International Journal of Applied Business and Economic Research
Volume 16, Number 4, 2018, ISSN : 0972-7302
available at http: www.serialsjournals.com

REJECT INFERENCE IN SCORING MODEL USING MACHINE LEARNING TECHNIQUES

Hasna Chamlal[1], Tayeb Ouaderhman[2], Mehdi Bazzi[1] and Youssef Tounsi[3]

[1] Département de Mathématiques et Informatique, Laboratoire de Modélisation, Analyse, Contrôle et Statistique, Faculté des Sciences Aîn Chock, Université Hassan II, Casablanca, Morocco
[2] Département de Mathématiques et Informatique, Laboratoire de recherche en Science des Matériaux des Milieux et de la Modélisation, FPK, Université Hassan premier, Khouribga, Morocco
[3] Laboratoire de Recherche en Réseaux, Informatique, Télécommunications et Multimédia, Ecole Supérieure de Technologie, Université Hassan II, Casablanca, Morocco
E-mail:chamlal@yahoo.com, t.ouaderhman@gmail.com, bazzimehdi@gmail.com, tounsi@gmail.com

Abstract: By developing the credit scoring models based on data of accepted applicants, the basic rule of statistics, having a random sample, is not respected. To remedy to this bias reject inference techniques can be applied to reintegrate rejected applications into the training sample. In this study, we applied machine learning technics, i.e., decision trees (DTs), logistic regression (LR), support vector machines (SVMs), random forests (RFs), and Bagging, to three public credit scoring datasets from UCI database and a real Moroccan private dataset that includes rejected applicants. The receiver operating characteristic (ROC) curve, used as accuracy indicator, show that the RFs technique performed the best with the three credit scoring datasets, and logistic regression (LR) showed the best performance with a direct marketing dataset.

Keywords: Reject inference; scoring model; Machine learning.

I. INTRODUCTION

Predictive analytics, which typically involve classification (discrimination) and regression operations [1], are used to forecast events by analyzing historical and transactional data using statistical, modeling, data mining, machine leaning, and artificial intelligence techniques. Predictive analytics are commonly used in the banking sector to provide various predictive scores [2] [3], such as application (credit), propensity, behavioral, collection, recovery, and attrition scores. In addition, predictive analytics can be used to detect fraudulent applications.

Application (credit) scoring refers to the assessment of the creditworthiness of new applicants. Application scoring considers the probability that an applicant will default on debt obligations based on answers to questions in application forms, e.g., current salary, number of dependents, and time at current residence [4].

A propensity score measures the probability that a customer will be interested in a given product or service. Propensity scores are calculated for existing customers based on their banking history, e.g., account activity and previously purchased banking products, and sociodemographic characteristics.

Behavioral scoring is performed for existing customers and is similar to application scoring. In fact, the decision about that how the lender has to deal with the borrower is in this area. Behavioral scoring models use historical data, such as account age, account activity, account balance, and past due payment history, to predict the time at which a borrower may default.

Collection scoring groups customers with different levels of insolvency. Customers who require more decisive action are separated from those who do not require immediate attention. Collection scoring models consider the degree of delinquency (early, middle, and late recovery) and facilitate better management of delinquent customers, i.e., from the first signs of delinquency (30–60 days) to subsequent phases and debt write-off.

A recovery score evaluates the amount that can be recovered relative to disputed accounts or loans. A recovery score can suggest the most effective recovery actions, while avoiding disproportionate actions against loyal, profitable, and low-risk customers.

Fraud detection models rank applicants according to the relative likelihood that an application may be fraudulent.

An attrition score measures the probability that a customer will leave the bank. Attrition scores are calculated for individuals who have been bank customers for at least several months. Attrition scores are based on account history, purchased banking products, general relations with the bank, and sociodemographic characteristics.

This study addresses reject inference by focusing on credit and propensity scores. Over time, the loan performance of accepted applicants can be categorized as good or bad. The probability that an accepted application will become a bad loan can be estimated using loan performance data, however, the estimated probability that a rejected application may in fact yield a good loan might be biased. The possibility of bias also exists for propensity scores because they are estimated using historical client data. However, these data are only available for existing customers or prospects selected in a previous advertising campaign, i.e., no data are available for new or rejected prospects.

