# GP predictions

*Rachid elkhayat*

*6/13/2020*

## The Outline

**1. Introduction**

-The project goal
-The key steps implemented

**2. Method and Analysis**

-Data preparation
-Data exploration and visualization
-Feature Engineering
-Creating function to process the predictions
-Creating train/test datasets
-Creating Our models

**3. The results**

-Visualizing in a table the accuracy of our models
-Plotting the cummlative returns

**4. Conclusion**

---

## 1. Introduction

Gold has been always viewed as the symbol of wealth since the ancient times of the human history, The demand of gold is raising day by day, Its price is affected by many factors which makes it hard to be predicted. Various studies have developed different predictive models based on different techniques and factors. Some studies try to make predictions based on historical prices while others has a different approach by explaining the correlations between the prices of gold with respect to a various of economic factors.

For a long period of time, the price of gold was fixed. After 1968, the price of gold began to be determined by the market. Gold trading has always been considered attractive due to its historic volatility. On the trading platforms the candles represent the open and close price in within the same day. Different people use different strategies in trading, usually individuals open and close positions within day or days. However, the bigger companies positions can be opened for months or years. So we can say that the duration of trade positions can vary from one day to years period, In our project here we will build a model that will aim to predict the next day position based on the previous data.

Defining The Candle Types: Candle types represents the difference between the open and close price, while the length of the candle represent the difference in the value between both positions. Candles help us to understand easly the fluctuation in the price and they are represenetd in three types:-

"Green candles" are the candles that represent the "bull" where the closing price is higher than the opening price. "Red candles" are the candles that represent the "bear" where the closing price is lower than the opening price. "Doji Candles" are the candles that don't have a candle body however they have an indicative wicks that can help in predictions.

Every candle we will see represents a day with its opening and closing price. For every day we have the following information: -Open price -High price -Low price -Close price

### The project goal

We will implement multiple methods to predict the "next day candle type" using differenet Machine learning models, We will train our models based on previous data that we will download from yahoo finance. Our goal is to achieve an accuracy over 50% based on which we will determine predictive capability of our models.

### The ket steps implement

```
 - Data exploratory and Analysis
 - Modeling and testing
  Naive bayes Model
  Support vector Model
  Random Forest Model
 - Models performance and accuracy
```

---

## 2. Method and Analysis**

### Data preparation

The data that we are working on in this project is a part of the dataset that we have downloaded from yahoo finanace, we decided to start the predictions starting from 2018 till May2020.

After installing the required packages and adding the libraries, we will manipulate the downloaded dataset to transform it to the format that will make it easy for us to work with.

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 3369 obs. of  5 variables:
##  $ Index     : Date, format: "2007-01-02" "2007-01-03" ...
##  $ GC=F.Open : num  635 640 624 624 608 ...
##  $ GC=F.High : num  635 640 624 624 608 ...
##  $ GC=F.Low  : num  635 628 624 617 608 ...
##  $ GC=F.Close: num  635 627 624 605 608 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Index = col_date(format = ""),
##   ..   `GC=F.Open` = col_double(),
##   ..   `GC=F.High` = col_double(),
##   ..   `GC=F.Low` = col_double(),
##   ..   `GC=F.Close` = col_double()
##   .. )
```


## Data exploration and visualization

Since 2018 the price of the gold is moving in an upward trend, with a huge plunge in march 2020, the plunge has been followed with a surge in the same month.

The Green candles are called "bull candles" and it represent the increase in the price, where the number of buyers are more than the number of sellers. while the orange candles are called "bear candles" which represent the descrease in the price, where the number of sellers are more than the number of buyers. while Doji, is a name for a session in which the candlestick has an open and close price that are virtually equal.

**Exponential moving average (EMA)** An exponential moving average (EMA) is a type of moving average (MA) that places a greater weight and significance on the most recent data points. It is a technical indicator that is used to produce buy and sell signals based on crossovers and divergences from the historical average. We can use several different EMA lengths of moving averages, in our project we will use 7-days and 20 days.

EMA gives more weight to recent prices and are calculated by applying a percentage of the current day closing price to the previous day(s) moving average.

The EMA is calculated in three steps: 1- Compute the SMA - simply the mean of closing price for the selected period 2- Calculate the multiplier for weighting the EMA- Equals to 2/timeperiod +1 3- Calculate the current EMA

$$EMA = (Close - previousEMA) * (2/WeightingMultiplier + 1) + previousEMA$$

In the below graphs we can clearly see the relation beween the price and the indicators, both indicators are moving with the price. EMA as one of the main indicator for predicting the price as it is more reactive to the price change compared to SMA. When the market is in a strong and sustained uptrend, the EMA indicator line will also show an uptrend and vice-versa.

The gold price started at around 1300 USD in 2018 and moved up the next two years to reach a value of more than 1700 USD, one of the most obvious observation that we can identify is the major drop in 2020 which was quickly recovered and the price continue its increase afterwards.

Plotting a subset of the data starting from 2020 in order to visualize the EMAs and track their performance.

