Investment advice based on market trends and financial distress of the company

Submitted in partial fulfilment of the requirements of the examination for MSc. Data Analytics, Dublin Business School, August 2020.

By:
Shaikh Sufiyan Ahmed

Period of completion: 09.06.2020 - 25.09.2020

Supervisor: Prof. Richard O'Callaghan

Dublin, 25.09.2020

DECLARATION

I hereby declare that apart from the explicit references made by others to the craft, the substance to the thesis is unique and has not been submitted in entire or to some extent in other colleges and some other degree. This project is my very own and contains nothing which is the result of work done in a group or a joint effort with others.

Student Name: Shaikh Sufiyan Ahmed

Student Number: 10530990

Date: 25.09.2020

Acknowledgment

I wish to express my deepest gratitude to my supervisor, Prof. Richard O'Callaghan for the guidance, remarks, and support through the learning process of my thesis. His, advice, insights, right from the beginning of my thesis, helped me very greatly in shaping what it is today. His constant support whenever I was stuck and constant encouragement whenever I made huge progress, kept me going steadily.

I would also like to thank Dr. Abhishek Kaushik, Dr. Shahram Azizi, Dr. Terri Hoare for their advice during the program and giving us all the necessary support and learning required to complete the MSc Data Analytics journey.

I would like to thank my parents, for being a constant source of love, support, and encouragement throughout this program. Forever indebted to them.

INDEX OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

FDP Financial distress prediction

SMOTE Synthetic Minority Over-sampling Technique

ADASYN Novel Adaptive Synthetic

LSTM Long Short Term Memory

RNN Recurrent Neural Network

SVM Support Vector Machine

LDA Linear Discriminate Analysis

RF Random Forest

XGBoost eXtreme Gradient Boosting

CRISP-DM Cross Industry Process for Data Mining

MLE Maximum Likelihood Estimation

ABSTRACT

In this day and age “Investment” has become a necessity and an important factor for companies and individuals. Investment is something which is called a monetary asset purchased with the idea that the investment will provide a profit in the future. Before investing people study Financial performance, Background and experience in the industry, Company uniqueness, Effective business model, Large market size of a particular company, but they do not focus on minute fingerprints of financial distress. The main of this research is to draw down the factors of investment under a single umbrella and generate and investment advice.

This study will mainly focus on developing two models to refine the investment process. The first model we have proposed is a stock market prediction based on deep learning techniques. Here we have used a Realtime dataset from Yahoo finance for a particular company where clients want to invest. Here we have used different deep learning techniques. But for this research sequential LSTM has outperformed all the models with minimum rmse (root mean squared error) score. With the help of this are able to predict the stock closing price of the company for up to one month. The second model will be of Financial Distress Predication based on various Bagging and Boosting techniques with the integration of various SMOTE techniques were used. But for this research Balanced Bagging Method with ADASYN has outperformed from all the models with an accuracy of 93%. ADASYN Adaptive Synthetic Sampling Method is a modified version of SMOTE which performed best with our bagging and boosting model, we have used ADASYN to deal with the class imbalance problem. This empirical research is carried out based on real world financial data of 3476 Chinees company with over 84 financial and non-financial features.

Keywords: Investment, Financial Distress Prediction, Stock Market Prediction, Deep learning, SMOTE, ADASYN, Bagging and Boosting, Ensemble Methods, Adaboost-SVM, LSTM.

Word Count: 10329

1 Introduction

1.1 Background

Investment is something which is called monetary asset purchased with the idea that the investment will provide a profit or capital appreciation in the future. Investments are mainly of four types growth Investment, stocks, mutual funds, and cryptocurrencies. Investments are all about future returns, and every investment comes with some degree of risk. We as humans, cannot completely dissipate the risk, but with the help of technology, we can minimize the risk as far as possible. In bygone days people used to take investment advice from many professionals like financial planners, bankers, and brokers which provide them investment advice based on the previous market trends and previous investment returns of that organization. As the technology has evolved, presently there are machine learning techniques available like, stock market prediction, which help the individual or a firm to make their decision of investments.

According to my research and understanding, the limitation I found in the investors and machine learning model is that they focus on the previous history of the stock's, previous trends of the stocks and other parameters, but they do not focus on minute fingerprints of financial distress in the data. However, due to the uncertainty of the business environment and strong competition, even companies with perfect operational mechanisms have the possibility of business failure and financial bankruptcy. For example, "COVID -19: The Pandemic" has created a huge loss to every growing sector from IT to Finance and many more. Predicting financial distress has become very important not only for the company but also for the investors and the customers. Enterprise has developed an initial control mechanism with the help of FDP (Financial distress prediction system), which will notify before the financial distress occurs. On the other hand, stock predication is also one of the major challenges. Till date, no technology is able to predict or forecast the actual stock price. With the advancement, with the help of machine learning and deep learning methods, researchers came very close to prediction and forecasting.

1.2 Research Project/Problem

In this research, we are trying to build two different models and drilled them down under one umbrella to refine the investment process and focus on every aspect one should consider before investing.

The first model will be, stock market perdition where we will be trying out different Deep learning models and evaluate their performance on a real-time dataset. Time series data set are difficult to predict due to its on linear and dynamic nature. So, we are applying advanced deep learning techniques to make the prediction and forecasting.

In the second model, we will be developing a Financial distress predictor using different bagging and boosting techniques with SMOTE. For Financial distress, many types of research came up with many different models. Some researchers worked on machine learning models and gained accuracy up 92% but the only limitation here was the research was performed on a small-scale dataset when they carried out this experiment on large dataset models accuracy fall to 30 – 40%. With the advancement of technology, researchers came up with hybrid models and started using multiple classifiers, which helped to boost the accuracy, performance, and stability of the model. The main issue these models faced that they were working on a balanced dataset. When this model got exposed to the class-imbalanced dataset the model underperformed. So, to deal with this issue we will be trying different SMOTE techniques to overcome the class-imbalance issue.

1.3 Research objectives

For the first model, we will be using a real-time dataset which will be obtained through yfinance API available in python. Through yfiannce API we can collect real-time stock price by using the company's ticker code. For predicting and forecasting the stock prices we will be using various deep learning models in which we will be comparing different LSTM models on real-time datasets and compare their rmse score. With the help of that, we will be predicting the future 30 days stock closing price of that particular company.

For the second model, I will apply various bagging and boosting techniques to improve the accuracy stability and performance of the model. I will be comparing the results with previously used machine learning and hybrid machine learning models and will compare and check their performance, accuracy, and stability of the model. Then I will test out the model's accuracy on a class-imbalanced dataset. If there is any major change in accuracy and f-score. We will try to overcome that issue with the help of SMOTE. Specifically, in smote we will be testing out our models on various SMOTE techniques and will note down the behavior of each model towards different balanced datasets generated by different SMOTE techniques.

1.4 Scope and Limitations

The process of investment can be improved by adding more scenarios that one look before investment under a single platform to ease the process of investment.

For Financial Distress prediction we have used an actual dataset of Chinese companies. To refine the FDP process, we can use the time ratio in the dataset that will help to make more accurate predictions.

For Stock market prediction we have only predicted the closing price of the company. If we consider the opening as well as the closing price of the company that will add more improvement in the investment process.

2 Literature Review and Related Work

2.1 Literature review on Financial Distress Prediction

Financial Distress is a financial term mainly used in corporate finance. Basically, Financial distress is a condition or situation where a company or individual cannot generate a certain amount of revenue or they are unable to pay their financial debt or obligations. Financial distress is the last step or a step before bankruptcy. Now a day's companies are taking major precautions towards Bankruptcy, to avoid this, Companies are implementing "Financial Distress Indicators" in their firm, which indicates the firm, that is the company currently running out of business or not, which will help them plan accordingly. Companies have advanced their technology and implemented Financial Distress Predictors in their firm. Financial distress predictor predicts that in future will the company will run out or business or not based on the historical data of the company. According to my research and study the early warnings of financial distress are as follows (a)Cash Flows, (b)Falling margins and poor profits, (c)Poor sales growth, (d)Defaulting on payments, (e)Difficulty in raising capital, etc. (Chong, 2020).

Back in the days when Financial distress prediction was introduced in the year 1966 the first model which was implemented was the "Beavers Univariant Model" which uses a single variable discriminant model. The research continued and presented a multivariant discriminant analysis model which was "Altman's Z-Score Model" in 1968 (Altman, 1968). The research was not stopped at this point in 1980 Ohlson's applied first Logistic Regression and created a "Logit Model" (Ohlson, 1980). These are the statistical models widely used to date. According to my research, the major advantage of using these statistical models is that it has a simple structure, it uses fewer parameters, less training time. (Jie Sun, 2019).

