Approach to Investment Portfolio Management with Machine Learning & Predictive Analytics

Project Report

Use Case – NIFTY BANK Index (2013-2018)

03/16/2019

Internal Guide
Prof. Sanjeet Singh
Professor, Operations Management Group
IIM Calcutta

Group V – Batch APDS 02
Sriniwas Mahapatro
Aditya S Prakash
Shweta Bhuwan
Sambita Chakraborty
Surjeet Ray
Amit Ahlawat
Anand Srinivasan
Arvind Kumar
Mayank Jain
Anuraag Thareja

Contents

Introduction

National Stock Exchange (NSE) is the leading stock exchange in India for trading equities. NIFTY BANK Index represents the 12 most liquid and large capitalized stocks from the banking sector which trade on the National Stock Exchange (NSE). It provides investors and market intermediaries a benchmark that captures the capital market performance of Indian Banking sector. Bank NIFTY is considered of paramount importance because it showcases the overall Banking sector's performance and is a major constituent of several funds and portfolios. It is also one of the important benchmarks to create ETF which can mimic the performance of the Banking sector.

Specialization

Financial Portfolio Management

Portfolio Analysed

NIFTY BANK Index (Large Cap, Market Capitalization: Rs. 1726 Cr

Equity Funds being a very widely used investment avenue, the prediction of stock prices is very much sought after. Although there can be several ways of predicting a stock price or fund value, it is challenging to find a highly accurate method to predict a stock value. Methodologies to predict broadly fall into two different categories: fundamental analysis and technical analysis. Fundamental analysis is the method to find the fair value or intrinsic value of a stock by looking at economic or financial factors that influence business whereas technical analysis attempts to find the future value of a stock using trends and patterns of the stock price from the past. Despite several methods and models employed in predicting stock market behaviour, accurate prediction still remains a problem which can be solved through newer and more efficient means. We chose a combination of Machine Learning methods to generate a successful prediction model for BANKNITY, by training these ML algorithms based upon economic relationship between fundamental macro-economic indicators and bank's stocks performance.

Literature Review

There are two distinct trading philosophies that are used while trading in stocks or index in the financial markets: Fundamental Analysis and Technical Analysis. Most of the reviewed literatures have focused on using ML techniques for Technical Analysis predictions for stocks and indices. In such analysis the input data into the algorithms usually consists of daily stock returns (time series) and daily volatility parameters. One of the recent research articles is the prediction of technical parameters for stocks and stock indices by comparing four ML prediction models: Artificial Neural network (ANN), Support Vector Machine (SVM), Random Forest and Naïve Bayes (Patel, Shah, Thakkar, & Kotecha, 6 August 2014). In this paper 10 technical parameters[1] were computed for a historical period of 10 years, which was used to learn through trend deterministic data.

In the Fundamental Analysis space, researches have been published for predicting the stock and indices values, using regression modelling on fundamental predictors. The most important literature result in this space was the CAPM (Capital Asset Pricing Model) model, which was introduced by Jack Treynor, William F. Sharpe, John Lintner and Jan Mossin (published in the Journal of Finance, 1962). In CAPM, return on assets is determined by the market factor (beta), which entails the systematic risk of the asset vis-à-vis broader market returns. Subsequently the Fama and French (1993) three factor asset pricing model was developed due to increasing empirical evidence that the CAPM performed poorly in explaining realised returns. Fama and French studied the role of size, Earnings/Price ratio (E/P), leverage and book-to-market equity ratio in explaining average returns for NYSE and NASDAQ stocks. Fama and French then extended the study to use a time-series regression approach, and constructed a three factor asset pricing model for stocks, that includes the conventional market beta factor, and two additional risk factors related to size and book-to-market equity. The expanded model paved way for several researchers to further analyse fundamental analysis techniques to explain stock returns. Most of these studies were targeted at analysing performance of stocks owing to its fundamental characteristics such as financial statement ratios (P/E, B/E etc.) and its comparison with the broader market indicators. Experiments have been conducted to study the impact of macro-economic financial indicators on the market returns. In one model prediction system for TOPIX (Tokyo Stock Exchange Prices Indexes), Kimoto & Asakawa et. al. postulated how a layered modular neural networks can be used to predict the profit of TOPIX by using Interest rate, Foreign exchange, Turnover and DJIA levels.

Despite the abundance of research literature available in the area of fundamental analysis predictions, and in the area of ML applications in technical analysis, there has been a knowledge gap in the applications of Machine Learning in the Fundamental analysis domain. Especially, the comparison of ML techniques for predictions, by using host of financial macro-economic indicators (Fundamental analysis) is under-explored.

[1] Simple moving average (SMA), Weighted moving average(WMA), Momentum, Relative Strength Index (RSI), Moving average convergence divergence (MACD), Larry William's R%, A/D Oscillator, Commodity Channel Index (CCI)

Additionally, it is observed that the review literatures have focused on individual stocks or the entire stock market index as a whole. There has been little emphasis on researching pattern for a sectoral sub-index portfolio such as Banking Index. Banking Index is a mix of banking and financial stocks, which is different from the stock market Index, because of specific business factors. Such an area of study demands specific macro-economic indicators, which are not direct predictors for the stock market as a whole.

