From Payment Services Directive 2 (PSD2) to Credit Scoring: A Case Study on an Italian Banking Institution

Roberto Saia, Alessandro Giuliani, Livio Pompianu and Salvatore Carta
Department of Mathematics and Computer Science,
University of Cagliari, Via Ospedale 72, 09124 Cagliari, Italy

Keywords: Business Intelligence, Decision Support System, Machine Learning, Credit Scoring, PSD2.

Abstract: The Payments Systems Directive 2 (PSD2), recently issued by the European Union, allows the banks to share their customer data if they authorize the operation. On the one hand, this opportunity offers interesting perspectives to the financial operators, allowing them to evaluate the customers reliability (Credit Scoring) even in the absence of the canonical information typically used (e.g., age, current job, total incomes, or previous loans). On the other hand, the state-of-the-art approaches and strategies still train their Credit Scoring models using the canonical information. This scenario is further worsened by the scarcity of proper datasets needed for research purposes and the class imbalance between the reliable and unreliable cases, which biases the reliability of the classification models trained using this information. The proposed work is aimed at experimentally investigating the possibility of defining a Credit Scoring model based on the bank transactions of a customer, instead of using the canonical information, comparing the performance of the two models (canonical and transaction-based), and proposing an approach to improve the performance of the transactions-based model. The obtained results show the feasibility of a Credit Scoring model based only on banking transactions, and the possibility of improving its performance by introducing simple meta-features.

1 INTRODUCTION

Nowadays, risk management has become a key factor in financial business scenarios. An appropriate credit risk management is crucial for supporting financial institutions that provide lending services, which may be affected by substantial economic losses due to loan defaults. Estimating the probability of default is a common way to assess the risk borrowers cannot repay their loans. In such a context, to support financial businesses in facing the competitive marketplace, defining a reliable and effective *Credit Scoring* model is essential to predict and avoid the aforementioned issue. Credit Scoring can be defined as the set of models and statistical methods aimed at automatically evaluating consumer credit, estimating the likelihood of default (Thomas et al., 2017; Gup, 2005).

Recently, Credit Scoring systems have rapidly grown due to the increase of both consumer credit requests and financial operators who flank the traditional credit circuits (i.e., banks) (Siddiqi, 2017). Indeed, the massive number of requests does not allow processing them manually, requiring effective automated systems to establish the reliability of a given user, according to a binary (by classifying the user as reliable or unreliable) or continuous (by assigning a solvency score to the user) criterion.

In such a context, many people can not access consumer credit because of the strict conditions required by financial institutions, e.g., in many countries, banks ask customers to provide formal documentation proving they have a permanent employment contract. Clients who do not meet these criteria can not access consumer credit, even if they have other financial incomes, allowing them to repay a loan. For these reasons, there has also been a growth of investments and efforts in Credit Scoring research, involving an ever-increasing number of researchers. The focus is on defining strategies and algorithms capable of correctly evaluating new cases (users) basing on the data previously collected.

A common way to define Credit Scoring models is to rely on Machine Learning techniques to quantitatively assess the default risk, typically basing on personal information (e.g., age, job, income, or outstanding debt debts) obtained from loan applicants (Lee and Sohn, 2017; Kim and Sohn, 2012). In details, the aim is to develop systems able to classify a user as *reliable* or *unreliable* by exploiting the available information, which from now on we define *in-*

Saia, R., Giuliani, A., Pompianu, L. and Carta, S.
From Payment Services Directive 2 (PSD2) to Credit Scoring: A Case Study on an Italian Banking Institution.
DOI: 10.5220/0010653200003064
In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021) - Volume 1: KDIR*, pages 164-171
ISBN: 978-989-758-533-3; ISSN: 2184-3228

stances.