2.1.1 Financial distress prediction using machine learning techniques

As the technology was evolved AI and Machine Learning came into existence, Machine Learning models were implemented to predict the financial distress. According to the research the AI and Machine Learning models has outperformed as compared to the previous statistical

model because AI models can fit non-linear relationship between financial distress and financial features easily and are not bound in the restriction of statistical assumptions (Jie Sun, 2019). According to the analysis of Kyung-Shik Shin, SVM outperforms from all the machine learning classification models because it works well with data having high dimensional space (Noviyanti Santoso, 2018). This paper mainly focusses on the diagnosis of financial distress in the early stages before the bankruptcy. This model made predictions on industrial companies in Indonesia. This model is made up using two machine learning algorithms LDA (Linear discriminant analysis) and SVM (Support vector machine) and created a hybrid stepwise-SVM. In this research, it was concluded that the model accuracy can significantly increase the prediction accuracy.

2.1.2 Financial Distress Prediction using hybrid and multiple classifiers

With the advancement of technology, researchers are also focusing on hybrid models in which they combine different techniques with machine learning models to improve their accuracy and performance of the model. These hybrid models are basically an upgraded version of simple machine learning models. In the year 2017 Nada Mselm, first implemented the hybrid model, this hybrid model was a combination support vector machine with the integration of partial least square. In this research, Nada Mselmi has considered three best performing machine learning models like the Logit model, Artificial Neural Networks, Support Vector Machine, and compared their results with the hybrid model to predict the financial distress in the company. For this research, they have considered a French small and medium scale industrial dataset. the results of the research were for the next one fiscal year the basic machine learning models have given approx. 88.57% accuracy and on the other hand the hybrid model outperforms all the machine learning models with 94.28% accuracy (Mselmi, 2017). In this research Nada Mselmi, has used the strong ability of partial least square of handling multicollinearity among predictors and the strength of Support vector machine of capturing nonlinearity relationship amongst the different variables (Mselmi, 2017) and with the help this model demonstrated that hybrid model has outperformed from the other machine learning models. According to my analysis and research hybrid models outperforms from normal machine learning models in all scenarios and circumstances. Another research took place in 2019 which focuses on using multiple classifiers

such as bagging and boosting techniques these methods are also called “Ensemble Methods”. The ensemble is basically a machine learning concept which is trained using simple algorithms. Two ways are “Bagging” and “Bosting”. Bagging is mainly used to avoid variance and helps to deal with overfitting issues. To execute this bagging method generates additional data or information to train the model from the dataset with the help of representation to generate multiple sets of valid data. Boosting is mainly used for predicting something or predictive models. In the year 2019 Jie Sun, has created a boosting model Adaboost-SVM and compared it with normal machine learning models and it leads to the result that hybrid models are 10% more stable and accurate (Jie Sun, 2019).

2.1.3 Review on performance of models without using smote

Many pieces of research on Financial Distress Prediction are performed on a self-modified or pre-assembled dataset, which means the following dataset contains an equal number of positive and negative value this type of dataset is considered as a balanced dataset. When we consider a real time scenario, it is nearly impossible to get a balanced dataset. In classification methods, the dataset is usually divided into class majority class which contains the maximum number of positive or negative samples and minority class which contains a smaller number of samples. Many cases consider the smaller number of samples as their prediction, for example as we are considering a distressed company to predict the Financial Distress Prediction. Louzada, Ferreira-Silva, and Diniz, has proposed research. In this research, they have considered a real time Brazilian bank dataset. They applied to best performing models’ naive logistic regression and logistic regression with state-dependent sample selection applied to simulated data. Which resulted that when performed with imbalanced dataset gives the same result and there no significant change in the performance and accuracy, but when considering the same dataset with balanced samples the results were superior and there was a huge change in the stability, accuracy, and performance in the models, where native logistic regression outperforms as compared to logistic regression with state-dependent sample selection when performed with the balanced dataset (Francisco Louzada, 2012). According to my research and understanding, while considering a financial case or considering the Financial distress prediction when applied using

machine learning models balanced data has an apparent impact on results because the imbalanced data make the model biased while training. So, with sensitive issues like finance, we should always consider a balanced dataset. As it is very difficult to get a complete balanced dataset there is the best solution to use SMOTE - Synthetic Minority Over-sampling Technique, this technique is widely used to deal with imbalanced data.

2.1.4 Financial Distress prediction using SMOTE technique

There are various methods and techniques to deal with an imbalanced dataset like Resample the training set, Use the right evaluation metrics, Use K-fold Cross-Validation in the right way. The best method amongst this is to Resample the training set. There are two different techniques to convert the dataset using resampling which is under-sampling and over-sampling techniques. Under-sampling reduces the size of the sample rate of the majority class equivalent to the minority class. Over-sampling increases the sample rate of the class of minority class up to the value of the majority class. While considering a financial case it is better to prefer over-sampling because there are chances of data loss in the under-sampling technique and while dealing with financial data, we cannot afford data loss, so researchers prefer over-sampling technique. According to the research of Sun, Li, Fujita, and Fu, SMOTE - Synthetic Minority Over-sampling Technique is best while dealing with finical data. SMOTE basically creates synthetic observations of the minority class. Sun, Li, Fujita, and Fu, proposed a model on how to effectively construct a dynamic financial distress prediction model based on class-imbalanced data streams with the help of using Adaboost-SVM with SMOTE and time weighting. This research was performed on the real-time dataset of 2682 chines companies, where 438 samples were financially distressed and 2190 samples were financially stable. In this research, they have created two different models. In the first model, they have considered Adaboost-SVM with time weighting (ADASVM-TW) in which they have used SMOTE to balance the imbalanced data. Where SMOTE is applied before the ADASVM-TW model is applied to make the data class balanced. The second model is designed in such a way that smote is embedded into the iteration of ADASVM-TW, which helps to create different patterns weighting mechanism. The main advantage of embedding smote into the process was it treats the old minority and new minority samples

differently. The results of this research show that the models underperform without applying SMOTE, it failed in recognizing the minority class of distressed company samples. They conclude that both models outperformed with the help of smote and the second model significantly outperforms the first model and Is more preferable while considering a Financial distress prediction (Jie Sun, 2019).

2.1.5 conclusion

The above literature review on different researchers helped us to understand the research work done by different researchers on Financial distress prediction. This helps us to understand how financial distress prediction previously performed using statistical analysis models like Beavers Univariant Model, Altman Z-score which were stationary models, to machine learning techniques like SVM, Decision tree, Naïve Bayes which performs well with small scale data and data with fewer features. As the technology evolved people moved towards creating hybrid models like a hybrid stepwise-SVM, LDA-SVM, to using multiple classifiers like Adaboost-SVM, Adaboost-Decision tree while using these techniques people are only focusing on increasing the performance using multiple classifier sand neglecting the class imbalance problem. As financial distress prediction related to a classification domain, many researchers are applying their own techniques and but didn't achieve success all time if they are archiving success, they are neglecting some important issues related to Financial Distress Prediction.

So, for this research, I will be applying various bagging and boosting techniques with the integration of SMOTE, where bagging and boosting techniques are also called as ensemble methods, and SMOTE is mainly used to deal with class imbalance problem. We will be comparing the results with previously used machine learning and hybrid machine learning models and will compare and check their performance, accuracy, and stability of the model. Along with that we will be comparing different SMOTE techniques and compare the results of how the model reacts to a different dataset created my different smote techniques.

2.2 Literature review on Stock Market Prediction

Stock Market which is also known as the equity market and share market, where a share is a financial instrument that represents ownership in a company. The stock market is a platform where aggregation of buyers and seller buys, sells, exchanges and issuance of stocks. The stock market is split into two parts like Primary market and secondary market. Where the primary market is part of the capital market which deals with issuance and sale of equity-backed securities to investors which are directly issued by the issuer. The primary market introduced new issues to the market through "Initial Public Offerings" (Hiransha M, 2018). Secondary Market is also known as the aftermarket which follows on public offering. It is a platform where previously owned or issued financial instruments are exchange are sold such as stock, bonds, and options (Hiransha M, 2018). The stock market is a collection of highly fluctuating time series data. Time series data is a sequence of numerical data in consecutive order. Most of the company, business, banking sector majorly depend upon these to make a profit and divide the risk (Kunal Pahwa, 2019). Stock prices and liquidity of the stock market are highly unpredictable and here comes Technology to help people out (Kunal Pahwa, 2019).

