Analysis of Distinct Feature Groups in the Credit Scoring Problem

Luiz F. V. Vercosa[1], Rodrigo C. Lira[1,2], 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. Registration and financial data have been traditionally used for the credit scoring problem. However, slight improvements in the reliability of the scores positively impacts financial companies. Therefore, exploring new features is a strategic task. This work analyzes the importance of new feature groups not commonly employed for the credit scoring task and others already used. We categorized features from open credit scoring datasets, such as German and Australian and compared their groups with the ones of a company dataset used in this work. Our dataset contains unusual feature groups, such as historical, geolocation, web behavior, and demographic data. In our analyzes, we first conducted bivariate tests with each feature-pair to assess their individual importance. Secondly, we ran XGBoost machine learning model with each feature group to evaluate each group importance. We also applied feature selection with binary Particle Swarm Optimization to assess the groups importance when combined. Next, we employed correlation tests to find inner and inter-correlation among the features groups. Finally, we used the company dataset and employed AdaBoost, Multilayer Perceptron, and XGBoost algorithms to find the best model for the task. Some of our main findings were that the unusual features added a slight improvement to registration features. We also detected reasonable inner correlation among some feature groups and found that all groups were relevant for the task with the Historical Group as the most promising. Lastly, XGBoost obtained the best performance over AdaBoost and Multilayer-perceptron for the task.

Categories and Subject Descriptors: H.4.0 [**Information Systems Applications**]: General; I.2.m [**Artificial Intelligence**]: Miscellaneous

Keywords: credit scoring, feature groups, machine learning, 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's perspective, money should be lent to 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.

The authors would like to thank the FACEPE and UPE for the financial support during the development of this work.

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, province, and age. However, we hypothesize that using alternative and unusual information, *e.g.*, geolocation and web-related, may improve credit scoring assertiveness. Geolocation data identifies the type of places surrounding the customer's house, whereas web-related features check the customer's preferences and habits on the web. These types of data can help to identify customer profiles that can lead to their paying habits.

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, only our previous work has studied the impact of using this kind of unusual information on the credit scoring problem.

In this paper, we extend our previous work that showed the impact of new information sources on credit scoring performance [Verçosa et al. 2020]. As new contributions, we analyzed the correlation among features of the same group and of distinct groups. We also added a second way of investigating each feature importance through XGBoost gain.

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 analyzes and the creation of credit scores by the models. Finally, Section 5 summarizes the paper and list possible future work 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]. However, 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, they may provide weak features that combined can become relevant and useful for the model. In addition, some of these features are relevant individually, such as *flag_social_network*. This particular feature indicates whether a person has a public profile in a social network and 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 may relate to its consuming habits that can ultimately tell if a person is willing to be indebted.

We used the features collected by web crawling in this work, 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 was extracted from mobile phones, and the authors concluded that those features helped improve the results. In [Liberati and Camillo 2018], the authors explore features that come from the customers' psychological traits and found that they decreased the employed models' error.

In [Paraíso et al. 2021], the authors explored data of network's topology extracted from mobile phones and used techniques such as node embeddings to increase the model precision in microfinance credit scoring problem. The authors claim satisfactory results with the addition of these features. In [Djeundje et al. 2021], the authors explored the features of mail usage, psychometric variables, and demographic variables to increase the precision to predict the credit risk of a new account with people without a credit history. 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 behaviour.

3. METHODOLOGY

3.1 Analysis of Other Credit Dataset Features

We have analyzed four other credit datasets' features to compare the features shown in the company dataset. The datasets include three well-known credit datasets available in the UCI Machine Learning Repository: 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 the Taiwan dataset is related to identifying 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 shown 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 essential 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. In our experiments, we balanced the dataset to contain equal number of good and bad payers. This dataset's features belong to different categories, which are described in Table II. This table also informs 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. In contrast, the company dataset presented in Table II has only a few features of the kind represented by the 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 province, 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 Feature Importance

In many datasets, only a few features have substantial importance to the problem [Hastie et al. 2001]. Because of that, it is essential to investigate the contribution of each feature in problem resolution. In this paper, we accessed feature importance in three distinct manners. The first approach used was to conduct a bivariate test between each feature and the ground truth using Kolmogorov–Smirnov test. The second one was to use XGBoost algorithm to measure the importance of the features. The third and final one was to perform a feature selection by using binary Particle Swarm Optimization (PSO) [Kennedy and Eberhart 1995] in conjunction with the machine learning algorithms.

