**Assessment of Investment Attractiveness in the world - First Progress Report**

Team | Accenture 4

Mentor | Pramanik Paritosh

Group Members | Di Mu(dm3686), Freddy Wong(ww2615), Hanlin Yan(hy2654),

Jace Yang (jy3174), Yuan Heng(yh3416)

[**Introduction**](#_qiw2bgcs14lm) **2**

[**Data**](#_7yrytqg7129w) **3**

[2.1 Data Description](#_qd94eyiorc3j) 3

[2.2 Target variable](#_jcw5jps28shf) 3

[2.3 Literature review](#_jecpbotxg3vt) 4

[2.4 Variable Selection](#_1mayr0m8wu0f) 4

[2.4.1 Manufacturing Variable Selection](#_4u21c7cp6p89) 6

[2.4.2 Service Variable Selection](#_27q7gvf24bnf) 8

[2.5 Country Selection](#_6vcoo4uilfs7) 10

[2.6 EDA](#_5y7b353931e8) 11

[2.7 Missing Value Imputation](#_laiplfgbaena) 11

[**Methods**](#_59f949a47o9i) **14**

[3.1 VAR model](#_ata4dwkvz2ua) 14

[3.2 ARIMA model](#_n3quxktug0un) 15

[3.3 LSTM model](#_1ik4girogyal) 16

[3.3.1 Architecture](#_fogmqdbodw0a) 16

[3.4 Prophet model](#_nosncy4bt0b6) 17

[**Results**](#_1uvbnl1lhb7n) **17**

[4.1 Metrics](#_g8di5nalxa07) 17

[4.2 Imputation Performance](#_3s2fqxtxlb59) 17

[4.3 Seasonality Check](#_dgu0m2lcrgax) 17

[**Goals and Next Steps**](#_ypuvlo4bre7f) **18**

[5.1 Univariate vs. Multivariate](#_ff9zc8h6623u) 18

[5.2 Hyper-parameter tuning](#_zbymd6rze3yd) 19

[5.3 Bayesian methods](#_br27b3p3n9ue) 19

[**Individual contributions**](#_h9lqkmoohgmr) **19**

### 

### **Introduction**

Business investment is a major pillar of economic growth. Attracting foreign investors and capitals leads to the increase of consumer spending, international trading, as well as the local investment, which could all have a great impact on production of goods and services in a certain economy. However there are a lot of influential factors that determine new investment. For instance, political factors like political stability and industrial regulations, economic condition factors like advanced level of science and technology, as well as the factors that are crucial to their direct profit, such as production cost and exchange rate.

Though factors that influence the attractiveness of business investment are countless, and there are a variety of definitions in business investment, we could roughly classify industry into manufacturing sector and service sector. Manufacturing industries primarily produce physical and tangible goods and typically have low customer contact. By contrast, service sectors generally produce intangible products that cannot be produced ahead of time. We use “*Manufacturing value added as % of GDP*” and “*Services value added % of GDP*” to represent the attractiveness of a country for business investment in Manufacturing and service, which are available in the World Bank.

That’s to say, for a given country, we trained two separate models: One for the manufacturing sector, the other for the service sector. We started from univariate modeling, and then moved to multivariate modeling. With initial variable choosing based on prior knowledge of the market and the economy, and variable selection according to the seasonality check and other variable selection techniques, we formed the predicted variables that can be used to fit the two industries. However, the missing data problem has consistently been the primary and impactful issue for all the models. To overcome the issue, we thus tried multiple imputation techniques.

To obtain the best model, we tried models from traditional statistical models to deep neural frameworks, including but not limited to ARIMA, VAR, LSTM and Prophet. For the evaluation methods, we chose popular metrics such as MAE, RMSE, and MAPE. We also performed seasonality check and residual analysis to validate the models.

