ASSESSMENT NUMBER: 02.

ASSESSMENT TITLE: CASE STUDY OF STOCK MARKET PREDICTION.

STUDENT NAME:

\_\_\_\_\_\_\_\_\_\_\_\_

STUDENT ID:

\_\_\_\_\_\_\_\_\_\_\_\_\_

STUDENT EMAIL:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **ABSTRACT:**

A stock market is a place where shares or ownership of a company are sold at some value fixed by the company owner. A stock exchange is a source to trade company stocks or a meeting place for the seller or buyer. A stock may be bought and sold many times and only if its listed on an exchange. Stock markets' high risk and high returns have attracted many people. Predicting stock market performance is a difficult task. we can use data science tools and machine learning models to predict its outcome with some good accuracy. Machine learning and data science techniques have the daring potential to decode patterns and insights we don't see on the first look, and these can be used to make amazingly trustworthy predictions.

1. **TOPIC AND BACKGROUND:**

The intention is to predict stock prices using information from a good dataset. Nowadays stock prediction is one of the most significant challenges because its data set is a financial time series data that becomes more difficult to predict due to the dynamic nature of it. There are so many factors involved in the prediction like physical, physiological, rational, and irrational behavior, etc. Stock prediction made by real analysts is made by using mathematical techniques of dividend yield, Insider Moves, interest rates any many more. Machine learning can prove to be a game-changer for the automation of this task we can pick imminent features from the data set and focus on its pattern and generalize on the basis of our learning. Machine learning is based on deductive reasoning in which first we learn on some data and then we generalize on other data which we have not seen. There are many machine learning algorithms that can help us in prediction tasks. We use different training and testing splitting techniques, in the training set we run the learning algorithm to make it learn the dataset, and the testing set is used to test the learning abilities of the algorithm and we can check the accuracy by comparing the actual result and the predicted result. This automated learning process is clearly making its mark in this current century.

1. **PROBLEM STATEMENT:**

We are intending to predict stock prices using the data set from [QUANDL](https://www.quandl.com/). The data set used is [TATA GLOBAL BEVERAGES](https://www.quandl.com/data/NSE/TATAGLOBAL-Tata-Global-Beverages-Limited) dataset. As shown:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DATA | OPEN | HIGH | LOW | LAST | CLOSE | TOTAL TRADE QUANTITY | TURNOVER IN HUNDRED THOUSANDS |
| |  | | --- | | 10/8/2018 | | 208 | 222.25 | 206.85 | 216 | 215.15 | 4642146 | 10062.83 |
| |  | | --- | | 10/5/2018 | | 217 | 218.6 | 205.9 | 210 | 209.2 | 3519515 | 7407.06 |
| |  | | --- | | 10/4/2018 | | 223.5 | 227.8 | 216.15 | 217.25 | 218.2 | 1728786 | 3815.79 |
| |  | | --- | | 10/3/2018 | | 230 | 237.5 | 225.75 | 226.45 | 227.6 | 1708590 | 3960.27 |
| |  | | --- | | 10/1/2018 | | 234.55 | 234.6 | 221.05 | 230.3 | 230.9 | 1534749 | 3486.05 |
| 9/28/2018 | 234.05 | 235.95 | 230.2 | 233.5 | 233.75 | 3069914 | 7162.35 |
| 9/27/2018 | 234.55 | 236.8 | 231.1 | 233.8 | 233.25 | 5082859 | 11859.95 |

We are intending to predict the target variable of “CLOSE” which is the closing price of the stock. By using different machine learning and data science techniques to train the machine learning model by using different features of the dataset these features are explained in the data set heading. We may be using one or two algorithms and then we will be picking up the best one to test on the test set of the data. The results will be compared with the actual data and the accuracies will be measured.

