MASA
Malaysian Actuarial Student Association

PPM-016-10-18102018

HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA

Research Topic: Impact of Covid-19 Pandemic on Banking Industry
Field: Banking Industry
Authors: Yap Zhi Ling, Gillian Goh Qiu Jin, Teo Jun Hong
Supervisor(s): Richard Lau Yee Heng
Published by: Malaysian Actuarial Student Association (MASA)
Date of Publication: 17th August 2020

Website : http://www.masassociation.org/
Facebook : http://www.facebook.com/MASAssociation

PPM-016-10-18102018
HERIOT-WATT UNIVERSITY MALAYSIA
1, JALAN VENNA P5/2
PRECINCT 5
62200 PUTRAJAYA

TABLE OF CONTENTS

1 ABSTRACT

In this research, we are going to discuss the impact of Covid-19 in the banking industry based on the control level of Covid-19 in a country. With the doubt, we come out with an argument -- "well-controlled countries tend to have a lower negative impact on the banking industry than those undercontrolled countries during the Covid-19 pandemic".

Here, we are considering Malaysia and China as the undercontrolled countries and USA as the not well-controlled countries. This is because Malaysia and China have drastically reduced the number of new cases, what is known as flattening the curve. In contrast, the coronavirus cases in the USA are still growing rapidly since the first week of March, and the USA now has the most confirmed cases and deaths compared to any other country worldwide.

2 INTRODUCTION AND STATEMENT OF PROBLEM

2.1 Introduction

The coronavirus (Covid-19) outbreak has snowballed into a global crisis and countries around the world are suffering in the economic recession and destabilizing effect of the pandemic. In order to cease the spread of Covid-19, countries around the world have implemented lockdown restrictions and the economic crisis has affected the entire world. According to the IMF, the world GDP has plunged 4.9% in 2020 and is expected to wipe out $12 trillion over two years due to the global coronavirus pandemic.

Without any doubt, the banking industry was unsurprisingly having an enormous impact and resulting in a reduction in business. For instance, TD Bank in the US is encouraging its customers to bank digitally and has also closed selected branches and reduced working hours and UniCredit of Italy has closed 70% branches while the remaining 30% are operating on alternate days.

Looking back to Malaysia, Bank Negara Malaysia (BNM) has also ordered all banks to grant an automatic six-month deferment of all loan/financing repayments starting from April 1, 2020. With the high proportion of household exposure at 58.2% of loan books, the income and reserve of the banks will also be affected.

2.2 Problem Statement

During the pandemic, banking stocks were highly impacted and most banks saw a price slump in mid-March during the period of 01 December 2019 to 30 April 2020. The high volatility in the stock market has depressed the banks' valuation, resulting in a drop of bank valuations in all counties around the world (P/NAV multiple experienced a severe downfall from 1.00x on 31 December 2019 to 0.69x on 30 April 2020).

On the other hand, the low interest rate and the significant impact of the COVID-19 has reduced the core banking profitability in mature markets. Hence during the economic recession, banks will need more reserve in hand and can't or won't take on additional debt. However, in order to continue financing the real economy and support the economy recovery, banks are called to identify the short-term phenomena and long-term impacts which would take into consideration of credit risk management and reclassification as well as loan loss provisions.

Hence, by mining the revenue of banks in a specific country, we will be able to discover trends in the banking industry. Creating content of the trending topics is one key for banks to compare their recovery rate and competitiveness with other banks around the world under the new norm and push them to accelerate their transformation in order to compete in the banking industry. In addition, this research will be developing a machine learning model to discover the future prospect of the banking industry in terms of revenue and identify the difference between undercontrolled countries and not-well controlled countries.

3 LIMITATIONS OF STUDY

3.1 Dataset

All dataset used in this research paper is collected from the internet. For each country, at least 4 banks will be considered throughout the whole research. In addition, a 5-year stock price data has been collected on a daily basis, including the daily high and low price as well as the adjusted close price. As for the revenue, we have collected the data for the past 4 years on a quarterly basis.

3.1.1 Revenue

The revenue dataset is collected from the financial report quarterly. Hence, due to the limited amount of data, it is hard to reflect the changes of the revenue of banking on a daily basis to reflect the impact of the pandemic. Some revenues of the bank are being predicted without any support of the news, thus it may affect the accuracy of the model.

4 METHODOLOGY

4.1 Introduction

This project is a data science project. The methodology which will be used is the data mining methodology, thus focusing on creating a machine learning model. In data mining projects, the methodology that we use is the CRISP-DM model. CRISP-DM stands for cross-industry process for data mining which provides a structured approach in planning a data mining project. (Crisp DM methodology-Smart Vision Europe, 2020)

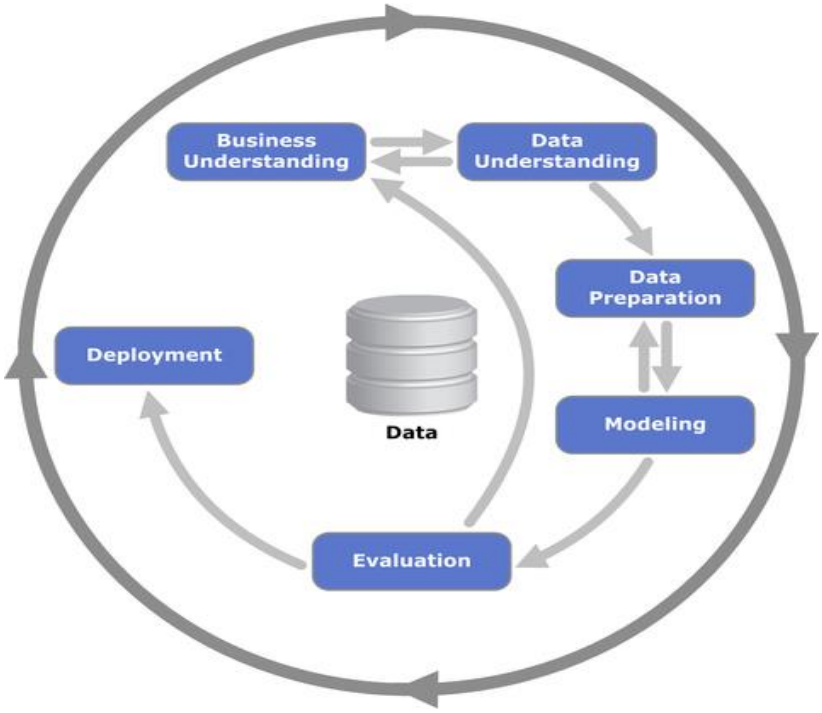


Diagram 1: CRISP-DM open standard process

4.2 Data Mining Method

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes (ori.hhs.gov, 2020). In this project, the data type will be numeric type as the data source is collected from Yahoo Finance, where the site has rich data.

