Investment Portfolio Management with Machine Learning & Predictive Analytics

APDS 2 Group V

Sriniwas M
Amit A
Aditya P
Anand S
Shweta B
Arvind K
Sambita C
Mayank J
Surjeet R
Anuraag T

Agenda

01

Problem Statement

Significance of the Study | Objectives | Data Enrichment

02

Methodologies Used

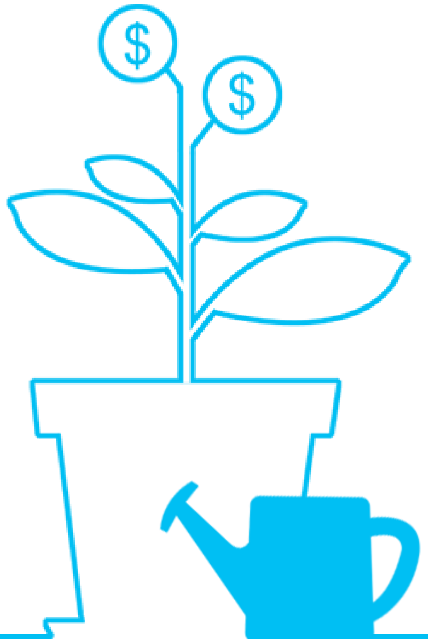
Correlation Analysis | Multiple Linear Regression |
Bayesian Model Average | K-Nearest Neighbors |
Support Vector Machine | Binomial Classification Tree |
Binomial Classification Tree after Pruning | Random Forest

03

Analysis of Results

Discussion of Findings | Limitations | Conclusion

01



Problem Statement

Significance of the Study | Objectives | Data Enrichment

Problem Statement

Predict the returns of a portfolio, accurately & reliably, using macroeconomic factors

Significance of the Study

- 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 **management of the companies** to prepare future road maps (expansion, capital market activities etc.)
- Useful for **researchers** and **scholars** working in the focus area

Objectives

- To investigate whether portfolio in study is dependent on **macroeconomic factors** using machine learning
- To build an efficient and usable decision system, which helps **predict the performance** of the portfolio in study

Feature Selection

- Based on existing economic theories
  - Index of Industrial Production (IIP)
  - Inflation (WPI)
  - Repo Rate
  - M3 Money Supply
  - DJIA Index
  - FII Flow
  - USD / INR FX Rate
  - Crude Oil Price
  - Foreign Exchange Reserves (FER)
  - Gold Price

Portfolio in Study

NSE: BANKNIFTY Index

Why have we chosen this portfolio?

1

Significant

Represents the 12 most liquid and large capitalized banking stocks from NSE

2

Prevalent

Major constitute of several Funds and Portfolios (as part / whole)