Based on the literature review, there exists an opportunity to study the applications of ML techniques on the Banking Index, by using fundamental macro-economic variables; and assess the performance of these ML algorithms for the data under study.

Research Objectives

The research intends to examine the sensitivity of BANKNIFTY index, towards several systematic and unsystematic factors, by using data science techniques. It aims to help the investors (or portfolio managers), practitioners, and researchers in this area, to gauge the impact of changes in several macroeconomic indicators on the Bank NIFTY portfolio. The detailed list of objectives and project deliverable are given below.

Objective # 1
To investigate whether NIFTY BANK is dependent on macroeconomic factors[2] using several supervised and unsupervised Machine Learning methods.

Objective # 2
To predict the performance of NIFTY BANK returns using machine learning algorithms.

Objective # 3
To compare the performance and accuracy of different machine learning models, and provide decision making suggestion for investors.

Hypotheses

- Is NIFTY BANK return dependent on the macroeconomic factors?
- What is the degree of relationship between NIFTY BANK and the macroeconomic factors?
- Can the ML algorithms provide an efficient and usable decision system, to predict BANKNIFTY returns?

Significance of the study

The study is useful for various users and practitioners who want to predict the returns of BANK NIFTY in an accurate and reliable manner.

- Assist investors looking to invest; or to manage risk of existing investments
- Aid portfolio managers in evaluating investment options
- Help regulators (RBI & SEBI) & policy makers to gauge health of the sector and formulate policies
- Help stakeholders & management of the companies to prepare future road maps (expansion, capital market activities etc.)
- Useful for researchers and scholars working in the focus area

Scope of Research

In Scope

- Project study is limited to NIFTY BANK Index
- Specific macroeconomic factors taken into consideration for ML algorithms

Out of Scope

- Technical analysis using charts and indicators, and time series prediction models
- Analysis of holdings of Index using Financial Statements

[2] Macroeconomic factors: Index of Industrial Production (IIP), Inflation (WPI), Interest Rate, M3 Money Supply, DJIA Index, FII flows, USD / INR FX Rate, Crude Oil Price, Repo Rate, Foreign Exchange Reserves and Gold Price.

Research Methodology

Sampling Technique	Stratified and Random Sampling
Sampling Frame	Time series normalized data is used for all variables for historical dates (November 2013 to October 2018). All the variables have been converted to monthly values (from daily values).
Sampling Size	• Monthly data for BANKNIFTY and predictor variables is selected for the period 01-Oct-2013 to 01-October-2018 (60 months). The period of five years is considered primarily to address the adequacy of sample size as well as to ensure that sufficient recent observations are available to train the algorithms. • The observed data (sample size 60 months) has been randomly partitioned using 75% and 25% weightage for training and test data respectively. • We standardize or normalize the data for these predictor variables, to achieve a common scale without losing information, and to improve the prediction accuracy using ML algorithms. Normalization is done by rescaling all the values of each variable to [0,1] range.
Data Collection Methods	• Historical data of macroeconomic and financial indicators • Historical values of NIFTY BANK Index
Statistical Methodologies	• Multiple Linear Regression and Correlation Analysis (Scatter Matrix) • Bayesian Model Averaging • K Nearest Neighbours (KNN) Clustering, • Decision Trees - Binomial Tree Classification, • Random Forest Classification, • Support Vector Machines (SVM)
Statistical Tools	R, Python and Microsoft Excel

Data Sourcing and Enrichment

Data Source and Time Series

The BANK NIFTY Index value is the indicator of level of the Bank Nifty portfolio. It represents the cumulative value of prices of individual stocks in the NIFTY Bank Index, computed using the free float market capitalization method.

The time series data of daily close of day Index value has been collected over the period of study (1-October-2013 to 01-October-2018). Monthly return of NIFTYBANK is calculated using the following formula:

$$r_t = \left[\frac{NIFTYBANK(month\ i)}{NIFTYBANK(month\ i-1)} - 1\right] * 100$$

It is a difficult exercise to predict the NIFTYBANK value with a high precision, due to prevalence of several noise factors. However, converting the returns into discrete variable with finite outcomes helps decrease the influence of noise and still predict the performance of the index with reasonable accuracy. For categorical analysis, the NIFTYBANK returns data was further bucketed into levels of return: High (gain of more than 1%), Medium (-1% < return < 1%), and Low (loss of more than 1%).

Normalization of Data

We standardize or normalize the data for these predictor variables, to achieve a common scale without losing information, and to improve the prediction accuracy using ML algorithms. Normalization is done by rescaling all the values of each variable to [0,1] range. Normalization formula is as shown below:

$$Normalized\ Value_i = \frac{(Observed\ Value_i - Minimum\ observed\ value)}{(Maximum\ observed\ value - Minimum\ observed\ value)}$$

Many of these predictor variables have been used with a lagged value, in the prediction algorithm because of the delayed impact of the variables in impacting the bank stock's performance.

Selection of features

Several variables are selected based on the existing literature on economic relationship between these variables and performance of Bank stocks. These lists of features are explained below.

