Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Reassessment and Monitoring of Loan Applications with Machine Learning

Zeynep Boz, Dilek Gunnec, S. Ilker Birbil & M. Kaan Öztürk

To cite this article: Zeynep Boz, Dilek Gunnec, S. Ilker Birbil & M. Kaan Öztürk (2018) Reassessment and Monitoring of Loan Applications with Machine Learning, Applied Artificial Intelligence, 32:9-10, 939-955, DOI: 10.1080/08839514.2018.1525517

To link to this article: https://doi.org/10.1080/08839514.2018.1525517

Published online: 04 Oct 2018.

APPLIED ARTIFICIAL INTELLIGENCE
2018, VOL. 32, NOS. 9–10, 939–955
https://doi.org/10.1080/08839514.2018.1525517

Taylor & Francis
Taylor & Francis Group

Reassessment and Monitoring of Loan Applications with Machine Learning

Zeynep Boz[a], Dilek Gunnec[b], S. Ilker Birbil[c], and M. Kaan Öztürk[d]

[a]Industrial Engineering Program, Sabancı University, Istanbul, Turkey; [b]Department of Industrial Engineering, Özyeğin University, Istanbul, Turkey; [c]Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands; [d]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

ABSTRACT

Credit scoring and monitoring are the two important dimensions of the decision-making process for the loan institutions. In the first part of this study, we investigate the role of machine learning for applicant reassessment and propose a complementary screening step to an existing scoring system. We use a real data set from one of the prominent loan companies in Turkey. The information provided by the applicants form the variables in our analysis. The company's experts have already labeled the clients as bad and good according to their ongoing payments. Using this labeled data set, we execute several methods to classify the bad applicants as well as the significant variables in this classification. As the data set consists of applicants who have passed the initial scoring system, most of the clients are marked as good. To deal with this imbalanced nature of the problem, we employ a set of different approaches to improve the performance of predicting the applicants who are likely to default. In the second part of this study, we aim to predict the payment behavior of clients based on their static (demographic and financial) and dynamic (payment) information. Furthermore, we analyze the effect of the length of the payment history and the staying power of the proposed prediction models.

Introduction

In banking and finance, credit scoring is one of the most important dimensions of decision-making process. To maximize their return and minimize their financial risk, loan companies employ certain credit scoring methods for evaluating the default risk of the applicants (Hand and Henley 1997). One basic and widely-adopted method is the use of score-cards. With these cards, a loan applicant receives a score according to her various features like home ownership, occupation and credit history. If the scores of the applicant are above some predetermined thresholds, then she is granted the loan (Anderson and Thompson 2009). As soon as the borrower is in the system, the company follows up her repayment performance and takes related actions when necessary. As

CONTACT Dilek Gunnec ✉ dilek.gunnec@ozyegin.edu.tr 🖃 Department of Industrial Engineering, Özyeğin University, Çekmeköy, 34794 Istanbul, Turkey

more data becomes available with the advent of large information systems, it is now possible to assist both the loan-granting and the payment-monitoring decisions with the recent machine learning tools (Zurada and Zurada 2011).

In this study, we collaborate with Koçfinans, a consumer financing company in Turkey. Koçfinans has around two million active clients using car, mortgage and home equity loans. To evaluate loan applications, Koçfinans also uses their own score-cards. Applicants are classified according to their demographic, personal and financial information that they provide when they file their application. Moreover, Koçfinans also purchases the credit score of an applicant from the Credit Bureau (KKB), which is a countrywide private company assigning credit scores to the individuals (KKB 2017). The main objective of the current study is to apply machine learning methods to assist the reassessment of the applicants and the monitoring of the existing clients at Koçfinans. We first discuss several complementary tools that could be used for second screening of the applicants after the initial screening with the score-card system. Then, we add the time-series payment data of the clients to predict their future payment behavior. Overall, our study can be considered as a new decision support system for Koçfinans. Figure 1 illustrates how our study improves the current system with the addition of two new stages; namely the applicant reassessment and the client monitoring.

With this work, we make the following contributions. We argue that machine learning methods could play an important role for both applicant assessment and client monitoring. To support our argument, we illustrate our findings on a real-life data set. Other studies in the literature on applicant assessment propose a complete replacement or a major change of the existing systems (Mandala, Nawangpalupi, and Praktikto 2012; Yap, Ong, and Husain 2011). Such a drastic

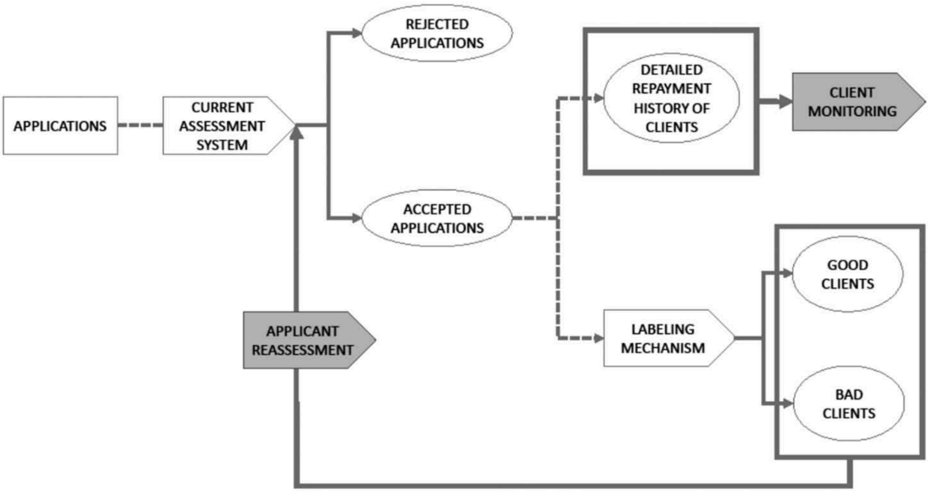


Figure 1. The flow chart of the current system and the proposed extensions (shaded).

change is mostly difficult to implement due to high costs and likely resistance within the institution. Our study, on the other hand, proposes a complementary stage for the ongoing applicant assessment process. Therefore, it is relatively easy to streamline it with the existing systems. Since the data for a second screening stage involves a selective set of customers, there is a significant difference between the default and the non-default rates on the loans. We further discuss several approaches to handle this imbalance in the labels when a supervised machine learning method is used. After a loan is granted to a client, machine learning methods can again be used to predict her payment status in the coming periods. When accurate, these predictions can easily lead to significant cost savings for the loan-granting institutions. As it is common for all prediction models, two questions remain to be answered for improving the accuracy: (i) How far in the past should one go for data collection? (ii) How frequently the models should be trained? In the last part of our study, we aim at answering both questions with our computational study.

