# Impact of Unusual Features in Credit Scoring Problem

Luiz F. V. Vercosa[1], Rodrigo C. Lira[2], Rodrigo P. Monteiro[3], Kleber D. M. Silva[1], Jailson O. L. Magalhaes[1], Alexandre M. A. Maciel[1], Byron L. D. Bezerra[1], Carmelo J. A. Bastos-Filho[1]

[1] Universidade de Pernambuco, Brazil
{lfvv,kdms,jolm,amam,byronleite,carmelofilho}@ecomp.poli.br
[2] Instituto Federal de Pernambuco, Brazil
rodrigo.lira@paulista.ifpe.edu.br
[3] Universidade Federal de Pernambuco, Brazil
rodrigo.paula@ufpe.br

**Abstract.** Standard features used for Credit Scoring includes mainly registration and financial data from customers. However, exploring new features is of great interest for financial companies, since slight improvements in the person score directly impact the company revenue. In this work, we categorize features from open credit scoring datasets and compare them with the features found in a real company dataset. The company dataset contains unusual feature groups such as historical, geolocation, web behavior, and demographic data. We performed bivariate tests using the Kolmogorov-Smirnov metric and features to assess the performance of the particular feature groups. We also generated a score of good payer by using AdaBoost, Multilayer Perceptron, and XGBoost algorithms. Then, we analyzed the results with different metrics and compared them with the real company results. Our main finding was that these features added a small improvement to current datasets. We also identified the most promising feature groups and noticed that the tuned XGBoost performed better than the company solution in three out of four deployed metrics.

CCS Concepts: • **Computing methodologies** → **Supervised learning by classification**; • **Information systems** → **Content analysis and feature selection**.

Keywords: credit scoring, feature groups, novel dataset, web crawling.

## 1. INTRODUCTION

The availability of financial credit is essential for financial agents, individuals, and companies. The agents make a profit through interests, while individuals and companies can pursue new investments to buy goods or expand their businesses. From the financial agent perspective, money should be lent for those willing to pay it back. The process of lending should also be simple, fast, and scalable. Consequently, the financial institutions present a trend of changing manual credit approval analysis to automatic and scalable alternatives [Mester et al. 1997].

In this context, credit scoring is a widely adopted technique, which allows a more reliable and scalable way of managing money lending risks [Thomas et al. 2002]. It mainly consists of applying computational techniques on customer data to generate a good payer score for each customer.

Open credit scoring datasets available in the literature (*e.g.*, German [Ekin et al. 1999], Taiwan [Yeh and Lien 2009], Australian and Japanese [He et al. 2018]) have reliable features traditionally used to tackle the credit scoring problem, *e.g.*, sex, marital status, previous payments, state, and age. However, we assume the hypothesis that using alternative and unusual information, *e.g.*, geolocation, and web-related may improve credit scoring assertiveness. Geolocation data identifies the type of places

This work was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - FinanceCode 001, FACEPE, and CNPq.

surrounding the customer's house, whereas web-related features check the customer's preferences and habits on the web. These type of data can help to identify customer profile what can lead to their paying habits.

In this paper, we analyzed the impact of new information sources on credit scoring performance. The dataset deployed in this work contains information about geolocation, web behavior, and demographic features. A company that operates in the information technology area provided the data, mainly obtained via web crawling. To the best of our knowledge, no work available in literature has studied the impact of using this kind of unusual information on the credit scoring problem.

In addition to the analysis of novel features, we employed boosting techniques such as AdaBoost and XGBoost to calculate the credit score. They have already been successfully applied in previous works [Zhou and Lai 2009] [He et al. 2018]. Finally, we used known metrics from literature as well as the company metric to analyze the results.

We organized the remaining of the manuscript as follows: we describe how the dataset was created in Section 2. In Section 3, we describe the methodology used in this work. Section 4 describes the results obtained through features analysis and the creation of credit score by the models. Finally, Section 5 summarizes the paper and list possible future directions.