Two problems arise because the scoring system must be applied to the entire population rather than only those who would be selected via a previous system. First, the number of customer observations available for analysis is reduced by excluding the rejected. Second, samples are not selected randomly. However, the former is not a significant problem because the number of existing observations is typically very large. However, the latter could cause biased results. Reject inference methodologies, which are typically used in credit scoring, can account for and correct such sample biases.

In a credit scoring system, a model is usually developed from a sample of clients who have obtained credit. However, selection bias occurs if rejected applications are not considered [8] [9] (Figure 1).

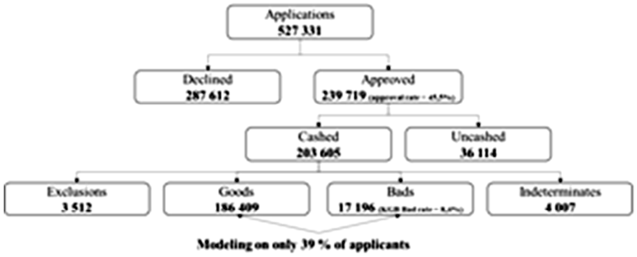

Figure 1: Selection bias illustration

This will be illustrated with another example concerning the applicant's activity. Based on economic theory, farmer applicants have a higher probability of default than other applicants do mainly in certain regions empirically turns out to be very risky. Say that a new bank starting a loan product is not aware of this risky category. After a certain period, the bank finds out that it lost a lot of money to this group and it creates a credit scoring in which it punishes the farmer group. After the implementation of this model, the bank refuses loans to many of the farmer potential clients but grants a loan to some of them that excel in other categories such as average yearly balance. When the bank now creates, a new credit scoring system it finds out that a very high percentage of the farmer clients paid back as agreed. If the bank based on this new fact now would decide to loosen the punishment of this group, it would end up in the same situation as in the first stage and find out again that farmer people default more than average. The problem arose when the second model was built. By only concerning the accepted agriculturalist people, the scoring system was based on the very elite sample of farmer that were granted a loan because their average yearly balance was exceptional high. Thus, a potential bias occurs. Therefore, reject inference techniques attempt to incorporate characteristics of rejected prospects into the process of calibrating the scoring.

The remainder of this paper is organized as follows. Section II reviews the related study. Note that few studies have considered reject inference relative to propensity scores. Commonly used machine learning techniques are described in Section III. In Section IV, machines learning techniques are applied to address the reject inference problem using credit scoring and direct marketing datasets. In addition, the performance of the obtained models is compared. Section V discusses the conclusions and suggestions for future study.

II. RELATED WORK

In order to structure the following discussion, we distinguish between the selection mechanism that determines whether an applicant is rejected or accepted by the bank, and the outcome mechanism that determines the response if good or bad loan of the applicants. Note that we refer to selection as a missing-data mechanism. According to Rubin (1976) and Little and Rubin (2002), missing data can be classified into three groups: missing completely at random, missing at random, and missing not at random (MNAR). In the first two cases, the missing-data mechanism is ignorable. Indeed, given all exogenous model variables, missing at random means that the probability of default is equal regardless of whether an application is accepted or rejected. In contrast, for MNAR data, the missing-data mechanism is not ignorable because additional information regarding future default obtained via human evaluation is considered, which can change the probability of defaulting. However, sample selection bias occurs in this case. Here, missing data must be included in the model to obtain proper outcome estimates.