Research on Credit Scoring has been heavily conditioned by the scarcity of publicly available datasets, due to the privacy policies of most financial companies. A key-point of state-of-the-art approaches trained with the canonical datasets is represented by the high level of class imbalance, as, typically, the unreliable cases (e.g., users that did not repay, fully or partially, a loan) occur less frequently than the reliable ones (e.g., users that repaid a loan). The data imbalance, which represents a very common scenario in several domains, e.g., in *Fraud Detection* (Carta et al., 2019b; Saia et al., 2017; Saia and Carta, 2017a) or in *Intrusion Detection* (Saia et al., 2018b; Saia et al., 2019b), is typically addressed by adopting several re-sampling strategies (Leevy et al., 2018; Junsomboon and Phienthrakul, 2017). It should be observed that such balancing techniques often cannot effectively face some types of problems, since the generation of synthetic unreliable instances based on the existing ones does not face the heterogeneity problem introduced before. Indeed, if there were similar instances related to both information classes (reliable and unreliable), they will continue to exist, even in more significant numbers, despite the data balancing. It means that very similar feature patterns can characterize both categories of users, reliable and unreliable. Consequently, the traditional information set may need to be enlarged.

To address the aforementioned issue, let us consider an emerging perspective. Currently, the Credit Scoring environment is strongly affected by the Payment Services Directive 2 (PSD2), i.e., the new regulatory framework, introduced by the European Commission, to regulate payment services and providers throughout the European Union (EU) and European Economic Area (EEA). The PSD2 aims to provide a more integrated European payments market, in which all players may be either banks or non-banks while ensuring a more secure and protected platform for consumers. In particular, the new directive encourages customers to exploit innovative online and mobile payments, such as through *Open Banking*, as the current rules better protect customers in online payments and make cross-border European payment services safer. In doing so, the PSD2 regulation allows third parties to obtain free access to client accounts and their payment transactions through bank APIs. This new regulation can be exploited in real-world scenarios, as models based also on bank transaction information may offer interesting employment prospects. However, there are still no clear policies on which technologies should be used and, in particular, which types of data the banks must share. To this end, there is the need to investigate the explanatory value of transaction data to estimate the potential usefulness in a Credit Scoring scenario.

This paper aims to exploit the transaction information to improve the characterization of the user instances for classic Credit Scoring models. Furthermore, basing on a state-of-the-art study, we also expand the feature space by adding a series of meta-information. To our knowledge, no peer-review publications exist on using transaction data to define Credit Scoring models. Our scientific contribution can be recapped as follows:

- the analysis of a real-world dataset provided by an Italian bank institution, built with client accounts information and their payment transactions data, compliant to the new PSD2 regulation;
- assessment of the usefulness of transaction data for Credit Scoring models;
- enrichment of the feature space with the introduction of a series of meta-features;
- evaluation of the related improvements for the Credit Scoring models.

This paper has been structured into the following sections: Section 2 presents the background and the related work of domain taken into account in this paper; Section 3 formalizes the notation adopted in this paper, providing also information about the exploitation of the additional meta-features; Section 4 describes the performed experiments in terms of environment, datasets, strategy, and adopted metrics, discussing the experimental results; remarks and future work where we are headed are given in Section 5, which also ends the paper.

2 BACKGROUND AND RELATED WORK

The literature identifies three different risk models based on the default condition (i.e., the failure to respect the legal obligations/conditions related to a financial service, such as a loan): (i) Probability of Default (PD), a model aimed to assess the likelihood of a default over a certain period; (ii) Exposure At Default (EAD), a model aimed to assess the total value a financial operator is exposed in case of default; (iii) Loss Given Default (LGD), a model aimed to assess the amount of money a financial operator loses in case of default.

For the purpose related to this paper, we will consider the PD model, since our objective is a dichotomous classification of the new instances into two classes, reliable or unreliable.

Approaches: The literature offers a considerable number of techniques and strategies, such as:

