

Understanding customer behavior on credit card receivables with the effect of macroeconomic variables

A Dissertation submitted for the degree of

Master of Science

In

Data Analysis for Business Intelligence

by

Sai Krishna Vadlamudi

Student number: 209054868

Under the guidance of

Alexander Gorban(prof.)

Department of Mathematics

University of Leicester

England

May 2022



**UNIVERSITY OF
LEICESTER**

Acknowledgment

I would like to acknowledge my profound gratitude and thank my academic supervisor, **Alexander Gorban(prof.)**, Department of Mathematics, University of Leicester, for his esteemed guidance, unending inspiration, and constant encouragement throughout my project work. My association with him has given me an excellent opportunity to satisfy my thirst for knowledge.

A massive thank you to industry supervisor, **Jeremy Levesley**, and the Program Director of DABI, **Andrew Morozov (Dr.)** for their efforts and for welcoming me into this real-time industry project.

Contents

List of Figures	5
List of Tables	7
Abstract	8
Introduction	9
Credit Cards.....	9
Credit card businesses and their operations	9
Scope of credit card spending in Hong Kong.....	11
Hong Kong's response to Corona Virus outbreak	14
Public response.....	14
Government's response.....	15
About variables used in the analysis	16
Gross Domestic Product.....	16
Private consumption expenditure.....	16
Government consumption expenditure.....	16
Gross domestic fixed capital formation.....	17
Changes in inventories.....	17
Exports and Imports of goods(f.o.b.).....	17
Exports and Imports of services	18
Unemployment rate (%)	18
Interest rate (%).....	19
Credit card – Total number of accounts	19
Credit card – Average total receivables	19
Credit card – Charge off amount	20
Credit card – Delinquent amount.....	20
Credit card – Rollover amount.....	20
Why is it important to know the credit card receivables?	21
Data Collection	22
About the data source.....	22
Data Pre-processing.....	23
Exploratory Data Analysis	24
Countplot.....	24
Line graphs showing the last 22 years trend of the variables	25
Correlation Plot	30

Feature Selection.....	32
Feature selection based on feature importance	33
Feature selection based on the p-value	35
Feature Engineering	36
Standardization using standard scalar	36
Modelling and Optimization.....	37
Multiple Linear Regression Analysis	37
Base Line Model.....	38
Principal Component Analysis (PCA)	38
Multiple Linear Regression Analysis	40
Decision Tree Regressor	40
Random Forest Regressor	41
XGboost Regressor	41
Results and discussion.....	42
Conclusion	44
Future Work.....	45
References.....	46
Appendix.....	49
Code	49
LinkedIn Posts.....	56
LinkedIn Post – 1.....	56
LinkedIn Post – 2	57
LinkedIn Post - 3.....	59

List of Figures

Figure 1 How revenue is generated through credit cards; Source: Moneygeek	10
Figure 2 Consumer spending in Hong Kong; Source: Trading Economics and Census and Statistics department of Hong Kong.....	11
Figure 3 E-Commerce payment method split by value, Source: J.P. Morgan 2019 Payments Trends ^[4]	12
Figure 4 Top e-commerce merchant segments, Source: J.P. Morgan 2019 Payments Trends ^[4]	12
Figure 5 Data Range Distribution (2000 – 2021)	24
Figure 6 Gross Domestic Product measures (2000 – 2021)	25
Figure 7 Private and Government Consumption Expenditure (2000 – 2021)	25
Figure 8 Gross Domestic Fixed Capital Formation (2000 – 2021).....	26
Figure 9 Changes in Inventories (2000 – 2021)	26
Figure 10 Comparison of Exports and Imports of Goods and Services (2000 – 2021)	27
Figure 11 Unemployment rate (2000 – 2021)	27
Figure 12 Interest Rate (2000 – 2021)	28
Figure 13 Total of Credit Card accounts (2000 – 2021).....	28
Figure 14 Credit Card Charge off Amount (2000 – 2021).....	29
Figure 15 Credit Card Rollover Amount (2000 – 2021).....	29
Figure 16 Average Total Receivables of Credit Card (2000 – 2021)	30
Figure 17 Correlation Plot: Multivariate Statistical Analysis.....	30
Figure 18 Graph of Feature Importance and values	34
Figure 19 P-values from OLS model	35
Figure 20 Head of the data after dropping values based on p-values	35
Figure 21 Head of the standardized data	36
Figure 22 Baseline model of Linear Regression, Model prediction comparison	38
Figure 23 Mean Squared Error and Mean Absolute Error graphs	39
Figure 24 Correlation plot of Principal Components	39
Figure 25 Linear Regression model with Principal components, Model prediction comparison	40
Figure 26 Decision Tree Regressor model with Principal components, Model prediction comparison.....	40
Figure 27 Random Forest Regressor model with Principal components, Model prediction comparison.....	41

Figure 28 XGboost Regressor model with Principal components, Model prediction comparison	41
--	----

List of Tables

Table 1 Multiple Linear Regression base line model results	42
Table 2 Multiple Linear Regression model with Principal components	42
Table 3 Comparison of results	43

Abstract

Under Basel II regulation banks need to understand the financial risk related to changes in the economy which happens due to several shifts in the economy, especially during global crisis. The customer behaviour on spending and spending methods tends to change whenever they experience the huge effect on their income, health, and safety. The period between 2000-2021 has seen all such crisis due to the SARS in 2002, global financial crisis in 2007 and corona virus in 2019. Compared to other parts of the worlds, the Hong Kong citizens are ones who closely went through all the above crisis periods.

The research focussed on behaviour of citizens of Hong Kong to understand how they learned from the epidemic outbreaks and identifying the reasons for them to switch from traditional cash payment methods to card payment and contactless payments. Credit card spending is one such payment method which has seen growth over the recent years and still expected to grow among the Hong Kong citizens.

A bank's income from credit cards depends significantly on the interest rates, fees on the cards. The growth in the credit card spending means the total receivables from credit card holders also raises up. To identify the changes in total receivables, a bank needs to understand how customers will respond to crisis and changes in macro-economic variables.

The core of the research identifies the macroeconomic variables data and credit card lending survey data to model the effect of changes and predict the 'total credit card receivables' from the individual credit cards holders who took credit cards from authorized institutions. Multiple Linear Regression, Decision tree regressor, Random Forest regressor and XGboost regressor are the various machine learning algorithms are implemented on the independent variables to predict the dependent(target) variable 'total credit card receivables. Finally, the comparison of these models is done to identify the best model in this study and scope of future work is suggested for bank.

Introduction

Credit Cards

Credit cards enable cardholders to borrow and spend up to their credit limits without incurring interest until the due date ^[1]. The monthly earnings and credit history are frequently used to set the credit limit, having a credit history can be both beneficial and detrimental. Individuals will be creating a positive credit history if they pay their credit card payments and other bills on time. In contrast failing to the debts may result in developing a bad credit history, which may limit customers capacity to obtain additional loans soon. ^[1]

Credit card businesses and their operations

"Credit card firms" refers to two different types of businesses: issuers and networks. ^[2] Banks and credit unions are issuers of credit cards. Customers borrow money from the issuer when they transact using a credit card. Retail credit cards are commonly referred to as "co-branded" credit cards ^[2] since they are typically issued by a bank under contract with a store, gas station, or another retailer. Companies that accept credit card transactions are referred to as networks. ^[2] Visa, Mastercard, and American Express are the three major networks in the Hong Kong. American Express is both issuer and networks.

The” below image (Figure 1) from ‘moneygeek’ ^[3] explains the breakdown of how revenue is generated through credit cards. The card issuers gain from both cardholders and retailers in a variety of ways. The card holders pay one or more kinds of additional fee such as interest, annual fee, late fee, cash advance fee, balance transfer fee, foreign transaction fee^[3] whereas the merchants pay fees for transaction called interchange fee, processor fee and assessment fee which has the network maintenance costs under it ^[3].

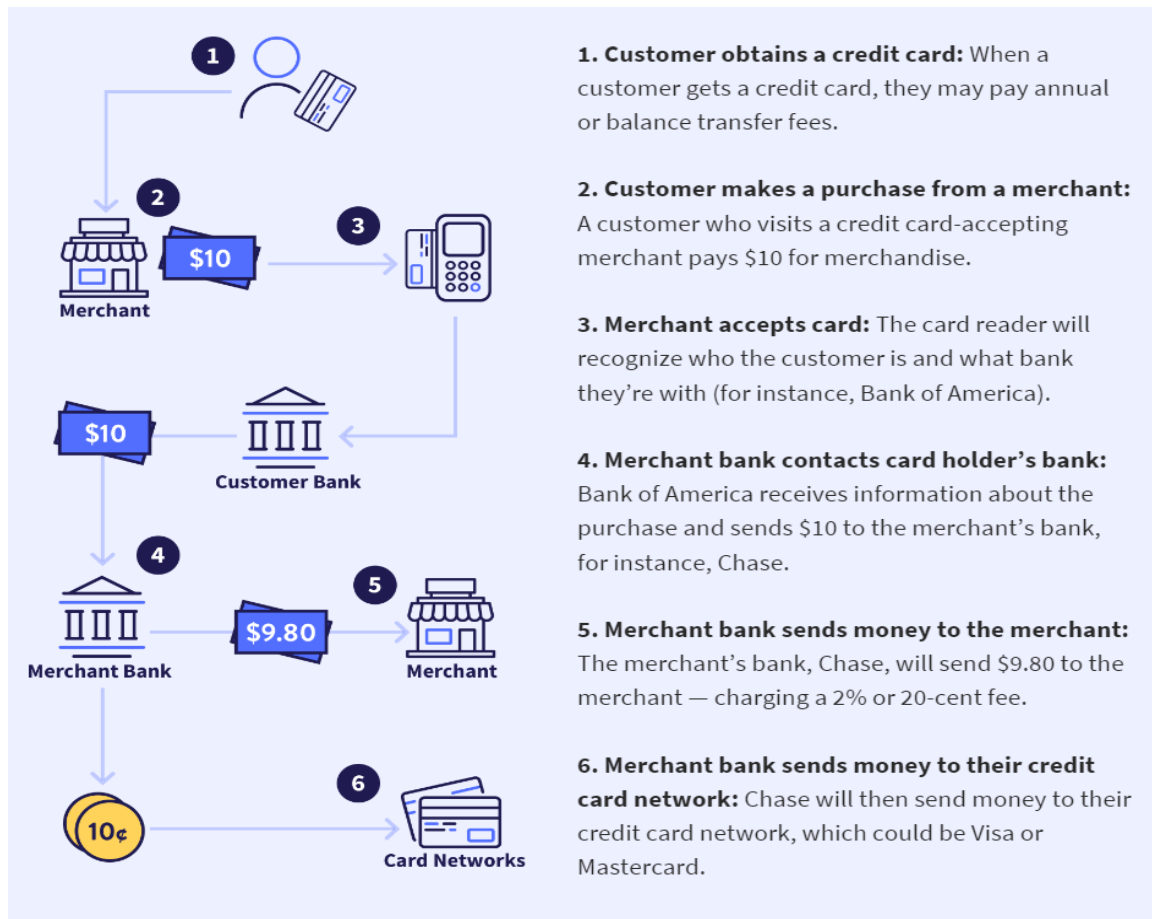


Figure 1 How revenue is generated through credit cards; Source: Moneygeek

Scope of credit card spending in Hong Kong

Hong Kong is a wealthy, well-banked region with good digital and physical infrastructure, which will enable future e-commerce growth. Because of Hong Kong's superior e-commerce infrastructure and strong e-commerce growth potential, the online shopping market is expected to rise ^[4]. The report from the trading economics and census and statistics department shows huge increase in the consumer spending over the recent years in the country as the spending went from 430000 HK\$ millions in February 2017 to up to 490000 HK\$ millions by the end of December 2018.



Figure 2 Consumer spending in Hong Kong; Source: Trading Economics and Census and Statistics department of Hong Kong

According to the JP Morgan 2019 payment trends (figure 3) in Hong Kong ^[4], cards are the most popular method of online shopping, accounting for little under half of all e-commerce payments, or \$1.8 billion in yearly sales. Among the card payment credit cards lead the payment market, with Visa®, Mastercard®, and American Express being well-known and widely used brands. Residents of Hong Kong are comfortable taking on debt, maybe because of their strong financial buffers: Hong Kong residents have one of the greatest net-worth-to-liabilities ratios ^[5] in the world.

Credit card ownership far outnumbers debit card ownership in Hong Kong. There are 2.67 credit cards for every 1,000 people, compared to only 0.81 debit cards. This is due to a culture of using credit cards to pay for household bills and everyday transportation, as well as a desire for credit card benefits like cashback and air miles programmes ^[6]. With 95.3 percent of the population having a bank account, there is a high level of financial inclusion. Domestic banks provide multi-currency cards, which allow payments to be settled in a variety of currencies, to profit on the local thirst for cross-border shopping with China ^[6].

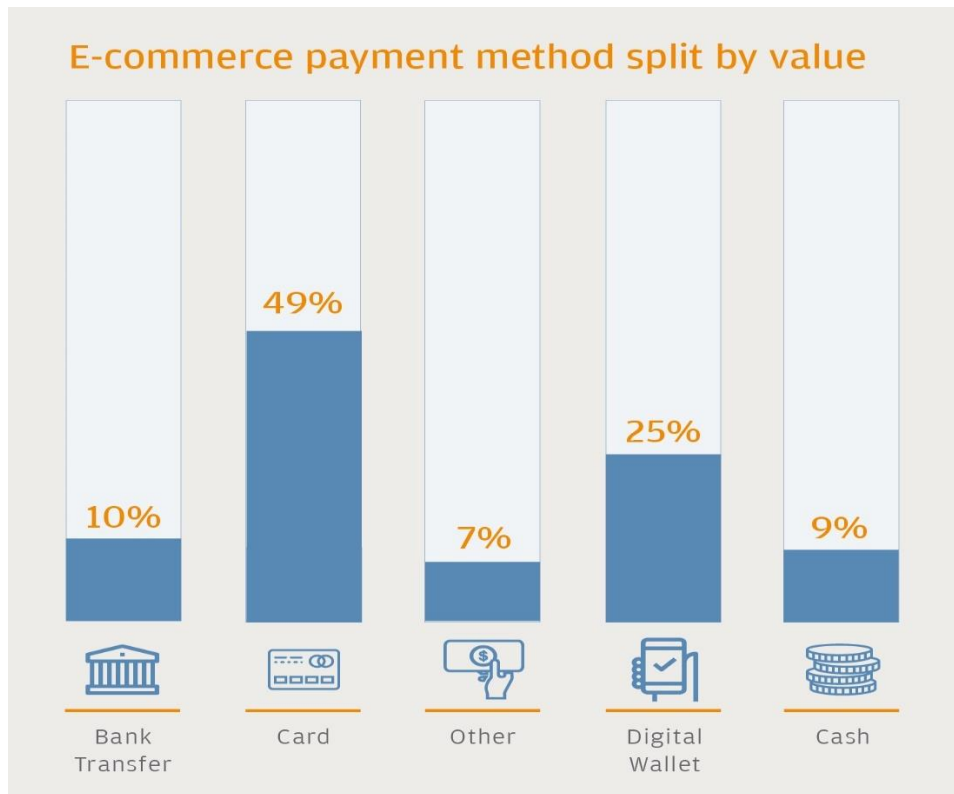


Figure 3 E-Commerce payment method split by value, Source: J.P. Morgan 2019 Payments Trends^[4]

Figure 4 illustrates the top e-commerce merchant segments in Hong Kong. Taking the percentage of total e-commerce value, customers spent highest value of purchases for travel, health and beauty, and consumer electronics^[4]. For improving the purchase value in other segments domestic merchants are also being enticed into the sector via online sales platforms that provide an entry point into the industry^[7].

Top e-commerce merchant segments



Figure 4 Top e-commerce merchant segments, Source: J.P. Morgan 2019 Payments Trends^[4]

The 'Visa consumer payment attitudes study 2.0'^[8] specifically examined the customer opinions and preferences in Hong Kong during the pandemic. The ground level study also identified that for the first time, credit and debit cards had surpassed cash as the most popular and preferred payment method ^[8] during the pandemic. This demand was driven by customers as two-fifths of Hong Kong customers utilised more digital banking services and would continue to do so afterward as they believed online purchases by credit cards, contactless payment in retailers as the safest way of making payments.