Reject inference techniques attempt to incorporate the characteristics of rejected applicants into the process of calibrating a scoring system based primarily on the behavior of accepted applicants or targeted consumers. Various reject inference techniques have been proposed in the literature or by consultancies [8]. The primary objective of credit scoring is to model the outcome mechanism. We assume that a vector of explanatory variables X = (X1, …., Xk) is observed completely for each applicant, and the class label Y (Y is the dependent variable) is observed for accepted applicants but missing for rejected applicants. Conventionally, a bad loan is labeled 0, and a good loan is labeled 1. Furthermore, we define an auxiliary variable a, where a = 1, if the applicant is accepted and a = 0, if rejected. Note that y is observed, if a = 1 and missing, if a = 0. Reject inference attempts to correct this inherent flaw using information about the rejected accounts. Various reject inference techniques have been proposed in the studies [5] [6] [8] [11] and has been empirically compares the predictive performance of algorithms that incorporate different possible reject inference techniques. Note that only a few studies have examined reject inference relative to propensity scores. Thus, we investigated reject inference relative to both credit scoring and propensity scores. Moreover, all previous studies into the reject inference problem simulated rejected applicants from the data of accepted applicants. In this study, we focus on the reject inference problem using a real-world Moroccan bank dataset that includes data about rejected applicants.

III. METHODOLOGY

A. The Machines Learning Methods

Several machine learning techniques can be used to construct and estimate scoring models, including SVM, logistic regression (LR), decision tree (DT), random forests (RFs), and bagging (BG) techniques. These techniques are introduced in the following sections.

1) Support vector machine (SVM)

The main idea of the SVM algorithm is that given a set of points belonging to one of the two classes, an optimal method is required to separate the two classes by a hyperplane (Figure 2). This is achieved by maximizing the distance from the closest points of either class to the separating hyperplane and minimizing the risk of misclassifying training samples and unseen test samples.

$$\begin{cases} \min_{\omega,b} & \frac{1}{2}||w||^2 \\ s.c & y_i(w^T x_i + b) \geq 1 \ \ \forall i = 1,...,p \end{cases} \tag{1}$$

Note that SVMs can be linear or nonlinear depending on how the given points are separated into the two available classes.

2) Logistic regression

LR is also known as logit regression or the logit model [17], is a regression model wherein the dependent variable

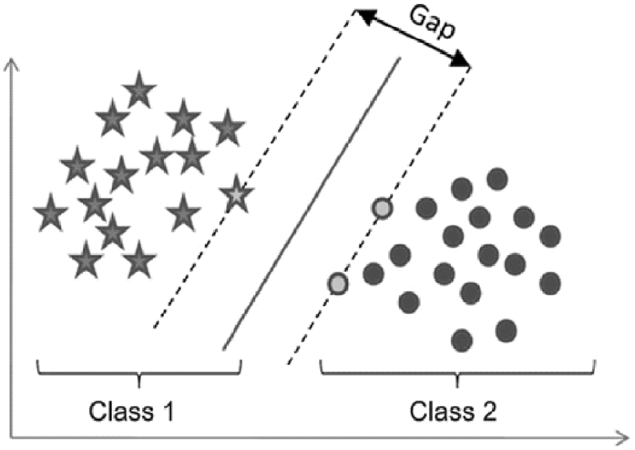

Figure 2: Support Vector Machines

is categorical. Here, we consider the case of a binary dependent variable, i.e., the variable can only take values of 0 or 1, which represent contrasting outcomes, such as good and bad payers, respectively. Cases wherein the dependent variable has more than two outcome categories can be analyzed using multinomial LR, or if the multiple categories are ordered, ordinal LR can be used [1]. In economics, LR is a representative example of a qualitative response/discrete choice model. LR can be implemented using logistic functions to predict the log odds ratios using the following formula.

$$logit(p) = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_n X^k \qquad (2)$$

The probability formula is expressed as follows.

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}} \qquad (3)$$

Here, *logit*(*p*) is a linear function of the explanatory variable $X = (X^1, ..., X^k)$, which is similar to linear regression.

LR was developed by statistician David Cox in 1958 [1] [4]. LR has been compared to other credit scoring techniques in the works [1] [11] [13] [14].

3) Decision trees

DTs are one of the most widely used machine learning classification and prediction methods. A DT can deal with both numerical and categorical data (e.g., gender); thus, it is easily applicable for determining personal credit ratings [3]. A DTs can be defined as a tree in which each branch node specifies a choice between the number of alternatives, and each leaf node distinguishes a classification or decision. Most algorithms for DTs induction, such as the ID3, C4.5, J4.8, CART and Credal Decision Tree algorithms, follow a greedy top-down recursive divide-and-conquer approach that begins with a training set and its associated class labels (King and Zhu, 1998). Kao *et al.* (2012) combined a Bayesian behavior scoring model and a CART-based credit scoring model.