- *Statistical*: in (Sohn et al., 2016) the authors exploit Logistic Regression (LR) to design a fuzzy Credit Scoring model aimed at assessing the default probability of a loan. In (Khemais et al., 2016) the authors use the Linear Discriminant Analysis (LDA) in order to achieve this result. A recent study (Roy and Shaw, 2021) proposes a model based on multiple-criteria decision-making (MCDM) that can be adopted as an internal scoring model to preliminary screen the loan applications, and it can be initially applied to reduce costs;
- *Transformed Domain*: in (Saia and Carta, 2017b) has been proposed a Credit Scoring approach based on the Fourier transform, similarly to the work done in (Saia et al., 2018a), where instead the Wavelet transform has been exploited;
- *Machine Learning (ML)*: in (Roy and Urolagin, 2019) a Credit Scoring approach based on Decision Tree (DT) and Support Vector Machine (SVM) algorithms has been defined, whereas in (Zhang et al., 2018) a Random Forest (RF) algorithm has been adopted. Another work is based on a survival gradient boosting decision tree (GBDT) approach (Xia et al., 2020);
- *Deep Learning (DL)*: recent works focus on exploiting DL models also in the field of Credit Scoring. In (Liu et al., 2019) the authors exploit an Artificial Neural Network (ANN) to perform the Credit Scoring task, as well as in (Lei et al., 2019), where has been defined an Imbalanced Generative Adversarial Fusion Network (IGAFN) based on a feed-forward neural network (FNN) and a Bidirectional Long Short-Term Memory (Bi-LSTM) network. An approach based on Generative Adversarial Networks (GAN) for data oversampling is proposed in (Engelmann and Lessmann, 2021);
- *Others*: approaches that focus on other factors. For example, the entropy factor has been taken into account in (Saia and Carta, 2016a), the linear-dependence of the involved data has been considered in (Saia and Carta, 2016c; Saia and Carta, 2016b), a discretized enriched technique has been considered in (Saia et al., 2019b), whereas some hybrid approaches that combine different methods have been proposed in (Tripathi et al., 2018; Zhang et al., 2019).

Let us remark that, in this preliminary work, we focus on developing a model aimed at highlighting the usefulness of transaction data rather than comparing the model with all state-of-the-art systems. In doing so, we compare our model with the most-known classical ML models.

Open Problems: Although there are numerous state-of-the-art approaches for credit scoring, they have to face some well-known problems, such as:

- *Data Scarcity*: the scarcity of real-world datasets has affected the research activity in the Credit Scoring domain. Otherwise, the development of this research area would undoubtedly have been more consistent and effective, and it is curious to observe how this kind of problem can be considered a side effect, given by the security and privacy policies that regulate many public and private companies (Sloan and Warner, 2018).
- *Data Imbalance*: the prediction models used in the Credit Scoring domain are commonly defined on the basis of datasets with a high degree of data imbalance, i.e., data characterized by unbalanced distributions of the events of interest (unreliable cases), which are significantly fewer than the other ones (reliable cases). Class imbalance is the most critical problem to face in developing Credit Scoring solutions, since a model trained by using unbalanced data underestimates the probability of rare events, tending to be biased towards the most common events (King and Zeng, 2001), reducing the performance dramatically. The literature offers several methods to balance the data, mainly according to one of the following strategies: introducing synthetic instances (oversampling); removing existing instances (undersampling); combining both the oversampling and undersampling strategies. In more detail, in the Credit Scoring context, the oversampling creates synthetic unreliable instances based on the existing ones, whereas the undersampling removes several reliable existing instances to balance their number with respect to the unreliable ones. Both oversampling and undersampling have their shortcomings. The former can lead to overfitting, as duplicating “bad” records may underestimate the likelihood of observations belonging to the minority class, whereas the latter may discard relevant cases from the majority class, overestimating the probability of “bad” samples (Weiss, 2004). An empirical study on the balancing techniques applied to Credit Scoring models highlighted that, albeit larger datasets require longer training times, oversampling significantly increases the accuracy relative to undersampling (Crone and Finlay, 2012). For this reason, in this paper, we adopted this technique to balance the data.
- *Cold Start*: the cold start problem concerns the evaluation model definition when such a process can not use samples of one of the classes of information involved, and it is a problem shared by many domains. In the Credit Scoring one, this typically happens when there are no unreliable cases avail-

able, and then the evaluation model training can not be performed by using only the reliable ones.

Evaluation Metrics: As already pointed out, we fall in a binary classification domain, in which the system should predict if a given user will be reliable or unreliable. In such a context, we are focused on confusion-matrix based metrics, such as, for instance, *accuracy*, *sensitivity*, *specificity*, and *Matthews correlation coefficient (MCC)*, which are largely used in literature in several domains. All the reported metrics are based on the *confusion-matrix*, i.e., a matrix of size $2x2$ that reports the numbers of True Negatives (TN), False Negatives (FN), True Positives (TP), and False Positives (FP). In the Credit Scoring literature, the confusion-matrix-based metrics are usually combined with other metrics based on the Receiver Operating Characteristic (ROC) curve (Green and Swets, 1966). One of the metrics primarily used is the *Area Under the ROC Curve (AUROC)*: the ROC curve plots the confusion-matrix-based *Sensitivity* against the confusion-matrix-based *Fallout*, respectively on the y-axis and the x-axis, giving us the separability measure of a binary classifier (i.e., the capability to discriminate the reliable and unreliable cases, correctly).

