Preprint of the paper:

Profit-Oriented Feature Selection in Credit Scoring Applications

by

Nikita Kozodoi, Stefan Lessmann, Konstantinos Papakonstantinou, Bart Baesens

Kindly find the final, revised paper from the publisher's website:

https://doi.org/10.1007/978-3-030-18500-8_9

Abstract

In credit scoring, feature selection aims at removing irrelevant data to improve the performance of the scorecard and its interpretability. Standard feature selection techniques are based on statistical criteria such as correlation. Recent studies suggest that using profit-based indicators for model evaluation may improve the quality of scoring models for businesses. We extend the use of profit measures to feature selection and develop a wrapper-based framework that uses the Expected Maximum Profit measure (EMP) as a fitness function. Experiments on multiple credit scoring data sets provide evidence that EMP-maximizing feature selection helps to develop scorecards that yield a higher expected profit compared to conventional feature selection strategies.

Keywords: feature selection, credit scoring, profit maximization

Profit-Oriented Feature Selection in Credit Scoring Applications

Nikita Kozodoi[1,2], Stefan Lessmann[1], Bart Baesens[3], and Konstantinos Papakonstantinou[2]

[1] Humboldt University of Berlin, Berlin, Germany
[2] Kreditech Holding, Hamburg, Germany
[3] Catholic University of Leuven, Leuven, Belgium

Abstract. In credit scoring, feature selection aims at removing irrelevant data to improve the performance of the scorecard and its interpretability. Standard feature selection techniques are based on statistical criteria such as correlation. Recent studies suggest that using profit-based indicators for model evaluation may improve the quality of scoring models for businesses. We extend the use of profit measures to feature selection and develop a wrapper-based framework that uses the Expected Maximum Profit measure (EMP) as a fitness function. Experiments on multiple credit scoring data sets provide evidence that EMP-maximizing feature selection helps to develop scorecards that yield a higher expected profit compared to conventional feature selection strategies.

Keywords: feature selection, credit scoring, profit maximization

1 Introduction

One of the most important tasks in credit risk analytics is to decide upon loan provisioning. Binary scoring systems are widely deployed to support decision-making and predict applicants willingness and ability to repay debt. Financial institutions face costs of gathering and storing large amounts of data on customer behavior used to score applicants. In addition, companies need to comply with regulation that enforces comprehensible models. Feature selection aims at solving this problem by removing irrelevant data, which can reduce costs and improve the scorecard performance and interpretability.

Recent literature criticized a widespread practice of using standard performance measures such as area under the receiver operating characteristic curve (AUC) for evaluating scoring models [6]. Relying on profit-based indicators may improve scorecard profitability [4, 15]. This finding stresses the importance of using value-oriented feature selection strategies that identify the optimal subset of variables in a profit-maximizing manner. The goal of this paper is to introduce the profit maximization framework to the feature selection stage to facilitate the business-driven model development.

We develop a wrapper-based feature selection framework that uses the Expected Maximum Profit measure (EMP) as a fitness function. EMP has been

previously used in credit scoring for model evaluation [15]. The advantage of the proposed approach is that it searches for variable subsets that optimize the business-inspired indicator. To validate the effectiveness of our method, we conduct an empirical experiment on multiple credit scoring data sets.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on profit-driven credit scoring and feature selection methods. Section 3 describes our experimental setup, whereas Section 4 presents empirical results. In Section 5, we discuss the main conclusions of our study.

2 Related Literature

2.1 Profit-Oriented Credit Scoring

The credit scoring literature has proposed several profit measures to improve the quality of scorecards. Serranco-Clinca et al. use the internal rate of return based on the loan interest [13]. Finlay proposes estimating a contribution of each applicant to the profit of the financial institution [4]. Both these measures imply replacing binary default indicator by a continuous target variable and therefore transform a classification problem into the regression task.

Recently, Verbraken and colleagues developed the EMP measure [15]. EMP is based on the costs of the incorrect classification of *bad* loans (defaults) and benefits of the correct prediction of *good* ones (repayers). It can be computed as:

$$\text{EMP} = \int_0^1 \Big[B \cdot \pi_0 F_0(t) - C \cdot \pi_1 F_1(t)\Big] f(B) d(B), \tag{1}$$

where B is the expected loss in case of default and C is the return on the investment, π_i are prior probabilities of *good* and *bad* loans, and $F_i(t)$ are predicted cumulative fractions of class i based on cutoff t. The return on investment is assumed to be constant, whereas the expected loss is a stochastic variable based on the loss given default and exposure at default (see [15] for details).

EMP can be interpreted as the incremental profit from deciding on credit applications using a scorecard compared to a baseline scenario where credits are granted without screening. In this paper, we use the EMP criterion to measure the scorecard profitability.