Due to all above discussed reasons and increase in demand there is huge growth potential identified in the credit card payments over the next few years. For merchants, this is both a problem and an opportunity to exceed consumer delivery expectations while also expanding their business easily. The growth potential ^[7] in this area is primarily identified in Hong Kong because

- a) Most of the people have never used e-commerce yet.
- b) Only around a quarter of Hong Kong residents purchase online.
- c) The country has the highest internet penetration rate among non-European countries, at 89.4%.
- d) 75% smartphone penetration, 95.3% Bank account penetration which is another encouraging fact.
- e) Announcement from Hong Kong chief executives in 2018 that they planned to invest 20% or more in their e-commerce and online sales activities.

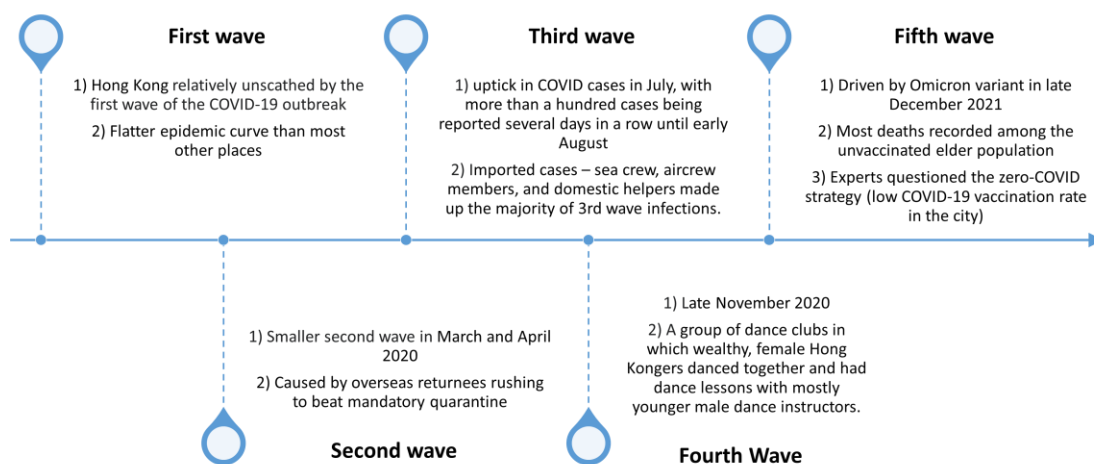
Fortunately, as the market for online purchasing in Hong Kong improves, the infrastructure to accommodate increased delivery numbers is already in place. This should allow retailers to satisfy consumer delivery expectations while expanding their business easily.

Hong Kong's response to Corona Virus outbreak

The line graphs plotted in the exploratory data analysis present in the later part of this report identified major changes in the variables during the SARS epidemic between 2002-2004. In contrast to the high downfall during that period, the corona virus didn't affect Hong Kong as much as the SARS did. Seeing different trends during similar kind of disease outbreaks raised a new question in during this research; Did Hong Kong learn anything after the SARS epidemic (2002-2004)? This research has found answer to this question as study the shows both Hong Kong government and citizens contributed their part to keep the severity of the coronavirus low during the initial days of pandemic.

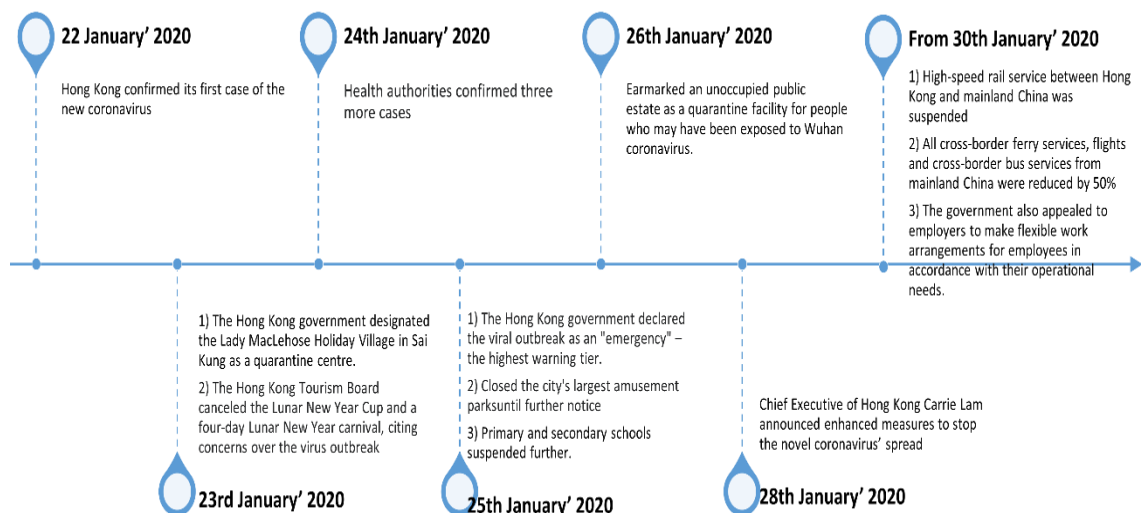
Public response

Being geographically close to China and the millions of mainland visitors who go there every year, and having status as an international transport hub, it is common to expect Hong Kong to have high vulnerability to corona the virus immediately during the initial outbreak but the behavioral changes in people such as using face masks in public during the SARS epidemic^[36]^[37] must be the reason for having confirmed corona virus infections low at 845, with only four deaths, by the April 2020. In addition to the knowledge of wearing marks, most people in the country already got habituated to the disease control measures like quarantine and isolation, social distancing, and border restrictions during SARS^[36]^[37]. Also, there are some citizens in Hong Kong who have grown distrusting the government, the World Health Organization, and the Chinese Communist Party^[38], for their skills and political motivations.



Government's response

During the beginning period of corona virus, when compared with the other countries, the Hong Kong government was very quick in responding after the first case of the new coronavirus confirmed in Hong Kong on 22nd 2020 [25]. Immediately on 23rd January a designated quarantine center is arranged [26] [27], and the tourism board cancelled the carnival by having concerns on the corona virus outbreak [28] [29]. After the health authorities confirming three more cases [30] on 24th January, on 25th January declared the situation as an "emergency" [31] by closing the famous amusement parks [32] and suspending the primary and secondary schools [33] [34]. The new measures were continued as on 26th January a public estate was designated as a quarantine center for the people who would have the direct exposure to Wuhan corona virus [35]. Upon the announcement of more enhanced measures by the chief executive on 28th January, later 30th January [35] between the mainland China and Hong Kong, the high-speed rail service was fully suspended, cross border ferry, flight and bus services were only 50% operated. Adding to the instantaneous actions taken at the beginning of outbreak, the "Zero-COVID" elimination strategy until February 2022 which led to close all the country's borders.



About variables used in the analysis

Gross Domestic Product

GDP is referred to as the "world's most powerful statistical indicator of national development and progress". It's a common indicator for international comparisons and a broad gauge of economic success

The Gross Domestic Product (GDP) of a country is estimated by calculating the total value of the goods and services, it has produced during a period, and it usually calculated quarterly^[11]. The two methods used to measure GDP usually are based on spending and by measuring income.

Private consumption expenditure

Consumer spending on goods and services is measured by private consumption expenditure, often known as personal consumption, consumer expenditure, or personal consumption expenditures^[12]. Food, housing, energy, clothing, health, recreation, education, communication, transportation, as well as hotel and restaurant services, are some examples of private consumption. It also includes durable goods like automobiles but excludes home purchases, which are considered as household investment.

In many nations, consumer expenditure accounts for about half to two-thirds of gross domestic product (GDP). In general, the greater the proportion of consumption, the poorer the nation, however there are exceptions in this case. China, for example, has a low ratio, while the United States has a high share^[12]. Since private consumption contributes for the majority of GDP, it is the primary source which drives economic growth.

Government consumption expenditure

The current expenditure by general government bodies on community services such as defence, education and other public order and safety is known as government consumption expenditure^[13]. The government is considered to be the consumer of its own output because these are offered either at free of cost or at prices that only pay a small percentage of expenditures. Because government output does not have a clearly observable market value, so it is valued at its cost of production in the national accounts. The value of any revenues from sales of government output like statistical publications is excluded from the value of government output to calculate government final consumption expenditure.

Gross domestic fixed capital formation

The gross domestic fixed capital formation is also referred as "investment" ^[14]. The acquisition of produced assets that includes purchases of second-hand assets, as well as the production of such assets by producers for their own use, minus disposals, is described as gross fixed capital formation (GFCF) ^[14].

Changes in inventories

Changes in inventories shows the difference between the quantity of ending inventory from the previous period and the amount of ending inventory for the current period ^[15] ^[16]. The closing inventories are fewer than the opening inventories if the changes in inventories of finished items and work in progress are negative. Because it is part of the cost of products sold, this negative sum is subtracted from revenue (from the income statement). When the changes in inventories of finished items and work in progress are positive, it means the closing stocks are more than the opening inventories.

Exports and Imports of goods(f.o.b.)

FOB ^[18] stands for Free on Board or Freight on Board. When a transaction's conditions of delivery are FOB, the seller is responsible for the cost of transporting goods on planes or ships whereas the buyer is responsible for the remainder of the costs associated with delivering the items to the buyer's location ^[18].

For example, there's a machinery vendor located in the London area of United Kingdom and the buyer is based in a city close to New York. Also, there is a product seller who has entered into an agreement with the buyer to sell the goods FOB London for USD 5300. The FOB London selling cost of items is USD 5300. As a result, the seller covers all costs associated with transporting the products to London port, as well as all costs associated with customs clearance in London, in order to get the commodities on board to airlines for shipping. As previously stated, the buyer is responsible for all additional costs associated with transporting the products to the buyer's location. The customer chooses the shipping firm or airline, and the seller ships the products according to the buyer's instructions. The buyer is responsible for paying the shipping firm or airlines for the cost of freight and he must make arrangements for the products to be insured and pays for the insurance.

Exports and Imports of services

Exports are a major topic these days, but most people refer to product exports when they talk about exports ^[19] ^[20]. More and more businesses are profiting from service exports, and various factors indicate that these exports will continue to increase fast. Exports of services are a growing trend in global trade. Many traditional manufactured product exports now include technology that necessitates setup, troubleshooting, maintenance, and repairs ^[19] ^[20].

A service exporter is someone who exports services that aren't physically visible, such as intangible items ^[19] ^[20]. A service export is any service given by a person in one country to individuals or businesses in another country. Exports of services are a growing trend in global trade. Tourism, software, health care, consulting, hotels, and other service industries are examples of industries where we cannot see the product physically, but they help in earning foreign exchange and the country will gain foreign exchange via selling services.

Unemployment rate (%)

People of working age who do not have jobs, are available for employment, and have made efforts to obtain work are considered to be unemployed ^[21]. Estimates of unemployment rates that are derived from the consistent application of this definition produce findings that are more internationally comparable than estimates that are derived from estimates based on national definitions of unemployment ^[21]. This indicator is computed by dividing the total number of jobless persons by the total number of people actively looking for work, and it is adjusted seasonally. The term "labour force"^[21] refers to both the total number of persons who are jobless as well as those who are already working.

Interest rate (%)

The interest refers to the base rate^[22] which is the annual percentage rate (APR) of interest that a central bank would charge commercial banks for borrowing. There are a few other names for the base rate, including the bank rate and the basic interest rate.

Although commercial banks are allowed to determine their own rates of interest for borrowing money, the rates that they use to determine the interest they charge on loans and the interest they pay savers are typically taken from the base rate^[22]. This indicates that central banks have the ability to utilise base rates to either stimulate or discourage consumer spending, depending on how the economy is performing at the time.

If the base rate is lowered by a central bank, it is likely that individual banks would likewise drop the interest rates they charge for loans and mortgages^[22]. This indicates that obtaining a loan may become less difficult, and that mortgage rates may become more favourable to purchasers as a result. However, this may also imply receiving fewer returns on the investments because the value of interest rate payments would decrease as a result of the lower base rates.

When the base rate of a central bank is raised, the cost of borrowing money and the interest rate on mortgages both rise, which is good news for lending institutions and for businesses that are in the business of selling goods and services^[22]. Nevertheless, if the funds were placed in interest-bearing accounts, the account holders may anticipate higher returns on the interest payments they receive as a result of the increase in the base rate^[22].

Credit card – Total number of accounts

Credit card – Total number of accounts refers to the total number of credit card account holders who obtained their credit cards from one of the authorised institutions

Credit card – Average total receivables

"Credit card receivables"^[17] in the credit card lending survey refer to amounts owed, or credit card receivables owed by individual credit card account holders. The average total receivable during the period is calculated^[17] by the sum of opening stock and closing stock divided by 2.

Credit card – Charge off amount

The charge-off amount ^[17] represents the total amount of credit card receivables written off the loan book during a given period (including principal, interest, and fees incurred of the charge-off accounts). This amount is set no matter when a charge is put on the profit and loss account, which could be before the amount is written off if the institution's policy is to make provisions before the amount is written off ^[17]. The policies regarding charge-offs can differ from one institution to another. When a receivable has been past due for more than 180 days or when it seems doubtful that the receivable will ever be repaid in full, a company will normally "write off" the account in question and remove it from their books (For example the account holder is bankrupt or cannot be traced) ^[17]. Both the charge-off amount and the delinquency amount together to provide a more complete picture as the first method takes into consideration all the credit card receivables that, as of the reporting date, had been past due for more than ninety days but had not yet been discharged as bad debt ^[17]. As a result, it serves as a leading signal of upcoming charge-offs. The latter method takes into account receivables that were written off within the stipulated time because they were substantially overdue (for example, for more than 180 days), as well as receivables that were written off earlier than 180 days because they were regarded irrecoverable ^[17].

Credit card – Delinquent amount

The entire amount of credit card receivables that have been past due for more than 90 days and were still unpaid as of the final day of the reporting month is used to calculate the delinquent amount ^[17]. This amount is expressed as a percentage of the total amount of credit card receivables. When a payment is more than one day late as of the last day of the reporting month, the associated credit card receivables are categorised as overdue ^[17]. The percentage of accounts that are delinquent is a useful early indicator of the overall quality of a credit card portfolio.

Credit card – Rollover amount

Rollover amount is the amount within total receivables in respect to which the cardholder has not fully repaid the statement balance but has at least made the minimum due required by the Authorized Institution ^[17]. This amount is referred to as "borrowing," and it is the amount that is included in rollover amount. Amounts that are past due are not included in this total. When the required minimum payment for a customer's account is not paid by the due date, the account is said to be past due ^[17].

Why is it important to know the credit card receivables?

Even though bank's earn income from credit card holders in several ways discussed above (fees, interest, late payment charges etc), they only receive this income when they receive the due amount from customers. The encouraging sign of growth in the credit card spending behaviour means the 'total receivables' from card holders will also go up after they make purchases using credit cards. Many credit card companies suffer every year when their customers do not make their payments on time or sometimes fully miss by going bankrupt. This means when the customers with due balance regularly miss payments or never make payments, the banks lose their money and income. The figures 13, 14, 15 and 16 in the exploratory data analysis part of this report clearly shows the increase in the total receivables at the end of 2021 and moreover increase in the rollover amount is also identified in the figure as some customers are only paying the minimum dues since several quarters. On top of all these the charge off amount reached to its highest during SARS and next highest during the global financial crisis and during the covid outbreak which led to huge losses for the banks. Hence it has become very much essential for banks to identify solutions by foreseeing the credit card receivables by learning from the past experiences from customers.

Data Collection

Data Collection is a primary step and important aspect of the research because any research in Machine Learning is strongly reliant on data. It's the most important factor that enables algorithm training and explains how machine learning has gained so much traction in recent years. The data obtained from publicly available sources must be information-rich and reliable to perform analysis.

About the data source

Hong Kong Monetary Authority (HKMA) in Hong Kong is in charge of ensuring monetary and banking stability. This research has relied on the official statistical data which was published in websites ^[9] ^[10] by the HKMA authorities as primary data sources. Several data files are collected from the areas of real sector and financial sector and identified the required macroeconomic variables for the research. The data sets of Gross domestic product, private consumption expenditure, government consumption expenditure, gross domestic fixed capital formation, changes in inventories, exports of goods(f.o.b.), exports of services, imports of goods(f.o.b.), imports of services, unemployment rate (%) and interest rate are eleven macroeconomic variables identified.