3 PROPOSED APPROACH

Before formalizing the proposed approach, we introduce the adopted formal notation: given a set of classified instances $I = \{i_1, i_2, \ldots, i_X\}$, composed by a subset of *reliable* ones $I^+ = \{i_1^+, i_2^+, \ldots, i_Y^+\}$ (then $I^+ \subseteq I$), and a subset of *unreliable* ones $I^- = \{i_1^-, i_2^-, \ldots, i_W^-\}$ (then $I^- \subseteq I$), we denotes as $\hat{I} = \{\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_Z\}$ another set of unclassified instances, each instance being characterized by a series of features $F = \{f_1, f_2, \ldots, f_N\}$, and a destination class $C = \{reliable, unreliable\}$.

Data Aggregation: In order to train the classification models of the involved algorithms, as the first step, we aggregate the transaction information of each user in a single data vector (record) on the basis of several criteria. In more detail, the adopted *Data Aggregation Approach* (DAA) aggregates the transaction of each user in terms of: *number of transactions*, *activity days*, *minimum handled amount*, *maximum handled amount*, *total handled amount*, *mean handled amount*, and *standard deviation* measured in all the transactions.

Meta-feature Addition: Similarly to other previous works (Saia et al., 2019b; Carta et al., 2020; Carta et al., 2019a; Saia et al., 2019a), which exploit preprocessing techniques in order to improve the performance of the machine learning algorithms, we propose a *Meta-features Addition Approach* (MAA) aimed to better characterize each instance in the sets I and $\hat{I}$, improving the credit scoring performance. In more detail, we added a series of meta-features $MF = \{mf_1, mf_2, \ldots, mf_4\}$, where $mf_1 = minimum$, $mf_2 = maximum$, $mf_3 = average$, and $mf_4 = standard\ \ deviation$, all of them calculated in the set of features F related to each instance, as formalized in Equation 1, then such a process involves both the set I and the set $\hat{I}$.

$$MF = \begin{cases} mf_1 = min(f_1, f_2, \ldots, f_N) \\ mf_2 = max(f_1, f_2, \ldots, f_N) \\ mf_3 = \frac{1}{N}\sum_{n=1}^{N}(f_n) \\ mf_4 = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(f_n - \bar{f})^2} \end{cases} \tag{1}$$

Instances Classification: Each instance $\hat{i} \in \hat{I}$ will be classified as *reliable* or *unreliable*, in accordance with the criteria formalized in the Algorithm 1. In order to simplify, with regard to each user, we have not used a different notation for the multiple instances (i.e., the related bank transactions) and those aggregated into a single vector through the DAA process.

Algorithm 1: Instance classification.

```
Require: $A$=Classifier, $I$=Set of classified instances, $\hat{i}$=Instance to evaluate
Ensure: $c$=Classification of the $\hat{i}$ instance
procedure GETEVALUATION($A$, $I$, $\hat{i}$)
    $I \leftarrow aggregateData(I)$                          ▷ DAA process on $I$ set
    $\hat{i} \leftarrow aggregateData(\hat{i})$                          ▷ DAA process on $\hat{i}$ instance
    $I \leftarrow addMetafeatures(I)$                        ▷ MAA process on $I$ set
    $\hat{i} \leftarrow addMetafeatures(\hat{i})$                        ▷ MAA process on $\hat{i}$ instance
    $model = trainModel(A, I)$                         ▷ Classification model training
    $c = getPrediction(model, \hat{i})$                      ▷ Instance classification
    return $c$
end procedure
```

4 EXPERIMENTS

All the involved code has been developed in the *Python* language, exploiting the *scikit-learn* [1] library. In order to ensure the reproducibility of the experiments carried out, the seed of the *pseudo-random number generator* has been fixed to *1*. We also performed independent-samples two-tailed Student's t-tests, highlighting no statistical difference between the results ($p > 0.05$).

