Improving Credit Risk Analysis with Cluster Based Modeling and Threshold Selection

Ajay Byanjankar
Abo Akademi University
Turku, Finland
ajay.byanjankar@abo.fi

Abstract

Credit risk has been an integral part of financial industry and is a challenging and difficult risk to manage. The diverse behavior of borrowers adds challenges to the risk analysis. Failing to accurately identify the borrowers' risk can lead to huge investment losses. Credit scoring is a popular and commonly used technique to analyze credit risk. A single credit scoring model may not be capable of generating a common rule to classify borrowers and hence segmented modeling can be applied to create more specific classification rules for achieving higher classification accuracy. In this study segmented modeling is applied with threshold selection for each segment to reduce relative cost of misclassification. The results from the study show that threshold selection based on the segmented modeling can give improvement over a single credit scoring model.

1. Introduction

Credit risk is an integral part of financial industry, where the focus in on accurately identifying credit risk associated with a borrower in preventing defaults. It is perceived as a difficult risk to manage and therefore is given high importance. The varying nature of borrowers make the risk analysis difficult and the difficulty is further increased by the dynamic financial environment and growing credit volumes. With credit risk analysis being a vital part of credit decision, its precision is an important determinant of credit management [1]. The result of poor credit decisions can lead to huge losses and thus the emphasize should be on to analyze the risk as accurately as possible. Most of the losses are due to the failure of borrowers to payback the credit [2].

The significance of credit risk management has created the need for more sophisticated techniques and tools in correctly identifying credit risk. Data mining tools have been a popular and great importance in building predictive models by applying pattern recognition methods [3]. Credit scoring model has been a popular and a standard tool in financial industry in predicting credit risk and selecting loan portfolios [4]. Ghatge and Halkarnikar [5] define credit scoring as statistical method that predicts the creditworthiness of a borrower. Credit scoring systems are developed using historical information on borrowers to decide whether to approve or decline a loan. It is used to predict the likelihood of a borrower to default on a loan [6] [7].

In most cases a single credit risk model is developed based on the historical information of past borrowers to predict the likelihood of default for a new borrower and decide whether to approve or decline a loan. However, with the presence of varying nature of borrowers a single classification rule may not be sufficient to capture the behavior pattern of various individual borrowers [8]. To overcome this problem segmented modeling can be applied, where borrowers are segmented based on their similarities and a separate model is developed for each segment of borrowers. As borrowers in each segment show similar behavior pattern, classification rules for each segment can be more specific and hence can contribute in increased accuracy of risk identification [9] [10] [11].

Credit scoring generally is applied as a binary classification model that classifies borrowers as good or bad. When classifying borrowers, a cut-off point (threshold) is applied to decide if a borrower should be classified as bad or good based on their likelihood of belonging to a class predicted by the model. Similar to classification rule, a single threshold may not be appropriate in classifying new applicants. Borrowers from low risk group have relatively low likelihood of default than high risk group, presenting positive correlation of likelihood of default with the assigned risk group [12] [13]. Therefore, setting separate thresholds for each segment based on the level of risk can help in identifying more risky borrowers. A small improvement in the accuracy of the credit decision might be helpful in reducing credit risk which can provide more savings [2].

The objective of this paper is to segment the

URI: https://hdl.handle.net/10125/63913
978-0-9981331-3-3

borrowers into different segments based on their similarities and model each segment separately for better risk assessment. The aim is to select a best threshold for each segment to minimize the overall risk. The segmentation is done with K-means clustering. The threshold selection for the classification is an imperative part of a classification task and is controlled by the business objective. Credit scoring being a cost sensitive task, where cost of misclassifying a bad borrower as good is more costly than misclassifying good borrowers as bad. Hence, we apply the idea of relative cost of misclassifcation to select the best threshold that minimizes the misclassification cost [14]. By treating each risk group separately and setting thresholds based on the segmented models, the objective is to obtain higher precision in risk identification and lower misclassification cost.

The rest of the paper is structured as follows. In section 2, related literature in the field of credit scoring and segmented modeling is summarized. Section 3 describes the data and research methods used in the study. The experimental process is presented in section 4 and the results are discussed in section 5. Finally, the conclusion is presented in section 6.