In order to extrapolate our observations and insights into the global world. We picked nine countries based on the data availability and the representativeness of the country. Though depending on factors like the countries’ economic stability and economic conditions, the performance of the models vary. We still achieved insightful conclusions. In the following sections, we would introduce our data, methods, and results. Lastly, we would list some of our ideas regarding what we would improve in the next steps.

### **Data**

#### **2.1 Data Description**

Statistics and data are a key part of that knowledge and are easily accessible on the web for all users. The World Bank provides free and open access to a comprehensive set of data about development in countries around the globe, together with other datasets cited in the data catalog.We get a great amount of public data from the dataset of the World Bank: <https://data.worldbank.org/>

The world bank stored their data in separated files according to their country and other indicators. We have to choose one target variable which can reflect the attractiveness of the country and various predictor variables to obtain the future target value. What’s more, there are some missing values in some specific years. Under such cases, we should impute the missing data programmatically.

#### **2.2 Target variable**

After careful consideration, we decided to divide the attractiveness of one country into two parts: manufacturing and service.

* For the manufacturing, we choose the predictor “Services\_value\_added\_%\_of\_GDP”. Manufacturing value added is the net output of a sector after adding up all outputs minus intermediate inputs. It also ignores the depreciation of goods. “Manufacturing\_value added\_%\_of\_GDP” reflects a country’s attractiveness for business investment in Manufacturing.
* As for service, we pick the predictor “Services\_value\_added\_%\_of\_GDP”. It is the value added in wholesale and retail service. Service value added is the net output of a sector after adding up all outputs minus intermediate inputs. It reflects the attractiveness of business investment in Service.

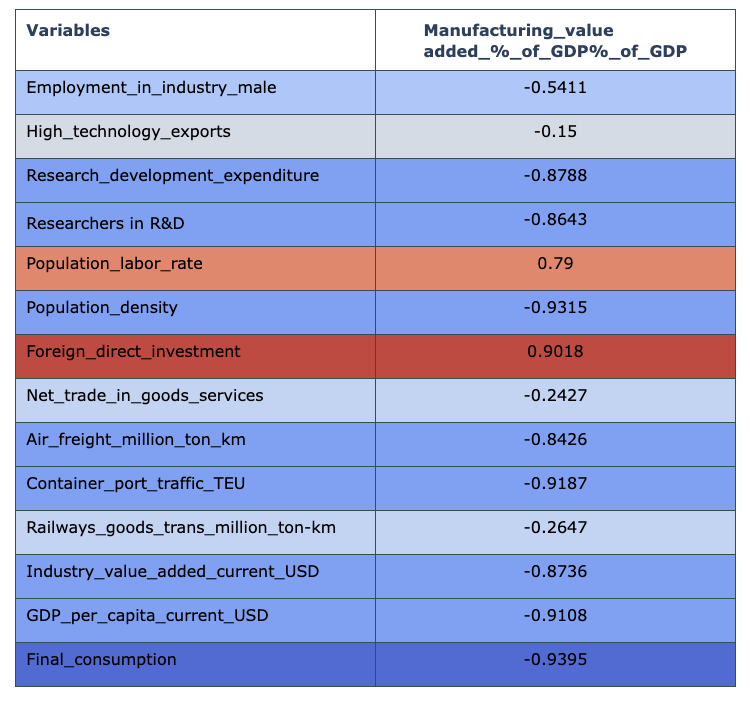
One of our final deliverables will be the trend analysis of manufacturing attractiveness *“Manufacturing\_value\_added\_%\_of\_GDP”* and service industry attractiveness *“Services\_value\_added\_%\_of\_GDP”*. It is anticipated that these two variables will have a close connection with the attractiveness for business investment in each country.