4) Bagging

Bagging is known as bootstrap aggregation, which is a machine ensemble learning method proposed by Breiman (1996), is used herein to obtain robust and accurate landslide models. Bagging is useful in landslide susceptibility models because it is sensitive to small changes in the training data, thus, it can improve the prediction capabilities of the model. The Bagging algorithm comprises three steps. First, bootstrap samples are obtained via random resampling from the training dataset to form a set of training subsets. Then, multiple classifier-based models are constructed based on each of the subsets. The final model is then formed by aggregating all classifier-based models.

5) Random Forests

The RFs technique refers to an extension of Bagging that applies to the particular case of decision trees [1]. Indeed, RFs is an ensemble method that combines the results of tree predictors, which are built after introducing two levels of randomization. Each tree in the forest is grown as follows.

First, subjects are sampled randomly from the data. The same number of subjects are sampled randomly with replacement from the original data and used as a training dataset to grow trees. Note that this sampling leaves out approximately one-third of the subjects. These samples are referred to as out-of-bag samples and are used as the test dataset to obtain an unbiased estimation of the prediction rate and variable importance. Thus, the RFs techniques require no external testing samples. Candidate variables are selected randomly to determine splitting criteria at each node of the tree. If there are M variables in the dataset, a number m much less than M is specified, such that at each node, only m variables are selected at random for evaluation, and the variable that most differentiates the predicting trait is selected to split the node. This splitting procedure is repeated to obtain all the nodes of the tree. The above steps are repeated to grow a predetermined number of trees to form a RFs. The prediction results of all trees are then pooled to "vote" for the best overall prediction result.

The RFs technique provides excellent prediction accuracy, and the importance of each variable can be measured using the difference in prediction accuracy before and after randomly permuting the values of the given variable. This measure includes both the marginal effects of the variable and the effects of interacting with other factors.

B. Machine Learning Evaluation

Many evaluation measurements are available for evaluating the predictive performance of models relative to scoring system, including receiver operating characteristics (ROC) curves, average accuracy, and Type I and Type II errors.

The ROC curve was first applied to assess how well the radar equipment used in World War II distinguished random interference or "noise" from the signals that were truly indicative of enemy planes [12].

The ROC curve plots the sensitivity or the true positives (TP) of a model on the vertical axis against 1-specificity or false positives (FP) on the horizontal axis. The area under the ROC curve (AUC) is a convenient way to compare different predictive models for binary outcomes.

Table I
Confusion Matrix for Credit Scoring

	Predicted class (%)	
Actual class (%)	Good loans	Bad loans
Good loans	TP	FN (Type II error)
Bad loans	FP (Type I error)	TN

From the confusion matrix table, the following calculations are defined:

$$\mathit{Average\ Accuracy}\,(ACC) = \frac{TP + TN}{TP + FN + TN + FP} \quad (4)$$

$$\mathit{Type\ I\ error} = \frac{FP}{TN + FP} \quad (5)$$

$$\mathit{Type\ II\ error} = \frac{FN}{TP + FN} \quad (6)$$

A TP is a good applicant that is correctly classified as good, and a true negative (TN) is a bad applicant that is correctly classified as bad. A false negative (FN), i.e., a Type II error, is a good applicant incorrectly classified as a bad application, and a FP, i.e., a Type I error, is a bad customer that is incorrectly classified as a good customer, which is a high-risk case

IV. RESULTS, ANALYSIS AND DISCUSSIONS

A. Data and variables

Four datasets are used in our evaluation. The first dataset is a private credit scoring dataset obtained from a Moroccan bank that contains 10 417 applicants including 7550 accepted and 2867 rejected, and it has observations relative to 10 variables. The second dataset is a famous German credit dataset, the third dataset is an Australian credit approval dataset, and the fourth is a direct marketing campaign dataset.