Table III: Computational time and settings for PSO feature selection models

Model	Parameters	GPU	Time (hours)
MLP	*hidden_ layer_ size*: 50 *activation*: *relu* *solver*: *adam* *alpha*: 0.0001 *learning_ rate*: *constant*	*no*	126
AdaBoost	*learning_ rate*: 1 *algorithm*: *SAMME.R* *number_ of_ estimators*: 50	*no*	80
XGBoost	*learning_ rate*: 0.1 *booster*: *gbtree* *max_ depth*: 3 *n_ estimators*: 100	*yes*	4

In XGBoost, the feature's contribution is calculated by changing the feature by random noise and measuring the improvement in the classification [Zheng et al. 2017]. It can be measured using three metrics: weight, coverage and, gain. Weight is the number of times the feature appears in a tree, and coverage is defined as the number of samples affected by the split [Shi et al. 2019]. In our paper, we used gain, which is the main factor of the importance in the tree branches [Zheng et al. 2017]. The gain captures the contribution of the corresponding feature by calculating its importance in each branch of each created tree.

PSO is a swarm intelligence algorithm that optimizes a search by gradually improving a candidate solution. To assess the quality of the solution, we used a machine learning algorithm. The binary version of PSO [Kennedy and Eberhart 1997] allows its application in feature selection problem where a portion of the features from the dataset are selected and a machine learning algorithm is applied. Therefore, it can decide which features are most promising together, regardless of their group. We used binary PSO in three distinct settings being PSO + MLP, PSO + AdaBoost and PSO + XGBoost.

In regards of the PSO settings used, we used 18 particles and the parameters were selected following the Clerk Constraint [Clerc and Kennedy 2002] being $c1 = 1.5$, $c2 = 1.5$, $w = 0.7$. With this parameters, the PSO is guaranteed to converge. Furthermore, we used 300 iterations and star topology. This topology provides fast convergence what was important given the high computational cost of the task. Table III shows the computational cost and settings used for all three machine learning techniques in the feature selection process.

3.4 Feature Correlation

The feature correlation analysis can help select the most promising features to be used by the model, allowing for reduction in the training phase and increase in model accuracy [Yu and Liu 2003]. This type of analysis can also guide companies in the pursuit of new features. It happens because each group of features is obtained by different means, e.g., Web features are obtained by web crawling, Demographic by the national census, Historical by other datasets, among others. Features group of interest may be those where features present good performance and low internal correlation.

When finding a correlation between two features, it is important to apply the right technique for each feature type. Table IV depicts the method used to find correlation among different feature categories, e.i. binary, multi-class, and continuous. This test was performed between two features at a time. Each feature can belong to only one category.

We chose the Spearman test for detecting correlation for continuous feature pair as shown in Table IV. It is a rank correlation test chosen because it does not require previous assumption of the features data distribution. We used the Jaccard Distance method for binary feature pair, since it presents

Table IV: Techniques used to find correlation for each feature category

Category Feature 1	Category Feature 2	Method	Range
Binary	Binary	Jaccard Distance	$[-1, 1]$
	Multi-class	XGBoost	$[-1, 1]$
	Continuous	XGBoost	$[-1, 1]$
Multi-class	Binary	XGBoost	$[-1, 1]$
	Multi-class	XGBoost	$[0, 1]$
	Continuous	XGBoost	$[0, 1]$
Continuous	Binary	XGBoost	$[-1, 1]$
	Multi-class	XGBoost	$[0, 1]$
	Continuous	Spearman	$[-1, 1]$

the percentage of times when the two binary values were equal, and it is easy to interpret. For the remaining pairs of features, we employed XGBoost model. This technique is suitable for this task based on the assumption that correlated features should also be a good predictor of one another. In addition, the XGBoost algorithm can identify non-linear patterns being superior to a Linear Regression for this purpose.