2.2.1 Stock Market Prediction Using Machine Learning Models

There are two machine learning techniques Supervised learning and Un-supervised learning techniques. According to the research of Nirbhey Pahwa, Neeha Khalfay and Vidhi Soni Unsupervised learning is not a good approach to predict the future stock price because unsupervised learning is used particularly when labels are not defined, there is no particular way to divide the dataset we have to perform iterations to find the best fit with the model (Nirbhey Singh Pahwa, 2017). According to my understand and research Stock market data is completely time series data. Time series data is a sequence of numerical data in consecutive order. and when we deal with time-series data the future results are always dependent on previous iterations. So, in this case, the future stock price is always dependent on previous iterations. Therefore, for stock market predication supervised learning is the best option. According to the experimental research, classification models outperform regression models. Nirbhey Pahwa, Neeha Khalfay, and Vidhi Soni performed stock prediction using SVM, Bayesian's Classifier, Decision Tree SVM

outperforms from the remaining models in terms of forecasting the data. This research was carried out on a pre-existing dummy dataset. (Paul D. Yoo, 2005) has proved that when we consider a huge dataset neural network models outperform other models including SVM. Neural networks have a better capability to understand the non-linear relationship from the input data to model non-linear dynamic data. According to my research and understanding, all these researches are performed on a statics dataset but the stock prediction is dependent on a real-time dataset. Researchers are taking stock prediction on another level by not just predicting the stock price based on previous data, but by predicting the stock price by social media information (Meghna Misra, 2018) has proposed a methodology where they predict the stock price by social media information. They found the limitation with prediction using previous data was that due to the changing pattern which results in a lack of accuracy (Meghna Misra, 2018). According to there, evaluation SVM outperformed all the models with an accuracy of 96%. Another research by (Zhen Hu, 2013) has proved that SVM performance was best amongst the other model.

2.2.2 Stock Market Prediction Using Deep Learning Models

Predicting the stock market and identifying its behavior and patterns has put up a challenge to the researcher. Researchers are trying their level best with the help machine learning to crack this. With the advancement of technology, researchers had moved towards deep learning and adapted deep learning models for predicting stock prices. The main reason to shift towards deep learning models is nonlinearity and nonstationarity of financial series make their prediction complicated (Mahla Nikou, 2019). This research was performed on the UK stock exchange dataset from the previous two years where different models like an artificial neural network, SVR, RF, and LSTM were used. Amongst these model's LSTM model outperformed by predicting the closest closing market price (Mahla Nikou, 2019). (Poonam Somani Shreyas Talele, 2014) has tried their hands on and has proposed a stock prediction using a hidden Markov model. The hidden Markov model is a statistical model that is used to model sequential data. For this research, the author has tested this model with the SVM model. The hidden Markov model outperformed the SVM model. Reacher's main aim to use this model was, the regular machine learning models were lacking in covering the stock price fluctuation. So, with the help of the Hidden Markov model author was able to cover it up (Poonam Somani Shreyas Talele, 2014).

Another research proposed by (Saleh Alhazbi, 2020) In this research they have used, Qatar stock exchange data for predicting the stock price. To execute this process author has used a Deep learning approach Convolutional Neural Networks (CNNs). In this research to improve the accuracy author is using external factors like the S&P index, Nikkei index, and oil price. Using the external factors as an input to the model author got a 10% hike in the accuracy of the model.

2.2.3 Conclusion

The above literature review on different researchers helped us to understand that while doing a stock market prediction real-time dataset are always helpful to train our model accurately. Stock market data consist of many nonlinearities and nonstationarity. According to my research Deep learning models always outperforms machine learning models. In deep learning, models outperform every other deep learning and machine, learning model. LSTM model works well with time-series data.

So, for this research. I will be using a real-time dataset, which I will be generating using "yfinance" API in python which allows you to access real-time datasets of any company using their ticker codes. I will be working specifically on the close price of the share of that particular company. I will be using advanced LSTM models and will be comparing their results and performance. With the help of these, we will be predicting as well as forecasting the next 30 days closing price of a particular company.

3 Design and Methodology

This section focuses on a detailed description of all the steps carried out during this research. This research is divided into two parts which are FDP prediction and Stock market prediction and forecasting. For this research we will be following CRISP-DM methodology, CRISP-DM (Cross Industry Standard Process for Data Mining) is an iterative process, which is a robust and well-proven methodology for machine learning (SPSS, 2000). The following fig 1.1 and 1.2 will give a clear outline of how this research was implemented.