2.2 Feature Selection

Feature selection methods split into filters, wrappers and embedded methods [5]. Filters rank and select features based on some general data characteristics. Popular measures include correlation, information gain and others [12]. Filters are fast and efficient but they were shown to perform poorly compared to wrappers and embedded methods [5]. Embedded methods conduct feature selection simultaneously with the model training. One of the popular approaches is recursive feature selection using the SVM framework [12]. The drawback of embedded methods is that they can only be applied within a specific model.

Wrappers go through different feature subsets and select the optimal subset based on the model performance. Since evaluating all possible feature combinations is computationally expensive, research has suggested heuristic search strategies. Popular approaches are sequential forward selection (SFS) and sequential backward selection (SBS) [5]. SFS starts with an empty model and iteratively adds features, selecting the one which brings the largest performance gain, whereas SBS starts with a full set of features and eliminates those contributing the least to the model performance. The search is continued until there is no further improvement. Another strategy relies on evolutionary algorithms such as genetic algorithms (GA) [16]. GAs operate on a population of individuals, where each individual represents a model with binary genes indicating inclusion of specific features. At each generation, a new population is created by selecting individuals according to their fitness (model performance), recombining them together and undergoing mutation. The individual with the highest fitness is selected after running multiple generations.

The literature on profit-oriented credit scoring focuses on model selection and parameter estimation but does not consider the feature selection stage. Existing studies on value-driven feature selection focus on feature costs. Some researchers suggest using a budget constraint that limits the maximal cost of the selected features [11]. Another approach is to use cost-adjusted ranking criteria when applying filter methods [3].

To the best of our knowledge, research on value-driven feature selection in credit scoring is currently limited to the embedded regularization framework for SVM [9, 10]. Recent benchmarking studies in credit scoring have shown that SVM performs poorly in comparison with other classifiers [7]. Given these results, developing a profit-driven feature selection approach that is not limited to SVM contributes to the literature. In this paper, we focus on wrappers due to their flexibility and better performance compared to filters.

3 Experimental Setup

3.1 Data Sets

The empirical evaluations are based on ten retail credit scoring data sets. Data sets *australian* and *german* stem from the UCI Repository [8]. The data sets *pakdd*, *lendingclub* and *gmsc* were provided by different companies for the data mining competitions on PAKDD and Kaggle platforms. Datasets *bene1*, *bene2* and *uk* were collected from financial institutions in the Benelux and UK [2]. The *thomas* data set is provided by [14], whereas *hmeq* was collected by [1].

Each of the data sets has a unique set of features describing the loan applicant (e.g., gender, income) and loan characteristics (e.g., amount, duration). Some data sets also include information on previous loans of the applicant. The target variable is a binary indicator whether the customer has repaid the loan or not. Table 1 summarizes the main characteristics of the data sets.

Table 1. Credit Scoring Data Sets (After Preprocessing)

Data Label	Sample Size	Num. Features	Default Rate
australian	690	42	0.4449
german	1,000	61	0.3000
thomas	1,225	28	0.2637
bene1	3,123	83	0.3333
hmeq	5,960	20	0.1995
bene2	7,190	28	0.3000
uk	30,000	51	0.0400
lendingclub	43,344	206	0.1351
pakdd	50,000	373	0.2608
gmsc	150,000	68	0.0668

3.2 Modeling Framework

The modeling pipeline consists of several stages. First, each data set is preprocessed in the same way. We impute missing values with means for continuous features and with modes for categorical features. Next, we encode categorical variables with $k-1$ dummies, where k is the number of unique categories.

The data sets are randomly partitioned into two subsets: training (70% cases) and holdout data (30%). On the training set, we use 5-fold cross-validation to perform feature selection. Then, we use the whole training set to train classification models with the identified feature subsets and evaluate results on the holdout data. The partitioning is repeated 10 times on each data set.

On a feature selection stage, we use three wrappers: SFS, SBS and GA. The parameters of GA were selected based on the grid search on a subset of training data: number of generations and number of individuals were set to 200. For each of the feature selection algorithms, we use two performance measures as objective functions. Relying on EMP as a fitness function is a central element of our approach, whereas using AUC serves as a benchmark for the standard feature selection techniques. Logistic regression is used as a base classifier.

4 Empirical Results

The performance of different methods is compared by the mean model ranks in terms of AUC and EMP. To compute the ranks, we order all algorithms by AUC and EMP values within each modeling trial on each of the data sets, and average the model positions. The results are depicted in Figure 1.