## 2. DATA ACQUISITION

We developed this work through a partnership with an IT company that performs data mining according to the Cross-Industry Standard Process for Data Mining (CRISP-DM) [Wirth and Hipp 2000]. Many companies in this field use registration data (*i.e.*, filled out the information in the application form)[Mester et al. 1997], but as a differential, our partner enriches their datasets in two manners.

At first, they record customers' behavioral history using data from other solutions of their own and data provided by third-party companies. By doing so, they obtain data that inform the frequency their products have recorded a person. The second way is by web crawling, *i.e.*, to search in public websites for data about people. In this manner, they collect other data related to the person, *e.g.*, geolocation, web behavior, registration data exposition, and census. The web crawling does not provide the essential features, such as gender, salary, previous payments, or age. However, it may provide insights into the habits and preferences of the customer. For example, knowing if a person has a public profile in a social network may help understand how this person is acquainted with the technology. Therefore, it helps the machine learning algorithm to create a profile of that person. The hypothesis we state is that a personal profile is linked to its consuming habits that can ultimately tell if a person is willing to be indebted.

In this work, we used the features collected by web crawling, which we will reference as the company dataset. We found a few works that presented unusual features for credit scoring as our dataset. In [Niu et al. 2019], the authors explored social network features coded as social stability, social exposure, and social quality. Their data were extracted from mobile phones and the authors concluded that those features were useful to improve the results. In [Liberati and Camillo 2018] the authors explore features that come from the psychological trait of the customers and found that they decreased the error of the employed models. Encouraged by recent success obtained by the novel features explored in other works, our article goes further in this endeavor and investigates novel feature groups obtained mainly through web crawling and including categories such as demographic, social networks, social programs, and web.

## 3. METHODOLOGY

### 3.1 Analysis of Other Credit Dataset Features

We have analyzed the features of four other credit datasets to compare the features showed in the company dataset. The datasets include three well-known credit datasets available in UCI Machine Learning Repository being German, Taiwan, and Japanese datasets. We also have included a Brazilian credit scoring dataset provided in a data mining competition [PAKDD Conference 2009]. The German, Japanese and Brazilian datasets regard credit granting for individuals, whereas Taiwan dataset is related to the identification of defaults in payments of credit cards.

We have grouped the features by similarity according to their sources. Therefore, we realized that most features could be described in three categories: Personal, Financial, and Lending information, as we show in Table I. The personal category contains individual information, *e.g.*, age, marital status, job type, and gender. Financial regards banking and property information, *e.g.*, credit history, incomes, and properties. Lending contains information about the financial product that the customer is requiring, *e.g.*, product type, purpose, amount, and duration. It is important to notice that Taiwan and Japanese datasets do not present Financial features.

Table I: Features categories found in analyzed credit datasets.

| | Features categories | | |
|---|---|---|---|
| Dataset | Personal | Financial | Lending |
| German | age, gender, marital status, residence time, employment time, job type, telephone ownership, housing, foreign worker | banking account, property, installment rate, credit history, debtors and guarantors, savings amount | purpose, credit amount, installment rate and, duration |
| Brazilian | age, gender, marital status, residence time, employment time, job type, housing, spouse profession, address, education level, birth address, etc | banking account, property (cars), exclusive account, month income, additional income, credit card (type) | product type (lending), payment day, submission (type) |
| Taiwan | age, gender, marital status, and education | | amount, bill statements, past payments, payment history |
| Japanese | employment status and time, gender, marital status, age, and housing area | | purchased item, deposit, monthly payment, and payment period |

### 3.2 Description of the Company dataset

The company dataset contains information about real customers. It has 175 features of 500 thousand customers. We suppressed the customers' identification to ensure their privacy. Among those features, six are categorical, and 169 are numerical. The ground truth is provided by the company's clients, *i.e.*, businesses that grant credit. This dataset presents approximately 246 thousand good payers and 254 thousand bad payers, being previously balanced. The features of this dataset belong to different categories, which are described in Table II. In this table, we also inform the number of features per category, and a short identification (code) for each group.