4.2.1 Dataset

Dataset is a collection of data that can be used by the computer for statistical and inference purposes (Anderson et al.,2017). In this project, the dataset will be the revenue and stock price that had been collected via Yahoo Finance. When developing an algorithm, more than 70% of the total data is used in the

project. A large quantity of datasets are used to train models at the best level for the best results. Therefore, datasets are crucial in machine learning (Medium, 2020).

The dataset will be stored as a comma separated file (csv) when it is scraped. The comma separated file is being chosen to store the data because Pandas can read csv files easily compared to the other file types. Then the list is converted into the Pandas dataframe. After collecting the revenue and stock prices from Malaysia, China and US banking industries from Yahoo Finance, the data frames are concatenated into one data frame,and output as a csv file via Pandas built-in function, to_csv().

4.2.2 Data Pre-Processing

Data Preprocessing is a technique used to convert the raw data into a clean data set. As data is collected from different sources, the data is in raw format. Hence, it is not feasible for analysis. (GeeksforGeeks, 2020) In this project, the data that will be used is numeric data, hence string change to numeric preprocess is needed. Then follow by making all the text data consistent. For example, different variations in input capitalization such as 'America' and 'america' will cause the different types of output or no output in the machine learning model. This is due to mixed-case occurrences of the word 'America' and insufficient evidence for the machine learning model to learn and update the weight and bias for the better result. Therefore, the text needs to be consistent before feeding into machine learning.

4.2.3 Transformation

According to Alley (2018), data transformation is the process to convert the format or structure of a data into another format to suit the required data format for the data modeling. The five techniques are normalisation, aggregation, discretization, attribute construction (GeeksForGeeks, 2019; Yaghini, 2009).

Data normalisation is a process to manipulate the data by scaling down or scaling up the range of data before it becomes used for further stages (Patro and Sahu, 2015; Yaghini, 2009). According to Patel (2019) and Patro and Sahu (2015) the example of normalization techniques are Min-Max normalization, Z-score normalization and Decimal scaling normalization. While data aggregation is the process to gather and express the raw data in a summary for statistical analysis (IBM,2020). Data discretization is a process to convert the continuous data attribute value into a set of finite intervals with minimal data loss (Jin, Breitbart, and Muoh,2009). Next will be attributed construction, it is a technique that inferred the existing attributes and replace or add new attributes (cs.ccsu.edu,2020). While the text transformation will be a bag of words, term frequency-inverse document frequency and one hot encoding.

4.2.4 Modeling (Algorithm)

The topic modeling will be done by using the LDiA library from scikit-learn and gensim. The LDiA library provided by scikit-learn and gensim are unsupervised learning will group the word together in a cluster without inferring the vector of the topic. The classification will be using linear regression, SVM and SGD.

4.2.5 Evaluation Metrics

Normally the evaluation metrics for the predictive machine model is the model accuracy, RMSE, MAE and r-square. Model accuracy is the measurement of determining the best model at identifying patterns and relationships between variables in a dataset based on the input data. The better a model can generalise to 'unseen' data, the better insights and predictions can be produced, which deliver even higher business value. (Accuracy, 2020) Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. It is a measure of how far the data points are from the regression line. Another metric used is the Mean Absolute Error (MAE). It is the amount of error in the measurements. It is the difference between the measured value and actual value. (Statistics How To, 2020) Furthermore, R-squared (R^2) is used to represent the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model. It explains the extent of the variance of one variable explaining the variance of the second variable. (R-Squared, 2020)

5 LITERATURE REVIEW

5.1 Domain LR -> Impact of pandemic on banking industry

Covid-19 is not the single pandemic observed in human history, but it has potential serious implications towards the way the world works. It is a paramount challenge for our modern societies and healthcare systems. The consequences of the pandemic for our global economy and financial sector are unpredictable. The impact of the pandemic towards the economy has become sophisticated as the outbreak continues. The consensus by economists worldwide is that we are heading for a significant economic downturn sooner or later although the duration is arguable. Fortunately, we have seen public authorities take immediate and different measures to address the issues to sustain the economy, the banking sector and, ultimately, the people.

The pandemic has forced the financial institutions including the banking sector, to form contingency plans to tackle various scenarios should the crisis spiral out of control. Banks actively construct different measures to mitigate the negative impact on its health. The most significant affected segment has been the debt market as the main profit of banks are generated from loans. The leveraged loan market, in particular, came under remarkable stress during the month of March. Bank-loan mutual funds, among the main holders of leveraged loans, suffered massive outflows that were reminiscent of the outflows they experienced during the 2008 crisis. If drastic action and policy are not taken promptly, the economy will be suffering from unwanted consequences especially the vulnerable population.

Banking flows to emerging economies remain important as a source of financing to banks and directly to the non-financial private sector. Global banking system has shown its fragilities during the global financial crisis of 2008. That may prevail again in the ongoing Covid-19 crisis. While financial reforms aimed to strengthen capital and liquidity buffers, bank balance sheets are under severe pressure once again as the demand for liquidity has grown and credit risk has soared. The unprecedented falls in economic activity and steep rise in unemployment will likely prompt significant rises in non-performing loans. The financial markets of emerging and developing economies are particularly exposed. Due to the global liquidity crunch and greater risk aversion, emerging and developing countries have suffered significant capital outflows. Maintaining financial stability will be critical to avoid an amplification of the costs of the crisis.

The financial sector may be particularly vulnerable to shocks given the close interaction between banks, so shocks to some institutions may propagate through the banking network. In the literature to date, most papers have been concerned with actual credit or derivatives signed between banks. In this paper, we explore how Covid-19 impacts the banking industry using revenue and stock price as the indicators.

In addition, should recovery not be rapid, there is a risk of a bankruptcy wave. This means that they are effectively bankrupt due to policy forbearance. In other crises (such as Japan's 'Great Recession' in the 1990s), such problems suppressed a robust recovery for decades. Finally, in a worst-case economic scenario, the economic recession in the region could be deeper or more exceptionally prolonged than expected, and possibly result in a systematic collapse in many of the country's banking systems.