In addition to above, the credit card lending survey results provided the useful credit card data for the analysis. This survey focused on authorised institutions (AIs) and their credit card-related subsidiaries. It excludes credit card issuers unrelated to authorised institutions. The data from this survey contains customer data on credit card spending such credit card total number of accounts, delinquent amount, charge off amount, rollover amount and average total receivables.

Data Pre-processing

Data pre-processing is also an important step because the data mining algorithms cannot be directly applied to the real-world data sets which contain raw data of poor quality with missing values, inconsistent information, and unnecessary data points called outliers. In a word, data preparation is a series of methods that aid in making the dataset more machine learning friendly. This implies describing what is to be forecasted, and then acquire the evidence that would best assist in making predictions. Hence the issues with the data are handled in the pre-processing step to transform the data into a meaningful format.

Most projects that use predictive modelling use "structured data" or "tabular data". The raw data which is collected in the previous step is available in different formats and available for different duration of periods. Hence to make it useful for the research, the extra information is removed from the data set, the structuring of the rows and columns is carried out. Finally, the different data files are merged into a single data file to have the quarterly data from 2000 – 2021 which is of past 22-year period.

Exploratory Data Analysis

Exploratory data analysis (EDA) is used to study and investigate data sets and describe their primary properties, which commonly includes the use of data visualization techniques. It assists data scientists in discovering how to best manipulate data sources and obtain the answers they require, making it easier to find patterns, test hypotheses, and double-check assumptions using visualizations like histograms, boxplots, scatter plots, correlation plot and line charts etc. Univariate non-graphical, univariate graphical, multivariate nongraphical and multivariate graphical are the majorly used EDA techniques.

Countplot

The countplot in figure 5 illustrates the number of times an observation appears in the categorical variable (counts). The distribution of the data is seen after the pre-processing the data.

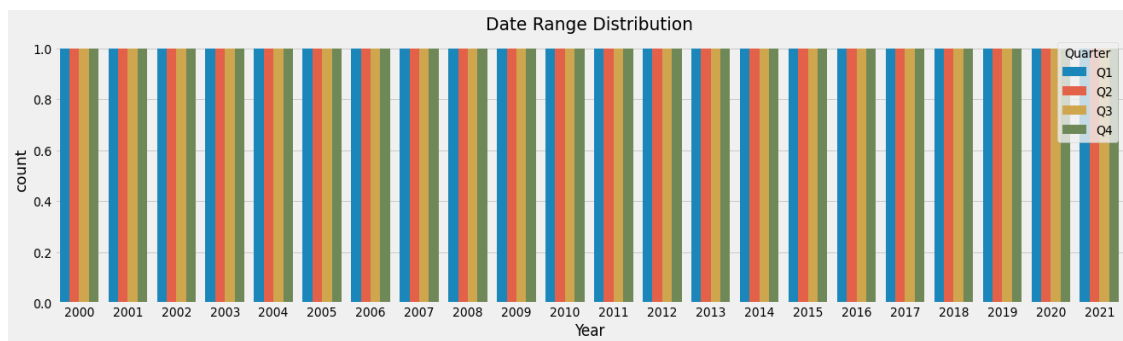


Figure 5 Data Range Distribution (2000 – 2021)

Line graphs showing the last 22 years trend of the variables

This graph in figure 6 illustrates how the GDP values increased every year starting from 2000 till 2021 in Hong Kong. When had a close glance, the big downfall has taken place in the period of 2020-2021 due to covid situations and fast increase during the period of 2004-2008.

From 2000 to 2021, there is a total of about HK\$450,000 million increase of their GDP that clearly shows that this country's economy is intensively increasing over the last two decades.

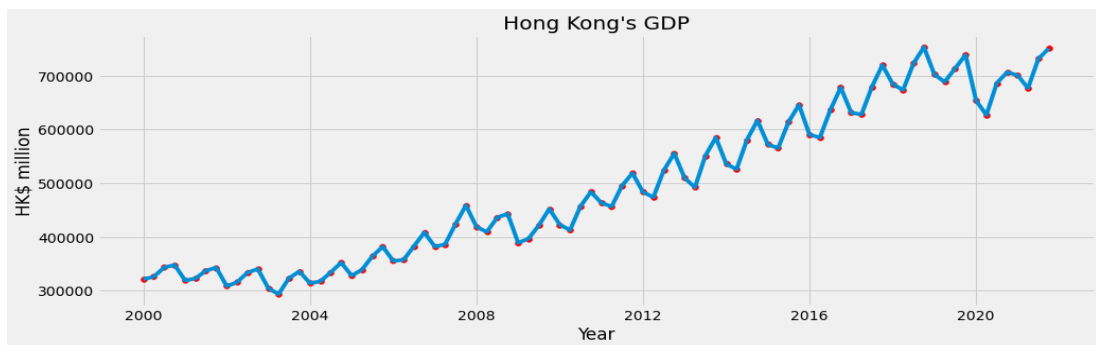


Figure 6 Gross Domestic Product measures (2000 – 2021)

Figure 7 shows the Hong Kong's Total consumption expenditure by comparing both private and government consumption expenditure. As we can see the private consumption always stayed highest than the government consumption expenditure. The red line indicating Govt consumption just got doubled from 2000 to 2021 whereas the private consumption in blue line shows that it got 40% increase within the same period. The expenditure for government consumption mostly remained stable till 2010 and slow heap till 2021 and for private consumption too maintained stable till 2007 and from there it started to strike high and got observable downfall during period of 2019-2020.

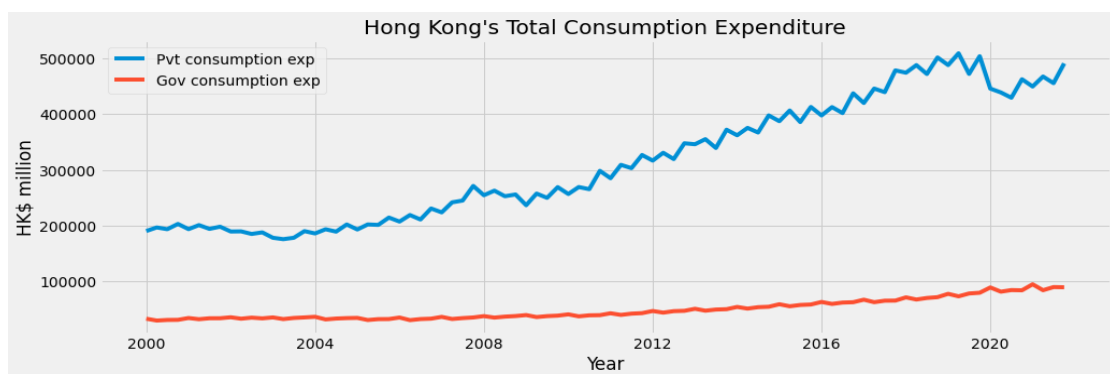


Figure 7 Private and Government Consumption Expenditure (2000 – 2021)

Looking from the year 2000 in figure 8, the Hong Kong's gross domestic fixed capital formation has more downfall even though it tries to upstream for certain time and reached closely to HK\$ 60000 million in 2003. After 2003, it started to upstream again and in 2019 it started fall back by 40% in 2020. At the end of 2020, it started to increase again and maintaining its upstream safely.

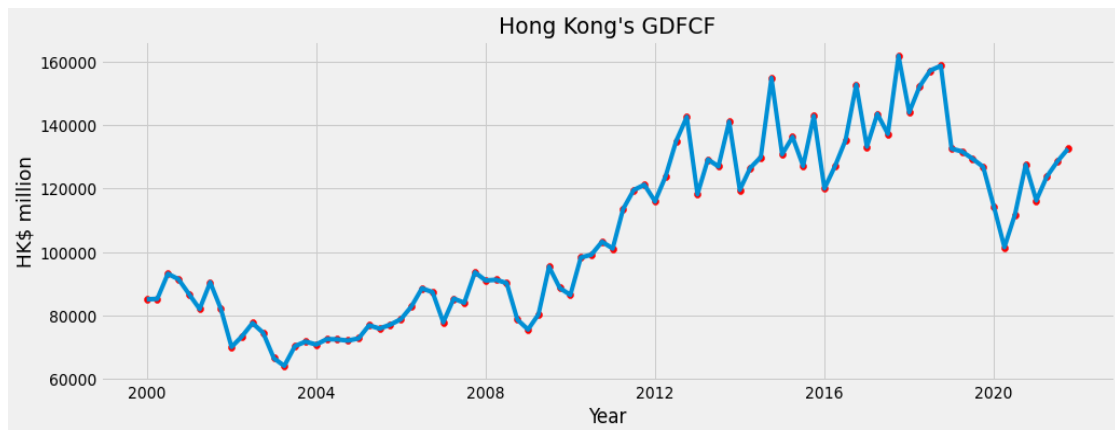


Figure 8 Gross Domestic Fixed Capital Formation (2000 – 2021)

This graph in figure 9 shows how the changes effected in inventories over the last two decades. From the plot, the highest peak reached in 2010 and 2020 and lowest in 2021. From 2020, the inventory changes raised, and it reached peak points 3 times within a year.

During the period 2011-2018, the changes were mostly switched over positive and negative to reach closely to zero.

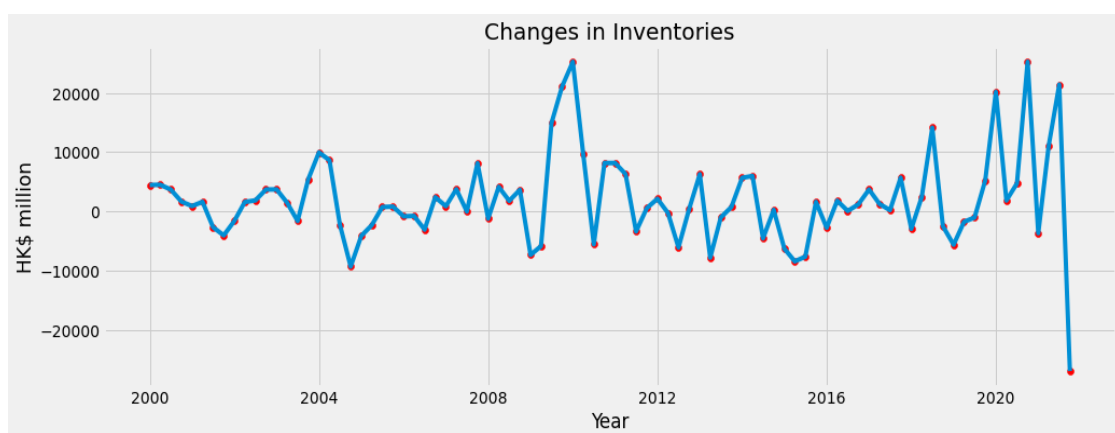


Figure 9 Changes in Inventories (2000 – 2021)

Figure 10 shows the comparison of exports and imports of goods and services. Beginning in 2000, Hong Kong's Exports of goods remained higher than the Imports of goods while service exports remained lower than the service imports. In the period between 2007 -2010, both service exports and imports almost levelled up to the same amount. Post 2010, Hong Kong sees growth in its service exports while the good exports went down compared to good imports. Both exports and imports of goods had drastically fallen during 2009-10 and 2020 which is due to global financial crisis and covid outbreak respectively.

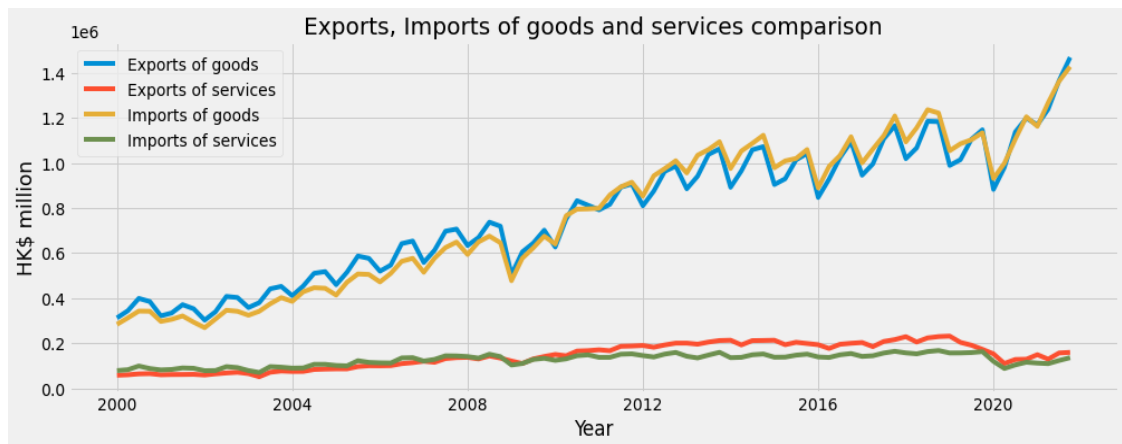


Figure 10 Comparison of Exports and Imports of Goods and Services (2000 – 2021)

Starting from 2000, the rate started at 5.2% and reached the peak value of 8.5% in 2003.

From the graph, the lowest rate is achieved in 2018,2019 but unfortunately it increased the next two years and reached 3.9% by the end of 2021

During 2012 and 2017, the graph is stable and remained between 3-4 % which can be considered as best lengthy period that has lowest unemployment rate.



Figure 11 Unemployment rate (2000 – 2021)

The graph in figure 12 shows that the interest rate was below 1 % during the period 2009-2015 and has highest rate of 8% in 2001. For several years after the global financial crisis of 2008, for instance, most central banks maintained relatively low base interest rates. Because of this, most commercial banks decided to charge their clients low interest rates on loans but also provide low interest rates on money stored in interest-based accounts^[22].

Although the graph is simple with few deviations, it can be concluded that from 2000-2008, the rate lies between 4 and 8% and from 2009 to 2021, the interest rate decreased to be between 3.5 and 0.75.

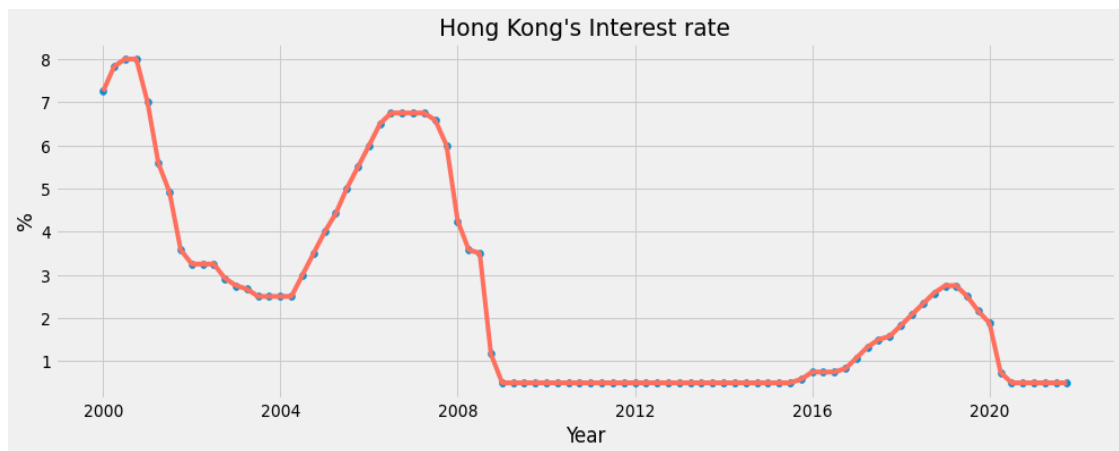


Figure 12 Interest Rate (2000 – 2021)

Figure 13 illustrates the credit card ownership accounts in Hong Kong which has drastically increased over the last 20 years. Overall, the lowest number of accounts is struck in 2000 and highest number has hit in 2020. During 2008-2012, the number of accounts has 3.8million increase that shows highest growth point over two decades.

