Contents

ISyE 6740 – Spring 2022
Project Final Report

Team Members

1. Suneet Taparia (GTID: 903661244),
2. Yoages Kumar Mantri (GTID: 903660372)
3. Saurabh Sinha (GTID: 903747945)

Project Title

Stock Classification based on predicted price and financial ratios using analytical models, leveraging Technical and Fundamental factors.

Problem Statement

Stock Market share price prediction is an age-old problem. We see this also a major struggle for Indian Investors. Over the years many algorithms and software have been built to solve this problem. There are numerous factors, including and not limited to a variety of subjective/sentiment factors, which can be used to predict stock value. Key challenge is that stock market data has variety, volume, velocity and is time series based. As such this problem poses a lot of analytical potential. For Indian Markets we could not find rich analytical models to help out with this problem.

The oldest and most well-known model of stock returns is the Capital Asset Pricing Model (CAPM)[1]. Factors [6] that drive stock returns and have stood the test of time are:

- **Size**: Smaller firms tend to have higher returns on average as compared to larger firms.
- **Value**: Inexpensive stocks tend to outperform expensive ones. It was documented by Fama and Fench in 1993.
- **Momentum**: Stocks that have performed well over the past years continue to perform well.
- **Profitability**: Stocks with robust operating performance tend to outperform those with weak performance.
- **Risk effect (volatility)**: Low beta assets tend to outperform high beta assets.

Another prevalent theory around Stock market price is the Efficient Market Hypothesis (EMH)[2]. This hypothesis states that the market is extremely efficient in reflecting individual information about individual stocks and about the market itself. There are three versions of Efficient Market Hypothesis:

a. **Weak Form**: Future prices cannot be predicted by analyzing historical prices.
b. **Semi-strong Form**: Prices adjust rapidly to new public information.
c. **Strong Form**: Prices reflect all information, public and private.

Due to technological advances, the information gap has reduced to ashes, and weak form is almost non-existent. Most markets have become Strong to Semi-strong. As per EMH, stock performance is thus impossible to predict as future price changes represent random departures from previous prices as information arrives randomly and prices adjust quickly.

This makes predicting stock price a complex problem, as we expect a lot of noise and a considerable number of possible factors. However, based on the large amounts of data and processing speeds at our disposal, our intension is to build a product which will help small scale investors, who don't have much knowledge about stock market, make data driven decisions on their portfolio.

Novelty

We couldn't find much literature on price prediction on Indian stock market. As such we wanted to do some modelling & analysis for this market. Also, most of the research that we've seen considers either Technical Factors or the Fundamental Factors. The goal of the project is to leverage both and classify stocks predicted price movement into 2 categories in the **short term** and **long term**.

a. **Up** – After a given time-period if the price of a share is greater than or equal to its current price.
b. **Down** – If the price of a share is less than its current price after a given time-period.

For this classification, we have done 2 types of analysis on the stock price historical data.

Technical analysis

This type of analysis focuses on changes in price, volume, and related statistics, with a forward-looking nature through the inferences gathered with technical indicators, developed through heuristics or mathematical calculations. The scope of this analysis will be limited to a **short time-period** (**T+3 days, T+7 days, T+30 days, T+90 days**).

Fundamental analysis

This type of analysis focuses on stock's intrinsic value using publicly available information. It uses factors based on the overall economy in relation to industry performance and a company's financial factors such as earnings, profit margin, assets, liabilities etc. These financial factors will become the variables from which our models classify the stocks. Fundamental analysis can be used for **long term** investment decision (**T+1 year or less**) since the factors used evolve slowly compared to technical factors analysis.

A good portfolio for an investor is thus a combination of both technical and fundamental analysis, which balances the risks associated with both methods.

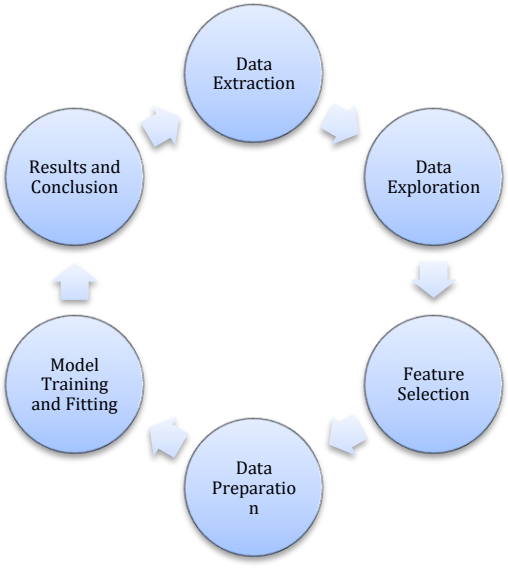
This project is an attempt to solve the problem of making investment decisions in the short and long term by applying various modelling techniques on technical and fundamental predictors. We have used Machine learning models discussed in ISYE-6740 to find out which combination works best for this problem.

Data Source

Stock data for Indian Markets from 2010 to 2021 was used for this project. We have focused our analysis on NIFTY 50 for Technical and NIFTY 50 and NIFTY Midcap 50 stocks for Fundamental analysis. The NIFTY50 is a benchmark Indian stock market index that represents the weighted average of 50 of the largest Indian companies listed on the National Stock Exchange of India. The idea behind choosing NIFTY 50 and NIFTY 50 Midcap stocks was to focus the analysis on fundamentally strong and highly traded stocks.

Methodology

We have followed the standard prescribed methodology for a data science project.



Data Extraction

For technical analysis, data was extracted using nsepy library.
For fundamental analysis, we used web scrapping on a reputed Indian stock market website www.moneycontrol.com.

Exploratory Data Analysis

We have used Pearson's correlation and variance inflation vector between various predictors to filter out the effect of multicollinearity. Below is the output of this analysis:

Pearson's Correlation coefficient, where x and y are factors evaluated for correlation

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Technical Analysis Features - Correlation Matrix using Pearson's correlation