```
#Plotting a subset of the data starting from 2020 in order to visualize the EMAs.
min2020<- min(gold_xts['2020']$Close)
max2020<- max(gold_xts['2020']$Close)
#Plotting the EMAs
chartSeries(gold_xts, subset = "2020-01::",theme = "white")
```


```
addEMA( n = 7, col = "purple")
```


```
addEMA( n = 20, col = "red")
```


the graphs above shows that the EMA7 line is more reactive to the changes compared with the EMA20. as it accomodate the changes faster and gives that a better indication for the short term trade.

### Feature Engineering

We will use feature engineering to extract features from data that will help us entrepret, discover new findings, create the data that would be useful for our machine learning models.

Our main goal in this project is to predict whether the next day candle is bull/bear,for that purpose we will create the below features in a separate data frame that will help us with our predictions:

- Current candle type - the candle type for the current day
- Candle previous day - the candle type of the previous day
- Doji - a type of candles with minimum movement in the price within the day(neutral)
- Position of close price to ema7 -(above/below)
- Position of close price to ema20 -(above/below)
- Candle next day - our main target (model predictions)
- Current return (close price - open price)
- Next day return (-/+)

```
## 'data.frame':    599 obs. of  9 variables:
##  $ CurrentCandle  : Factor w/ 2 levels "bear","bull": 2 1 2 1 1 1 1 2 2 2 ...
##  $ PreviousCandle : Factor w/ 2 levels "bear","bull": 2 2 1 2 1 1 1 1 2 2 ...
##  $ Doji           : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 2 1 1 ...
##  $ PositionToEma7 : Factor w/ 2 levels "above","below": 1 2 2 2 2 2 2 1 1 1 ...
##  $ PositionToEma20: Factor w/ 2 levels "above","below": 1 1 1 2 2 2 2 2 1 1 ...
##  $ Ema7ToEma20_   : Factor w/ 2 levels "above","below": 1 1 1 1 2 2 2 2 2 1 ...
##  $ DailyReturn    : num  0.5 14.8 1.9 12.1 13.8 ...
```

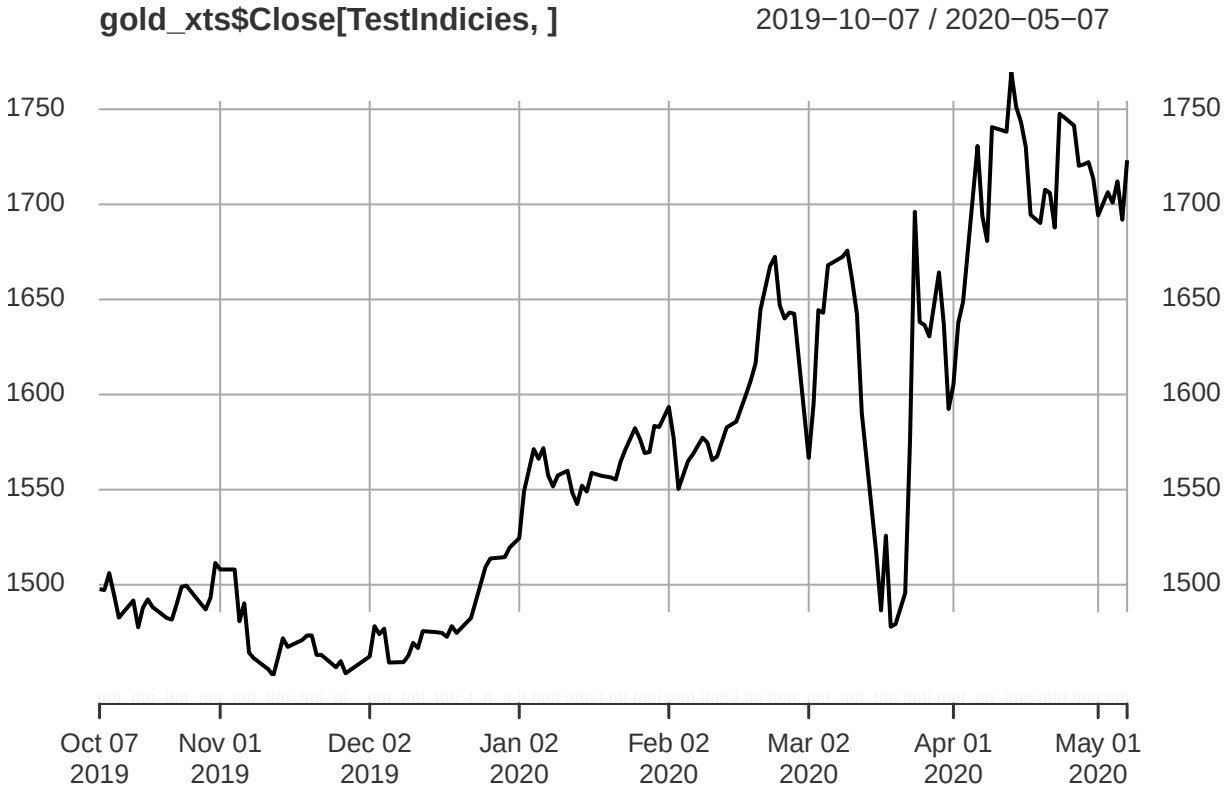
```
##  $ NextDayReturn  : num  14.8 1.9 12.1 13.8 0.2 ...
##  $ NextDayCandle  : Factor w/ 2 levels "bear","bull": 1 2 1 1 1 1 2 2 2 2 ...
```

## Creating train/test datasets

Splitting the data in to train and test datasets, Test set indicies 30%,#Training set 70%, we need to keep in mind that the data when dealing with time series cannot be splitted in an arbitrary way. we need to take the chronological order into consideration.