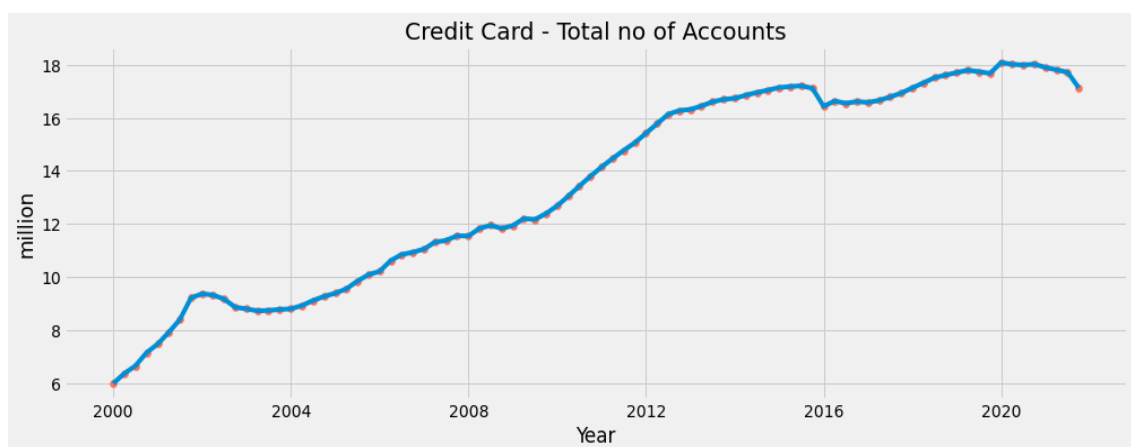


Figure 13 Total of Credit Card accounts (2000 – 2021)

The graph in figure 14 shows how much credit card charge off amount is charged over year by year. Highest value of HK\$2250 million is reached between 2002-2003 and has steep downfall to HK\$500million in 2021 and lowest is stroked in 2011.

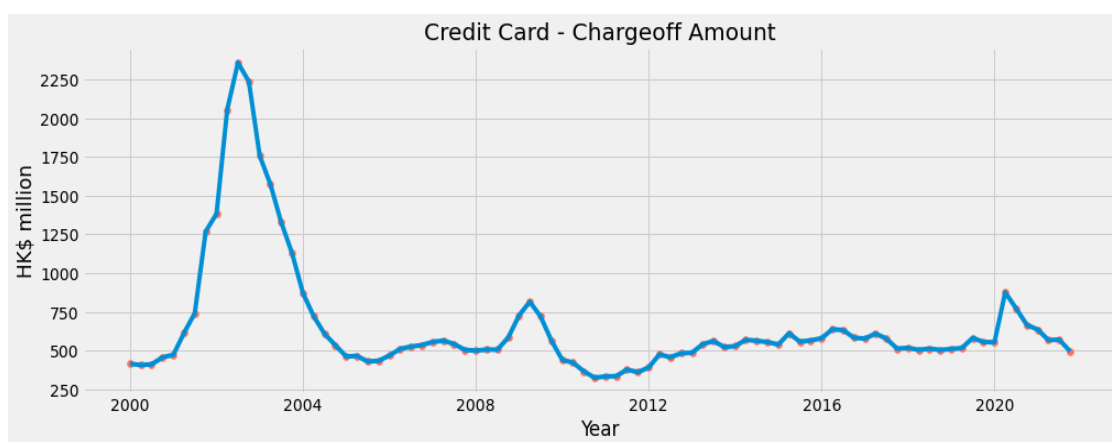


Figure 14 Credit Card Charge off Amount (2000 – 2021)

As seen in figure 15, starting from 2000, Credit card rollover amounts started to spike and has hit a maximum of HK\$34000 million in 2002 and then slowly it started to downfall and stabilised between HK\$24000 million and HK\$26000 million for about 3 years.

Following, it again started to decrease further and reached its lowest amount of \$18000 and from there again it started to increase back close to HK\$26000 in 2020 and it fallen back to hit HK\$22000 in 2021.

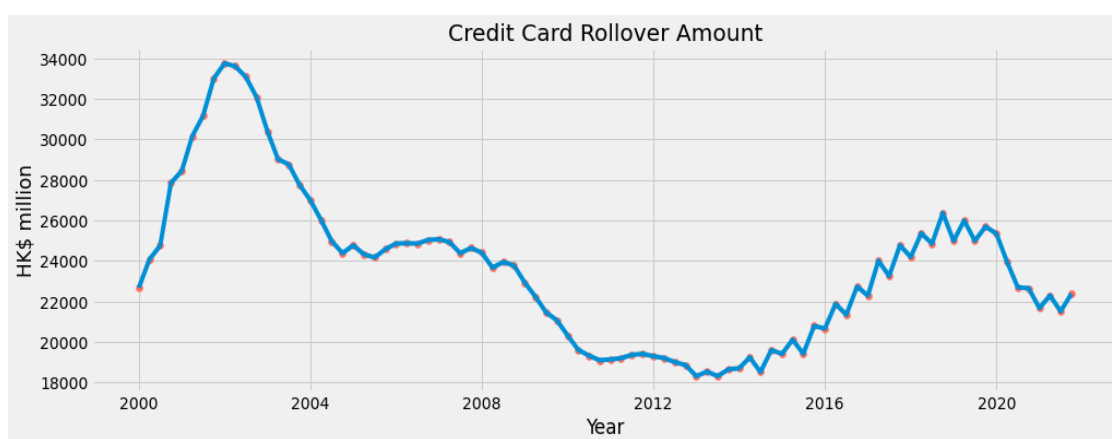


Figure 15 Credit Card Rollover Amount (2000 – 2021)

Figure 16 clearly shows above 8% the increase in the total receivables at the end of last quarter of 2021 which nearly went upto HK\$135 billion. It also shows a trend of total

receivables go up to its highest during SARS and next highest during the global financial crisis and during the covid outbreak

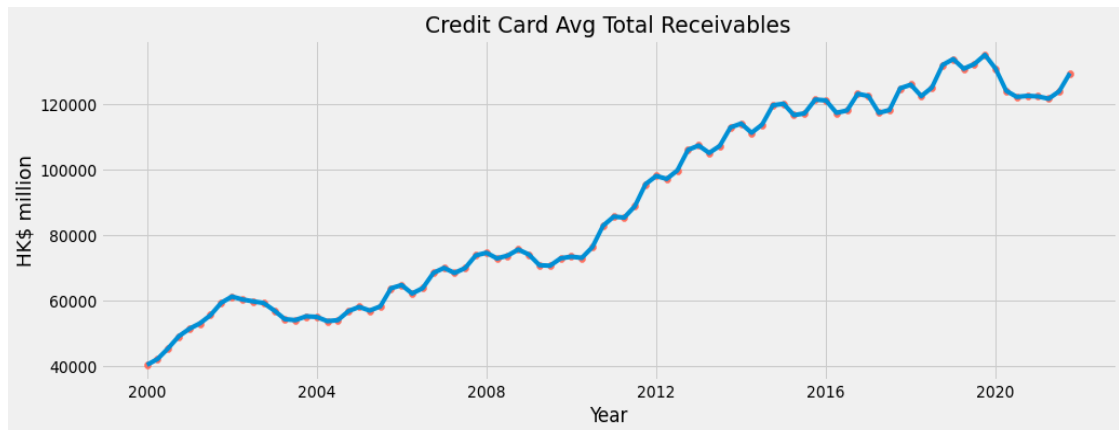


Figure 16 Average Total Receivables of Credit Card (2000 – 2021)

Correlation Plot

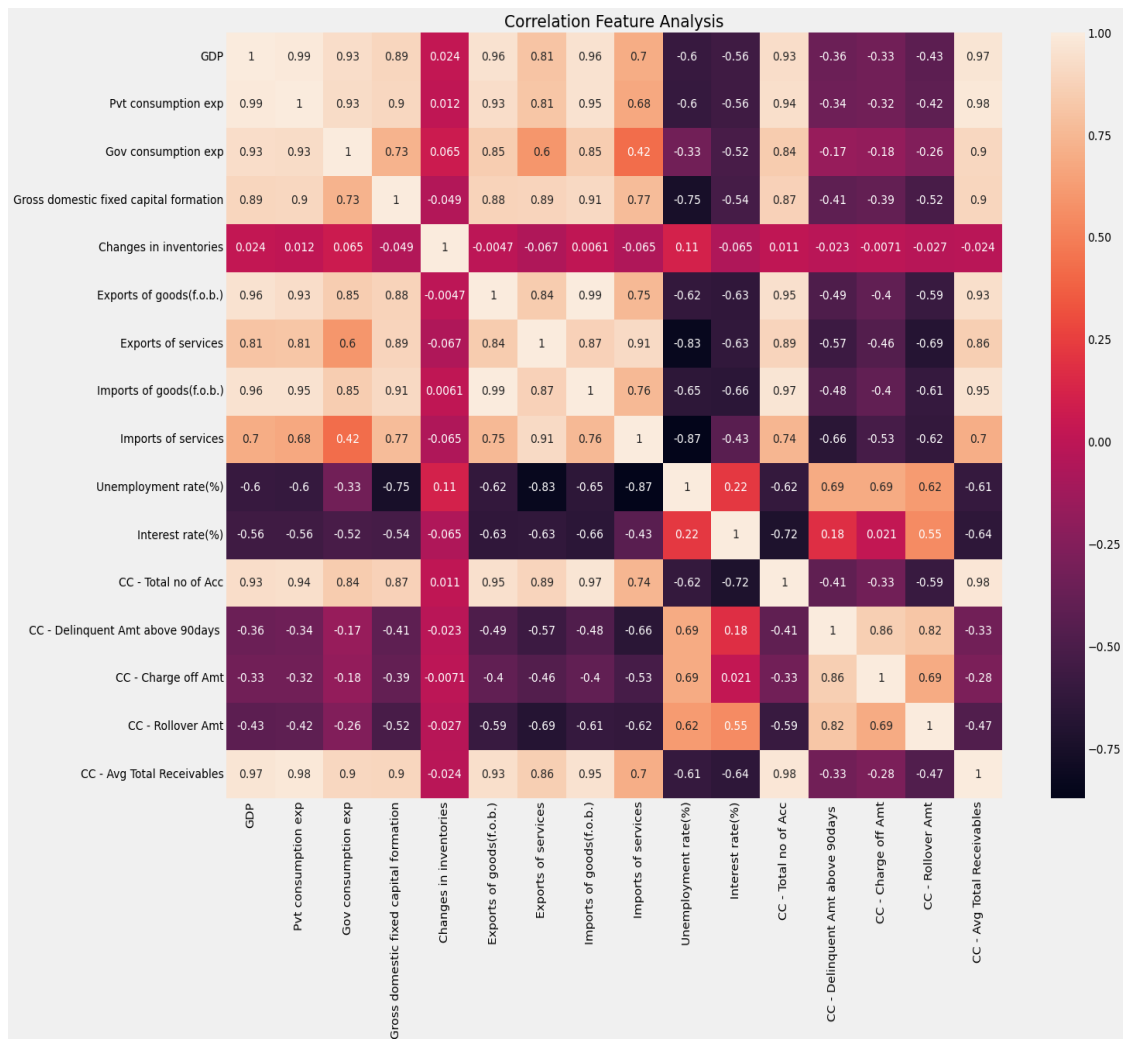


Figure 17 Correlation Plot: Multivariate Statistical Analysis

A correlation matrix is used in EDA step as a part of multivariate statistical analysis. It is a typing process for comparing correlation coefficients across different features in a dataset. It enables researchers to see how much connection there is between various variables. The correlation matrix helps researchers to minimize the number of features in a dataset by allowing us to find variables with high degrees of connection. This is known as dimensionality reduction, and it may be utilised to speed up and increase the model performance.

As seen in the heat map of figure 17 the dataset contains the same number of rows that the correlation matrix has, and it also has the same number of columns. As the dataset contains 16 variables, the correlation matrix will also include 16 rows as well as 16 columns. Each row and column represent a different variable (or column) in the dataset. The value in the matrix reflects the coefficient of correlation between the row and column that correspond to the row's and column's respective variables in the dataset. The values in the matrix along the diagonal line always remains same and will be equal to '1'. This is because the diagonal elements show the correlation between the variable and itself. The numbers in the upper right corner of the matrix are a mirror copy of the ones in the lower left corner. This is due to the fact that the connection between the two variables that comprise each row-column pair will, at all times, remain unchanged. When we observe the correlation between the different variables, we can see that there is a good correlation between the variables in the whole data. The variables Gross domestic product, private consumption expenditure, gross domestic fixed capital formation, exports of goods, exports of services, imports of goods, imports of services and credit card – total number of accounts have correlation coefficients above 0.7 which indicates that these are highly correlated with the target variable 'Credit card – Average total receivables' The next step is to perform feature selection and engineering for selecting inputs for the modelling.

Feature Selection

Feature selection is the process of identifying variables that are beneficial in the process of predicting the response in the context of machine learning. When developing predictive models, it is generally accepted as a good practise to first determine which properties are the most significant. For discussing more in detail, it is not uncommon for datasets taken from the real world to have columns that consist of nothing but noise. It is best to eliminate variables of this kind due to the amount of memory space they take, the amount of time and the computing resources it is going to cost, and this is especially true for large datasets. Sometimes, having a variable makes perfect sense from a financial standpoint, but we will be unsure whether it truly assists in predicting the target variable. It is also necessary to take into consideration the possibility that a characteristic that may be valuable in one machine learning method (for example, a decision tree) can be underrepresented or underutilised by another machine learning algorithm (like a regression model). Having said that, even though the variable in question does not appear to be particularly helpful in explaining the target variable, it is still possible that, in the presence of other predictors or in combination with other predictors, the variable in question will prove to be significantly useful. This mean variable might have a low correlation value with target variable but when combined with other factors, it may be used to assist explain some patterns and phenomena that can't be explained by the other variables alone. In situations like these, it might be difficult to decide whether to include such variables in the analysis. Hence it is always advisable to have variables that have solid business rationale behind the inclusion of a variable and to rely only on the variables themselves. This is because having variables that have sound business logic backing the inclusion of a variable will help you make better decisions

Understanding the importance of feature selection and for determining if a certain variable is significant or not, as well as the extent to which it is contributing to the model; two feature selection methods namely 'feature selection based on feature importance' and 'Feature selection based on the p-value' are used in this research for identifying the useful variables that will help in predicting the target variable.

Feature selection based on feature importance

The relevance of each data feature is rated using the feature importance score. The higher the score, the more significant or pertinent the characteristic is in relation to the variable being evaluated. There is a built-in class that is included with Tree-Based Classifiers called Feature Importance. The selection of features in this approach is done with the assistance of Extratree Classifier (from `sklearn.ensemble import ExtraTreesClassifier`). `ExtraTreesClassifier` is an ensemble learning approach that uses decision trees as its foundation.

The Extra Trees Classifier ^[24] is a form of ensemble learning approach that, in order to output its classification result, combines the outcomes of numerous de-correlated decision trees that have been gathered together in a "forest." It is only distinct from a Random Forest Classifier in the method in which decision trees within the forest are constructed, but conceptually, it is quite comparable to a Random Forest Classifier ^[24]. The data from the initial training sample is used to create each Decision Tree contained inside the Extra Trees Forest. Then, at each test node, each tree is given a random sample of features from the feature set, and from those features, each decision tree must choose the feature that would partition the data the most effectively based on some mathematical criterion (typically the Gini Index) ^[24]. This random sampling of characteristics results in the production of numerous decision trees that are independent of one another. During the construction of the forest, the normalised total reduction in the mathematical criteria is used in the decision of feature of split (Gini Index, if the Gini Index is used during the construction of the forest) is calculated for each feature. This value is known as the feature's Gini Importance ^[24]. To execute feature selection, each feature is ranked in descending order based on its Gini Importance, and the top features are picked based on the preferences ^[24].

By using Extra Tree Classifier in this research, the top 12 features of the dataset are extracted using the feature importance metrics. The figure 18 reveals these 12 features which are ‘Credit card – Charge off Amt’, ‘Changes in inventories’, ‘Imports of goods(f.o.b.)’, ‘Exports of services’, ‘Government consumption expenditure’, ‘Imports of services’, Gross domestic Product’, ‘Private consumption expenditure’, Gross domestic fixed capital formation’, ‘Credit Card Delinquent Amount’, ‘Exports of goods(f.o.b.)’, ‘Credit Card Rollover Amount’ and their importance scores respectively.