Related literature

Credit Scoring research has a 40 years history with major developments in recent years. In a comprehensive literature survey Thomas (2000) provide a broad perspective on how economic conditions can be considered in forecasting the financial risk of lending to customers and of maximizing profit rather than minimizing default risk. They review statistical and operational research based techniques for both credit and behavioral scoring. In a later review, Crook, Edelman, and Thomas (2007) focus on banking sector and review the basics of credit scoring and classification approaches. They suggest that statistical methods, linear programming and neural networks models are the most commonly used methodologies. There is a wide range of literature on credit scoring applications by machine learning techniques as there is access to abundant information. Such studies identify necessary information to assess a credit application by using decision tree models (Mandala, Nawangpalupi, and Praktikto 2012; Yap, Ong, and Husain 2011), logistic regression (Yap, Ong, and Husain 2011), *k*-means clustering and support vector machines (Chen et al. 2012), stochastic gradient boosting and random forest model (Thornhill 2008). Applications are usually problem specific and methods perform distinctly in compliance with the data.

Behavioral scoring, i.e., assessing the likelihood of client default by dynamic information, requires larger amount of data compared to credit scoring, thus there is less attention to it in literature from both financial institutions and scholars (Kennedy et al. 2013). For traditional methods, Thomas, Ho, and Scherer (2001) provide a review on behavioral, customer and profit scoring methodologies including Markov chain processes, segmentation models and survival analysis. More recently, machine learning and data mining techniques are used for behavioral scoring. Hsieh (2004) focus on understanding behavioral patterns of credit card customers of a bank. Using segmentation analysis, they are able to cluster customers and assign association rules for each cluster. Sarlija, Bensic, and Zekic-Susac (2009) compare performances of survival analysis and neural network method for a data set on a 6-months period. They identify importance of misclassification cost sensitivity for Type I and Type II errors in both models with different cost ratios, and conclude that for earlier observation periods and for the equal cost ratios for both types of errors, survival models work better than neural network models, whereas neural network models are more effective for the longer time periods. Kennedy et al. (2013) focus on changing performance monitoring periods in behavioral scoring. They conclude that it is easier (in terms of data preparation) and better to consider a 12-month period compared to using altering time periods. They predict default probabilities by employing logistic regression model.

Class imbalance is common when working with data sets where the majority of data inherently will belong to one particular class. For example, with churn data (similarly, with fraudulence or disease related data), the proportion of churners is expected to be smaller than non-churners. To overcome problems with such imbalance, most of the studies focus on the sampling methods (over- or under-sampling) for generating suitable data sets for the standard classification algorithms (Burez and Van den Poel 2009; Crone and Finlay 2012; Zhou and Wang 2012). Other studies use random forest algorithm (Chen, Liaw, and Breiman 2004), adapting the random forest algorithm by assigning weights to decision trees in the forest (Zhou and Wang 2012), and adaptive boosting and bagging algorithms (Galar et al. 2012). Knowledge discovery with learning algorithms using skewed data is another challenge. He and Garcia (2009) review existing solutions on sampling, cost-sensitive learning models, and kernel-based and active learning models. They propose that instead of the singular assessment measures, such as accuracy and error rate which may be misleading due to sensitivity to data changes and distribution of classes, receiver operating characteristic curve, precision-recall curves and cost curves are more efficient and unbiased with imbalance learning. Similarly, measures focusing on the incorrect classification of the minority class, such as recall and negative predictive value (NPV) are also found to be more suited for imbalanced data sets (Burez and Van den Poel 2009; Yap, Ong, and Husain 2011). More specifically, King and Zeng (2001) focus on rare events such as wars and epidemiological infections which have tens of times smaller number of observations than the non-events. They use logistic regression model together with data collection strategies to overcome the drawback of the model underestimating the rare events.

Data set

The data set of Koçfinans consists of 110,000 unique accounts for personal car loans. Around 60% of the accounts are closed. An applicant provides demographic, personal and financial information during application. In addition, Koçfinans acquires a score (KKB score) about the past financial activities of an applicant through the Credit Bureau. If the applicant has been granted a loan in the past, then the data in her application folder is also used to create some additional features. We include only this initial set of data (no payment information) in the applicant reassessment part of our study. We use the labels "good" and "bad" as our target variables for each account. These labels are given by the experts at Koçfinans. As the clients are already screened by the score-cards these labels are significantly imbalanced in favor of the good clients. That is, approximately 94% of all clients are labeled as good where the remaining 6% of the clients are labeled as bad. Figure 2 illustrates this imbalanced nature of the data for several features. Although we know that approximately one-fourth of the applications are accepted, unfortunately, we do not have access to the data of the rejected applications as this data cannot be stored by law.

The payment periods of the considered accounts are between 2012 and 2015. In the client monitoring part of our study, we include the payment information data in addition to the data in the application folder. This data brings in additional features, such as; the approved loan amount, the loan term, the advance payment amount, the return rate (calculated as the ratio of the loan on interest to loan amount) and the number of guarantors. Table 1 lists in detail the set of attributes that we have used. For client monitoring, we use the payment statuses of the accounts as our target variables. There are in total 15 different labels corresponding to the payment status codes indicating time of payment for customers. We consider a subset of the clients who have been active for at least 12 payment periods. Therefore, the size of the data set considered in the second part of our study involves approximately 60,000 clients.

