Using Machine Learning Algorithms to Detect Earnings Manipulation

Context

1. Earnings management occurs when managers use their judgment in financial reporting and in structuring transactions to alter financial reports to mislead stakeholders about the underlying economic performance of the company

2. As on December 2016, the total number of listed companies in Indian stock exchange was approximately 5622.

3. Securities and Exchange Board of India (SEBI) reported that approximately 3.14% of the Indian companies are involved in earnings manipulation.

Literature Review

Beneish Model	Data Used	Outcome
The earliest work carried out on predicting the earnings manipulation was done by Beneish (1997, 1999). Devised an earnings manipulation model based on Probit regression.	The model used eight financial ratios to build M – Score (manipulation score) to identify companies who are likely to have manipulated financial books.	Model looked into the data reported by U.S companies. The data used was not imbalanced and thus probit regression model was able to give a better accuracy in classification.

Financial Ratios

Days Sales to Receivables Index (DSRI)

$$DSRI = \frac{\frac{Receivable_{(t)}}{Sales_{(t)}}}{\frac{Receivable_{(t-1)}}{Sales_{(t-1)}}}$$

DSRI greater than 1 implies revenue inflation

Gross Margin Index (GMI)

$$GMI = \frac{\frac{Sales_{(t-1)} - Cost\ of\ Goods\ Sold_{(t-1)}}{Sales_{(t-1)}}}{\frac{Sales_{(t)} - Cost\ of\ Goods\ Sold_{(t)}}{Sales_{(t)}}}$$

GMI greater than 1 means gross margin is deteriorating

Asset Quality Index (AQI)

$$AQI = \frac{\frac{1 - (Current\ Assest_{(t)} + netPPE_{(t)})}{Total\ Assests_{(t)}}}{\frac{1 - (Current\ Assest_{(t-1)} + netPPE_{(t-1)})}{Total\ Assests_{(t-1)}}}$$

AQI greater than 1 may indicate the tendencies of capitalizing and deferring costs that should have been expensed

Sales Growth Index (SGI)

$$SGI = \frac{Sales_{(t)}}{Sales_{(t-1)}}$$

SGI greater than or less than 1 may indicate that the firm is under possible pressure to manipulate earnings to keep up appearances

Financial Ratios

Depreciation Index (DEPI)

$$DEPI = \frac{\dfrac{Depreciation\ Expense_{(t-1)}}{(Depreciation\ Expense_{(t-1)} + netPPE_{(t-1)})}}{\dfrac{Depreciation\ Expense_{(t)}}{(Depreciation\ Expense_{(t)} + netPPE_{(t)})}}$$

DEPI greater than 1 may indicate tendencies of the assets being depreciated at a slower rate to boost earnings

Sales and General Administrative (SGAI)

$$SGAI = \frac{\dfrac{SGAIExpense_{(t)}}{Sales_{(t)}}}{\dfrac{SGAIExpense_{(t-1)}}{Sales_{(t-1)}}}$$

SGAI less than 1 may indicate that the company may manipulate earnings to defer costs

Accruals to Asset Ratio (ACCR)