- **Index of Industrial Production (IIP)** is a number (ratio) which measures the growth of production in several sectors in the economy. An increase in IIP usually indicates increase in production levels and increased economic activity. A positive rate of change of IIP – ceteris paribus – leads to increase in financial borrowing and lending, which can subsequently result in improved earnings for Banks.
- **Dow Jones Industrial Average (DJIA)** is a US stock market index that replicates the value of 30 large public owned companies in US. In other words, DJIA is one of the prominent stock market indexes of US. Based on empirical evidence, DJIA has a significant impact on the movements of several large stock markets of the world, including NIFTY. Significant increase in DJIA levels, usually boost the NIFTY level, and the stock prices of major stocks. DJIA has been incorporated as a feature, by considering its monthly return as a predictor.
- **Foreign Institutional Investment (FII)** refers to the investment in India in securities, made by an investor established or incorporated outside of India. The Indian stock market (NIFTY) sees participation from both domestic as well as foreign investors, and the level of participation from FIIs is substantial. FII has been incorporated into the study by taking into consideration the monthly FII flow measured in crores of Rupees. FII over a month can either be positive (net FII inflow) or negative (net FII outflow) number. Ceteris paribus, the stock market level goes up when there is a substantial FII inflow, and vice-versa.
- **USD INR Exchange Rate (FX Rate)** denotes the value of one US dollar in terms of Indian rupees. An increase in the FX rate denotes depreciation of INR w.r.t. USD, and vice-versa. FX rate has been considered in the exercise by calculating the percentage change in FX rate for each month. FX rate can have complex effects on the Indian Stock Market. Although a high FX rate is beneficial for the exporters, but it can also lead to inflation because of costly imports (especially oil imports) and can be detrimental for the stock market.
- **Repo Rate (Repo)** is the interest rate at which Reserve Bank of India lends money to commercial banks in the event of any shortfall of funds. During periods of high inflation, RBI sets a high repo rate to increase the cost of funds, and thereby decrease the money supply in the market. And during periods of low inflation, RBI can decrease the repo rate to increase the money supply in the market. Usually, a decrease in repo rate, boosts the economic activities due to low cost of funds, and provides an impetus to the stock markets. However, it has been observed that a change in repo rate does not translate into change in cost of funds immediately, because the transmission of rate cuts may take longer time for commercial banks. For our experiment, a lagged repo rate can be considered to include the effect of impact of change in repo rates in subsequent months.
- **M3 Money Supply (M3)** is a measure of money supply in the economy, which is an aggregate of sum of currency with public, demand deposits in all commercial banks, deposits with RBI and time deposits with commercial banks. M3 measure is also known as 'Broad Money' or 'Money aggregate'. A higher M3 value indicated surplus availability of liquidity and lower cost of funds, which can provide a stimulus to the economic activity and GDP of the economy. In our study, we have tried to measure the impact of monthly change in M3 on the monthly returns of BANKNIFTY.
- **Wholesale Price Index (WPI)** is an indicator of country's level of inflation, and is measured as the average price change of representative basket of wholesale goods. A high inflation index, decreases the real rate of return on investments, and can be detrimental for investors seeking returns from the stock market. Additionally, high WPI can also lead to monetary tightening by RBI, which can subsequently have an impact on the GDP and IIP of the country. Monthly WPI has been recorded along with the change in month-on-month WPI, to gauge an impact of change in inflation on the BANKNIFTY returns.
- **Foreign Exchange Reserves (FER)** also called as Forex reserves, is a measure of money or other assets held by the central bank of an economy. Forex reserves held by RBI are measured in US dollars. It includes foreign currency assets, fold reserves, Special Drawing Rights (SDR), and IMF reserve positions. Higher reserves, ensures a stable USD/INR exchange rate, and can favourably impact the balance of payments of the economy. An increase in FER is seen favourable for trade balances and foreign investments, and can provide a boost to the GDP of the economy. FER month end reserves have been converted to monthly change in FER, for analysis purposes.
- **Crude Oil Price (Brent)** is the leading global benchmark for oil prices, and serves as an indicator of oil price change. It plays an important role in the Indian economy, because India is highly dependent on crude oil imports for its energy needs, and the volatility of oil price impacts the current account balances. Higher oil price, also results in higher cost of production for industries and can lead to lower IIP. The oil price has been included in the analysis in the form of monthly change in oil price as a predictor variable.

Analysis of Results

Scatter Plot of Variables and Correlation Analysis

The collected historical data of all the ten variables are used to create the scatter matrix, which contains scatter plots between each pair of variables. Each pair-wise scatter plot helps to visualize any pattern of relation between the variables.