As KKB score is also received from the Credit Bureau, one may consider whether this score by itself is sufficient for the assessment of a client as good or bad. Figure 3 illustrates the KKB score distributions within the good and bad classes. A two-sided Kolmogorov–Smirnov test shows that the good and bad KKB scores are almost certainly from different distributions.

Our simulation experiments are conducted on a workstation with Intel (R) i5-5200U CPU@2.2 GHz and 4 GB RAM running on Windows 10 operating system. We employ R programming language and software version 3.2.3 (Wooden Christmas-Tree). We used a set of R functions that we have listed in Appendix A.

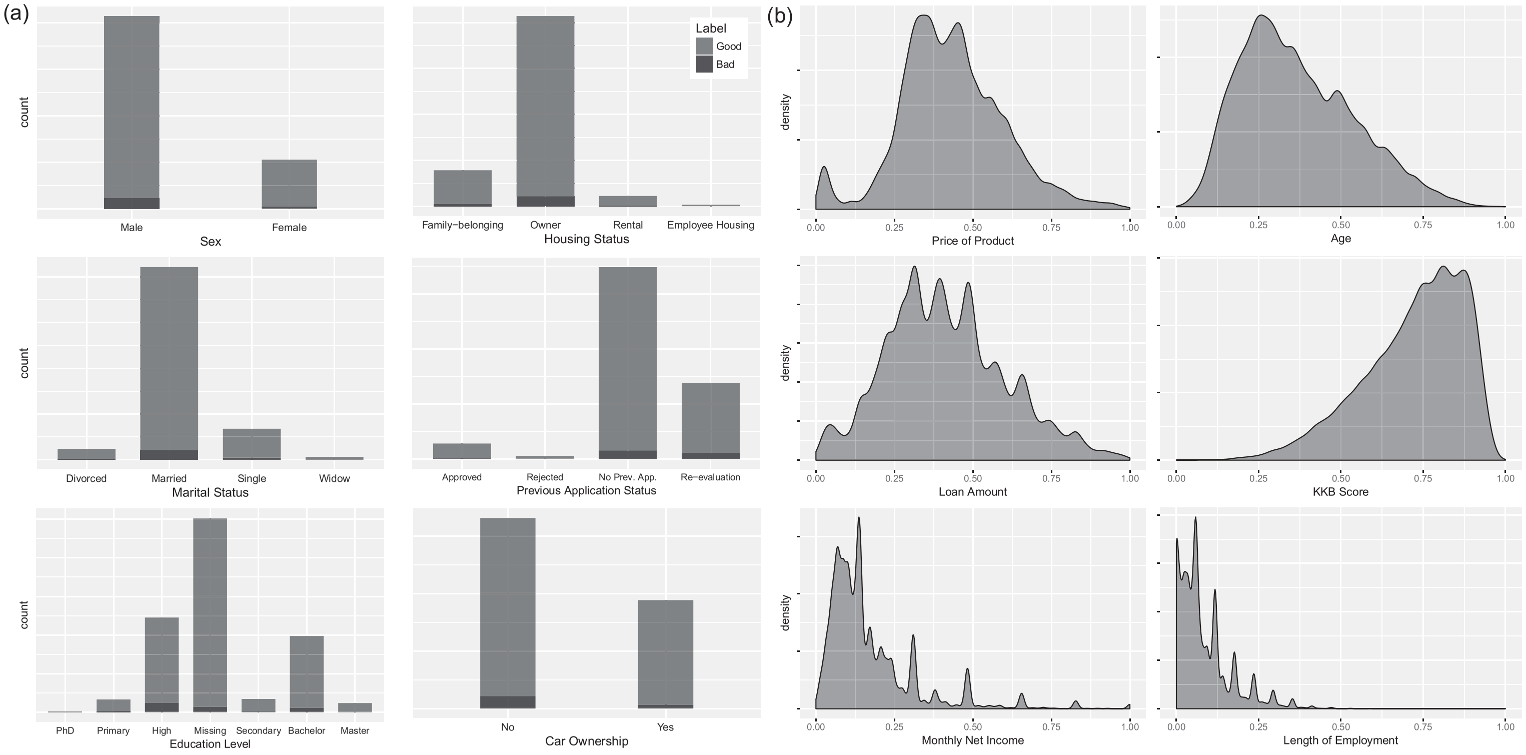


Figure 2. (Left) Normalized distributions of some of the numerical variables. (Right) Good and bad label distributions for some of the categorical variables. Count numbers are not disclosed due to the privacy concerns of the company.

Table 1. List of variables used in the analysis.

Variable type	Attribute name	Applicant reassessment	Client monitoring
Numerical	Price of the Product	■	■
	Loan Amount	□	■
	Monthly Net Income	■	■
	KKB Score	■	■
	Age	■	■
	Length of Employment	■	■
	Number of Guarantors	□	■
	Return Rate	□	■
	Advance Payment Amount	□	■
	Loan Term	□	■
Categorical	Sex	■	■
	Marital Status	■	■
	Education Level	■	■
	Housing Status	■	■
	Previous Application Status	■	■
	Product Type Code	■	■
	Payment Status Code[a]	□	■
Binary	Car Ownership	■	■
	Good-Bad Labels[b]	■	□

[a]Target variable for Client Monitoring.
[b]Target variable for Applicant Reassessment.