```
##  CurrentCandle PreviousCandle  Doji     PositionToEma7 PositionToEma20
##  bear:253      bear:252       no :328   above:224      above:221
##  bull:166      bull:167       yes: 91   below:195      below:198
##
##
##
##
##  Ema7ToEma20_  DailyReturn      NextDayReturn     NextDayCandle
##  above:209    Min.   : 0.0000   Min.   : 0.0000   bear:253
##  below:210    1st Qu.: 0.6999   1st Qu.: 0.6999   bull:166
##               Median : 3.1000   Median : 3.1000
##               Mean   : 4.9938   Mean   : 5.0112
##               3rd Qu.: 6.8000   3rd Qu.: 6.8000
##               Max.   :35.6000   Max.   :35.6000
```

```
##  CurrentCandle PreviousCandle  Doji     PositionToEma7 PositionToEma20
##  bear:77       bear:77        no :151   above:98       above:106
##  bull:83       bull:83        yes:  9   below:62       below: 54
##
##
##
##
##  Ema7ToEma20_  DailyReturn     NextDayReturn     NextDayCandle
##  above:108    Min.   :  0.00   Min.   :  0.000   bear:77
##  below: 52    1st Qu.:  2.90   1st Qu.:  2.875   bull:83
##               Median :  6.90   Median :  6.800
##               Mean   : 10.98   Mean   : 10.947
##               3rd Qu.: 13.05   3rd Qu.: 13.050
##               Max.   :105.20   Max.   :105.200
```


### Creating function to process the predictions

Creating a function for processing predictions, With this function we will create three features (pred , prediReturn, cumReturn) that we will include in the new dataframe of the models.

The three feature are: pred: Our model predictions predireturn: to determine wether our predictions lead to positive or negative returns. cum return: to calculate the cummulative returns for each model.