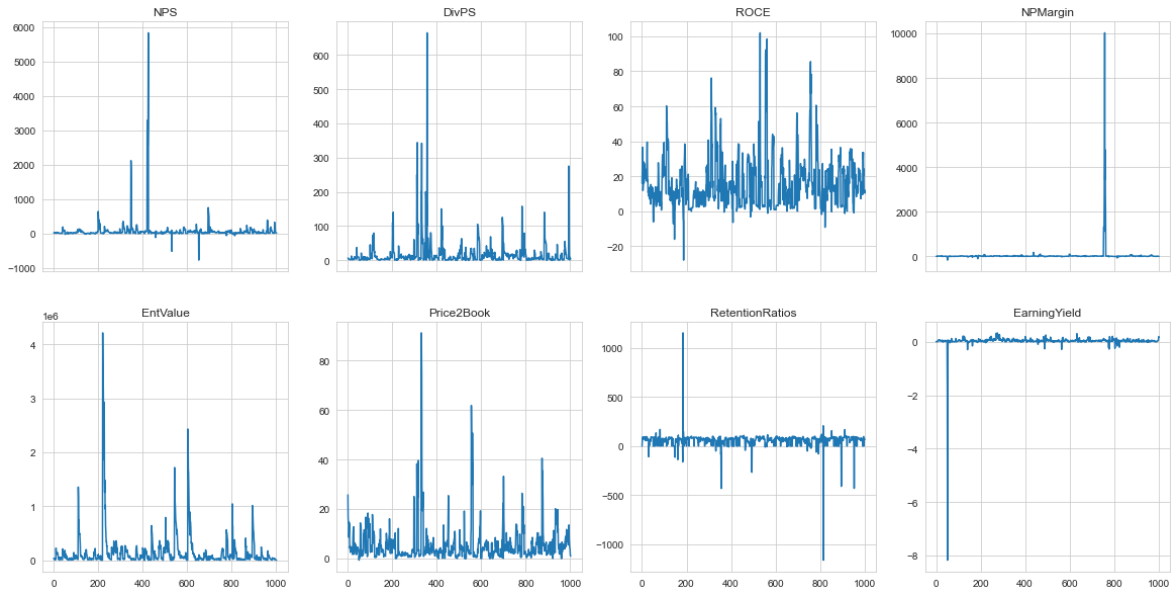
	Prev Close	Open	High	Low	Last	Close	VWAP	Volume	Turnover	Trades	Deliverable Volume	%Deliverble
Prev Close	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.022	-0.156	-0.190	-0.011
Open	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.022	-0.156	-0.190	-0.011
High	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.021	-0.155	-0.190	-0.012
Low	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.175	-0.022	-0.157	-0.190	-0.010
Last	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.021	-0.156	-0.190	-0.011
Close	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.021	-0.156	-0.190	-0.011
VWAP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	-0.174	-0.021	-0.156	-0.190	-0.011
Volume	-0.174	-0.174	-0.174	-0.175	-0.174	-0.174	-0.174	1.000	0.624	0.710	0.820	-0.215
Turnover	-0.022	-0.022	-0.021	-0.022	-0.021	-0.021	-0.021	0.624	1.000	0.853	0.557	-0.210
Trades	-0.156	-0.156	-0.155	-0.157	-0.156	-0.156	-0.156	0.710	0.853	1.000	0.573	-0.247
Deliverable Volume	-0.190	-0.190	-0.190	-0.190	-0.190	-0.190	-0.190	0.820	0.557	0.573	1.000	0.015
%Deliverble	-0.011	-0.011	-0.012	-0.010	-0.011	-0.011	-0.011	-0.215	-0.210	-0.247	0.015	1.000

We also observed some stocks like UNIPHOS, TATATEA were delisted and as such we have excluded such stocks. Some stocks like SBILIFE, HDFCLIFE were listed quite recently and had very less data against them. These stocks were also excluded from Technical Analysis.

Fundamental Analysis Features - Correlation Matrix using Pearson's correlation

	BasicEPS	DilutedEPS	CashEPS	DivPS	OpRev	NPS	ROCE	NPMargin	ROA	ROE2Networth	EntValue	EntValuePerNetSales	Price2Book	Price2Sales	RetentionRatios	EarningYield
BasicEPS	1	0.993	0.957	0.27	0.908	0.97	0.051	-0.013	0.129	0.033	-0.032	-0.028	0.021	-0.024	0.039	0.027
DilutedEPS	0.993	1	0.956	0.27	0.913	0.966	0.048	-0.013	0.13	0.032	-0.034	-0.029	0.019	-0.025	0.036	0.027
CashEPS	0.957	0.956	1	0.261	0.963	0.987	0.049	-0.014	0.046	0.029	-0.037	-0.029	0.005	-0.025	0.05	0.023
DivPS	0.27	0.27	0.261	1	0.218	0.282	0.404	0.013	0.034	0.193	-0.033	-0.01	0.39	-0.001	-0.203	0.018
OpRev	0.908	0.913	0.963	0.218	1	0.941	0.02	-0.02	0.008	0.018	-0.04	-0.037	-0.018	-0.033	0.054	0.024
NPS	0.97	0.966	0.987	0.282	0.941	1	0.074	-0.013	0.067	0.045	-0.033	-0.029	0.023	-0.025	0.049	0.028
ROCE	0.051	0.048	0.049	0.404	0.02	0.074	1	0.296	-0.136	0.478	-0.104	0.221	0.586	0.244	-0.1	0.062
NPMargin	-0.013	-0.013	-0.014	0.013	-0.02	-0.013	0.296	1	0.03	0.193	0.006	0.876	0.063	0.88	-0.07	0.028
ROA	0.129	0.13	0.046	0.034	0.008	0.067	-0.136	0.03	1	-0.057	0.007	0.029	0.048	0.029	-0.022	0.014
ROE2Networth	0.033	0.032	0.029	0.193	0.018	0.045	0.478	0.193	-0.057	1	-0.027	0.141	0.32	0.147	0	0.837
EntValue	-0.032	-0.034	-0.037	-0.033	-0.04	-0.033	-0.104	0.006	0.007	-0.027	1	0.031	-0.024	-0.001	0.041	0.005
EntValuePerNetSales	-0.028	-0.029	-0.029	-0.01	-0.037	-0.029	0.221	0.876	0.029	0.141	0.031	1	0.167	0.998	-0.079	0.005
Price2Book	0.021	0.019	0.005	0.39	-0.018	0.023	0.586	0.063	0.048	0.32	-0.024	0.167	1	0.175	-0.08	-0.017
Price2Sales	-0.024	-0.025	-0.025	-0.001	-0.033	-0.025	0.244	0.88	0.029	0.147	-0.001	0.998	0.175	1	-0.084	0.004
RetentionRatios	0.039	0.036	0.05	-0.203	0.054	0.049	-0.1	-0.07	-0.022	0	0.041	-0.079	-0.08	-0.084	1	0.053
EarningYield	0.027	0.027	0.023	0.018	0.024	0.028	0.062	0.028	0.014	0.837	0.005	0.005	-0.017	0.004	0.053	1

Fundamental factors data distributions do not show any patterns:



Feature Selection

Volume, Turnover, Trades and Deliverable Volume were shortlisted for Technical Analysis. In addition, there were some companies which were listed recently and hence did not have complete data for past 11 years such companies were excluded from analysis.

For Fundamental analysis after removing the highly correlated features, the following features got selected: Net Profit / Share, Dividend / Share, Net Profit Margin, Return on Capital Employed, Price to Book ratio, Retention Ratio, Enterprise Value, Earning Yield.

Data Preparation

Labels were created for 4 different time horizons in technical analysis – T+3 days, T+7 days, T+30 days and T + 90 days based on comparison between actual stock price at the T + *n* day and the current day price.

For e.g. *If the price for a stock on 6th Oct 2020 is 550 and the price on 9th Oct 2020 is 570, then the T+3 label is* ***Up***

For Fundamental analysis purposes, the labels or predictors were created as 1 or 0 indicating Up or Down as in whether to invest in a particular stock for long term T + 1 year or less.

To create the label mentioned above we've used a threshold-based methodology, where if the % change in the feature values have increased than the threshold value, it is considered as positive indicator.
Only if more than 50 % of the features have been reported positive then that becomes a candidate for 1 or "Up" else 0 or "Down".