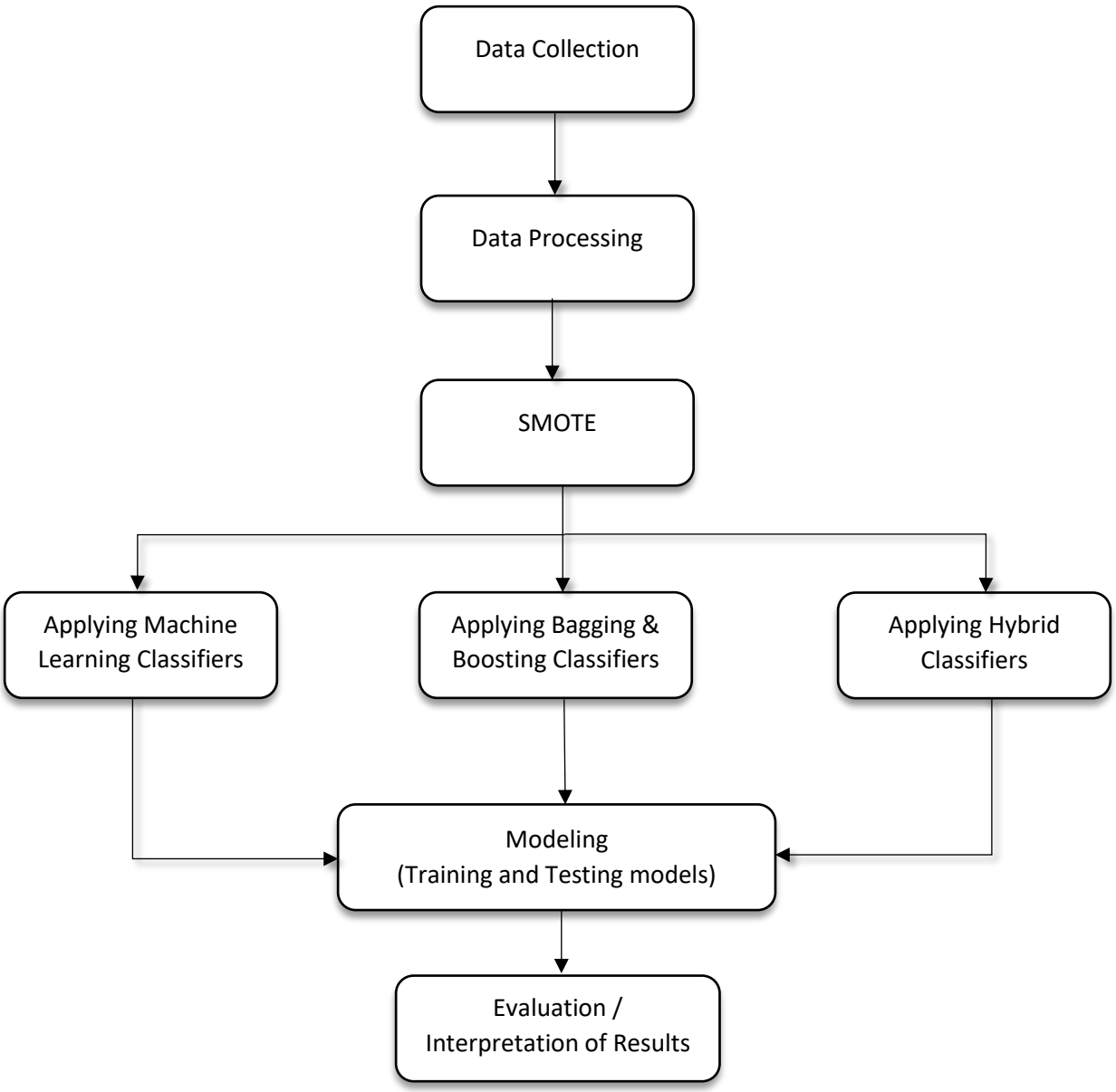

Fig 3.1 Thesis implementation diagram for FDP

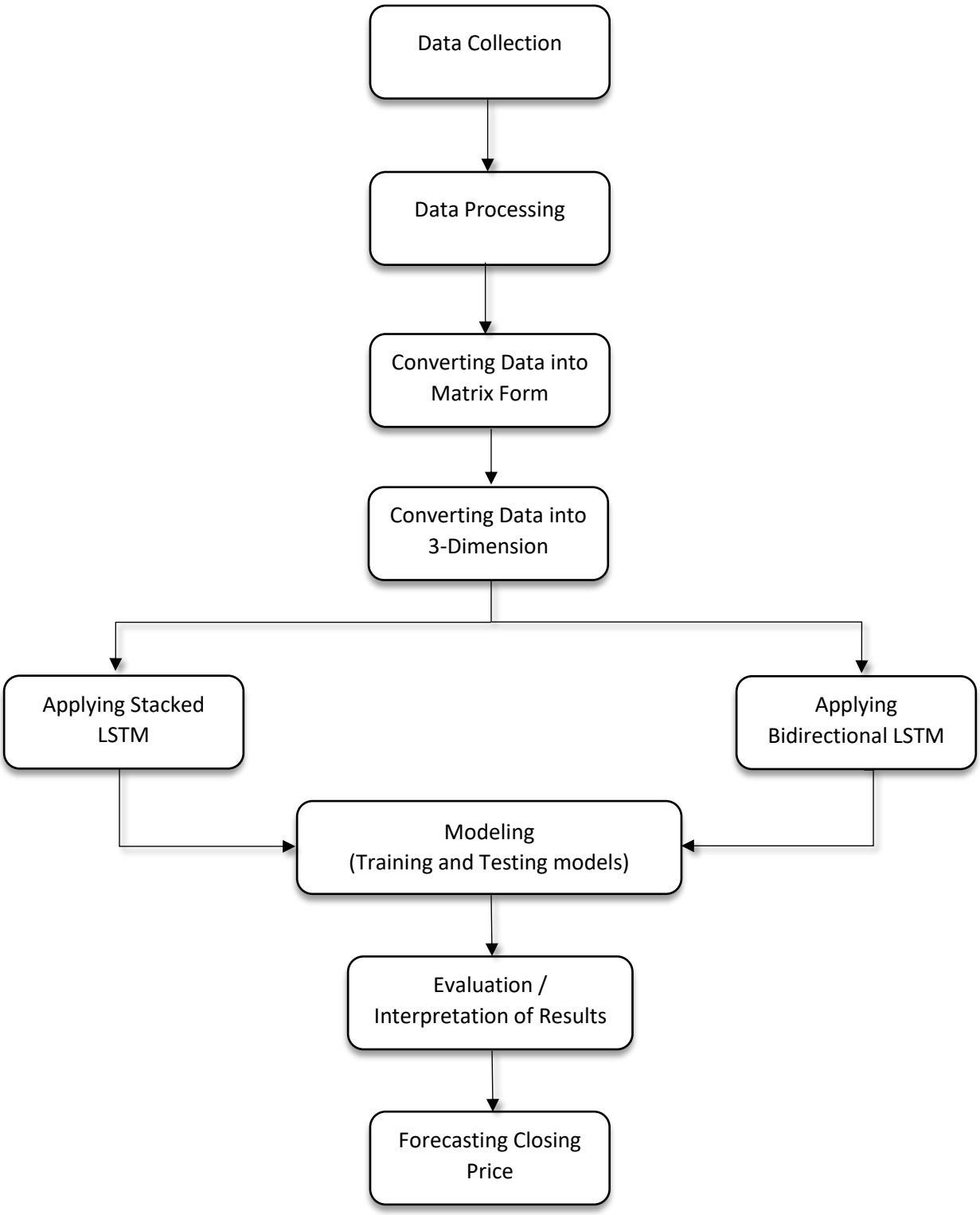

Fig 3.2 Thesis implementation diagram for Stock Market Prediction

3.1 Business Understanding

In this phase, overall business and project requirements are taken into consideration from a business point of view. This information is then converted into the definition of machine learning problems, to construct a project plan. This research mainly focuses on generating investment advice based on the FDP of a company and stock market trends. In this research, we will be using various bagging and boosting techniques for FDP and deep learning models for stock market prediction and forecasting.

3.2 Data Understanding

In this phase, the main objective is to collect the initial data according to the business requirement. To gain this understanding different activities will be performed, such as identity data quality problems, discover first insights into the data, and detect interesting subsets. In this research, we will be working on two different data set for two different models.

3.2.1 FDP data set description

For financial distress prediction, we got the dataset from Kaggle. This dataset contains actual values from 3673 chines companies. This dataset contains 83 financial and non-financial features like net profit / total assets, working capital / total assets, current assets / short-term liabilities, etc. The dataset is labeled dataset which contains the result column which describes the company is healthy or distressed. For this research, we have a class imbalanced dataset where there are 136 records of distressed companies and 3536 records of a healthy company. Table 3.1 shows the distribution of the company's financial status based on their scores

Company Status	Label	Score Range
Healthy	0	Score > -0.5
Distressed	1	Score < -0.5

Table 3.1 distribution of healthy and distressed companies based on their scores

3.2.2 Stock Prediction data set description

For the stock market prediction dataset, we are using the real-time data of a particular company. To get this real time dataset we have used yfinance API. yfinaicne is a “Yahoo Finance” API that aggregates the real-time stock prices of a particular company using the ticker symbol of the company. A ticker symbol is an abbreviation for a company to uniquely identify the specific company in the stock market. With the help of yfinaicne API, we can set the period of the data that is to be generated. yfinance API generates the following seven columns Date, open, close, adjacent close, volume. Specifically, for this research, we are only considering the closing price. On which we will be predicting and furcating the future closing price.

3.3 Data Preparation

Data preparation is the most important step of the CRISP-DM life cycle because the is the main information on which the models learn the patterns and generates the results accordingly. The data collected from the real world is not always in the machine learning format they are always raw. Real-time data always contains missing values noisy data and these kinds of data are not suitable for any machine learning model.

The following section describes the data preparation steps involved in preparing the dataset for both models:

3.3.1 Data Preparation for Financial Distress Prediction

3.3.1.1 Transforming the target variable from continuous values to categorical values

The target variable in our dataset is a score generated through the various financial and non-financial features present in the dataset. There are around 3536 records and each record has its score. These target variables were in the form of continuous data and classification models don’t work with continuous data they best perform with categorical data. So, to convert the continuous data into categorical data, we have categorized the data into two parts based on their scores. Companies with a score greater than -0.5 are considered as healthy companies and companies with a score less than -0.5 are considered as distressed companies. Healthy companies are represented with 0 and distressed companies are represented with 1.

3.3.1.2 Balancing the class-imbalanced dataset

When we consider an actual dataset, it is nearly impossible to get a perfectly balanced dataset. Considering the FDP scenario, it is not possible to get a balanced dataset. So, to convert a class-imbalanced dataset into a balanced dataset there are various methods available. Fist method is under-sampling, where the majority class is drilled down to the ratio of the minority class. In the under-sampling method, data is randomly deleted up until the majority class is equivalent to the minority class. The second method is oversampling, where minority class is extended up to the ratio of the majority class. In the oversampling method, data is randomly duplicated and generated in the minority class up until the ratio of the majority class. For this research, we have chosen to go with oversampling techniques. The reason for not choosing under-sampling was that when we deal with financial data, we cannot afford the loss of data. From the oversampling techniques, we have tried and tested the models with different techniques of SMOTE where we found the best results with the ADSYN Adaptive Synthetic Sampling Method. There are two ways to implement ADASYN first one is before splitting the dataset into train and test set and the second one is after splitting the dataset. For this research, we have implemented ADASYN after splitting the dataset because if we apply it before the split data will get into the validation or testing set, and when we apply the model it performs good, but when we apply the model on the different dataset it underperforms. So, to avoid this we first split the dataset into train and validation set and then apply ADASYN only on the train set so that the validation set can be raw and we can get accurate results.

3.3.2Data Preparation for Stock Market Prediction and Forecasting

3.3.2.1 Feature scaling the data

There are various machine learning algorithms that are highly sensitive when it comes to features. We use feature scaling to normalize the range of the features of data. If we don't use feature scaling then there are chances that the model can get biased towards the higher number. There are various methods of feature scaling like z-score or standardization and Min-Max scaling. The main difference between these two scaling methods is standardization scale the data in fix

range and in MinMax scaler we can define the custom range according to the model. For this research, we have used Min-Max scaling to scale the features.

3.3.2.2 Converting the data from the array to matrix

For stock market prediction we need time series data and time-series data in an array format. So, convert it into matrix form we have created and definition in which we have used the concept of timestamp for example if we are considering the closing price data from 1 January 2010 to 1 January 2020. To divide this data for prediction, we use timestamps. This means that if we put time stamp 4 so model will look for the stock price of the previous four and then generate the output and for the next row it will shift from one date and repeat the same process.

3.3.2.3 Splitting the time series data into train and test data set

When we are dealing with time-series data, we cannot directly apply the splitting methods like normal train random splitting and cross-validation. For time-series data future prediction or data is completely based on previous values. So, this time series data is split according to the date format for example from 10 Jan 2010 to 10 Jan 2017 will be training set and from 1 Feb 2017 to 1 Jan 2019 will be our validation set. We only have to give the percentage of splitting and the dates are automatically gets sorted accordingly. For this research, we have considered 100-time stamps for better results.