```
predictedReturn <- function(df, pred) {
  #pred is our predictions from the machine learning model
  df$pred <- pred
  # transforming to negative value if our prediction is wrong
  df$prediReturn <- ifelse(df$NextDayCandle != df$pred, -df$NextDayReturn, df$NextDayReturn)
  # calculating the cummulative sum
  df$cumReturn <- cumsum(df$prediReturn)
  return(df)
}
```

## Creating Our models

Before we start creating the Model we will define the Accuracy that we will take it as reference to evaluate the different models performance.

- Model Performance and Accuracy - We will be assessing our data buy the percentage of accuracy that results from each model, the accuracy will be calculated by obtaining the error and subtracting the outcome by one. below is the Accuracy equation:

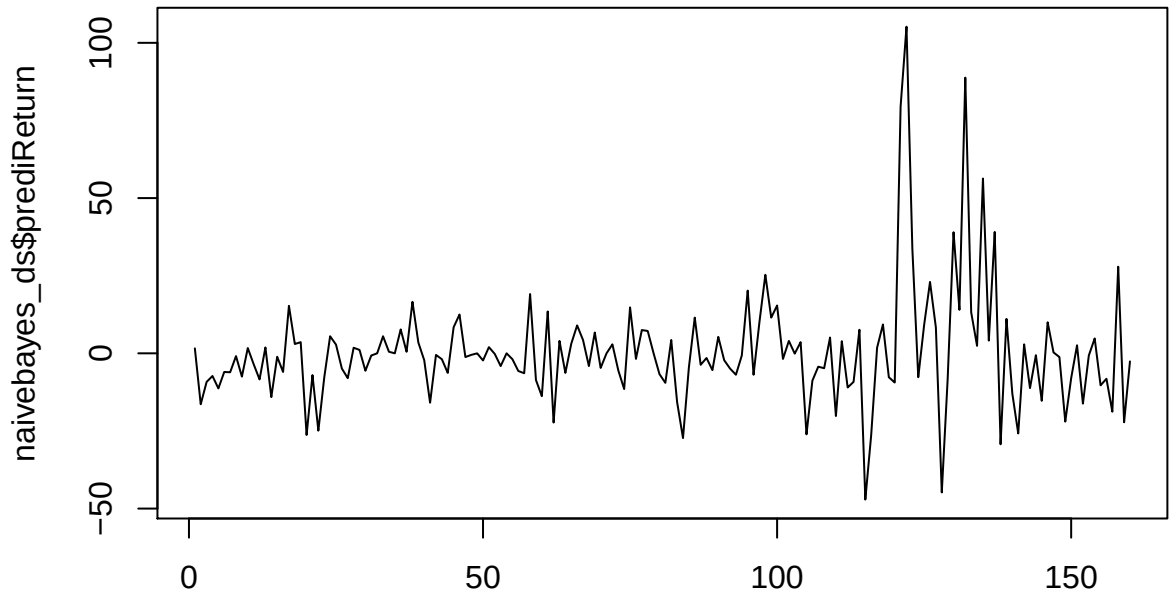
$$Accuracy = 1 - ModelError$$

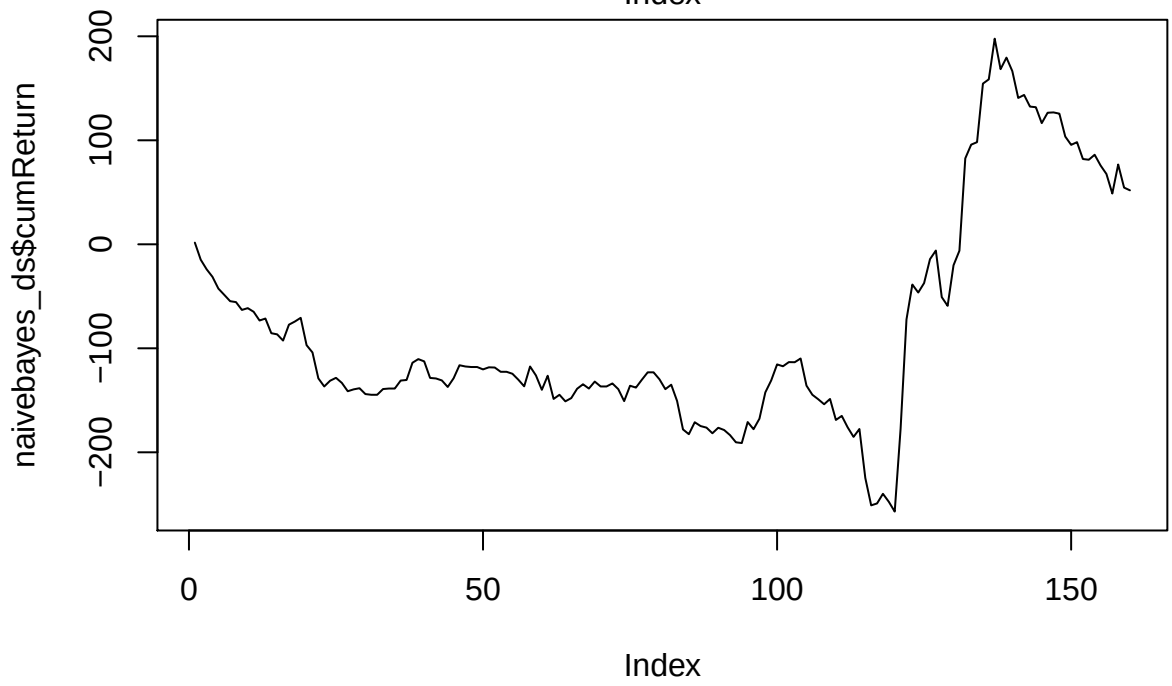
### Model 1-Naive Bayes Model-

In brief, The principle behind Naive Bayes is the Bayes theorem. It is used to calculate the conditional probability, which is the probability of an event occurring based on information about the events in the past. Mathematically, the Bayes theorem is represented as:

$$P(A_i|B) = P(B|A_i)P(A_i)/P(B)$$

$P(A|B)$ is the posterior probability of C given B $P(A)$ is the prior probability of class $P(B|A)$ is the likelihood which is the probability of predictor given class. $P(B)$ is the prior probability of the predictor



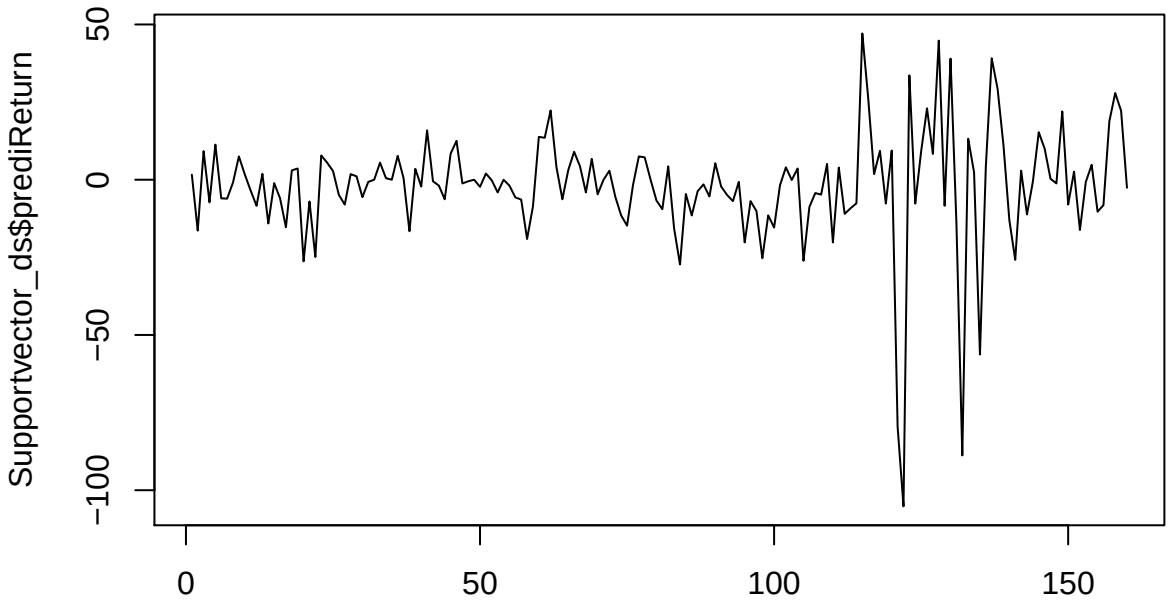



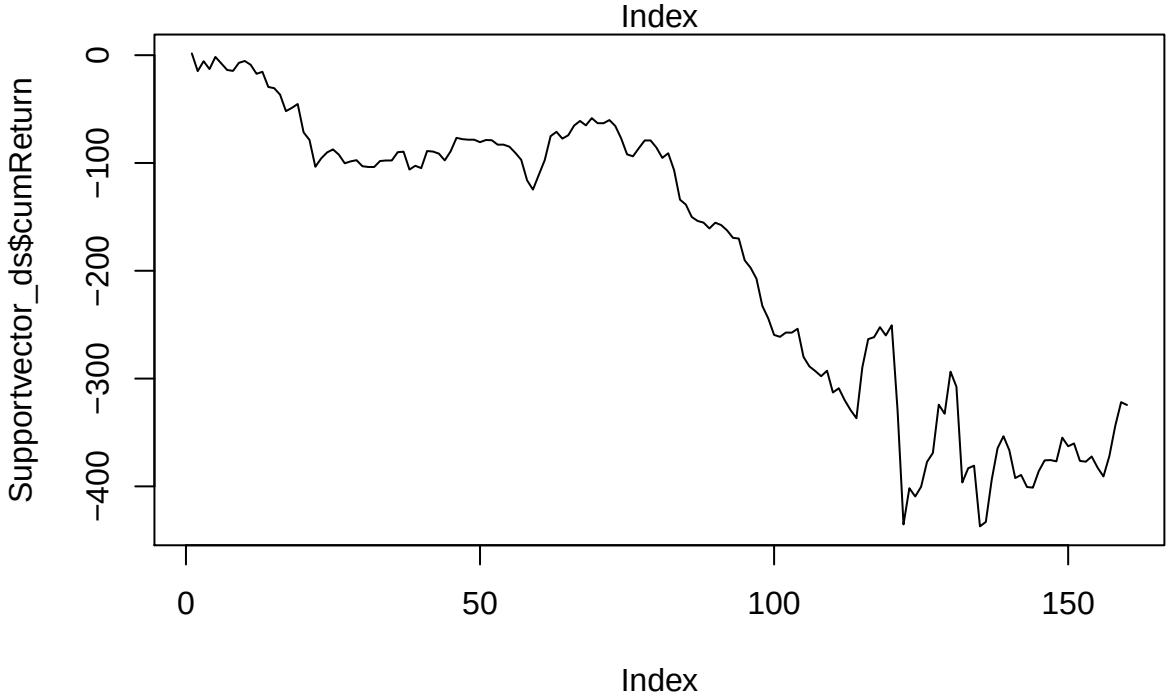

| | bear | bull |
|---|---|---|
| bear | 49 | 28 |
| bull | 60 | 23 |