The German credit dataset (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)) contains observations relative to 21 variables for 1000 past credit applicants. In this dataset, each applicant is rated as having good (700 cases) or bad (300 cases) credit. The Australian credit approval data originates from quinlan (http://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval)) (Moro et al., 2011) and contains information about credit card application. The attribute names and values in this dataset have been changed to meaningless symbols to protect confidentiality. Herein, the number of instances is 690, and the number of attributes is 14. Note that this dataset has a few missing values. We also used a public dataset collected from a Portuguese bank [10] (https://archive.ics.uci.edu/ml/datasets/bank+marketing) for the propensity score (Table 2). The bank used its own contact center to conduct direct marketing campaigns to motivate and attract deposit clients would be "yes" or not; "no" subscribed. This dataset (Bank-full.csv) is ordered by date and contains 17 attributes and various examples corresponding to 45211 objects.

B. Research design and results

The approach used for the private dataset comprises building a model, i.e., Model1, using the SVM, LR, CART, RFs and Bagging for the good/bad known population. Then, the rejected population was inferred and scored using the method that obtained the best ROC performance. Note that the same discriminant threshold was used to select good or bad for the rejected population. Subsequently, we constructed a new model, i.e., Model2, for the entire population using the same machine learning methods used previously. The dataset was split into training (70%) and testing (30%) sets for each model based on stratified sampling (accepted/rejected).

For the private dataset, a simulation study was conducted following the steps proposed by Anderson and Hardin (2013) [5].

- Step 1: Build models of LR, CART, RFs, Bagging and supervised SVM using the sample of all the accepted applicants: Split the data into a training set and a test set with a ratio 7:3.
- Step 2: The model with the best ROC performance from Step 1 is used to assign good/bad labels to rejected applicants, seen as the truth.
- Step 3: The pooled data are divided into training and test sets at a ratio of 7:3 based on stratified sampling (accepted/rejected).
- Step 4: LR, CART, RFs, Bagging and supervised SVM models are constructed using the training set.
- Step 5: The classification rules derived in Step 4 are applied to the test set.
- Step 6: Performance is evaluated using the test set.

We then assigned good/bad labels to the rejected applicants based on the RFs model, which demonstrated the best ROC performance (79.49%; Table 2). Then, the pooled data were divided into training and test sets at a ratio of 7:3 based on stratified sampling (accepted/rejected). Stratification is the process of dividing members of a population into two subgroups, i.e., good or bad, before sampling (ratio: 7:3) and later merging into a training set: 70% good and 70% bad and a test set: 30% good and 30% bad. We then constructed LR, CART, RFs, Bagging, and supervised SVM models using the training set. These models were then evaluated using the test set.

For the public datasets, we conducted a simulation study following the steps proposed by Anderson and Hardin (2013).

- Step 1: Build LR, CART, RFs, Bagging, and supervised SVM models using the entire dataset.
- Step 2: The model with best ROC performance from Step 1 is used to create rejected applicants (20% of the applicants with a lower score).
- Step 3: The pooled data are divided into training and test sets (ratio: 7:3) based on stratified sampling (accepted/rejected).
- Step 4: Build LR, CART, RFs, Bagging, and supervised SVM models using the training set.
- Step 5: The classification rules derived in Step 4 are applied to the test set.
- Step 6: Performance is evaluated using the test set.

The proposed approach first creates the rejected population (20% with a lower score) by building a model in Step 1.

Table II
Results of the step 1-case of Moroccan bank dataset

ML Technique	*Area under ROC*
LR	72,42%
CART	61,31%
Bagging	78,72%
SVM	54,09%
RFs	79,49%

Table III
Results of the step 1

ML Technique	*Area under ROC*		
	German Credit data	*Australian credit*	*Direct Marketing Campaigns*
LR	81,01%	92,49%	**91,89%**
CART	79,77%	89,17%	66,76%
Bagging	84,33%	94,00%	73,63%
SVM	80,12%	90,85%	66,70%
RFs	**88,63%**	93,93%	73,25%

NB: The model with highest AUC is highlighted in bold.