In order to apply XGBoost for correlation tests, we balanced the classes of the testing features. For features from Web group, this decreased the precision of the test in some extent, since we noticed huge class predominance for some features. For example, the feature *flag_web* has 99% of its values as “1” and only 1% as "“0” values. Finally, the results were normalized and put on the same scale to improve visualization. The Range column in Table IV presents the range of correlation values after normalization. For most cases, we were able to detect a negative correlation, i.e. $[-1, 1]$. However, for some feature pairs, such as between Multi-class and Continuous features, the algorithms and techniques used only allowed for identifying positive correlation, i.e. $[0, 1]$.

3.5 Models

In our experiment, we employed three machine learning models: XGBoost [Chen and Guestrin 2016], 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]. Next, MLP was used since it is a well-known technique for classification problems.

3.6 Evaluation Metrics

The binary classification of good and bad payers is not the credit companies’ main objective, 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][Huang et al. 2019], Area Under ROC Curve (AUC) [Fawcett 2005], and the Kolmogorov-Smirnov (KS) test [Neuhauser 2011]. 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.

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

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

3.7 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 assessed the resulting probability for each class that is provided by the machine learning libraries used. The ground truth "1" represents 100% good payer and "0" represents 100% bad payer. The first group of experiments 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. We also used XGBoost gain to rank the most important features. Finally, we employed binary Particle Swarm Optimization in conjunction with a machine learning algorithm to find most promising combined features, regardless of their group.

The second group of experiments uses well-known metrics and all three models to generate credit scores. 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. Our experiment used random subsampling to split the dataset into 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.

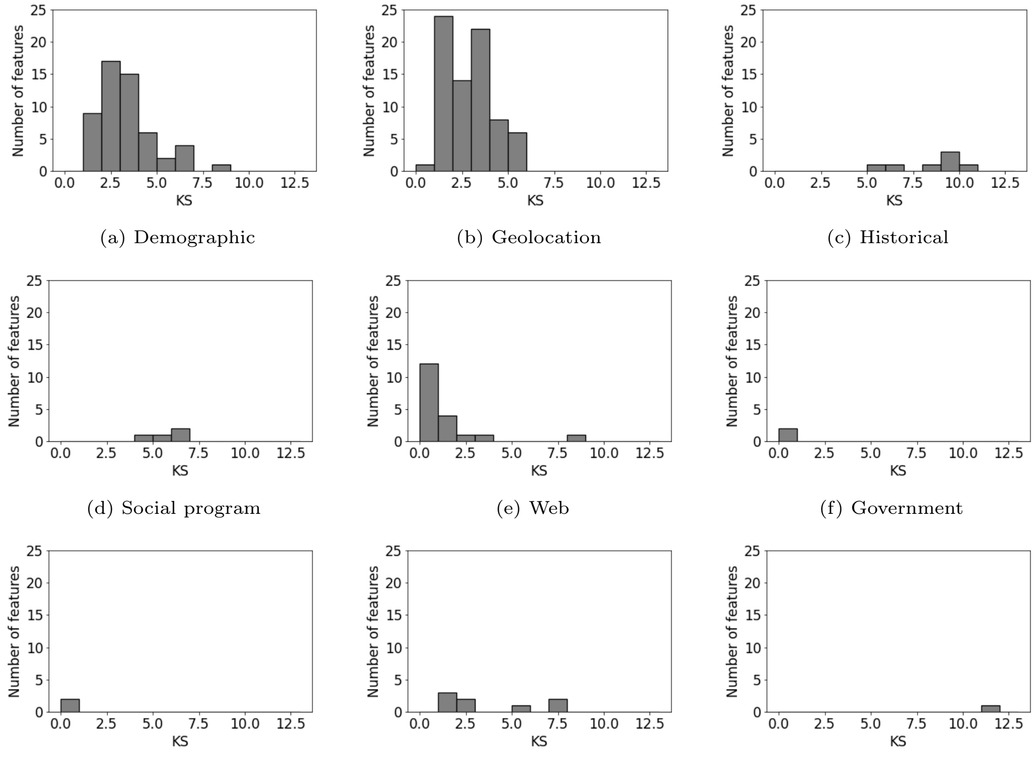
The MLP configuration presented the best results using two hidden layers with 400 neurons each, invscaling as the learning rate schedule, learning rate 0.2, adam as the optimizer, ReLU as activation function, and 400 iterations without loss improvement to interrupt the learning process. The best XGBoost 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 Feature Importance