Datasets: The experiments have been performed using real-world datasets provided by a huge Italian

[1] http://scikit-learn.org

bank. In this regard, it should be noted that the data provided by the bank allow us to compare the performance of the canonical credit scoring models (i.e., those based on the information usually exploited in literature) with those based only on bank transactions, as both the datasets used during the experiments refer to the same set of users. Each user has been labeled as *reliable* or *unreliable*. In details, two types of datasets have been provided by the bank, one containing transactions made by a set of bank clients and one containing, for the same clients, the canonical information used in the literature for the Credit Scoring tasks (e.g., age, gender, job, incomes, or loans). Similar to other works that use real data from private companies, due to confidentiality reasons, the dataset can not be made public by us, not even in an anonymous form, except for its characteristics, which should allow the reproducibility of the experiments on other similar datasets.

The first dataset, named Credit Card Transactions (CCT), contains $3,376,573$ transactions related to the use of a revolving credit card, carried out by $49,847$ bank customers in the period from $01/01/2017$ to $31/12/2018$. The CCT dataset has been preprocessed according to the criteria described in Section 3, which leads toward the features reported in Table 1.

In addition, concerning the same users in the CCT dataset, we use another dataset, named Bank Customers Information (BCI), that contains the Credit Scoring information usually exploited in the literature. It refers to the same $49,847$ customers of the CCT dataset, and its features are reported in Table 2.

Each dataset contains $49,459$ reliable cases (99.23%) and 388 unreliable ones (0.77%). Therefore, they are characterized by a high level of data imbalance, similarly to the other datasets typically available in the Credit Scoring field. In order to evaluate the performance as correctly as possible, avoiding the influence of the minority class in defining the evaluation model, both datasets have been preprocessed using an oversampling technique, the *Adaptive Synthetic Sampling Approach* (ADASYN) (He et al., 2008), keeping its default parameters (i.e., *sampling_strategy='auto'*, *n_neighbors=5*, *n_jobs=None*).

Table 1: CCT Dataset Features.

Feature	Description	Type
F01	Unique identifier of the user	Integer
F02	Number of user transactions	Integer
F03	Days of user activity	Integer
F04	Minimum amount handled by the user	Real
F05	Maximum amount handled by the user	Real
F06	Total amount handled by the user	Real
F07	Average amount handled by the user	Real
F08	Standard deviation of the user transactions	Real

Table 2: BCI Dataset Features.

Feature	Description	Type	Feature	Description	Type
F01	user from	Date	**F08**	Employed from	Date
F02	Family members	Integer	**F09**	Real estates	Real
F03	Family income	Real	**F10**	Annual income	Real
F04	Resident from	Date	**F11**	Other incomes	Real
F05	House type	String	**F12**	Loans amount	Real
F06	Job type	String	**F13**	Mortgages amount	Real
F07	Job sector	String	**F14**	Rent amount	Real

Metrics: In order to evaluate the performance, we rely on three different metrics. Two of them, *Sensitivity* and *Specificity*, are both based on the *confusion matrix*, indicating, respectively, the *true positive rate* and the *true negative rate*. They estimate the capability to classify the *reliable* and *unreliable* instances, correctly. The other one, the *Area Under the Receiver Operating Characteristic* curve (AUC), is instead derived from the *Receiver Operating Characteristic* (ROC) curve, and provides us information about the predictive performance of an evaluation model, regardless of the balancing of classes (reliable and unreliable) in the dataset (number of examples available for the two classes).

The Equation 2 formalize the Sensitivity and Specificity metrics, where *TP* indicates the instances classified as *reliable* correctly, *TN* indicates the instances classified as *unreliable* correctly. *FN* and *FP* indicate, respectively, those *unreliable* wrongly classified as *reliable*, and those *reliable* wrongly classified as *unreliable*. These metrics give us the measure of how many instances have been correctly classified by an evaluation model.

$$Sensitivity = \frac{TN}{(TN+FP)},\ Specificity = \frac{TP}{(TP+FN)} \quad (2)$$

The *Area Under the Receiver Operating Characteristic* curve (AUC) metric is largely used in the Credit Scoring literature, since it allows us to assess the capabilities of an evaluation model, regardless of the data balancing. Formally, considering the subsets of *reliable* (I_+) and *unreliable* (I_-) instances in the set I, Equation 3 formalizes all possible comparisons α of the scores of each instance i, and the AUC value (in the range $[0,1]$, where 1 indicates the best performance) is given by averaging over them.

