INTERNATIONAL BURCH UNIVERSITY
FACULTY OF ENGINEERING AND NATURAL SCIENCES
DEPARTMENT OF INFORMATION TECHNOLOGIES

CEN 359 INTRODUCTION TO MACHINE LEARNING

STOCK PRICE PREDICTION

Prepared by:
Amina Mehić

Proposed to:
Alen Bošnjaković, Assist. Prof. Dr.
Dželila Mehanović, Teaching assistant

Sarajevo, June 2021

TABLE OF CONTENTS

ABSTRACT

Share market prices depend on various factors such as physical factors, psychological factors and global factors which cause volatility in prices and hence making it very difficult to predict the performance of the stock market with high accuracy. With machine learning getting applied to different domains, recently research has been carried out to predict the stock prices using machine learning algorithms. To implement this project, different NLP (Natural language processing) techniques in addition with machine learning algorithms will be applied on stock market data. Two types of ML model: Random Forest Classifier and Naive Bayes will be trained. Later, results of these two models will be compared as to which performed better. Random Forest classifier gave better performance with 85% score on the testing set and less number of misclassified samples as can be seen from confusion matrix in the later section in comparison to naive bayes classifier which gave 84% score on the testing set with more number of misclassified samples.

INTRODUCTION

Predicting stock market prices accurately is a very challenging task due to the volatile nature of the financial stock markets. With the introduction of machine learning and artificial intelligence, algorithms with increased performance and computation have been efficient in predicting stock prices. For this project, Naive Bayes and Random Forest techniques have been utilized for predicting whether the stock price is going to increase or not depending on the top 25 headlines of a newspaper. This is an example of a binary classification algorithm. The input to our machine learning algorithms is the top 25 news headlines data with labels of whether the stock price will increase or not next day. Trained ML models will be able to predict either label 0 or label 1. Target value of 0 indicates that the stock price is not going to increase the next day and target value of 1 indicates that the stock price is expected to increase the following day. This can be very helpful in stock markets to make timely entry and exit in a trade.

RELATED WORK

A Lot of work has been carried out previously to predict stock prices as part of financial analysis. [2] developed a model for stock market forecasting using News Headlines. They too worked on the same dataset where the target variable is a binary variable which takes value of 1, when the Dow Jones Indus-trial Average (DJIA) increases, and zero otherwise. Their main step was to pre-process daily news, to generate features for the prediction problem. They trained a deep learning model with Long Short-Term Memory (LSTM) architecture to improve the prediction of the changes in DJIA. [1] have utilized Artificial Neural Network and Random Forest techniques for predicting the next day closing price for five companies belonging to different sectors of operation. New variables were created using the given financial data: Open, High, Low and Close prices of stock. These new variables were used as inputs to the model which was later evaluated using standard strategic indicators: RMSE and MAPE. [3] is an article which gives an introduction to To Text Classification And Sentiment Analysis. Sentiment analysis, one of applications of text classification, is used to guess the positive or negative attitude of a user towards a topic given a sentence. Here our project is somewhat related to this, as depending on the headline, the price of stock will either be positively impacted or negatively impacted.

METHODS

There are broadly two types of machine learning algorithms classified as supervised and unsupervised machine learning algorithms. In supervised learning, we deal with the labelled data, that is the correct answer is already known whereas in unsupervised learning, we deal with unlabelled data and the goal is to find some similarities or differences in the data. Regression and classification constitutes 2 types of supervised learning algorithms. In regression, output variable is a real value whereas in classification, output variable is a category. For this project, since our goal is to predict whether the

price of stock will increase or not, we will be using classification algorithms, and that too binary classification for its implementation.

1. First algorithm is the Random **Forest Classifier.** Decision trees as these are the building blocks of random forest, therefore it becomes important to understand it before learning about random forest. Decision trees try to learn some rules using the features from the dataset and later use these decision rules to arrive at a decision and make a prediction for an input. In some scenarios, decision trees may lead to overfitting. In such cases, a random forest is used which is an ensemble and consists of a large number of decision trees. Majority vote from predictions from decision trees is used to get the final prediction. Below are the steps performed in the algorithm:
    1. Random samples are selected from the dataset.
    2. Decision tree is constructed for each sample and prediction is obtained from each decision tree.
    3. A vote for each predicted result.
    4. Prediction result with the most votes is selected as the final prediction.