5.1.1 -> Bank

Bank is a financial institution licensed to receive deposits and make loans. It acts as an intermediary between depositors and borrowers with taking in funds, called deposits as its primary role. It will then use these deposits to create money in the economy by making loans such as home mortgages, business loans, and car loans. This process enables the banks to derive a profit by charging higher interest rates on the loans than they pay for deposits. There are several types of banks which may provide various financial services ranging from offering safe deposit boxes and currency exchange to retirement and wealth management. In most countries, banks are regulated by the national government or central bank. While in Malaysia, the banking system that comprises commercial banks, investment banks, and Islamic banks is the main source of financing to support economic activities.

Banks provide a secure place to deposit money. In addition, banks will pay savers a small percent of the deposited amount based on an interest rate. It may be very low for current accounts but the interest rate can be significant for saving accounts. Especially during the period of inflation, interest rates on deposits help to maintain the real value of the savings to avoid the money from losing value against inflation.

Furthermore, banks lend money to firms, customers and homebuyers in both upturns and downturns. The bank lending varies from unsecured personal loans to secured mortgage lending. Higher interest rates may be charged for unsecured lending due to the risk factor. Unsecured lending includes personal loan and business loan. Personal loans are issued for a huge purchase such as a car or funding a career or educational improvement while business loans are used for investment and business expansion. On the other hand, secured mortgage lending tends to be at a lower rate as the loan is secured against the value of the house, so it can be over 30 years or more. Moreover, a bank can also accept an overdraft with customers. Customers can borrow money in the short term immediately despite the amount allowed mostly being quite small.

Not only that, banks offer others features to consumers as well. For example, it enables instant access to cash, provides advice on financial matters, offers several ways to make international payments and allows electronic transfer of money. Customers can arrange travel insurance through banks too.

Healthy banking systems are essential for a well-functioning financial system to a modern economy. The security and well-being of banks are the top priority to ensure the economic activities remain robust. During upturn, banks are important to ensure the capital market is flourishing with capital to allow for economic expansion. While in downturn, banks are important to ensure the capital market is reserved with adequate loans for liquidity of the market. If a bank fails to perform these tasks, the consequences for the entire economy could quickly become so wide-reaching that the economy may collapse through a series of ripple effects. Therefore, strict regulatory requirements are in force to ensure the banks meet their current payment obligations.

5.1.2 -> Impact of the pandemic on the bank, and the impact to the country and the people

The impact of the pandemic towards banks can be examined using different segments of the banking operation. Among all, the most significant consequence would be the surging of the non-performing loan. Increased defaults resulting from curtailed economic activities, lower recoveries, higher credit exposures

and credit rating downgrade of customers in heavily affected industries. Subsequently, The expected reduction in the value of the loan portfolio from impairment reasons may impact the banks' ability to meet certain ratios such as Capital Adequacy Ratio (CAR), Loan to Deposit Ratio (LDR), Non Performing Loan (NPL) ratios. Banks can expect a large migration of loans to stage 2 and the accompanying sharp increase in reduced cash inflows from loan repayment may impact banks' liquidity position, increased cash withdrawals by depositors to meet their own funding needs and flight to quality as customers may move their funds from banks expected to be significantly impacted by the pandemic to less impacted banks. Besides, the combined effect of low business activities, higher impairment and possible operational and fair value losses may result in reduced profit levels and capital depletion. Capital adequacy may drop below the regulatory threshold as a result of increase in credit exposure.

At the same time, the pandemic has a substantial impact on the strategy, business and operating model of the banking industry. Covid-19 has presented an unprecedented live test of the bank's ability to work in a completely different manner from home, shifting more services online. Concerns around continuity of processes which have direct touchpoints with the customer but are still manually operated in bank branches that have been shutdown. Straight-through processing will become an imperative given the remote working conditions.

As the digital transformation programme arises, the pandemic should act as the catalyst to drive the rapid growth of digital interactions and transactions between the people as physical meeting is restricted by policies. Even though the digital transaction during the pandemic has increased significantly as most of the consumers opt for online transactions, the aftermath effect of this new norm remained unpredictable once the economy resumed opening. The surge in demand of digital assets will be a challenge for banks as banks will need to innovate to provide solutions to the clients within a short period of time. The transformation will benefit banks which are more mature and progressive in digital banking. However, there exists cyber risk as adoption of digital channels and remote working have expanded all of a sudden. Cyber criminals may seek to exploit any failures in the maintenance of IT systems during the pandemic, banks more than ever need to assess whether they are in a position to guarantee at a minimum the technical and application security measures to ensure as strong IT environment as possible, and let's not forget the remote access facilities of the staff. At the early phase of the pandemic, banks have cautioned about attacks of banks IT networks with cyber threats that are trying to exploit any remote access with new techniques. It is extremely important to be specifically aware of the cyber threat to avoid leaking important information of the banks and the customers.

On the other hand, the virus outbreak has caused most of the countries declared lock down mode or movement restriction control. Travel restriction and closure of all airports inhibits tourism and travel activities, hurting SMEs in tourism business. While the emerging leisure and entertainment sectors are cancelled due to movement restrictions. Globally, air traffic has been significantly reduced as compared to last year. Almost all the flights are cancelled which account for 70% decline of air traffic compared to last year. As for ground transportation, the demand reached a very low level with the strict curfew in major cities and intercity movement restriction as the usage of e-hailing apps has declined significantly. With the exception of the segments of retail stores that sell necessary products like groceries, hygiene kits and healthcare products, trade is heavily impacted due to decreased spending due to closure of the shopping

facilities and movement restriction. In this difficult time, most of the countries have formulated drastic monetary and fiscal policies to assist its people. The government and central banks announced various fiscal stimulus, that includes buying bonds, fiscal aids, deferment of loan schedule, lowering interest rate etc in the hope to prevent collapse of the economy.

If banks are under stress, the credit crunch will affect small and medium-sized enterprises (SMEs) and microbusinesses badly – reducing low-income household's ability to maintain livelihoods including in the informal sector. Policy measures to address this have included reducing capital ratios and interest rates, which may help to balance financial stability and finance for growth – although their effectiveness will also be subject to the bank's risk appetite. Meanwhile, the slowing of the economic activities signify that the immature business will be facing the risk of closure due to insufficient reserve for the payment of fixed cost. That will lead to dismissal of employees and hiring freeze. Without stable income and adequate savings, most of the loans will be progressed to stage 2 or even stage 3. The failure of repaying the loans will have a negative impact on borrowers such as lower creditworthiness, bankruptcy, seizure of mortgage and even closure of their own business due to insolvency.