1. **AIMS AND OBJECTIVES:**

The aim is to predict the stock prices of a certain data set by making this possible. We will be capable to automate the tiring process of serious analysis. Our aim is to predict the closing price variable of the TATA GLOBAL BEVERAGES dataset by using machine learning algorithms to train the data and predict the prices with trustworthy accuracy .the skill which we will be developing will help us in understanding the data exploration by plotting different charts using python libraries of matplot or seaborn. This graphic exploration will help us in understanding the trends and patterns of the data set and will help us build our skill of understanding the patterns from visualizations and exploratory data analysis. We will be able to model the data by using data analysis algorithms. It helps us understand there working and also there use cases

1. **DATASET USED:**

The data set used is from [QUANDL](https://www.quandl.com/) and we are focusing on [TATA GLOBAL BEVERAGES](https://www.quandl.com/data/NSE/TATAGLOBAL-Tata-Global-Beverages-Limited) dataset. As shown:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DATE | OPEN | HIGH | LOW | LAST | CLOSE | TOTAL TRADE QUANTITY | TURNOVER IN HUNDRED THOUSANDS |
| |  | | --- | | 10/8/2018 | | 208 | 222.25 | 206.85 | 216 | 215.15 | 4642146 | 10062.83 |
| |  | | --- | | 10/5/2018 | | 217 | 218.6 | 205.9 | 210 | 209.2 | 3519515 | 7407.06 |
| |  | | --- | | 10/4/2018 | | 223.5 | 227.8 | 216.15 | 217.25 | 218.2 | 1728786 | 3815.79 |
| |  | | --- | | 10/3/2018 | | 230 | 237.5 | 225.75 | 226.45 | 227.6 | 1708590 | 3960.27 |
| |  | | --- | | 10/1/2018 | | 234.55 | 234.6 | 221.05 | 230.3 | 230.9 | 1534749 | 3486.05 |
| 9/28/2018 | 234.05 | 235.95 | 230.2 | 233.5 | 233.75 | 3069914 | 7162.35 |
| 9/27/2018 | 234.55 | 236.8 | 231.1 | 233.8 | 233.25 | 5082859 | 11859.95 |

* **OPEN:** The starting price of the stock on a particular day.
* **CLOSE:** The closing price of the stock on a particular day.
* **HIGH:** The maximum price of the share for the day.
* **LOW:** The minimum price of the share for the day.
* **LAST:** The last price of the share for the day.
* **TOTAL TRADE QUANTITY:** The number of share bought or sold in that day.
* **TURN OVER IN HUNDRED THOUSAND:** The turnover for the day.

**DATA GATHERING:**

Thankfully the data set is publically available from the website QUANDL. And we are focusing on TATA GLOBAL BEVERAGES dataset. We are simply downloading it and reading it as a CSV (comma-separated values) file in our python jupyter notebook using the pandas library.

**DATA PREPARATION:**

We will deal with missing values by using pandas function of draw missing data table which tells about missing values we will be filling them with the fillna function which replaces the missing values by the mean and median of the column.

There is no need for feature engineering methods because we will only be using the features provided no need to create new features.

There is no categorical variable handling needed as all the columns are numerical already.

As we can see that our dataset does not need much preparation. So we mainly focus on its patterns by using exploratory data analysis.

**DATA EXPLORATION:**

For exploring each columns data visually we will be using matplot or seaborn library. We will be using histograms to check distributions of values in each columns. We will be using dot plot functions to check the time series data pattern. To check the modeling result we will be using dot plot function to plot the time series data and compare the result visually.

1. **METHODS AND TECHNIQUES TO BE USED:**

Here we will be using the real data science algorithms for modeling the data set and finally training it and using it to generalize on other test cases.

**DATA MODELLING:**

Firstly we will be using the linear regression statistical method in which we will be using an “X” column from the data set and a “Y” column from the data set which will be our target variable which we will predict. The linear regression method fits an equation to the dataset then this equation is used to predict the “Y” variable. We calculate the slope and then the y-intercept. The slope tells us how much “Y” changes by a unit increase in “X” and the y-intercept tells us the average value of “Y” when “X” is zero. The equation looks like this:

**Yi = b0 + b1Xi.**

The slope is b0 and intercept is the b1. We determine the constants and then use different “X” values to calculate “Y” which is the predicted value.