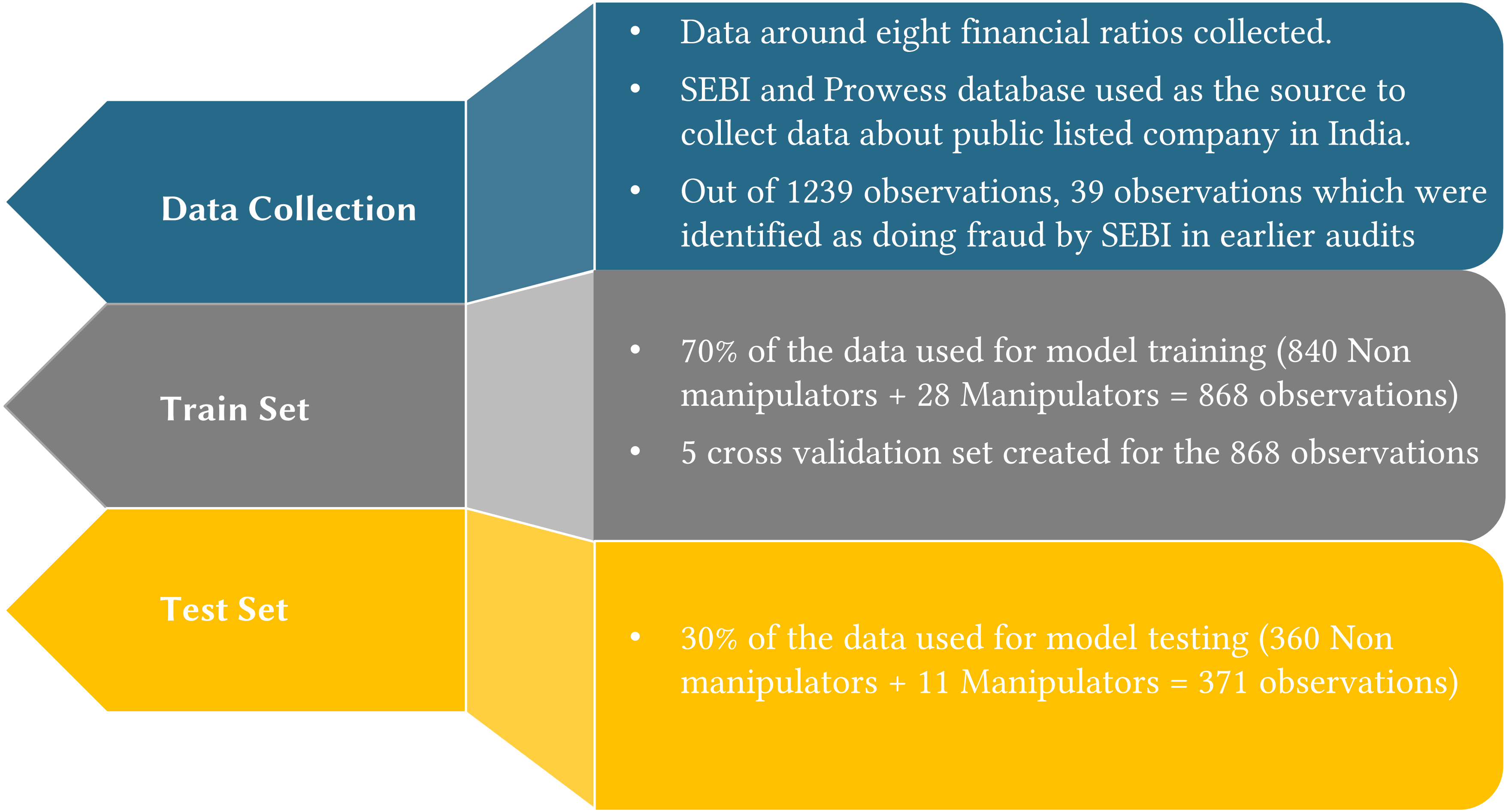
$$ACCR = \frac{Profit\ after\ Tax_{(t)} - Cash\ from\ Operations_{(t)}}{Total\ Assests_{(t)}}$$

ACCR greater than 1 may indicate that the accruals can possibly be used to manipulate earnings

Leverage Index (LEVI)

$$LEVI = \frac{\dfrac{(LTD_{(t)} + CurrentLiabilities_{(t)})}{Total\ Assests_{(t)}}}{\dfrac{(LTD_{(t-1)} + CurrentLiabilities_{(t-1)})}{Total\ Assests_{(t-1)}}}$$

Data Collection



Data Collection

- Data around eight financial ratios collected.
- SEBI and Prowess database used as the source to collect data about public listed company in India.
- Out of 1239 observations, 39 observations which were identified as doing fraud by SEBI in earlier audits