#### **2.3 Literature review**

***Data consolidation***

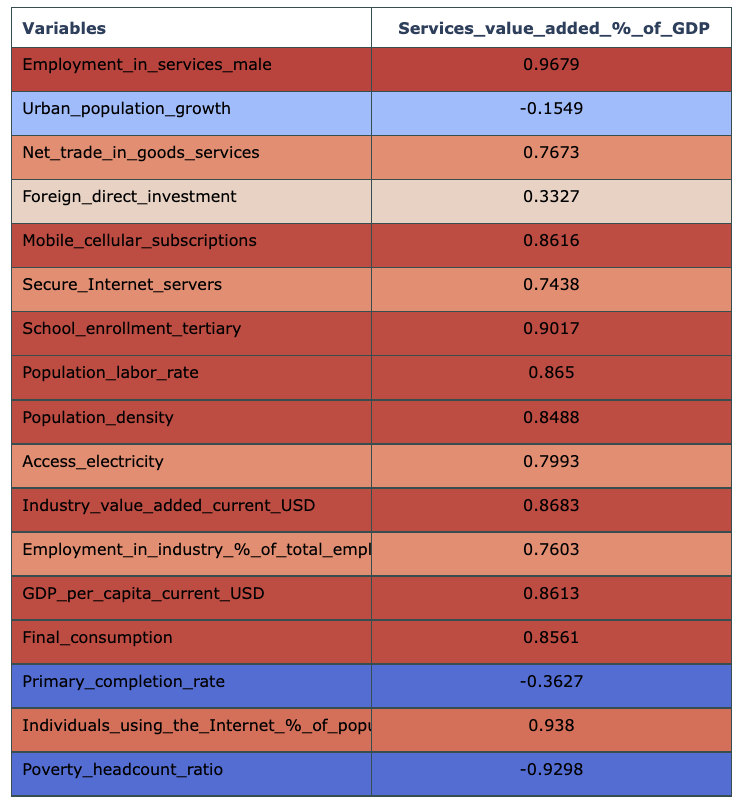
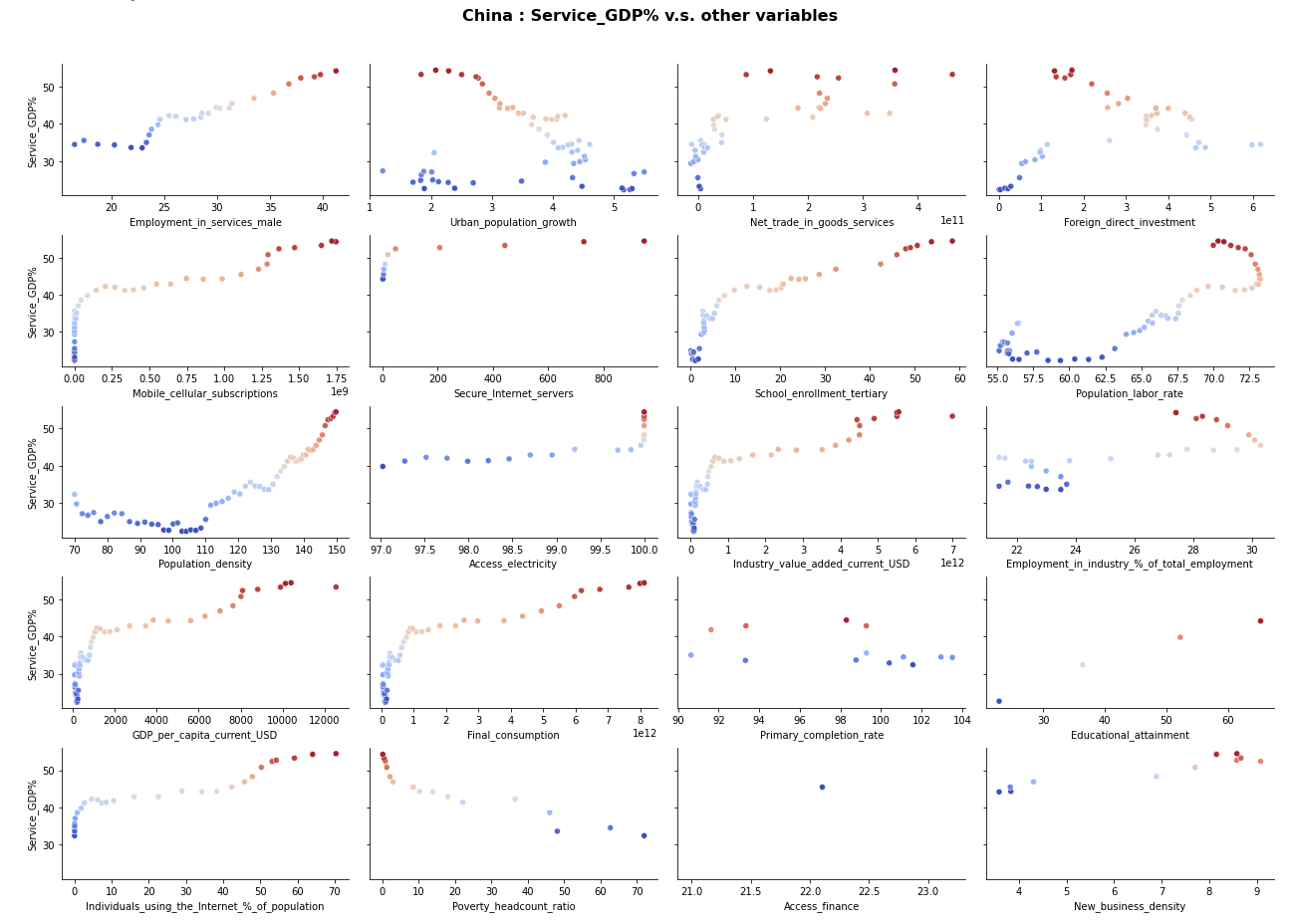
#### **2.4 EDA & Variable Selection**

