Predicting the likelihood of company bankruptcy using machine learning

Qiwei Lu

Aug 2020

Too big to fail? Not any more!

SEARS

Bankruptcy impacts everyone

- **The management**

- **Financial service providers**

  e.g. auditors

- **Investors**

- **Regulators**

Can we predict it?

- Bankruptcy prediction has been a subject of formal analysis since at least 1932, when FitzPatrick published a study of 20 pairs of firms, one failed and one surviving, matched by date, size and industry, in *The Certified Public Accountant*. He did not perform statistical analysis as is now common, but he thoughtfully interpreted the **ratios and trends in the ratios**.

--Wikipedia

About the research dataset

- 10,000 Polish companies from UCI Machine Learning Repository
- The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.
- 64 financial attributes are given for each company(features) and whether it is bankrupted(labels)

Financial ratios explained

1. **liquidity** ratios, which demonstrate a company's ability to pay its short-term debts and other liabilities.

2. **Activity** ratios, which demonstrate a company's efficiency in operations.

3.**Leverage** ratios, which demonstrate a company's ability to pay its long-term debt.

4. **Performance/profitability** ratios, which tell investors about a company's profit.

5. Valuation ratios, which rely on a company's current share price, provide a picture of whether or not the stock makes a compelling investment at current levels.

Features collinearity

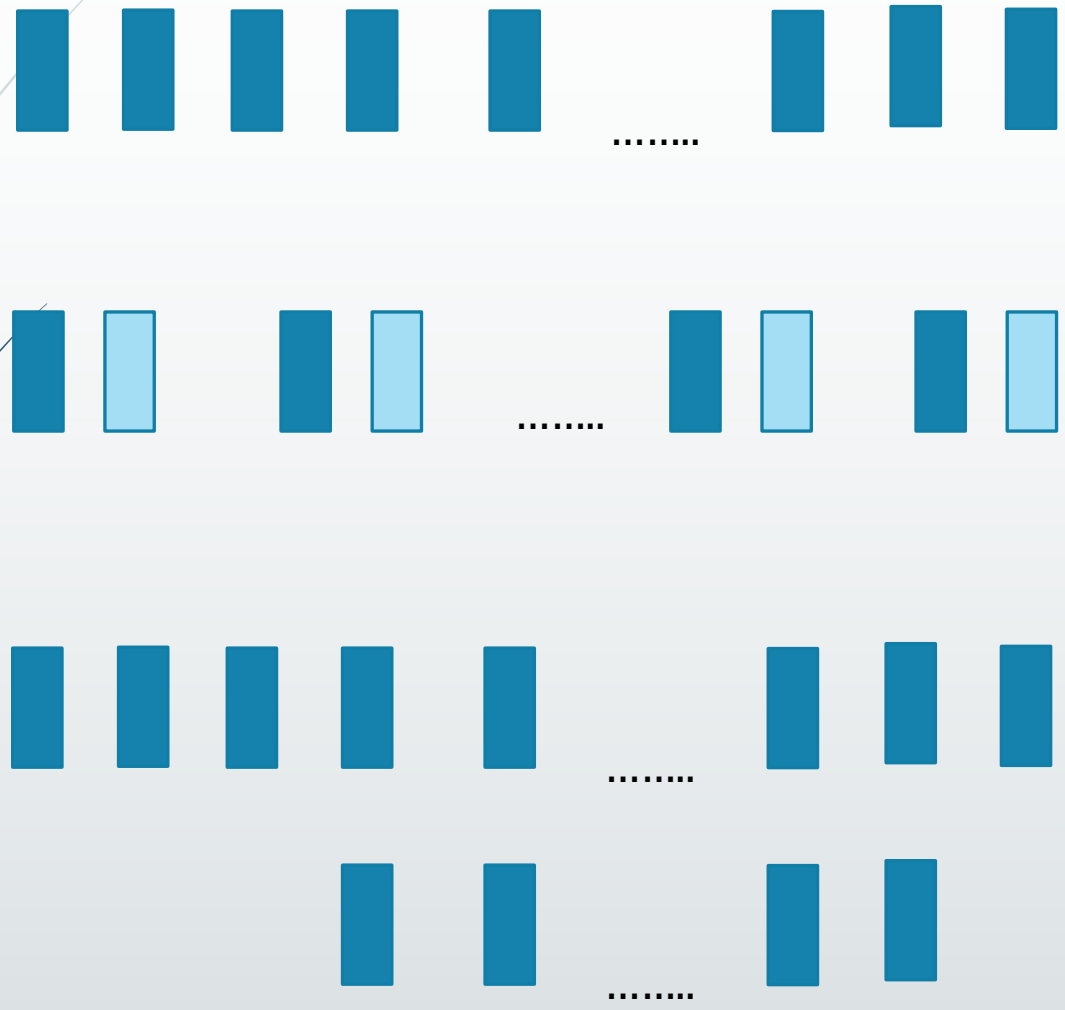
Calculate Pearson correlation coefficients of each feature pair

Calculate the correlations between the target and features

For each highly correlated feature pair, drop the feature with weaker relationship with the target

Drop the weak features among all remaining features

Features collinearity



Original 64 features

Find the pairs that has Pearson correlation coefficient over 0.95

Within each pair, drop the feature that has less strong relationship with the target

47 features, set the ANOVA F-value threshold between feature and target to be 0.1 and find the ones below it