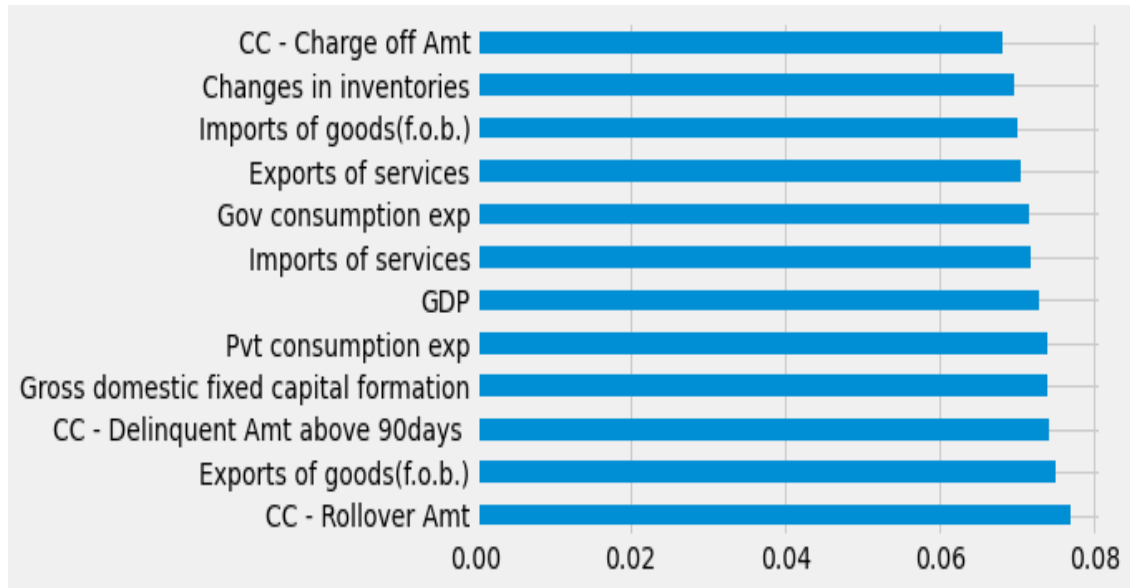


Figure 18 Graph of Feature Importance and values

```
[0.07275862 0.07390805 0.07149425 0.07390805 0.06965517 0.07505747
0.07034483 0.07011494 0.07183908 0.04448276 0.03942529 0.04793103
0.07402299 0.06816092 0.07689655]
```

Feature selection based on the p-value

The target variable column is removed in this step, then the top 12 important variables that selected from the previous step are now examined using the p-values which are generated by developing the Ordinary Least Squares (OLS) model whose results can be seen figure 19. In this step, the feature selection for predicting the target variable is determined by identifying how the features affect the p-value. These calculated p-values can be used to determine whether to preserve a feature or not. If the p values are higher than 0.05, then the variables are not statistically significant to keep in the model. As the seen in the OLS results of figure 19, the 'Credit card – Charge off Amt', 'Imports of services' and 'Private consumption expenditure' have the p-values '0.215', '0.146' and '0.059' respectively and variables are not necessary to keep in the later part of modelling. Hence as seen in the head of the data in figure 20, these three variables are manually dropped before proceeding to the later step of modelling.

	coef	std err	t	P> t	[0.025	0.975]
const	1.167e+04	6132.989	1.902	0.061	-548.133	2.39e+04
CC - Rollover Amt	-0.7628	0.312	-2.446	0.017	-1.384	-0.142
Exports of goods(f.o.b.)	-0.0675	0.019	-3.470	0.001	-0.106	-0.029
CC - Delinquent Amt above 90days	21.2900	6.228	3.419	0.001	8.886	33.694
Gross domestic fixed capital formation	-0.1196	0.047	-2.562	0.012	-0.213	-0.027
Pvt consumption exp	0.0720	0.038	1.913	0.059	-0.003	0.147
GDP	0.0508	0.022	2.334	0.022	0.007	0.094
Imports of services	0.1135	0.077	1.470	0.146	-0.040	0.267
Gov consumption exp	0.4122	0.131	3.144	0.002	0.151	0.673
Exports of services	0.1023	0.040	2.585	0.012	0.023	0.181
Imports of goods(f.o.b.)	0.0716	0.020	3.620	0.001	0.032	0.111
Changes in inventories	-0.1634	0.049	-3.312	0.001	-0.262	-0.065
CC - Charge off Amt	2.7705	2.214	1.251	0.215	-1.640	7.181

Figure 19 P-values from OLS model

In [41]:

Dropping variable manually where p value is >.05
stand_df = stand_df.drop(['CC - Charge off Amt','Imports of services','Pvt consumption exp'],axis=1)
stand_df.head()

Out[41]:

	CC - Avg Total Receivables	CC - Rollover Amt	Exports of goods(f.o.b.)	CC - Delinquent Amt above 90days	Gross domestic fixed capital formation	GDP	Gov consumption exp	Exports of services	Imports of goods(f.o.b.)	Changes in inventories
0	40300.9615	22675	312039	331	84973	321039	33220	57703	283493	4481
1	42101.9885	24078	344122	319	85143	325717	29911	59704	312303	4521
2	45369.2640	24753	398739	333	93068	342978	30986	63795	342454	3710
3	49065.3320	27846	384735	388	91332	347767	31177	64835	341752	1687
4	51360.6910	28444	321174	480	86566	318843	34450	59710	296590	928

Figure 20 Head of the data after dropping values based on p-values

Feature Engineering

Feature engineering is an important phase in the predictive modelling using machine learning [23]. It entails the alteration of a given feature space, usually with the use of mathematical functions, with the aim of reducing the modelling error for a specified aim [23]. This process includes transforming raw tabular data into a format that the machine learning model can understand and involves enhancing some columns to provide machine learning models with additional information in the anticipation of obtaining more accurate results.

Standardization using standard scalar

Figure 21 displays the head of the standardized data performed using the standard scalar function in python.

The approach of standardizing the data is performed as a part of feature engineering process in our research. Standardization of a dataset is a fundamental need for many machine learning algorithms because they may perform poorly if the individual features do not resemble standard normally distributed data. Hence by using standard scalar, the distribution is shifted such that it will have mean zero and scaling to unit variance. For this, every input variable is scaled individually by centering the mean and dividing by the standard deviation.

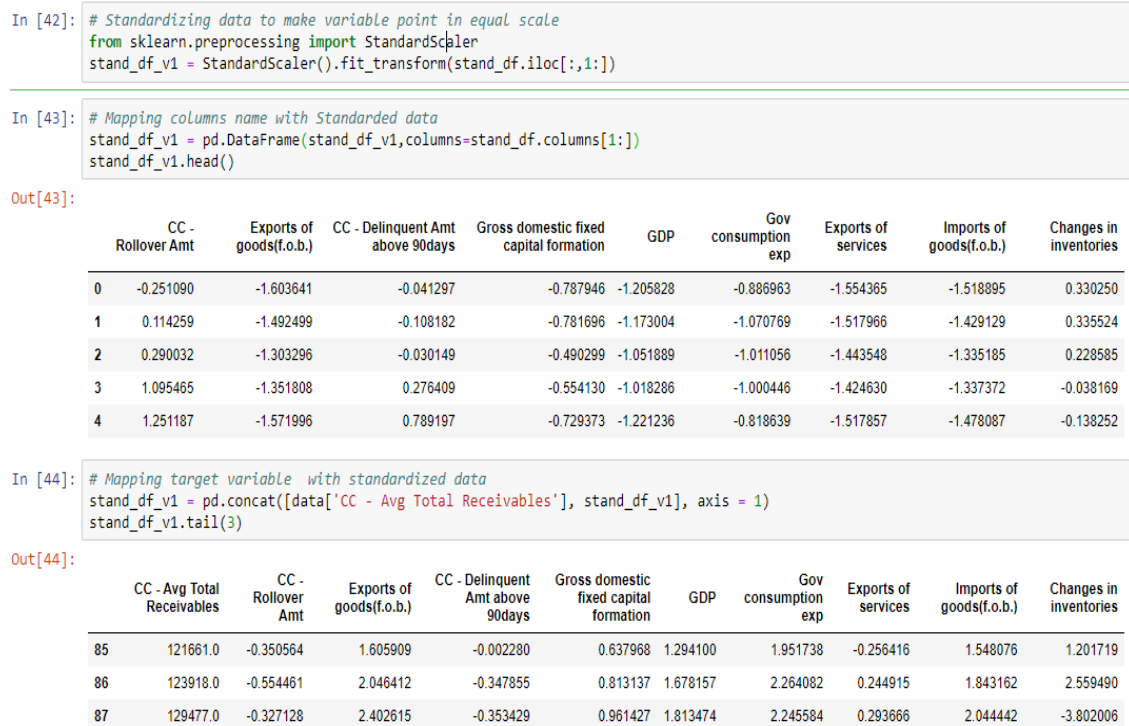


Figure 21 Head of the standardized data

Modelling and Optimization

In this process, the prepared data set which is taken from output after performing feature selection and feature engineering is divided into data for training and data for testing. The majority of the data in the dataset is set aside for use as training data, and in addition to this, a subset of the dataset that will be used for testing is also prepared.

For the machine learning models used in this research, the data is split into 80:20 ratio, which means 80% of the total data is used to create the train data and 20% of the data is used to create the test data.

Now, the machine learning models are trained and constructed based on the training data, and then evaluated based on the testing data. The testing data can be considered new and undiscovered data, which enables the correctness of the model to be evaluated together with its degree of generalisation.

Multiple Linear Regression Analysis

A multiple linear regression model roughly calculates the relationship between dependent variable and two or more independent variables which is an explanatory variables using a straight line ^[39]. In other words, it is usually used to predict the value of one variable based on the value of two or more variables. The variable whose value to being predicted is called as dependent variable and the variable's that we are using to predict the value of dependent variable is called independent variable's ^[39]. This method is used to see the strength of relationship between the dependent variable and independent variables ^[39].

Multiple linear regression formula ^[39]:

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$$

Where,

y = the predicted value of the dependent variable

β_0 = the y-intercept (It will be value of y when all other parameters are set to 0)

$\beta_1 X_1$ = the regression coefficient β_1 of the first independent variable (X_1)

$\beta_n X_n$ = the regression coefficient of the last independent variable

ε = model error (how much variation there is in our estimate of y)

The multiple linear regression calculates three things to best fit line for each independent variable. Those are as follows:

1. Regression co-efficient
2. T-statistics of the overall model
3. The associated p-values

Base Line Model

Figure 22 shows the comparison of test data and train data distribution which is obtained by building a base line linear regression model on the 80% the train data and 20% test data. Here (70, 9) is the dimension of training data set and (18,9) is the dimension of test data.

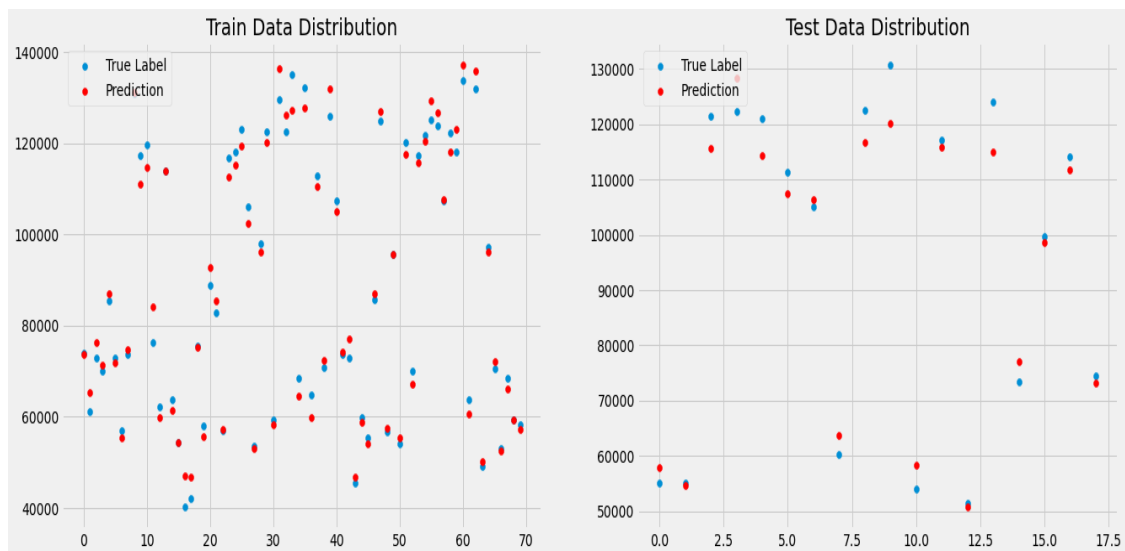


Figure 22 Baseline model of Linear Regression, Model prediction comparison

Principal Component Analysis (PCA)

Principal Component Analysis is a type of unsupervised learning algorithm that is utilised in the field of machine learning for the purpose of dimensionality reduction. Using orthogonal transformation, this statistical method transforms the observations of correlated characteristics into a set of linearly uncorrelated data. This is done by converting the correlated features into orthogonal coordinates. The output features of PCA are called as principal components. The PCA technique works by taking into account the variance of each characteristic. This is done because a high attribute reveals a good split between the classes, and as a result, it minimises the dimensionality.

This research uses the application of PCA in order to reduce the correlation between the pair of variables. After applying the PCA, the module produced 9 principal components as output of PCA. The mean squared error and the mean absolute error graphs are generated. As seen in figure 23, the model with one PCA component has a high error rate. If we keep adding PCA components the error rate starts reducing when moving towards right. Hence the first 6 Principal components are chosen as these still contains most of the information from the data set. The correlation between these 6 principal components is checked by generating a heat map in figure 24 which clearly confirms that there is no correlation between the pair of variables.

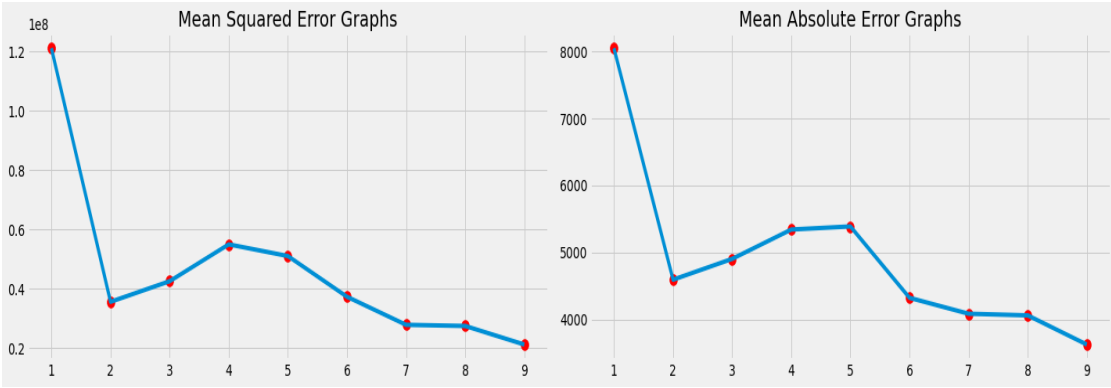


Figure 23 Mean Squared Error and Mean Absolute Error graphs

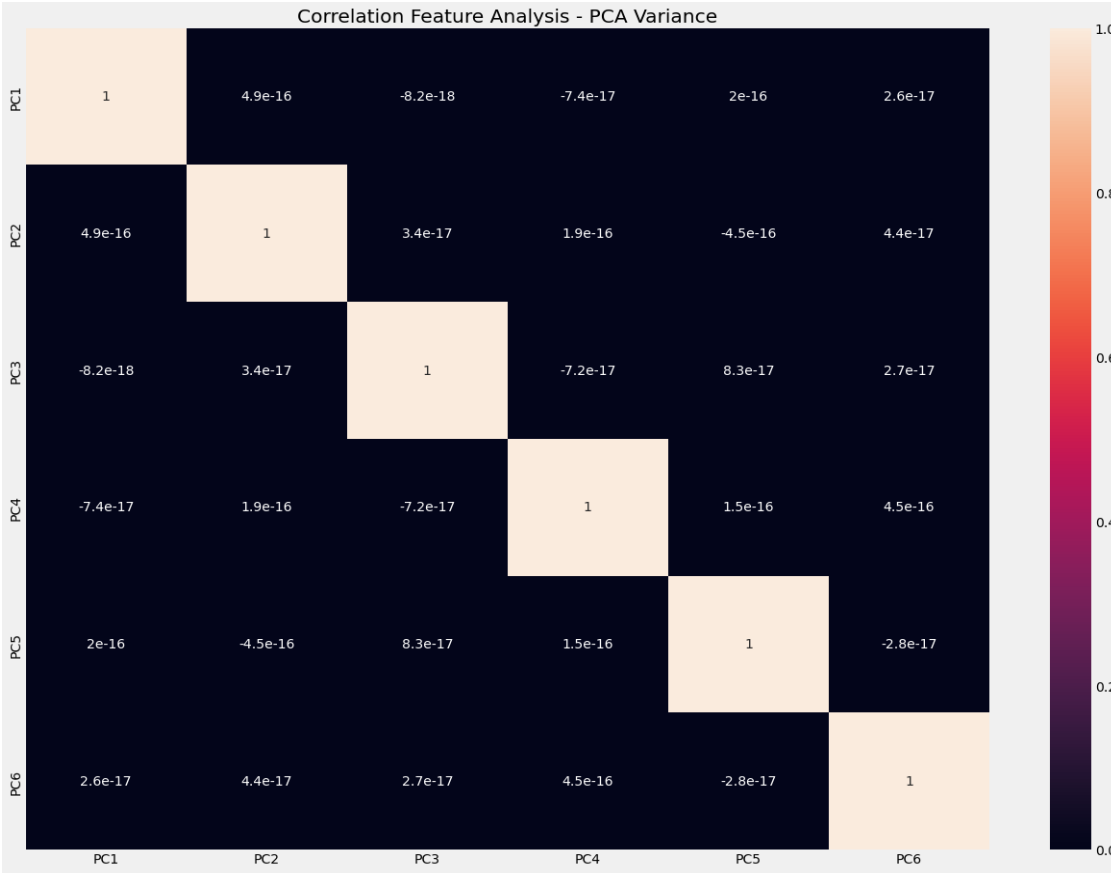


Figure 24 Correlation plot of Principal Components

Multiple Linear Regression Analysis

Figure 25 shows the comparison of test data and train data distribution which is obtained by building a linear regression model with 6 principal components on the 80% the train data and 20% test data.