Train Set

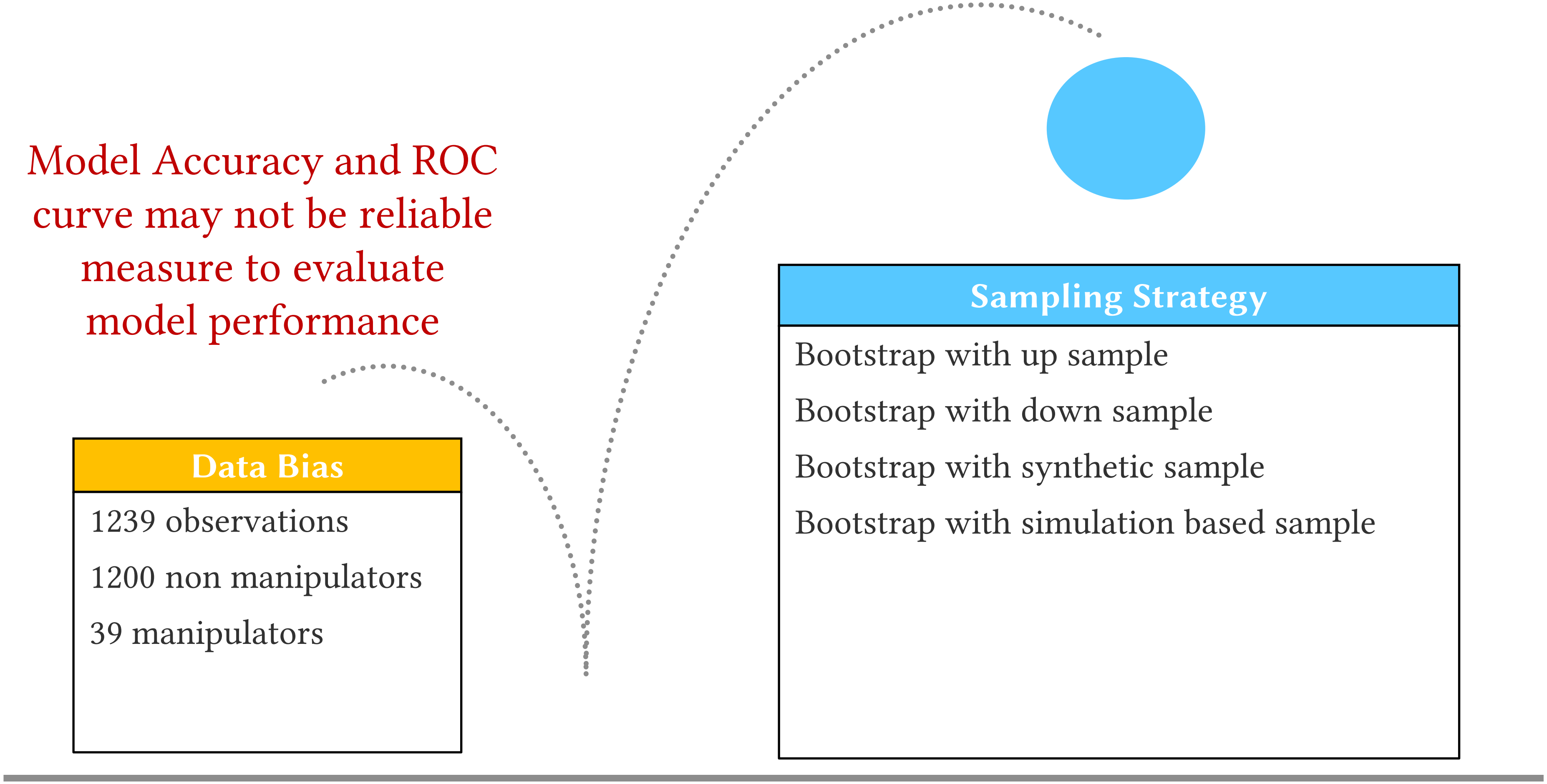
- 70% of the data used for model training (840 Non manipulators + 28 Manipulators = 868 observations)
- 5 cross validation set created for the 868 observations