The datasets listed in Table I have features directly related to the customer and the lending, whereas the company dataset presented in Table II has only a few features of the kind represented by

Registration group. On the other hand, the company dataset presents categories not found in other datasets and indirectly related to the customers, *e.g.*, web, geolocation, and historical. For instance, the web category checks the customer web exposition to selected subjects, *e.g.*, interests in politics, arts, and books. The purpose of this group of features is to create a profile that might indicate good payers. The geolocation category, in turn, checks the customer's home exposition to specific places, *e.g.*, churches, police stations, shopping centers. This group of features can reveal characteristics of the neighborhood as nobility and preferences of the customer. Another impressive group of features is the Historical group. The company has solutions in other market areas as credit card granting and car insurance. Therefore, this group of features reveals the frequency and time that the customer registration data, *e.g.*, e-mail, and phone, appears in these other datasets.

Table II: Categories of features of the Company dataset

| **Category** | **Code** | **Description** | **Count** |
|---|---|---|---|
| Key | - | ID of the customer | 1 |
| Classification | - | classification as good (1) or bad client (0) | 1 |
| Government | GOV | indicates the customer as a civil or military government employee | 2 |
| Politics | POL | check customer relation to politics | 2 |
| Registration | REG | monthly income, customer state, and social class | 3 |
| Social Program | SOC | indicates whether the customer benefits from social programs | 4 |
| Financial | FIN | financial data from the customer | 8 |
| Historical | HIS | exposure of customer registration data in another company datasets | 8 |
| Web | WEB | check customer web exposition and interests to previously selected subjects | 19 |
| Demographic | DEM | features based on demographic census | 53 |
| Geolocation | GEO | check customer's home geographic exposition to previously selected places | 76 |

### 3.3 Models

In our experiment, we employed three machine learning models: XGBoost [Chen et al. 2015], Adaboost [Ying et al. 2013], and Multilayer Perceptron (MLP) [Nazzal et al. 2008]. The first one, XGBoost, is already used by the company. Adaboost was chosen for being, as XGBoost, a boosting technique that has been successfully employed for the credit scoring task [Zhou and Lai 2009]. Finally, MLP was used, since it is a well-known technique for classification problems.

### 3.4 Evaluation Metrics

The binary classification of good and bad payers is not the main objective of the credit companies, but the payer score is. Therefore, we chose metrics that attend to those requirements. Those metrics were the Lift, Mean Squared Logarithmic Error (MSLE) [Massmann and Holzmann 2012], Area Under ROC Curve (AUC) [Fawcett 2005], and the Kolmogorov-Smirnov (KS) test [Neuhauser 2011]. The AUC and KS metrics allow us to observe how well the scores given by the models can split the groups of good and bad payers. To do so, they access the whole range of probabilities of good payers given by the models. On the other hand, the MSLE metric penalizes greater errors than smaller ones, allowing to measure the variance of the scores errors given by the models. Finally, the metric lift checks for errors in the extremes range of scores. In these ranges, the score should be more reliable for identifying good payers, *i.e.*, in the highest range and bad payers, *i.e.*, in the lowest range. Therefore, this metric checks the credibility of the best and worst model scores. The Lift metric is presented in Eq. 1 and comprehends the percentage of good payers (GP) in 90-100 percentile range, *i.e.*, 10% best scores, over the percentage of GP in the 0-10 percentile range, *i.e.*, 10% worst scores plus one, to prevent

division by zero. The company uses this metric. It can vary from zero to one where one is its best value.

$$lift = \frac{\text{GP in 90-100 percentile range}}{(\text{GP in 0-10 percentile range}) + 1} \quad (1)$$

The best and lowest value for MSLE is zero, while the other metrics range from 0 to 100%.

### 3.5 Experimental Methodology

We have performed two groups of experiments. For both groups we trained the models as classification problems. However, when analyzing the results, we employed metrics that considers the resulting probability for each class. The first experiment verifies which groups of features are of most importance when creating a credit score. We employed the XGBoost technique because it yielded the best results in our experiments. We used the metric already adopted by the company, the KS, which can correlate the score given by the model with the ground truth. The second group of experiments uses well-known metrics and all three models to generate credit scores. We also contrasted our results with those obtained by the company machine learning model, but they did not specify which model they use.