We then assigned good/bad labels to the rejected applicants based on the RFs model for the German credit data (ROC performance: 88.63%), Bagging for the Australian credit dataset (ROC performance: 94%), and LR for the direct marketing campaign dataset (ROC performance: 91.89%; Table 3). The pooled data were then divided into training and test sets (ratio: 7:3) based on stratified sampling (accepted/rejected). Using the training set, LR, CART, RFs, Bagging, and supervised SVM models were constructed, these models were then evaluated using the test set.

The performance of these different methods was evaluated by comparing the AUC values. For Step 1, each method was executed only once, and for Step 2, each technique was executed for 100 runs.

To evaluate the discriminant power of the five methods, we calculated the average and standard deviation of the AUC values based on 100 simulations. The corresponding box plots are shown in Figure 3. Table 4 shows that the RF technique demonstrated the best reject inference for the Moroccan bank, German credit, and Australian credit datasets. For the direct marketing dataset, LR demonstrated the best performance. Table 5 summarizes the accuracy obtained using the five methods relative to the accepted, rejected, good, bad, and Type I and II errors. For the Moroccan dataset, RF showed the best overall accuracy relative to accepted, rejected, and both. For the other datasets, the RF and Bagging techniques performed better relative to both rejected and accepted cases. In terms of Type I errors (i.e., good applicants are predicted as bad), the RFs technique consistently performed better compared with the other four methods. In contrast, for Type II errors (i.e., bad applicants are predicted as good), the RFs technique showed the lowest performance in most cases.

Note that we found very few Type II errors, which, in practice, helps reduce business losses.

Table IV
Overall performance by AUC

Dataset	*Avg/SD*	*LR*	*CART*	*Bag*	*SVM*	*RFs*
Private dataset	Avg of AUC	76,28%	70,23%	89,90%	75,33%	91,59%
	SD	1,06%	2,46%	0,94%	2,25%	0,84%
German Credit	Avg of AUC	81,76%	76,19%	85,85%	80,34%	89,25%
	SD	2,01%	2,72%	2,05%	2,26%	1,76%
Australian credit	Avg of AUC	94,02%	92,08%	94,97%	93,44%	95,30%
	SD	1,22%	1,86%	1,17%	1,48%	1,18%
Direct marketing	Avg of AUC	91,89%	66,65%	72,97%	77,87%	73,05%
	SD	0,30%	1,45%	0,68%	1,77%	0,61%