Test Set

- 30% of the data used for model testing (360 Non manipulators + 11 Manipulators = 371 observations)

Applying Supervised Learning Algorithm to detect fraud

Data challenges and remedy



Model Accuracy and ROC curve may not be reliable measure to evaluate model performance

Data Bias

1239 observations

1200 non manipulators

39 manipulators

Sampling Strategy

Bootstrap with up sample

Bootstrap with down sample

Bootstrap with synthetic sample

Bootstrap with simulation based sample

Outcome of traditional models

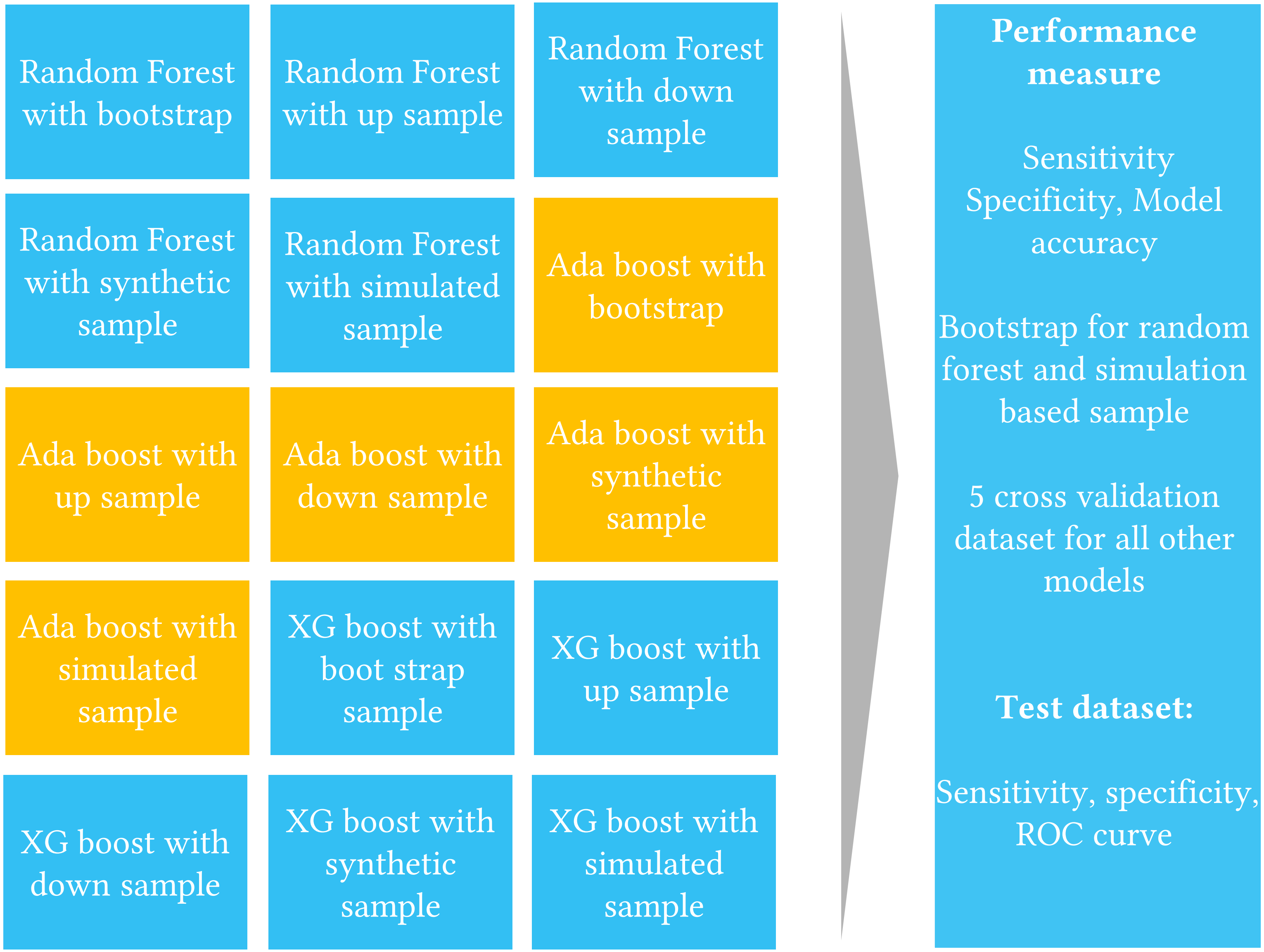
Low specificity reported by logistic regression and Neural network