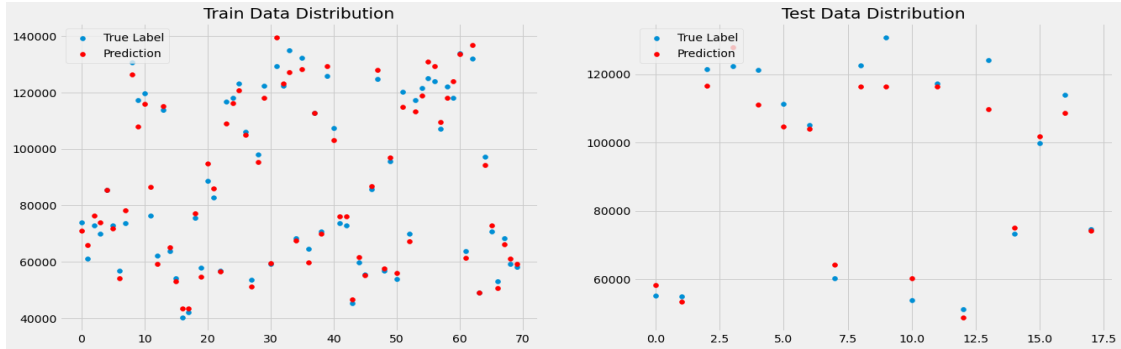


Figure 25 Linear Regression model with Principal components, Model prediction comparison

Decision Tree Regressor

Decision tree regressor is commonly used in data mining as supervised machine learning technique which builds regression or classification model in the form of tree structure. The goal is to create a model that predicts the value of a target variable based on several input variables [40] [41]. In decision tree regressor the data set is broken down into smaller and smaller parts according to a certain parameter. The output of the final parameter is predicted by the subset of the set of data. A final tree consists of a decision node and a leaf node [40] [41]. A decision node has several branches which represents the value for each attribute that is tested. A leaf node represents the decision on numerical target value.

Figure 26 shows the comparison of test data and train data distribution which is obtained by building a decision tree regressor model with 6 principal components on the 80% the train data and 20% test data.

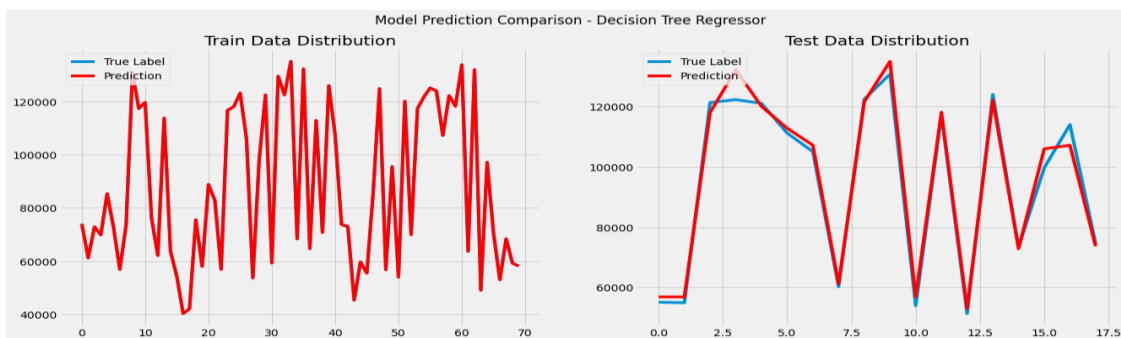


Figure 26 Decision Tree Regressor model with Principal components, Model prediction comparison

Random Forest Regressor

Figure 27 shows the comparison of test data and train data distribution which is obtained by building a Random Forest regressor model with 6 principal components on the 80% the train data and 20% test data.

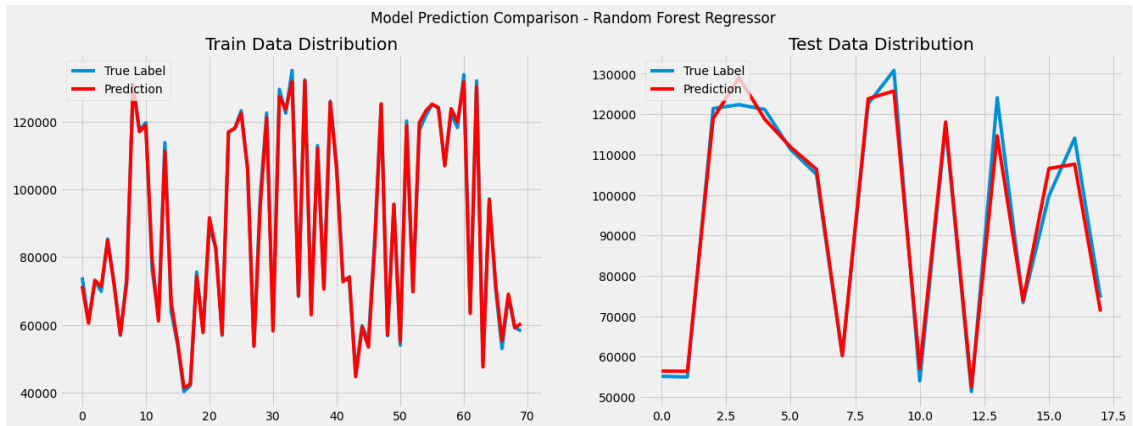


Figure 27 Random Forest Regressor model with Principal components, Model prediction comparison

XGboost Regressor

Figure 28 shows the comparison of test data and train data distribution which is obtained by building a XGboost regressor model with 6 principal components on the 80% the train data and 20% test data.

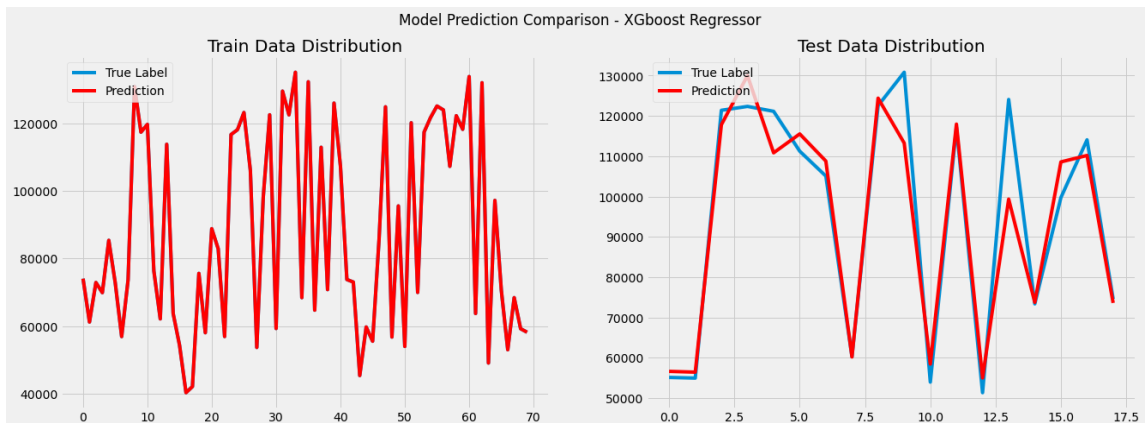


Figure 28 XGboost Regressor model with Principal components, Model prediction comparison

Results and discussion

The regression models which are trained in the modelling step are now evaluated using the metrics '**Mean Absolute Error (MAE)**', '**Root Mean Square Error (RMSE)**', and '**R²**' values. These values convey the accuracy the predictions and how far off they are from the actual data.

The MAE and RMSE values predicted by regression model indicate the differences between the predicted values and the actual values of a dependent variable.

R-squared is a statistical term that measures the amount of variation explained by an independent variable or variables in a regression model for a dependent variable. The R² score indicates the percentage of correct predictions returned by the regression model.

Table 1 Multiple Linear Regression base line model results

Type	MAE	RMSE	R²
Train data	2580.15	1.040439e+07	0.9877
Test data	3921.58	2.366561e+07	0.9718

Table 2 Multiple Linear Regression model with Principal components

Type	MAE	RMSE	R²
Train data	3046.14	1.450893e+07	0.9829
Test data	5049.91	4.216659e+07	0.9498

As seen in the table 1, the test results of Linear regression baseline model shows that 97% of model predictions are correct whereas in table 2, the test results of Linear regression model with principal components still managed to obtain 95% of correct predictions. This 2% drop in the accuracy would be due to the lowering the data's dimension using PCA without losing too much information but this 95% of accuracy is still good as implementing PCA has resolved the correlation problems in our machine learning model.

Table 3 Comparison of results

		Linear Regression	Decision Tree	Random Forest	Xgboost
Train data	MAE	3046.14	0	1068.511749	0.010148
	RMSE	1.450893e+07	0	1.896120e+06	2.010253e-04
	R²	0.9829	1.00	0.9977	1.000000
Test data	MAE	5049.91	2758.9526	3018.0768	5537.9536
	RMSE	4.216659e+07	1.370270e+07	1.615802e+07	7.026508e+07
	R²	0.9498	0.9836	0.9807	0.9163

The final machine learning models in this analysis are developed using the output data from the principal component analysis. Table 3 compares MAE, RMSE and R² values of the Linear regression model with the values of other three machine learning models after implementing the PCA. As it can read from the above table the decision tree, random forest and XGboost models are not performing well on the training data which could be due to the problem of overfitting. The linear regression model achieves over 98% accuracy in train data and 94 %accuracy in test data by having all the collinearity problems resolved. Hence it can be said that Linear regression is more sustainable model for predicting the target variable in this analysis.

Conclusion

In this research, the various factors that are changing customers thinking towards spending methods are identified, the last twenty-year trends of macroeconomic variables and credit card repayments are studied using the exploratory data analysis. The appropriate feature selection methods are identified, and the final machine learning models are developed and optimized using the data from the out of principal components analysis as it helped in analysis to solve several problems in our machine learning model. The key benefit identified from using the PCA technique is that it aided in the reduction of correlation by reducing the number of variables and adding new independent variables to the models. The core of this research focussed on developing a sophisticated machine learning model that can predict credit card receivables by utilizing a variety of machine learning methods, such as Decision Tree, Random Forest, and XGboost method. The results of all four machine learning models are compared and out of all four machine learning models, the Multiple Linear Regression method performed better on both training and test data as it achieved 98% and 95% accuracy respectively. Hence it is accurate method in this analysis for predicting the 'Total Credit Card Receivables'

Future Work

By using the core research performed on customer behaviour during several crisis outbreaks in Hong Kong

- 1) This analysis can be taken as a reference to study the cultural factors and spending behaviours of other countries where banks identify future potential of growth in their business.
- 2) Banks could survey their credit card customers to know the real-time factors which are affecting the payments to the bank. The availability of surveyed customer data will help researchers provide useful insights to banks for resolving customer problems towards making the payments
- 3) More availability of data can help in implementing multiple machine learning models to predict customer behaviour on credit card receivables
- 4) Helps banks to be aware of the future risk of drop in credit spending as by the end of 2021, the famous digital wallets attracted millions new users and thousands of new businesses in Hong Kong. Currently the digital wallet transactions stand next to the credit card payments and already growing trend in usage digital wallets is sign that this mode of payment would have surpass the use of credit cards in Hong Kong within next three years.
- 5) Banks can also start looking more towards analyzing the digital wallet transactions data for expanding their business into this area.

References

- [1] “How Credit Cards Really Work, And How Banks And Credit Card Companies Make Money From Us” published by Dollars and Sense
- [2] “How Do Credit Card Companies Make Money?” by Melissa Lambarena, lead writer on credit cards team at NerdWallet
- [3] “How Do Credit Card Companies Make Money?” by Nathan Paulus, moneygeek.
- [4] “J.P. Morgan 2019 Payments Trends” – Global Insights Report: Data has been provided to J.P. Morgan Merchant Services by Edgar, Dunn and Company, 2018
- [5] HKMA.gov.hk, September 2018. ‘Understanding Household Indebtedness in Hong Kong.’ Accessed February 2019.
- [6] EJInsight, June 2018. ‘Contactless payments seen growing in popularity in HK’.
- [7] KPMG.com, November 2017. ‘Outlook for e-commerce in Hong Kong.’ Accessed February 2019.
- [8] ‘How the pandemic has changed the payments and banking ecosystems in Hong Kong, Macau and Taiwan’ VISA CONSUMER PAYMENT ATTITUDES STUDY 2.0
- [9] ‘Data Source’ Hong Kong Monetary Authority (HKMA) <https://www.hkma.gov.hk/eng/data-publications-and-research/data-and-statistics/economic-financial-data-for-hong-kong/>
- [10] ‘Data Source’ Hong Kong Monetary Authority (HKMA) <https://www.hkma.gov.hk/eng/data-publications-and-research/data-and-statistics/monthly-statistical-bulletin/>
- [11] Investopedia ‘How to Calculate the GDP of a Country’, by POONKULALI THANGAVELU
- [12] ‘Consumption (annual variation in %)’ by Focus Economics, Economic forecasts from the worlds leading economists
- [13] ‘Government final consumption expenditure’, Australian System of National Accounts: Concepts, Sources and Methods
- [14] OECD (2022), Investment (GFCF) (indicator). doi: 10.1787/b6793677-en (Accessed on 3 April 2022)
- [15] ‘Changes in inventories’, Statistics Explained by Eurostat

- [16] Glossary of Statistical terms, OECD 'CHANGES IN INVENTORIES' System of National Accounts (SNA), 2008, 10.118 SNA 10.7 and 10.28.
- [17] 'Credit Card Lending Survey Results for Fourth Quarter 2021' Hong Kong Monetary Authority, variables in explained in Annexure.
- [18] 'What is FOB price in Exports and Imports and how it works' published as an online Export Import Training the audience who want to learn international trade of export and import.
- [19] TRADE READY, Blog for International Trade Experts; 'What are service exports, and why are they suddenly so important?' By Doris Nagel ,CEO of Globalocity LLC
- [20] 'Who is a Service Exporter?' published as an online Export Import Training the audience who want to learn international trade of export and import.
- [21] OECD (2022), Unemployment rate (indicator). doi: 10.1787/52570002-en (Accessed on 4 April 2022)
- [22] 'Base rate definition' published by Spread Bets and CFDs under Glossary of trading terms
- [23] Khurana, U., Samulowitz, H., & Turaga, D. (2018). Feature Engineering for Predictive Modeling Using Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/11678>
- [24] 'ML | Extra Tree Classifier for Feature Selection' by Alind Gupta; GeekforGeeks
- [25] Cheung, Elizabeth (22 January 2020). "China coronavirus: death toll almost doubles in one day as Hong Kong reports its first two cases". South China Morning Post.
- [26] "Hong Kong holiday camps become quarantine zones as virus fears spike". Yahoo! News.
- [27] AFP (23 January 2020). "Hong Kong turns holiday camps into quarantine zones as virus fears spike".
- [28] "Lunar New Year carnival canceled". The Standard. 23 January 2020
- [29] Chan, Kin-wa (23 January 2020). "Wuhan coronavirus: Lunar New Year Cup cancelled by government just hours after HKFA promotes the event". South China Morning Post.