3.3.2.4 Converting the test data into 3 -dimensions

In this research for stock market prediction, we are using deep learning LSTM model. LSTM long short-term memory is a deep learning model based on RNN (recurrent neural network) architecture. LSTM consists of recurrent layers, where it requires 3- D data as input. Where the three dimensions are batch size, timestamp and input dimension. For our research, the batch size is 2624. The timestamp is 100 and the input dimension is 1. For converting the data, we have used reshape function to change the dimension from 2-D to 3-D.

3.4 Modeling

After the data is properly pre-processed and ready to serve as input to a machine learning model for better results and performance.

3.4.1 Models used in FDP

3.4.1.1 XGBoost Classifier

XGBoost (eXtreme Gradient Boosting) algorithm is a boosting algorithm. XGBoost works on the gradient boosting framework (Tianqi Chen, 2016). Boosting is a technique which works sequentially on the principles of ensemble learning. Boosting is a combination of weak learners which provides results with improved accuracy. The boosting technique uses a sequential approach to train the model. XGBoost is a decision tree-based algorithms. XGBoost performs great with structured and tabular data. In simple machine learning models, a single model gets trained with the dataset. we can only do some parameter tuning to make the performance better. On the other hand, the boosting technique uses a more iterative approach that combines multiple models sequentially. Where each model learns from the error made by the previous model and improves the same and serves as input to another model. In this sequential manner, we get the best result out of all. XGBoost does not require any scaling of data. XGBoost has fewer chances to get overfit. Using XGBoost we can find the feature importance.

Following diagram represents the working cycle of XGBoost model

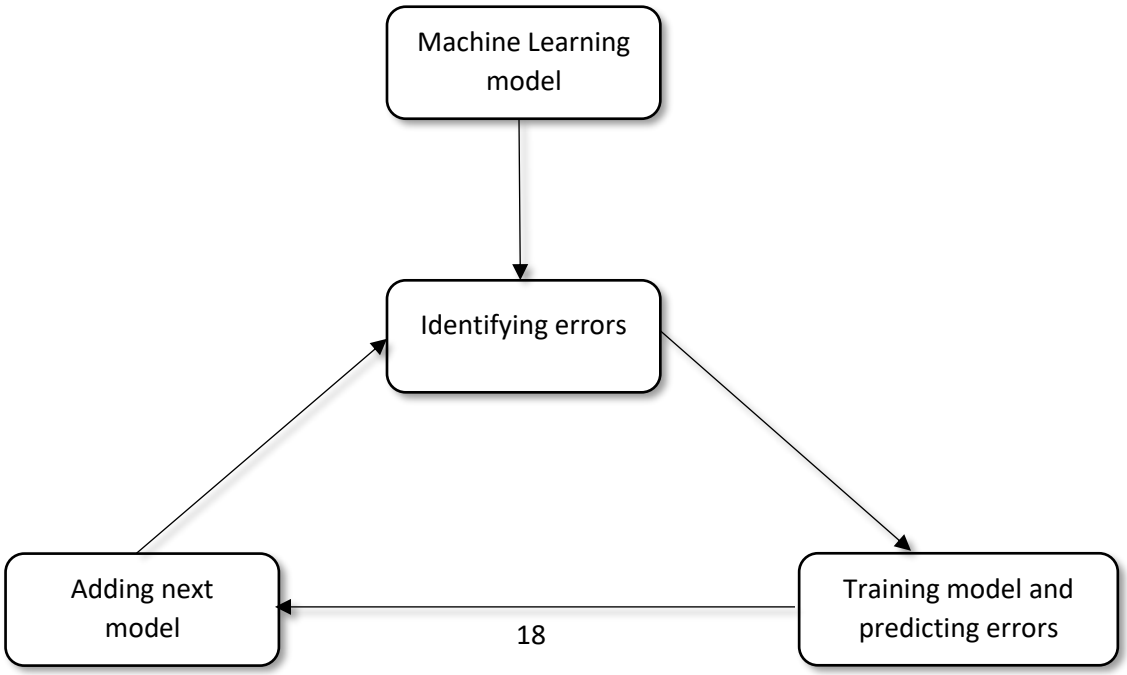

Fig 3.3 Working cycle of XGBoost

Following are the steps involved in building an XGBoost model

Step 1: Installing XGBoost library for python

Step 2: load and prepare the data set for modeling

Step 3: Training the XGBoost model on the training set

Step 4: making prediction

Step 5: evaluating the model

3.4.1.2 Balanced Bagging Classifier

A balanced bagging classifier is a bagging classifier. Bagging is an ensemble algorithm that uses multiple models on various training subsets of the dataset. Bagging is a technique which works parallelly on the principles of ensemble learning. After the parallel execution, all the results are combined from all the models to generate the final result using the voting method. The balanced Bagging classifier approach is mainly used with imbalanced data. Balanced Bagging classifier works on a random under-sampling strategy. Each model in bagging is trained parallelly.

3.4.1.3 Random Forest Classifier

Random forest is also known as a random decision forest. Random forest is a supervised machine learning algorithm Random forest is an ensemble learning method that is used for both classification and regression. In this research, we have used a random forest classifier. Random forest classifier is an ensemble tree based classifier. Ensemble algorithms are those algorithms that combine multiple different classifiers or same classifiers multiple times. Random forest classifier is a combination of a decision tree which are randomly selected from the subset of the training set. Random forest generates the result based on the aggregation votes provided by each decision tree. Random forest is considered as the most accurate machine learning classifier for many datasets. The random forest creates randomness inside the trees. The random forest model does not suffer from overfitting problem. The random forest can deal with missing values on its own. Random forest uses a rule-based approach.

The following figure demonstrates the working cycle of Random forest

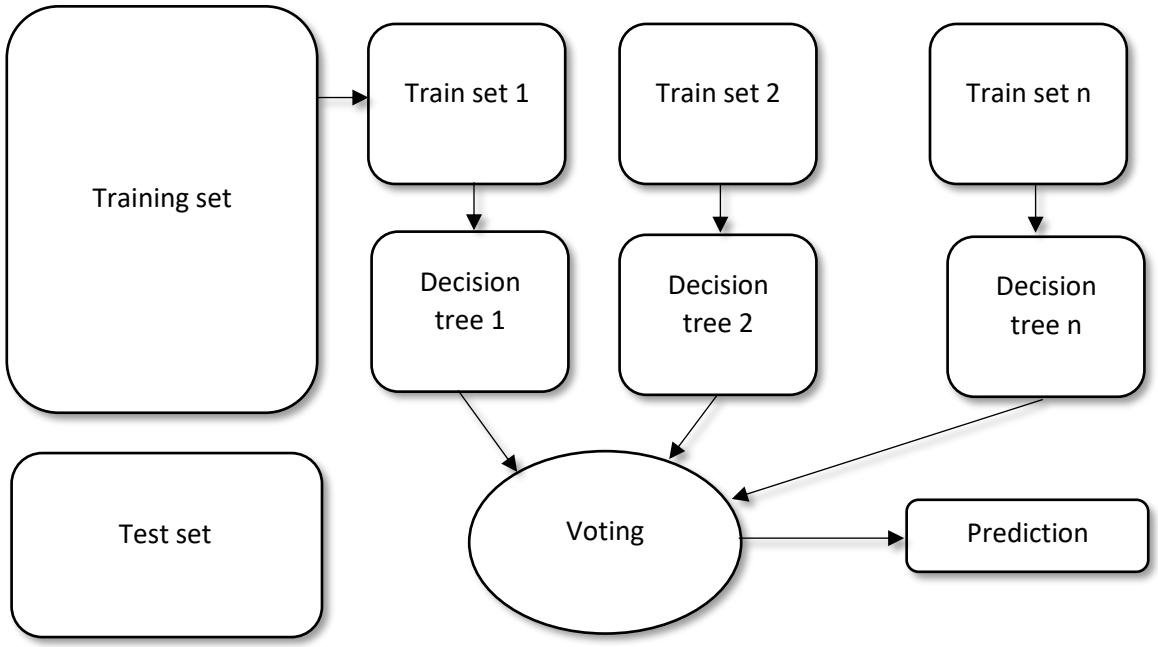

Fig 3.4 Working cycle of random forest classifier

Following are the steps for working of the algorithm

Step 1: Random samples are selected from the training dataset

Step 2: decision tree is constructed for each sample and the results are collected from each decision tree.

Step 3: Voting system is used to get the best result

Step 4: According to the voting system prediction results with the highest votes is considered as the final result.