4.1.1 *Individual Feature Importance.* We created a histogram (Fig. 1) for each group of features based on KS from a bivariate test. The histograms exhibits the importance of each feature in the dataset and also displays them inside their respective group. The feature importance is given by its KS in x-axis. The y-axis shows the number of features with same KS. From this analyzes, one can notice that Historical and Registration groups presented features with the best individual KS, whereas Politics and Government groups obtained the worst results. Also, 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.

We also analyzed feature importance by employing XGBoost gain as mentioned in Section 3.3. In this analysis, we obtained the 15 most relevant features from the XGBoost classification as shown in



(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.

Fig. 2. REG1 was the feature that individually added more gain. It is a feature from the registration group, and it represents the customer's income. The second most relevant feature was FIN6 that represents the financial institution (bank) where the customer received income tax refund, information obtained via web crawling. The third most significant feature was HIS3 that represents the frequency with which the customer's phone was found in others company datasets. An example of an unusual feature found in the top 15 is GEO73, which is related to the number of taxi stops nearby customer's home.

The fifteenth most important features ranking exhibited in Fig. 2 included features from all groups, except for Government (GOV). In addition, the groups that appeared most were Historical (HIS), Social Program (SOC), Registration (REG), and Geolocation (GEO). The XGBoost gain reinforced the KS test is shown in Fig. 1. In both tests, some of the most important groups are Registration (REG), Financial (FIN), Social Program (SOC) and Geolocation (GEO). However, the relevance given to Politics (POL) group by XGBoost gain is higher than the KS test. In fact, the feature POL1 ranked as the twelfth most important whereas in KS test it obtained one of the worst results. This may happen because XGBoost gain measures the importance of the features when associated with others present in the same tree, whereas KS bivariate test analyses the feature alone, directly with the ground truth.

4.1.2 *Feature Group Importance.* We employed XGBoost model with each feature group separately and used KS as metric. This allowed us to measure the importance of each group as a whole and check progressive gains by individually adding the groups. These results are shown in the bars of

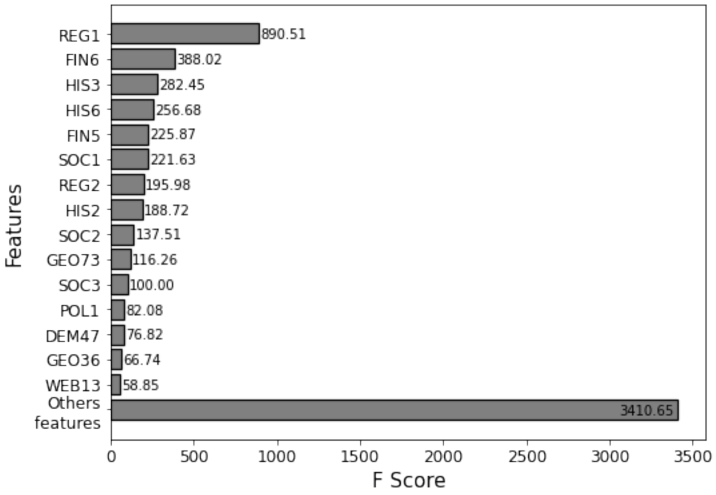


Fig. 2: Comparison between features using Gain metric from XGBoost.

Fig. 3, where each bar represents a group of features by its code, as depicted in Table II. In contrast, the increasing line presents the cumulative results of adding each group sequentially to the model.