$$\alpha(i_+,i_-) = \begin{cases} 1, & if\ i_+ > i_- \\ 0.5, & if\ i_+ = i_- \\ 0, & if\ i_+ < i_- \end{cases} \quad (3)$$

$$AUC = \frac{1}{I_+ \cdot I_-} \sum_{1}^{|I_+|} \sum_{1}^{|I_-|} \alpha(i_+,i_-)$$

Strategy: The experiments involve five of the most performing algorithms in Credit Scoring literature, i.e., *Gradient Boosting* (GB) (Chopra and Bhilare, 2018), *AdaBoost* (AB) (Freund and Schapire, 1999), *Random Forests* (RF) (Malekipirbazari and Aksakalli, 2015), *Multilayer Perceptron* (MP) (Luo et al., 2017), and *Decision Tree* (DT) (Damrongsakmethee and Neagoe, 2019). Table 3 reports their parameters.

Table 3: Algorithms Parameters.

Algorithm	Parameter	Value
Gradient Boosting (*GB*)	*n_estimators* *learning_rate* *max_depth*	100 0.1 3
AdaBoost (*AB*)	*n_estimators* *learning_rate* *algorithm*	50 0.1 SAMME.R
Random Forests (*RF*)	*n_estimators* *max_depth* *min_samples_split*	10 none 2
Multilayer Perceptron (*MP*)	*alpha* *max_iter* *solver*	0.0001 200 adam
Decision Tree (*DT*)	*min_samples_split* *max_depth* *min_samples_leaf*	2 none 1

The performance comparison has been made by taking into account the average value of the three considered metrics (i.e., *Sensitivity*, *Specificity*, and *AUC*). In addition, in order to reduce the impact of the data dependency, all the experiments have been performed according to a *k-fold cross-validation* criterion, with *k=10*.

The strategy adopted for the experiments follows four steps: (i) evaluation of a model trained using the canonical users information (BCI dataset); (ii) evaluation of a model trained using only the bank transactions (CCT dataset); (iii) comparison of the performance of the two aforementioned evaluation models, both applied to the same set of users; (iv) assessing the benefits of the proposed preprocessing approach (meta-features enrichment) in the context of the model trained only with the bank transactions.

Results: According to our experimental strategy, the first set of experiments has been aimed at evaluating the state-of-the-art algorithms in the context of an evaluation model trained using the canonical Credit Scoring information, then without using the bank transactions (BCI dataset). The results are reported in Table 4. In Table 5 are instead reported the performance related to the same users of the BCI dataset, but evaluated on the basis of a model trained using the CCT dataset. It should be observed how, apart from some algorithms (i.e., AB and MP), all the other ones reach good performances, especially DT and GB, in accordance with many literature works in this domain. The average values reported in Table 5, related to the model trained using the credit card transactions, have been compared in Table 6 to the results obtained after adding the meta-features (the best performance are highlighted in bold). Furthermore, to better highlight the impact of meta-features, the last column in the Table reports the increment, expressed in percentage, of the average values.

Table 4: Canonical Credit Scoring Model Performance.

Algorithm	Dataset	Sensitivity	Specificity	AUC	Average
GB	BCI	0.9320	0.9295	0.9307	0.9307
AB	BCI	0.8933	0.8815	0.8873	0.8874
RF	BCI	0.9560	0.9992	0.9776	0.9776
MP	BCI	0.4836	0.7899	0.6367	0.6367
DT	BCI	0.9613	0.9712	0.9715	0.9715

Table 5: Transactions-based Credit Scoring Model Performance.

Algorithm	Dataset	Sensitivity	Specificity	AUC	Average
GB	CCT	0.7920	0.8477	0.8173	0.8190
AB	CCT	0.6907	0.7123	0.7008	0.7013
RF	CCT	0.9803	0.8253	0.8883	0.8980
MP	CCT	0.6311	0.6157	0.6231	0.6233
DT	CCT	0.9703	0.9054	0.9356	0.9371

Table 6: Transactions-based Credit Scoring Model Performance, Before and After the Meta-features Addition.