2. Literature Review

In this chapter we present the related studies to motivate the relevance of our study. Credit scoring is a popular analytical tool used by financial institutions to evaluate credit risk of borrowers. It is applied as a classification task to classify new loan applicants as good or bad borrowers based on their information provided. Credit scoring as a popular tool for predicting credit worthiness of loan applicants have been successfully applied using statistical and machine learning models [6][15][16]. In addition to being a classification task, credit scoring is a cost sensitive task and the cost of misclassification varies across the groups. The cost of misclassiying a bad borrower as good borrower is very high compared to misclassiying a good borrower as bad borrower. Hence, these costs need to be considered while evaluating the effectiveness of a model. Considering the different costs of misclassification Relative Cost of misclassification can be applied to measure the effectiveness of a model results which focuses on reducing the misclassifcation cost [17][14].

Several studies have shown the effectiveness of segmented modeling over a single credit risk model. Scitovski et al. [10] used adaptive Mahalanobis clustering algorithm for segmenting retail clients of a Croatian bank. They proposed the use of separate credit scoring models for each segments for better risk assessment and customized business strategy to each cluster. Correa et al. [9] applied cluster analysis as a part of a predictive algorithm, where they first determined to which cluster a client belongs to. Then they calculated a specific credit risk scorecard for each cluster and compared the result with the traditional method of developing a single scorecard. The results from the clustering showed a sign of improvement.

Ghanbari et al. [11] performed a cluster-based classification in retail banking data. They first developed credit scoring model with three classification techniques logistic regression, decision tree and support vector machine. The scoring models were re-built with clusters from cluster analysis as an additional classifier input. The results showed increased classification accuracy for the models with clustering data. Similarly, Bakoben et al. [18] used the outcomes of cluster analysis of behavior of credit card accounts for behavioral scoring. The cluster analysis was performed with dissimilarity measure of statistical model parameters. Behavioral model was built using logistic regression with clustering results and aggregated behavior which outperformed the behavioral model with only aggregated behavior. Peng et al. [19] in their paper investigated the applicability of clustering in credit card accounts classification. Their results show that clustering as a single classification model has a low classification rate and classification results can be improved considerably by combining clustering results with supervised methods.

Polena and Regner [20] studied the borrowers behavior by segmenting borrowers according to their risk groups. They studied the determinants of borrowers' default in P2P lending for separate loan risk classes. Their study suggest that the significance of variable determining default vary according to the risk class and only few of the variables are consistently significant over all the risk groups. The studies so far have shown that the segmented modeling can result in increased accuracy in classifying default and not-default borrowers. However, the studies have applied clustering for segmenting the borrowers and consider a constant threshold across all the segments in classifying the borrowers. They do not show the effect of setting different threshold.

3. Data and Research Methods

This section describes the data used for the study, preprocessing of the data and feature selected for the modeling. Further, the method applied for the modeling are also presented.

3.1. Data

Data for the study is from the peer-to-peer lending platform Prosper and includes information on loans issued between 2005 and 2014. In its original state there are 113937 loans described by 81 features, which is further processed in relation to the study requirements. Features with high missing values were removed along with the redundant features. Similarly, features describing post loan approval behavior were also removed. Furthermore, literature review and domain knowledge were applied for screening relevant features. The final set of features selected are as follow:

- LoanStatus: Status of the loan
- BorrowerRate: Interest rate on the loan
- Term: Loan term
- ProsperRating: Rating assigned to the loan
- EmploymentStatus: Current employment status
- EmploymentStatusDuration: Length of employment status
- IsBorrowerHomeOwner: Type of home ownership
- CurrentCreditLines: Number of credit lines
- OpenRevolvingAccounts: Number of open revolving accounts
- ListingCategory: Purpose of loan
- InquiriesLast6Months: Number of inquiries in past 6 months
- CurrentDelinquencies: Number of account delinquent
- PublicRecordsLast10Years: Number of public records last 10 years
- BankUtilization: Percentage of revolving credit utilized
- AvailableBankcardCredit: Total available credit via bank card
- Investors: Number of investors that funded the loan
- OpenRevolvingMonthlyPayment: Monthly payment on revolving accounts
- StatedMonthlyIncome: Monthly income
- MonthlyLoanPayment: The scheduled monthly loan payment
- DebtToIncomeRatio: Debt to income ratio of borrower
- creditscore_average: Average of lower and upper credit scores