For e.g. *If for a given year the Features mentioned above are above the good values for threshold[9] for at least half of the Features, then the Label is marked as 1 (Up) else 0 (Down)*

Model Training and Fitting

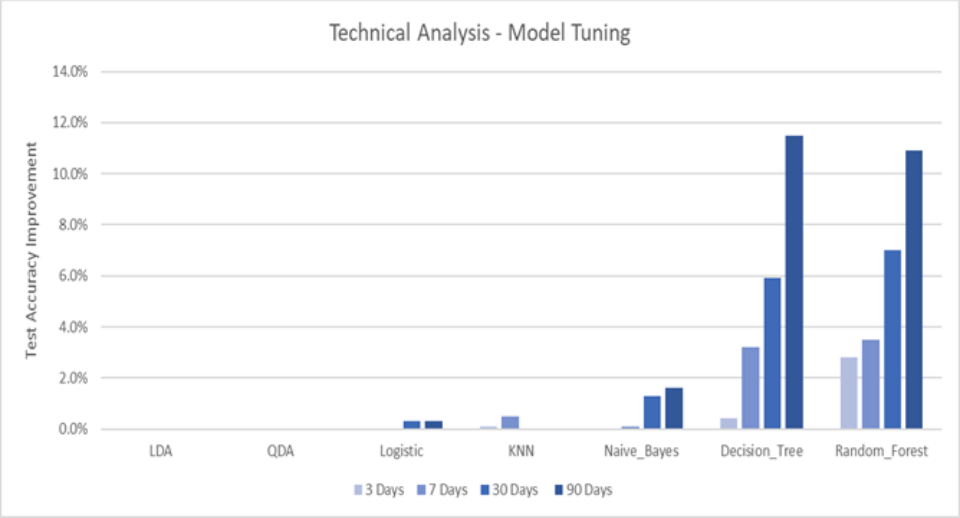
As this is a typical classification and prediction problem, we have applied the following models for trend prediction (**Up / Down**)– Logistic Regression, Naïve Bayes, K Nearest Neighbor, Decision Tree, Random Forest, SVM, Neural Network, AdaBoost, LDA and QDA.

More details of this step are described below in 2 parts – Technical and Fundamental Analysis

Technical Analysis

Model training for technical analysis was done on the entire dataset with 80:20 split between train and test.

Post this different model were tuned using Grid Search with 5-fold cross validation using accuracy as the scoring criteria. Below is graphical represent of accuracy improvement observed.



Comments

For LDA and QDA we observed no change in improvement because the default parameters were the most optimal already. For Logistic, KNN and Naïve Bayes, we observed minor improvement in accuracies with KNN focusing more on 3 Days and 7 Days, while the other 2 on 30- and 90-Days predictions. These changes were minor improvements so nothing concrete can be concluded. For Decision Tree and Random Forest, the changes observed are the most significant ones.

In Decision Tree, 'Gini' was seen to be the most optimal criteria and major difference was observed because of variations in max_leaf_nodes. For different stocks, different max_leaf_nodes provided optimal results. Similarly for Random Forest, a lot of variation was observed in max_depth and n_estimators parameters. Optimizing on these resulted in significant improvement in accuracies.

Fundamental Analysis

For fundamental analysis, we had relatively lesser data. Hence, model training for Fundamental analysis was done on a single training and test set with 80:20 split. As part of applied Machine Learning strategies along with Grid search and 5-fold cross validation, we've included several models and did hyper parameter tuning.

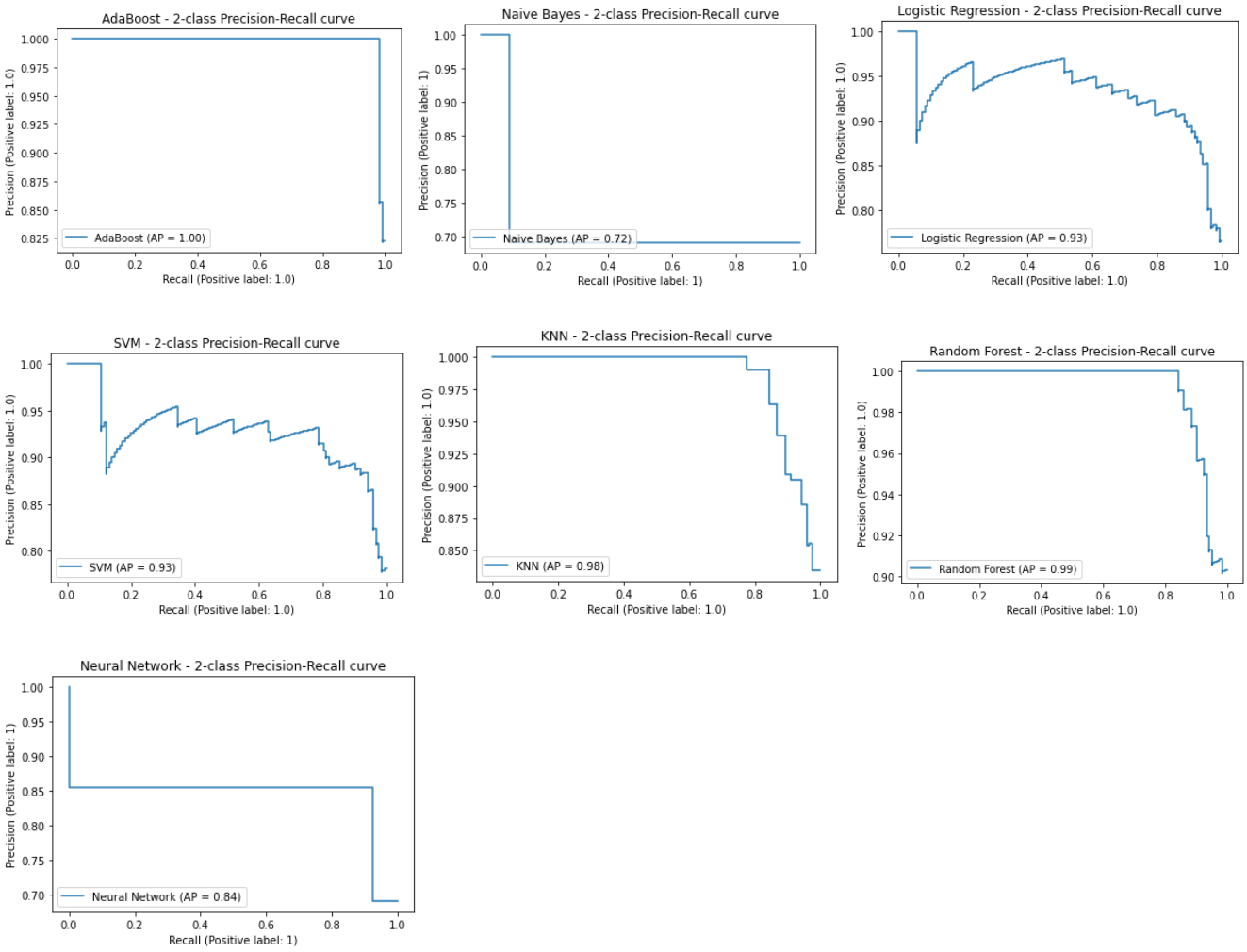
We've made use of Precision-Recall curve to measure the model performance.

Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

High scores for both show that the classifier is returning accurate results (high precision), as well as returning most of all positive results (high recall).

Below are the captured Precision – Recall curves for the applied Models, which are the result of GridSearch, 5-fold CV and Hyper parameters tuning like – gamma, criterion, n_estimators, min_samples_leaf, epochs, batch_size, penalty etc.



Average precision (AP) summarizes such a plot as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight:

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) P_n$$

Where P_n and R_n are the precision and recall at the nth threshold. A good model would have high value of the Average Precision.

Evaluation and Final Results

We are primarily trying to classify different shares into Buy and Don't Buy categories. The confusion matrix looks something like below for all the classification models used.

Hyperparameter tuning of models has been done using various metrics, each resulting in different kinds of optimization:

1. **Precision**: Precision score is the ratio of actual positives versus predicted positives. Maximizing this will help an investor to minimize wrong buys more thus its more suitable for risk-averse investors

   Precision Score = True Positive / True Positive + False Positive

2. **Recall**: Recall is the ratio of actual positives versus predicted positives. Maximizing this will help an investor to reduce missed opportunities thus it is more suitable for greedy and risk-loving investors.

   Recall = True Positive / True Positive + False Negative

3. **Accuracy**: Maximizing this will help an investor in the overall sense by making sure that he gets only the right buys and avoids wrong buys, thus this is a balance between the above two metrics.

   Accuracy = True Positive + True Negative / True Positive + False Negative + False Positive + True Negative

4. **F1 score**: This is another metric which is calculated as Harmonic Mean of Precision and Recall. Its interpretation is too complicated to comprehend in this context. Hence, we are not using this metric at all.

Each of these 3 criteria will give different results as they tend to optimize different things.

For Technical and Fundamental Analysis, we chose overall Accuracy to be the optimizing metric since it optimizes prediction in the overall sense.

Confusion Matrix Explained

For all models, we have detailed the elements of a confusion matrix from the viewpoint of an investor.

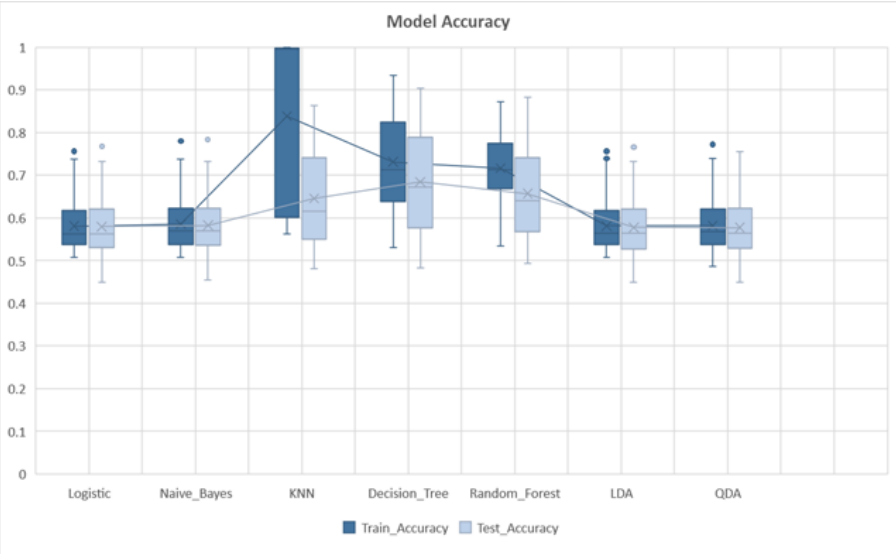
		Predicted	
		Share Price Increased (Buy Share)	**Share Price didn't Increase (Don't Buy Share)**
Actual	**Share Price Increased**	**True Positive** Correct price increase prediction so investor makes profit	**False Negative** Investor misses an opportunity to make profits
	Share Price didn't Increase	**False Positive** Wrong price increase prediction so investor suffers losses	**True Negative** Correct price non-increase prediction so investor avoids losses

There's a heavy influence of Ensemble Machine Learning way of solving problem in this project to solve problem of choosing the right stocks. Each Model we've created makes its own prediction, but if an Investor chooses to make use of them, we suggest to cross-validate the results with other series of Models that we have.

More models predicting the same output is likely to be the right choice for Investors to pick the stocks to invest.

Technical Analysis Results

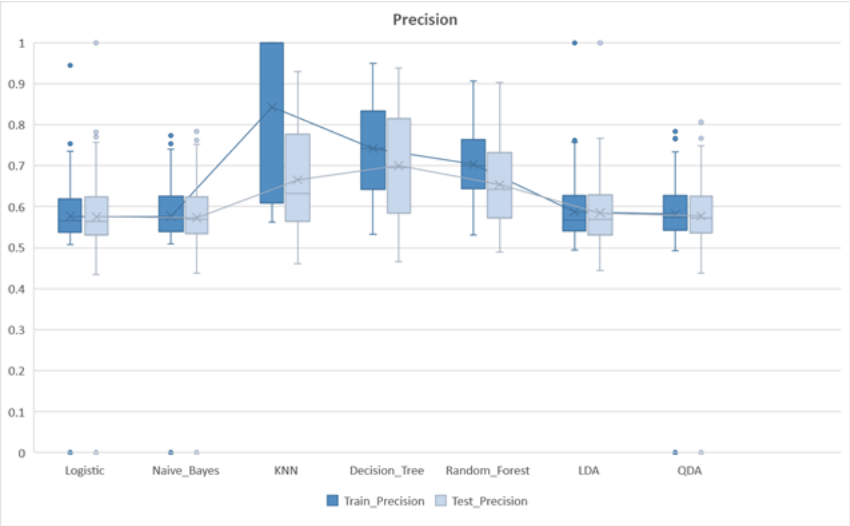
The result of various models in terms of prediction accuracy is as follows:



Based on the above we can see that KNN, Decision Tree and Random Forests have performed better on training set, but the test accuracy of these models is not so good. Logistic Regression, Naïve Bayes, LDA and QDA have consistent performance on both training and test sets. The best accuracy on test set was observed for **Decision Tree** and **Random Forest**. We can clearly see a case of overfitting for KNN as there is a stark difference in train and test accuracy. This has happened because of close in the cluster centers in Train resulting in a lot of misclassifications in Test.

