on the time **reviewing articles** to **select** the best **algorithms** and **tools**.

- Currently, I am **still on track** with my Gantt Chart as I have **developed basic models** for all of the algorithm selected.

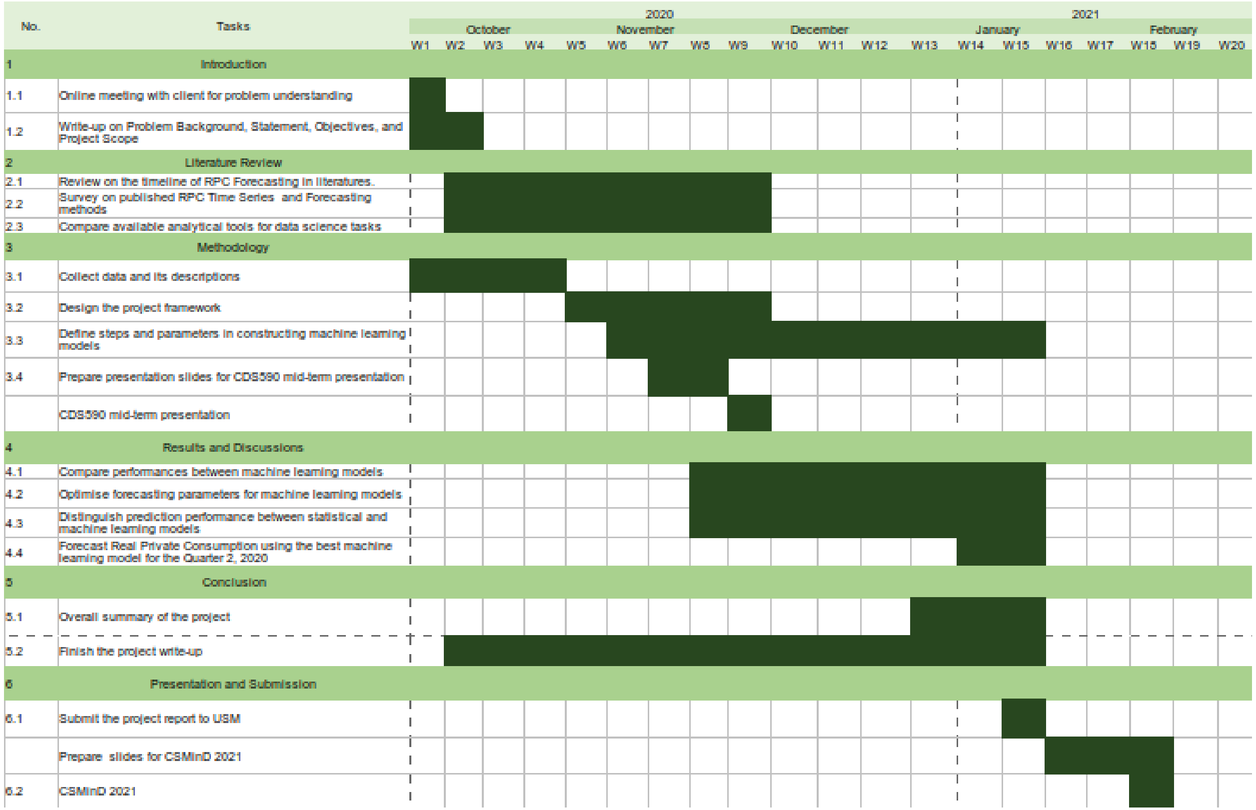


Figure 1 Gantt Chart of Project Consultancy and Practicum

Research Methodology: DS Lifecycle

- As a **data scientist**, my lifecycle has been **rolling throughout** the **data science lifecycle**.
- I spent **most of the time** in **modeling phase** back-and-forth between **Feature Engineering** and **Model Evaluation**.
- After satisfied with model evaluation, I will **propose** the **machine learning** models for the **new approach** seek by Fiscal and Economics Division, Ministry of Finance, Malaysia.