3

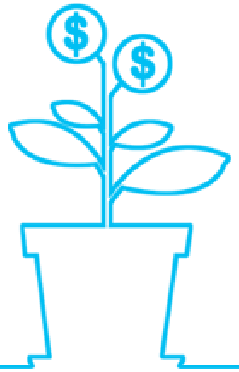
Standard

Widely mimicked benchmark of performance for investments

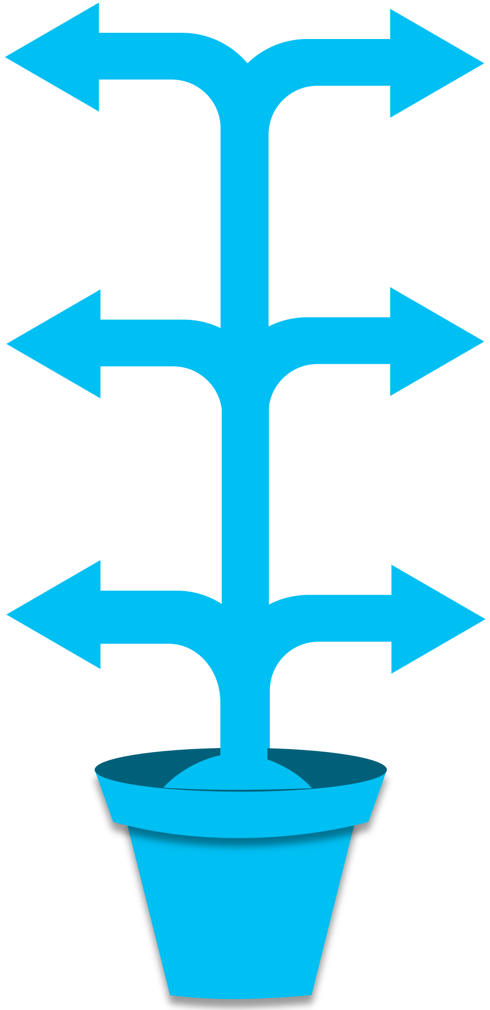
4

Representative

This study can be easily expanded to other portfolios as well



Data Enrichment



Time Series Data

For Nifty Bank and all features
(EoD values from 01-Oct-2013 to 30-Sep-2018)

Rate of Returns

Nifty Bank month end data converted to monthly rate of returns

Categorization

Categorize Nifty Bank returns into
- HIGH (> 1% return)
- LOW (< -1% return)
- MEDIUM (-1% to 1% return)

Normalization

Normalize all variables to rescale into [0,1] range

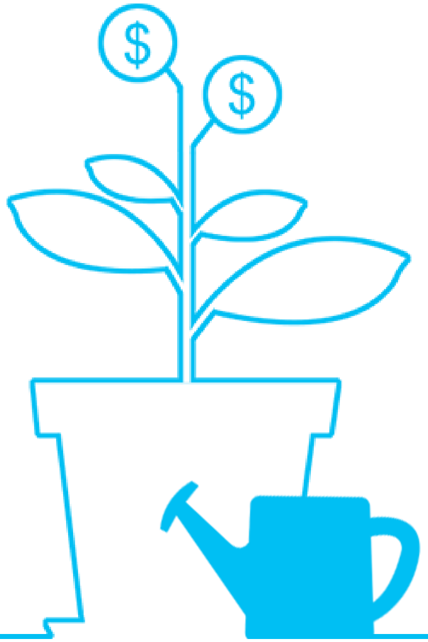
Lagged Data

Create time series for lagged effect of variables

Partitioning

Random training (75%) and test data (25%) partition (repeated test sampling)