```
## [1] 0.45
```

The accuracy value shows that our predictions are not reliable enough, we also could see that by looking at the cumulated returns graph, the big plungs has affected our prediction as it makes it very challenging for the model to predict the right candle type. We will try to create a model that can accomodate the changes in a better way and result in better outcome.

**Model 2 -Support Vector Machine-** Ee will use SVM to classify our data points into either bull/ bear, We will need to find the Hyperplane or in other words the line between the two classes, the hyperplane is an (n minus 1)-dimensional subspace for an n-dimensional space The Hyperplane separates the features that share the same property, In our case we will predict the features wethers it is bull/bear.



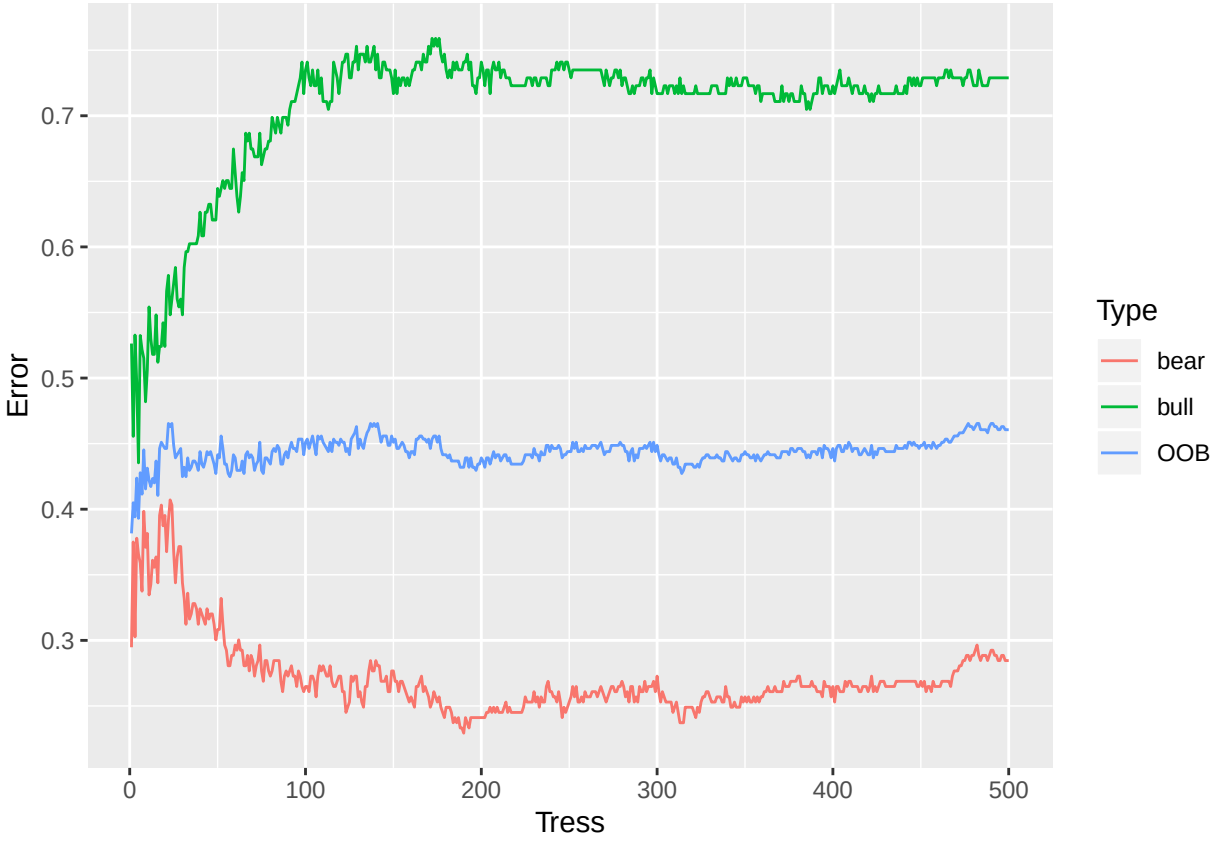