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 cards in retail companies and requests for car insurance. The appearance of 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 might indicate a fraudster. Finally, these features have time-related information that indicates frequency and activity.

Another group with high performance is the Registration group (REG), even though it has only three features. This is the group that represents the traditional features in the dataset. 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 following 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 detail 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.

The line displayed in Fig. 3 shows the credit scoring performance by sequentially adding the features groups to the XGBoost model and performing the bivariate analysis afterwards. For example, the GEO bar line point 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 the models' performance since traditional datasets only present features from the REG group. We can also notice that the historical group's performance was

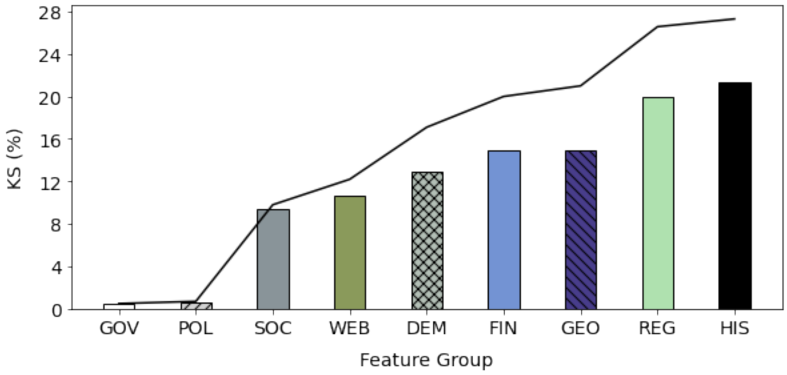


Fig. 3: KS obtained by each feature group (bars) and by the groups cumulatively (line).

higher than the REG Group, indicating that this is the most promising group.

4.1.3 *Feature Inter Group Importance.* We applied PSO in conjunction with each machine learning algorithm in order to find the most promising features regardless of their group. The dataset contains a total of 175 features from which MLP selected 88, AdaBoost 90 and XGBoost 99 features. In addition, the three algorithms selected 25 features in common.

Fig. 4 exhibits the percentage of feature by group selected by PSO in conjunction with each machine learning technique employed. It is possible to notice that all three settings selected at least 40% of features from each group. In addition, all PSO settings selected all features from Registration group (REG) and 60% or more of the Historical (HIS) group. Furthermore, the PSO + MLP gave preference to Web (WEB) group whereas PSO + ADA preferred Government (GOV) and Financial (FIN) groups. In its turn, PSO + XGB attributed more importance to GOV and Social Program (SOC) groups. The relatively low percentage of GEO and DEM groups may be due to the high inner correlations among features of these groups.

When analysing the average of features preferred by the three PSO settings showed as "AVERAGE", the most selected were REG, GOV and HIS. Finally, it is interesting to notice that GOV group was always selected what indicates that this feature is important when combined to others.

This analysis indicates that all features groups are relevant for the final results with REG and HIS groups standing out for all techniques.

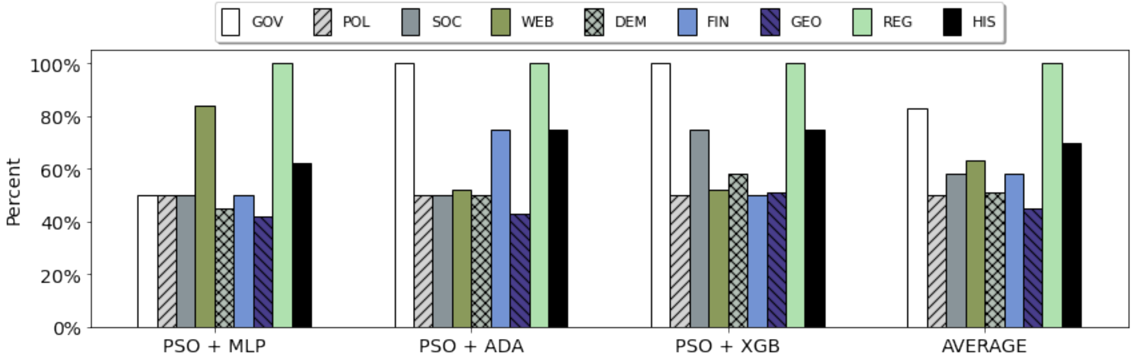


Fig. 4: Percentage of features by group selected by PSO and machine learning algorithms. We also show the percentage of features in common found by all three algorithms.