5.2 Technical LR

ML is defined as a machine that learns whenever it changes its structure, program, or data in such a manner that its expected future performance improves (Nilsson, 2005). Therefore, the accuracy of the ML model can be improved by continuing collecting new data and feed into the model. In addition, J.D. Kelleher and Tierney, (2018) define ML as a field of computer science that focuses on developing and evaluating algorithms that can extract useful patterns from datasets. ML can be divided into four categories namely supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Kotsiantis, Zaharakis, and Pintelas, 2006; SAS Institute Inc., 2020). However, the widely adopted machine learning methods are supervised and unsupervised learning due to their existence longer than reinforced and semi-supervised learning (SAS Institute Inc., 2020). Supervised learning is a form of ML technique in which the goal is to learn a function that maps from a set of input attribute values for an instance to an estimate of the missing value for the target attribute of the same instance (J.D. Kelleher and Tierney, 2018). Supervised learning is the most popular machine learning technique that is being used in the industry due the outcome of the model is easier to be interpreted than unsupervised learning. On the other hand, unsupervised machine learning is the opposite of supervised learning where it draws inferences from datasets without the labeled data (The MathWorks Inc., 2020). Hence, the output from the unsupervised learning is difficult to be interpreted by humans, therefore, normally the result of the unsupervised learning will be used as the input to train the supervised learning, which is known as semi-supervised learning. Natural language processing is part of machine learning. The details of natural language processing will be discussed in the subsection 2.3.2. In this report, both supervised and unsupervised learning techniques will be explored to perform natural language processing to classify the topic of the documentation.

5.2.1 ML Algorithm

Throughout this pandemic, multiple works have been done to keep track of the coronavirus pandemic. Many scholars have developed a number of predicting methods for the trend forecasting of COVID-19, in

some severe countries and global, debating about mathematical models, infectious disease models, and artificial intelligence models. The models based on mathematical statistics, machine learning and deep learning have been applied to the prediction of time series of epidemic development. There is an urgent need for innovative solutions to monitor, develop and manage big data on the growing networks of the infected subjects, patient details, their community movements, and integrate with clinical trials and pharmaceutical, genomic and public health data. Machine learning technique offers the way to efficiently analyse the growth of infection with community behaviour through text messages, online communications and social media. Many studies have suggested that the COVID-19 spread follows exponential distribution. Following the evaluation of previous datasets on SARS virus pandemic, many studies have shown that data corresponding to new cases with time has a large number of outliers and may or may not follow a standard distribution like Gaussian or Exponential. In recent study by Data-Driven Innovation Laboratory, Singapore University of Technology and Design (SUTD), the regression curves were drawn using the Susceptible-Infected-Recovered model and Gaussian distribution was deployed to estimate the number of cases with time. However, in the previously reported studies on the earlier version of the virus, namely SARA-CoV-1, the data was reported to follow Generalized Inverse Weibull (GIW) Distribution better than Gaussian. This fits the following function to the data:

$$f(x) = k \cdot \gamma \cdot \alpha^{\beta} \cdot \beta \cdot x^{(-1-\beta)} \cdot exp\,(-\gamma(\alpha/x)^{\beta}$$

In the studies, f(x) denotes the number of cases with x, where $x > 0$ is the time in number of days from the first case, and $\alpha, \beta, \gamma > 0, \in R$ are parameters of the model. Now, the appropriate values of the parameters α, β and γ can be found to minimize the error between the predicted cases ($y = f(x)$) and the actual cases ($\hat{y}$). This can be done using the popular Machine Learning technique of Levenberg-Marquardt (LM) for curve fitting. However, as various sources have suggested, in the initial stages of COVID-19 the data has many outliers and noise. This makes it hard to accurately predict the number of cases. Thus, an iterative weighting strategy is proposed and called fitting technique as "Robust Fitting".

The main idea is maintaining weights for all data points (i) in every iteration (n, starting from 0) as w_i^n. A curve is fixed using the LM technique with weights of all data points as 1, thus $w_i^0 = 1\ \forall$ i. Then, the weight is computed corresponding to every point for the next iteration w_i^{n+1} as:

$$w_i^{n+1} = \frac{\frac{exp\,(1-(d\_i\hat{}n - tanh\,(d\_i\hat{}n))}{(max_i\, d_i^n - tanh\,(d_i^n))}}{\frac{(\Sigma_i\, exp\,(1-(d_i^n - tanh\,(d_i^n))}{(max_i\, d_i^n - tanh\,(d_i^n))}}$$

Simply, in the above equation, tanhshrink function is defined as tanhshrink = x - tanh(x) for the distances of all points along the y axis from the curve (di). This is to have a higher value for points far from the curve and near 0 value for closer points. This is then standardized by dividing with max value over all points and subtracted from 1 to get a weight corresponding to each point. This weight is then standardized using softmax function so that the sum of all weights is 1. The curve is fit again using the LM method, now with the new weights w_i^{n+1}. The algorithm converges when the sum total deviation of all weights becomes lower than a threshold value.

To find the best fitting distribution model for the data corresponding to COVID-19, the data on daily new confirmed COVID cases are retrieved and studied. Five sets of global data on daily new COVID-19 cases were used to fit parameters of different types of distributions. Finally, the best performing 5 distributions are identified.

While there are a lot of existing findings concerning the pandemic using different models, the impact towards the stock price attracts the attention of the researchers too. Predicting stock price has been considered as an audacious goal of many as the stock market is unpredictable, dynamic and volatile at times. It is a challenging task as it depends on various factors including political affection, economic indicators, investors sentiment, speculation, company earnings etc. Most of the previous work use classical algorithms like linear regression, Random Walk Theory (RWT), Moving Average Convergence and Divergence (MACD) and also linear model like Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) for predicting stock price. Studies have shown that predicting stock price using Machine Learning can be more efficient and accurate if the right training model and inputs are applied. Some techniques based on neural networks such as Artificial Neural Network (ANN), Convolutional NEURAL Network (CNN), Recurrent Neural Network (RNN) and deep neural networks like Long Short Term Memory (LSTM) also have shown promising results.

ANN is capable of finding hidden features through a self-learning process. These are good approximators and are able to find the input and output relationship of a very large complex dataset. Thus, ANN proves to be a good choice for predicting stock price for an organization. Some works have been proposed which use Random Forest (RF) for forecasting purposes. RF is an ensemble technique. It is normally capable of performing both regression and classification tasks. It operates by constructing multiple decision trees at training time which outputs mean regression of individual decision trees. Two techniques i.e. ANN and RF have been used for predicting the closing price of an organization. The models use a set of new variables created using the financial dataset with Open, High, Low and Close of a particular company. These new indicators will play a crucial role in terms of improved accuracy of the models in predicting the next day closing price of a particular company.