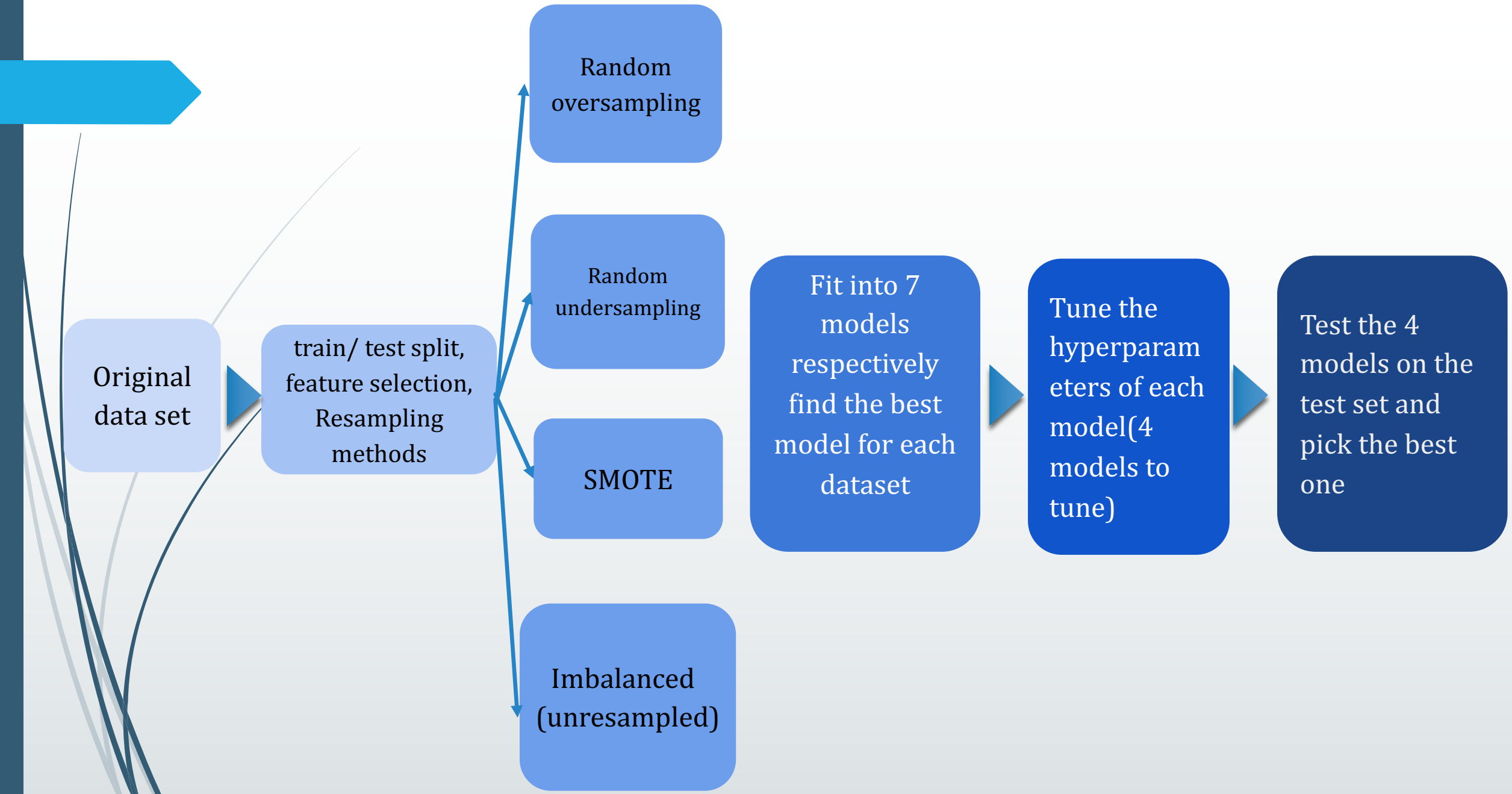
Drop the weak ones, finally got 35 features

Data imbalance

- Use stratified split for the train test split and cross validation
- Use random oversampling, random undersampling and Synthetic Minority Oversampling Technique (SMOTE) methods to resample the data
- Use multiple evaluation metrics to assess models(accuracy, recall, precision, roc auc scores)
- Tune the threshold for classification