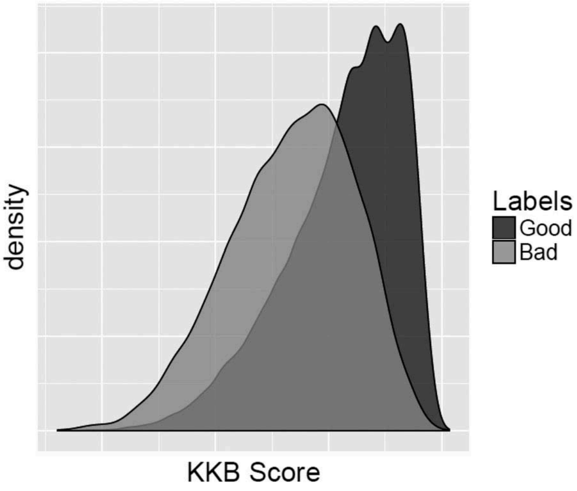


Figure 3. KKB scores and the labels.

Applicant reassessment

In the first part of our study, we discuss a secondary screening system for classifying the applicants as good or bad. This effort is significant for the company as it provides a complementary process that minimizes false positive errors (treating bad clients as good). Such a process also plays an important role, particularly when applicants with limited information need to be classified. Overall, the proposed applicant reassessment system aids in reducing the risk for the company.

As mentioned in Section “Data set”, the imbalanced nature of the data causes biased classification results with some of the standard methods when simple random sampling is preferred. To overcome this obstacle, our first attempt is to use *cost sensitive decision tree models* (Chen, Ribeiro, and Chen 2015; He and Garcia 2009), which give higher importance to false classification of minority classes (bad labels in our case). We also select two classification models, random forests and adaptive boosting, which are known to be sensitive to the imbalanced class distribution (Chen, Liaw, and Breiman 2004; Galar et al. 2012).

To report the performances of different methods, we focus on the NPV measure which focuses on the incorrect classification of the minority class as it is more suited for studies on imbalanced data sets (Burez and Van den Poel 2009; Yap, Ong, and Husain 2011). NPV is the ratio of the number of true negative (bad) predictions to the total number of negative predictions. From the company’s point of view, NPV simply reflects the trade-off between the gain from rejecting the bad customers and the loss of turning down good customers.

In our analysis, we have used a set of standard machine learning methods. Logistic regression (LR) and decision tree (DT) are among the frequently used methods for classification in credit scoring (Yap, Ong, and Husain 2011). Support vector machine (SVM) and random forest (RF) methods have also been studied in the context of credit scoring (Baesens et al. 2003; Huang, Chen, and Wang 2007; Sharma 2009; Thomas, Oliver, and Hand 2005). Adaptive boosting (AB) is another ensemble method that aims to improve the DT method by correcting the errors to get higher classification performance. Our last method is the Naïve Bayes (NB) classifier. Table 2 shows the results that we have obtained with these methods. We have reserved 70% of the data for training and the rest for the testing. Note that we have two more methods in the table, DTCS(3:1) and DTCS(5:1). These are cost-sensitive DT methods, where predicting the bad clients incorrectly (false negative) is penalized and given a three and five times higher cost than predicting the good clients incorrectly (false positive). The first three rows of Table 2 show that assigning costs to

Table 2. Performance results for applicant reassessment.

	Accuracy	Sensitivity	Specificity	F1	AUC	NPV	PBN[a]
DT	0.945	1.000	0.000	0.972	0.500	–	0
DTCS(3:1)	0.942	0.995	0.018	0.970	0.507	0.178	185
DTCS(5:1)	0.907	0.952	0.139	0.951	0.545	0.142	1771
RF	0.944	0.999	0.004	0.972	0.502	0.412	16
NB	0.938	0.991	0.040	0.968	0.515	0.192	373
AB	0.926	0.976	0.078	0.962	0.527	0.153	919
LR	0.944	0.999	0.005	0.972	0.502	0.209	40
SVM	0.945	1.000	0.000	0.972	0.500	–	0

[a] Number of applicants predicted as bad.

misclassification improves the performance of the regular DT method, which simply labels all customers as good and obtains high accuracy due to the label imbalance in the data.

The results in Table 2 show that DT and SVM methods can be easily ruled out as they are not capable of predicting the bad applicants. RF model results in relatively high NPV performance. However, it classifies only 16 applicants as bad among 1800 applicants in the test set. If the company is willing to reassess a large number of applications, then the other models may be used. In that case, NB seems like a reasonable choice as almost 20% of the applicants that it labels as negative turn out to be bad clients.

From a practitioner's point of view, another important point is to learn the key variables, which can be used to identify the bad applicants. Figure 4 represents the relative importance of the variables that we have obtained from the RF method, which is the most successful method in terms of NPV. KKB score and age are the two most important variables, whereas the sex variable has the least importance. We note that the distribution of good and bad labels among female applicants is close to those of male applicants (see also Figure 2).

Koçfinans is already aware of the key role KKB score plays in credit scoring. Therefore, with the intention to further eliminate bad customers they usually focus on a subset of customers with KKB scores below a certain threshold. When they do (for a given threshold), the percentage of bad labels in the filtered subset becomes almost twice the original and raises up to 12%. We have run all the methods with this subset and obtained the results in Table 3. These results show that although the accuracies of the methods may deteriorate, NPV measures increase for most of the methods. Moreover,

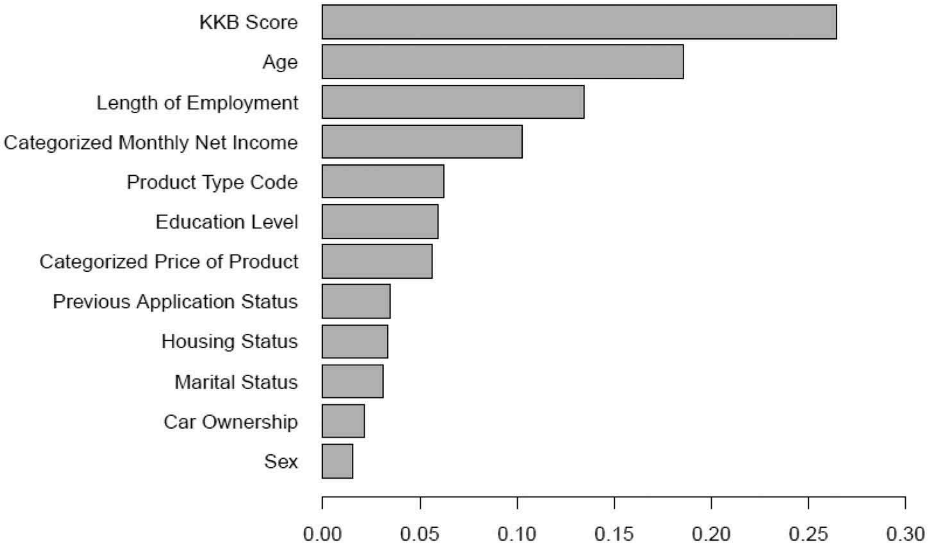


Figure 4. Relative importance of the variables.