Algorithm	Dataset	Average before	Average after	Increment (%)
GB	CCT	0.8190	**0.8208**	+0.22%
AB	CCT	0.7013	**0.7056**	+0.62%
RF	CCT	0.8980	**0.9118**	+1.54%
MP	CCT	0.6233	**0.6354**	+1.94%
DT	CCT	0.9371	**0.9397**	+0.28%

Discussion: The experimental results lead toward the following considerations:

- the adoption of the canonical customers information for the model training leads toward better performances of those obtained using only the bank transactions. In any case, to face those scenarios where the canonical information is not available, the latter evaluation model offers an interesting opportunity, as it allows the financial operators to extend the set of potential customers, with all the related advantages;
- according to the previous observation, we experimented how the addition of simple meta-features can improve the Credit Scoring performance of all the classification algorithms taken into account during the experiments;
- it is therefore clear that both models can be exploited in real-world scenarios, and that those based on bank transactions offer interesting employment prospects, albeit presenting lower performances than those obtained by canonical models;
- although the improvements introduced by the addition of the meta-features appear slight, they can be considered an interesting result, given their impact in real-world scenarios, where are usually involved

a huge number of customers;
- the previous observation is supported by the fact that in the context of the used dataset (*49,847* bank customers), the improvements related to the *GB*, *AB*, *RF*, *MP*, and *DT* algorithms lead toward, respectively, *60*, *169*, *837*, *613*, and *135* further correctly classified customers;
- summarizing, the results demonstrate both the feasibility of a model based only on banking transactions, and the possibility of improving its performance by introducing simple meta-features.

5 CONCLUSIONS AND FUTURE WORK

The recent Payments Systems Directive 2 (PSD2) issued by the European Union, which enables the banks to share the user's data with their prior consent, makes it essential to revise the canonical methodologies for defining Credit Scoring models. Indeed, the state-of-the-art Credit Scoring models are usually trained using different information about the customers, mainly based on personal data such as, for instance, age, gender, current job, total incomes, or previous loans. To this end, we investigated the feasibility of defining a Credit Scoring model based on bank transactions of customers instead of using the canonical information. Transaction data has also been enhanced through the introduction of suitable meta-features. The performed experiments indicate a performance reduction when the Credit Scoring models have been trained using the bank transactions only, as reported in Table 4. However, as shown in Table 5, they indicate that it is possible to get acceptable performance also by using the bank transactions, grouped according to simple criteria, and that these performances can be improved by adding meta-features (Table 6). These results open up interesting scenarios, since they allow the financial operators to extend their potential customers in many contexts such as, for instance, in the consumer credit one, by virtue of the fact that it is possible to evaluate them even in the absence of the canonical information used up to now.

As future work, we plan to extend the proposed approach, testing more sophisticated methodologies able to further improve the Credit Scoring performance, according to the available user's data.

ACKNOWLEDGEMENTS

This research was partially funded by the "Bando 'Aiuti per progetti di Ricerca e Sviluppo'– POR FESR 2014-2020 – Asse 1, Azione 1.1.3. Project Intelli-Credit: AI-powered digital lending platform".

REFERENCES

Carta, S., Fenu, G., Ferreira, A., Recupero, D. R., and Saia, R. (2019a). A two-step feature space transforming method to improve credit scoring performance. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 134–157. Springer.

Carta, S., Fenu, G., Recupero, D. R., and Saia, R. (2019b). Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. *Journal of Information Security and Applications*, 46:13–22.

Carta, S., Podda, A. S., Reforgiato Recupero, D. R., and Saia, R. (2020). A local feature engineering strategy to improve network anomaly detection. *Future Internet*, 12(10):177.

Chopra, A. and Bhilare, P. (2018). Application of ensemble models in credit scoring models. *Business Perspectives and Research*, 6(2):129–141.

Crone, S. F. and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1):224–238.

Damrongsakmethee, T. and Neagoe, V.-E. (2019). Principal component analysis and relieff cascaded with decision tree for credit scoring. In *Computer Science On-line Conference*, pages 85–95. Springer.

Engelmann, J. and Lessmann, S. (2021). Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *Expert Systems with Applications*, 174.

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Gup, B. E. (2005). *Commercial banking : the management of risk*. J. Wiley.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE.