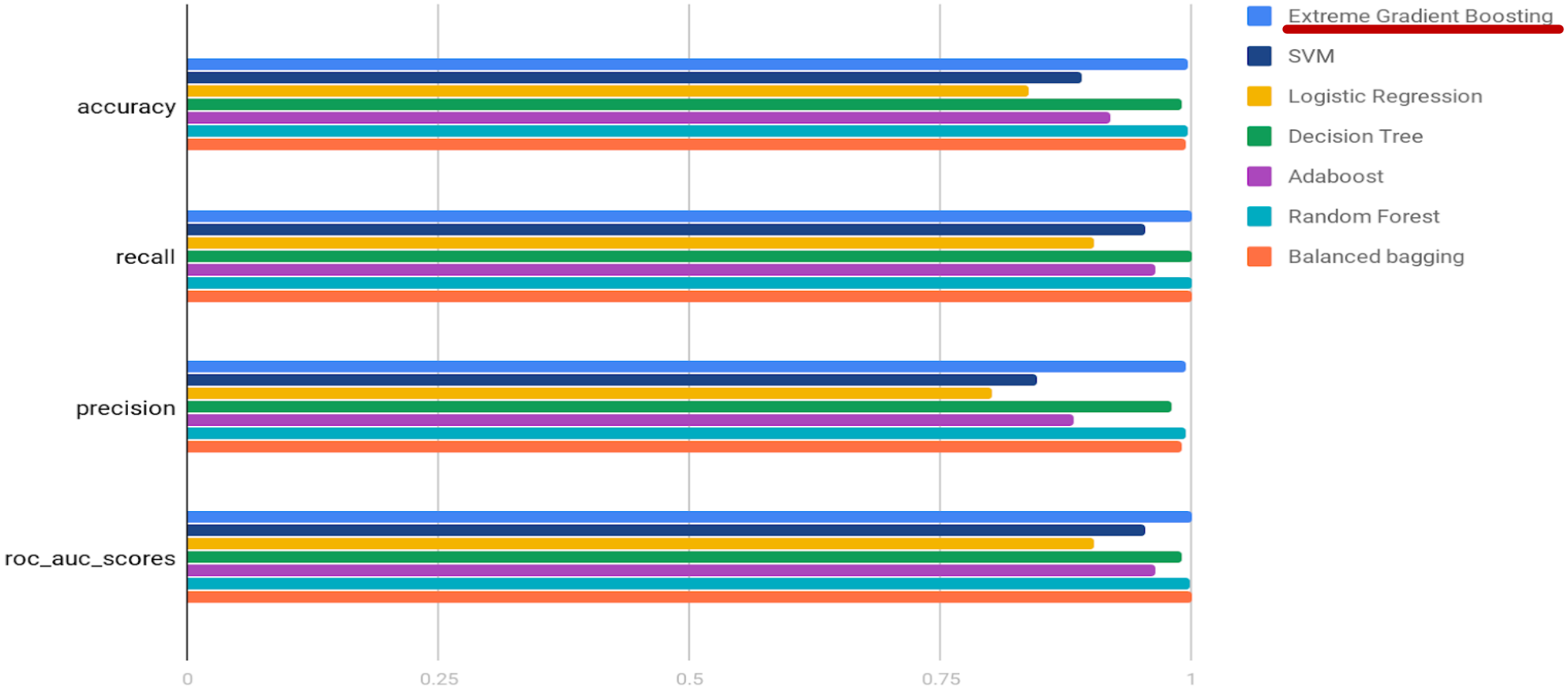
Model selection and training

- Extreme Gradient Boosting classifier (Xgboost)
- Support Vector Machine(SVM)classifier
- Logistic Regression classifier
- Decision Tree classifier
- Adaboost classifier
- Random Forest classifier
- Balanced Bagging classifier


Original data set
train/ test split, feature selection, Resampling methods
Random oversampling
Random undersampling
SMOTE
Imbalanced (unresampled)
Fit into 7 models respectively find the best model for each dataset
Tune the hyperparam eters of each model(4 models to tune)
Test the 4 models on the test set and pick the best one

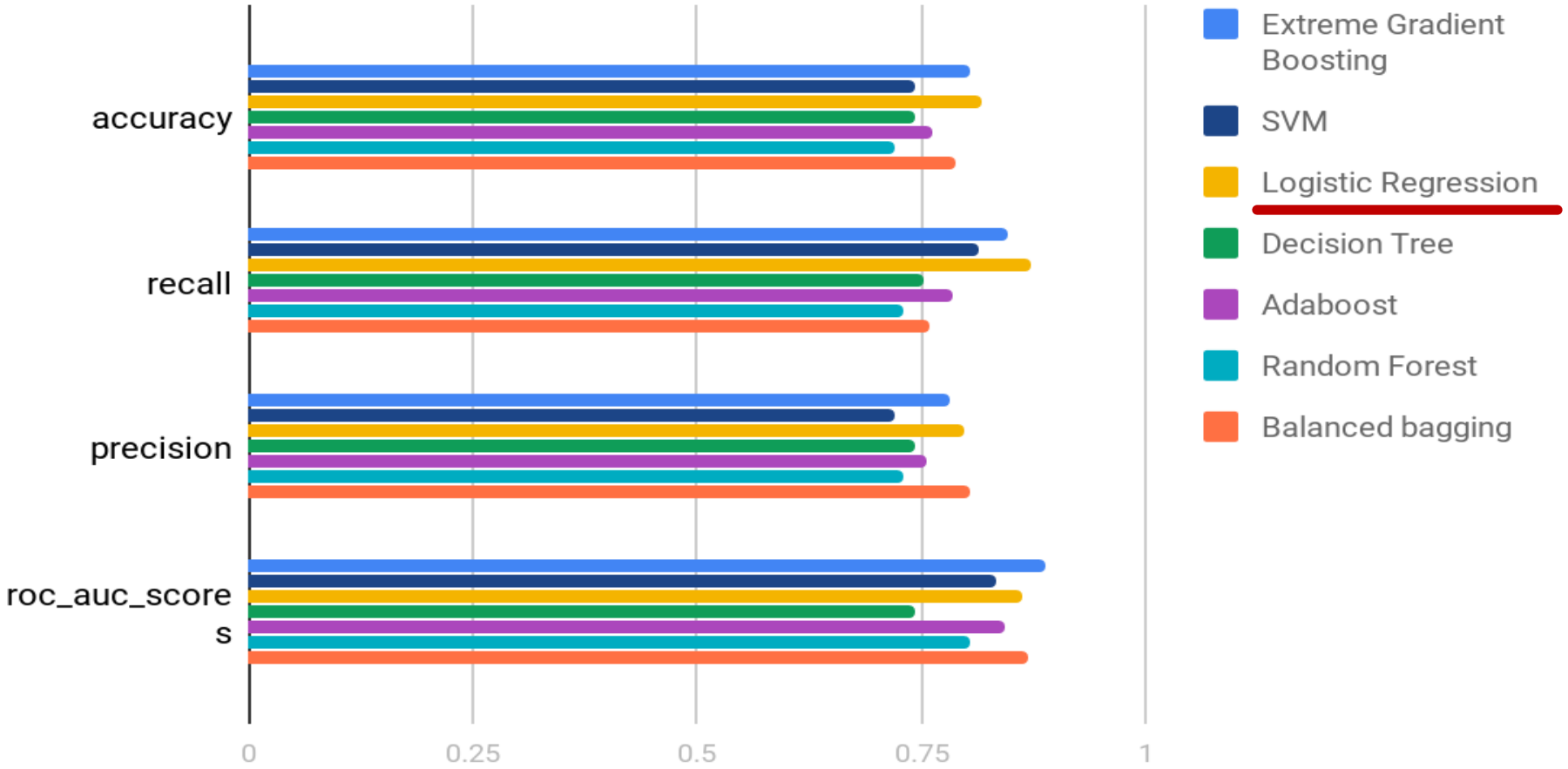
Modeling results and analysis – Oversampling