Model Performance on Company Financial Ratios

Model **(70% Train, 30% Test)**	Performance on 5 CV Data (28 – Yes, 840 – No)			Performance on Test Set (11- Yes, 360 – No)			ROC on Test Set
	Sen	Spec	Overall	Sen	Spec	Overall	
Logistic Regression	0.99	0.03	0.97	1.00	0.18	0.97	0.94
Neural Network	0.99	0.19	0.97	0.99	0.36	0.97	0.88
Sensitivity = Non-Manipulator (No), Specificity = Manipulator(Yes)							

Overall accuracy and ROC curve suggests a good performance of the model

Ensemble methods with sampling–Models applied



Random Forest with bootstrap	Random Forest with up sample	Random Forest with down sample
Random Forest with synthetic sample	Random Forest with simulated sample	Ada boost with bootstrap
Ada boost with up sample	Ada boost with down sample	Ada boost with synthetic sample
Ada boost with simulated sample	XG boost with boot strap sample	XG boost with up sample
XG boost with down sample	XG boost with synthetic sample	XG boost with simulated sample

Performance measure

Sensitivity Specificity, Model accuracy

Bootstrap for random forest and simulation based sample

5 cross validation dataset for all other models

Test dataset:

Sensitivity, specificity, ROC curve

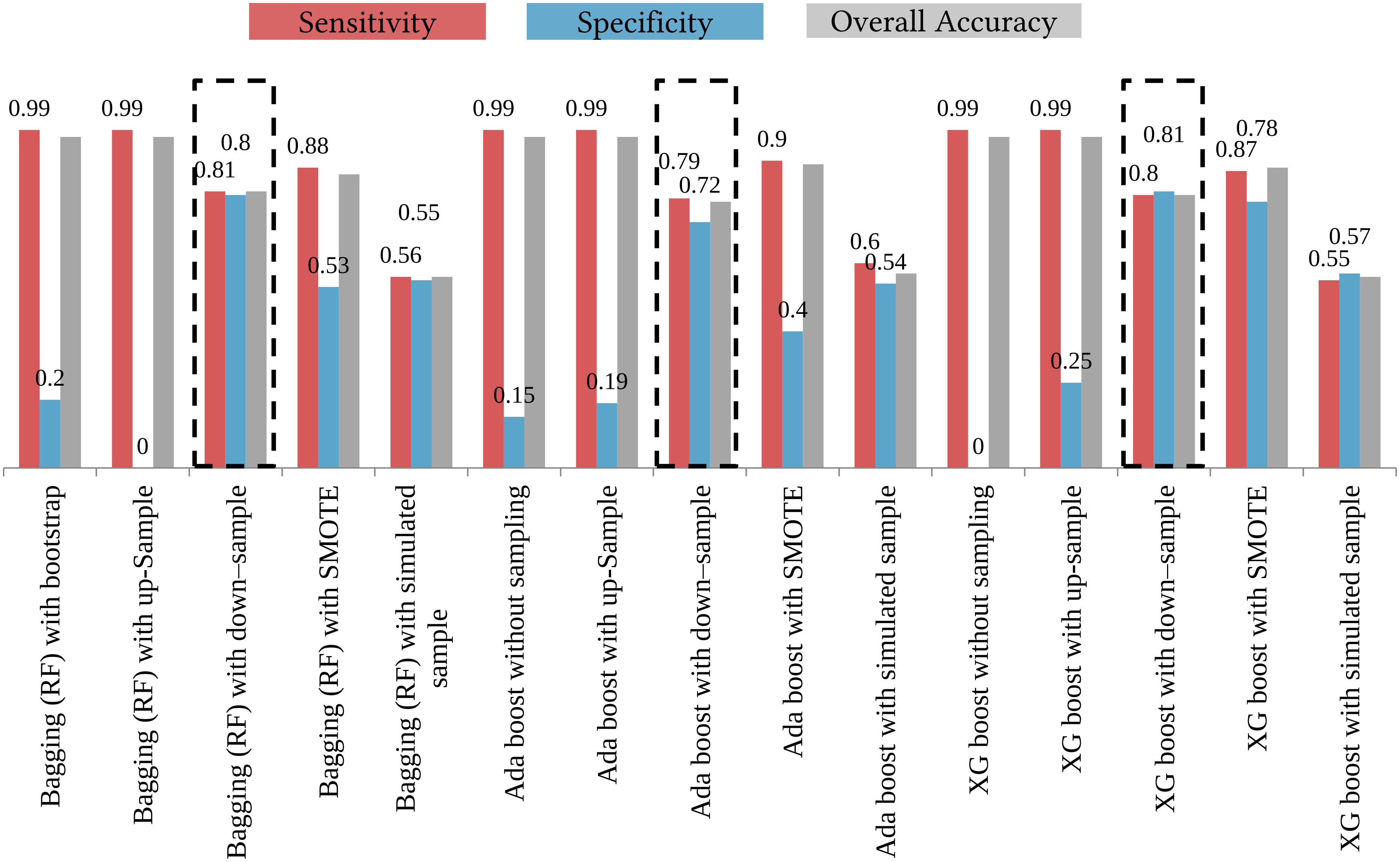
Performance measures from ensemble methods

