Predicting Bankruptcy of Polish Companies

Andrew Foresi

November 25, 2018

Problem

- Classification - Determine the best model for classifying bankrupt (1) or healthy (0) companies using their financial metrics from the models below.
- This would have real world value for a variety of stakeholders

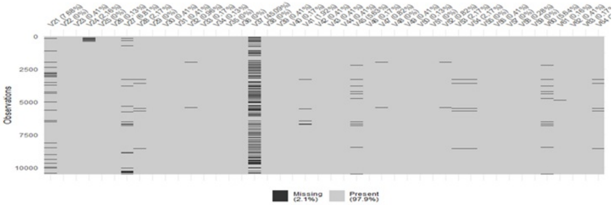
- Bagging
- Random Forests
- Gradient Boosting Machine
- K-Nearest Neighbors
- Neural Network

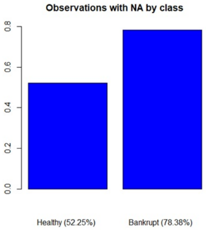
Data

- 10,503 observations (companies) sourced from UCI ML library
- 64 predictors which are various financial metrics from each company's financial statements
- Binary target variable of bankrupt or healthy after 3 years of observation
- Data preprocessing will play an important role in producing accurate results

Missing Data

- 53% of observations included at least one NA
- NA more common in minority class
- Options: Delete or impute missing values





Solution

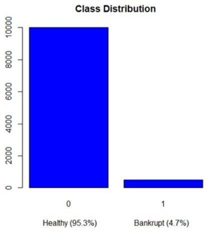
- Removal of V37 (current assets - inventories) / (lt liabilities)
- Data deletion, KNN-imputation and new NAcount variable were evaluated

Class Imbalance

- Common for bankruptcy/fraud/disease datasets to be highly imbalanced
- Only 493 (4.7%) of 10,503 observations were bankrupt
- More extreme cases are common

Solution

- Under sample the majority class while including >95% of minority class in train set
- Another approach would be to use SMOTE, but results were insignificant

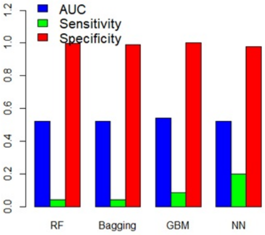


Test Results with Minimal Data Preprocessing

NA Deletion

Model	Mis	AUC	Sensitivity	Specificity
GBM	2.25%	0.5417	0.0833	1.00
Bagging	2.46%	0.5203	0.0417	0.9900
Random Forest	2.45%	0.5203	0.0417	0.9989
NN	2.46%	0.5217	0.2000	0.9794

Low misclassification, low sensitivity as expected



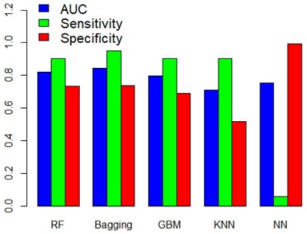
Test Results with Data Preprocessing

NA Deletion, majority class under-sampling

Model	Mis	AUC	Sensitivity	Specificity
GBM	30.53%	0.7974	0.9048	0.6900
Bagging	25.72%	0.8453	0.9523	0.7382
Random Forest	26.02%	0.8204	0.9048	0.7361
KNN	47.44%	0.7110	0.9048	0.5173
NN	34.73%	0.7555	0.0563	0.9941

Increased sensitivity and decreased specificity

Likely a welcomed result as determining bankruptcy may be more important

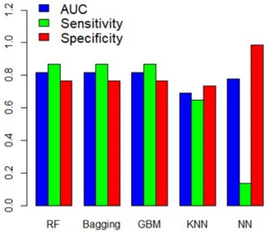


Test Results with Data Preprocessing (Cont.)

KNN Imputation -V37, under-sampling, #NA column

Model	Mis	AUC	Sensitivity	Specificity
GBM	22.81%	0.8179	0.8687	0.7671
Bagging	22.90%	0.8174	0.8687	0.7661
Random Forest	23.09%	0.8164	0.8687	0.7641
KNN	27.00%	0.6903	0.6465	0.7341
NN	25.38%	0.7556	0.1347	0.9874

- KNN Imputation, removed V37, new variable which counts NAs per column (addresses unequal NA distribution between classes)
- More of a trade-off between sensitivity and specificity



Discussion/Conclusion

- Sensitivity more important than specificity in this application and therefore bagging produces a much better result when missing values are deleted instead of imputed with a sensitivity of 95.23%
- Therefore 95.23% of bankrupt companies were accurately classified as bankrupt. Misclassifying a bankrupt observation as healthy is much more costly than misclassifying a healthy observation as bankrupt
- Much value came in the data preprocessing stage. Confident that more work and techniques applied to the data could increase sensitivity even further.
    - Many imputation techniques yet to be tested
    - Parameter tuning SMOTE