3.4.1.4 Adaboost-SVM

AdaBoost-SVM is an ensemble learning method used commonly for the classification. AdaBoost is an ensemble learning classifier which works on the principle of combining a series of the low-performance classifier with an objective of creating an improved classifier. AdaBoost is a non-linear classifier, it is robust to overfitting as this method minimizes the upper bound of the training error by properly selecting the optimal week classifier and voting weight. It is one of the meta algorithm that combines several machine learning methods into a single predictive model

for enhancing the performance of the model. SVM (Support Vector Machine) is a supervised machine learning algorithm that can comfortably work with the textual and numerical datasets that are it can be used for both classification and regression problems. It is also used for solving real-world problems as this classifier is capable of classifying two-group classification problems. AdaBoost-SVM is created by merging two algorithms AdaBoost algorithm and SVM algorithm, this method parallelizes by allocating each base learner to different-different machines for achieving better performance in classification on an imbalanced dataset. AdaBoost-SVM demonstrates better generalization performance than SVM on the imbalanced classification problem.

3.4.1.5 Naïve Bayes Classifier

Naïve Bayes algorithm is a classification algorithm. Naïve Bayes algorithm follows Bayes theorem. It is used to classify different objects based on certain features. Naïve Bayes works well with a huge dataset with a minimum number of features. It has on dependent feature and all rest of the features are independent features. Naïve Bayes is a combination of multiple algorithms. Naïve Bayes can be used for real time prediction, it is very scalable with a huge dataset. Naïve Bayes is good at handling irrelevant features. Naïve Bayes is also considered a bad estimator in some scenarios. Naïve Bayes is mainly used for spam filtering, text classification, stock prediction.

Following steps represents the working of naïve bayes algorithm

Step 1: Dataset is converted into a frequency table
Step 2: Likelihood tables are created by finding the probability.
Step 3: Now naïve bayes equation is used to calculate the posterior probability for each class.
Step 4: class with the highest priority is the final result of the prediction.

Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig 3.5 Bayes Theorem

3.4.1.6 Decision Tree Classifier

Decision tree classifier is a supervised learning algorithm. Decision tree splits the data continuously according to the parameter. Decision tree algorithm works for both classification and regression problems. Decision tree consists of nodes, edges, and leaf nodes. Where nodes are the value of the attribute. Edges are the outcomes of the test which further splits int node or leaf. Decision tree builds a tree-like structure which is obtained by binary recursive partitioning. Decision tree performs best when classifying unknown records. The only problem with the decision tree is overfitting. Decision trees are mainly used in financial analysis, Astronomy, etc.

The following figure represents the decision tree.

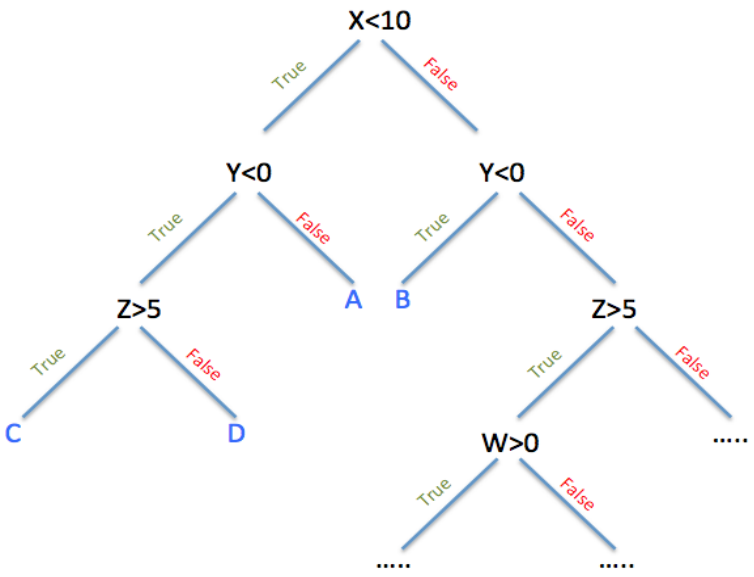

Fig 3.6 Decision tree structure

3.4.1.7 Logistic Regression

Logistic regression is one of the simple and popular algorithms used for performing classification. It is used for predicting the continuous dependent variables by using a given set of independent features. It uses different methods for estimating the parameters which give an enhanced result means the model will result as unbiased with lower variance. It is a supervised learning classification algorithm that is used to predict the probability of the target variable. To achieve model coefficients that are related to the predictors for predicting target it utilizes the maximum likelihood estimation (MLE). One of the advantage of using logistic regression is that it not only gives the measure of how relevant a predictor is but also its direction(positive/negative) of association, this algorithm is easier to implement, interpret and very efficient to train the model

3.4.2 Models used in Stock Market Prediction

3.4.2.1 Stacked LSTM

LSTM Long short-term memory is a deep learning model that works on the principles of artificial recurrent neural network architecture. LSTM is used to learn order dependencies in a prediction problem where data is present in a sequential format. LSTM models are comprised of hidden layers where the original model consists of only a single hidden layer. Stacked LSTM is an upgraded version of the original LSTM. Stacked LSTM is comprised of multiple hidden LSTM layers. Each layer in stacked LSTM consists of multiple memory cells.

Working of Stacked LSTM model.

The architecture of stacked LSTM is designed in such a way that a single LSTM model is combined with multiple hidden LSTM layers one upon another. First, the input is provided to the first LSTM model then after processing the first LSTM model provides a sequential output rather than creating a single output from series of data which will serve as an input to the second LSTM layer. After all the LSTM layers a dense layer is included which will generate the final output. In stacked LSTM layers the next layer learns from the techniques used by previous layers and provides the improved output.

Following figure demonstrates the architecture of stacked LSTM

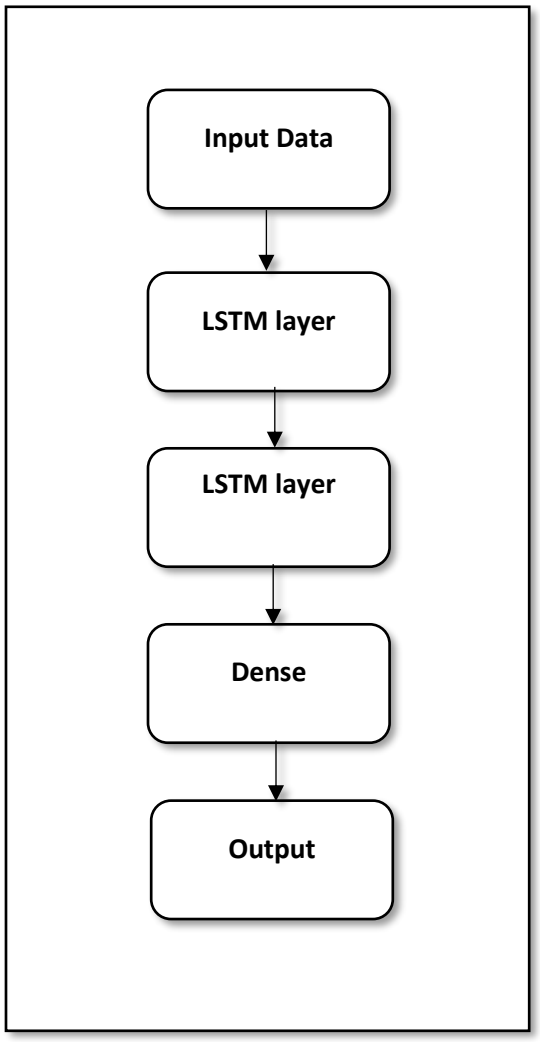

Fig 3.7 Stacked LSTM Architecture

3.4.2.1 Bidirectional LSTM

LSTM Long short-term memory is a deep learning model that works on the principles of artificial recurrent neural network architecture. Bidirectional LSTM is another advanced version of traditional LSTM. Bidirectional LSTM works best with sequential data. It is mainly used for classification problems that contain sequential data. Bidirectional LSTM consists of two hidden layers that are connected to each other in the opposite direction to gain the same output.

Working of Bidirectional LSTM

Bidirectional LSTM trains two models during its process as compare to traditional LSTM which trains only a single model. It trains the model in a sequential format. Input given to both the models are the same, the only difference between the input provided is that the first model gets the exact input and the second model gets the reversed version of the input. This technique is used to learn from every aspect of the input and generate additional information that will help to get better results.

3.5 Evaluation

Evaluation of the models we have applied is done in this step. All the models perform good but to find the best amongst it is a challenge. This can be done using various evaluation techniques available. The evaluation depends on what techniques and data you are using. For example, for this thesis, we have an imbalanced dataset so for imbalanced data we do not rely on accuracy but we also compare the precision, recall, and f-score to calculate the final result.

3.5.1 Accuracy

Accuracy is used to evaluate the classification model. Accuracy calculated the correct predictions done by the model. Accuracy is only reliable for a balanced dataset.