| | bear | bull |
|---|---|---|
| bear | 65 | 12 |
| bull | 76 | 7 |

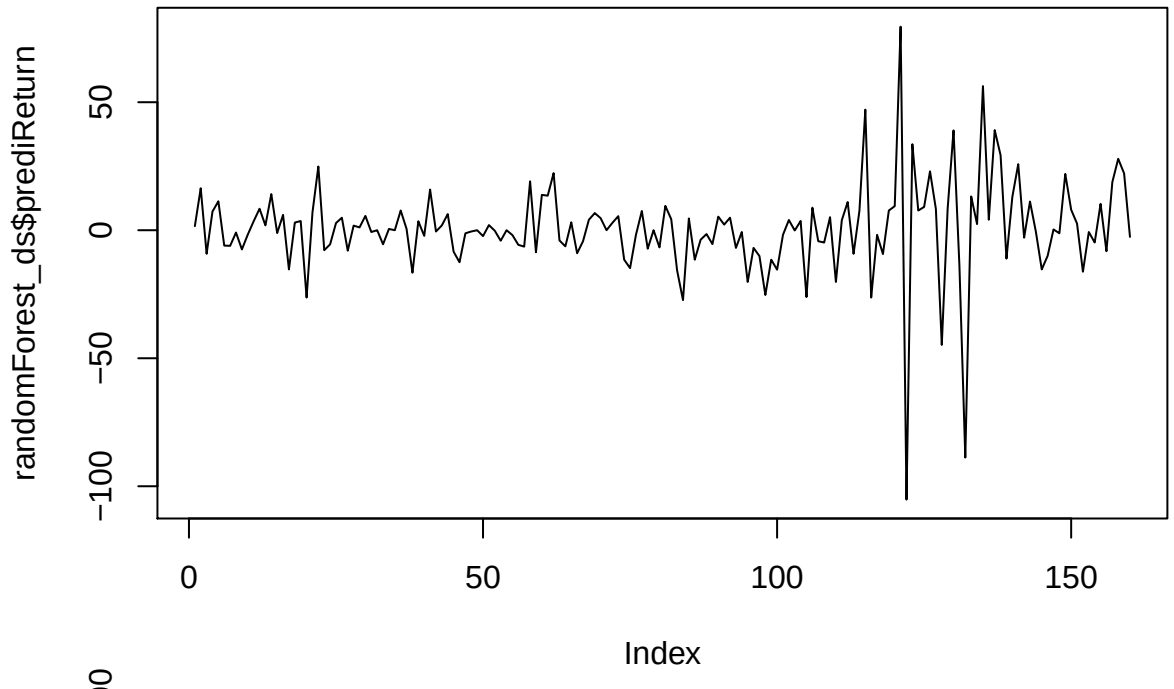
```
## [1] 0.45
```

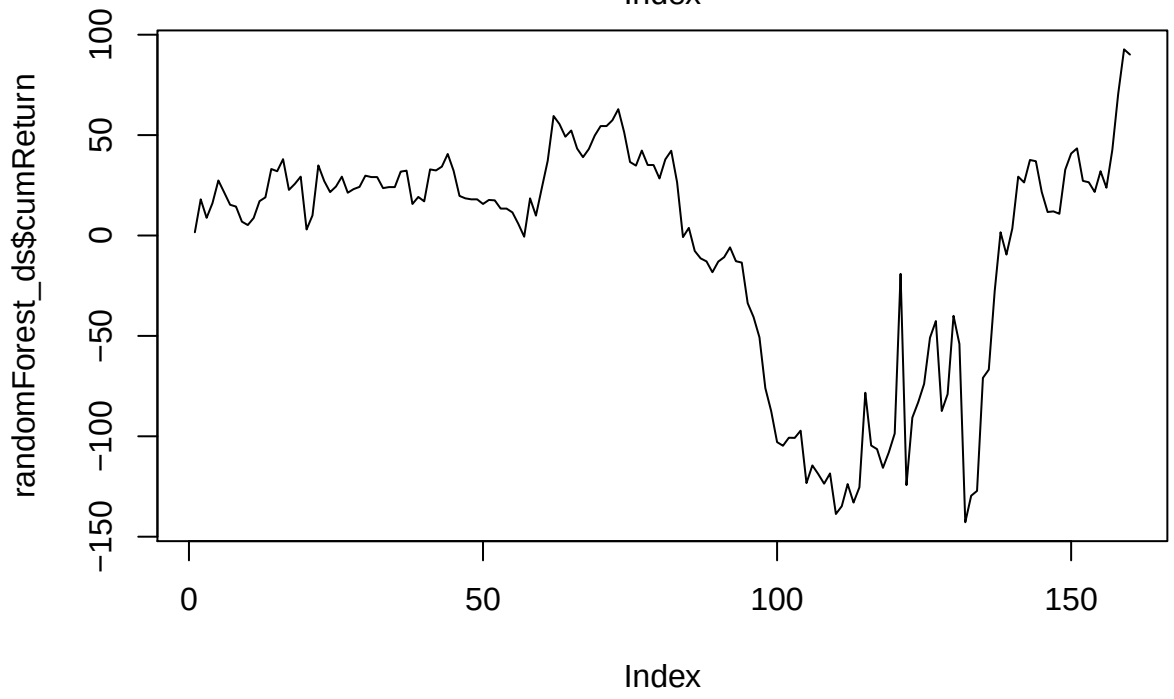
The SVM model accuracy shows that it is not the right model as it is still under our target of 50%. In the next section will use a different machine learning method to try to achieve our Goal

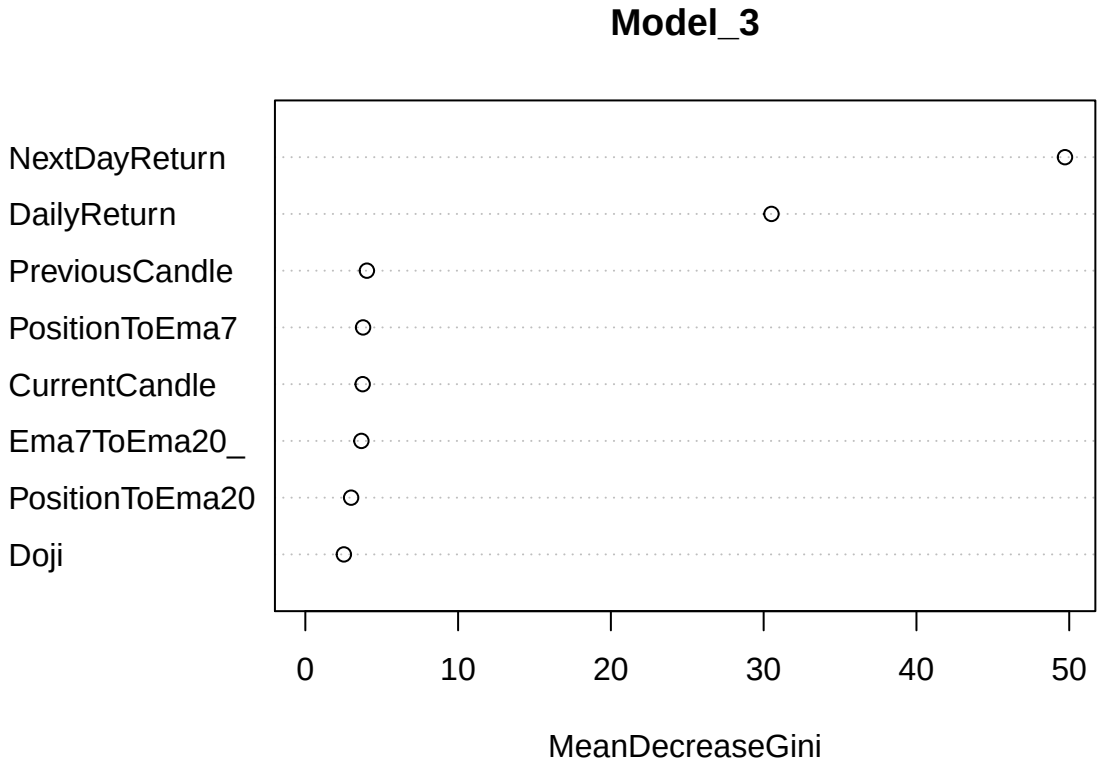
**Model 3 -Random Forest-**

In a nuttshell, Random Forests grows many classification trees, It is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.




randomForest_ds$prediReturn
50
0
−50
−100
0
50
100
150
Index


randomForest_ds$cumReturn
100
50
0
−50
−100
−150
0
50
100
150
Index




| | bear | bull |
|---|---|---|
| bear | 52 | 25 |
| bull | 51 | 32 |