Figure 6. Oversampling metrics



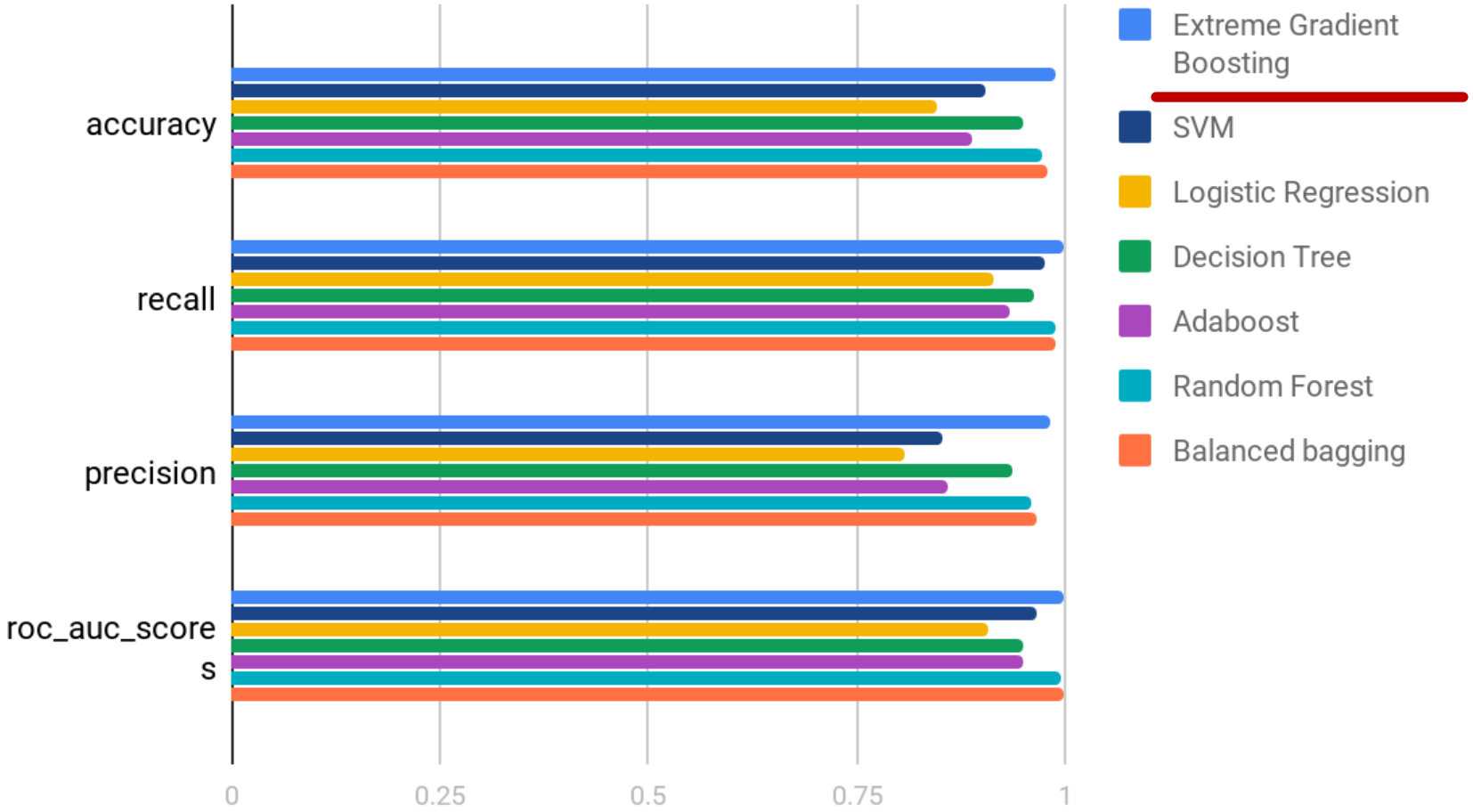
Modeling results and analysis – undersampling