The loans are either 12, 36 or 60 months loans whose current status are described as *Cancelled, Chargedoff, Completed, Current, Defaulted, FinalpaymentInProgess, Past Due(>120 days), Past Due(1-15 days), Past Due(16-30 days), Past Due(31-60 days), Past Due(61-90 days) and Past Due(91-120 days)*. For our analysis, the loans with status *Chargedoff* and *Defaulted* were treated as Bad loans (Defaults) and loans with status *Completed* were treated as Good (Non-Defaults) loans . Rest of the loans were removed from the analysis as they are in payment process and their final state is not known. Loans are assigned the grades ranging from AA to HR, where AA is the best and HR is the worst. The feature *ListingCategory* was re-grouped to reduce the number of categories to avoid cardinality problem. Similarly, numerical features *PublicRecordsLast10Years and CurrentDeliquencies* were binned and converted to categorical. Outliers in some of the features were replaced with their median values. After all the preprocessing, the final data is composed of 55084 loans and 21 features. 30.8% of the loans are Default loans and 69.2% of them are Non-Defaults.

Majority of the loans (32.44%) are borrowed for the purpose of Debt consolidation. Most of the loans are assigned the rating C (17.18%) and D (20%). The HR rating consists of 13.12% of the loans. Table 1 shows the summary of the loans across the Ratings. As depicted in Table 1, the default risk varies across the Ratings and higher interest rate is set for the riskier groups to compensate for the risk. Furthermore, the default risk for the low risk group with the Rating AA is significantly low compared to HR Rating loans. Hence, with such huge difference in the default risk a single common threshold may not be suitable to accurately classify loans with varying nature and risk. Therefore, the study aims at applying separate threshold for loans based on their similarities with the objective of reducing the misclassification cost.

Table 1. Ratings summary

Prosper Rating	Default Rate(%)	Average Interest(%)	Average creditscore
AA	12	9	796
A	18	12	746
B	26	16	710
C	30	19	678
D	34	24	660
E	41	28	627
HR	49	29	607

3.2. Research Methods

3.2.1. Logistic Regression Logistic regression is a widely used statistical modeling technique that explains the probability of outcome in relation to explanatory features similar to linear regression. Unlike linear regression, logistic regression is applied to classify data, where the dependent variable is binary. It is widely used in developing credit risk scorecards and predicts the probability of an applicant belonging to one of the predefined class. Logistic regression is still considered to be a suitable method for credit scoring due the simplicity in model building and easiness in the interpretability of the model results [9][11].

3.2.2. Random Forest Random forest is an ensemble approach proposed by Breiman [21]. Random forests are non-parametric statistical method that allows the computation of regression and classification problems with a single versatile structure [22]. It is a combination of classification or regression trees that are created using bootstrap samples and random feature selection of the training data. Each individual tree votes for one class for each observation and the final prediction of random forest is the aggregate result of the ensemble trees. Random forest is more efficient than a single decision tree and can be trained in less time than a single decision tree when there are large number of predictors [21][23][24].

3.2.3. Gradient Boosting Model Gradient boosting is an ensemble algorithm that combines both the bagging and boosting approaches. It builds additive regression models by sequentially fitting a base learner (decision tree) at each iteration and applies gradient descent algorithm to minimise the loss function [25]. During each iteration, the base learner is built using a random sub-sample of the train data (without replacement) and weights are assigned to the data, where incorrectly classified data are given higher weights. The weights forces the new classification tree to put more emphasize on correctly classifying the incorrectly classified data points in the next iteration [26].