```
## [1] 0.525
```

By plotting the OOB error rate graph, we can observe that the error rate doesnt have significant fluctuations or up/down trend by increasing the number of trees, Hence we will keep the default tress number of 500. The variable of importance plot we can identify that the variable with the most significant impact on our predictions is "NextDayReturn" and the "DailyReturn"

The Accuracy of the model exceeded the 50% which was our target for this project.

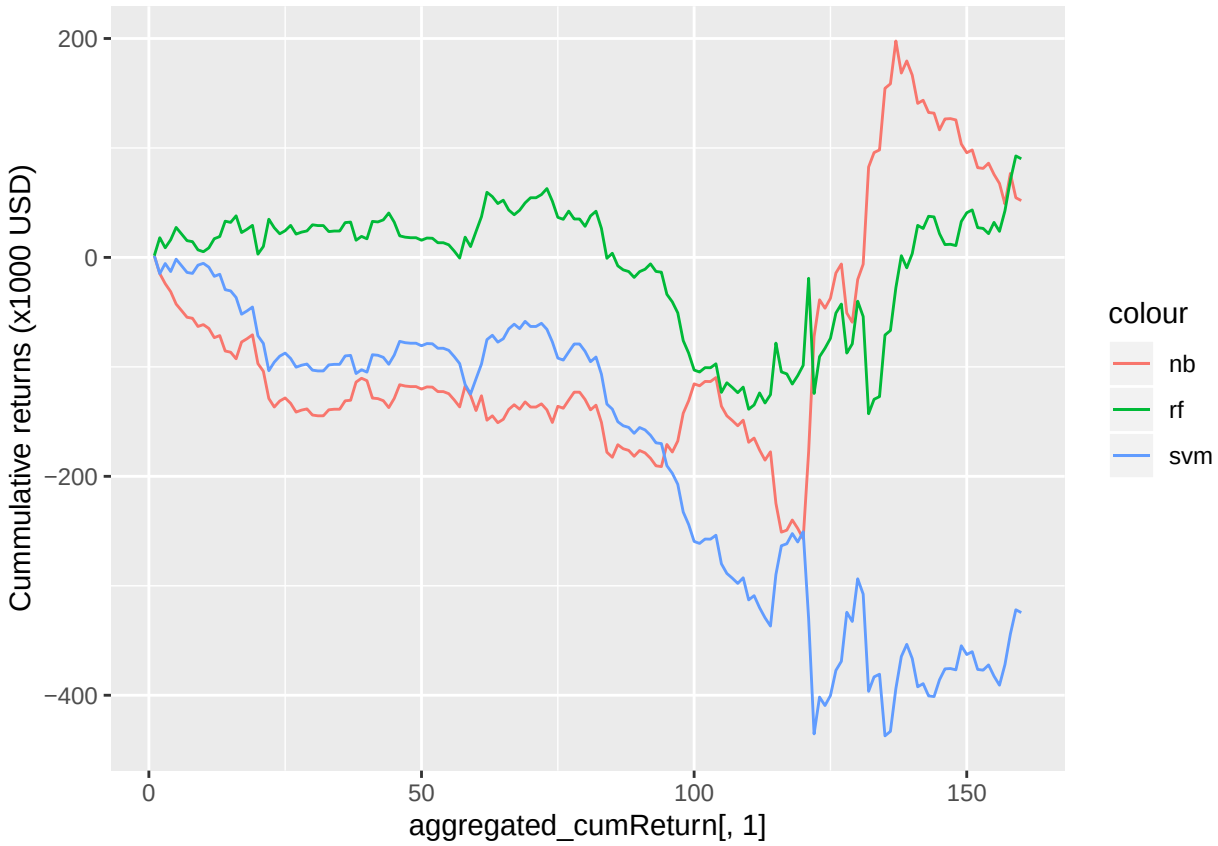
# 3- The Results

### Visualizing in a table the accuracy of our models

| Method | Accuracy |
|---|---|
| Naive bayes Model | 0.450 |
| Support vector Model | 0.450 |
| Random Forest Model | 0.525 |

### Plotting the cummlative returns

Creating a data frame that includes all the cummulative returns, in order to know how the equity is fluctuating and wether we are going to make gain by using the models.




## 4-Conclusion

In the above chart we can see the equity curve for the different algorithems , with Naive bais method the risk of losing 200k as we can see in the graph and then end up with 50k on the positive end, we can conclude that it is a very risky and unreliable model in this case. while SVM model for sure will lead to bigger loses. Random forest doesnt look like the perfect model as well however it managed to sustain the equity in the positive end more than the other two models and it lead to better results

#————————————————————————Thank you ————————————————————————
—