We performed each experiment 30 times, since the models results are non-deterministic. However, the company provided us the result of only one running of their non-deterministic model. In our experiment, we used random subsampling to split the dataset in 75% training slice and 25% test slice. We also fine-tuned the parameters of the models for the second group of experiments. For this task, we employed grid search (GridSearchCV) [Bergstra and Bengio 2012], an exhaustive search over specified parameters available in the scikit-learn library. As the search was exhaustive, it was essential to select only the most promising parameters.

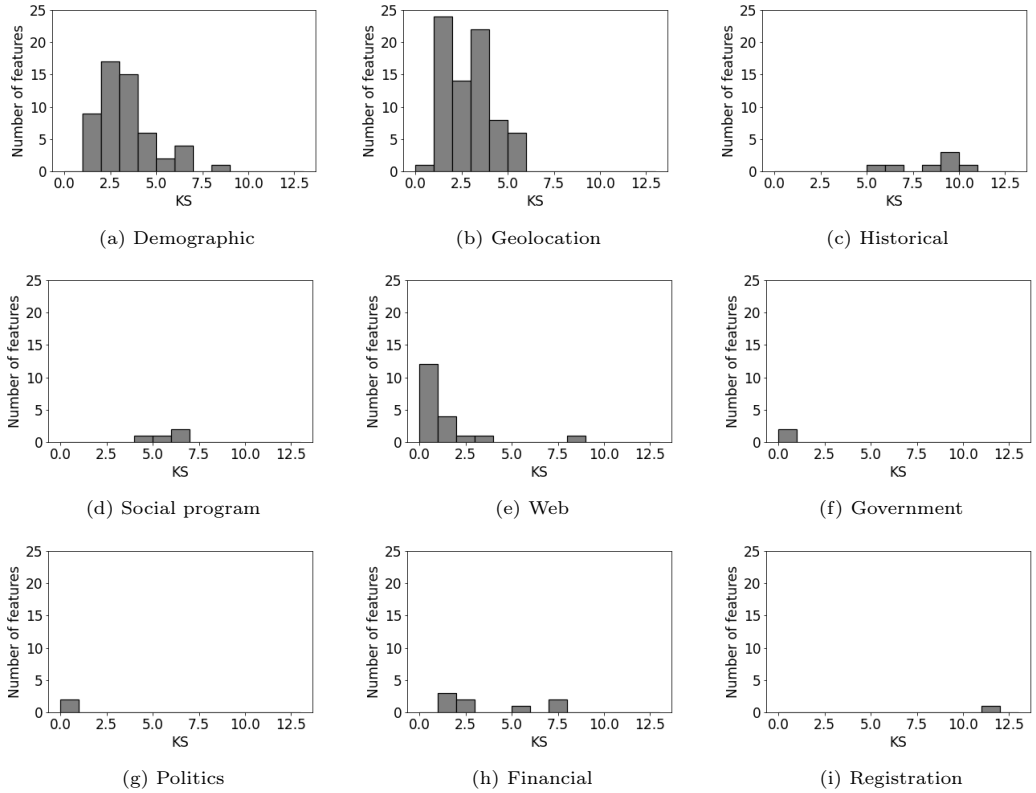
The MLP configuration presented the best results using two hidden layers with 400 neurons each, invscaling as the learning rate schedule, learning rate 0.2, adam as optimizer, ReLU as activation function, and 400 iterations without loss improvement to interrupt the learning process. The best XGBoost configuration presented the following parametric configuration: learning rate 0.05, gbtree booster, max-depth 4, and 400 estimators. The AdaBoost most prominent parameter values were: learning rate 0.3, SAMME.R learning algorithm, and 2700 estimators. We performed the simulations using Google Colab and Amazon Web Services (AWS) infrastructures.

## 4. RESULTS

### 4.1 Analysis of the Features Groups

First of all, we have analyzed the company dataset. We have created a histogram (Fig.1) for each group of features based on KS from a bivariate test. The histograms shows the importance of each feature inside a group. From this perspective, one can notice that Historical and Registration groups presented features with the best individual KS, whereas Politics and Government groups obtained the worst results. In addition, in Fig.1 we can contrast the number of features of each group. One can notice that the Geolocation and Demographic groups contain the majority of features.