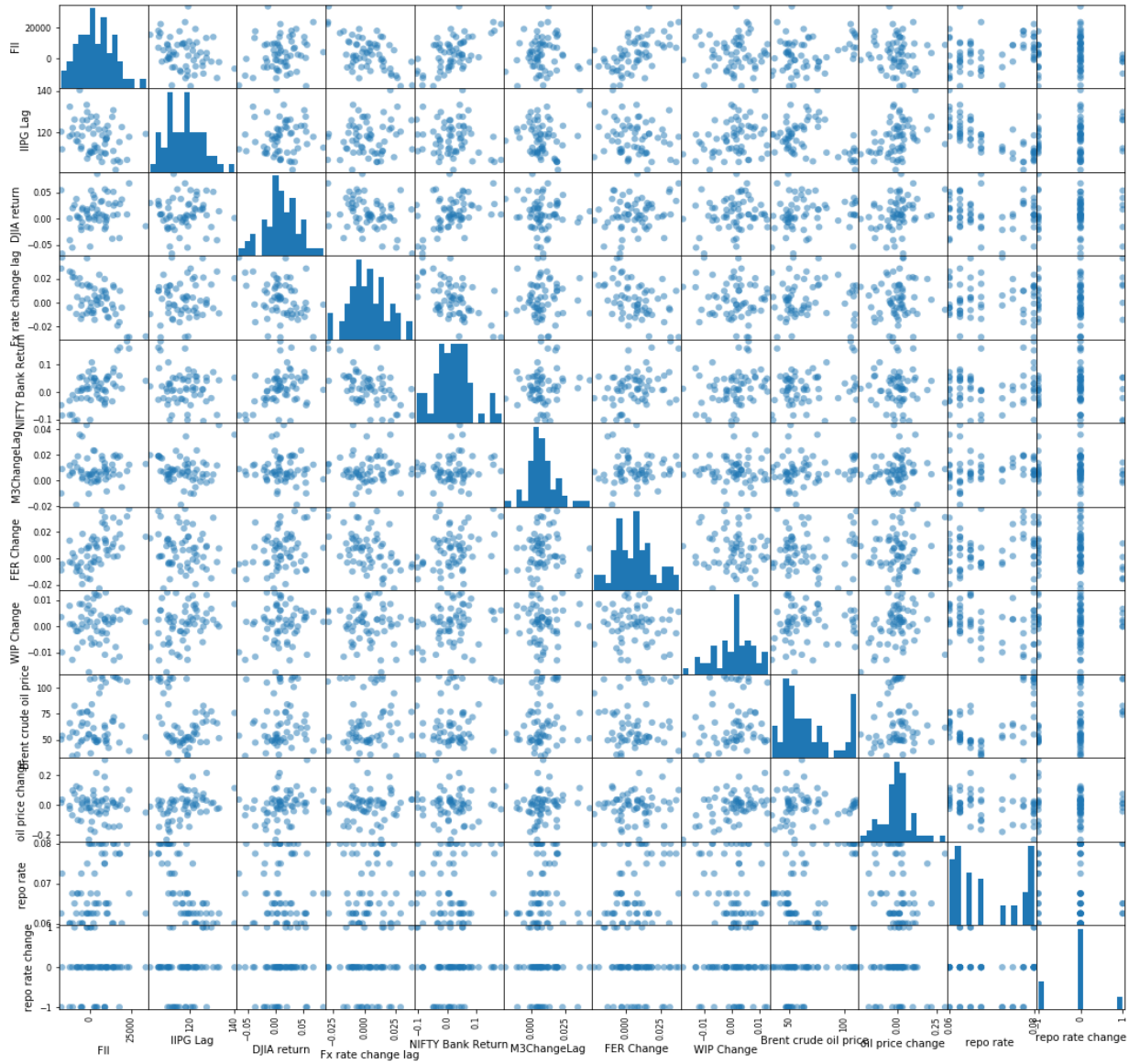


Figure 1: Scatter Matrix for each Input Variable

Observations

- Positive relation between FII and Bank NIFTY returns (Pearson Correlation Coefficient ρ = 0.6)
- Positive correlation between change in FER and Bank NIFTY returns (ρ = 0.37); and between DJIA returns and Bank NIFTY returns (ρ = 0.4)
- No correlation between change in Fx rate and Bank NIFTY reuturns. But there is a negative relation between lagged value of change in Fx rate and Bank NIFTY returns (ρ = - 0.54)
- No visible correlation of IIP index or Oil Price or M3 money supply with Bank NIFTY returns

Multiple Regression

$$Bank\ NIFTY\ return = \beta_0 + \beta_1 * FII - \beta_2 * Fx\ Rate\ Change\ Lag + \beta_3 * DJIA\ return$$

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.052e-03  7.365e-03   0.958  0.34240
## FII              2.402e-06  7.053e-07   3.406  0.00123 **
## FXRateChangeLag -9.697e-01  5.089e-01  -1.905  0.06186 .
## DJIAReturn       5.100e-01  2.134e-01   2.390  0.02025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04675 on 56 degrees of freedom
## Multiple R-squared:  0.4649, Adjusted R-squared:  0.4363
## F-statistic: 16.22 on 3 and 56 DF,  p-value: 1.046e-07
```

Figure 2: R output after dropping insignificant variables (Multiple Regression)

Observations

- After doing several iterations of multiple regression and dropping the insignificant variables, we find that FII, Fx rate change lag and DJIA return are significant variables in determining the Bank NIFTY returns.
- FII and Return on DJIA have positive slope coefficients. An Increase in FII or positive returns on DJIA, leads to higher returns on NIFTY Bank. Lagged change in USD/INR has a negative coefficient. When INR depreciates w.r.t. USD, the return on NIFTY Bank decreases.
- The R square value is 46%. Rest of the variation in Bank NIFTY return is explained by several other factors, which are mostly micro and firm specific factors. Inclusion of all the factors in a closed form equation is highly infeasible.

Bayesian Model Averaging