The historical data for the companies to be examined have been collected. Six new variables have been created for the prediction of stock closing price. These variables have been used to train the model. The new variables are as follows:
1. Stock High minus Low price (H-L)
2. Stock Close minus Open price (O-C)
3. Stock price's seven days' moving average (7 DAYS MA)
4. Stock price's fourteen days' moving average (14 DAYS MA)
5. Stock price's twenty one days' moving average (21 DAYS MA)
6. Stock price's standard deviation for the past seven days (7 DAYS STD DEV)

ANN, is one of the intelligent data mining techniques that identify a fundamental trend from data and to generalize from it. ANN is capable of simulating and analysing complex patterns in unstructured data as

compared to most of the conventional methods. The model uses the basic structure of the Neural Network having neurons with different layers. The model works with three layers. It consists of an input layer, hidden layer and the output layer. The input layer consists of new variables which are H-L, O-C, and 7 DAYS MA, 14 DAYS MA, 21 DAYS MA, 7 DAYS STD DEV and Volume. The weights on each input load is multiplied and added and sent to the neurons. The hidden layer or the activation layer consists of these neurons. The total weight is calculated and is moved to the third layer which is the output layer. The output layer consists of only one neuron which will give the predicted value in terms of closing price of the stock. The Fig.1 shows a detailed representation of ANN architecture with the new variables acting as input. Fig. 1. Detailed architecture of Artificial Neural Network (ANN) for stock price prediction

Besides, Random Forest Random Forest (RF) is an ensemble machine learning technique. It is capable of performing both regression and classification tasks. The idea is to combine multiple decision trees in order to determine the final output instead of relying on individual decision trees which in order reduce the variance in the model. In this work, new created variables are provided for the training of each decision tree which in turn determines the decision at the nodes of the tree. The noise in stock market data is usually quite high because of its huge size and can cause the trees to grow in a completely different manner as compared to the expected growth. It aims at minimizing forecasting error by treating the stock market analysis as a classification problem and based on training variables predicted the next day closing price of the stock for a particular company

Correlation research aims at calculating and understanding the impact of a linear or nonlinear relationship between two continuous variables. Coefficients of association assume values ranging from negative correlations (-1) to uncorrelated (0) to positive correlations (+1). The sign of the coefficient of correlation (i.e, positive or negative) determines the direction of relation. The ab-solute value shows the strength of the linear relationship (Tables 2) which is very close to +1.

	Positive	*Recovered*	*Deceased*	*Active*
Positive	1			
Recovered	0.985837589	1		
Deceased	0.988199348	0.950408963	1	
Active	0.988313465	0.948769793	0.9985863	1

Initially the data cleaning process is performed on the two datasets to remove any missing values. Then a correlation analysis is performed on the data sets using Python programming through Spyder of Anaconda Navigator App. Then Linear regression model is used to evaluate the relative impact of active cases due to daily positive cases in Odisha as well as in case of India data. The key goal of linear regression is to fit a straight line with the data forecasts Y for X where Y is the total number of daily active cases and X is the total number of positive cases. The least squares method is commonly used to estimate the intercept and slope regression parameters which define the line. The below Fig.1 and 2 shows the average peak values of active cases in part of Odisha as well as India. The model can be expressed as in Eq.(1) where Y and X are

dependent and independent variable, α is the intercept and b is the regression parameter as slope and ε is the random error respectively.

$$Y = \alpha + bX + \varepsilon$$

The limitations of Linear regression are that it often explores a relation between the mean of the input variables and output variables. Just as the mean is not a full description of a single variable, linear regression is just not a clear understanding of variable relationships. Therefore, an analysis of the various factors is done using Multiple Linear Regression (MLR) models. The dependent variable (target variable) is dependent on many independent variables, in this case. You can describe a regression equation involving multiple variables as:

$$Y = \beta + \beta_0 x_1 + \beta_1 x_2 + \beta_2 x_3$$

Where Y is the predictor or target variable and $x_1; x_2; x_3$ are the independent variables. β is the y-intercept and$\beta_0, \beta_1, \beta_2$ and ε are the coefficients and error terms respectively.

Besides, there are other papers focused on using models to identify the infected people so that prevention of spread can be taken. The support vector machine classifies the corona affected X-ray images from others using the deep feature. The methodology is beneficial for the medical practitioner for diagnosis of coronavirus infected patients. Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

Deep feature extraction is based on the extraction of features acquired from a pre-trained Convolutional neural network (CNN). The deep features are extracted from fully connected layers and fed to the classifier for training purposes. The deep features obtained from each CNN network are used by SVM classifiers. After that, the classification is performed, and the performance of all classification models are measured. The rice leaf disease identification model based on deep features by SVM classifier is shown in Figure 2.

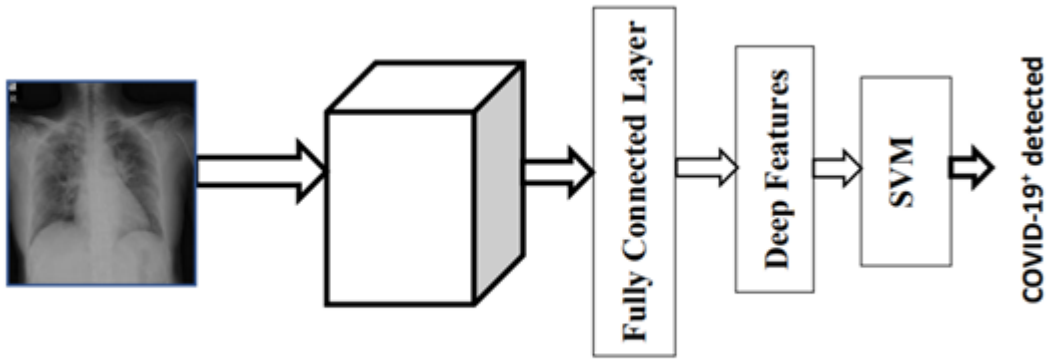


Figure 2. Detection Corona Virus by SVM based on Deep Feature using X-ray images.

The deep features of CNN models are extracted from a particular layer and feature vector is obtained. The features are fed to the SVM classifier for rice disease identification purposes. The feature layer and feature vector are detailed in Table 3.

Table 3. Details of feature layer and feature vector of CNN models.

