The 5th FINACT-IRAN National Conference on Financial and Actuarial Mathematics
December 22-25, 2018

Credit risk of bank loan case study in Iran

Amirsalar VahidiAsgari[1]
MSc Student, Amirkabir University of Technology, Tehran, Iran

Seyede Niloofar Ebrahimi
MSc Student, Amirkabir University of Technology, Tehran, Iran

Dr.Erfan Salavati
Assistant professor, Amirkabir University of Technology, Tehran, Iran

Abstract

One of the most common indicators that used to identify credit risk is the ratio of non-performing loans (NPL). One major issue with granting loans is whether the borrowers could fulfill their obligation or not.The purpose of the present study is to identify the factors affecting the NPL of a bank for the period 2013-2017. We test machine learning methods against traditional methods on a collected dataset and the results proves that Random Forest method works better among other techniques such as logistic regression, SVM, ANN and etc. And also decision tree is used as a rule extraction method.

Keywords: Non-performing loan, Credit risk, Data mining.

AMS Mathematical Subject Classification [2018]: 13D45, 39B42

1 Introduction

Non performing loan indicator is a crucial indicator for banks. To ensure that bank gives loans to right persons, credit assessment is a critical decision making process.

Credit risk evaluation decisions are crucial for banks due to high risks associated with inappropriate credit decisions. It is an even more important task today as the banks have been experiencing serious financial problems during the past few years.

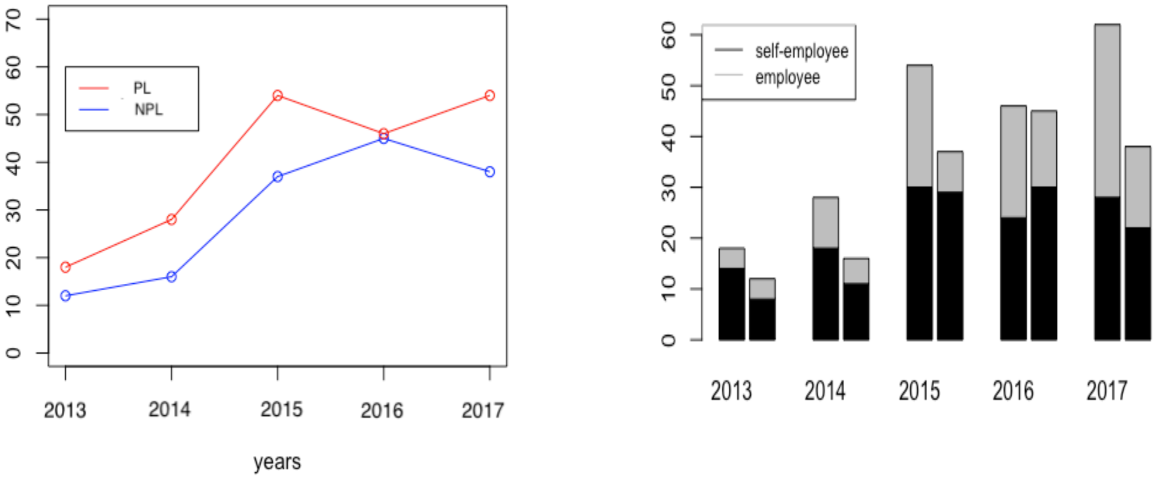
In this research statistical society including all the customers of one of the branches of Saderat bank in Tehran covering the time period from 2013-2017 who have either returned their facilities which they have borrowed from the bank or not. Each application case has 13 variables describing the information we will discuss later. The application data set is grouped into 148 customers who does not honor their loan obligation to service debt (non-performing) and 208 customers who fulfilled their loan obligation (performing). We defined nonperforming loans as loans whose payments have been delayed by more than 60 days. As 148/356 of customers are in non-performing group so that we can consider this dataset as a balanced dataset. This statistical society in terms of credit status are classified in two groups of non-performing and performing.

[1] speaker

Each application case has 13 variables describing the information such as applicant's Gender, Age, Marital status and so on as listed below: Gender, Age, Job, Annual income, Housing, Hometown, Marital status, Types of loans, Loan amount, Loan rate, Year of receiving loan, Minimum amount of mortgage and Number of scheduled payments.

Descriptive Statistics

This section is aimed to visualize the dataset. A visual interactive plot can alleviate the burden for us as decision makers by selecting an appropriate parameter set by giving an insight into our data.
As shown in Fig1(a) number of performing loans(PL) are always more than non-performing loans(NPL), also performing loans have an increasing trend during the first three years and in (2016) the number of performing and non-performing loans are almost equal and it expresses the importance of the problem.
A bar-plot of job vs frequency of PLs and NPLs during the time period of (2013-2017) is given in Fig1(b) and it demonstrates that the number of self-employees are more than the number of employees and in the first four years, number of customers who do not honor their loan obligations are more in self-employees group than the other one.



(a)

(b)

Figure 1: (a)frequency of PLs and NPLs Plot (b)bar-plot of job vs frequency of PLs and NPLs during the time period of (2013-2017)

Fig2 shows the correlation matrix of some of the features. For example as we see, the parameters Annual Income and Loan Amount have high correlation.