```
  11  models were selected
 Best  5  models (cumulative posterior probability =  0.7987 ):

                p!=0    EV         SD         model 1     model 2     model 3     model 4     model 5
Intercept       100.0   4.441e-03  2.656e-02   6.857e-04   7.052e-03   1.277e-02   5.173e-04  -5.587e-03
FII             100.0   2.759e-06  7.746e-07   3.136e-06   2.402e-06   2.458e-06   2.297e-06   3.065e-06
IIPG_Lag          6.8  -6.463e-06  2.124e-04           .           .           .           .           .
DJIA_return      83.5   4.749e-01  2.905e-01   6.085e-01   5.100e-01           .   5.365e-01   6.366e-01
M3ChangeLag      20.2   1.536e-01  4.017e-01           .           .           .   8.041e-01   7.433e-01
FxRateChangeLag  52.1  -5.484e-01  6.485e-01           .  -9.697e-01  -1.264e+00  -1.008e+00           .
FER_Change        7.0  -7.418e-03  1.515e-01           .           .           .           .           .

nVar                                             2           3           2           4           3
r2                                           0.430       0.465       0.410       0.484       0.446
BIC                                     -2.556e+01  -2.524e+01  -2.350e+01  -2.328e+01  -2.318e+01
post prob                                    0.282       0.240       0.101       0.090       0.086
```

Figure 3: R output (Bayesian Model Averaging)

Observations

- Several models are feasible, each of which considers a mix of predicted variables from the list of 10 features. The best 5 models account for a high level of cumulative posterior probability.
- The model averaging provides a range of coefficient estimates that can be used to explain the regression model.

K Nearest Neighbour

K = 1

```
table(KNN_model_1,actual_returns)

mean(KNN_model_1==actual_returns)

##          actual_returns
## KNN_model High Low Medium
##    High      5   1      0
##    Low       3   3      0
##    Medium    1   1      1

## [1] 0.6
```

K = 5

```
table(KNN_model_5,actual_returns)

mean(KNN_model_5==actual_returns)

##          actual_returns
## KNN_model High Low Medium
##    High      8   2      1
##    Low       1   3      0
##    Medium    0   0      0

## [1] 0.7333333
```

Figure 4: Confusion Matrix & Accuracy (K Nearest Neighbour)

Observations

- The KNN classification algorithm was used on the data set containing all the normalized predictor variables (10) and the labelled 'BANKNIFTY return'.
- Several iterations were conducted for k-means ranging from 1 to 7.
- Our trained models provided a high level of accuracy of 60% and 73%, for k-means of 1 and 5 respectively.

Binomial Classification Tree

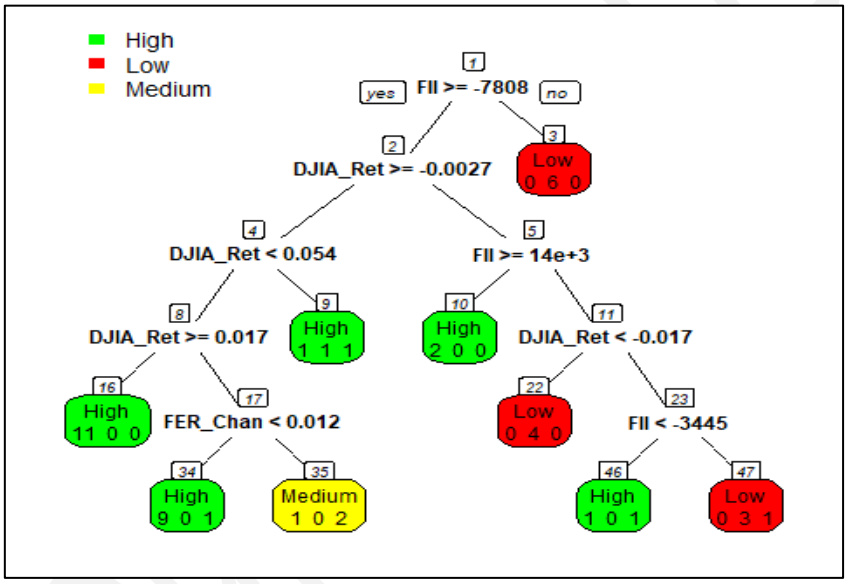


Figure 5: Classification Tree with Categories High, Medium, & Low

Observations

- Classification tree (decision tree) algorithm is used to create a model for classifying NIFTYBANK returns into three different categories High, Low and Medium.
- Only select variables are considered for this technique, to keep the decision tree from becoming complex: DJIA Returns, Repo rate, FII, and FER Change. The model returned a high accuracy of 80%.
- Pruning is required to eliminate nodes which have less than 4 observations. Thus three nodes, i.e. 10, 25 and 46 are removed, to avoid over fitting of data, and generate rational and practical decision rules, as shown below.

#	Rule	Classification Label
1	FII outflow > Rs. 7,808 Cr	Low
2	FII inflow > Rs. 13,522 Cr & DJIA Returns < 0.27%	High
3	FII inflow < Rs. 13,522 Cr & DJIA Returns < -1.7%	Low
4	FII outflow > Rs. 3,445 Cr & DJIA Returns > -1.7%	Low

5	FII outflow < Rs. 7,808 Cr & DJIA Returns > 5.4%	High
6	FII outflow < Rs. 7,808 Cr & DJIA Returns is in [1.7%, 5.4%)	High
7	FII outflow < Rs. 7,808 Cr & DJIA Returns < 1.7% & FER Change < 1.2%	High
8	FII outflow < Rs. 7,808 Cr & DJIA Returns < 1.7% & FER Change > 1.2%	Medium

Table 1: Decision Tree Rules with Classification Labels High, Medium, & Low

Random Forest