Model Performance on Company Financial Ratios

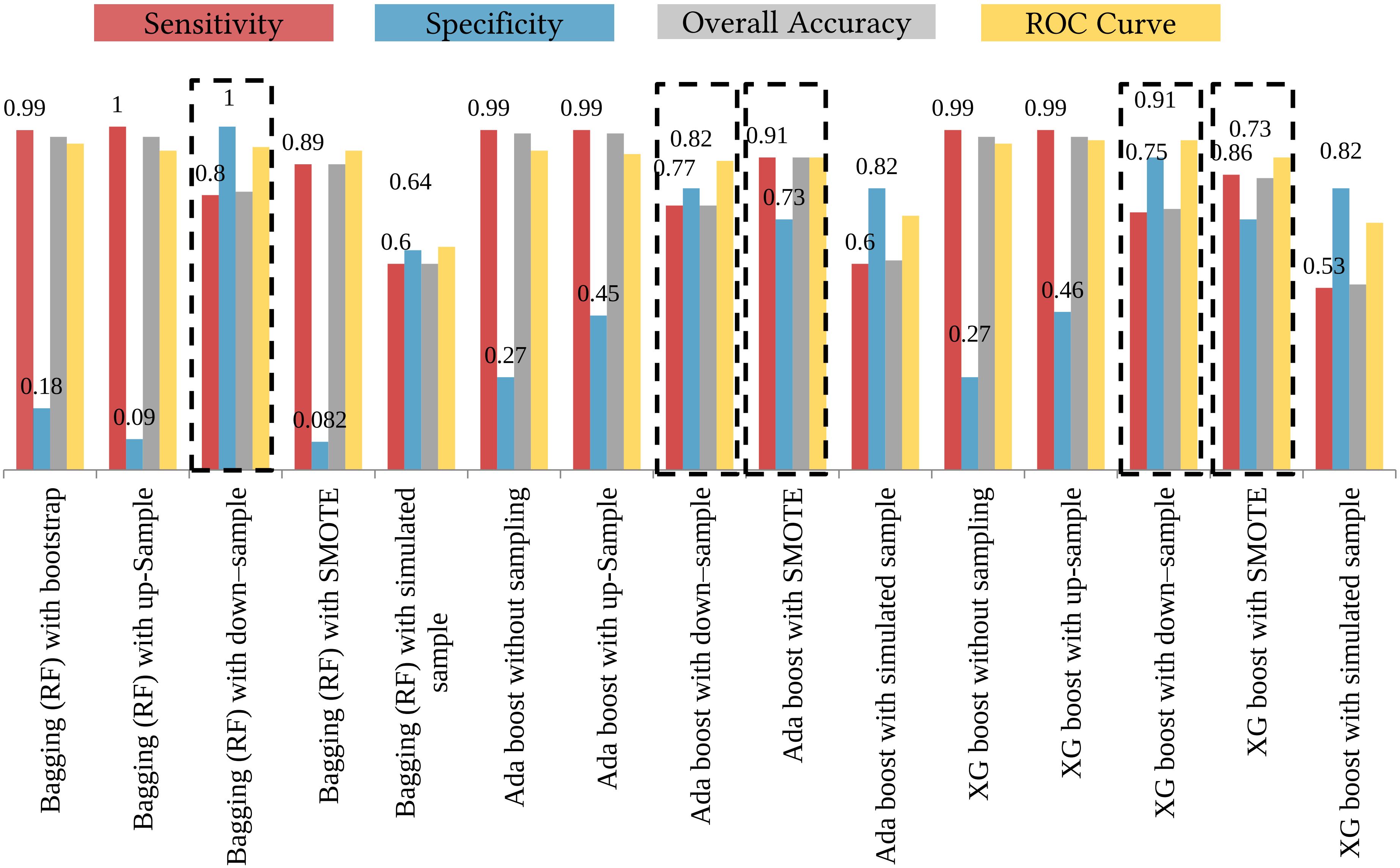
Model (70% Train, 30% Test)	Performance on 5 CV Data (28 – Yes, 840 – No)			Performance on Test Set (11- Yes, 360 – No)			ROC on Test Set
	Sen	Spec	Overall	Sen	Spec	Overall	
Bagging (RF) with bootstrap	0.99	0.20	0.97	0.99	0.18	0.97	0.95
Bagging (RF) with up-Sample	0.99	0.00	0.97	1.0	0.09	0.97	0.93
Bagging (RF) with down–sample	0.81	0.80	0.81	0.80	1.0	0.81	0.94
Bagging (RF) with SMOTE	0.88	0.53	0.86	0.89	.082	0.89	0.93
Bagging (RF) with simulated sample	0.56	0.55	0.56	0.60	0.64	0.60	0.65
Ada boost without sampling	0.99	0.15	0.97	0.99	0.27	0.98	0.93
Ada boost with up-Sample	0.99	0.19	0.97	0.99	0.45	0.98	0.92
Ada boost with down–sample	0.79	0.72	0.78	0.77	0.82	0.77	0.90
Ada boost with SMOTE	0.90	0.40	0.89	0.91	0.73	0.91	0.91
Ada boost with simulated sample	0.60	0.54	0.57	0.60	0.82	0.61	0.74
XG boost without sampling	0.99	0.00	0.97	0.99	0.27	0.97	0.95
XG boost with up-sample	0.99	0.25	0.97	0.99	0.46	0.97	0.96
XG boost with down–sample	0.80	0.81	0.80	0.75	0.91	0.76	0.96
XG boost with SMOTE	0.87	0.78	0.88	0.86	0.73	0.85	0.91
XG boost with simulated sample	0.55	0.57	0.56	0.53	0.82	0.54	0.72

Sensitivity = Non-Manipulator (No), Specificity = Manipulator(Yes)

Performance on 5 CV set–Graphical View



Performance on Test set–Graphical View



Thank you