Results suggest that EMP-based wrappers identify feature subsets that yield a higher EMP on the holdout data, whereas using AUC as objective leads to models with a higher AUC. For all three wrappers, EMP optimization during feature selection generalizes to a higher expected profit on the new data. Therefore, our results emphasize the importance of selecting the appropriate fitness function in the early stages of scorecard development. If the goal of the scoring model is to maximize the expected profit, this measure should also be used

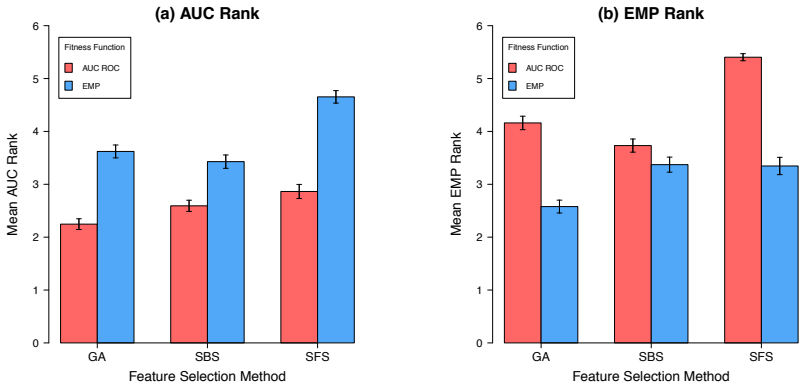

Fig. 1. Mean ranks of different feature selection methods in terms of AUC (left) and EMP (right). The ranks are computed for 6 methods and aggregated across 10 trials on 10 data sets. Whiskers refer to intervals of one standard error around the mean.

as an objective for feature selection. Relying on AUC as one of the standard performance measures results in a suboptimal scorecard with a lower EMP.

We also observe differences in terms of the model complexity. For the sequential methods, EMP-driven wrappers reach the stopping criteria earlier, resulting in a lower average number of the selected features for SFS (14 compared to 25) and a higher number of features for SBS (89 compared to 81). These results provide evidence that using profit-driven fitness function may also lead to a faster convergence, which is important for reducing the computational time.

5 Conclusions and Future Work

This paper presents a profit-driven framework for feature selection in credit scoring. We use the recently developed EMP measure as a fitness function for wrapper-based feature selection. The effectiveness of our approach is evaluated on ten real-world retail credit scoring data sets.

Empirical results indicate that the proposed profit-maximizing feature selection framework identifies variable subsets that yield a higher expected profit compared to methods based on standard performance measures. These results stress the importance of using the business-inspired metrics for feature selection. Relying on a standard practice of using statistical measures such as AUC may lead to scorecards with a lower profitability, which motivates implementing the profit maximization on different stages of the model development.

Future research could pursue several directions. For practitioners, it would be important to extend the profit-driven framework to other stages of model development. A benchmarking study with a rich set of EMP-based wrappers would help identifying the optimal search strategy for profit-driven feature selection. Another direction would be to use the developed approach in other business applications such as customer churn.

References

1. Baesens, B., Roesch, D., & Scheule, H. (2016). Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. John Wiley & Sons. doi:10.1002/9781119449560
2. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627-635. doi:10.1057/palgrave.jors.2601545
3. Boln-Canedo, V., Snchez-Maroo, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, 86, 33-45. doi:10.1016/j.knosys.2015.05.014
4. Finlay, S. (2010). Credit scoring for profitability objectives. European Journal of Operational Research, 202(2), 528-537. doi:10.1016/j.ejor.2009.05.025
5. Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing): Springer-Verlag. doi:10.1007/978-3-540-35488-8
6. Hand, D. J. (2005). Good practice in retail credit scorecard assessment. Journal of the Operational Research Society, 56(9), 1109-1117. doi:10.1057/palgrave.jors.2601932
7. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136. doi:10.1016/j.ejor.2015.05.030
8. UCI Machine Learning Repository http://archive.ics.uci.edu/ml/
9. Maldonado, S., Bravo, C., Lopez, J., & Perez, J. (2017). Integrated framework for profit-based feature selection and SVM classification in credit scoring. Decision Support Systems, 104, 113-121. doi:10.1016/j.dss.2017.10.007
10. Maldonado, S., Prez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. European Journal of Operational Research, 261(2), 656-665. doi:10.1016/j.ejor.2017.02.037
11. Min, F., Hu, Q., & Zhu, W. (2014). Feature selection with test cost constraint. International Journal of Approximate Reasoning, 55(1), 167-179. doi:10.1016/j.ijar.2013.04.003
12. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28. doi:10.1016/j.compeleceng.2013.11.024
13. Serrano-Cinca, C., & Gutirrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, 8, 113-122. doi:10.1016/j.dss.2016.06.014
14. Thomas, L. C., Edelman, D. B., Crook, J. N. (2002) Credit Scoring and its Applications. Philadelphia: SIAM. doi:10.1137/1.9780898718317
15. Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, 238(2), 505-513. doi:10.1016/j.ejor.2014.04.001
16. Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In Feature Extraction, Construction and Selection (pp. 117-136). Springer, Boston, MA.