Classification Models for Predicting Credit Risk

Suhail F Sheikh
Dept. Of Computer Science
PES University
Bangalore,India
sahilsheikh1761@gmail.com

Shreyas Bhaktaram
Dept. Of Computer Science
PES University
Bangalore,India
bhaktharamshreyas@gmail.com

Srikar S
Dept. Of Computer Science
PES University
Bangalore,India
srikars2001.official@gmail.com

***Abstract*—Credit risk management is important to financial institutions which provides loans to individuals and businesses.Credit loans and finances have risk of being defaulted if they exceed a certain limit (also known as NPA).It becomes a major problem for credit providers to continue their operations , so credit risk analysis plays a major role for credit providers . To understand risk levels of credit users,credit providers normally collect vast amount of information about borrowers . This paper gives insights about the data and some predictive analytics techniques that can be used to analyze and determine credit risk.**

***Index Terms*—credit risk,predictive analytics,NPA**

I. Introduction

Customer data and credit-history in banks are considered to be extremely important. Data integrity,traceability and confidentiality of customer credit is of key importance. Therefore, most of the banks build their own data-center to protect important information and store it in a data warehouse/repository. Important information and credit card information can be used for fraud detection, marketing and business decisions , even for credit-approval. Credit approval is the operation in which the credit history of a customer is checked in order to decide whether this customer is approved for a service or a loan. The historical credit data generally has information of the customers of which the most important is financial data. The difficulties faced by credit risk approval include credit-card defaults, a wide-ranging expression for fraud and theft.

During the last decade, risk management in credit has gained importance for both lenders and borrowers, especially, in developing countries. Therefore, banks and financial institutions have started to revise their lending policies. This study aims to examine the relationship between the performance of consumer credit clients' payment and some demographic variables (such as residential status and occupation) and some financial variables (such as loan size,income, interest rate,credit category,maturity).

A conceptual model is constructed to explain the relationship between performance of consumer credit clients' payment and interest rate,credit category, loan size,income, maturity, residential status and occupation. The equation of the model is as follows:

Payment performance $=\beta 0 + \beta 1 * CreditType + \beta 2 * Interest + \beta 3 * Income + \beta 4 * Principal + \beta 5 * NumberofPayments + \beta 6 * ResidentialStatus + \beta 7 * Occupation + \epsilon$

II. Review of Literature

A. Can we predict the credit risk using the classification models?

In this paper[1] they have classified credit card frauds broadly into two categories behavorial fraud and application fraud .Paper[1] mainly focuses on credit approval/denial . The paper compares the classification method such as decision trees,Support Vector Machines(SVM) and logistic regression and measure the performance metrics such as accuracy,precision and other metrics.

Classification models used for this research are : C5.0 which builds decision trees from a set of training data in the same way as ID3, using the concept of Information entropy.

CART(classification and regression tree) is a binary decision tree which is constructed by splitting a node into two child nodes repeatedly,beginning with the root node that contains the whole learning sample.

SVM(support vector machine) which finds a hyper-plane that creates a boundary between the types of data.

Logistic Regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. Instead of fitting the data to a straight line, logistic regression uses a logistic curve for prediction.

Various performance metrics against which models were compared:

a. Classified Instances - The importance performance measure is correctly classified instances and incorrectly classified instances.

b. ROC – ROC(Reciever operating characteristic) curve shows the relationship between TPR and FPR c. Confusion Matrix – The confusion matrix illustrates the accuracy of the solution to a classification problem.

The credit risk dataset was published by UCI machine learning. The dataset is interesting because there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with large numbers of values. There are 690 instances in this dataset, with 307 (44.5%) being positive(credit approved) and 383 (55.5%) being

negative(credit denied. In this research different models performance metric attached below is a snapshot

Classification Method	Success Rate	Time taken to build model
C5.0	95.1691	0.02
CART	84.058	0.19
SVM	84.7826	0.31
Logistic Regression	87.9227	0.08

In this study, four classification methods were used to build fraud detecting models. The work demonstrates the advantages of applying the data mining techniques including decision trees, SVM and Logistic Regression to the automatic credit approval problem for the purpose of reducing the bank's risk. The results show that the proposed classifiers of CART outperform other approaches in solving the problem under investigation.