```
Forest_model

##
## Call:
##  randomForest(formula = Categorical_NIFTYBank_Return ~ ., data = forest
_data, method = "class", ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 26.67%
## Confusion matrix:
##        High Low Medium class.error
## High     32   1      0  0.03030303
## Low       6  12      1  0.36842105
## Medium    5   3      0  1.00000000
```

Figure 6: Confusion Matrix and Out of Bound (OOB) Estimate of the Model

Observations

- The application of Decision tree algorithm generated good results. However, there are several other macro factors that had not been considered for identifying tree structure, to avoid over fitting. Hence an ensemble technique is used to aggregate results from several trees.
- Random Forest, which is a collection of tree structured classifiers, was modelled to include 500 trees in the ensemble methods, as well to include all the available features.
- The predictions have a high accuracy of 73.33%, with an out-of-bound error estimate of 26.67%.

Support Vector Machine

		Predicted		
		HIGH	MEDIUM	LOW
Actual	HIGH	8	1	0
	MEDIUM	4	2	0
	LOW	3	0	0
Accuracy	61.11%			

Table 2: Confusion Matrix (Support Vector Machine)

	precision	recall	f1-score	support
High	0.53	0.89	0.67	9
Low	0.67	0.33	0.44	6
Medium	0.00	0.00	0.00	3
avg / total	0.49	0.56	0.48	18

Figure 7: Classification Report (Support Vector Machine)

Observations

- The SVM classification algorithm was used on the data set containing 7 normalized predictor variables and the labelled 'BANKNIFTY return'.
- It gives us a satisfactory result, whereas other models, which we have used for prediction, are producing better results.
- It is not at all able to predict Class "Low" correctly.

Comparative Study of Effectiveness of Methodologies

Machine Learning Method	Accuracy (Range for 20 test samples)
Multiple Regression	46.49% (41% - 48%) R square
KNN	73.33% (53% - 80%)
Binomial Tree	80.00% (67% - 80%)
Random Forest	73.33% (60% - 80%)
SVM	61.11% (40% - 70%)

Table 3: Summary of Machine Learning Methods

Discussion of Findings

- Regression and model averaging have limited success in predicting the returns of the portfolio. However they do provide the coefficient estimates which can be used to determine the sensitivity of the returns to movements in individual factors (provided other factors remain constant)
- All the ML models are able to provide an average to high level of accuracy in determining the category of Bank NIFTY returns (High, Medium and Low). Particularly, KNN, Binomial Tree and Random Forest have predicted the class of returns more consistently and accurately as compared to other models.
- The models (KNN, Tree and Random Forest) performed consistently, with a high accuracy range obtained from repeated training samples from the observations (20 iterations).
- From the test results, it was observed that the ML models work best for high (return > 1%) and low returns (return < 1%) categories, as compared to medium returns category. Also the performance of the model was adversely effected, when the categories of returns was increased from 3 labels to 5 labels. This indicated that it is difficult to predict the returns with high accuracy due to level of noise prevalent in returns data.

Limitations

- Prediction capabilities of the models may change when data from different time periods are considered. For instance, model would have lower accuracy if we consider the recessionary period from 2008-2011.
- The model fails to accurately predict the impact of unsystematic factors and news items. For instance, the recent change of RBI governor, or corruption allegations on Chanda Kochhar, the MD of ICICI Bank can impact the BANK NIFTY but might not be predicted
- Model will depend on accurate prediction of the selected factors
- Model is not immune to firm specific isolated risk factors.

Conclusion