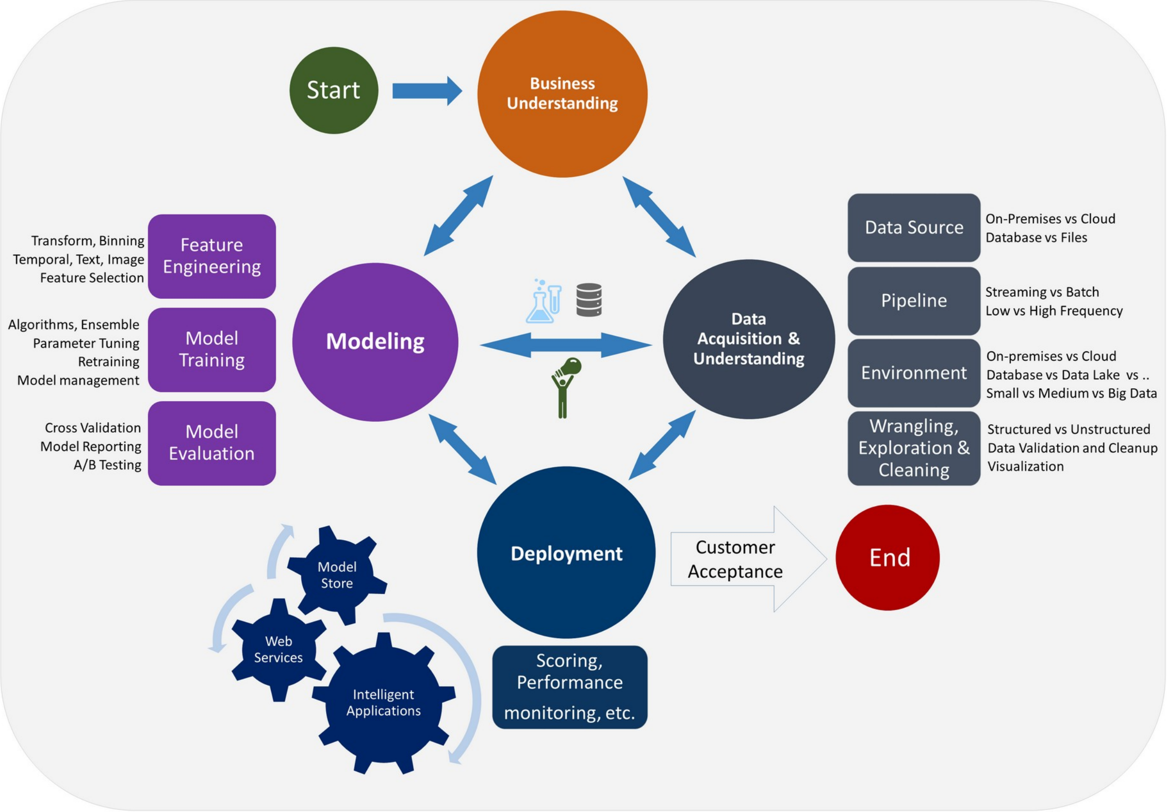


Figure 2 Data Science Lifecycle Projects

Results and Discussion: Data Cleansing

Transform Data

Monthly to Quarterly

Date	FS	ICG	TSP	SOM
1995Q1	1714.600000	2448.700000	2036.70000	75641.0
1995Q2	1771.000000	2756.300000	2223.90000	81561.0
1995Q3	1775.800000	3022.800000	2618.60000	86385.0
1995Q4	1782.200000	3203.100000	2501.60000	71463.0
1996Q1	1882.600000	2890.800000	2330.40000	64806.0
...	...	...	...	...
2019Q2	19154.941736	18837.769720	2501.83600	124764.0
2019Q3	19038.573194	18504.265237	2552.02500	142484.0
2019Q4	19217.321673	19704.177128	2602.99800	146849.0
2020Q1	19148.158828	17777.865886	2822.28515	125747.0
2020Q2	6220.979782	5027.206052	700.14000	1.0

Change Dataset Timeline

2015 - 2020

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

Remove Insufficient Attributes

SSW, & MSV are discarded

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

Figure 7 Output of each process in Data Cleansing phase

Results and Discussion: Data Preparation

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

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

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