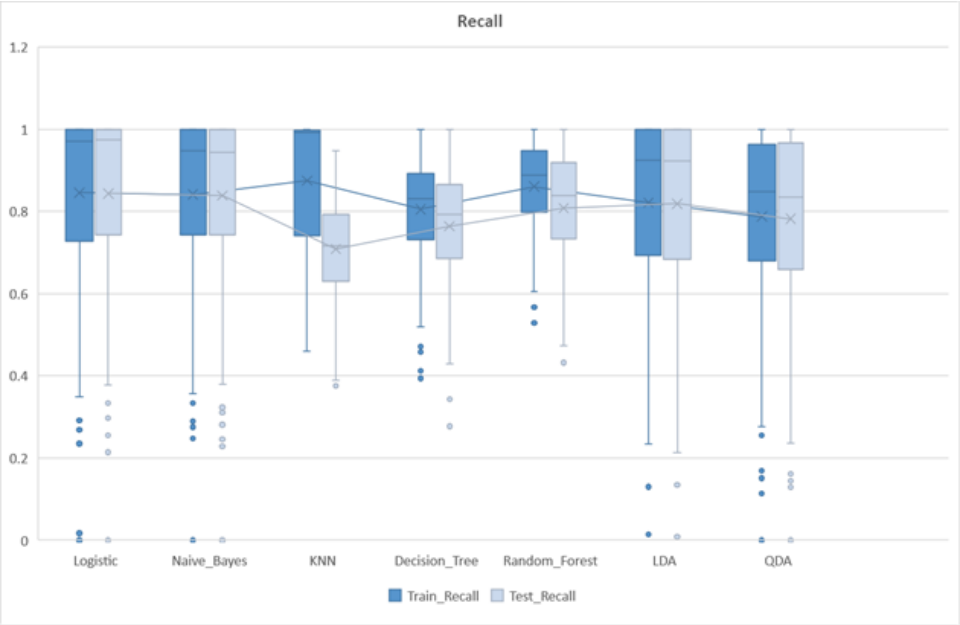
For some of the stocks we can see that there are outliers observed in accuracy. These were BAJFINANCE, HINDUNILVR and EICHERMOT. All 3 of these stocks have very high train and test accuracies across models. We could not find a definite reason behind this. The only directional reason is use of just 4 features may have resulted in spurious accuracies for these stocks.

Apart from classification accuracy, we have also captured other metrics (**Precision and Recall**) to evaluate model performance. The results are as shown below:



It is a good measure to use when the cost of False positives is high. As detailed above low precision will result in investors investing in wrong stocks and thereby losing money. The highest precision score on test sets is displayed by Decision Tree and Random

Forests. This is in line with what was observed for accuracy. For a lot of stocks, we observed high precision in Train and Test as these stocks had a larger proportion of Ups in both these datasets. HDFCBANK and DIVISLABS is one of the examples having 80% ups in Train and Test datasets.

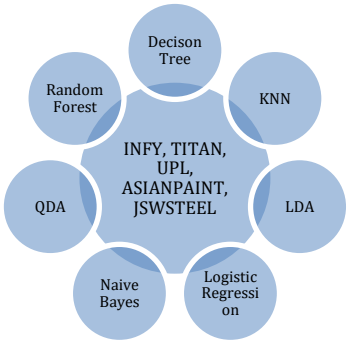


This measure is used to select model when there is high cost associated with False Negative. We would like this number to be as high as possible. Since majority of these stocks were in increasing mode, we observed a very high recall overall. A higher recall score on the test set is displayed by Logistic Regression and Naïve Bayes.

Technical Analysis Recommendations

We have used accuracy of trend prediction at individual stock level. Using different models, we have identified the 10 example stocks with good classification accuracy. If a stock is being classified correctly with high classification accuracy by all or most models, then this gives us high confidence on the result. Our final recommendation would be these common stocks.

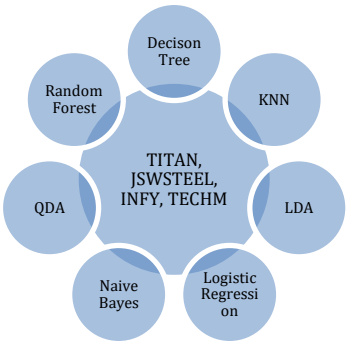
The output is as shown below



Common Predictions

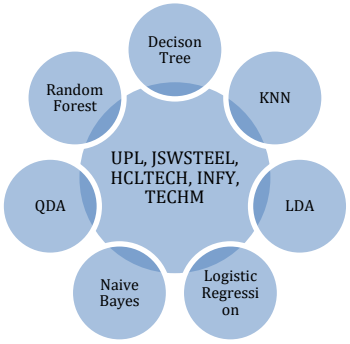
T+3 days (Example Stocks to buy based on Model Prediction)						
Decision Tree	**KNN**	**LDA**	**Logistic**	**Naïve Bayes**	**QDA**	**Random Forest**
INFY	ASIANPAINT	TECHM	TECHM	TECHM	INFY	INFY
TITAN	WIPRO	INFY	INFY	INFY	TITAN	TECHM
UPL	BAJAJ-AUTO	TITAN	TITAN	HCLTECH	UPL	TITAN
MARUTI	AXISBANK	UPL	UPL	TITAN	TECHM	UPL
POWERGRID	HCLTECH	HCLTECH	HCLTECH	WIPRO	JSWSTEEL	ASIANPAINT
ASIANPAINT	KOTAKBANK	JSWSTEEL	JSWSTEEL	JSWSTEEL	DIVISLAB	HCLTECH
JSWSTEEL	IOC	DIVISLAB	DIVISLAB	UPL	HCLTECH	POWERGRID
TECHM	INFY	ASIANPAINT	ASIANPAINT	DIVISLAB	ASIANPAINT	MARUTI
NESTLEIND	ULTRACEMCO	BAJAJ-AUTO	WIPRO	TCS	MARUTI	EICHERMOT
AXISBANK	ONGC	POWERGRID	BAJAJ-AUTO	BAJAJFINSV	BPCL	HINDUNILVR

Based on the above we can see that the highlighted stocks are predicted by at least 5 of the 7 models.



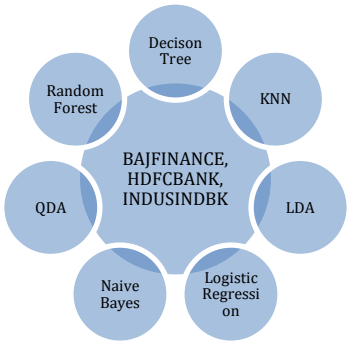
Common Predictions

T+7 days (Example Stocks to buy based on Model Prediction)						
Decision Tree	**KNN**	**LDA**	**Logistic**	**Naïve Bayes**	**QDA**	**Random Forest**
MARUTI	MARUTI	TECHM	TECHM	TECHM	INFY	INFY
BRITANNIA	HCLTECH	INFY	INFY	INFY	TITAN	TITAN
TITAN	UPL	WIPRO	WIPRO	WIPRO	ASIANPAINT	ASIANPAINT
DIVISLAB	HEROMOTOCO	HCLTECH	JSWSTEEL	BAJAJFINSV	JSWSTEEL	UPL
JSWSTEEL	SUNPHARMA	JSWSTEEL	HCLTECH	JSWSTEEL	UPL	HCLTECH
INFY	BPCL	TITAN	TITAN	TITAN	DIVISLAB	BAJFINANCE
POWERGRID	BRITANNIA	GRASIM	GRASIM	GRASIM	TECHM	JSWSTEEL
BAJFINANCE	TECHM	DIVISLAB	DIVISLAB	ADANIPORTS	HDFC	GRASIM
NESTLEIND	IOC	BAJAJ-AUTO	ASIANPAINT	DIVISLAB	BAJFINANCE	POWERGRID
TECHM	HDFC	ASIANPAINT	BAJAJ-AUTO	HINDUNILVR	MARUTI	TECHM