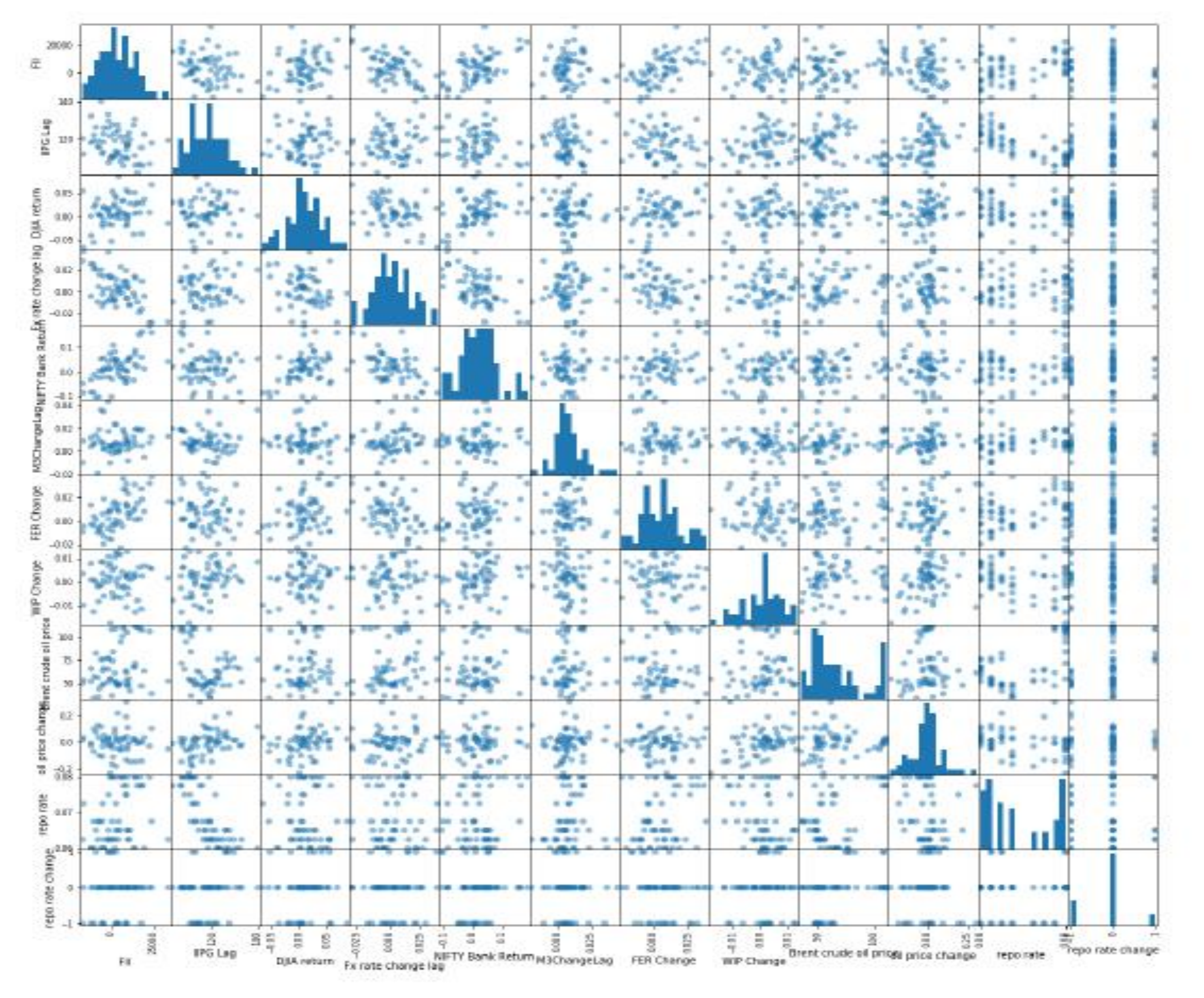
02



Methodologies Used

Correlation Analysis | Multiple Linear Regression | Bayesian Model Average | K-Nearest Neighbors | Support Vector Machine | Binomial Classification Tree | Binomial Classification Tree after Pruning | Random Forest

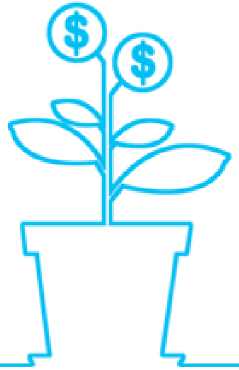
Correlation Analysis



Observations

- **Positive correlation** between Nifty Bank Returns &
  - FII (ρ = 0.6)
  - Change in Foreign Exchange Reserves (ρ = 0.37)
  - DJIA Returns (ρ = 0.4)
- **No correlation** between Nifty Bank Returns &
  - Change in FX Rate (ρ = 0.0)
  - But negative correlation with lagged value of Change in FX rate (ρ = – 0.54)
- **No visible correlation** between Nifty Bank Returns &
  - IIP
  - Crude Oil Price
  - M3 Money Supply

ρ = Pearson's Correlation Coefficient



Multiple Linear Regression

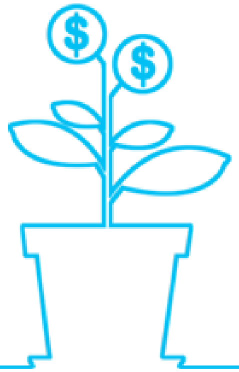
```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.052e-03  7.365e-03   0.958  0.34240
## FII             2.402e-06  7.053e-07   3.406  0.00123 **
## FXRateChangeLag -9.697e-01  5.089e-01  -1.905  0.06186 .
## DJIAReturn      5.100e-01  2.134e-01   2.390  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
```

Observations

- Increase in Nifty Bank Returns led by
  - An Increase in FII
  - Positive Returns on DJIA
- Decrease in Nifty Bank Returns when
  - INR depreciates w.r.t. USD

Drawback

Other macro factors ignored by the model due to constraints of regression



Bayesian Model Average

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

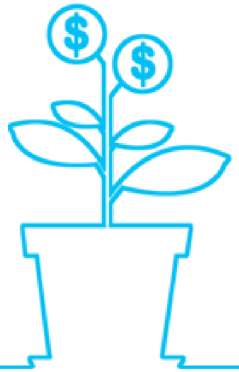
	p!=0	EV	SD	model 1	model 2	model 3	model 4	model 5
Intercept	100.0	4.441e-03	2.656e-02	6.857e-04	7.052e-03	1.277e-02	5.173e-04	-5.587e-03
FII	100.0	2.759e-06	7.746e-07	3.136e-06	2.402e-06	2.458e-06	2.297e-06	3.065e-06
IIPG_Lag	6.8	-6.463e-06	2.124e-04	.	.	.	.	.
DJIA_return	83.5	4.749e-01	2.905e-01	6.085e-01	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

