Performance of ML in Bankruptcy Classification

Nicole Blyuss
nika24@my.yorku.ca

Aleksandr Tsybakin
al.tsybakin@gmail.com

Vanessa Bridge
vane.m.bridge@gmail.com

Omesh Chandan
omeshchandan@gmail.com

YORK UNIVERSITÉ UNIVERSITY

Introduction

Credit lenders benefit from understanding their client's economic health and with the predominance of *data collection*, these institutions can benefit from applying ML algorithms to current and prospective clients to determine their ability to pay back their loans[1]. By being able to determine whether a current or prospective client may default on credit, financial institutions can greatly increase their revenues.

Multiple methods are available to assist in the *classification* of clients and their likelihood of defaulting.

This study investigates the performance of several of these methods on a common data set.

Objective

- Use different machine learning models to accurately predict the likelihood of borrowers defaulting on their loans. Our focus is on classification algorithms:
- **SVM**, **Logistic Regression**, **K-nearest Neighbors**, **Decision Trees**, **Random Forest** and **Boosting Models**
- Model performance is compared under the **F1 Score measure.**

Data

- Data on Polish companies from 2000-2012 [2].
- The data set consists of over 60 numerical attributes extracted from the financial statements of Polish companies.
- More than 50% of observations have at least one missing value.
- The dataset is highly unbalanced (bankruptcy and non-bankruptcy) with approximately 4% and 96% , respectively.

Data Processing

- **Correlation analysis:** attributes with a collinearity > 0.9 were removed.
- **Data splitting**: 80/20 ratio is used for training and testing, respectively.
- **Missing values**: attributes with > 20% of missing data are removed.
- **Data imputation:** used MissForest to predict missing data.
- **Scaling**: Standard Scaler is applied to make the variables in the same range.
- **Oversampling & Undersampling**: rebalanced with SMOTE, Edited-Nearest Neighbours and K-Means.

Models

The best parameters for models are chosen using **Grid Search** with 5-Fold **Cross validation.**

Model	Definition	Pros	Cons
Logistic Regression	A model used when then response is binary. Produces the probability of success which is used as a classification threshold	Simple and Interpretable	Assumes linearity of the data
Support Vector Machine	A supervised ML algorithm that uses a hyperplane (line) for classification	Versatile	computing constrained
Random Forest	A model which uses a collection of decision trees for classification	Robust to Outliers & Non Linear Data	Prone to overfitting
Boosting Models	A set of algorithms sequentially built by increasing influence in high performing models while simultaneously minimizing errors in models that performed poorly	Adequate for unbalance method	Sensitive to outlier

Table 1:Description of Models.

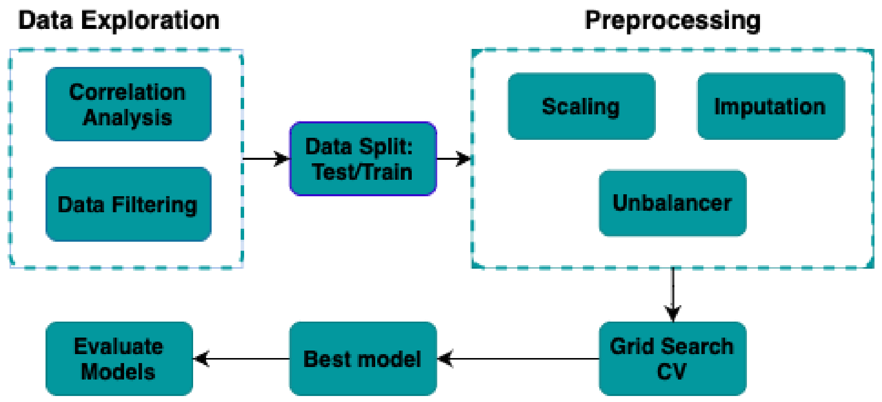

Figure 1: Pipeline.

Results

Outcome: Each of the employed model increased its performance and overall scores across all of the adequacy statistics, after the data unbalancing was handled. For unbalanced data, the more reliable metric for model comparison iis the F1 Score.

Measures	LR	SVM	RF	AdaB	GB	XGBoost
F1-score	0.6619	0.6892	0.6374	0.7544	0.6667	0.7438
Accuracy	0.5648	0.5741	0.6944	0.7407	0.6944	0.713
Precision	0.5412	0.5426	0.7838	0.7167	0.7333	0.6716
Recall	0.8519	0.9444	0.537	0.7963	0.6111	0.8333

Table 2: Model Comparison acoss four metrics on Balanced Test Data .

Measures	LR	SVM	RF	AdaB	GB	XGBoost
F1-score	0.09	0.087	0.2044	0.1834	0.1579	0.1614
Accuracy	0.3305	0.2297	0.8431	0.7243	0.7696	0.8279
Precision	0.0475	0.0456	0.1273	0.1036	0.092	0.0996
Recall	0.8519	0.9444	0.5185	0.7963	0.5556	0.4259

Table 3: Model Comparison acoss four metrics on Unbalanced Test Data

Conclusion

- By leveraging econometric features banks and financial institutions are able to better predict which corporate or commercial clients are likely to go bankrupt thus increasing their probability of default. Based on the results obtained from testing on balanced date, the models better suited for such analysis are Adaptive Boosting and XGBoost models.
- Inherent in every model are assumptions. Under the assumption that the data structures of the past are a likely indicator to future events, then there is a clear application of the Adaptive Boosting and XGBoost models in industry.

References

[1] D. Bacham and D. J. Zhao, "Machine learning: Challenges, lessons, and opportunities in credit risk modeling." [Online]. Available: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling. [Accessed: 24-Apr-2021].

[2] UCI Machine Learning Repository, Polish companies bankruptcy data, data set, [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data