4.2 Features Correlation

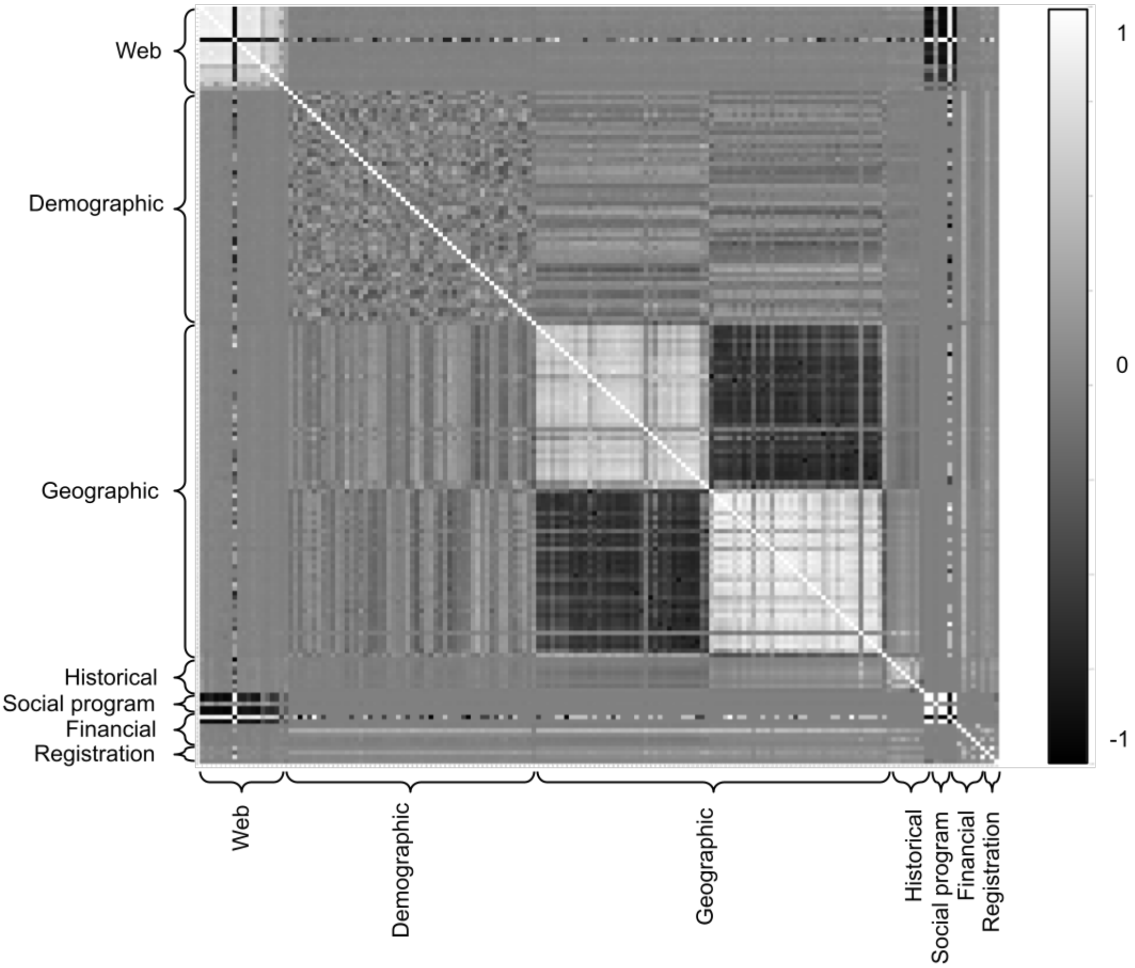


Fig. 5: Correlation of feature groups

Fig. 5 shows the results of our correlation tests through a heat map. We showed the results in the level of feature groups. However, each small square presents the normalized value found for each feature pair. Also, the heat map is symmetric over its white diagonal.