Observations

- Several models generated, from a mix of feature set
- High cumulative posterior probability of ~ 80% for the top 5 models
- Provides range of coefficient estimates to explain the relationship

Drawback

Difficult to establish a closed form equation due to variety of factors and impact of noises



K-Nearest Neighbours

Confusion Matrix & Accuracy for 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
```

Confusion Matrix & Accuracy for 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
```

Observations

- All the 10 factors were used to train the model (75% of training data)
- High levels of accuracy for k = 1 and k = 5
- Consistent accuracy range (53% to 80%) and a high mean accuracy level of 67% from random training samples (20 iterations)

Support Vector Machine

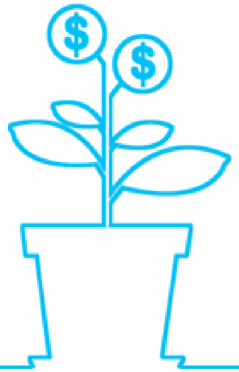
		Predicted		
		High	Medium	Low
	High	8	1	0
Actual	Medium	4	2	0
	Low	3	0	0
Accuracy	**61.11%**			

Observations

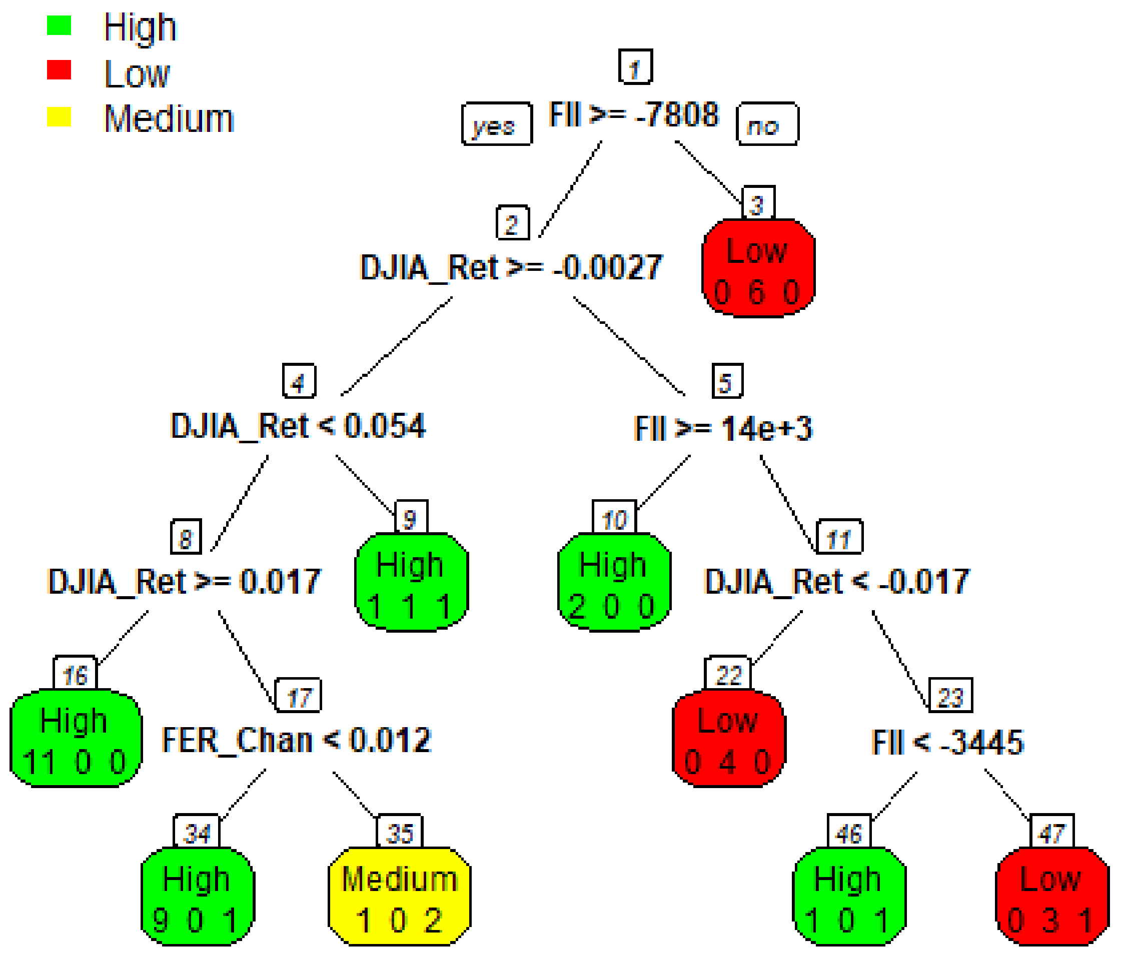
- SVM classification algorithm was used to train data set containing 7 normalized predictor variables and the labelled Nifty Bank Returns
- Satisfactory accuracy level of 61.11%