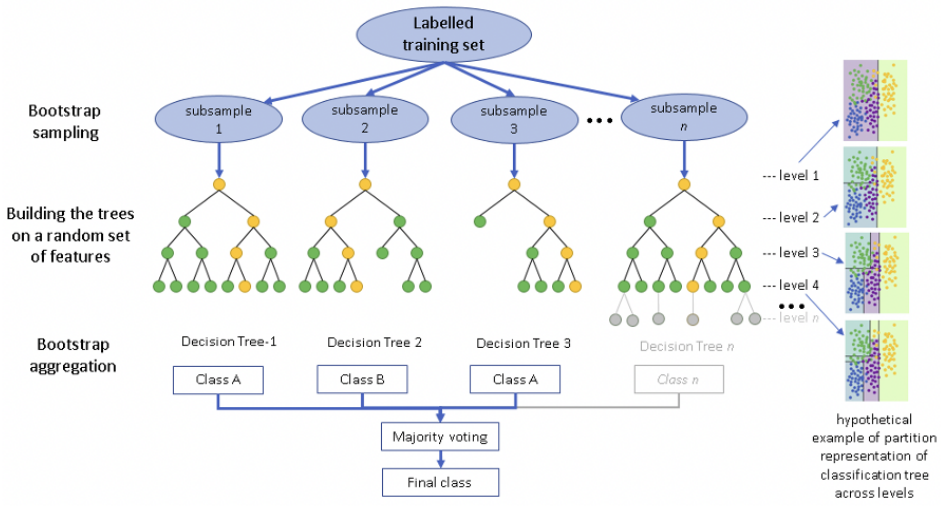


Fig. 1: Random Forest Classifier

2. Second algorithm used is **Naive Bayes Classifier.** Naive Bayes is a supervised machine learning algorithm and is based on the bayes theorem of probability. There is an underlying assumption associated with naive bayes algorithm that there exists an independence among predictors. In other words, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This algorithm is extensively used for larger datasets. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

where,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of the predictor given class.
- P(x) is the prior probability of the predictor.

  There are three types of Naive Bayes model under the scikit-learn library: gaussian, multinomial and bernoulli. Multinomial Naive Bayes is most commonly used for text classification, so even for this project, we will be using a Multinomial Naive Bayes classifier.

DATASET

We use a dataset containing news headlines and stock market index available on Kaggle [4]. There are two channels of data provided in this dataset:

- News data: - historical news headlines are crawled from Reddit WorldNews Channel. They are then ranked by reddit users' votes, and only the top 25 headlines are considered for a single date. This data is considered from 8th August, 2008 to 1st June, 2016.
- Stocks data: Dow Jones Industrial Average (DJIA) is used to "prove the concept" and is present in between the range 2008-08-08 to 2016-07-01.

There are three data files in .csv format:

- RedditNews.csv: consisting of two columns. The first column is the "date", and the second column is the "news headlines". All news is ranked from top to bottom based on how hot they are. Hence, there are 25 lines for each date.
- DJIA_table.csv: Downloaded directly from Yahoo Finance
- CombinedNewsDJIA.csv: consisting of 27 columns. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25". It contains the top 25 daily news headlines every sin-gle weekday starting from August 8, 2008, all the way till June 2016.

In our project implementation, we will be working with CombinedNewsDJIA.csv dataset to predict whether the stock price is going to increase the next day or not given the top 25 headlines of the day. Dataset was splitted into train and test data using 80%-20% ratio. Training data consisted of samples from 2008-08-08 to 2014-12-31 and Test Set consisted of samples from 2015-01-02 to 2016-07-01. Since null values were present in the dataset, after removing rows with null values, training data consisted of 1860 samples and testing set consisted of 378 samples.

Snippet of dataset:

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top16	Top17	To
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatent' trials are nothing but...	...	b'Georgia Invades South Ossetia - if Russia ge...	b'Al-Qaeda Faces Islamist Backlash'	b'Condolee Rice: "The would no to
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	...	b'Israel and the US behind the Georgian aggres...	b'"Do not believe TV, neither Russian nor Geor...	b'Riots are going c Mont (Canad
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b"If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...	b'Stratfor: The Russo-Georgian War and the Bal...	b"I'm Trying to Get a Sense of This Whole Geor...	...	b'U.S. troops still in Georgia (did you know t...	b'Why Russias response to Georgia was right'	b'Gorbac accuses of maki "seriou
3	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	b"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain\'s policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 *million* quake survivors...	b"Bush announces Operation Get All Up In Russi...	b'Russian forces sink Georgian ships '	...	b'Elephants extinct by 2020?'	b'US humanitarian missions soon in Georgia - i...	b"Georg DDOS c from soun

Fig. 2: Snippet of the dataset

As part of more preprocessing, text headlines were first cleaned to remove everything else so that it contains only alphabets and then 25 headlines were merged into one. Later, text was converted to lowercase and after splitting the text to get words, stopwords were removed from the dataset. Lemmatization was performed to find the base or dictionary word of a word depending on its meaning and context. These lemmatizated collection was then converted to a vector of token counts using CountVectorizer to get the final training and testing sets for model training.

TOOLS

This project has been implemented in the python programming language. Python3 version was installed and a Jupyter notebook was installed using “pip3 install jupyter notebook”. Jupyter notebook makes it easy to write the code with text and equations in between using the code and markdown cells respectively. Following libraries are used for implementation of this project:

- Pandas: for reading the csv in the form of dataframe
- Sklearn: for machine learning algorithms
    - Random Forest Classifier from sklearn.ensemble.RandomForestClassifier
    - Multinomial Naive Bayes classifier from from sklearn.naive_bayes import MultinomialNB
    - Lemmatization using from nltk.stem import WordNetLemmatizer
    - Count Vectorization using from sklearn.feature_extraction.text import CountVectorizer
    - Metrics (Accuracy score, classification report and confusion matrix) from sklearn.metrics
    - Hyperparameter tuning using sklearn.model_selection.GridSearchCV
- Matplotlib: used for plotting
- Seaborn: a graphics library

RESULTS AND DISCUSSION

We performed classification using 2 algorithms: Random Forest Classifier and Naive Bayes. Evaluation metrics such as accuracy score (on both training and testing data), classification report listing precision, recall and f1 score and confusion matrix were used for both the algorithms.