3.2.4. K-Means Clustering K-means is clustering algorithm introduced by MacQueen [27] and is popular due to its simplicity and fast computation that has the ability of efficiently partitioning huge amount of data [28]. K-means follows an iterative process to segment data into k mutually excessive clusters, where each cluster is represented by an adaptively-changing centroid, which starts from some initial value assignments. It computes the squared distances between the inputs and the centroids and the inputs are assigned to the nearest centroids. In the iterative process K-means minimizes the sum of squared distance from each data point to its cluster. The measure of distances is generally Euclidean distance [28][29]. The ”Elbow Method” was applied to select the optimal number of clusters, where the total within-cluster sum of squares of distances was used as the criteria to select the optimal number of clusters.

4. Experiment

In this section the analysis process in achieving the study objective is illustrated. It includes the description of the modeling process and the process of threshold selection for the optimal result.

4.1. Modeling

To begin the modeling process the data was partitioned into train and test sets in the ratio 80:20 to perform the modeling and validation. The whole training set was first used to build a credit scoring model using Logistic regression (LR), Random Forest (RF) and Gradient Boosting model (GBM) with 10 fold cross validation. The model performance was evaluated with Area Under the ROC curve, AUC score and Area under precision and recall curve (PR-AUC) on the test data. The performance evaluation of the models are reported in Table 3 and Table 4. For the segmented modeling, K-means clustering was first applied on the train set to segment the data. With the K-means results the train set was segmented into 5 clusters. The loans in the test set was then assigned to the clusters by calculating minimum euclidean distance to the clusters. A summary of the clusters is presented in Table 2.

Table 2. Cluster summary

Cluster	Deafult Rate(%)	Average Interest(%)	Average creditscore
1	19	13	717
2	50	20	674
3	33	17	733
4	44	27	626
5	21	18	710

The cluster summary shows that clusters 2 and 4 have a very high default rate with high interest and low credit scores and hence can be considered as risky groups. Clusters 1 and 5 have low default rates and high credit scores that suggest they have low risk. Similarly, cluster 3 has a moderate default risk but have high credit scores. Hence, with the help of clustering we are able to segment the loans into similar segments and it is visible that the clusters have varying risk. When applying a single model and a common threshold to classify the loans of such varying risk, it may not result in an accurate classification. Therefore, treating loans of similar risk separately for threshold selection could provide improvement in the classification results.

After obtaining the data for the segments, for each segment Logistic Regression, Random Forest and Gradient Boosting was applied for the modeling with 10 fold cross validation. The performance of each of the models for the segments are reported in Table 3 and Table 4. From the evaluation results with AUC and PR-AUC scores, GBM models are performing better for most of the cases. However, these results are not considered as the final evaluation results as the objective is obtaining the lowest misclassification cost. Therefore, the models are further evaluated based on the relative cost of misclassification.

Table 3. Model evaluation with AUC

AUC			
Segment	**LR**	**RF**	**GBM**
Full	0.743	0.757	**0.761**
Cluster 1	0.721	0.710	**0.723**
Cluster 2	0.623	**0.678**	0.598
Cluster 3	0.761	**0.774**	0.770
Cluster 4	0.688	0.698	**0.713**
Cluster 5	0.724	**0.732**	0.727

Table 4. Model evaluation with PR-AUC

PR-AUC			
Segment	**LR**	**RF**	**GBM**
Full	0.557	0.585	**0.591**
Cluster 1	0.364	0.372	**0.379**
Cluster 2	0.669	**0.710**	0.689
Cluster 3	0.581	0.611	**0.612**
Cluster 4	0.633	0.638	**0.656**
Cluster 5	**0.421**	0.417	0.404

4.2. Thresholding

A common practice in credit scoring model is to use a threshold in classifying loans to a class. A threshold acts as a cut off point, where loans having a likelihood of belonging to a class greater than the threshold is assigned to the class for which the threshold was applied. Depending on the business need, there are multiple ways to set a threshold that would provide an optimal result. Credit scoring is a cost sensitive problem, where the misclassification costs for the classes are not the same. The cost associated with misclassifying bad borrowers as good borrowers and approving their loan applications can be more costly than failing to correctly classify good borrowers. In practice, the cost of misclassifying bad borrowers is very high compared to misclassifying good borrowers as bad. Therefore, the objective of the study is to find the optimal threshold that would provide classification results with minimum misclassification cost.