Drawback

Unable to predict the class ‘Low Returns’

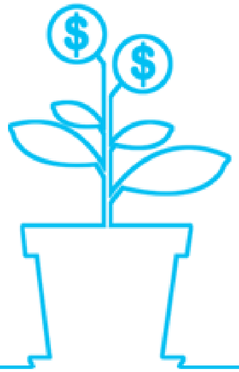


Binomial Classification Tree



Observations

- 4 prominent features were identified to construct the tree, to reduce complexity
- Simpler rules which are inline with existing macroeconomic theory
- Pruning is required to eliminate overfitting
- Nodes having less than 4 observations can be pruned (nodes: 10, 35, 46)



Binomial Classification Tree after Pruning

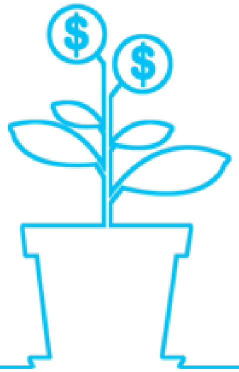
#	Rule	Classification Label
1	FII outflow > Rs. 7,808 Cr	Low
2	FII inflow > Rs. 13,522 Cr & DJIA Returns < 0.27%	High
3	FII inflow < Rs. 13,522 Cr & DJIA Returns < -1.7%	Low
4	FII outflow > Rs. 3,445 Cr & DJIA Returns > -1.7%	Low
5	FII outflow < Rs. 7,808 Cr & DJIA Returns > 5.4%	High
6	FII outflow < 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

Observations

- Set of decision rules which classify the category of Nifty Bank Returns
- High accuracy of 80%

Drawback

Only 4 features were used. Several iterations of the algorithm may generate different sets of rules.



Random Forest

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

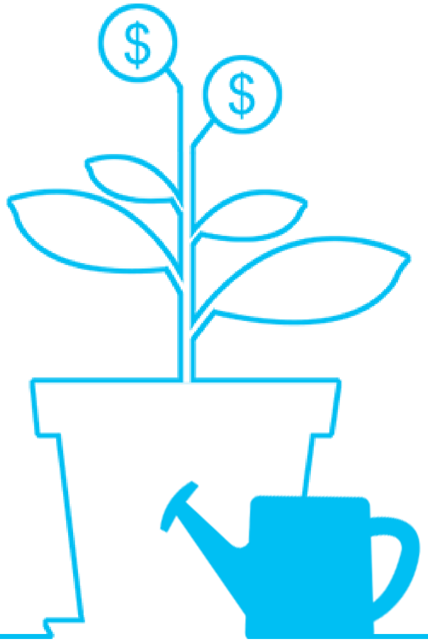
```
Forest_model

##
## 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
```

Observations

- An ensemble method to aggregate results from 500 trees, and all the available features
- High prediction accuracy of 73.33%
- Out of bound estimate of 26.67%
- Better prediction for High and Low classes

03



Analysis of Results

Discussion of Findings | Limitations | Conclusion

Discussion of Findings

Methods →	Regression	Model Averaging	KNN (K = 5)	SVM	Binomial Tree	Random Forest
Accuracy / Goodness of Fit	46%#	48%*	73%	61%	80%	73%
Algorithm	$E(Y) = X\beta$	Ensemble	Distance Measure	Maximize Margin	ID3 CART	Ensemble
Model Complexity Reduction	OLS	BIC	Increase K	Reduce C	Pruning	Pruning
Range of Accuracy^	41% - 48%	-	53% - 80%	40% - 70%	67% - 80%	60% - 80%

R Square
* Best Model
^ Range for 20 test samples

Prediction Capabilities

KNN, Binomial Tree and Random Forest have good prediction capabilities

Degree of Relationship

Regression (& model averaging) can augment the other methods to find degree of relationship.