- [30] "HK probes 3 more infection cases". Hong Kong's Information Services Department (in Chinese (Hong Kong)).
- [31] Chan, Kin-wa (25 January 2020). "Hong Kong declares Wuhan virus outbreak 'emergency' – the highest warning tier". Hong Kong Free Press.
- [32] Chan, Thomas (26 January 2020). "China coronavirus forces temporary closure of Hong Kong Disneyland, Ocean Park for indefinite period". South China Morning Post
- [33] "China coronavirus: Hong Kong leader hits back at criticisms of being slow". South China Morning Post. 25 January 2020.
- [34] ^ "LNY school holiday extended". Hong Kong's Information Services Department (in Chinese (Hong Kong)).
- [35] "Quarantine centre sites explained". Hong Kong's Information Services Department (in Chinese (Hong Kong)).
- [36] "To mask or not to mask: WHO makes U-turn while US, Singapore abandon pandemic advice and tell citizens to start wearing masks". South China Morning Post. 4 April 2020
- [37] Cowling, Benjamin; Ali, Sheikh Taslim; Ng, Tiffany; Tsang, Tim; Li, Julian; Fong, Min Whui; et al. (17 April 2020). "Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study". The Lancet. 5 (5): e279–e288. doi:10.1016/S2468-2667(20)30090-6. PMC 7164922. PMID 32311320.
- [38] Jump up to:^{a b c} Tufekci, Zeynep (12 May 2020). "How Hong Kong Did It". The Atlantic. MSN.
- [39] Scribbr "An introduction to multiple linear regression" by Rebecca Bevans. Revised on May 6, 2022.
- [40] Decision Tree – Regression
- [41] Rokach, Lior; Maimon, O. (2014). Data mining with decision trees: theory and applications, 2nd Edition. World Scientific Pub Co Inc. doi:10.1142/9097. ISBN 978-9814590075. S2CID 44697571

Appendix

Code

```
In [1]: # Load Libraries
import os
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

plt.style.use('fivethirtyeight')
```

```
In [2]: from warnings import filterwarnings

filterwarnings("ignore")
```

```
In [3]: # read dataset
path = "C:/Work_otherfile/Junk files/"
data = pd.read_excel('C:\\Users\\saikr\\Desktop\\Report\\Hongkong final data.xlsx')
```

```
In [4]: # remove quarterly column
#data = data.drop("Quarterly", axis = 1)
```

```
In [5]: data.head()
```

```
In [6]: data.describe()
```

Out[6]:

	Year	GDP	Pvt consumption exp	Gov consumption exp	Gross domestic fixed capital formation	Changes in inventories	Exports of goods(f.o.b.)	Exports of services	Imports of goods(f.o.b.)	Imports of services	Unemployment rate(%)
count	88.000000	88.000000	88.000000	88.000000	88.000000	88.000000	8.800000e+01	88.000000	8.800000e+01	88.000000	
mean	2010.500000	492890.829545	314163.204545	49187.636364	106402.454545	1976.465909	7.749581e+05	143152.125000	7.709768e+05	125972.420455	
std	6.380646	143334.473531	108332.385845	18105.770302	27352.443839	7627.210819	2.903218e+05	55288.689968	3.227856e+05	26393.009027	
min	2000.000000	293356.000000	175897.000000	29911.000000	63954.000000	-26857.000000	3.024920e+05	51014.000000	2.687970e+05	69870.000000	
25%	2005.000000	354555.500000	202966.250000	34527.250000	81972.750000	-2288.000000	5.169688e+05	93688.500000	4.713942e+05	103455.500000	
50%	2010.500000	481168.500000	291799.500000	40536.000000	101251.000000	1258.500000	8.108935e+05	143694.000000	7.976015e+05	135066.500000	
75%	2016.000000	627541.750000	412833.250000	62336.000000	129246.250000	4578.500000	1.013813e+06	193742.000000	1.055382e+06	147529.750000	
max	2021.000000	753690.000000	509272.000000	94933.000000	161919.000000	25235.000000	1.468515e+06	231465.000000	1.427133e+06	167457.000000	

```
In [7]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88 entries, 0 to 87
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   Year                                88 non-null    int64
 1   Quarter                            88 non-null    object
 2   Quarterly                          88 non-null    datetime64[ns]
 3   GDP                                88 non-null    int64
 4   Pvt consumption exp                88 non-null    int64
 5   Gov consumption exp                88 non-null    int64
 6   Gross domestic fixed capital formation 88 non-null    int64
 7   Changes in inventories             88 non-null    int64
 8   Exports of goods(f.o.b.)           88 non-null    int64
 9   Exports of services                88 non-null    int64
10   Imports of goods(f.o.b.)           88 non-null    int64
11   Imports of services                88 non-null    int64
12   Unemployment rate(%)               88 non-null    float64
13   Interest rate(%)                  88 non-null    float64
14   CC - Total no of Acc               88 non-null    float64
15   CC - Delinquent Amt above 90days  88 non-null    float64
16   CC - Charge off Amt                88 non-null    float64
17   CC - Rollover Amt                  88 non-null    float64
18   CC - Avg Total Receivables          88 non-null    float64
dtypes: datetime64[ns](1), float64(7), int64(10), object(1)
memory usage: 13.2+ KB
```

```
In [9]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Pvt consumption exp'], label = 'Pvt consumption exp')
plt.plot(data['Quarterly'],data['Gov consumption exp'], label='Gov consumption exp')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Hong Kong's Total Consumption Expenditure")
plt.legend()
plt.show()
```

```
In [10]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Unemployment rate(%)'], color = '#F6F5FE')
plt.scatter(data['Quarterly'],data['Unemployment rate(%)'])
plt.xlabel('Year')
plt.ylabel('%')
plt.title("Hong Kong's Unemployment rate")
plt.show()
```

```
In [11]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Gross domestic fixed capital formation'])
plt.scatter(data['Quarterly'],data['Gross domestic fixed capital formation'], color = 'red')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Hong Kong's GDFCF")
plt.show()
```

```

In [12]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Changes in inventories'])
plt.scatter(data['Quarterly'],data['Changes in inventories'], color = 'red')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Changes in Inventories")
plt.show()

In [13]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Exports of goods(f.o.b.)'], label='Exports of goods')
plt.plot(data['Quarterly'],data['Exports of services'], label='Exports of services')
plt.plot(data['Quarterly'],data['Imports of goods(f.o.b.)'], label='Imports of goods')
plt.plot(data['Quarterly'],data['Imports of services'], label='Imports of services')
#plt.scatter(data['Quarterly'],data['Gross domestic fixed capital formation'], color = 'red')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Exports, Imports of goods and services comparison")
plt.legend()
plt.show()

In [17]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Unemployment rate(%)'], color = '#FF6F5E')
plt.scatter(data['Quarterly'],data['Unemployment rate(%)'])
plt.xlabel('Year')
plt.ylabel('%')
plt.title("Hong Kong's Unemployment rate")
plt.show()

In [18]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['Interest rate(%)'], color = '#FF6F5E')
plt.scatter(data['Quarterly'],data['Interest rate(%)'])
plt.xlabel('Year')
plt.ylabel('%')
plt.title("Hong Kong's Interest rate")
plt.show()

In [19]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['CC - Total no of Acc'])
plt.scatter(data['Quarterly'],data['CC - Total no of Acc'], color = '#FF6F5E')
plt.xlabel('Year')
plt.ylabel('million')
plt.title("Credit Card - Total no of Accounts")
plt.show()

In [20]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['CC - Charge off Amt'])
plt.scatter(data['Quarterly'],data['CC - Charge off Amt'], color = '#FF6F5E')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Credit Card - Chargeoff Amount")
plt.show()

In [21]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['CC - Rollover Amt'])
plt.scatter(data['Quarterly'],data['CC - Rollover Amt'], color = '#FF6F5E')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Credit Card Rollover Amount")
plt.show()

In [22]: plt.figure(figsize = (15, 5))
plt.plot(data['Quarterly'],data['CC - Avg Total Receivables'])
plt.scatter(data['Quarterly'],data['CC - Avg Total Receivables'], color = '#FF6F5E')
plt.xlabel('Year')
plt.ylabel('HK$ million')
plt.title("Credit Card Avg Total Receivables")
plt.show()

In [23]: # remove quarterly column
data = data.drop("Quarterly", axis = 1)

In [24]: # select continous value only without datetime variable
continous_value = data.iloc[:, 2:]

continous_value.sample(3)

In [25]: plt.figure(figsize = (20, 15))
sns.heatmap(continous_value.corr(), annot = True).set_title("Correlation Feature Analysis");

```

▼ Feature selection

```

In [38]: # Fearture selection with help of ExtratreeClassifier
X = combine_data.drop(["CC - Avg Total Receivables", 'Quarter'],axis=1) #independent columns
y = combine_data[["CC - Avg Total Receivables"]] #target column i.e price range
from sklearn.ensemble import ExtraTreesClassifier
import matplotlib.pyplot as plt
model = ExtraTreesClassifier()
model.fit(X,y)
print(model.feature_importances_) #use inbuilt class feature_importances of tree based classifiers
#plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(12).plot(kind='barh')
plt.show()

[0.07287356 0.07045977 0.07413793 0.07471264 0.07206897 0.07471264
 0.0716092  0.07126437 0.07172414 0.04091954 0.03827586 0.04850575
 0.07528736 0.07517241 0.06827586]

```

```

In [39]: # Top 12 feature which based on random DT
selected_fea = feat_importances.nlargest(12).index.to_list()
selected_fea

Out[39]: ['CC - Delinquent Amt above 90days ',
          'CC - Charge off Amt',
          'Gross domestic fixed capital formation',
          'Exports of goods(f.o.b.)',
          'Gov consumption exp',
          'GDP',
          'Changes in inventories',
          'Imports of services',
          'Exports of services',
          'Imports of goods(f.o.b.)',
          'Pvt consumption exp',
          'CC - Rollover Amt']

In [40]: # selected_fea.append('CC - Avg Total Receivables')
# selected_fea.append('Quarter')

In [41]: # Selecting only variable which we have selected from extratree classifier process
stand_df = combine_data[selected_fea]
stand_df.head()

In [42]: # Mapping target variable
stand_df = pd.concat([data['CC - Avg Total Receivables'], stand_df], axis = 1)
stand_df.head(2)

In [44]: # Building a OLS model so we can feature which are most significant to the model >=0.5
import statsmodels.api as sm

x = stand_df.drop(['CC - Avg Total Receivables'],axis=1)
y = stand_df[["CC - Avg Total Receivables"]]

# Statsmodels.OLS requires us to add a constant.
x = sm.add_constant(x)
model = sm.OLS(y,x,random_states = 10)
results = model.fit()
print(results.summary())

In [45]: # Dropping variable manually where p value is >.05
stand_df = stand_df.drop(['CC - Charge off Amt','Imports of services','Pvt consumption exp'],axis=1)
stand_df.head()

```

▼ Feature Engineering

```

In [46]: # Standardizing data to make variable point in equal scale
from sklearn.preprocessing import StandardScaler
stand_df_v1 = StandardScaler().fit_transform(stand_df.iloc[:,1:])

In [47]: # Mapping columns name with Standardized data
stand_df_v1 = pd.DataFrame(stand_df_v1,columns=stand_df.columns[1:])
stand_df_v1.head()

In [48]: # Mapping target variable with standardized data
stand_df_v1 = pd.concat([data['CC - Avg Total Receivables'], stand_df_v1], axis = 1)
stand_df_v1.tail(3)

Out[48]:
   CC - Avg Total  CC - Delinquent  Gross domestic  Exports of  Gov  GDP  Changes in  Exports of  Imports of  CC -
   Receivables    Amt above       fixed capital  goods(f.o.b.)  consumption  Exp  inventories  of services  goods(f.o.b.)  Rollover
   85    121661.0      -0.002280      0.637968      1.605909      1.951738  1.294100      1.201719    -0.256416      1.548076    -0.350564
   86    123918.0      -0.347855      0.813137      2.046412      2.264082  1.678157      2.559490      0.244915      1.843162    -0.554461
   87    129477.0      -0.353429      0.961427      2.402615      2.245584  1.813474      -3.802006      0.293666      2.044442    -0.327128

In [49]: # Saving with original data so rest of the code work fine
combine_data = stand_df_v1.copy()
combine_data.head(2)

Out[49]:
   CC - Avg Total  CC - Delinquent  Gross domestic  Exports of  Gov  GDP  Changes in  Exports of  Imports of  CC -
   Receivables    Amt above       fixed capital  goods(f.o.b.)  consumption  Exp  inventories  of services  goods(f.o.b.)  Rollover
   0    40300.9615    -0.041297    -0.787946    -1.603641    -0.886963  -1.205828      0.330250    -1.554365    -1.518895    -0.251090
   1    42101.9885    -0.108182    -0.781696    -1.492499    -1.070769  -1.173004      0.335524    -1.517966    -1.429129      0.114259

```

Regression Analysis

```
In [50]: #!pip install statsmodels
```

```
In [51]: import statsmodels.api as sm
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
```

```
In [52]: # review dataset used
combine_data.head()
```

```
Out[52]:
```

	CC - Avg Total Receivables	CC - Delinquent Amt above 90days	Gross domestic fixed capital formation	Exports of goods(f.o.b.)	Gov consumption exp	GDP	Changes in inventories	Exports of services	Imports of goods(f.o.b.)	CC - Rollover Amt
0	40300.9615	-0.041297	-0.787946	-1.603641	-0.886963	-1.205828	0.330250	-1.554365	-1.518895	-0.251090
1	42101.9885	-0.108182	-0.781696	-1.492499	-1.070769	-1.173004	0.335524	-1.517966	-1.429129	0.114259
2	45369.2640	-0.030149	-0.490299	-1.303296	-1.011056	-1.051889	0.228585	-1.443548	-1.335185	0.290032
3	49065.3320	0.276409	-0.554130	-1.351808	-1.000446	-1.018286	-0.038169	-1.424630	-1.337372	1.095465
4	51360.6910	0.789197	-0.729373	-1.571996	-0.818639	-1.221236	-0.138252	-1.517857	-1.478087	1.251187

```
In [53]: # apply one hot encoder into our quarter categorical variable
# regression_data = combine_data.join(pd.get_dummies(combine_data['Quarter'])).drop("Quarter", axis = 1)
regression_data = combine_data.copy()
```

```
In [55]: # define feature & target
feature = regression_data.drop('CC - Avg Total Receivables', axis = 1)
target = regression_data['CC - Avg Total Receivables']
```

```
In [56]: # split dataset
X_train, X_test, y_train, y_test = train_test_split(feature, target, random_state = 180, test_size = 0.2)

# check dimension
print(X_train.shape, X_test.shape)

(70, 9) (18, 9)
```

```
In [57]: # baseline model multiple linear regression
model = LinearRegression()

# training model
model.fit(X_train, y_train)

# evaluate model
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)
```

```
In [58]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))

ax[0].scatter(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].scatter(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].scatter(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].scatter(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [59]: # metrics evaluation model

#baseline = {
#    'Type': ['Train', 'Test'],
#    '#MSE': [mean_squared_error(y_train, train_preds), mean_squared_error(y_test, test_preds)],
#    '#MAE': [mean_absolute_error(y_train, train_preds), mean_absolute_error(y_test, test_preds)]
#}

def get_metrics(true, preds):
    return {
        'MAE': mean_absolute_error(true, preds),
        'RMSE': mean_squared_error(true, preds),
        'R2': r2_score(true, preds)
    }

result = pd.DataFrame(['train'] + list(get_metrics(y_train, train_preds).values()),
                      ['test'] + list(get_metrics(y_test, test_preds).values()),
                      columns = ['Type', 'MAE', 'RMSE', 'R2'])

result

#metrics_evaluate_baseline = pd.DataFrame(baseline)

#metrics_evaluate_baseline
```

```
Out[59]:
```