A confusion matrix with the misclassification costs is described in Table 5. Cost of misclassifying bad loans as good loans is represented as cost of False Negatives, C(FN) and Cost of misclassifying good loans as bad loans is represented as cost of False Positives, C(FP). The costs of correctly classifying the bad and good loans are ignored as the focus for the study is on only the misclassification costs.

Table 5. Confusion matrix

	Actual Bad	Actual Good
Predicted Bad	C(TP)	C(FP)
Predicted Good	C(FN)	C(TN)

The relative cost of misclassification approach is applied to select the best threshold. It is an appropriate approach to cost sensitive problems as relative cost takes into consideration the cost of Type I (False Negatives) and Type II (False Positives) errors that allows for risk based performance measure. The relative cost of misclassification is calculated as [14] [17]:

$$RC = \alpha(P_I C_I) + (1-\alpha)(P_{II} C_{II}) \qquad (1)$$

where α is the probability of belonging to the bad class, P_I is the probability of being False Negatives and C_I is the relative cost of False Negative. Similarly, P_{II} is the probability of being False Positive and C_{II} is the relative cost of False Positive. Assigning misclassification costs is a challenging and important task in real world [17]. Hence, for this study, to keep the analysis simple cost ratios are used to represent misclassification costs. Given the cost ratios, relative cost is calculated at different thresholds for the models. The best model and threshold is selected as the one that provides the lowest relative cost.

5. Results

As discussed above, a single threshold may not precisely classify the loans with the varying level of risk. Assigning a separate threshold for loans with similar risk behavior could increase the precision in classification. To validate the approach threshold optimization is performed on the full data and on each of the segments obtained from the clustering to obtain lowest relative cost. The threshold optimization is performed with each of the models developed in the modeling stage with different cost ratios. The threshold optimization was performed by selecting 20 random thresholds to select the threshold that gives the lowest relative cost of misclassification. In Figure 1 we can see the effect of threshold selection on relative cost of misclassification for the full data and the segments. Figure 1 depicts the results obtained with GBM model with a cost ratio of 1:3, which represents the cost of False Positives to False Negatives.

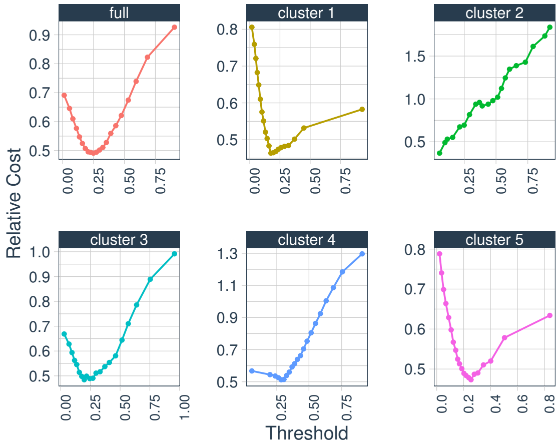

Figure 1. Threshold selection with GBM

Results from Figure 1 presents that the optimal threshold for the lowest relative cost varies across the segments. For the full data, the optimal threshold is at around 0.25 with a relative cost of 0.491. With the individual threshold selection for the segments, except for segment 4, all other segments have lower relative cost than the full data. Hence, threshold selection based on the segments has helped in achieving lower relative cost. For segment 2, the relative cost is very low compared to the full data and other segments and also its optimal threshold is very low. With comparatively very low relative cost for segment 2, it justifies that segmented modeling and threshold selection based on the segments can be significant in lowering the misclassification cost which adds savings to the investment. In addition, the low threshold of 0.055 for segment 2 states that it has high default risk as also seen in cluster summary from Table 2.

The threshold optimization was performed with six different cost ratios for the full data and the segments with all the models developed. The best model was selected as the one that gave the lowest relative cost at each cost ratios. Figure 2 is the summary of the threshold optimization process, where the best model with the lowest relative cost and the corresponding threshold is shown for all the cost ratios. In Figure 2, we can see that GBM is the best model in majority of the cases followed by RF models.