The ML models depict a high success rate in predicting the class of NIFTY BANK returns. It is noteworthy, that a single model cannot consistently perform well, but a host of models can be used subjectively to arrive at the best prediction. The study can be used practically in several fields for benefits of different users, such as Investors, stake holders and portfolio managers. Particularly, it can be an effective tool to augment the existing fundamental/technical models used by an investment manager to make investment decisions. The study can also be expanded to cover more variables, and other algorithms (such as learning algorithm, ANN etc.) to provide a holistic decision support system for the practitioners.

However the models should not be treated as a one stop solution. They need to be trained and recalibrated periodically to account for changes in macro environment and improve the prediction efficiency. Specifically, any paradigm shift or disruptive change in the macro factors should be studied independently and then incorporated into the model. Moreover, the algorithms cannot replace the wisdom of an experienced analyst or portfolio manager; but it can supplement the capabilities of existing tools to make faster decisions with higher conviction.

Appendix I: Overview of Methods and Models

The project intends to analyse the Performance and Risk-metrics of the index in study using various statistical methodologies outlines below.

Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. It is a technique to model "Effect" using "Cause(s)" or Proxy of "Cause(s)". The overall idea of regression is to examine two things: First, does a set of predictor variables do a good job in predicting an outcome (dependent) variable? Second, which variables in particular are significant predictors of the outcome variable, and in what way do they-indicated by the magnitude and sign of the beta estimates-impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. A form of a multiple regression with two independent variables is shown below

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

There are 3 terms related to Linear Regression in order to answer 3 important questions.

- *Correlation* - Is there a linear relationship between cause and effect variables? If it exists, how strong is it?
- *Regression Equation* - If linear relationship is strong, give an estimated linear relationship (ß) of cause and effect variable based upon data.
- *Coefficient of Determination (R^2)* - How good is the regression fit? Proportion of variation in Y explained by variation in explanatory variable(s) through regression relation.

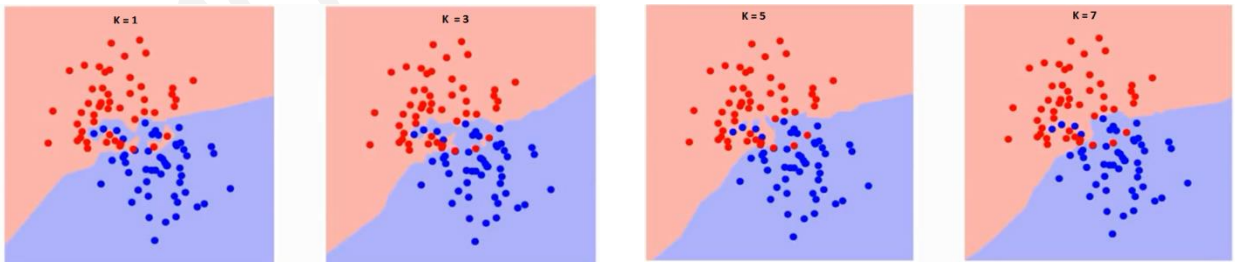
K Nearest Neighbour Classification (KNN)

The k-nearest neighbors (kNN) is one of the oldest and simplest methods for pattern classification. Nevertheless, it often yields competitive results, and in certain domains, when cleverly combined with prior knowledge. In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. Similarity is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

K in KNN is a hyperparameter that you, as a designer, must pick in order to get the best possible fit for the data set. Intuitively, you can think of K as controlling the shape of the decision boundary we talked about earlier.

When K is small, we are restraining the region of a given prediction and forcing our classifier to be "more blind" to the overall distribution. A small value for K provides the most flexible fit, which will have low bias but high variance. Graphically, our decision boundary will be more jagged.



On the other hand, a higher K averages more voters in each prediction and hence is more resilient to outliers. Larger values of K will have smoother decision boundaries which means lower variance but increased bias.

Decision Trees – Binomial Tree Classification (DT)

Decision Tree is one of the most widely used and practical methods for classification through inductive inference. It is a method for approximating discrete-valued target functions. It uses a greedy, top-down recursive approach to bucket each observation under a classification (predicted value).

The goal of classification trees is to predict or explain responses on a categorical dependent variable. Starting from the root of a tree, the feature space 'X', containing all examples is split recursively into subsets, usually

two at a time. Each split depends on the value of only a unique variable of input *x*. If *x* is categorical, the split is of the form $x \in A$ or $x \notin A$ where *A* is subset of X.