Secondly, we will be using the K-Nearest Neighbor algorithm to predict the target variable. Now the KNN algorithm memorizes the data set and makes the calculation at the test time. When it gets the test case it takes the distance of this test case from all of the target variable’s entries then sets out particular close distant values. The distance formula used is Euclidean distance.



The number of closely distant values is determined by the value of “K” which is determined randomly or by testing out different values of “K” and coming up with the best one. So after setting out “K” nearest distant target value entries we see to which class most of the nearest values corresponds. Whichever class has the most values that class is selected for the test case and the result is predicted.

So we will mainly be focusing on these two algorithms to predict the closing price of the stock. Then we will be comparing their accuracies to come up with the best model. We may also look over some more algorithms if we had time.

**PRESENTATION AND AUTOMATION:**

We may be data science tools such as google colabaratory or jupyter notebook. We will be labeling each step with mark up headings. Proper and beautiful EDA (Exploratory Data Analysis) will be displayed. The modeling results will be compared and the best one will be picked for further testing.

1. **ETHICAL ISSUES:**

The automated data science predictive analyses are prone to some ethical issues such as the author or the owner of this system may gain unauthorized access to important or private data of some company or individuals of the company. There is the issue of integrity as well. As for the design of these predictive systems or algorithms, there is a need for a good amount of data. Sometimes the data is not available so we may use different ways to get data like web scraping or anything as such which may be a breach of the integrity of some company. One issue with these predictive systems that the algorithms used in these systems may have many unintended and unexplained effects. In other words, these predictive analyses only give a limited idea to the users as to why a certain decision was taken which does not prove the justification or the legitimateness of the decision. One great issue of these artificially intelligent systems is the problem of the bias influence which may lead to unfair or inaccurate predictions which causes a lack of reliability in the system. For example, a low profile company is selling its share very regularly and a high profile company that sells its share rarely the demand is of the high profile company shares but due to their lack of participation, the predictive systems may decide to say that the low profile company has more demand. So there is a bias in the data set. These biases are removed by many difficult techniques of machine learning. All of these ethical issues remain in these kinds of systems. As these systems only provide the hint of a certain decision it is mainly up to us how we use that information for our benefit.

1. **REFERENCES**
2. Ethical Implications of Predictive Risk Intelligence | Sherpa Project. Sherpa Project. (2020). Retrieved 31 October 2020, from <https://www.project-sherpa.eu/ethical-implications-of-predictive-risk-intelligence/>.
3. YOUTUBE. (2020). Retrieved 31 October 2020, from <https://www.youtube.com/watch?v=6kZ-OPLNcgE&t=713s>.
4. GLOBAL BEVERAGE, T. (2020). *Quandl*. Quandl.com. Retrieved 31 October 2020, from <https://www.quandl.com/data/NSE/TATAGLOBAL-Tata-Global-Beverages-Limited>.
5. **GLOSSARY:**

|  |  |
| --- | --- |
| STOCK MARKET | It is a collection of markets where shares are bought, sold, or issued regularly by many companies. |
| MACHINE LEARNING | It is the domain of artificial intelligence which provides the ability to automatically learn and improve by looking at some examples and no explicit programming is needed |
| DEDUCTIVE REASONING | The method of reasoning in which similar cases are looked over and learned and then generalized on unseen but similar cases |
| ALGORITHMS | Sequence of steps to solve a particular problem |
| TRAINING SET | The data used by the algorithm to learn about the problem. |
| TESTING SET | The data used to test the learning abilities of an algorithm |
| MACHINE LEARNING MODELS | It is an algorithm that is trained to recognize certain types of patterns. |
| MATPLOT LIBRARY | Open source programming language of python has the inbuilt implementation of making graphs by providing data sets. |
| SEABORN LIBRARY | Open source programming language of python has the inbuilt implementation of making graphs by providing data sets. |
| EXPLORATORY DATA ANALYSIS | It is used for visualizing data set by making certain graphs to understand the pattern of the data set. |
| FEATURE ENGINEERING | Make more columns in the data set by using the existing columns |
| CATEGORICAL VARIABLES | The features which belong to a certain class |