CNN models	Feature Layer	Feature Vector	CNN models	Feature Layer	Feature Vector
AlexNet	fc6	4096	Xception	predictions	1000
			Resnet18	Fc1000	1000
			Resnet50	Fc1000	1000
Vgg16	fc6	4096	Resnet101	Fc1000	1000
			Inceptionv3	predictions	1000
			Inceptionresnetv2	predictions	1000
Vgg19	fc6	4096	GoogleNet	loss3-classifier	1000
			Densenet201	Fc1000	1000

The classes SGDClassifier and SGDRegressor provide functionality to fit linear models for classification and regression using different (convex) loss functions and different penalties. E.g., with loss="log", SGDClassifier fits a logistic regression model, while with loss="hinge" it fits a linear support vector machine (SVM).

The optimizer is a robust algorithm that helps to reduce the loss of a deep neural system by changing some attributes such as learning rate and changing weight and enhance the overall performance of the system. An optimizer can improve the performance of a neural system. It is essential to use an optimizer to reduce loss functions. In the existing research, they used two extensively used optimizers such as Stochastic Gradient Descent. They used Stochastic Gradient Descent (SGD) in the DenseNet121 and Base CNN model. Stochastic Gradient Descent is a widely utilized optimizer, much of the time, it is utilized in traditional CNN models to streamline. It is an updated form of Batch SGD. SGD gets rid of this repetition by performing each update in turn. It is subsequently generally a lot quicker and can likewise be utilized to learn on the web. SGD performs visit refreshes with a high change that prompt the target capacity to change intensely.

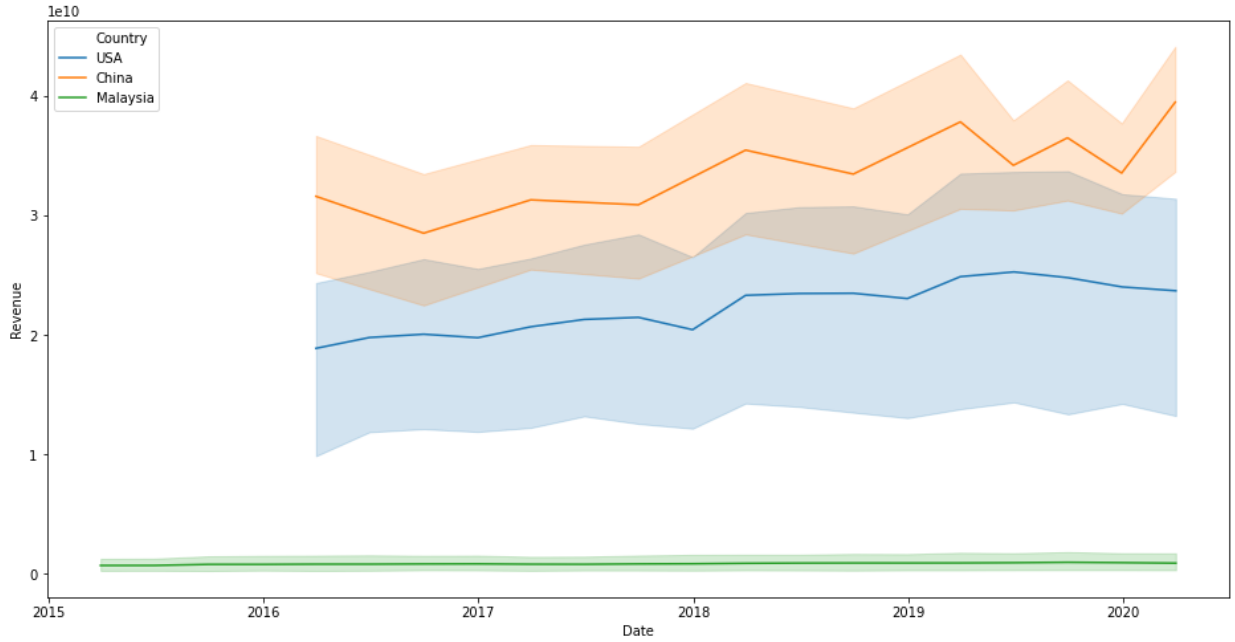
5.3 Summary(Summary of the LR)

Two implications follow for the future of banking. First, banks will operate in a financial system that is awash with liquidity and interest rates are extremely low. Second, the government will be a key player in the financial sector, both as a borrower (to fund its deficit) and as a “risk absorber” providing guarantees, back-stops and more direct fiscal support for borrowers whose businesses and cash flows bear the brunt of the virus. This brings us to the related issue of how banks' loan books are likely to look in the future since

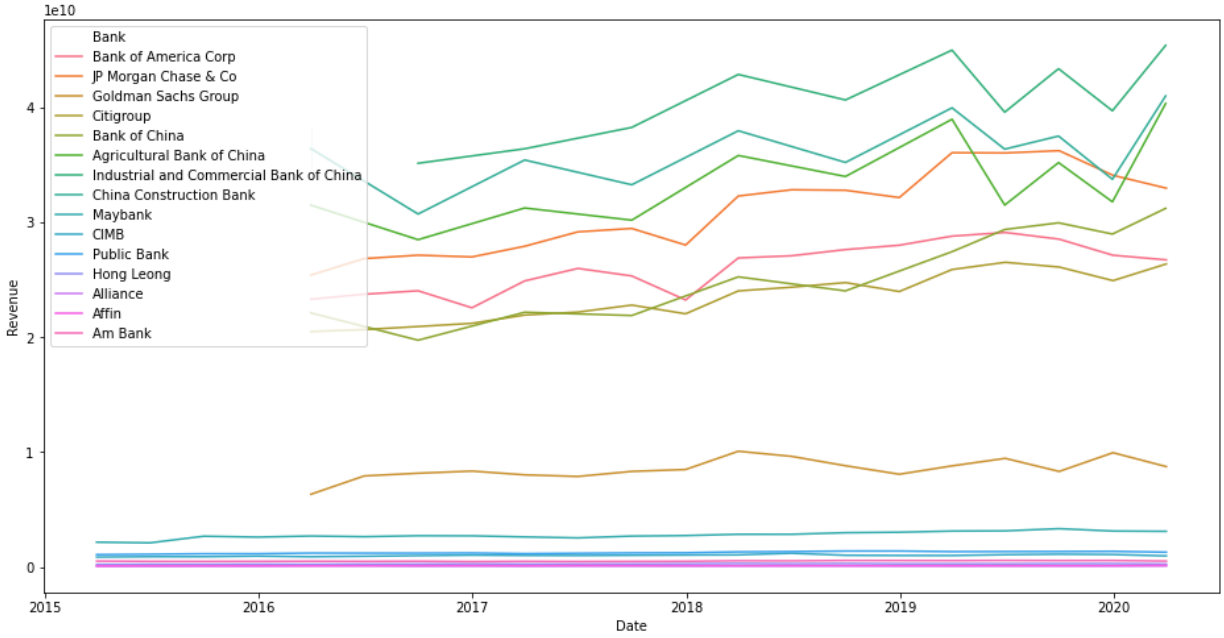
major economic upheavals invariably lead to an escalation in risk perception and a flight to quality. This means that banks will prefer to give loans to borrowers whose cash flows are visible and strong, while avoiding borrowers whose cash flows and incomes run the risk of being disrupted. A thumb rule that banks often follow is that size matters. Bigger companies on average are less likely to default than smaller ones and the flight to quality could translate into flight to size. Banks remain highly risk averse and the consensus among industry leaders is that most companies in consumer-oriented sectors at the moment are now operating with less than 70 percent of their capacity. The banking sector's health depends on how soon the economy recovers. All eyes are now on the government's upcoming fiscal stimulus package. As banks grapple with the many challenges posed by the COVID-19 crisis it becomes clear that, whatever the eventual outcome, they will learn many valuable lessons about their customers, their own capabilities, and the market as a whole.