We made scatter plots of the target variable v.s. our initial selected variables based on prior knowledge for each country. For example, the plot above shows the relationship between manufacturing industry’s attractiveness (*Manufacturing\_value\_added\_%\_of\_GDP* ) and other potential predicted variables.



The severe problem of missing data on the *Access\_Finance* *(% of firms receiving bank investment)* is reflected in the graph, making us decide to drop this variable. It remains unknown regarding the relationship between the manufacturing attractiveness and *Employment\_in\_industry\_rate*, *High\_technology\_exports, Net\_trade\_in\_goods\_services, and railways\_goods\_trans\_million\_ton-km,* as there is no pattern in the scatter plot.

On the contrary, the other variables all have an obvious linear relationship, either positive or negative, with the target. On the other hand, the correlation table plots the correlation coefficient between the target and the predicted variables after removing those with a lot of missing data. Most of the variables indicate a strong correlation, suggesting that we have chosen the right candidate variables.



For the service industry, we performed a similar analysis. The *educational\_attainment, access\_finance* are decided to be dropped due to the severe issue of missing data. The variable *primary\_completion\_rate* and *New\_business\_density* seems not very useful in explaining the attractiveness of the service industry. Other remaining variables all have some relationship with the target. The correlation coefficient table on the left side further confirms that our selection of variables are useful and suitable.

##### **2.4.1 Manufacturing Variable Selection**

There are a lot of variables that have an impact on the attractiveness of business investment in manufacturing. The following factors will significantly affect the attractiveness:

* **Infrastructure**: If the manufacturing industry of one country is attractive, it must have a great scale and a promising future. More specifically, the country should have a suitable infrastructure for the growth of the manufacturing companies.The transportation system should be stable and effective so that the company can sell their products all over the country. And the output of the manufacturing should exceed the input in them
* **Labor Force**: The manufacturing companies can produce many products and make a profit only if they employ enough workers. Thus the labor force is a vital session which is considered by all companies. They will consider whether the country has an abundant labor force, and whether the country will have a stable labor force in the long term.
* **Consumption**: As the manufacturing company needs to sell their products, the customers in this country must be able to afford the products. The more products they sell, the more profit they can obtain.
* **Education**: Some manufacturing companies may need more advanced and professional employees, so the education situation of the workers in this country is another important factor that can attract the investment. As the foreign labor is less and more expensive, the manufacturing industry is more likely to grow their production in the country with more educational and professional workers.
* **Technology**: The company with more advanced technology is more popular. If the company invests a lot of money into their research and development department, it is going to produce the products with higher standards. And the number of researchers in a country also reflects the basic level of the manufacturing industry in it.
* **Financial**: As the investment is essential to the growth of a manufacturing company, the ability to get more investments and some loans is a plus factor for most companies. Thus they’d like to choose some countries with more banks or lower taxes to establish their branch companies.Under such a situation, it will attract more investment in their manufacturing industry.

| **No** | **Index** | **Short Description** | **Unit** |
| --- | --- | --- | --- |
| 1 | Air\_freight\_million\_ton\_km | Air freight in million ton-km measured in metric tons times kilometer traveled. | Million ton-km |
| 2 | Container\_port\_traffic\_TEU | Container port traffic (TEU:20-foot equivalent units) is the shipment of containers from port; inbound and outbound. | Number of containers |
| 3 | Railways\_goods\_trans\_million\_ton-km | Goods transport by railways in million tons measured in metric tons times kilometer traveled. | Million ton-km |
| 4 | Logistic\_performance | It is the quality of the logistic related infrastructure (1 being low and 5 being high). In this index, trade, and transport related infrastructure are evaluated. | Numeric index (1=lowest to 5=highest) |
| 5 | Industry\_value\_added\_current\_USD | Total output in manufacturing, industry, construction, power, water, mining etc. in USD. It is the net output of a sector after adding up all outputs minus intermediate inputs and assets depreciation. | USD |
| 6 | GDP\_per\_capita\_current\_USD | It is the GDP divided by the total midyear population of a particular country. | USD |
| 7 | Final\_consumption | Final consumption or total consumption expenditure is the expenditure for household final consumption plus the government final consumption. | USD |
| 8 | Access\_finance | Access to finance is the percentage of firms using banks to finance investments. | Percentage of firms using bank |
| 9 | New\_business\_density | The number of new limited liability companies registered in a year per 1000 people (ages between 15-64). | Number of business entity per 1000 people per year |
| 10 | Employment\_in\_industry\_male | The number of male employees in industries, divided by the number of all employees | Percentage |
| 11 | High\_technology\_exports | The sum of high technology products which be exported, divided by the sim of exported manufacturing products | Percentage |
| 12 | Research\_developement\_expenditure | It is the research and development expenditure divided by the total GDP. | Percentage |
| 13 | Researchers in R & D | It is the number of researchers in research and development. | Per Million People |
| 14 | Population\_labor\_rate | The percentage of population ages 15-64 | Percentage |
| 15 | Population\_density | The population density | Number of people pre sq.km land area |
| 16 | Foreign\_direct\_investment | It refers to the direct investment equity flow in an economy. It is the sum of equity capital, reinvestment of earnings and other capital. | USD |
| 17 | Net\_trade\_in\_goods\_and\_service | It is derived by offsetting imports goods and service against exports goods and service. | USD |

##### **2.4.2 Service Variable Selection**

The variables which can have a great impact on the attractiveness in service can also be divided into several categories:

* **Infrastructure**: If the service industry of one country is attractive, it must have a great scale and a promising future. More detailed information, the country should have a suitable infrastructure for the growth of the service companies. They should obtain access to electricity. And it should have a stable and effective internet system.
* **Labor Force**: The service companies can provide a lot of services and make a profit only if they employ enough workers. Thus the labor force is a vital session which is considered by all companies. They will consider whether the country has an abundant labor force, and whether the country will have a stable labor force in the long term.
* **Consumption**: As the service company needs to sell their products, the customers in this country must be able to afford the service. The more services they provide, the more profit they can obtain.
* **Education**: Some service companies may need more advanced and professional employees, so the education situation of the workers in this country is another important factor that can attract the investment. As the foreign labor is less and more expensive, the service industry is more likely to grow their production in the country with more educational and professional workers.
* **Internet**: The Internet is essential to the manufacturing industry. They can provide a great number of services on the internet and enhance their product and management system. A stable and fast internet environment is a catalyst for the growth of the service industry.
* **Financial**: As the investment is vital to the growth of a service company, the ability to get more investments and some loans is a plus factor for most companies. Thus they’d like to choose some countries with more banks or lower taxes to establish their branch companies.Under such a situation, it will prefer the country which can provide many domestic investments and appeal to lots of foreign investors.

| **No** | **Index** | **Short Description** | **Unit** |
| --- | --- | --- | --- |
| 1 | Access\_electricity | Percentage of population with access to electricity. | Percentage of a country's population |
| 2 | Industry\_value\_added\_current\_USD | Total output in manufacturing, industry, construction, power, water, mining etc. in USD. | USD |
| 3 | Employment\_in\_industry\_%\_of\_total\_employment | It is the person’s occupation in an industry to produce goods or services for an earning. The industry sector includes mining and quarrying, manufacturing, construction, and public utilities. | Percentage of total employment |
| 4 | GDP\_per\_capita\_current\_USD | It is the GDP divided by the total midyear population of a particular country. | USD |
| 5 | Final\_consumption | Final consumption or total consumption expenditure is the expenditure for household final consumption plus the government final consumption. | USD |
| 6 | Primary\_completion\_rate | The number of enrollments minus the number of people who are repeating in the last grade of the primary school, divided by the population at the last grade. | Percentage |
| 7 | Educational\_attainment | The percentage of people ages 25 and above who completed education in post-secondary non-tertiary level. | Percentage |
| 8 | Individuals\_using\_the\_Internet\_%\_of\_population | Internet users are individuals who have accessed the internet recently using a computer, cell phone, or by any device. | Percentage of population |
| 9 | Poverty\_headcount\_ratio | It is the percentage of the population who are earning less than $1.90 per day at 2011 international price. | Percentage of population |
| 10 | Access\_finance | Access to finance is the percentage of firms using banks to finance investments. | Percentage of firms using bank |
| 11 | New\_business\_density | The number of new limited liability companies registered in a year per 1000 people (ages between 15-64). | Number of business entity per 1000 people per year |
| 12 | Employment\_in\_industry\_male | The number of male employees in industries, divided by the number of all employees | Percentage |
| 13 | Urban\_population\_growth | It is the growth of the percentage of people living in the urban areas | Percentage |
| 14 | Mobile\_Celluar\_subscription | It is the subscriptions to a public mobile telephone service that provide access to PSTN using cellular technology. | Number per 100 people |
| 15 | Secure\_Internet\_Service | The number of distinct public-trusted internet server | Number Per Million People |
| 16 | School\_enrollment\_tertiary | The percentage of people with successful completion of education at the secondary level | Percentage |
| 16 | Population\_labor\_rate | The percentage of population ages 15-64 | Percentage |
| 17 | Population\_density | The population density | Number of people pre sq.km land area |
| 18 | Foreign\_direct\_investment | It refers to the direct investment equity flow in an economy. It is the sum of equity capital, reinvestment of earnings and other capital. | USD |
| 19 | Net\_trade\_in\_goods\_and\_service | It is derived by offsetting imports goods and service against exports goods and service. | USD |

#### **2.5 Country Selection**

When choosing the countries, we consider mainly two aspects: whether the country has a lot of unique characters and can represent a range of countries; whether most of the variables of the country are available in the world bank.