Figure 7, undersampling metrics



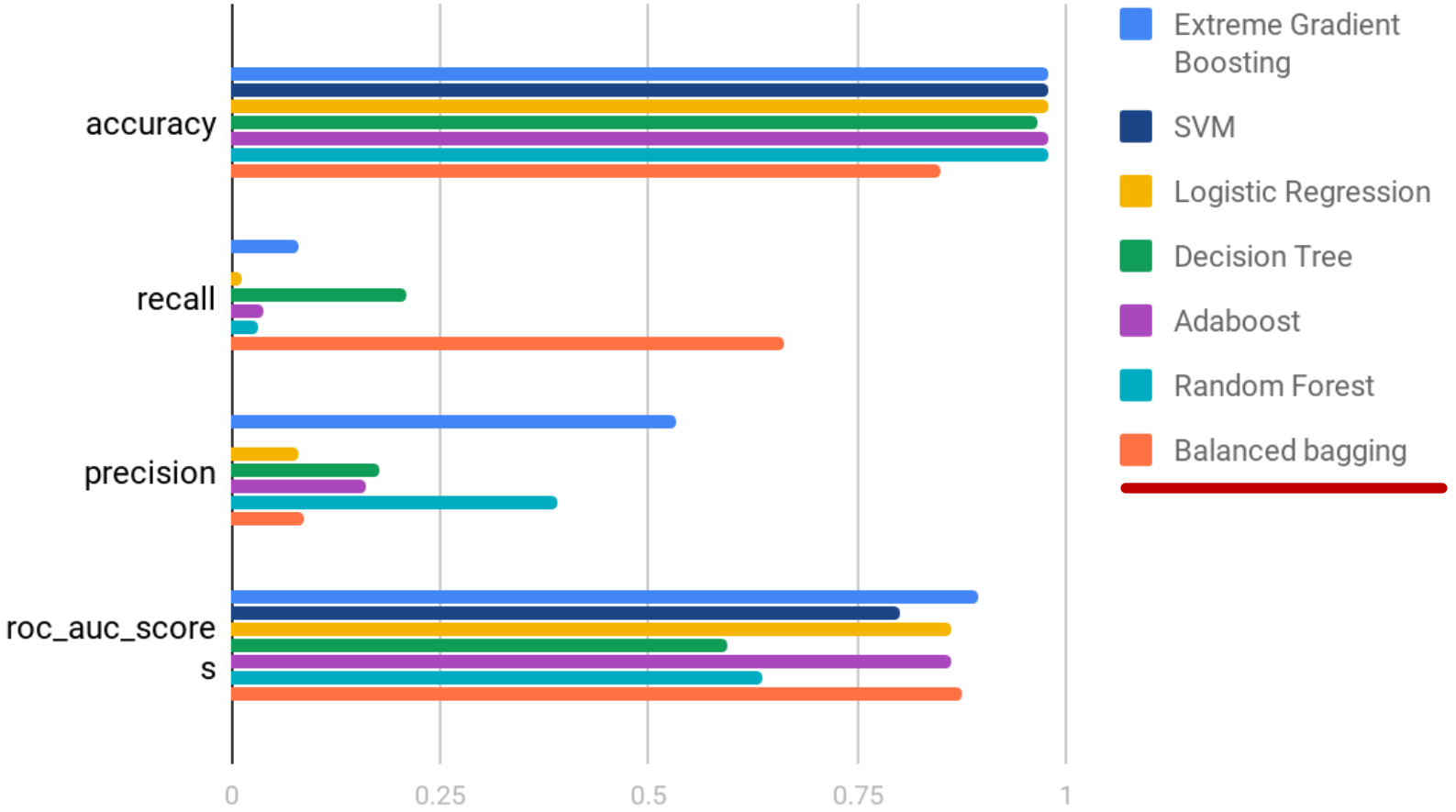
Modeling results and analysis – SMOTE

Figure 8. SMOTE metrics



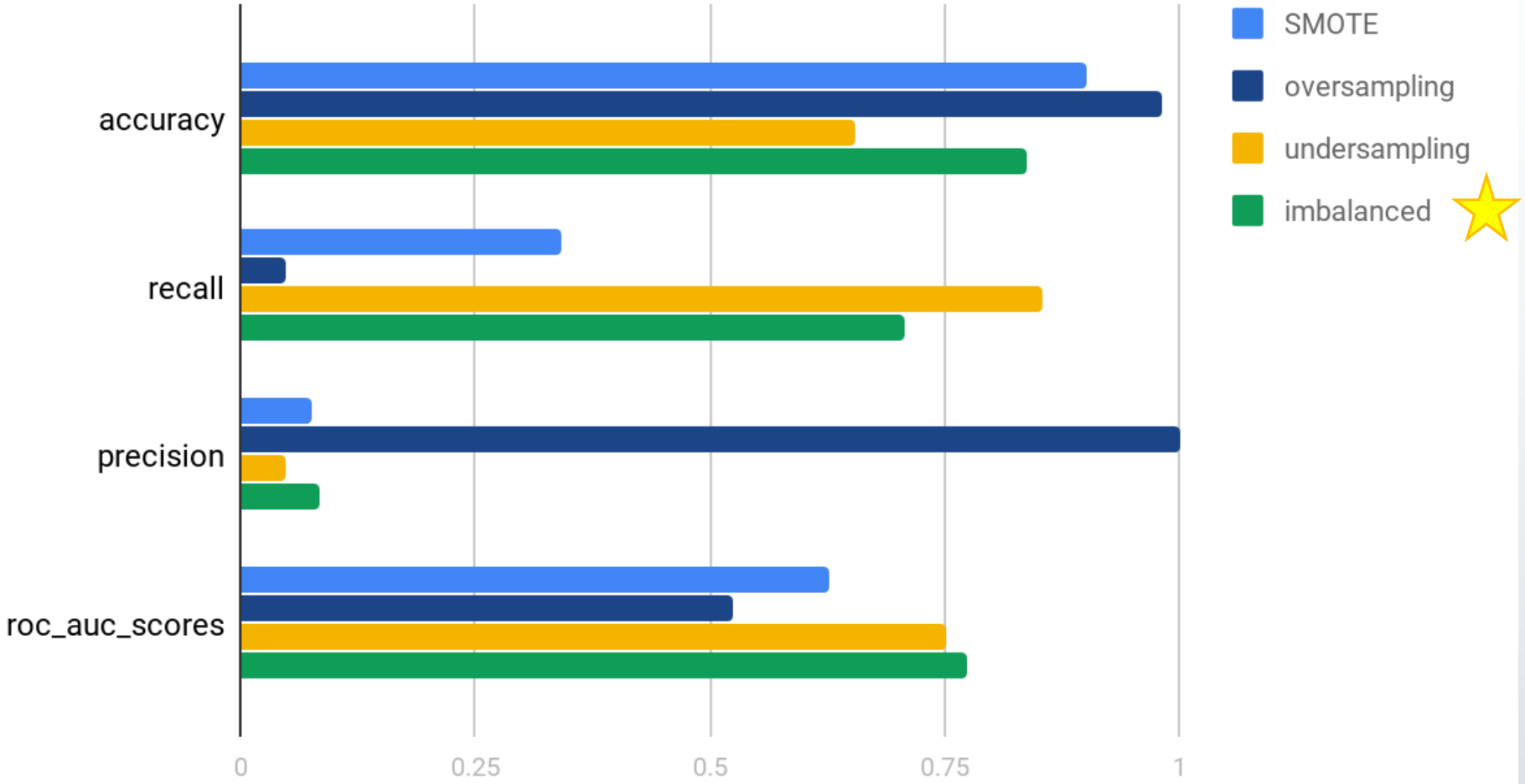
Modeling results and analysis – imbalanced