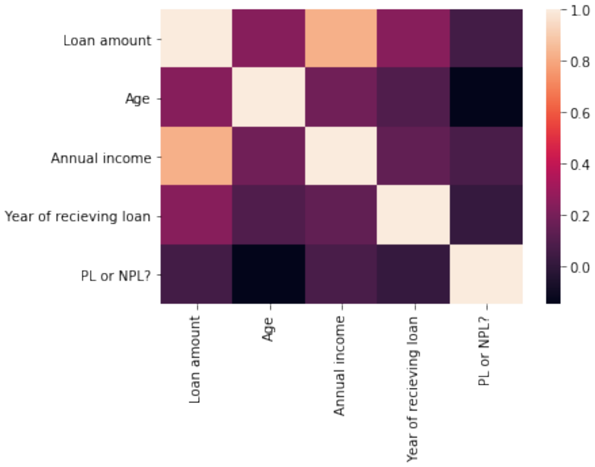


Figure 2: Correlation matrix of some of the features

2 Main results

In this paper, we aim to predict the target variable (PL, NPL) using machine learning methods such as Logistic regression, SVM, Random Forest and Artificial Neural Networks.

It's notable that in order to fit the categorized attributes in methods mentioned before, categorized features are encoded by One Hot Encoder, price related features have been transformed by log-scale and other numerical features have been standardized.

To evaluate the generalization performance of the methods, the k-fold cross-validation technique is used, which is a well-known resampling approach for determining the network parameters.

In this study, the data set is partitioned into five folds of which four of them (%80) are used as the training set to build up the forecast model and the remaining fold (%20) as the test set to justify the generalization performance of the model. Each fold is randomly constructed and contains the equivalent number of PL and NPL.

A grid and randomized search is used to determine hyper-parameter values that maximize the accuracy on the test set.

One of the interesting and challenging problems of this data is that there is a bijective mapping between the following features : Types of loans, Loan amount, Loan rate, Minimum amount of mortgage and Number of scheduled payments so its necessary to choose the most effective one.

As mentioned earlier, machine learning methods are used to predict target variables.

Experimental results and analysis

Logistic Regression

Using Logistic Regression gives us a good insight about features and their relations.

Features with the p-value less than 0.05 and the risk ratio intervals do not include 1, have the most effect on the prediction.

Fitting Logistic Regression to our dataset and according to the p-values and the risk ratio intervals we obtain that four significant independent variables, Housing, Marital status and Minimum amount of mortgage were included in the final regression model.

As mentioned earlier since Logistic Regression is a traditional method for classification, We don't expect the accurate results from it, so we consider it's results as a benchmark.

Table 1: Table of Results

Accuracy	**Precision**	**Recall**	**AUC**
%65	%61	%45	%69

Support Vector Machine

Experiments using SVM for credit risks are relatively new, however, several papers have recently been published assessing the performance of SVM for credit risks.
Unlike most neural networks, one of the major advantages of SVM is its fewer parameter settings.
It was originally designed for binary classification in order to construct an optimal hyperplane so that the margin of separation between the negative and positive data set will be maximized.
The grid search is used to find hyper-parameters.
It's interesting to know that linear kernel gives us the best result among other kernels.

Table 2: The hyper-parameters optimized based on grid search

C	**Gamma**	**Kernel**	**Shrinking**
1	0.01	**Linear**	**True**

Table 3: Table of Results

Accuracy	**Precision**	**Recall**	**AUC**
%65	%65	%57	%72

To compare the diagnostic performance of experiments we use ROC curves, its interesting that among other methods we used, SVM have the most area under the curve (AUC) and as we see in Fig3 it's clear that SVM shows the better diagnosis.

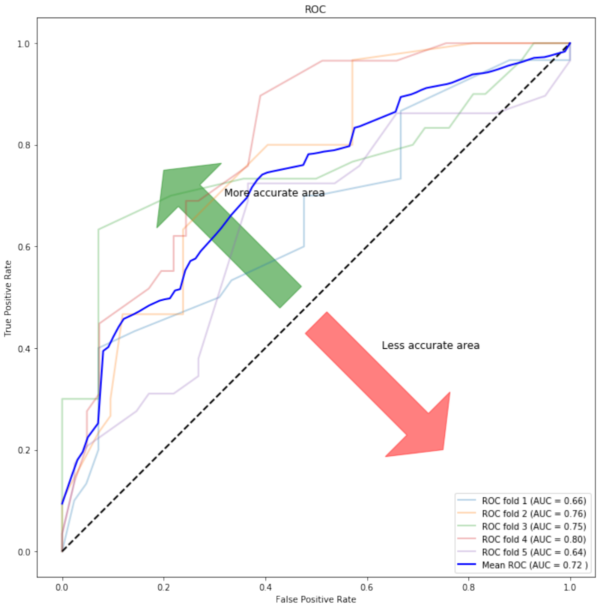


Figure 3: ROC curve of SVM

Artificial Neural Networks

Artificial Neural Networks are one of the main tools used in machine learning. As the neural part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn.

ANNs gather their knowledge by detecting the patterns and relationships in data and learn (or are trained) through experience, not from programming.