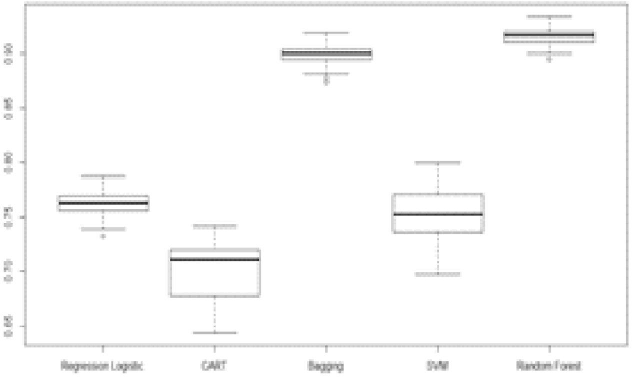

Figure 3: AUC Moroccan bank dataset box plot

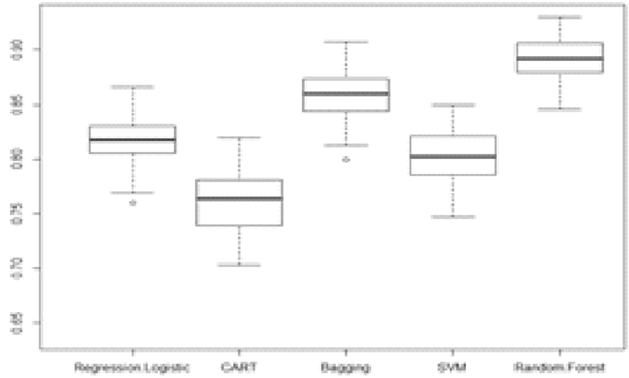

Figure 4: AUC German dataset box plot

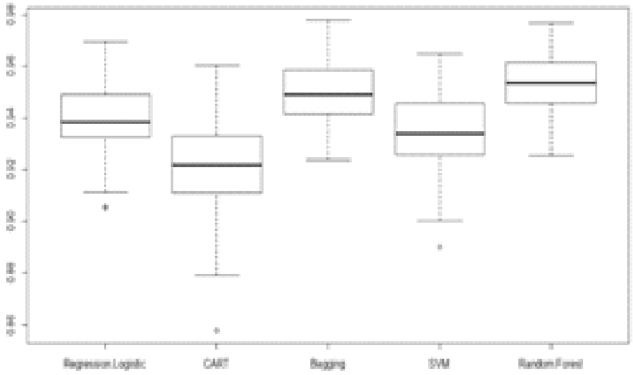

Figure 5: AUC Australian dataset box plot

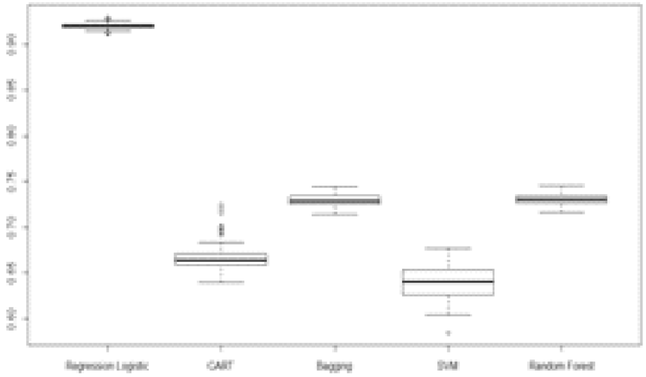

Figure 6: AUC Direct marketing campaigns dataset box plot

Table V
Accuracy Results Table

		Overall accuracy					Type I errors					Type II errors				
		Logit	CART	Bag	SVM	RF	Logit	CART	Bag	SVM	RF	Logit	CART	Bag	SVM	RF
PRIVATE DATASET	**Accepted**	0,893	0,894	0,924	0,896	**0,930**	0,895	0,774	0,698	0,960	*0,659*	0,048	0,067	0,040	0,055	*0,036*
	Rejected	0,745	0,724	0,853	0,620	**0,884**	0,428	0,928	0,248	0,640	*0,196*	0,171	*0,000*	0,104	0,270	0,083
	Both	0,859	0,848	0,911	0,800	**0,923**	0,599	0,595	0,368	0,829	*0,321*	0,078	0,091	0,050	0,113	*0,044*
GERMAN CREDIT	**Accepted**	0,681	**0,804**	0,775	0,780	0,764	0,500	*0,322*	0,348	0,371	0,400	0,170	*0,141*	0,168	0,157	0,167
	Rejected	0,980	0,970	**0,990**	0,980	0,985	0,010	0,010	*0,000*	0,010	0,005	1,000	1,000	*0,500*	1,000	0,667
	Both	0,780	0,730	0,785	0,759	**0,819**	0,253	0,316	0,247	0,277	*0,206*	0,195	0,233	0,191	0,214	*0,162*
AUSTRALIAN CREDIT	**Accepted**	0,825	0,787	0,829	0,810	**0,839**	0,205	0,296	0,202	0,225	*0,191*	0,154	*0,126*	0,148	0,164	0,139
	Rejected	0,978	**0,986**	0,978	0,971	**0,978**	1,000	1,000	1,000	1,000	1,000	0,015	*0,014*	0,015	0,015	0,015
	Both	0,887	0,891	**0,989**	0,880	0,891	0,139	0,144	*0,121*	0,148	0,131	0,095	*0,084*	0,088	0,101	0,094
DIRECT MARKETING	**Accepted**	0,896	0,876	**0,903**	0,876	0,876	0,436	0,518	*0,426*	0,518	0,518	0,059	0,052	*0,040*	0,052	0,052
	Rejected	0,997	0,992	**0,998**	0,992	0,992	1,000	0,969	*0,818*	0,969	0,969	*0,001*	*0,001*	*0,001*	*0,001*	*0,001*
	Both	0,911	0,908	**0,921**	0,908	0,908	0,455	0,467	*0,377*	0,467	0,467	*0,049*	0,059	0,054	0,059	0,059