Figure 9, imbalance metrics

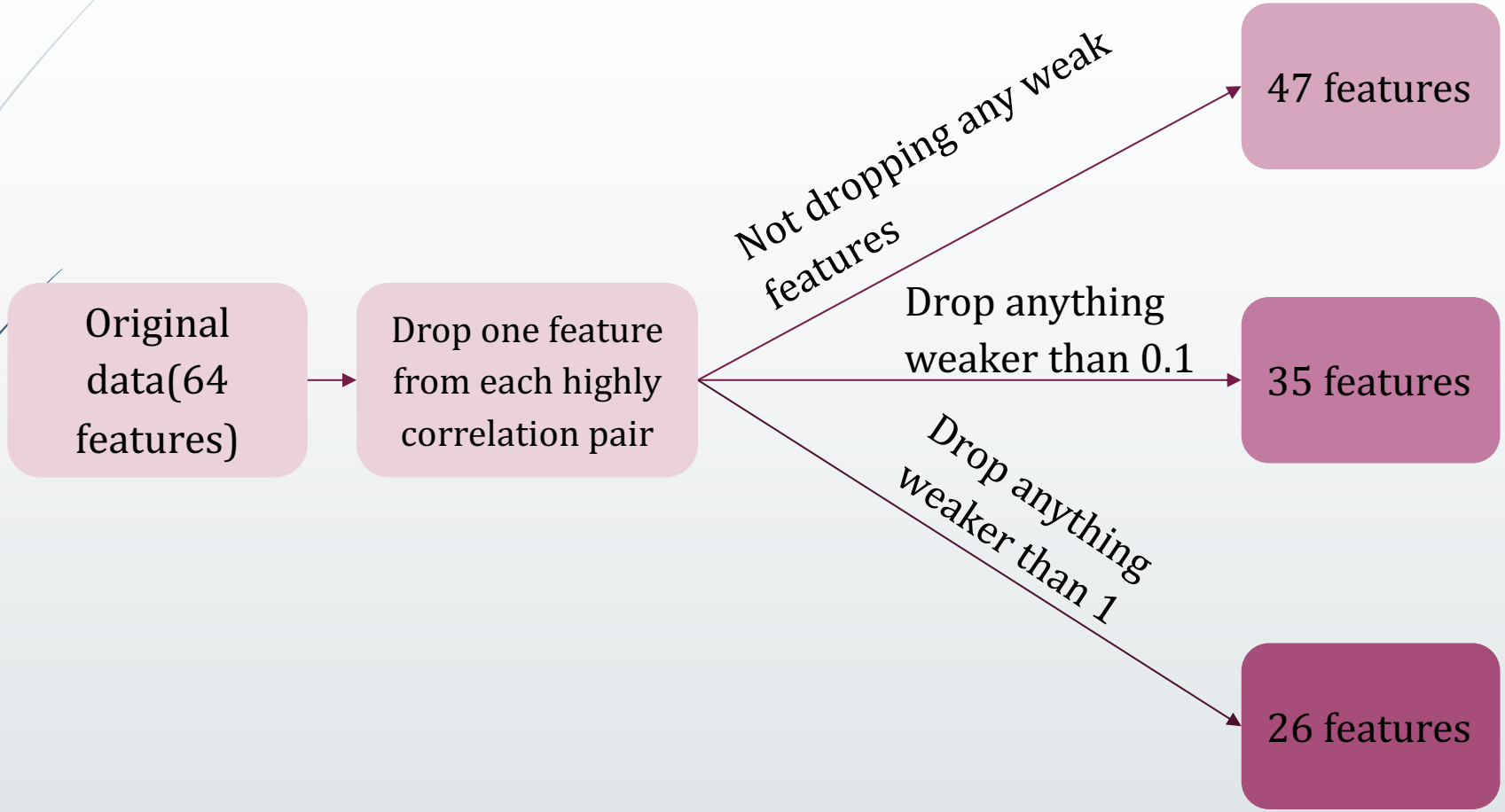


Final pick made by test set

Figure 10, final test set metrics

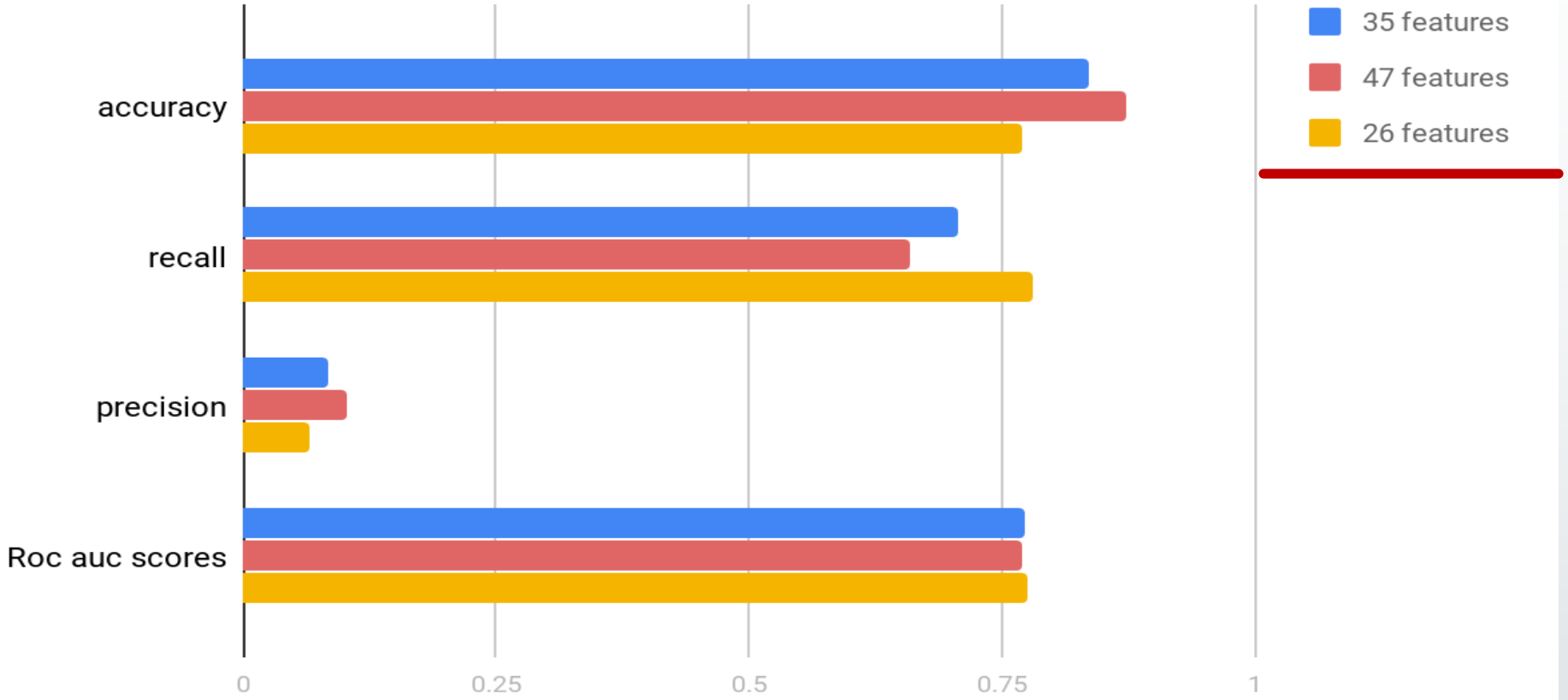