They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

To finding number of hidden layers and neurons, we use grid search also Adam optimizer is used to minimizing cost function.

The activation function Relu, is used for all layers expect the last layer and the Softmax is used for the last layer.

Due to the sample size, we didn't expect high accuracy from this method. ANN will show better results in samples with more number of data.

Table 4: Table of Results

Accuracy	**Precision**	**Recall**	**AUC**
%66	%65	%45	%69

Random Forest

In this section, we describe the Random Forest model, which has been designed to led us to the highest accuracy. The Random Forest algorithms are form a family of classification methods that rely on the combination of several decision trees.

The particularity of such Ensembles of classifiers is that their tree-based components are grown from a certain amount of randomness. Based on this idea, Random Forest is defined as a generic principle of randomized ensembles of Decision trees. This algorithm estimates the importance of a variable by looking at how much prediction error increases.

As the grid search has high computational complexity, the hyper-parameters have been optimized based on randomized search.

Table 5: The hyper-parameters optimized based on randomized search

n-estimators	**Max-leaf nodes**	**Max-features**	**Max-depth**	**Criterion**
155	7	2	4	**Gini**

Table 6: Table of Results

Accuracy	**Precision**	**Recall**	**AUC**
%70	%66	%53	%70

Results

We show the results of experiments in Table 1. as we see, Random Forest has the highest accuracy among other methods and also in comparison to others.

to compare the diagnostic performance of experiments we use ROC curves and it's clear that SVM shows the better diagnosis.

As we are looking to select a model based on a balance between precision and recall, we used F-measure. The F-measure is defined as the weighted harmonic mean of the precision and recall.

As we see, SVM have the highest F-measure among others.

Table 7: Results of the performed experiments

	Accuracy	**Precision**	**Recall**	**AUC**	**F-measure**
Logistic regression	%65	%61	%45	%69	%51.8
SVM	%65	%65	%57	%72	%60.7
ANN	%66	%65	%45	%69	%53.2
Random Forest	%70	%66	%53	%70	%58.8

Rule Extraction from Trees

A Decision tree does its own feature extraction. The univariate tree only uses the necessary variables, and after the tree is built, certain features may not be used at all. We can also say that features closer to the root are more important globally. Another main advantage of decision trees is interpretability. The decision nodes carry condition that are simple to understand. A grid search is used to determine hyper-parameter values that maximize the recall in order to extract the rule. For example as we see in this tree, most of the customers who are under 47.5 years old with annual income more than 40.2, fulfilled their loan obligation. Other thing we didn't expect was the tree shows that, most of the customers who are under 47.5 years old with annual income less than 40.2 who are single, fulfilled their loan obligation.

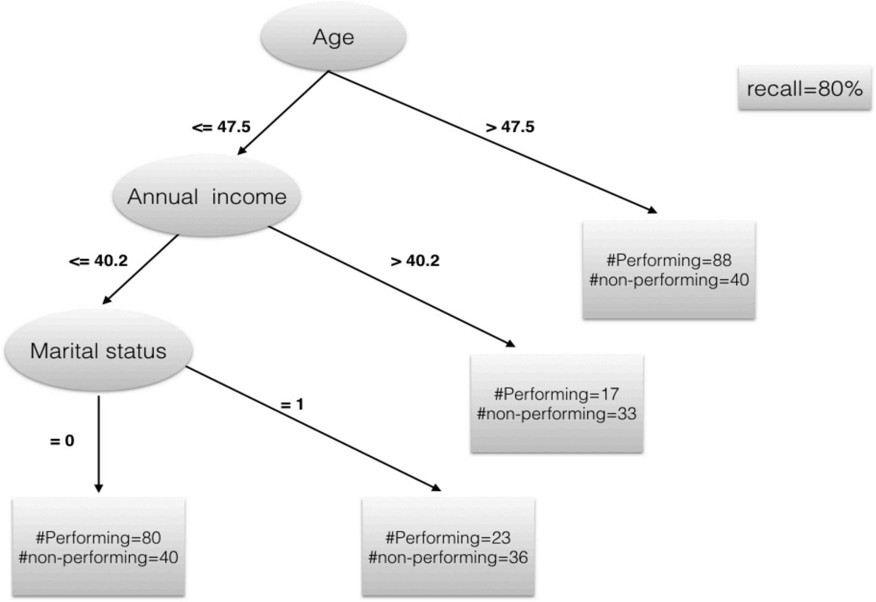


Figure 4: Rule extraction from DT

References

[1] M. Huag, S. Li, and W. Shiueb, *The evaluation of consumer loans using support vector machines*, Elsevier, Expert Systems with Applications, 2006.

[2] M. Carey and W. Treacy, *Credit risk rating systems at large US banks*, Ph.D. Thesis, Imperial College, University of London, 2001.

[3] T. Bellotti and J. Crook, *Support vector machines for credit scoring and discovery of significant features*, Edinburgh EH8 9JY, UK, 2007.

e-mail: amirsalar.vahidi@aut.ac.ir
e-mail: niloofarebrahimi@aut.ac.ir
e-mail: erfan.salavahi@aut.ac.ir