6 MAIN BODY OF ARTICLE

6.1 Revenue Trends



Due to the market volume, we can't really compare the revenue in Malaysia to those in the USA and China. As shown in the graph, China seems to have a higher revenue compared to the USA. In addition, the revenue in China has an upward trend during the Covid-19 pandemic. This is because China has successfully controlled the spread of Covid-19 and flattened the curve. Hence, the banks in China were able to sustain under the new norm and continue growing their businesses. On the other hand, the daily Covid-19 cases have increased dramatically and the pandemic is not well-controlled in the USA. Therefore, the graph has shown that the revenue in the USA has slightly dropped during the pandemic. As for Malaysia, although the pandemic is undercontrolled, the revenue also seems to be the same and did not increase as much as China. This is because of the sharp slowdown in economic activity due to the Covid-19 and the slow recovery rate in Malaysia. Another possible reason is that Malaysia is still a developing country while China is a developed country with a strong economy and rapid productivity growth. In conclusion, only China has shown a growth in revenue during the pandemic while the revenue growth in Malaysia remains the same and the USA shows a slightly drop.



As shown in the line graph above,all 4 banks show an upward trend during the pandemic in China. There are Bank of China, Agricultural Bank of China, Industrial and Commercial Bank of China and China Construction Bank. However in the USA, the revenue of Citigroup and JP Morgan Chase & Co has decreased whereas the revenue of the Bank of America Corp and Goldman Sachs Group has increased.

6.2 Revenue Prediction

Root Mean Squared Error (RMSE) is an absolute measure of fit and it is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data. In other words, how close the observed data points are to the model's predicted values. Hence, lower values of RMSE indicates better fit.

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts, without considering their direction. All the individual differences are weighted equally in the average. It is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation.

R-squared is a goodness-of-fit measure for linear regression models. It measures the strength of the relationship between the model and the dependent variable on a 0-100% scale.

6.2.1 Revenue

Country	RMSE	MAE	R squared
USA	0	0	1

China	0	0	1
Malaysia	0	0	1

From the table above, the USA, China and Malaysia shared the value for RMSE, MAE and R squared, which are 0,0,1 respectively. In this case, RMSE=MAE, which means that all the errors are of the same magnitude. In addition, MAE=0 also shows that the model gives a high accuracy on the predicted value.

7 CONCLUSION

In conclusion, we can observe that in China where the pandemic is well-controlled, banks are showing an increase in their revenues. In addition, Malaysia as the country where the pandemic is undercontrolled, the revenue of banks remains the same as the past years. However in the USA, some of the banks in our research have experienced a decrease in their revenue because of the poor control level of the pandemic. In a nutshell, we can conclude that the level of negative impact on the banking industry can be classified by the control level of the pandemic. Well-controlled countries tend to have a lower negative impact on the banking industry than those undercontrolled countries during the Covid-19 pandemic. However, since there are still active cases around the world and all businesses are adapting under the new norm, it may need 2-3 more years to fully investigate the real impact of the Covid-19 to the banking industry around the world.

