Credit scoring using machine learning algorithms

Abdykalykova A E

Kazakh National Technical University after K.I.Satpaeva, Almaty

Corresponding author's e-mail: asselya081296@gmail.com

Abstract

Credit scoring it is a way to predict borrower's behavior or future possibility of delay using input and historical information. It affects government's economy, thus financial companies should give loans only to responsible and solvent a part of population. Every financial institution has its own scorecard models. Usually, those models are based on logistic regression and decision tree, because of their simple interpretability. Since the data volume grows, variety and types of modern predictive methods develop the possibility of increasing the predictive power of models is growing too. This thesis is going to be about process of building qualitative model and modern optimization methods.

Keywords: Scorecard; Credit; Machine learning.

1 Introduction

Recently, consumer spending has become one of the key factors in macroeconomic conditions worldwide. Therefore, it is important to focus on credit scoring in order to better predict consumer behavior.

Credit scoring is a method that helps to decide whether to provide loans to consumers, it is a probability of person's debt repay in a timely manner, based on person's credit history. People are considered financially reliable when their score is higher. Credit scoring eliminates the human factor and uses only reliable data.

There are two main problems in credit scoring: giving a loan to a bad borrower and refusing to a good one. However, there are many ways to accomplish this problem, and some of them are more effective than others.

Advanced statistical and mathematical methods provide fast and automatic tools that help to make effective decisions. Models based on machine learning algorithms and artificial intelligence are believed to be more effective to support approval process in finance companies. The combination of machine learning methods can make a big contribution to the lending system and will be much more complicated in terms of use. Because machine learning forecasts are more adaptive and flexible to change, they can produce more accurate results. Therefore, the purpose of this thesis is to identify the most reliable and effective method.

2 Methods

The dataset consist of data from personal information, credit bureau, transactions and so on. Because of consumer privacy protection laws, all individual identification data were encrypted. The aim of any machine-learning model is the identification of statistically reliable relationships between input data features and the target variable. The target variable is a binary value, indicating whether an account is delinquent by 90 days or more within 12 months.

The dataset was preprocessed before the final feature selection using several classical methods like Chi-squared, Information Gain and new methods such as Lime/Shad (understanding of parameter's influence on model prediction), PCA (reducing data dimension). The selected features contain information about type of job, experience, number of credits, incomes & expenses, age, and marital status and so on.

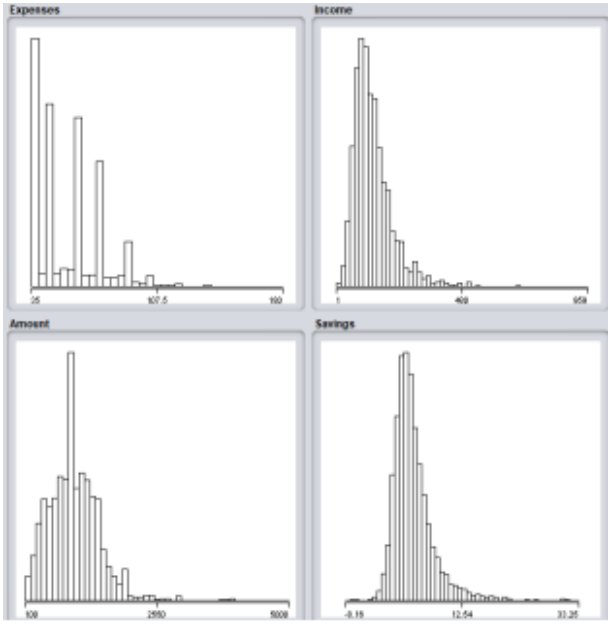


Figure 1 Example of features' distribution

Then data was trained and tested with multiple number of supervised machine learning models, such as Logistic Regression, Decision Tree, Random Forest, Naïve Bayes,

XGBoost, Support Vector Machine. Cross-validation technique was performed and parameters of algorithms were correctly tuned for models' better work. The performance was evaluated using several performance measurements such as accuracy, AUC-score, ROC-curve, Gini, confusion matrix, recall, precision, F-score.

3 Results

The result of the research should propose the best model to predict Credit Defaults.

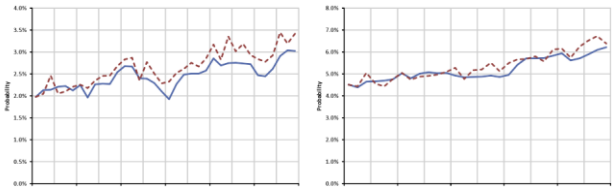
Models were compared by several metrics and after all of the comparisons was selected the best model.

There are given main metrics' results:

Accuracy: XGBoost – 81,2%, SVM – 77%, Random Forest – 75%, Logistic Regression – 74.6%,Naïve Bayes - 69% and Decision Tree - 70%.

AUC: XGBoost – 0.854, SVM – 0.751, Random Forest – 0.774, Logistic Regression – 0.698, Naïve Bayes – 0.685 and Decision Tree – 0.704.

F-score: XGBoost – 87,2%, SVM – 86.3%, Random Forest – 84.9%, Logistic Regression – 82%,Naïve Bayes 76% and Decision Tree - 80%.



SVM XGBoost

Figure 2 Comparing prediction and fact of target

4 Conclusion

XGBoost and SVM have shown the best results comparing with such a popular machine learning models used in credit scoring as Decision tree and Logistic Regression. Applying such methods as preprocessing dataset to avoid imbalance (and as a consequence incorrect result of the models), new optimizing methods in feature selection (reducing over-fitting, improving accuracy, reducing training time) helped to achieve such a good results, when most of the models have a high values of metrics.

From risk management perspective, the aggregation of machine-learning forecasts may have much to contribute to the management of systemic risk.

References

[1] Stein R M 2005 The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing *Journal of Banking & Finance*

[2] Baesens B, Gestel T V, Viaene S, Stepanova M, Suykens J, Vanthienen J 2003 *Benchmarking state-of-the-art classification algorithms for credit scoring*

[3] Hand D, Henley W 1997 Statistical classification methods in consumer credit scoring: A review *Journal of the Royal Statistical Society*

[4] Galindo J, Tamayo P 2000 *Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications*

[5] Anne Kraus 2014 *Recent Methods from Statistics and Machine Learning for Credit Scoring*

[6] Foster D, Stine R 2004 *Variable selection in data mining: Building a predictive model for bankruptcy*

[7] Loh W-Y *Fifty Years of Classification and Regression Trees*