In our first experiment, we have analyzed the importance of each feature group from the company dataset. To attain that, we employed the XGBoost model with each feature group separately. Next, we performed a bivariate test between the ground truth and the model's good payer scores. These




(a) Demographic (b) Geolocation (c) Historical

(d) Social program (e) Web (f) Government

(g) Politics (h) Financial (i) Registration

Fig. 1: Histogram of bivariate test with each feature and ground truth using KS metric. The features are shown in their respective groups being (a) Demographic, (b) Geolocation, (c) Historical, (d) Social Programs, (e) Web, (f) Government, (g) Politics, (h) Financial and (i) Registration.

results are shown in the bars of Fig. 2, where each bar represents a group of features by its code, as depicted in Table II. The HIS group, *i.e.*, the Historical group, obtained the best results. It may be explained by the fact that the company has many datasets about other products they have in the market. Some of the datasets include the requirement of credit card in retail companies and request for car insurance. The appearance of a customer registration data in those datasets suggests customer purchasing behavior, *e.g.*, a customer that tries to create credit cards in many different stores might be more likely to be a person with debts. This group of features also identifies whether the customer has multiple phone numbers and e-mails that also might indicate a fraudster. Finally, these features have time-related information that indicates frequency and activity.

Another group with high performance was the Registration group (REG), even though it has only three features. It is essential to notice that all the datasets analyzed included that kind of information since they are directly related to the customer. The Geolocation group is the third most important group. This group of features may indicate the customer's purchase power since it contains information about the analyzed person's neighborhood. Also, it suggests the customer's preference, which is linked to the idea of profiles. In the next position in the importance ranking, appears the Financial group, FIN. These features are also created based on other companies' datasets and are indirectly related to the customer. It explains the group not being at the top of the rank. Next, came the Demographic group. These features are based on census information that describes with more details the geographic region the person resides. The Web group is ranked sixth and indicates customer interest in certain subjects from the web, such as politics, books, and culture. Social program comprehends only four features, and it has similar performance to the Web category. Finally, the Government and Politics

groups had poor performance when comparing to the other groups.

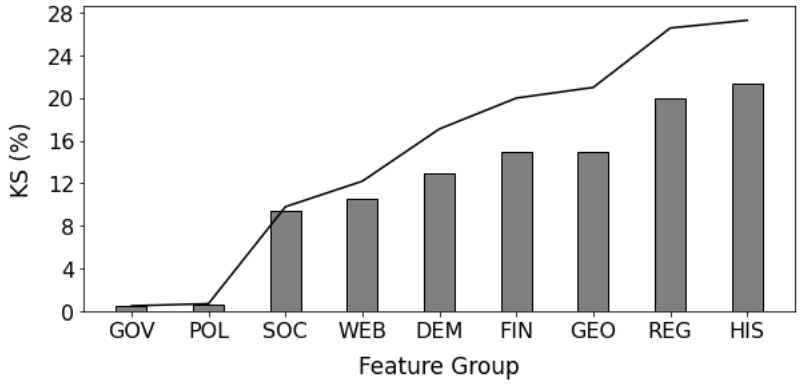



Fig. 2: KS obtained by each feature group (bars) and by the groups cumulatively (line).

The line displayed in Fig. 2 shows the credit scoring performance by sequentially adding the features groups to the XGBoost model and performing the bivariate analysis afterwards. For example, the line point over the GEO bar indicates that the model created a score using all groups except for REG and HIS. By contrasting the performance of REG features alone against its performance united with the other groups, one can see that there is a substantial improvement of over five KS percentual points. This suggests that the novel features can improve performance of the models, since traditional datasets only present features from REG group. We can also notice that the performance of the Historical group alone was higher than the REG Group, indicating that this is the most promising group.