Table 3. Performance results for applicant reassessment on a subset of data filtered by KKB score.

	Accuracy	Sensitivity	Specificity	F1	AUC	NPV	PBN[a]
DT	0.885	1.000	0.000	0.939	0.500	–	0
DTCS(3:1)	0.829	0.915	0.171	0.904	0.543	0.210	592
DTCS(5:1)	0.713	0.748	0.448	0.822	0.598	0.190	1719
RF	0.884	0.999	0.009	0.938	0.504	0.472	17
NB	0.881	0.993	0.025	0.937	0.515	0.300	61
AB	0.848	0.944	0.119	0.917	0.532	0.216	397
LR	0.877	0.986	0.045	0.934	0.516	0.297	110
SVM	0.885	1.000	0.000	0.390	0.500	–	0

[a]Number of applicants predicted as bad.

except for DT and SVM, these figures are all higher than simple random selection (which would have yielded 0.12 NPV performance, on average).

Client monitoring

In the second part of our study, we discuss how to predict the payment behavior of clients in the upcoming periods. In addition to the data used in the first part, we also use the time-series payment data of each client (see also Table 1).

In our data, clients are already categorized into 15 types of status code records, such as; *late payment, normal payment, early account closure, legal proceedings, advice note*, etc. These categories are time-dependent and updated dynamically at each payment period. As shown in Figure 5, we develop three modeling approaches to predict the payment behavior in the future. In (i), we employ the complete payment history of a client. We consider in (ii) the effect of the length of payment history by using the previous status labels consecutively (Ballings and Van den Poel 2012). This analysis shows how many past payment periods we should consider to build well-performing models for prediction. In (iii), we aim to measure the staying power of the models. That is, we investigate how far into the future we can predict by using the same model without a decline in its performance (Risselada, Verhoef, and Bijmolt 2010).

We apply machine learning methods by reserving 70% of our data as the training set and the rest as the testing set. Table 4 shows our results for model (i) in terms of accuracy and AUC. In this respect, SVM and RF models have the best performance for predicting the status code of the 13th period by using the status codes of the first 12 periods.

Figure 6 illustrates the relative importance of the variables. As expected, the status codes in the most recent two periods have higher weights. These are followed by the loan amount and the financial means of the client like net income and KKB score.

The results of model (ii), where we investigate the effect of the length of the event history, are presented in Figure 7. Recall that we use the first 12

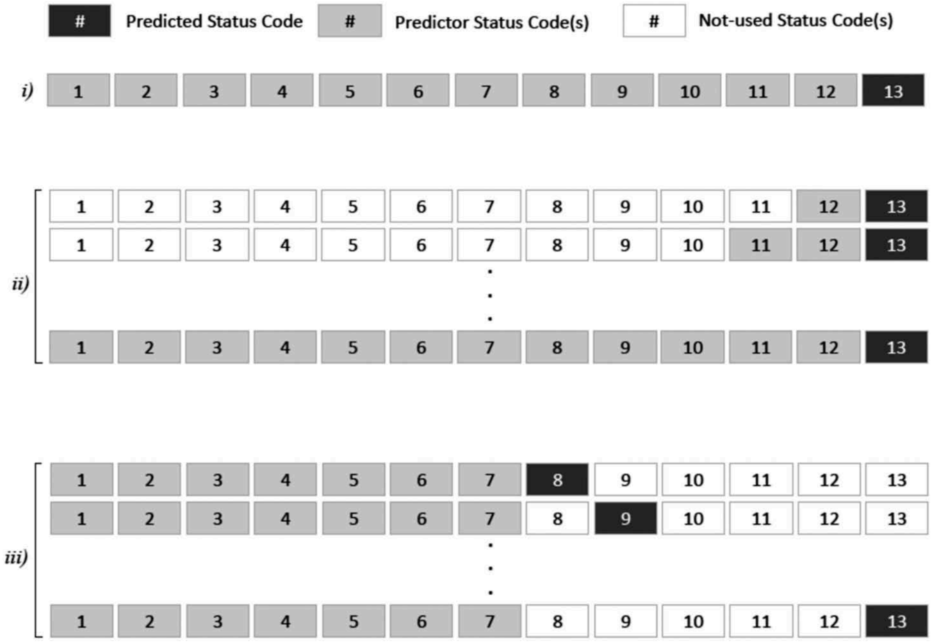


Figure 5. Time-line illustrations of the client monitoring models. (i) Predicting 13th status code by using the previous 12 status codes. (ii) Predicting the 13th status code by using the previous n status codes with $n \leq 12$. (iii) Predicting the $(7 + k)$th status code with $1 \leq k \leq 6$.

Table 4. Performance results for the client monitoring model (i).

	Accuracy	AUC
DT	0.647	0.735
SVM	0.652	0.761
NB	0.539	0.652
RF	0.651	0.770
AB	0.591	0.695
LR	0.636	0.715

payment status code records and by going cumulatively backwards in time, we try to figure out how many time periods are needed to obtain accurate results. Using less data into the past is crucial for the company as they prefer to reduce the time spent for considering the complete history of the client, which requires time- and resource-consuming data retrieval process. Figure 7 shows that it is sufficient to use the past four periods for DT, RF and SVM methods. For these three methods, it may not be worth going further into the past as the increase in the performance is quite limited. On the other hand, performances of LR and NB methods drop, if the data from further in the past is included. AB model shows a slightly fluctuating behavior within the observed time interval. These figures imply that RF method is the best choice for the client monitoring model (ii), since it reaches the highest AUC value by only using the past three periods.

