****

**CAPSTONE PROJECT**

**STOCK MARKET PREDICTION USING MACHINE LEARNING**

**NOVEMBER 2021**

**BY**

**ANTHONY NNAMANI**

**INTRODUCTION**

Stock trading has been in existence for more than two centuries. People invest or trade stock for a variety of reasons, ranging from wealth creation and accumulation, to investing in stock as a retirement plan. Knowing the right stock to buy can be very difficult and time consuming as the market is very volatile and risky and full of uncertainties. Any method or technique that can be applied to understand stock market and identify the right stock to invest in will be very beneficial as this will the investors to minimise loss and make good profit and return on their investment. In this study, machine learning algorithm was trained and applied to study stock market data and subsequently predict the stocks that are buy worthy. Logistic regression, RandomForest, Naïve Bayes, Support Vector Machine, Gradient Boosting and Extreme Gradient Boosting were the supervised machine learning models that were trained in this project. Accuracy, precision, recall, AUC, F-1 score were the metrics used to evaluate the models and RandomForest was selected as the best model based on these metrics

**BUSINESS QUESTION**

How do investors know and identify the right stock to invest in? As the stock market is very volatile and risky, investors lose huge amount of money daily on stocks, so it is very important that investors perform their due diligence by conducting appropriate research on the market before investing their money or risk losing their investment.

**DATA QUESTION**

Can a machine learning model be trained to study the stock market data and then predict if a stock is worth buying or not worth buying?

Machine learning is a technique in data science and artificial intelligence in which a computer application is trained to study dataset to learn pattern and understand the dataset and then apply its understanding by making prediction on a new but similar dataset. If a model can be trained to study the financial indicators in a company’s historical financial dataset and then use the knowledge on a new data to predict if the company’s share is good to invest in, this will go a long way in helping investors identify those stocks that have potentials of giving investors good return.

**IMPORTANCE OF THIS PROJECT**

The significance of this project can not be over emphasis. Stock market is a huge market. It worth trillions of dollars globally. It is one of the biggest investment strategies for a lot of people for over two centuries now. The table below show the market capitalization of some of the major stock market around the world.

| **Domestic Market Capitalization (USD millions)** | | |
| --- | --- | --- |
| **Exchange** | **Location** | **Market Cap.\*** |
| NYSE | U.S. | 22,987,587 |
| Nasdaq | U.S. | 13,286,825 |
| Japan Exchange Group | Japan | 6,000,171 |
| Shanghai Stock Exchange | China | 5,037,349 |
| Euronext | France | 4,821,103 |
| Hong Kong Exchanges and Clearing | Hong Kong | 4,595,366 |
| LSE Group | U.K. | 4,024,164 |
| Shenzhen Stock Exchange | China | 3,454,965 |
| TMX Group | Canada | 2,386,066 |
| Saudi Stock Exchange (Tadawul) | Saudi Arabia | 2,333,838 |
| BSE India Limited | India | 2,181,351 |
| National Stock Exchange of India Limited | India | 2,162,693 |
| Deutsche Boerse AG | Germany | 2,020,041 |
| SIX Swiss Exchange | Switzerland | 1,775,268 |
| Nasdaq Nordic and Baltics | Sweden | 1,594,481 |
| Australian Securities Exchange | Australia | 1,497,599 |
| Korea Exchange | South Korea | 1,402,716 |
| Taiwan Stock Exchange | Taiwan | 1,143,210 |
| B3 - Brasil Bolsa Balcão | Brazil | 1,118,281 |
| Moscow Exchange | Moscow | 772,189 |

**Source**; Investopidia.com

Stock market is very volatile, price of stock fluctuates, and investors can lose huge amount of money very quickly. When people invest in stock without conducting a proper and indebt research on financial market and companies they are interested in, they are likely to lose their money as their investment and judgement will be influenced mainly by their sentiment.

**STAKEHOLDERS FOR THIS PROJECT**

Stocks are traded on various stock exchange platforms like the Australian Security Exchange through majorly by stockbrokers who act as the middlemen between the seller and the buyer. The stockbrokers offer a range of services to the investors. They mediate and facilitate the buying and selling of shares, financial advice as well as wealth and portfolio management to their clients. Some of the stockbroking firm here in Australia include CommSec of the Commonwealth Bank, CMC market, Wespac, NAB trade, Kalkine etc. the brokers and researchers in these financial institutions mentioned above are all stakeholders in this project.

**PROJECT PROCESS OVERVIEW**

**DATA AQUIZATION AND OVERVIEW**

The data was collected from Kaggle. According Kaggle, the data was originally collected by leveraging Financial Modelling Prep API and Pandas\_datareader. The data is a five-year financial data. Starting from 2014 to 2018. I dropped the 2014 data as it has almost same value with 2015 data, so I used 2015 to 2018 data for this project. The dataset has a combined 17957 observations and 224 features

Data Source: <https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018>

**EXPLORATORY DATA ANALYSIS**

Some analysis was conducted on the dataset to grab insights on the data. Because each observation on the dataset represents a different company, it was not possible to analyse and explore the data as whole. Instead, distribution of companies across business sectors was used to visualize the data. Secondly, one company (Proter and Gambel; P&G) was selected for quick financial analysis. Some of the visualizations are shown below.