Common Predictions

T+30 days (Example Stocks to buy based on Model Prediction)						
Decision Tree	**KNN**	**LDA**	**Logistic**	**Naïve Bayes**	**QDA**	**Random Forest**
TECHM	KOTAKBANK	UPL	UPL	UPL	UPL	AXISBANK
ONGC	MARUTI	TECHM	TECHM	TECHM	INFY	HCLTECH
AXISBANK	BRITANNIA	BAJAJFINSV	INFY	GRASIM	TECHM	INDUSINDBK
INDUSINDBK	UPL	GRASIM	GRASIM	HCLTECH	GRASIM	ASIANPAINT
HCLTECH	INDUSINDBK	TATAMOTORS	TATAMOTORS	INFY	NTPC	JSWSTEEL
BAJAJ-AUTO	BAJAJ-AUTO	HCLTECH	JSWSTEEL	JSWSTEEL	JSWSTEEL	KOTAKBANK
BRITANNIA	HINDALCO	JSWSTEEL	TITAN	TITAN	DIVISLAB	HDFCBANK
INFY	AXISBANK	NTPC	BAJAJFINSV	HINDUNILVR	TITAN	SHREECEM
JSWSTEEL	HCLTECH	INFY	AXISBANK	BAJAJFINSV	BAJFINANCE	TCS
KOTAKBANK	ONGC	TITAN	BAJFINANCE	BAJFINANCE	INDUSINDBK	INFY



Common Predictions

T+90 days (Example Stocks to buy based on Model Prediction)						
Decision Tree	**KNN**	**LDA**	**Logistic**	**Naïve Bayes**	**QDA**	**Random Forest**
BAJFINANCE	AXISBANK	UPL	UPL	HCLTECH	HDFCBANK	HDFCBANK
HDFCBANK	BRITANNIA	TITAN	HDFCBANK	HDFCBANK	BAJFINANCE	BAJFINANCE
INDUSINDBK	KOTAKBANK	HDFCBANK	BAJFINANCE	UPL	HEROMOTOCO	INDUSINDBK
MARUTI	TECHM	BAJFINANCE	INDUSINDBK	BAJFINANCE	INDUSINDBK	POWERGRID
BAJAJ-AUTO	MARUTI	HCLTECH	HCLTECH	INDUSINDBK	UPL	DRREDDY
TECHM	INDUSINDBK	INDUSINDBK	GRASIM	NESTLEIND	HCLTECH	JSWSTEEL
DRREDDY	HDFCBANK	ITC	ITC	BAJAJFINSV	ITC	AXISBANK
ALL	BAJAJFINSV	BRITANNIA	POWERGRID	ITC	ALL	BAJAJ-AUTO
NESTLEIND	DIVISLAB	HEROMOTOCO	ASIANPAINT	HDFC	TECHM	TECHM
HEROMOTOCO	UPL	ASIANPAINT	DIVISLAB	MARUTI	MARUTI	HCLTECH

Significance

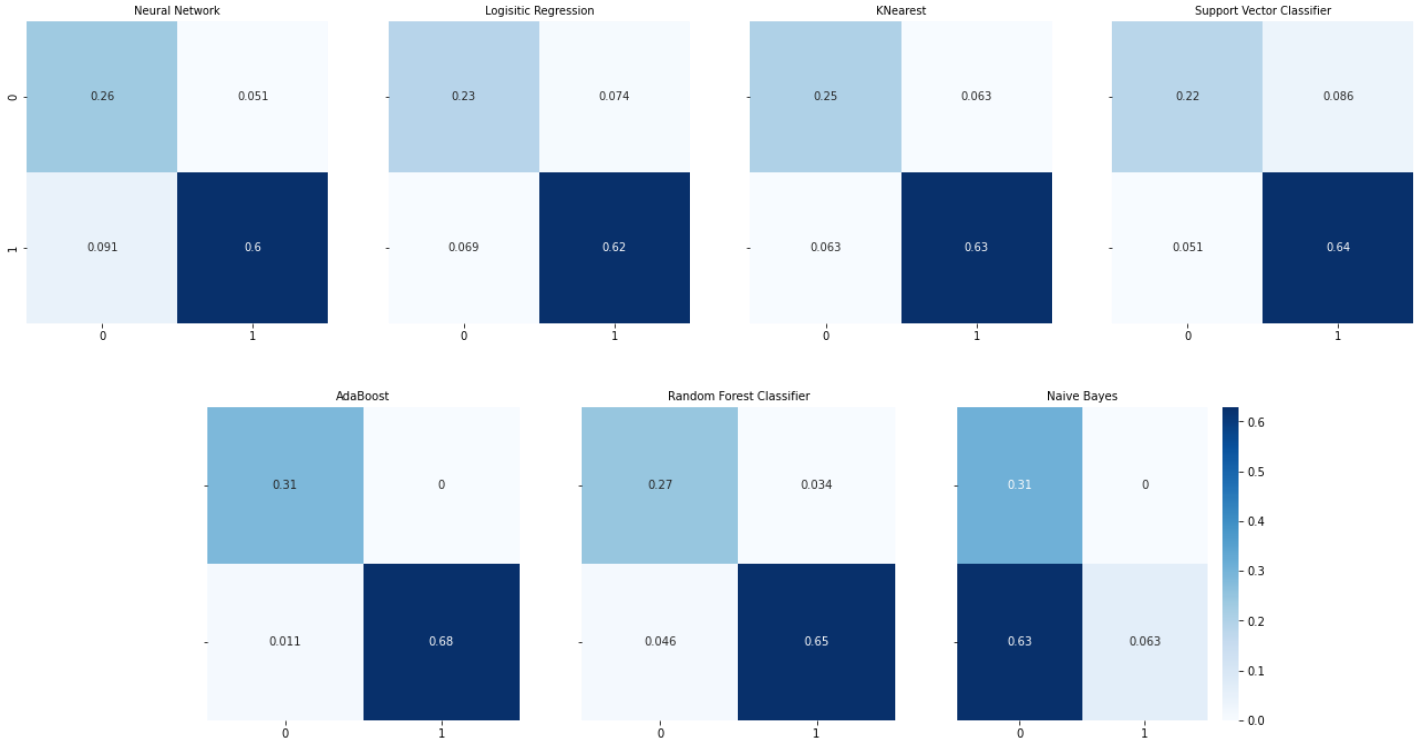
We can clearly see that almost all the stocks predicted have gone up. This makes our modelling approach significant for an investor.

Fundamentals Analysis Results

We've evaluated each of the Models for their ability to make the maximum possible correct predictions that is True Positives and True Negatives.