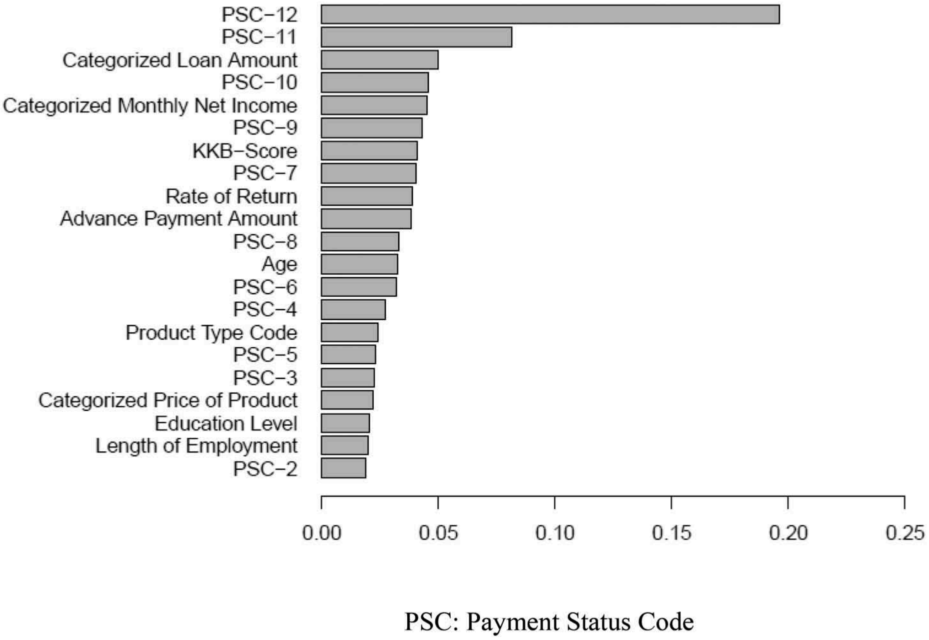
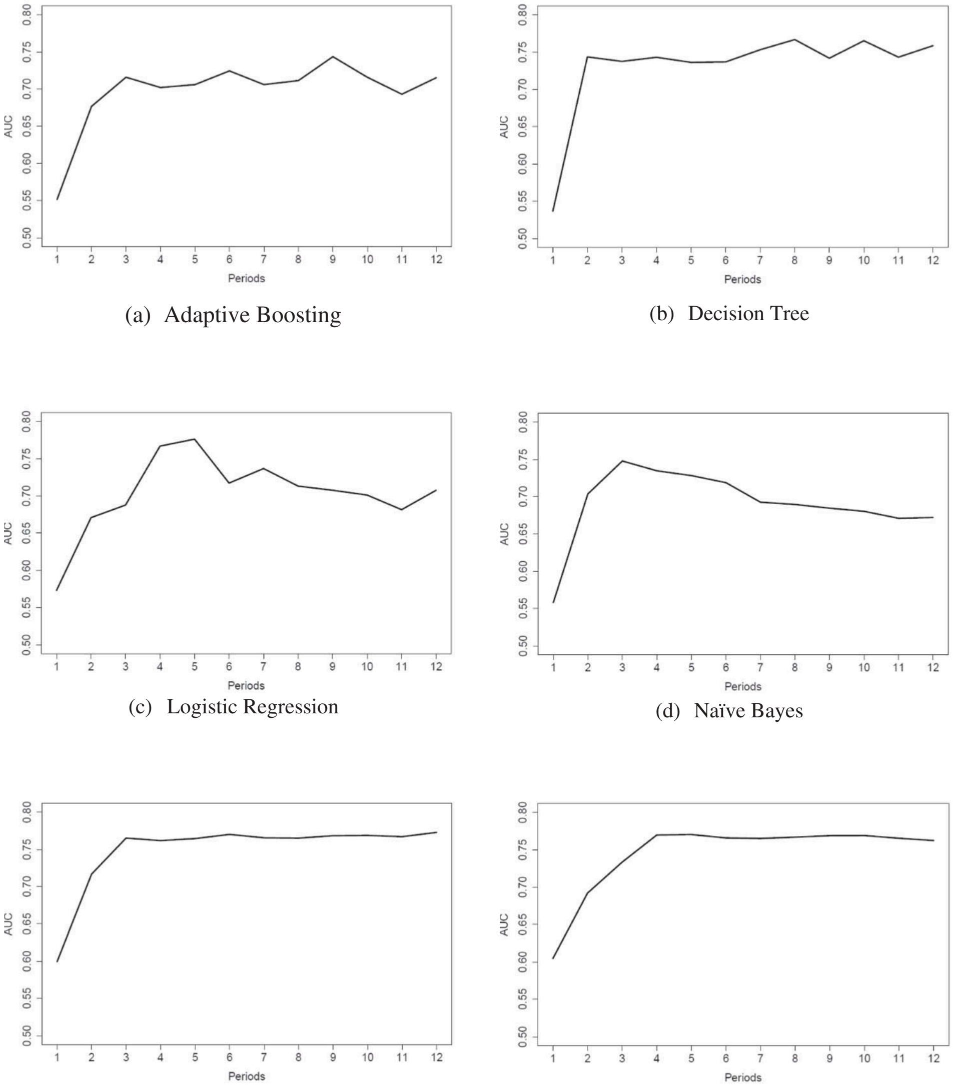


Figure 6. Relative importance of the variables for the client monitoring model (i).