The inner correlation among the Web group features is high and can be identified as a lighter square in the top-left corner of Fig. 5. When investigating the reason, we notice that most of the Web features are binary, i.e., flags, and they have predominantly the zero class, which accounts for 96% of the cases. It makes these features significantly positively correlated to each other. An important exception for this rule is the feature *flag_web* in which 99% of the cases belong to class one. Therefore, it presents strong negative correlation with other web features and can be identified in the heat map as a black cross inside the light square. When comparing the Web group to Demographic, Geographic, and Historical groups, it is possible to notice little or no correlation except for the *flag_web* feature once more. This probably happened because this feature counts only with 16 negative cases making the balanced dataset too small for proper analysis. When comparing the Web group with Social Program and Financial groups, we notice a strong negative correlation. This is also because the binary features presented in those groups also have a largely predominant class, particularly the class with value one. Finally, the registration group did not present a significant correlation with the Web group.

The Demographic group presents some inner correlation, but it is hard to distinguish which features are more correlated with the others. In contrast, the correlation between Demographic and Geographic groups is given mainly by some specific features of the Demographic group exhibited as lighter lines. We also noticed a significant correlation with few financial features. When investigating further, we noticed that this feature is *neighborhood_ income* that presents the income of people from the same region what explains the correlation with Demographic and Geographic features.

The Geographic group presents a surprising correlation among its features that is perceptible through four squares. For this analysis, it is important to mention that the geographic features are split into two main subcategories being *least distance* features and *concentration* features. The *least distance* features capture the smallest distance from a specific place to the person's home, e.g., closest shopping mall. In contrast, the *concentration* presents the number of occurrences in the person's home surroundings, e.g., nearby shopping malls. Therefore, an inverse correlation between these two types of features is reasonable, since when the concentration of a place increases, the closest place's distance tends to decrease. This explains the two black squares that depicts negative correlation between the *concentration* and *smallest distance* subgroup features. On the other hand, the white squares present the correlation between the same subgroups. This behavior suggests that wherever it has a high concentration of a specific place, it also tends to have a high concentration of most other places, e.g., a place with many supermarkets also has many bars in its surroundings. Exceptions for that rule are Geographic features that do not map places but regions. As an example, the feature *smallest distance* to the sea. In that case, their distribution is linked to the country's geography.

The Historical group has a high inner correlation but a low correlation with other groups. This group tells how frequently customer data such as e-mail and telephone appear in other company datasets. Therefore, the results suggest that when one type of customer data appears in those datasets, other data tend to appear as well and vice-versa.

The Social Program and part of Financial groups present a very high correlation, i.e., mostly one. This happens because these group features present mostly flags that are once more very unbalanced, e.g., mostly class one. It does not mean that they tell the same information, but they add little to the other. They add some information because they have a different number of null values. Moreover, their non-null values also may belong to different records.

Finally, in the registration group, the feature *province*, which tells the region in the country where the person lives, has some correlation with Demographic and Geographic groups, which is also reasonable.

In this analysis of the correlation of the features, it was possible to notice some interesting behavior. Firstly, the Web and Social Program features mainly have a heavily predominant zero class, except for *flag_ web* that has a predominant one class. Next, we notice that the Demographic group has some inner correlation and inter-correlation with the Geographic group. We also noticed that inside Geographic groups, two subcategories being *smallest distance* and *concentration* that have a strong negative correlation.

4.3 Comparison of Machine Learning Techniques

We analyzed the performance of different models with well-known metrics described in the literature. Table V depicts the results of the models for 30 simulations with their respective standard deviation. We can notice that XGBoost obtained the best results for all metrics used. 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. The best performance in the MSLE metric indicates that XGBoost presents smaller errors than the other solutions. For metrics Lift, AUC, and KS, AdaBoost obtained the second-best results, followed by MLP. However, for metric MSLE, MLP obtained the second best results. The results suggests a superiority of tree-based techniques over Multilayer-perceptron with XGBoost standing out as the best technique.