The goodness of split is measured by an *impurity function* (i(t)) defined for each node (t). The basic idea is to choose a split such that the child nodes are purer than their parent node. The split continues till the end subsets (leaf nodes) are 'pure'; that is till one class dominates. The prominent impurity functions that are used are:

- **Entropy** is a measure of impurity or homogeneity of data. During the construction of the tree, attribute selection is done, based on the attribute that leads to maximum reduction in entropy, i.e. it provides the highest *information gain*. Entropy is measured as shown below

$$i(t) = \sum_{j=1}^{K} -Pr(j/t)\log[Pr(j/t)]$$

  Where, i(t) is the impurity measure, and $Pr(j/t)$ is the estimated probability of class *j* within node *t*,

- **Gini Index** is another measure of the impurity, used in building the decision tree using CART algorithm.

$$i(t) = \sum_{j=1}^{K} Pr(j/t)[1-Pr(j/t)]$$

Random Forest

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

Random forest aims to reduce the correlation issue by choosing only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by setting a stopping criteria for node splits

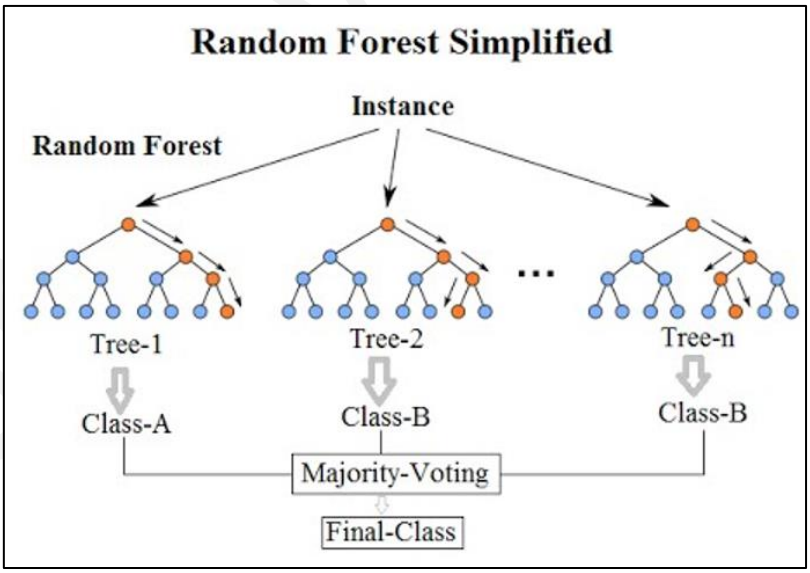


Figure 8: Random Forest

Support Vector Machine

Support Vector Machine are supervised machine learning algorithms used mainly for classification and regression tasks. When a SVM is used for classification, it's called Support Vector Classifier (SVC). Similarly, for regression it's called Support Vector Regressor (SVR).

A support vector machine, works to separate the pattern in the data by drawing a linear separable hyperplane in high dimensional space.

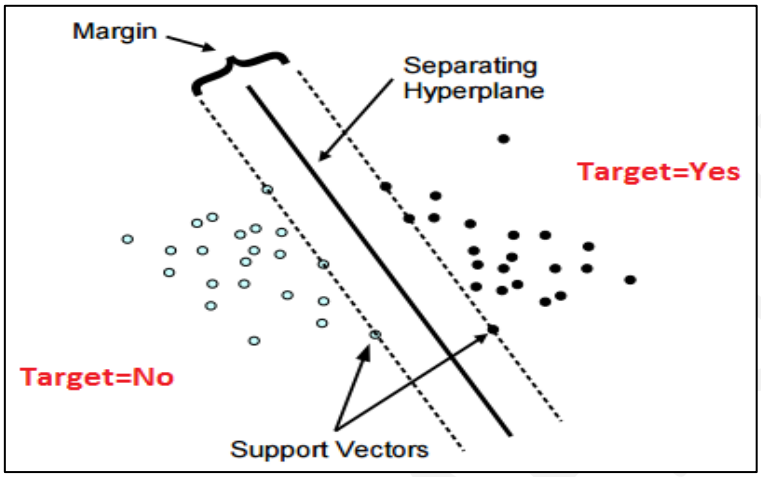


Figure 9: Support Vector Machine

Appendix II: Data and R Program Code Files

Project Data.csv

KNN.R

Decision Tree.R

Regression with Model Averaging.R

Random Forest.R

Appendix III: Data Sources

#	Data	URL
1	NIFTY 50, NIFTY Bank, NIFTY Auto	https://www.nseindia.com/products/content/equities/indices/historical_index_data.htm
2	FII Trading Activity	https://www.moneycontrol.com/stocks/marketstats/activity.php?flag=FII&month=201501201810
3	USD INR FX Rate	https://in.investing.com/currencies/usd-inr-historical-data
4	Historical Fund NAV	https://www.amfiindia.com/nav-history-download
5	Index of Industrial Production	https://data.gov.in/catalog/all-india-index-industrial-production-base-2011-12100
6	Dow Jones Industrial Average	https://finance.yahoo.com/quote/%5EDJI/history/
7	Repo Rate	https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home
8	Brent Crude Oil Daily EOD Price	https://www.macrotrends.net/2480/brent-crude-oil-prices-10-year-daily-chart

Appendix IV: References

#	Title	Author	Publication
1	An augmented Fama and French model for emerging stock market	Sunil Bundoo, University of Mauritius	Applied Economic Letters - Dec 2008
2	News Analytics and Dual Sentiment Analysis for Stock Market Prediction	Naren, Sangavi, Revathy & Srimathi, SASTRA Deemed University, Thanjavur, India	Conference Paper - Dec 2017
3	Volatility and stock prices: Implications from a production model of asset pricing	Parantap Basu, Prodyot Samanta, Department of Economics, Fordham University, Bronx, NY 10458, USA	Elsevier Economics Letters 70 (2001) 229-235
4	Textual Analysis of Stock Market Prediction Using Financial News Articles	Robert Schumaker, Hsinchun Chen, University of Arizona	Americas Conference on Information Systems - Dec 2006 Proceedings
5	Application of machine learning techniques for stock market prediction	Bin Weng, Auburn University, Auburn, Alabama	PhD Dissertation - May 2017
6	Stock Trend Prediction Using News Sentiment Analysis	Kalyani Joshi, Prof. Bharathi H. N., Prof. Jyothi Rao, Department of Computer Engineering, KJSCE, Mumbai	International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 3, June 2016
7	Stock Market Prediction System with Modular Neural Networks	Kimoto, T., & Asakawa, K.	The Nikko Securities Co., Ltd., 3-1 Marunouchi 3-Chome, Chiyoda-Ku Tokyo 100, Japan
8	Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques	Patel, J., Shah, S., Thakkar, P., & Kotecha, K.	Elsevier - August 2014