Class Predictions

ML models work best for predicting High Return and Low Return classes.

Limitations

01 Prediction accuracy falls when different time periods are considered

Use Case

Recessionary period 2008-2010 needs a separate model

02 Does not account for isolated firm specific risk factors

Use Case

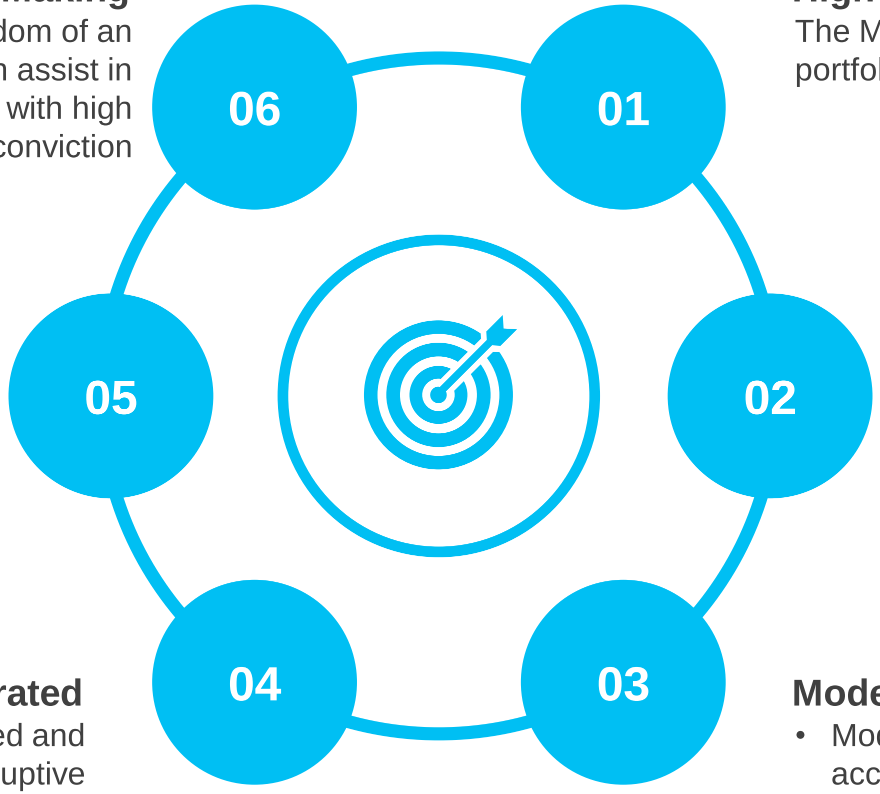
Impact of news of corruption allegations on MD of ICICI BANK

03 Fails to predict impact of unsystematic factors

Use Case

Impact of change of RBI Governor or impact of election results

Conclusions



High Accuracy of ML methods
The ML models predict the returns of a portfolio with high accuracy

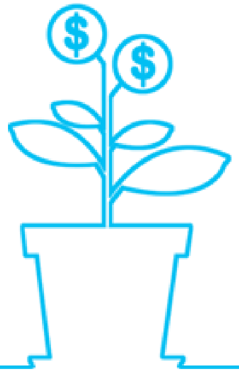
Practical Usage
They can be used practically in areas of portfolio management for investors and investment managers

Model can be expanded
- Model can be expanded to accommodate more features, and other algorithms
- A set of models provides better prediction than a single model

Model needs to be recalibrated
Model needs to be trained and recalibrated periodically for disruptive changes

Use along with Existing Models
Should be used in conjunction with existing fundamental / technical models

Support Decision Making
Cannot replace wisdom of an experienced analyst, but can assist in making faster decisions with high conviction



Thank you

Q&A