Junsomboon, N. and Phienthrakul, T. (2017). Combining over-sampling and under-sampling techniques for imbalance dataset. In *Proceedings of the 9th International Conference on Machine Learning and Computing*, pages 243–247.

Khemais, Z., Nesrine, D., Mohamed, M., et al. (2016). Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *International Journal of Economics and Finance*, 8(4):39.

Kim, Y. and Sohn, S. (2012). Stock fraud detection using peer group analysis. *Expert Systems with Applications*, 39(10):8986–8992.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 9(2):137–163.

Lee, B. K. and Sohn, S. Y. (2017). A credit scoring model for smes based on accounting ethics. *Sustainability*, 9(9).

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1):42.

Lei, K., Xie, Y., Zhong, S., Dai, J., Yang, M., and Shen, Y. (2019). Generative adversarial fusion network for class imbalance credit scoring. *Neural Computing and Applications*, pages 1–12.

Liu, C., Huang, H., and Lu, S. (2019). Research on personal credit scoring model based on artificial intelligence. In *International Conference on Application of Intelligent Systems in Multi-modal Information Analytics*, pages 466–473. Springer.

Luo, C., Wu, D., and Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65:465–470.

Malekipirbazari, M. and Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631.

Roy, A. G. and Urolagin, S. (2019). Credit risk assessment using decision tree and support vector machine based data analytics. In *Creative Business and Social Innovations for a Sustainable Future*, pages 79–84. Springer.

Roy, P. and Shaw, K. (2021). A credit scoring model for smes using ahp and topsis. *International Journal of Finance and Economics*.

Saia, R. and Carta, S. (2016a). An entropy based algorithm for credit scoring. In *International Conference on Research and Practical Issues of Enterprise Information Systems*, pages 263–276. Springer.

Saia, R. and Carta, S. (2016b). Introducing a vector space model to perform a proactive credit scoring. In *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, pages 125–148. Springer.

Saia, R. and Carta, S. (2016c). A linear-dependence-based approach to design proactive credit scoring models. In *KDIR*, pages 111–120.

Saia, R. and Carta, S. (2017a). Evaluating credit card transactions in the frequency domain for a proactive fraud detection approach. In *SECRYPT*, pages 335–342.

Saia, R. and Carta, S. (2017b). A fourier spectral pattern analysis to design credit scoring models. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, page 18. ACM.

Saia, R., Carta, S., et al. (2017). A frequency-domain-based pattern mining for credit card fraud detection. In *IoTBDS*, pages 386–391.

Saia, R., Carta, S., and Fenu, G. (2018a). A wavelet-based data analysis to credit scoring. In *Proceedings of the 2nd International Conference on Digital Signal Processing*, pages 176–180. ACM.

Saia, R., Carta, S., and Recupero, D. R. (2018b). A probabilistic-driven ensemble approach to perform event classification in intrusion detection system. In *KDIR*, pages 139–146.

Saia, R., Carta, S., Recupero, D. R., Fenu, G., and Saia, M. (2019a). A discretized enriched technique to enhance machine learning performance in credit scoring. In *KDIR*, pages 202–213.

Saia, R., Carta, S., Recupero, D. R., Fenu, G., and Stanciu, M. (2019b). A discretized extended feature space (defs) model to improve the anomaly detection performance in network intrusion detection systems. In *KDIR*, pages 322–329.

Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons.

Sloan, R. H. and Warner, R. (2018). When is an algorithm transparent? predictive analytics, privacy, and public policy. *IEEE Security & Privacy*, 16(3):18–25.

Sohn, S. Y., Kim, D. H., and Yoon, J. H. (2016). Technology credit scoring model with fuzzy logistic regression. *Applied Soft Computing*, 43:150–158.

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit Scoring and Its Applications, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Tripathi, D., Edla, D. R., and Cheruku, R. (2018). Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification. *Journal of Intelligent & Fuzzy Systems*, 34(3):1543–1549.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19.

Xia, Y., He, L., Li, Y.-G., Fu, Y., and Xu, Y. (2020). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy*, pages 1–24.

Zhang, W., He, H., and Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, 121:221–232.

Zhang, X., Yang, Y., and Zhou, Z. (2018). A novel credit scoring model based on optimized random forest. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 60–65. IEEE.