Birkbeck Institute for Data Analytics (BIDA)

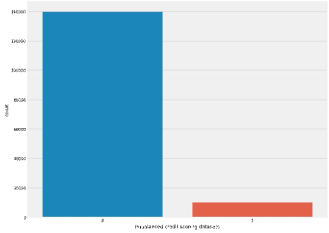
Knowledge Lab

Birkbeck
UNIVERSITY OF LONDON

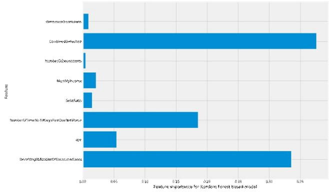
An application of classification techniques for credit scoring

Research Student
Seongil Han

Supervisors
Andrea Cali
Alessandro Provetti



Imbalanced credit scoring dataset



Relative feature importance

Birkbeck
UNIVERSITY OF LONDON

Department of Computer Science and Information Systems

Research Aims

The primary rationale is to understand and improve credit scoring models using machine learning techniques. The performance of techniques is compared with open source financial datasets and the existing models can be improved by novel computer science techniques in realistic problem of the areas, where;

- binary and multi-class credit classification
- balanced and imbalanced data sets
- complete data sets without missing values and incomplete dataset with missing values

Research Methodology

The architectural design for credit scoring is described as Figure 1. Firstly, the open source financial data is pre-processed. Secondly, feature engineering is employed to decide feature subsets that have the most effective and least redundant attributes. Then, the engineered subsets are used to train machine learning models. Finally, evaluation metrics are applied to compare the experimental results among models for credit scoring.

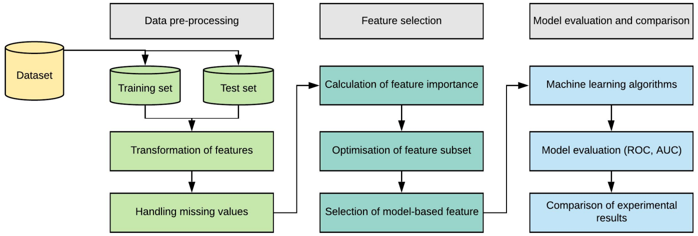

Dataset
Data pre-processing
Training set
Test set
Transformation of features
Handling missing values
Feature selection
Calculation of feature importance
Optimisation of feature subset
Selection of model-based feature
Model evaluation and comparison
Machine learning algorithms
Model evaluation (ROC, AUC)
Comparison of experimental results

Figure 1. System Architecture

Research Approach

In the experiment, credit scoring data sets from Kaggle competition called ‘Give me some credit’ is used to build classifiers. To evaluate the classification performance of credit scoring model, it is compared with logistic regression as benchmark model and two ensemble methods (random forest and extreme gradient boosting) are employed. Receiver operating characteristic (ROC) and Area under the curve (AUC) are chosen to measure the performances among the models. Figure 2 shows that the random forest and extreme gradient boosting(XGBoosting)-based models perform well in imbalanced credit scoring context as Brown and Mues (2012) mentioned in the paper.

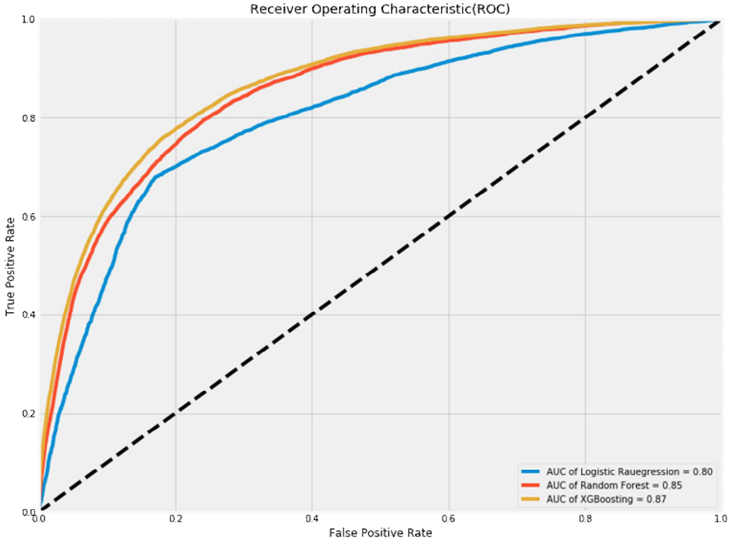


Figure 2. ROC and AUC for models

References

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with applications*, 39, 3446-3453