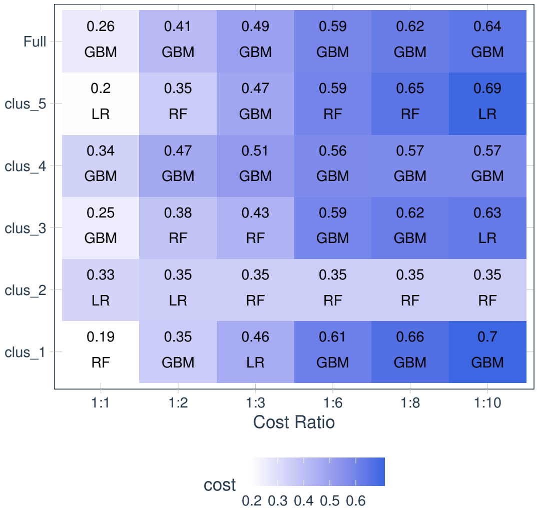

Figure 2. Threshold optimization summary

As the cost ratios increases the threshold decreases to obtain lowest relative cost, which shows high importance given to identify more of the bad loans. The difference in the optimal threshold for each segment at different conditions is visible from Figure 2, that states the relevance of threshold selection based on segmentation. In addition, in most of the cases the segments have lower relative cost than the full data, showing the applicability of the procedure. Furthermore, segment 2 behave very differently than other segments across the cost ratios and has comparatively low relative cost, that can add value to the overall portfolio.

Comparing the relative costs of full data with the relative costs of the segments, there is a decrease in majority of the cases. By taking the average of the relative cost of the individual segments, the average relative cost is lower than the relative cost of the full data, except for the case of cost ratio 1:1. The comparison between the relative cost of full data with respect to the average of the relative costs of the individual segments is presented in Table 6. With the decrease in the relative cost of misclassifcation, the segmented modeling was successful in reducing the relative cost and hence adding some savings in the overall investment. Therefore, the improvement in the relative cost of misclassification through segmented modeling over a single credit scoring model justifies the applicability of segmented modeling and threshold selection.

Table 6. Relative cost comparison

Cost Ratios	Full Data	Segment Average	Reduction (%)
1:1	0.256	0.261	-1.7
1:2	0.406	0.379	6.51
1:3	0.491	0.445	9.46
1:6	0.587	0.539	8.2
1:8	0.617	0.568	7.96
1:10	0.639	0.587	8.19

6. Conclusion

Credit risk analysis has been an imperative part of financial industry and there have been needs for more sophisticated techniques and tools for the accurate evaluation of the risk. Credit scoring is a popular and commonly used technique for evaluating credit risk. It is a common practice to develop a single credit score model from the historical available information to predict the likelihood of default of new applicants for making loan decisions. However, studies have shown that segmented modeling can add precision to overall accuracy of classifying borrowers. Hence, in this study borrower segmentation is performed with the help of K-means clustering and a separate credit scoring model is developed for each segment.

In addition, a separate threshold for each segment is selected to obtain the minimum relative cost of misclassification. The results show that each segments have different risk behavior and hence the optimal threshold varies according to the risk. Furthermore, with the individual threshold selection for the segments there was improvement in the relative cost of misclassification compared to the relative cost for a single credit scoring model. Therefore, the results have shown the relevance of applying segmented modeling and individual threshold selection for improved decision and investment savings. For ensuring the applicability of the study and achieving better results, future research will focus on applying different methods of segmentation along with different models for credit scoring.

References

[1] H. A. Bekhet and S. F. K. Eletter, "Credit risk management for the jordanian commercial banks: a business intelligence approach," 2012.

[2] A. Lahsasna, R. N. Ainon, and Y. W. Teh, "Credit scoring models using soft computing methods: A survey.," *Int. Arab J. Inf. Technol.*, vol. 7, no. 2, pp. 115–123, 2010.