Formula: Accuracy = No of correct predictions / Total no of predictions

3.5.2 Confusion Matrix

Confusion matrix or error matrix is a table representation that is used to describe the performance of the model. Confusion matrix is used to evaluate classification models. The table value represents the correct prediction and incorrect prediction made by the model.

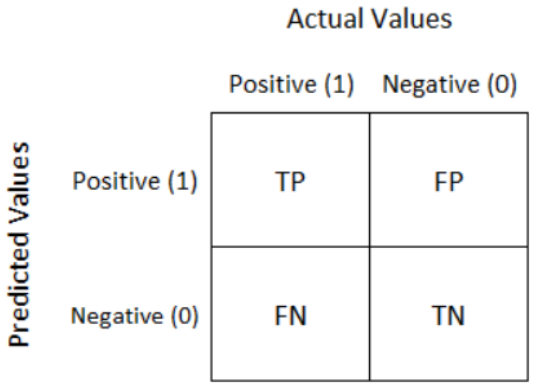

Fig 3.8 Confusion Matrix

3.5.3 Recall

Recall or true positive rate is used to find the ratio between true positive prediction from the actual number of positive predictions. Recall is mainly used when the dataset has an imbalanced data.

Formula to calculate recall.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Fig 3.9 Recall Formula

3.5.4 Precision

Precision is a ratio between predicted positive values and total predicted positive. Precision is mainly used when the dataset has an imbalanced value. Below is the formula to calculate the precision.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Fig 3.10 Precision Formula

3.5.5 F1 Score

F1 score is a measure of test accuracy. F1 score is the harmonic mean of precision and recall. Higher the f1 score better the model. Below is the formula to calculate the F1 score of the model.

$$F = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

Fig 3.11 F1-Score Formula

3.5.6 RMSE

RMSE root-mean-squared error is a measure that is used to fin the difference between the values predicted by the model and the actual values. RMSE is mainly used in the evaluation of regression and deep learning models.

4 IMPLEMENTATION AND RESULTS

4.1 Introduction

This phase is a continuation of chapter 3 where we have seen all the steps and models we are going to implement in this research. This phase particularly focusses on actual working and implementation of each step-in chapter 3.

4.2 Modeling

4.2.1 Models used in FDP

4.2.1.1 XGBoost Classifier

XGBoost is a boosting method which is mainly used for imbalanced dataset. In this research, we have combined XGBoost with ADSYN SMOTE which gave an accuracy of 94.34% and F-Score of 33.70%.

4.2.1.2 Balanced Bagging Classifier

Balanced Bagging classifier is an ensemble learning algorithm. Where balanced bagging is a bagging algorithm which is mainly used for imbalanced data. With a combination of ADASYN SMOTE, it gives an accuracy of 94.1% and an F-Score of 43.10%.

4.2.1.3 Random Forest Classifier

Random forest which is also called as traditional bagging classifier which uses multiple decision trees to find the output. In this research random forest in combination with ASASYN SMOTE has performed and gave an accuracy of 93.28% and F-Score of 39.28%. These results were obtained using 5 estimators.

4.2.1.4 Adaboost-SVM

Adaboost-SVM is a hybrid classifier. It is a combination of an ensemble learning algorithm and machine learning algorithm, where Adaboost is an ensemble learning algorithm and SVM is a machine learning algorithm. Adaboost SVM when performed with the integration of ADASYN SMOTE it results in 95% accuracy and 17.91% F-Score.

4.2.1.5 Naïve Bayes

Naïve Bayes is a machine learning algorithm that is used mainly with problems with multiple classes. In this research, it has underperformed with an accuracy of 15.60% and an F-Score of 7.94%.

4.2.1.6 Decision Tree

Decision tree is machine learning which is considered as the best model when it comes to classification problems with many features. After implementing the decision tree, we got an accuracy of 92.91 % and an F-Score of 19.81%. This is the best accuracy we got from the traditional machine learning model.

4.2.1.7 Logistic Regression

Logistic regression is a machine learning classifier, in this research logistic regression is the worst performer form all the listed models. Logistic regression results inaccuracy of 15% and F-Score of 7.9%

4.2.2 Models used in Stock Market Prediction

4.2.2.1 Stacked LSTM

Stacked LSTM is an upgraded version of LSTM where it uses multiple LSMT layers in a stacked format. In this research, we used 2 layers of stacked LSTM with a timestamp of 100 days and batch size accordingly. Stacked LSTM model resulted with a minimum error rate. RMSE of the training set is 0.0181 and RMSE of validation set is 0.0276

4.2.2.2 Bidirectional LSTM

Bidirectional LSTM is an advanced version of traditional LSTM. In this research, Bidirectional LSTM has underperformed as compared to stacked LSTM. The RSME of bidirectional LSTM is 0.0455 for the training set and 0.193 for the validation set. In bidirectional LSTM the model is been overfitted because RMSE of validation set is greater than RMSE of the training set.

4.3 Evaluation

4.3.1 Evaluation of FDP Models

To evaluate a classification model we check the accuracy, precision, recall, f1 score, and performance metrics which is calculated using the confusion matrix. The following table demonstrates the test results we have performed on various models.

The following table represents the comparison of different models based on various evaluation methods.

Models	Accuracy %	Precision %	Recall %	F1 Score%	TP	TN	FP	FN
XGBoost Classifier	94.34	31.25	36.58	33.70	1028	16	33	25
Balanced Bagging Classifier	**94.1**	**33.33**	**60.97**	**43.10**	**1013**	**24**	**48**	**17**
Random Forest	93.28	30.98	53.65	39.28	1005	23	56	18
Adaboost-SVM	**95**	23.07	14.63	17.91	1041	6	20	35
Naïve Bayes	15.60	41.40	97.50	7.94	132	40	929	1
Decision Tree	92.19	15.71	26.82	19.81	999	17	62	24
Logistic Regression	15	41	97.5	7.9	130	42	928	2

Table 4.1 Performance Table of FDP Models

For FDP we have used some machine learning models like naïve Bayes, decision tree the accuracy of these models was so bad except decision tree with an accuracy of 92.19% but when we are dealing with an imbalanced dataset, we do not consider accuracy as the main evaluating feature because it provides misleading information. To evaluate a model performing on imbalanced data we consider the F1-score which a harmonic mean of precision and recall. When comparing the F1-Score of machine learning models every model underperformed. We have used various bagging and boosting techniques where we found good results with XGBoost Classifier, Balanced Bagging Classifier and Random forest all the models were having accuracy above 90% but among these three only balanced aging f1-Score was good. We have tested this model against a hybrid model Adaboost-SVM which gave an accuracy of 95% but the F1-score was very low as compare to the bagging and boosting model.

Below fig 4.1 and 4.2 demonstrate the comparison of accuracy and f1-score of all the models

Models:

1 – XGBoost

2 – Balanced Bagging

3 –Random Forest

4 – Logistic Regression

5 – Decision tree

6 – Naïve Bayes

7 – Adaboost-SVM

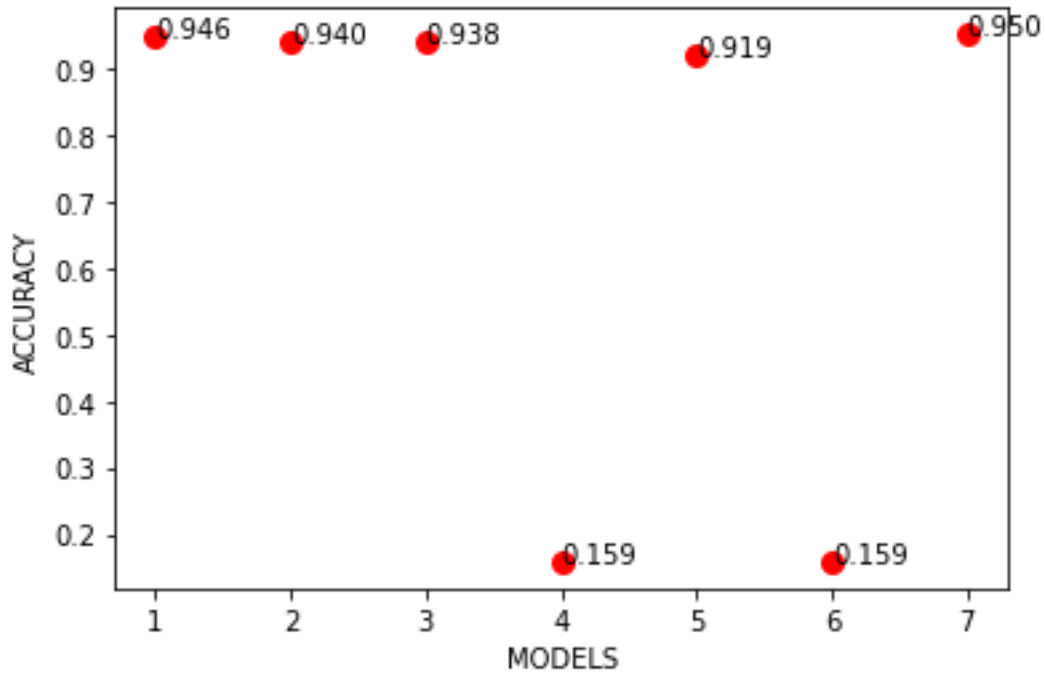

Fig 4.1 Accuracy Comparison Chart

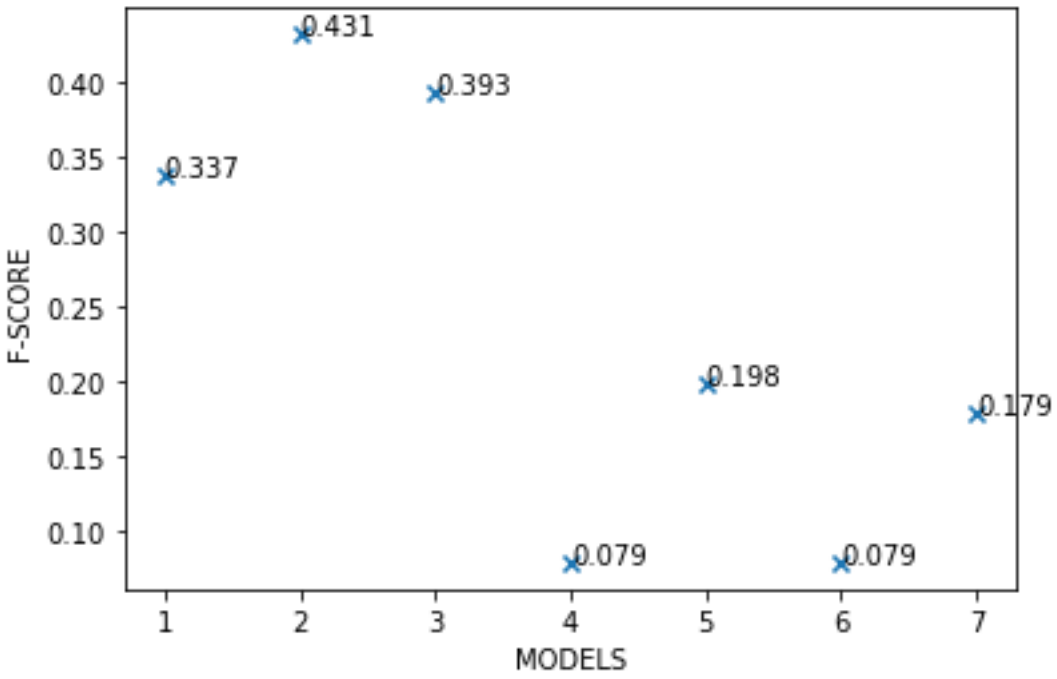

Fig 4.2 F1-Score Comparison Chart

Finally, from the above evaluation, we can conclude that Balanced Bagging with the integration of ADASYN is the best model for FDP

4.3.2 Evaluation of Stock Prediction Models

Models	**RMSE (Train Set)**	**RMSE (Test Set)**
Stacked LSTM	0.0181	0.0276
Bidirectional LSTM	0.0455	0.1934

Table 4.2 Performance Table of Stock Prediction Model

The following table demonstrates the results of stock prediction model. For stock prediction, we have used the time-series dataset. to check the performance of time series dataset we calculate the RMSE of train seta and validation set to check the performance of the model. Based on the performance of the model Stacked LSTM with two stacked LSTM performs best as compare to the other model with minimum rmse score.

5 CONCLUSIONS

How to invest and get effective results is always an important research topic for financial management. Considering the investment scenario many investment-related models are created with the help of machine learning based on market trends. The only thing these investments depend on the previous stock records of the company. However, they almost ignore the minute fingerprints of financial distress. This paper tries to consider this problem and tries to propose two models that will refine the investment process.

The first model will predict the financial health of the company and will give advice for investment based on the outcome of the model that will predict that the company is healthy or distressed using Balanced Bagging ensemble classifier with the integration of ADASYN. Where Balanced Bagging model simply applies ADASYN to balance all the variables. ADASYN is applied only to training data to get accurate results. The empirical experiment for the FDP model is carried out based on the real-world data a total of 3672 Chinese companies, where the data set consists of 136 financial distressed samples and 3536 healthy samples. Experimental results show that the Balanced Bagging classier with ADASYN increased accuracy by 2-3% as compare to the Balanced Bagging classier with SMOTE. Results also show that Balanced Bagging Classifier outperformed from all the bagging, boosting, ensemble, and traditional machine learning algorithms.

The second model will do stock perdition and forecasting of stock prices up to 100 days of a particular company using Stacked LSTM. Where stacked LSTM is an advanced version of traditional LSTM where multiple LSTM layers are stacked into a single LSTM model which increases the accuracy of the model. The empirical experiment for the stock prediction and forecasting is carried out on real-time datasets with time-series data. The real-time data was retrieved from Yahoo Finance using Finance API in python. Using yfiance previous 5 years data of that particular company was retrieved. This research particularly focused on the closing price of the company. Experimental results show that traditional LSTM and bidirectional LSTM underperforms as compare to the results of the stacked LSTM model.

6 FUTURE WORK

In future research, these models can be enhanced using more aspects of investments. The first model can be implemented in various directions like default diagnosis, health sector, fake news detection, etc. These FDP models can be further explored with deep learning algorithms and can be compared with the performance of various ensemble learning models. The seconds model can be further enhanced using hybrid deep learning models and for casting more clear and accurate predictions. The research can be more refined by considering both opening and closing price prediction and forecasting.

References

Altman, E. I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance.*

Chen, M.-Y., 2011. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Elsevier.*

Chong, Z., 2020. Analysis of Financial Crisis Early Warning Model of Listed Enterprises in China. *Journal of Physics: Conference Series.*

Francisco Louzada, P. H. F.-S. C. A. D., 2012. On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. *Expert Systems with Applications.*

Haibo He, Y. B. E. A. G. a. S. L., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence,* pp. 1322-1328.

Hiransha M, G. E. ,. V. K. M. S. K., 2018. NSE Stock Market Prediction Using Deep-Learning Models. *International Conference on Computational Intelligence and Data Science (ICCIDS 2018).*

Jie Sun, H. L. ,. H. F. ,. B. F. ,. W. A., 2019. Class-imbalanced dynamic Þnancial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion.*

Kunal Pahwa, N. A., 2019. Stock Market Analysis using Supervised Machine Learning. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con).*

Kyung-Shik Shin, T. S. L. H.-j. K., 2005. An application of support vector machines in bankruptcy prediction model.

Mahla Nikou, G. M. J. B., 2019. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting Finance & Management .*

Meghna Misra, A. P. Y. H. K., 2018. Stock Market Prediction using Machine Learning Algorithms: A Classification Study. *International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering - (ICRIEECE).*

Mselmi, N. L. A. H. T., 2017. Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis.*

Nirbhey Singh Pahwa, N. K. V. S. D. V., 2017. Stock Prediction using Machine Learning a Review Paper. *International Journal of Computer Applications (0975 – 8887).*

Noviyanti Santoso, W. W., 2018. Financial Distress Prediction using Linear Discriminant Analysis and Support Vector Machine. *Journal of Physics.*

Ohlson, J. A., 1980. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research.*

Paul D. Yoo, M. H. K. T. J., 2005. Machine Learning Techniques and Use of Event Information for Stock. *International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on.*

Poonam Somani Shreyas Talele, S. S., 2014. Stock Market Prediction Using Hidden Markov Model. *IEEE 2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) .*

Saleh Alhazbi, A. B. S. A. A.-M., 2020. Using Deep Learning to Predict Stock Movements Direction in Emerging Markets: The Case of Qatar Stock Exchange. *IEEE.*

Samuel Olusegun Ojo, P. A. O. M. M. a. J. A. A., 2019. Stock Market Behaviour Prediction using Stacked LSTM. *IEEE.*

SPSS, 2000. CRISP-DM. In: *CRISP-DM.* s.l.:SPSS.

Tianqi Chen, C. G., 2016. XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785-794.

Zhen Hu, J. Z. ,. a. K. T., 2013. Stocks Market Prediction Using Support Vector Machine. *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering.*

Zhi Xiao, X. Y. Y. P. X. D., 2011. The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster–Shafer evidence theory. *Knowledge Based System.*