Table V: Models performance on the used metrics.

Models	**Lift**	**MSLE**	**AUC (%)**	**KS (%)**
MLP	0.5995 ± 0.004	0.1133 ± 0.001	66.04 ± 0.17	23.55 ± 0.3
AdaBoost	0.6312 ± 0.003	0.1242 ± 0.0001	68.08 ± 0.14	26.30 ± 0.27
XGBoost	**0.6496 ± 0.003**	**0.11 ± 0.0001**	**68.81 ± 0.14**	**27.24 ± 0.28**

5. CONCLUSION

The credit scoring problem is vital to financial companies because it is directly related to profit increase. In this work, we analyzed the importance of diverse feature groups for generating credit score using a real dataset. Some of these groups presented unusual features when compared to other datasets from the literature. Among those features, there are groups related to geolocation, historical, web behavior, and demographic data. We used KS test to access the performance of the features individually. To collectively check feature performance, we employed XGBoost gain and feature selection with binary PSO. All the tests pointed Registration and Historical as most promising groups followed by Financial and Social Program groups. In addition, XGBoost gain and binary PSO revealed that some groups that seem to not be relevant, are in fact important when combined with others, as in the case of Government group and some Geolocation features. Therefore, it was possible to conclude that all used groups were relevant to improve the results. In addition, the XGBoost test with isolated feature groups showed that the performance of the standard dataset, represented by the registration group alone, could be improved with the addition of the unusual groups specially with Historical features. We detected high inner correlation in the groups Web and Geolocation and moderate inner correlation in the Demographic group. This suggests that the removal of some of the features in those groups may not affect substantialy the results as also suggested by the feature selection with binary PSO, where only about 40% of features from groups Geolocation and Demographic were selected. Finally, we noticed that XGBoost obtained the best performance over MLP and AdaBoost for all metrics used. A limitation of our work is that we did not have access to the company's full dataset, which contains more registration features, and would help to contrast even further the results obtained by the addition of the novel groups. Nonetheless, the results indicate that these new feature groups are promising and deserve further attention and effort.

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. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '16. ACM, New York, NY, USA, pp. 785–794, 2016.

Clerc, M. and Kennedy, J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE transactions on Evolutionary Computation* 6 (1): 58–73, 2002.

Djeundje, V. B., Crook, J., Calabrese, R., and Hamid, M. Enhancing credit scoring with alternative data. *Expert Systems with Applications* vol. 163, pp. 113766, 2021.

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.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

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.

Huang, Z., Wang, Z., and Zhang, R. Cascade2vec: Learning dynamic cascade representation by recurrent graph neural networks. *IEEE Access* vol. 7, pp. 144800–144812, 2019.

Kennedy, J. and Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks.* Vol. 4. IEEE, pp. 1942–1948, 1995.

Kennedy, J. and Eberhart, R. C. A discrete binary version of the particle swarm algorithm. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation.* Vol. 5. pp. 4104–4108 vol.5, 1997.

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. *Geophysical Research Abstracts* vol. 14, pp. 12354, 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.

Paraíso, P., Ruiz, S., Gomes, P., Rodrigues, L., and Gama, J. Using network features for credit scoring in microfinance. *International Journal of Data Science and Analytics*, 2021.

Shi, X., Wong, Y. D., Li, M. Z.-F., Palanisamy, C., and Chai, C. A feature learning approach based on xgboost for driving assessment and risk prediction. *Accident Analysis and Prevention* vol. 129, pp. 170–179, 2019.

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

Verçosa, L. F., Lira, R., Monteiro, R., Silva, K., Magalhaes, J., Maciel, A., Bezerra, B., and Bastos-Filho, C. Impact of unusual features in credit scoring problem. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning.* SBC, Porto Alegre, RS, Brasil, pp. 81–88, 2020.

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.

Yu, L. and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03).* pp. 856–863, 2003.

Zheng, H., Yuan, J., and Chen, L. Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies* 10 (8): 1–20, August, 2017.

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.