[3] A. Khashman, "Credit risk evaluation using neural networks: Emotional versus conventional models," *Applied Soft Computing*, vol. 11, no. 8, pp. 5477–5484, 2011.

[4] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," in *2014 13th International Conference on Machine Learning and Applications*, pp. 263–269, IEEE, 2014.

[5] A. Ghatge and P. Halkarnikar, "Ensemble neural network strategy for predicting credit default evaluation," *International Journal of Engineering and Innovative Technology (IJEIT) Volume*, vol. 2, pp. 223–225, 2013.

[6] R. Malhotra and D. K. Malhotra, "Evaluating consumer loans using neural networks," *Omega*, vol. 31, no. 2, pp. 83–96, 2003.

[7] T.-S. Lee and I.-F. Chen, "A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines," *Expert Systems with Applications*, vol. 28, no. 4, pp. 743–752, 2005.

[8] M. K. Lim and S. Y. Sohn, "Cluster-based dynamic scoring model," *Expert Systems with Applications*, vol. 32, no. 2, pp. 427–431, 2007.

[9] A. Correa, A. Gonzalez, C. Nieto, and D. Amezquita, "Constructing a credit risk scorecard using predictive clusters," in *SAS Global Forum*, vol. 128, 2012.

[10] S. Scitovski and N. Šarlija, "Cluster analysis in retail segmentation for credit scoring," *Croatian Operational Research Review*, vol. 5, no. 2, pp. 235–245, 2014.

[11] S. Ghanbari, S. Pashazadeh, and H. Bevrani, "Credit risk prediction using clustered classification," 2014.

[12] C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios, "Determinants of default in p2p lending," *PloS one*, vol. 10, no. 10, p. e0139427, 2015.

[13] R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu, "Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending," *Applied Economics*, vol. 47, no. 1, pp. 54–70, 2015.

[14] S. Oreski, D. Oreski, and G. Oreski, "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment," *Expert systems with applications*, vol. 39, no. 16, pp. 12605–12617, 2012.

[15] H. Ince and B. Aktan, "A comparison of data mining techniques for credit scoring in banking: A managerial perspective," *Journal of Business Economics and Management*, vol. 10, no. 3, pp. 233–240, 2009.

[16] A. Blanco, R. Pino-Mejías, J. Lara, and S. Rayo, "Credit scoring models for the microfinance industry using neural networks: Evidence from peru," *Expert Systems with applications*, vol. 40, no. 1, pp. 356–364, 2013.

[17] S. Oreški and G. Oreški, "Cost-sensitive learning from imbalanced datasets for retail credit risk assessment," *TEM JOURNAL-Technology, Education, Management, Informatics*, vol. 7, no. 1, pp. 59–73, 2018.

[18] M. Bakoben, T. Bellotti, and N. Adams, "Identification of credit risk based on cluster analysis of account behaviours," *arXiv preprint arXiv:1706.07466*, 2017.

[19] Y. Peng, G. Kou, Y. Shi, and Z. Chen, "Improving clustering analysis for credit card accounts classification," in *International Conference on Computational Science*, pp. 548–553, Springer, 2005.

[20] M. Polena and T. Regner, "Determinants of borrowers default in p2p lending under consideration of the loan risk class," *Games*, vol. 9, no. 4, p. 82, 2018.

[21] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, and N. Villa-Vialaneix, “Random forests for big data,” *Big Data Research*, vol. 9, pp. 28–46, 2017.

[23] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, “Random forest: a classification and regression tool for compound classification and qsar modeling,” *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[24] M. Pal, “Random forest classifier for remote sensing classification,” *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[25] J. H. Friedman, “Stochastic gradient boosting,” *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[26] R. Lawrence, A. Bunn, S. Powell, and M. Zambon, “Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis,” *Remote sensing of environment*, vol. 90, no. 3, pp. 331–336, 2004.

[27] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[28] K. R. Žalik, “An efficient k-means clustering algorithm,” *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1385–1391, 2008.

[29] M. Erisoglu, N. Calis, and S. Sakallioglu, “A new algorithm for initial cluster centers in k-means algorithm,” *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1701–1705, 2011.