Rethinking on the feature selection



Rethinking on the feature selection

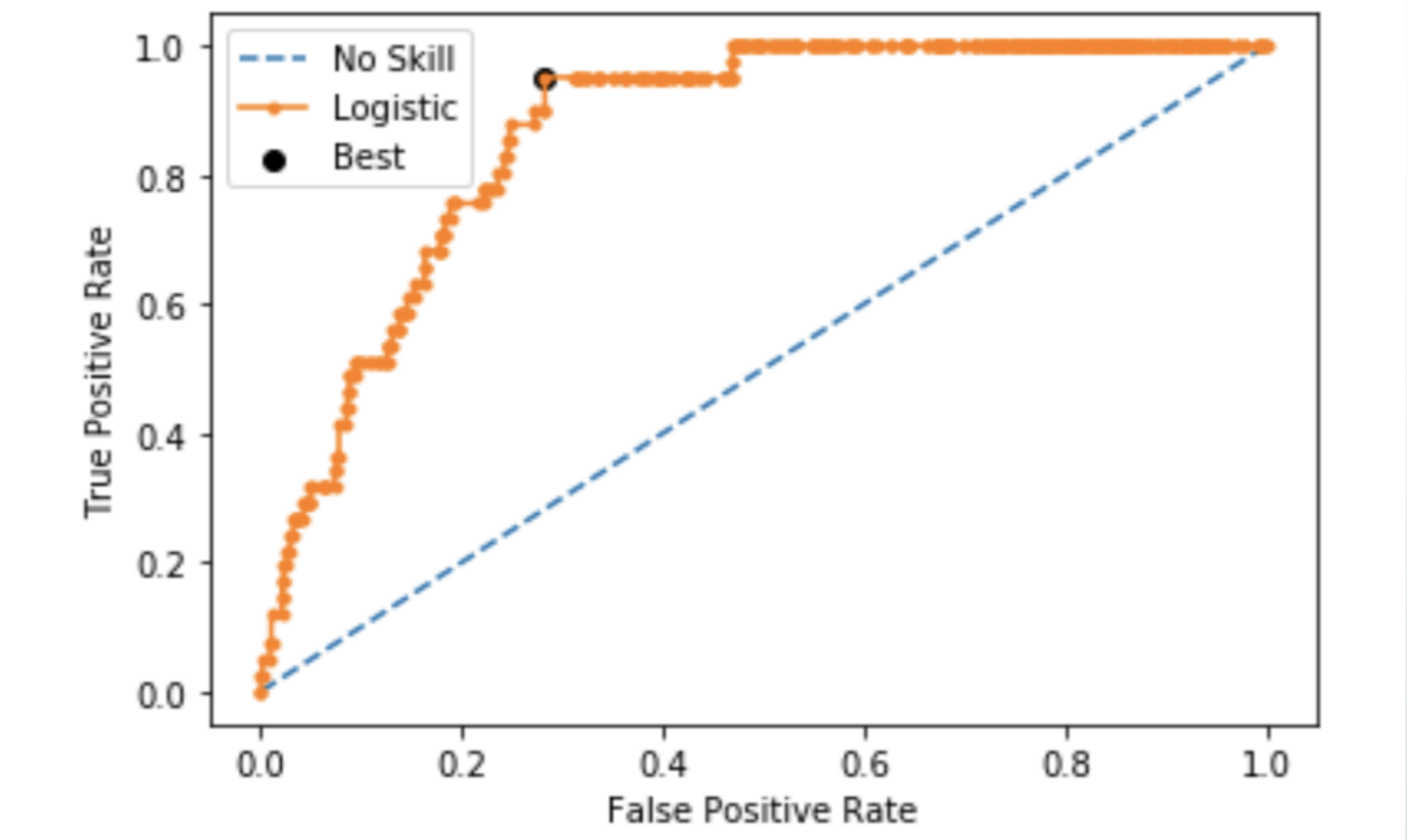
Figure 12, score comparison between different feature sets