In our last experiment, we consider the staying power of the machine learning methods through client monitoring model (iii). Unlike model (ii), this analysis takes predictor set as constant but changes the predicted status code one by one. Our results are summarized in Figure 8. We observe that most of the methods can only stand for at most two months before they need to be trained again. Though NB method seems to have a more stable predictive power, its performance in terms of AUC measure is not as good as most of the other methods. The performance of LR method stays almost constant for three months with relatively high AUC values. These results suggest that RF and SVM can be used with acceptable performances when the models are re-trained every other 2 months. Recall from Figure 6 that the recent payment behaviors are the most important variables. This is in line with our current observation regarding the rather short length of the staying power.

Conclusion

With abundance of customer data, banks and financial institutions are ambitious about using data analytics to minimize their overall risk. In this study we develop a decision support system that can support such institutions for both application assessment and client monitoring. For the new loan applications with limited and static data, it is difficult but critical to correctly identify the risk of defaults. In the first part, using a set of machine learning methods, we propose



(a) Adaptive Boosting

(b) Decision Tree

(c) Logistic Regression

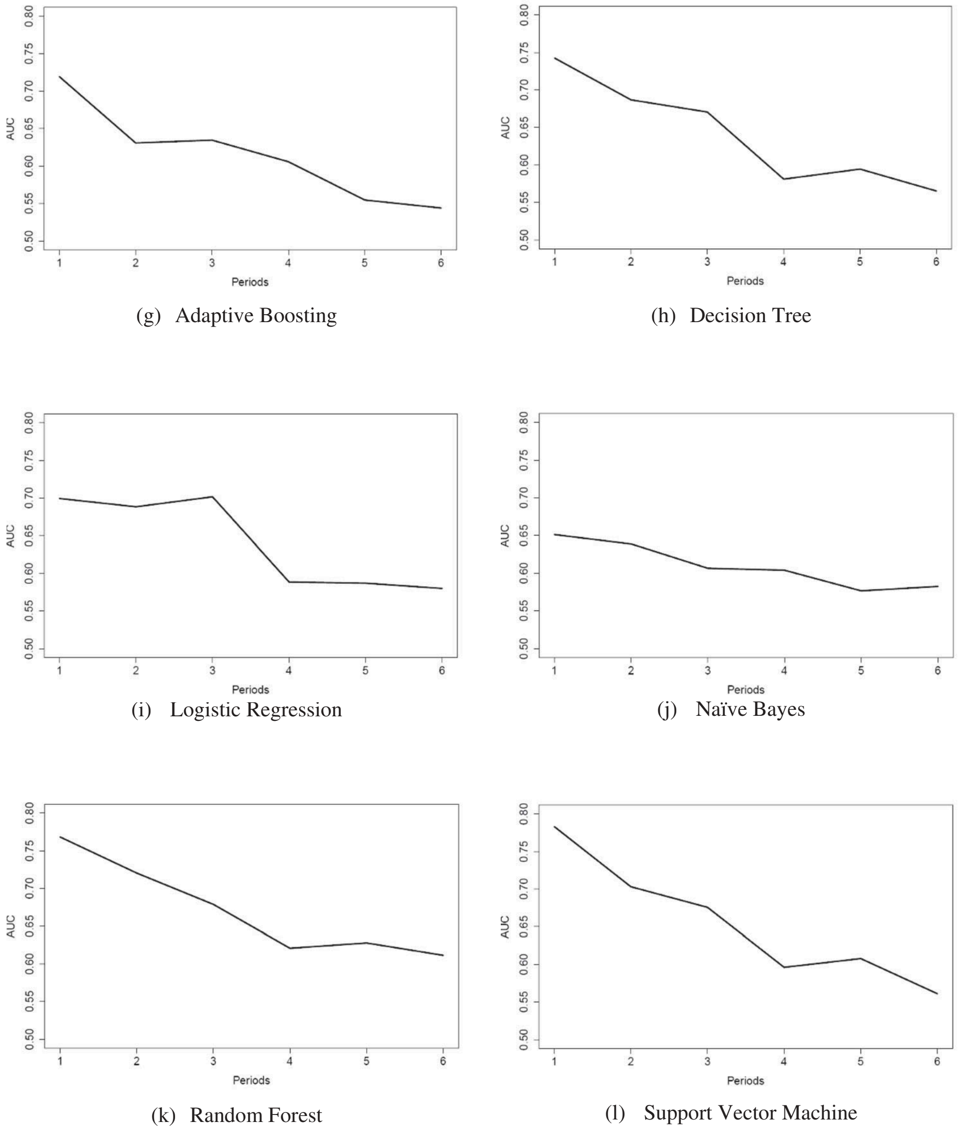
(d) Naïve Bayes

(e) Random Forest

(f) Support Vector Machine

Figure 7. The effect of the length of payment history on the performances of the models for the client monitoring model (ii).

models that are able to point the institution toward re-evaluating a set of prescreened customers. Among these models, a suitable model can be chosen depending on the institution's willingness (or available resources) for re-evaluation. Monitoring existing customers allows taking early actions toward potential default risks. This involves working on dynamic information on customer payment data. In the second part, we propose modeling approaches to predict the customer behavior using previous customer status records. We also investigate how far into the past one should look for data collection and how often the



(g) Adaptive Boosting

(h) Decision Tree

(i) Logistic Regression

(j) Naïve Bayes

(k) Random Forest

(l) Support Vector Machine

Figure 8. The staying powers of the models for the client monitoring model (iii).

models need to be re-trained. Overall, we observe that support vector machine and random forest are two high performing and efficient methods for customer monitoring.

References

Alfaro, E., M. Gámez, and N. García. 2013. Adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software* 54 (2):1–35. doi:10.18637/jss.v054.i02.

Anderson, B. S., and R. W. Thompson. 2009. *Developing credit scorecards using SAS credit scoring for enterprise miner 5.3*. Cary: SAS Institute Inc.

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* 54 (6):627–35. doi:10.1057/palgrave.jors.2601545.