### 4.2 Comparison with the Company Solution

Our second experiment comprehends an analysis of the performance of different models with metrics described in the literature. Table III depicts the results of the models. We can notice that XGBoost obtained the best results for metrics Lift, AUC, and KS metrics. By performing best in AUC and KS metrics, XGBoost presents the best score distribution. On the other hand, the Lift metric's best result indicates that our tuned XGBoost is the most confident technique when presenting extreme scores. For metrics Lift, AUC, and KS, AdaBoost obtained the second best results, followed by the Company Solution and MLP. However, for metric MSLE, the Company Solution yielded the best results being closely followed by XGBoost. The best performance in the MSLE metric indicates that the Company Solution presents smaller errors than the other solutions. For this same metric, MLP attained better results than the AdaBoost algorithm.

Table III: Models performance on the used metrics using full dataset and including the company results.

| **Models** | **Lift** | **MSLE** | **AUC (%)** | **KS (%)** |
|---|---|---|---|---|
| MLP | 0.6044 | 0.12 | 66.93 | 24.02 |
| AdaBoost | 0.6308 | 0.1240 | 68.08 | 26.34 |
| XGBoost | **0.6496** | 0.1099 | **68.81** | **27.24** |
| Company Solution | 0.6296 | **0.1095** | 68.30 | 26.61 |

## 5. CONCLUSION

The credit scoring problem is vital to financial companies because it is directly related to profit increase. In this work, we presented a real dataset with unusual features compared to other datasets from the literature. Among those features, there are groups related to geolocation, historical, web behavior, and demographic data. We noticed that these particular groups were able to improve the results of the registration features alone substantially. Also, the Historical group presented some of the best KS among all groups, followed by Registration, Geolocation, and Financial groups. We believe that these groups of features can identify different customer profiles based on consuming habits, preferences, and behavior. Thus, companies need to direct their attention and efforts to these features group. A limitation of our work is that we did not have access to the company's full dataset, which would help to contrast even further the results obtained the traditional features with the novel groups.

We also have tested the dataset using three well-known models for credit scoring. We noticed that XGBoost obtained the best performance than MLP and AdaBoost for metrics Lift, AUC, and KS. The Company Solution obtained the best performance for the MSLE metric.

We have ensured that the company results were included in our final experiment, but the comparison with their result did not provide statistical evidence because they provided only one running. However, it shows that our results are quite similar and have consistency. As future work, we intend to analyze the correlation of the features from the dataset and perform feature selection to find more promising features from distinct groups.

## REFERENCES


Bergstra, J. and Bengio, Y. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research* 13 (1): 281–305, 2012.

Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 2015.

Ekin, O., Hammer, P. L., Kogan, A., and Winter, P. Distance-based classification methods. *INFOR: Information Systems and Operational Research* 37 (3): 337–352, 1999.

Fawcett, T. An introduction to roc analysis tom. *Irbm* 35 (6): 299–309, 2005.

He, H., Zhang, W., and Zhang, S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Systems with Applications* vol. 98, pp. 105 – 117, 2018.

Liberati, C. and Camillo, F. Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society* 69 (12): 1994–2005, 2018.

Massmann, C. and Holzmann, H. Analysing goodness of fit measures using a sensitivity based approach. *EGUGA*, 2012.

Mester, L. J. et al. What's the point of credit scoring? *Business review* 3 (Sep/Oct): 3–16, 1997.

Nazzal, J. M., El-Emary, I. M., and Najim, S. A. Multilayer perceptron neural network (mlps) for analyzing the properties of jordan oil shale 1, 2008.

Neuhauser, M. *Nonparametric statistical tests: A computational approach*. Chapman and Hall/CRC, 2011.

Niu, B., Ren, J., and Li, X. Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information* 10 (12): 397, 2019.

PAKDD Conference. 13th Pacific-Asia Knowledge Discovery and Data Mining Conference (PAKDD 2009) - Data Mining Competition, 2009.

Thomas, L. C., Crook, J., and Edelman, D. *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

Wirth, R. and Hipp, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK, pp. 29–39, 2000.

Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36 (2): 2473–2480, 2009.

Ying, C., Qi-Guang, M., Jia-Chen, L., and Lin, G. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica* 39 (6): 745–758, 2013.

Zhou, L. and Lai, K. K. Adaboosting neural networks for credit scoring. In *The Sixth International Symposium on Neural Networks (ISNN 2009)*. Springer, pp. 875–884, 2009.