NB: The model with highest accuracy is highlighted in bold and the one with the smallest error rates was highlighted in italics

V. CONCLUDING REMARKS

The reject inference problem has a long history in credit scoring; however, this problem has not been resolved adequately yet. Since the repayment behavior information of rejected applicants is unavailable, the reject inference problem can be considered as a statistical problem with missing data. Various statistical techniques are used by credit scorers depending on whether the data that are missing at random or not. In addition, given the wide application of machine learning techniques and the increasing use of RFs as an efficient classification algorithm, reject inference can also be considered a machine learning problem, wherein algorithms learn to use information from a rejected group to optimize an objective gradually. In this study, we tested the predictive performance using several datasets covering Risk and Marketing fields. To the best of our knowledge, this study was the first to focus on real-world data with real rejected applicants relative to the reject inference problem. The results showed that applying the RFs technique to the reject inference problem demonstrated the best performance. Compared to the LR, Bagging, CART, and supervised SVM techniques, the RFs approach showed better performance for both the German credit and the Australian datasets. In addition, LR showed the best performance relative to the propensity score with the direct marketing campaign dataset. In a simulation, we proved that using the rejected applicant information is valuable in practice. We found that these machine learning methods can be used as an effective reject inference technique for credit scoring applications. In future, we plan to introduce deep learning and other ensemble algorithms to address the reject inference problem.

REFERENCES

W. Chen, C. Ma and L. Ma (2009). "Mining the customer credit using hybrid support vector machine technique", Expert Systems with Applications 36(4):7611-7616.

S. Tuffery (2012). Data mining et statistique decisionnelle: l'intelligence des donnees, in: Editions Technip, 4eme edition.

L. Breiman (1996). "Bagging Predictors", Machine Learning, 26(2), 123-140.

L. Thomas, D. Edelman and J. Crook (2002). Credit Scoring and its Applications. SIAM: Philadelphia, USA.

Z. Li, Y. Tian, K. Li, F. Zhou and W. Yang (2017). "Reject inference in credit scoring using Semi-supervised Support Vector Machines" Expert Systems with Applications 74, 105-114.

J. Banasik and J. Crook (2007). "Reject inference, augmentation, and sample selection", European Journal of Operational Research, , 183(3), 1582-1594.

H. Chamlal, T. Ouaderhman, M. Bazzi and Y. Tounsi (2017). "Reject Inference in Credit Scoring: Classical vs Machine Learning Approach", 61st ISI word statistics congress, Marrakech, Morocco.

J. Crook and J. Banasik (2004). "Does Reject Inference Really Improve the Performance of Application Scoring Models?", *Journal of Banking & Finance* 28, 857–874.

A.J. Feelders, S. Chang and G.J. McLachlan (1998). "Mining in the Presence of Selectivity Bias and Its Application to Reject Inference", AAAI Press.

Y. Freund and R.E Schapire (1997). "A Decision Theoretic Generalization of On Line Learning and an Application to Boosting", journal of computer and system sciences 55, 119-139.

D.J Hand and W.E Henely (1993). "Can Reject Inference Ever Work?", IMA Journal of Mathematics Applied in Business and Industry, 5(1), 45-55.

T. Hastie, R. Tibshirani and J. Friedman (2001). "The Elements of Statistical Learning", in: Springer-Verlag, NewYork.

S. Moro, P. Cortez and P. Rita (2014). "A Data-Driven Approach to Predict the Success of Bank Telemarketing". Decision Support Systems, 62, 22-31.

S. Moro, R. Laureano and P. Cortez (2011). "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology", Proceedings of the European Simulation and Modelling Conference - ESM'2011, Portugal.

N. Sun, J.G Morris, J. Xu, X. Zhu and M. Xie (2014). "iCARE: A Framework for Big Data-Based Banking Customer Analytics", IBM Journal of Research and Development, 58(5/6), 4-1.

C. Tsai and J. Wu (2008). Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring. Expert Systems with Applications, 34(4), 2639-2649.