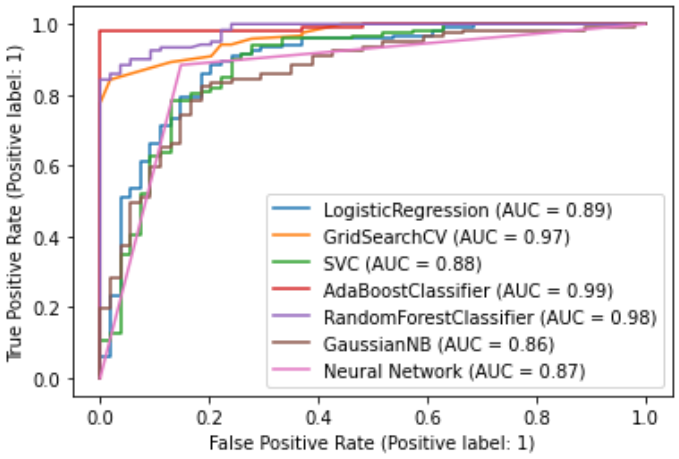
Well, it is a performance measurement for machine learning classification problem where output can be two or more classes.
It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

Confusion matrix captured for all the classifiers are captured below:



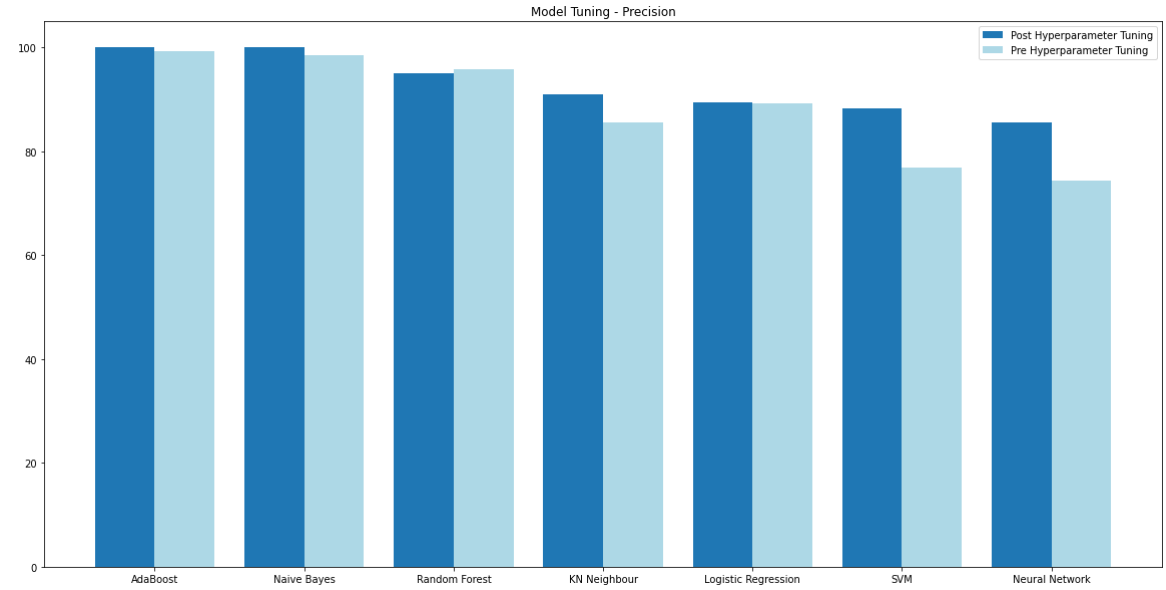
An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means the model has no class separation capacity whatsoever.

The ROC-AUC curve for all the Models is displayed as follows:

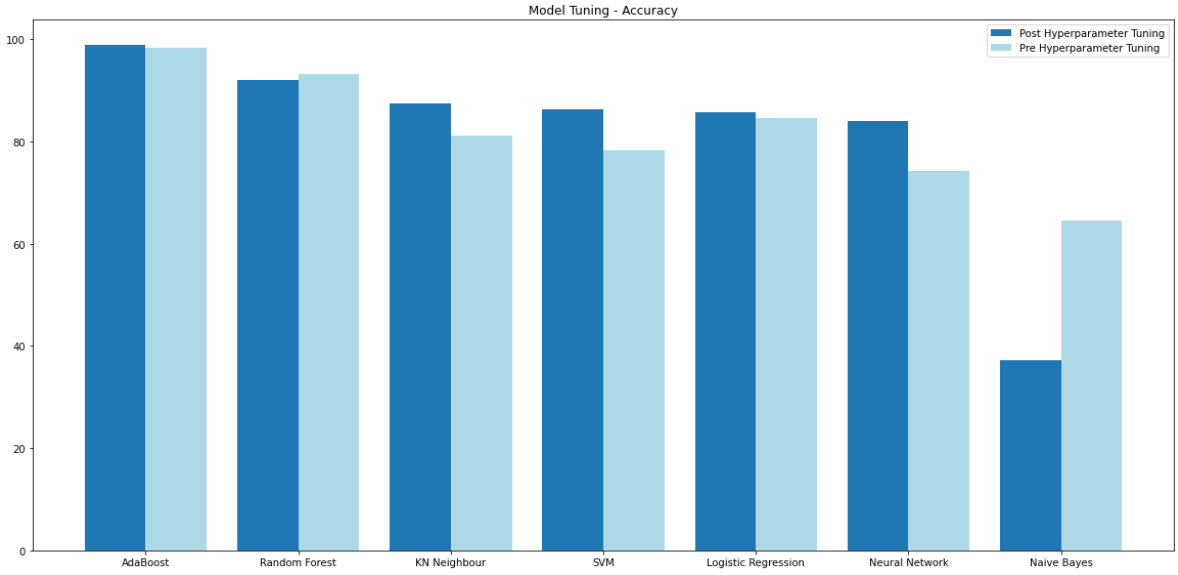


Below is a comparison of Precision and Accuracy of different models before and after tuning.

For a Model to be best performing Model, Precision should be high as possible.



For a Model to be best performing Model, Accuracy should be high as possible.



AdaBoost, Random Forest, K-Nearest Neighbor, SVM, Artificial Neural Network, Logistic Regression have resulted in high accuracy, and are selected to make a predict the results.

We have dropped the worst performing Model in terms of Accuracy i.e., Naïve Bayes from the final evaluation.

Fundamental Analysis Recommendations

We took the last fiscal year FY21 fundamentals data and made a prediction for all of the stocks, using all the tuned Models.