8 APPENDICES

```
#trend
from google.colab import files
uploaded = files.upload()

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("Revenue.csv")
df.head()
df.dtypes
df["Date"]=pd.to_datetime(df["Date"])
df.dtypes
plt.figure(figsize=(16,8))
ax = sns.lineplot(x="Date", y="Revenue", hue="Bank", data=df)
plt.figure(figsize=(16,8))
ax = sns.lineplot(x="Date", y="Revenue", hue="Country", data=df)


#modelling
from google.colab import files
uploaded = files.upload()

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df=pd.read_csv("Revenue.csv")
df.head(10)
df.dtypes
df['Date']=pd.to_datetime(df["Date"])
df.dtypes
df_us=df.loc[df['Country']=="USA"]
df_us.head()
df_us.drop(["Country","Bank"],axis=1,inplace=True)
df_us.head()
df_us.nunique
df_us.sort_values(by='Date')

x_us=df_us.loc[df_us['Date']<'2019-03']
y_us=df_us.loc[df_us['Date']>='2019-03']
```

```
x_us_train=x_us.drop('Date',axis=1)
y_us_train=x_us.Revenue
x_us_test=y_us.drop('Date',axis=1)
y_us_test=y_us.Revenue

#checking the shape of the train and test data
x_us_train.shape,y_us_train.shape,x_us_test.shape,y_us_test.shape
from sklearn.preprocessing import StandardScaler

#initiating standard scaler
scaler_us=StandardScaler()

#fit the scaler in training features
scaler_us.fit(x_us_train)

#Rescale both sets using the trained scaler
x_us_train_scaler=scaler_us.transform(x_us_train)
x_us_test_scaler=scaler_us.transform(x_us_test)

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
lin_us = LinearRegression()

lin_us.fit(x_us_train_scaler, y_us_train)
predictions_lin_us = lin_us.predict(x_us_test_scaler)

print('RMSE: {0:.3f}'.format(mean_squared_error(y_us_test, predictions_lin_us)**0.5))
print('MAE: {0:.3f}'.format(mean_absolute_error(y_us_test, predictions_lin_us)))
print('R^2: {0:.3f}'.format(r2_score(y_us_test, predictions_lin_us)))

df_china=df.loc[df['Country']=="China"]
df_china.head()
df_china.drop(["Country","Bank"],axis=1,inplace=True)
df_china.head()
df_china.nunique
df_china.sort_values(by='Date')

x_china=df_china.loc[df_china['Date']<'2019-03']
y_china=df_china.loc[df_china['Date']>='2019-03']

x_china_train=x_china.drop('Date',axis=1)
```

```
y_china_train=x_china.Revenue
x_china_test=y_china.drop('Date',axis=1)
y_china_test=y_china.Revenue

#checking the shape of the train and test data
x_china_train.shape,y_china_train.shape,x_china_test.shape,y_china_test.shape

from sklearn.preprocessing import StandardScaler

#initiating standard scaler
scaler_china=StandardScaler()

#fit the scaler in training features
scaler_china.fit(x_china_train)

#Rescale both sets using the trained scaler
x_china_train_scaler=scaler_china.transform(x_china_train)
x_china_test_scaler=scaler_china.transform(x_china_test)

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
lin_china = LinearRegression()

lin_china.fit(x_china_train_scaler, y_china_train)
predictions_lin_china = lin_china.predict(x_china_test_scaler)

print('RMSE: {0:.3f}'.format(mean_squared_error(y_china_test, predictions_lin_china)**0.5))
print('MAE: {0:.3f}'.format(mean_absolute_error(y_china_test, predictions_lin_china)))
print('R^2: {0:.3f}'.format(r2_score(y_china_test, predictions_lin_china)))

df_my=df.loc[df['Country']=="Malaysia"]
df_my.head()
df_my.drop(["Country","Bank"],axis=1,inplace=True)
df_my.head()
df_my.nunique
df_my.sort_values(by='Date')

x_my=df_my.loc[df_my['Date']<'2019-03']
y_my=df_my.loc[df_my['Date']>='2019-03']

x_my_train=x_my.drop('Date',axis=1)
y_my_train=x_my.Revenue
```

```
x_my_test=y_my.drop('Date',axis=1)
y_my_test=y_my.Revenue

#checking the shape of the train and test data
x_my_train.shape,y_my_train.shape,x_my_test.shape,y_my_test.shape
from sklearn.preprocessing import StandardScaler

#initiating standard scaler
scaler_my=StandardScaler()

#fit the scaler in training features
scaler_my.fit(x_my_train)

#Rescale both sets using the trained scaler
x_my_train_scaler=scaler_my.transform(x_my_train)
x_my_test_scaler=scaler_my.transform(x_my_test)

from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
lin_my = LinearRegression()

lin_my.fit(x_my_train_scaler, y_my_train)
predictions_lin_my = lin_my.predict(x_my_test_scaler)

print('RMSE: {0:.3f}'.format(mean_squared_error(y_my_test, predictions_lin_my)**0.5))
print('MAE: {0:.3f}'.format(mean_absolute_error(y_my_test, predictions_lin_my)))
print('R^2: {0:.3f}'.format(r2_score(y_my_test, predictions_lin_my)))
```

9 GLOSSARY

CRISP-DM: Cross-industry Process for Data Mining

SVM: Support Vector Machine

RMSE: Root Mean Squared Error

MAE: Mean Absolute Error

10 REFERENCES

Accuracy, M., 2020. Machine Learning Model Accuracy | Model Accuracy Definition. [online] DataRobot. Available at: https://www.datarobot.com/wiki/accuracy/.

Alley, G., 2018. What is data transformation? Available at: https://www.alooma.com/blog/what-is-data-transformation.

Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D., and Cochran, J.J., 2017. Modern business statistics with microsoft office excel (with xlstat education edition printed access card). Cengage Learning.

cs.ccsu.edu,2020. Data preprocessing [Online]. Available from: https://cs.ccsu.edu/~markov/ccsu_courses/DataMining-3.html.

GeeksforGeeks. 2020. Data Preprocessing For Machine Learning In Python - Geeksforgeeks. [online] Available at: https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/.

GeeksForGeeks, 2019. Data transformation in data mining. Available at: https://www.geeksforgeeks.org/data-transformation-in-data-mining/.

IBM,2020. Data aggregation [Online]. Available from:https://www.ibm.com/support/knowledgecenter/en/SSBNJ7_1.4.2/dataView/Concepts/ctnpm_dv_use_ data_aggreg.html.

Investopedia. 2020. R-Squared. [online] Available at: https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20(R2),variables%20in%20a%20regression%20model..

Jin, R., Breitbart, Y., and Muoh, C., 2009. Data discretization unification. Knowledge and information systems,19(1),p.1.

KPMG. 2020. COVID-19: Impact On The Banking Sector. [online] Available at: <https://home.kpmg/xx/en/home/insights/2020/07/covid-19-impact-on-banking-m-and-a-2020.html#:~:text=depressed%20banks'%20valuation%E2%80%A6-,COVID%2D19%20has%20generated%20significant%20instability%20and%20high%20volatility%20in,x%20on%2030%20April%202020>.

Limited, B., 2020. COVID Crisis Sinks Global Economy In 2020, Collapsing GDP 4.9%: IMF. [online] https://www.bangkokpost.com. Available at: <https://www.bangkokpost.com/world/1940428/covid-crisis-sinks-global-economy-in-2020-collapsing-gdp-4-9-imf>.

Medium. 2020. What Is Training Data Its Types And Why It Is Important?. [online] Available at: https://becominghuman.ai/what-is-training-data-its-types-and-why-it-is-important-f998424c3c9.

Ori.hhs.gov. 2020. Data Collection. Available at: https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html#

Patel, P.C., 2019. Data transformation in data mining [Online]. Available from: https://www.geeksforgeeks.org/data-transformation-in-data-mining/.

Patro,S. and Sahu,K.K., 2015. Normalization: a preprocessing stage. Arxiv preprint arxiv:1503.06462.

Smart Vision Europe. 2020. Crisp DM Methodology - Smart Vision Europe. [online] Available from: https://www.sv-europe.com/crisp-dm-methodology/.

Statistics How To. 2020. Absolute Error & Mean Absolute Error (MAE) - Statistics How To. [online] Available at: https://www.statisticshowto.com/absolute-error/.

Statistics How To. 2020. RMSE: Root Mean Square Error - Statistics How To. [online] Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviation%20of,the%20line%20of%20best%20fit
Tcs.com. 2020. [online] Available at: <https://www.tcs.com/content/dam/tcs/pdf/Industries/Banking%20and%20Financial%20Services/COVID-19-Crisis-Implications-for-financial-services-industry.pdf>.

Yaghini, 2009. Data mining: part 2. data preprocessing [Online]. Available from: http://webpages.iust.ac.ir/yaghini/Courses/Data_Mining_881/DM_02_04_ Data%20Transformation.pdf.