Ballings, M., and D. Van den Poel. 2012. Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications* 39 (18):13517–22. doi:10.1016/j.eswa.2012.07.006.

Burez, J., and D. Van den Poel. 2009. Handling class imbalance in customer churn prediction. *Expert Systems with Applications* 36 (3):4626–36. doi:10.1016/j.eswa.2008.05.027.

Chen, C., A. Liaw, and L. Breiman. 2004. *Using random forest to learn imbalanced data*. Berkeley: University of California. 110. http://statistics.berkeley.edu/sites/default/files/techreports/666.pdf.

Chen, N., B. Ribeiro, and A. Chen. 2015. Comparative study of classifier ensembles for cost-sensitive credit risk assessment. *Intelligent Data Analysis* 19 (1):127–44.

Chen, W., G. Xiang, Y. Liu, and K. Wang. 2012. Credit risk evaluation by hybrid data mining technique. *Systems Engineering Procedia* 3:194–200. doi:10.1016/j.sepro.2011.10.029.

Crone, S. F., and S. Finlay. 2012. Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting* 28 (1):224–38. doi:10.1016/j.ijforecast.2011.07.006.

Crook, J. N., D. B. Edelman, and L. C. Thomas. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research* 183 (3):1447–65. doi:10.1016/j.ejor.2006.09.100.

Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. 2012. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42 (4):463–84. doi:10.1109/TSMCC.2011.2161285.

Hand, D. J., and W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160 (3):523–41. doi:10.1111/rssa.1997.160.issue-3.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21 (9):1263–84. doi:10.1109/TKDE.2008.239.

Hsieh, N.-C. 2004. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert Systems with Applications* 27 (4):623–33. doi:10.1016/j.eswa.2004.06.007.

Huang, C.-L., M.-C. Chen, and C.-J. Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications* 33 (4):847–56. doi:10.1016/j.eswa.2006.07.007.

Kennedy, K., B. MacNamee, S. J. Delany, M. O'Sullivan, and N. Watson. 2013. A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications* 40 (4):1372–80. doi:10.1016/j.eswa.2012.08.052.

King, G., and L. Zeng. 2001. Logistic regression in rare events data. *Political Analysis* 137–63. doi:10.1093/oxfordjournals.pan.a004868.

KKB (2017). Credit bureau. Accessed September 25, 2017. http://www.kkb.com.tr.

Kuhn, M., S. Weston, and N. Coulter, and code for C5.0 by R. Quinlan, M. C. C. (2015). C50: C5.0 Decision trees and rule-based models. R package version 0.1.0-24.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomforest. *R News* 2 (3):18–22.

Mandala, I. G. N. N., C. B. Nawangpalupi, and F. R. Praktikto. 2012. Assessing credit risk: An application of data mining in a rural bank. *Procedia Economics and Finance* 4:406–12. doi:10.1016/S2212-5671(12)00355-3.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2015. e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. *R Package Version* 1.6–7.

R Core Team. 2015. *R: A Language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Risselada, H., P. C. Verhoef, and T. H. Bijmolt. 2010. Staying power of churn prediction models. *Journal of Interactive Marketing* 24 (3):198–208. doi:10.1016/j.intmar.2010.04.002.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. 2011. proc: An open-source package for R and S+ to analyze and compare roc curves. *BMC Bioinformatics* 12:77. doi:10.1186/1471-2105-12-77.

Sarlija, N., M. Bensic, and M. Zekic-Susac. 2009. Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications* 36 (5):8778–88. doi:10.1016/j.eswa.2008.11.042.

Sharma, D. (2009). Guide to credit scoring in R. https://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf.

Therneau, T., B. Atkinson, and B. Ripley. 2015. RPART: Recursive partitioning and regression trees. *R Package Version* 4.1–10.

Thomas, L., R. Oliver, and D. Hand. 2005. A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society* 56 (9):1006–15. doi:10.1057/palgrave.jors.2602018.

Thomas, L. C. 2000. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting* 16 (2):149–72. doi:10.1016/S0169-2070(00)00034-0.

Thomas, L. C., J. Ho, and W. T. Scherer. 2001. Time will tell: Behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics* 12 (1):89. doi:10.1093/imaman/12.1.89.

Thornhill, O. (2008). Credit risk evaluation of online personal loan applicants: A data mining approach. http://www.bisolutions.us/Credit-Risk-Evaluation-of-Online-Personal-Loan-Applicants-AData-Mining-Approach.php.

Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. fourth ed. New York: Springer.

Yap, B. W., S. H. Ong, and N. H. M. Husain. 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications* 38 (10):13274–83. doi:10.1016/j.eswa.2011.04.147.

Zhou, L., and H. Wang. 2012. Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering* 10 (6):1519–25. doi:10.11591/telkomnika.v10i6.1323.

Zurada, J., and M. Zurada. 2011. How secure are good loans: Validating loan-granting decisions and predicting default rates on consumer loans. *Review of Business Information Systems (RBIS)* 6 (3):65–84. doi:10.19030/rbis.v6i3.

Appendix A Software

Below is a list of R functions that we have used in our experiments:

- Logistic Regression: glm function with binomial family and logit link parameters (R Core Team 2015).
- Multinomial Logistic Regression: multinom function from nnet library (Venables and Ripley 2002).

- Decision Tree: rpart package (Therneau, Atkinson, and Ripley 2015).
- Cost Sensitive Decision Tree: C50 library (Kuhn, Weston, and Coulter 2015).
- Naïve Bayes: naiveBayes function from e1071 library (Meyer et al. 2015).
- Support Vector Machine: svm function from e1071 library (Meyer et al. 2015).
- Adaptive Boosting: boosting function from adabag library (Alfaro, Gámez, and García 2013).
- Random Forest: randomForest library (Liaw and Wiener 2002).

Moreover, the area under curve evaluations for multiple classes are obtained with multiclass. roc from pROC package (Robin et al. 2011).