In the first stage, we choose ten countries: China, the United Kingdom, the United States, Hong Kong, Iran, Russia, South Africa, India, Italy and Vietnam. These ten countries are located on several continents and contain both developed and developing countries. They also include countries which have been proven to be attractive to the investments in the past years and countries that are anticipated to be attractive to the investors. However, when we try to fit some models, we find that some countries have so many missing values that can not be ignored. And massive imputation does not make sense in this project. As we have tried various implementation methods, the predicted result seems not good. As a result, we decided to pick some other countries.

After the modification, we choose nine countries: China, the United Kingdom, the United States, Japan, India, Canada, Singapore, Australia and Brazil. These countries are representative as well and have much less missing values.For example, the variable “*Manufacturing\_value\_added\_%\_of\_GDP*” in Singapore, Brazil and India starts from 1960, which is suitable for our univariant prediction model.Therefore we decide to move forward our project using the current nine countries.

#### 

#### **2.6 Missing Value Imputation**

Even Though variables are selected with the consideration of data availability to minimize the amount of missing values, there are still missing values due to the timing of data collection. It is common to see that many metrics become available after a specific year. In most of the cases, the data is continuous after the year of first collection while some are collected every other year. Almost none of the variables are missing at random. Imputation was carried out with the following methods

| **Imputation Methods** | **Short Description** |
| --- | --- |
| Mean | Imputes with the average value of all the data already given in the time series |
| Median | Imputes with the median value of all the data already given in the time series |
| Last Observation Carried Forward (LOCF) | Previous non-missing values are carried foward and replaced with the missing values |
| Next Observation Carried Backward (NOCB) | The next non-missing values are copied with the previous missing values |
| Rolling Statistics  Simple Moving Average  Weighted Moving Average  Exponential (Weighted) Moving s Average |  |
| Interpolation  Linear interpolation    Spline interpolation | Estimates unknown values by assuming linear relation within a range of data points  Using a mathematical function, the method estimates values that minimize overall curvature, thus obtaining a smooth surface passing through the input points. |
| Multivariate Imputation by Chained Equation (MICE) | Calculates the mean of every column with missing values as a placeholder. A series of regression model is run to impute each missing value sequentially by constantly changing the placeholder with predictions from previous iteration. |

Data summary

|  | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Manufacture | | | | | | | | | |
| # of instances | 62 | 62 | 62 | 62 | 62 | 62 | 62 | 62 | 62 |
| # of  features | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| Missingness  in target  variable | Yes | No | Yes | Yes | Yes | No | Yes | No | Yes |
| Missing  rate | 41% | 38% | 38% | 36% | 38% | 36% | 40% | 36% | 40% |
| Service dataset | | | | | | | | | |
| # of  features | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Missingness  in target  variable | Yes | No | Yes | No | Yes | No | Yes | No | Yes |
| Missing  rate (%) | 41% | 37% | 41% | 39% | 38% | 37% | 43% | 40% | 40% |

Missing patterns

Comparison of performance based on experimental data

#### 

#### 

### 

### **Methods**

#### **3.1 VAR model**

#### 

#### **3.2 ARIMA model**

#### 

#### 

#### **3.3 LSTM model**

##### **3.3.1 Architecture**

The plot above illustrates the architecture of the LSTM model. It has two layers of LSTM cells, one dropout layer and one dense layer as the output. The input of the model

#### 

#### **3.4 Prophet model**

### **Results**

#### **4.1 Metrics**

#### **4.2 Imputation Performance**

#### **4.3 Seasonality Check**

### **Goals and Next Steps**

#### **5.1 Univariate vs. Multivariate**

#### **5.2 Hyper-parameter tuning**

#### **5.3 Bayesian methods**

### **Individual contributions**

* **Di Mu:**
* **Freddy Wong:**
* **Hanlin Yan:**
* **Jace Yang:**
* **Yuan Heng:**

### **Appendix**

### **Reference**