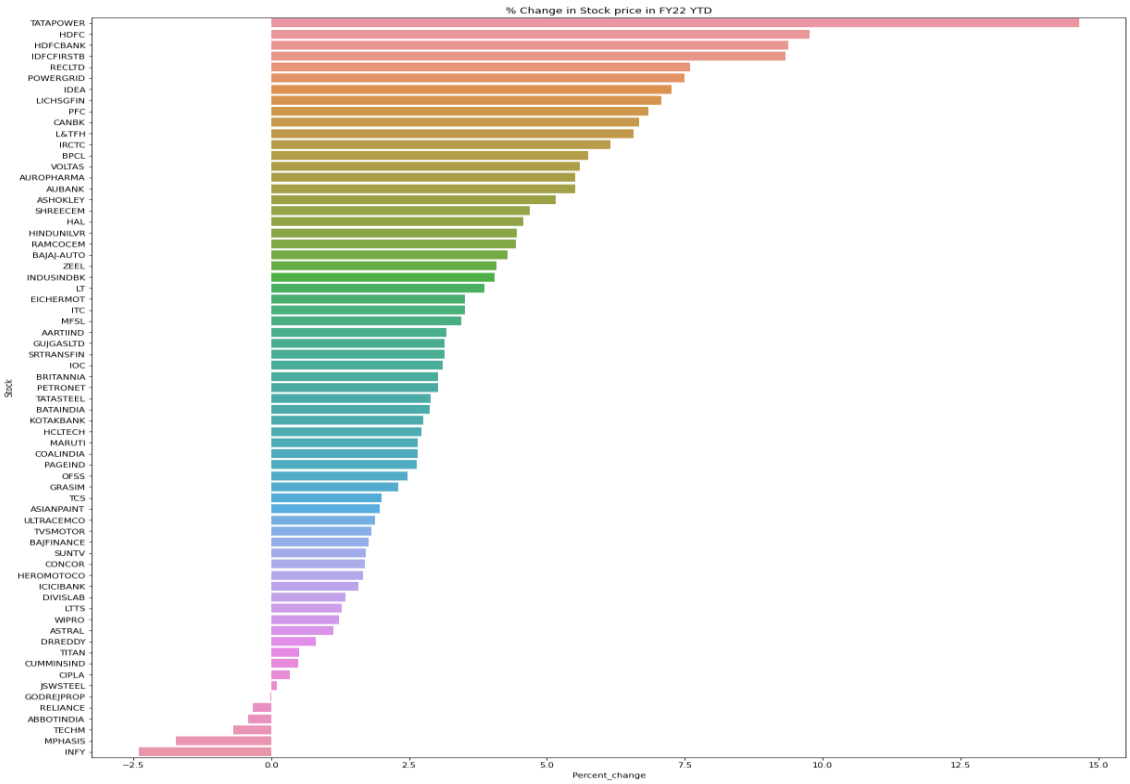
T+ 1 Year (Example Stocks to buy based on Model Prediction)					
AdaBoost	**Logistic Regression**	**SVM**	**Random Forest**	**KN Neighbors**	**Neural Network**
ZEEL	TORENTPOWER	TATASTEEL	PETRONET	TVSMOTOR	RELIANCE
WIPRO	ZEEL	RELIANCE	MARUTI	ULTRACEMCO	SHREECEM
VOLTAS	SHREECEM	TVSMOTOR	MPHASIS	AARTIND	SUNTV
ULTRACEMCO	RELIANCE	TCS	LICHSGFIN	ABBOTINDIA	POWERGRID
TVSMOTOR	TATASTEEL	AARTIND	ONGC	TATASTEEL	ASIANPAINT
ASIANPAINT	TCS	ULTRACEMCO	KOTAKBANK	ICICIBANK	IRCTC
ASHOKLEY	SUNTV	SHREECEM	ICICIBANK	SUNTV	ITC
ABBOTINDIA	POWERGRID	ABBOTINDIA	HDFCBANK	TCS	SBI
AARTIND	AARTIND	VOLTAS	HEROMOTOCO	HAL	TITAN
RELIANCE	ASIANPAINT	SUNTV	HCLTECH	HDFC	TCS

We've leveraged Ensemble Modeling approach. Our buy recommendation will be the common predicted stocks.

Significance

To check significance of our recommendation we did a reality check.
For the recommended stocks we've compared the stock price of FY21 and current price (FY22) and tested the hypothesis, whether the stock price have gone "Up" or "Down". We could clearly see that ~90% of the stocks recommended have gone up at an average of 7%. The stocks that haven't gone up have fallen by an average of 3%. So, in the overall sense, this recommendation will end up in making profits for an investor.

The recommended stocks as output have positive correlation with increase in stock prices. And if an Investor buys 1 share each of the recommended stocks, she / he would have positive investment returns.



Challenges and Future Enhancements

We faced the following key challenges in solving this modelling problem.

1. Identifying relevant variables for model building: Stock markets are mostly non-parametric, non-linear, noisy and deterministic chaotic system. There are number of variables at play, and it is difficult to find the most apt and relevant variable to determine future stock price. For e.g. Any big news causing political instability can impact stock prices in a big way both in the short and long term. It is difficult to factor this variable in a model.

2. Web scrapping for fundamental data: Web scrapping data from website was challenging due to inconsistent HTML structures for different stocks. We couldn't find data as easily as those available for US Markets.

3. Data size: In technical analysis, we had large data, but only a small set of predictors. Also, these predictors were largely correlated. In fundamental analysis, we had relatively small data in terms of number of records, but many predictors.

4. Modelling Time: Hyperparameter tuning along with model fitting for SVM took very long time to train on technical and fundamental data. Due to time constraints, we dropped this model from our project.

5. Small Caps: This analysis showed very unreliable results. There are many more subjective elements at play which we couldn't get data for.

To further refine the outcome of our project, we would want to do the following:

1. Time series modelling (ARIMA) and ANN for Technical analysis.

2. 10 years data for fundamental analysis isn't enough for model training, as the data changes only once every quarter. Using more data for training will result in a more robust model.

3. Access to multiple data sources, Automated data extraction pipeline and Model deployment automation can result in a full-fledged product that stock market investors could leverage for decision making.

4. Technical Analysis can be extended to Futures and Options.

5. Small Cap stocks or startup valuations can be done using fundamental analysis

Team Member Responsibilities

TASK	TEAM MEMBER
TOPIC SELECTION	All have contributed equally
DATA CLEANING	All have contributed equally
EXPLORATORY DATA ANALYSIS	All have contributed equally
MODEL FITTING	All have contributed equally
RESULT INTERPRETATION AND CONCLUSIONS	All have contributed equally

References

1. Eugene F. Fama and Kenneth R. French, The Capital Asset Pricing Model: Theory and Evidence
2. Alexandra Gabriela Ţiţan, The Efficient Market Hypothesis: Review of Specialized Literature and Empirical Research
3. Credit Suisse, Technical Analysis - Explained
4. Silpa K S, Arya Mol J, Dr. A S Ambily, A study on fundamental analysis of selected IT companies listed at NSE
5. Marwa Sharaf, Ezz El-Din Hemdan, Ayman El-Sayed & Nirmeen A. El-Bahnasawy , "StockPred: a framework for stock Price prediction", 12th Feb 2021
6. Jennifer Bender, Remy Briand, Dimitris Melas, Raman Aylur Subramanian, "Foundations of Factor Investing", Dec 2013
7. Mehr Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, Stock Closing Price Prediction using Machine Learning Techniques, International Conference on Computational Intelligence and Data Science (ICCIDS 2019)
8. Sreelekshmy Selvin; R Vinayakumar; E. A Gopalakrishnan; Vijay Krishna Menon; K. P. Soman, Stock price prediction using LSTM, RNN and CNN-sliding window model
9. Investopedia - https://www.investopedia.com/financial-ratios-4689817