Tuning classification threshold

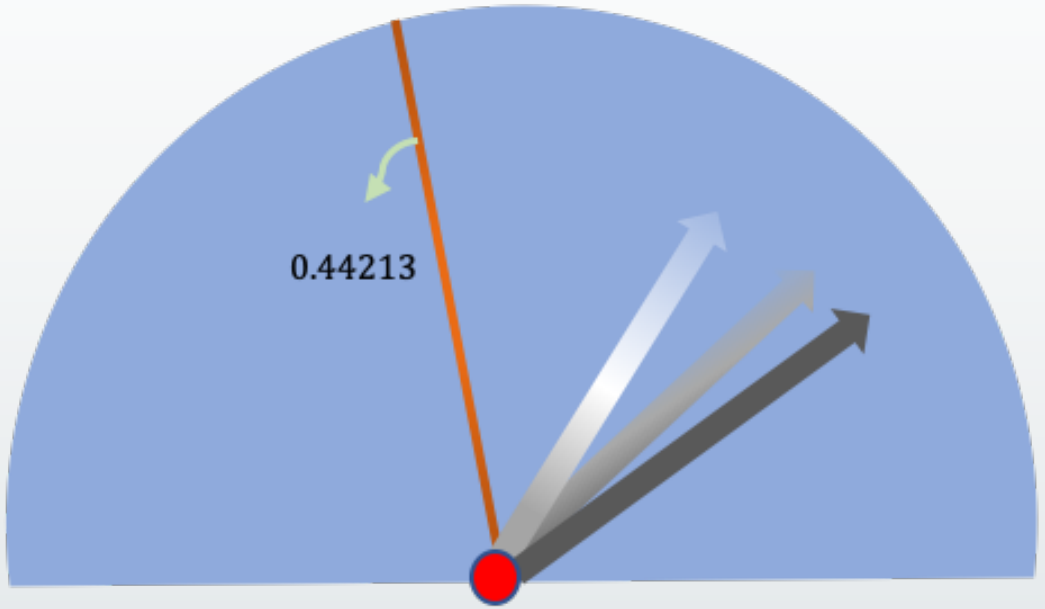
- Sensitivity = TruePositive / (TruePositive + FalseNegative)
- Specificity = TrueNegative / (FalsePositive + TrueNegative)
- Where:
- Sensitivity = True Positive Rate
- Specificity = 1 – False Positive Rate
- G-Mean = sqrt(Sensitivity * Specificity)

Tuning classification threshold



Application of the test result

The range of the possibility of client classified as bankruptcy using the ML model

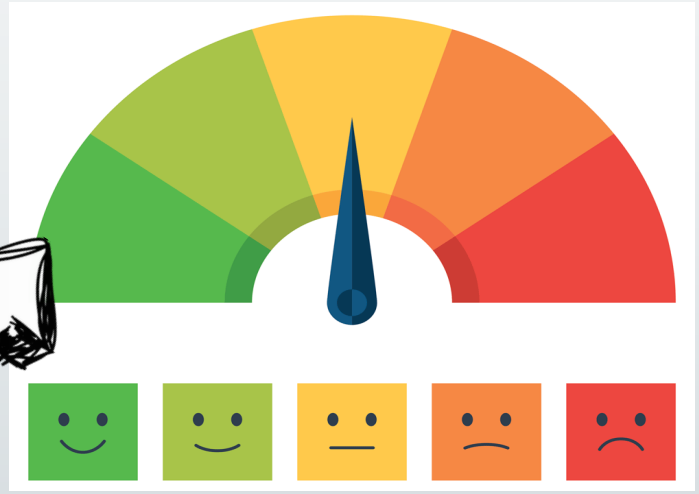


Auditors have to find the threshold of substantial doubt in going concern

Application of the test result

- The management

- The auditors



Recap

- We used domain knowledge to understand the given dataset
- We used machine learning tools to select features, deal with data imbalance, train and test 7 models and found the best one with good performance: balanced bagging classifier model
- We then adjusted the feature selection threshold and tuned classification threshold and boosted performance even more.
- We have also explored the possible application of the model to the real business world where the positive possibilities will be a better reference than the crisp predicted labels for management and auditors to judge for threshold of the substantial doubt for going concern.

Thanks!

- Github link to report and codes: https://github.com/lqw110/capstoneproject1
- Linkedlin profile: https://www.linkedin.com/in/qiweilu

- Let's connect and let me know your thoughts!