	Type	MAE	RMSE	R2
0	train	2580.154517	1.040439e+07	0.987757
1	test	3921.588164	2.366561e+07	0.971839

```
In [60]: train_stat = sm.add_constant(X_train)
train_sum = sm.OLS(y_train, train_stat).fit()
train_sum.summary()
```

```
In [61]: test_stat = sm.add_constant(X_test)
test_sum = sm.OLS(y_test, test_stat).fit()
test_sum.summary()
```

Principal Component Analysis

```
In [62]: from sklearn.decomposition import PCA
         from sklearn.preprocessing import StandardScaler
```

```
In [63]: def pca_score(n_comp, feature = feature, target = target):
         """
         params:
             - total PCA components
             - feature dataset
             - target dataset
         output:
             - mean square error, mean absolute error
         """

         # normalize feature
         norm = StandardScaler()
         feature_norm = norm.fit_transform(feature)

         # define PCA with n_coms
         pca = PCA(n_components = n_comp)

         # apply PCA
         pca_feature = pca.fit_transform(feature_norm)

         # split dataset
         X_train, X_test, y_train, y_test = train_test_split(pca_feature, target, test_size = 0.2, random_state=100)

         # check dimension
         print(X_train.shape, X_test.shape)

         # training model
         model = LinearRegression()
         model.fit(X_train, y_train)

         # prediction
         preds = model.predict(X_test)

         # evaluate model
         mse_score = mean_squared_error(y_test, preds)
         mab_score = mean_absolute_error(y_test, preds)

         return mse_score, mab_score
```

```
In [64]: regression_data.sample(3)
```

```
In [65]: list_mse, list_mae = [], []
```

```
         for n in range(1, 10):
             mse, mae = pca_score(n)
             list_mse.append(mse)
             list_mae.append(mae)
```

```
(70, 1) (18, 1)
(70, 2) (18, 2)
(70, 3) (18, 3)
(70, 4) (18, 4)
(70, 5) (18, 5)
(70, 6) (18, 6)
(70, 7) (18, 7)
(70, 8) (18, 8)
(70, 9) (18, 9)
```

```
In [66]: # plot metrics visualisation
         fig, ax = plt.subplots(1, 2, figsize = (20, 5))

         ax[0].plot(np.arange(1, len(list_mse)+1), list_mse)
         ax[0].scatter(np.arange(1, len(list_mse)+1), list_mse, s = 100, color = 'red')
         ax[0].set_title("Mean Squared Error Graphs")
         # ax[0].set_xticks(np.arange(1, 15), np.arange(1, 15))

         ax[1].plot(np.arange(1, len(list_mse)+1), list_mae)
         ax[1].scatter(np.arange(1, len(list_mse)+1), list_mae, s = 100, color = 'red')
         ax[1].set_title('Mean Absolute Error Graphs')
         # ax[1].set_xticks(np.arange(1, 15), np.arange(1, 15))

         fig.tight_layout()
         plt.show()
```

```
In [67]: # normalize feature
         norm = StandardScaler()
         feature_norm = norm.fit_transform(feature)

         # define PCA with n_coms
         pca = PCA(n_components = 6)

         # apply PCA
         pca_feature = pca.fit_transform(feature_norm)

         # split dataset
         X_train, X_test, y_train, y_test = train_test_split(pca_feature, target, test_size = 0.2, random_state=100)

         # check dimension
         print(X_train.shape, X_test.shape)
```

```
(70, 6) (18, 6)
```

```
In [68]: # wrapped dataframe
         data_pca = pd.DataFrame(data = pca_feature, columns = [f"PC{x+1}" for x in range(6)])

         data_pca.head()
```

```
Out[68]:
```

	PC1	PC2	PC3	PC4	PC5	PC6
0	2.816020	-1.033942	-0.434308	0.320234	0.191018	-0.632588
1	2.853401	-0.920537	-0.408986	0.235305	-0.117131	-0.601119
2	2.602392	-0.669171	-0.270664	0.104989	-0.239499	-0.721009
3	2.919652	-0.023669	0.040871	0.009209	-0.608141	-0.551136
4	3.348777	0.368225	0.142453	-0.009923	-0.337412	-0.439274

```
In [69]: plt.figure(figsize = (20, 15))
         sns.heatmap(data_pca.corr(), annot = True).set_title("Correlation Feature Analysis - PCA Variance");
```

```
In [70]: model = LinearRegression()

# training model
model.fit(X_train, y_train)

# evaluate model
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)
```

```
In [71]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))

ax[0].scatter(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].scatter(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].scatter(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].scatter(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [79]: # define model
model = LinearRegression()

# training
model.fit(X_train, y_train)

# evaluate
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)

# get metrics
result = pd.DataFrame(['train'] + list(get_metrics(y_train, train_preds).values()),
                      ['test'] + list(get_metrics(y_test, test_preds).values()),
                      columns = ['Type', 'MAE', 'RMSE', 'R2'])
result
```

```
Out[79]:
```

	Type	MAE	RMSE	R2
0	train	3046.142480	1.450893e+07	0.982928
1	test	5049.911323	4.216659e+07	0.949823

```
In [80]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))
fig.suptitle("Model Prediction Comparison - Multiple Linear Regression")

ax[0].plot(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].plot(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].plot(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].plot(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [81]: # define model
model = DecisionTreeRegressor()

# training
model.fit(X_train, y_train)

# evaluate
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)

# get metrics
result = pd.DataFrame(['train'] + list(get_metrics(y_train, train_preds).values()),
                      ['test'] + list(get_metrics(y_test, test_preds).values()),
                      columns = ['Type', 'MAE', 'RMSE', 'R2'])
result
```

```
Out[81]:
```

	Type	MAE	RMSE	R2
0	train	0.000000	0.000000e+00	1.000000
1	test	2972.952639	1.820088e+07	0.978342

```
In [82]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))
fig.suptitle("Model Prediction Comparison - Decision Tree Regressor")

ax[0].plot(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].plot(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].plot(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].plot(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [83]: # define model
model = RandomForestRegressor()

# training
model.fit(X_train, y_train)

# evaluate
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)

# get metrics
result = pd.DataFrame(['train'] + list(get_metrics(y_train, train_preds).values()),
                      ['test'] + list(get_metrics(y_test, test_preds).values())),
                      columns = ['Type', 'MAE', 'RMSE', 'R2'])
result
```

```
Out[83]:
```

	Type	MAE	RMSE	R2
0	train	1016.687908	1.675531e+06	0.998028
1	test	3159.583655	2.221176e+07	0.973569

```
In [84]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))
fig.suptitle("Model Prediction Comparison - Random Forest Regressor")

ax[0].plot(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].plot(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].plot(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].plot(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [85]: # define model
model = XGBRegressor()

# training
model.fit(X_train, y_train)

# evaluate
train_preds = model.predict(X_train)
test_preds = model.predict(X_test)

# get metrics
result = pd.DataFrame(['train'] + list(get_metrics(y_train, train_preds).values()),
                      ['test'] + list(get_metrics(y_test, test_preds).values())),
                      columns = ['Type', 'MAE', 'RMSE', 'R2'])
result
```

```
Out[85]:
```

	Type	MAE	RMSE	R2
0	train	0.010148	2.010253e-04	1.000000
1	test	5537.953686	7.026508e+07	0.916387

```
In [86]: fig, ax = plt.subplots(1, 2, figsize = (20, 7))
fig.suptitle("Model Prediction Comparison - XGboost Regressor")

ax[0].plot(np.arange(len(y_train)), y_train, label = 'True Label')
ax[0].plot(np.arange(len(y_train)), train_preds, label = 'Prediction', color = 'red')
ax[0].set_title("Train Data Distribution")
ax[0].legend(loc = 'upper left')

ax[1].plot(np.arange(len(y_test)), y_test, label = "True Label")
ax[1].plot(np.arange(len(y_test)), test_preds, label = "Prediction", color = 'red')
ax[1].set_title("Test Data Distribution")
ax[1].legend(loc = 'upper left')

plt.show()
```

```
In [87]: # re-define result
evaluate_before_tuning = pd.DataFrame({
    'Model' : ['Linear Regression', 'Decision Tree', 'Random Forest', 'Xgboost'],
    'Train - RMSE' : [1.450893e+07, 0.000000e+00, 1.872868e+06, 2.010253e-04],
    'Train - MAE' : [3046.142480, 0.000000, 1070.754737, 0.010148],
    'Train - R2' : [0.982928, 1.000000, 0.997796, 1.000000],
    'Test - RMSE' : [4.216659e+07, 1.643387e+07, 2.119435e+08, 7.026508e+07],
    'Test - MAE' : [5049.911323, 3025.730417, 2954.380666, 5537.953686],
    'Test - R2' : [0.949823, 0.980935, 0.980444, 0.916387]
})

# rounded result
evaluate_before_tuning.iloc[:, 1:] = evaluate_before_tuning.iloc[:, 1:].round(2)

# overview
evaluate_before_tuning
```


```
Out[87]:
```


	Model	Train - RMSE	Train - MAE	Train - R2	Test - RMSE	Test - MAE	Test - R2
0	Linear Regression	14508930.0	3046.14	0.98	42166590.0	5049.91	0.95
1	Decision Tree	0.0	0.00	1.00	16433870.0	3025.73	0.98
2	Random Forest	1872868.0	1070.75	1.00	211943500.0	2954.38	0.98
3	Xgboost	0.0	0.01	1.00	70265080.0	5537.95	0.92

LinkedIn Posts

LinkedIn Post – 1

Link to post -1: https://www.linkedin.com/posts/sai-krishna-vadlamudi_bankingindustry-banking-loans-activity-6925213777915203584-cIc?utm_source=linkedin_share&utm_medium=member_desktop_web



Sai Krishna Vadlamudi (He/Him) • You
MSc Data Analysis for Business Intelligence| Former Senior GIS Engineer at A...
1mo • Edited • 

...

MSc Data Analysis for Business Intelligence Project
Duration: March'2022 - May'2022

Post #1
Introducing the project:

I am very excited to announce that I have been working on Msc Project together with a multinational bank and the [University of Leicester](#)



The project primarily focuses on analyzing the customer's response to various changes in the economy, and interest rates in order to predict the future banking based on the macroeconomic variables.

The ultimate goal of predictions is to support the bank in developing the best solutions for clients rather than the best products in general and for considering customer behavior seriously to improve their satisfaction.

I thank my team members Vinay, [Manoj Ganji](#), and Rosmi for working together with me.









A massive thank you to my supervisors [Jeremy Levesley](#)(prof.), [Alexander Gorban](#) (prof.), and program director Andrew Morozov(Dr.) for their efforts and welcoming me into this project.

[#bankingindustry](#) [#banking](#) [#loans](#) [#creditcards](#) [#mortgages](#)
[#customerbehaviour](#) [#behavioranalysis](#)
[#innovation](#) [#technology](#) [#teamwork](#) [#growth](#) [#newadventures](#)
[#dataanalysis](#) [#businessintelligence](#) [#analytics](#)
[#datamining](#) [#data](#) [#machinelearning](#) [#datascience](#) [#regression](#) [#pca](#)
[#timeseriesanalysis](#) [#predictions](#) [#future](#) [#futureofbusiness](#) [#futurebanking](#)
[#unitedkingdom](#) [#hongkong](#)

  Daisy .. and 31 others

9 comments

Reactions



+24

LinkedIn Post – 2

Link to post - 2: https://www.linkedin.com/posts/sai-krishna-vadlamudi_datacollection-datapreprocessing-exploratorydataanalysis-activity-6928071689813540864-jfxk?utm_source=linkedin_share&utm_medium=member_desktop_web



Sai Krishna Vadlamudi (He/Him) • You

MSc Data Analysis for Business Intelligence | Former Senior GIS Engineer at A...

3w •

MSc Data Analysis for Business Intelligence Project

Duration: March'2022 - May'2022

Post #2

Overview of the methods implemented

a) **#Datacollection:**

This is a primary step and important aspect of the research because the data obtained from publicly available sources must be information-rich and reliable to perform analysis. Hence we relied on the statistical data which was published by the UK and Hong Kong government authorities

Hong Kong Monetary Authority (HKMA), Office for National Statistics Statista, etc as primary data sources to gather the Macroeconomic variables data.

b) **#Datapreprocessing** and **#ExploratoryDataAnalysis(#EDA)** :

This is also an important step because the data mining algorithms cannot be directly applied to the real-world data sets which contain raw data of poor quality with missing values, inconsistent information, and unnecessary data points called outliers. The issues with the data are handled in the preprocessing step to transform the data into a meaningful format. EDA is carried out to identify the important characteristics using various visualizations like boxplots, histograms, line charts, scatter plots, etc.

c) **#Regressionanalysis** :

This statistical method is used to examine the linear relationship between a dependent and independent variable(s). Typically this analysis predicts the target variable(dependent variable) whether and how some phenomenon influences the other or how several variables are related.

d) **#PrincipalComponentAnalysis(#PCA):**

This is a dimension reduction technique implemented to reduce the size of large data sets of variables into a smaller set that still contains most of the information from the large set. The new transformed features are called Principal Components, they do not exhibit a correlation between the pair of variables and the importance of the component decreases when moving from 1 to n.

#bankingindustry #banking #loans #creditcards #mortgages

#customerbehaviour #behavioranalysis #innovation

#technology #teamwork #growth #newadventures

#dataanalysis #businessintelligence #analytics #datamining

#data #machinelearning #datascience #regression #pca #timeseriesanalysis

#predictions #future #futureofbusiness #futurebanking #unitedkingdom

#hongkong



LinkedIn Post - 3

Link to post - 3: https://www.linkedin.com/posts/sai-krishna-vadlamudi_bankingindustry-banking-loans-activity-6928826167110062080-sEm9?utm_source=linkedin_share&utm_medium=member_desktop_web



Sai Krishna Vadlamudi (He/Him) • You

MSc Data Analysis for Business Intelligence | Former Senior GIS Engineer at A...
3w • 🌐

MSc Data Analysis for Business Intelligence Project
Duration: March'2022 - May'2022

Post #3

Learnings from the project:

- 1) Versatile web research skills for identifying reliable data sources from trusted organizations.
- 2) Checking sources for finding journal articles, academic publications and looking for citations and evidence.
- 3) Integrating previous work experience and knowledge of core data analysis modules for project initiation, planning, execution, monitoring, and closing within the given deadlines.
- 4) Working in a team with Vinay, [Manoj Ganji](#), and Rosmi, learning faster, understanding individual talent and strengths for working towards goals.
- 5) Planning and scheduling meetings for an open discussion and exchanging feedback helped me to showcase my responsibility and personal and professional commitment.
- 6) Seizing the opportunity to meet in person and build positive relationships with industry experts [Sandeep Birdie](#) and [Jeremy Levesley](#). Thanks to the [University of Leicester Leicester Innovation Hub](#) for bringing everyone together.

Once again a massive thank you to my supervisors [Jeremy Levesley](#)(prof.), [Alexander Gorban](#)(prof.), [Alec Odeide](#), and program director Andrew Morozov(Dr.), our team will be grateful for receiving your assistance and feedback.

[#bankingindustry](#) [#banking](#) [#loans](#) [#creditcards](#) [#mortgages](#)
[#customerbehaviour](#) [#behavioranalysis](#)
[#innovation](#) [#technology](#) [#teamwork](#) [#growth](#) [#newadventures](#)
[#dataanalysis](#) [#businessintelligence](#) [#analytics](#)
[#datamining](#) [#data](#) [#machinelearning](#) [#datascience](#) [#regression](#) [#pca](#)
[#timeseriesanalysis](#) [#predictions](#) [#future](#) [#futureofbusiness](#) [#futurebanking](#)
[#unitedkingdom](#) [#hongkong](#)

👍 Jeremy Levesley and 8 others

Reactions