For Random Forest Classifier:

- With a random forest classifier, we got a score of 100% on the training and 85% on the testing set.

```
#Training and testing accuracy
print("Training Accuracy:",accuracy_score(df_train['Label'], y_train_pred)*100)
print("Testing Accuracy:",accuracy_score(df_test['Label'], y_test_pred)*100)
```

```
Training Accuracy: 100.0
Testing Accuracy: 84.92063492063492
```

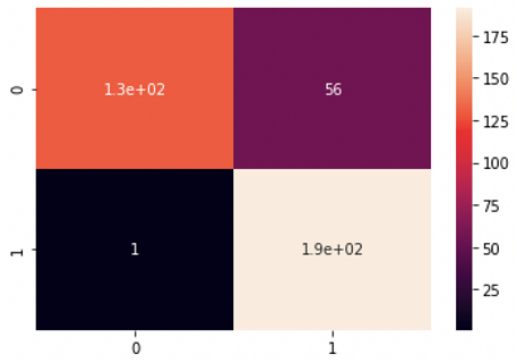
- Classification Report: 0.85 score for weighted precision, recall and f1 score

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.70      0.82       186
           1       0.77      0.99      0.87       192

    accuracy                           0.85       378
   macro avg       0.88      0.85      0.85       378
weighted avg       0.88      0.85      0.85       378
```

- Confusion Matrix:



For Naive Bayes Classifier:

- With a naive bayes classifier, we got a score of 100% on the training and 84% on the testing set.

```
Training Accuracy: 100.0
Testing Accuracy: 84.12698412698413
```

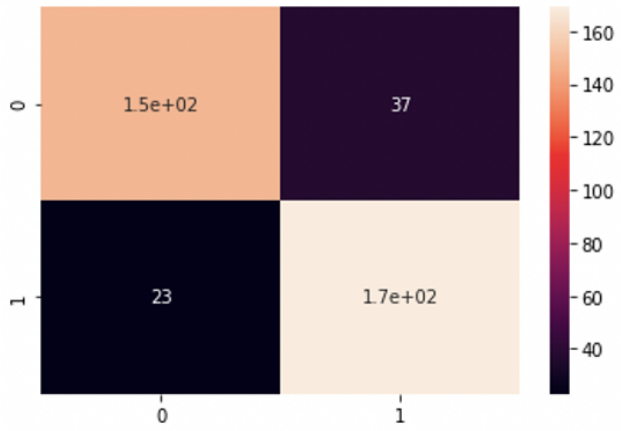
- Classification Report: 0.84 score for precision, recall and f1 score

```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.80      0.83       186
           1       0.82      0.88      0.85       192

    accuracy                           0.84       378
   macro avg       0.84      0.84      0.84       378
weighted avg       0.84      0.84      0.84       378
```

- Confusion Matrix:



CONCLUSION/FUTURE WORK

For this project, we used two machine learning classification algorithms named Naive Bayes classifier and Random Forest Classifier. After importing the dataset in pandas dataframe using pandas library, exploratory data analysis was performed to get better insights about the data. Later, preprocessing steps like removing null values, cleaning the text so that it contains only alphabets, removing stop words, performing lemmatization and doing count vectorization . Dataset was splitted in a ratio of 80-20% for the train and test split. We observed that compared to random forest classifiers with default parameters, multinomial naive bayes with default parameters gave less score on the testing set. Random forest classifier gave a score of 85% on the testing set which is better as compared to 84% given by naive bayes classifier.

For Future steps we would like to try more classification algorithms using embedding, tf-id vectorizers, N gram modelling, LSTM's and other ensemble methods. In addition to these, we would like to do some more EDA like word cloud etc.

REFERENCES

[1] Vijh, M., Chandola, D., Tikkiwal, V.A. and Kumar, A., 2020. Stock closing price prediction using machine learning techniques. Procedia Computer Science, 167, pp.599-606

[2] Hassanzadeh Kalshani, Ali & Razavi, Ahmad & Asadi, Reza. (2020). Stock Market Prediction using Daily News Headlines. SSRN Electronic Journal. 10.2139/ssrn.3685530.

[3] Miguel González-Fierro. Jan 31, 2017. A Gentle Introduction To Text Classification And Sentiment
Analysis. Retrieved from publication: https://miguelgfierro.com/blog/2017/a-gentleintroduction-to-text-classification-and-sentiment-analysis/

[4] Sun, J. (2016, August). Daily News for Stock Market Prediction, Version 1. Retrieved [June 6, 2021] from https://www.kaggle.com/aaron7sun/stocknews