**DATA VISUALIZATION HIGHLIGHT**

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, icon, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**PREPROCESSING**

The dataset was so messy and had a lot of missing values. Some of the columns were repeated while some other had different names but same value. In terms of handling the missing value, the dataset was grouped by sectors and then used the mean value of each sector to impute for the missing values in that sector. The sector column which was a categorical variable was changed to numeric by applying label encoding. This method was selected over others to reduce noise and overfitting in the model since there are many sectors in the dataset and the ‘’Class’’ column which is the target variable was used as the label in the label encoding process and finally, the data was standardised using StandardScaler(). The features used for the models were selected by eliminating columns that are highly correlated to each other and columns that were repeated as the data was so messy. This was done by using PyCrate, an automated machine learning package which ran the through the data and automatically pre-process the dataset and selected features that are important to training the models.

**MODELLING**

Five different supervised machine learning models were trained in this project. The models trained are discussed below.

**LOGISTIC REGRESSION** model works by modelling the probability of a distinct outcome. And it is used mostly for binary classification, and it is easy to implement.

**RANDOM FOREST**, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. It fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

**NAÏVE BAYES** is a [classification technique](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=6stepsnaivebayesarticle) based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**GRADIENT BOOST** works by weighting the observations, putting more weight on difficult to classify instances and less on those already handled well. New weak learners are added sequentially that focus their training on the more difficult patterns. A loss function to be optimized., a weak learner to make predictions and additive model to add weak learners to minimize the loss function.

**EXTREME GRADIENT BOOST (XGBOOT)** XGBoost is similar with gradient boosting but XGB applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Result** | | | | |
|  | ACCURACY | PRECISSION | RECALL | F1 SCORE |
| **REGRESSION** | 0.58 | 0.59 | 0.69 | 0.53 |
| **RANDOMFOREST** | 0.67 | 0.68 | 0.67 | 0.68 |
| **NAÏVE BAYES** | 0.52 | 0.64 | 0.16 | 0.65 |
| **GBOOST** | 0.64 | 0.65 | 0.68 | 0.67 |
| **XGBOOST** | 0.65 | 0.66 | 0.68 | 0.67 |

Chart, line chart

Description automatically generated

Using these metrics to evaluate the models it can be seen that RandoForest classifier is the better model as it higher metrics value. It is very important to note that for this project, precision is preferred over recall. This is because the focus to minimise false positive to the barest minimum. Since the project to train a model that will predict good stock investment, it will be dangerous to misclassify a bad stock as good stock as this will lead to investing in bad stock and subsequently losing the money invested.

**METRICS FOR MODEL EVALUATION**

Same metrics were applied to evaluate all the models from which the best model was selected. The metrics used are discussed below.

**ACCURACY**: This is the ratio of the total number of correct predictions and the total number of predictions.

**PRECISION** is a metric that quantifies the number of correct positive predictions made. valuates the fraction of correct classified instances among the ones classified as positive. Precision can be calculated as follows, Precision = TruePositives / (TruePositives + FalsePositives). This metric is preferred when the focus is to minimise false positive.

**RECALL**: Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. It is calculated as follows Recall = TruePositives / (TruePositives + FalseNegatives). In a project where false negative needs to be minimised, recall is the appropriate metric to be used in evaluating the model.

**F-SCORE** Is the harmonic mean of precision and recall. It combines the precision and recall metrics to give a better view of the model performance. It can be calculated using the formular

F-Measure = (2 \* Precision \* Recall) / (Precision + Recall).

**THE RECEIVER OPERATOR CHARACTERISTIC (ROC) CURVE** This is a probability curve that plots the True Positive Rate (TPR) against False Positive Rate (FPR) at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC)is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

**SUMMARY AND NEXT STEP**

Stock market and investment has been one of the major investment strategies for wealth creation over the years. A lot of people put their money in stock across the globe for many numbers of reasons. The benefit of conducting enough research to identify the good stock to invest in can not be over emphasised as failure to do so will likely result in huge loss for the investor. Unfortunately researching about the financial market can be very tedious and time consuming. Leveraging the power of artificial intelligence in this area will be very helpful to all the stakeholders in different financial institutions across the globe.

In this project financial data was analysed and several algorithms trained to predict if a stock is worth buying or not. The models were also tested on a new dataset and their performances recorded and evaluated using the appropriate metrics. RandomForest gave a better result of based on the metrics used for the evaluation and hence it is recommended over other models.

My next step will be to implement and deploy the model so that it can be used on a live and current financial dataset for stockbrokers and other stakeholders to use it in their daily stock trade and other financial services.

**REFERENCE**

* [**https://www.investopedia.com/articles/investing/082614/how-stock-market-works.asp**](https://www.investopedia.com/articles/investing/082614/how-stock-market-works.asp)
* [**https://towardsdatascience.com/understanding-random-forest-58381e0602d2**](https://towardsdatascience.com/understanding-random-forest-58381e0602d2)
* [**https://www.datacamp.com/community/tutorials/xgboost-in-python**](https://www.datacamp.com/community/tutorials/xgboost-in-python)
* [**https://scikit-learn.org/stable/modules/svm.html**](https://scikit-learn.org/stable/modules/svm.html)
* [**https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/**](https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/)
* [**https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification**](https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification)

**Data Source:**

* [**https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018**](https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018)