B. Can we predict credit risk based on performance metrics of different models?

This paper[2] focuses on building data-mining models for prediction and decision making(binary decision- approved or denied in our case). In this paper,the three algorithms that have been proposed are K-Nearest neighbours(KNN), Decision Trees(DT) and Logistic Regression(LG). The model is tested on a dataset to determine if the credit tested is approved or denied. Performance metrics like accuracy,sensitivity,etc are calculated for different models to evaluate which model performs better. The dataset used in this paper is related to credit card and financial area. It consists of variety of attributes (Integer,Real, Categorical) including missing values.The dataset has 15 attributes with 690 instances of which 383(55.5%) were classified as -ve (Denied) and 307(44.5%) as +ve(Approved)

Classification Models:
Decision Tree(DT): A decision tree is a flowchart-like structure in which each internal node represents a test on a feature and each leaf node represents a class label and branches represent conjunctions of features that lead to those class labels.
The different algorithms used by DT are ID3,C4.5 algorithm,CART
And the selection measures could be Information Gain,Gini Index,Gain Ratio

K-Nearest Neighbors Algorithm (KNN):
K-nearest-neighbor algorithm is a lazy learning algorithm often abbreviated KNN, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.
The distance measures used for classification are:
Euclidean distance,Manhattan distance,Hamming Function

Selecting the optimal K is best achieved by first analyzing the data. Mainly, a large K value is more accurate as it decreases the noise but there is no guarantee. There is another way to retrospectively determine a good K value which is called Cross-validation in which a part of the data is used to validate the K value. Historically, 1-NN is not the best K, for most datasets the value is confined between k=3 and k=10

Logistic Regression (LG):
Logistic regression (LG) is a sort of regression analysis used for forecasting the result of a categorical variable depending on one or more input variables. LG uses a logistic curve instead of a straight line for fitting the data

Comparison of different models:
The three models used are compared against performance metrics and results have been tabulated below.

	K-NN		*DT*		*LR*	
Measures	+	-	+	-	+	-
Accuracy	*0.8623*	*0.8623*	*0.8579*	*0.8579*	*0.8666*	*0.8666*
Sensitivity	*0.8436*	*0.8773*	*0.8241*	*0.8851*	*0.8925*	*0.8460*
Specificity	*0.8773*	*0.8436*	*0.8851*	*0.8241*	*0.8460*	*0.8925*
F-measure	*0.8450*	*0.8761*	*0.8377*	*0.8737*	*0.8563*	*0.8757*
Precision	*0.8464*	*0.8750*	*0.8519*	*0.8626*	*0.8228*	*0.9076*
Recall	*0.8436*	*0.8773*	*0.8241*	*0.8851*	*0.8925*	*0.8460*
Brier score	*0.2511*	*0.2511*	*0.2495*	*0.2495*	*0.1951*	*0.1951*
MCC	*0.7212*	*0.7212*	*0.7118*	*0.7118*	*0.7344*	*0.7344*

Conclusion:
The classification of the credit approval dataset is very significant to evaluate the security of the credit risk which will result in an efficient and effective credit approval system. In this research paper we examined three classification models (K-Nearest Algorithm), (Decision Tree) and (Logistic Regression). The three models classified the dataset with a promising result but with the best accuracy resulting from the Logistic regression model. DT and k-NN have a better result for other measures such as the precision result for DT with the "+" target class, and the K-NN have better result for the Brier score. It was noticed that that DT and LR have the higher sensitively which mean that the two models are very sensitive for any changing of parameters as it can be seen from Fig.5-9. For the future work we hope to try another classification models such as (Naive Bayes classifier and SVM) for an accurate comparison and to improve the risk approval credit system that can predict the approved class from the denied credit with the most accurate results.

C.

During the last decade, importance of risk management in credit has increased for both borrowers and lenders, especially, in developing countries. For this reason, banks and financial

institutions started to revise their lending policies. This study aims to examine the relationship between the consumer credit clients' payment performance and some demographic variables (such as marital status, sex, age, residential status, occupation) and some financial variables(such as income, loan size, interest rate, maturity, credit category).

A conceptual model is constructed to explain the relationship between consumer credit clients' payment performance and credit category, interest rate, sex, age, marital status,income, loan size, maturity, residential status and occupation.

The present paper is important for three reasons. First, many previous studies and financial institutions have focused on the relationship between lenders' decision and the characteristics of the consumer credit applicants rather than the relationship between payment performance of the consumer credit clients and their characteristics.

It is, of course, important to get some information about the relationship between characteristics of people apply for consumer credit (applicants) and to whom the credit will be given.

However, it is equally beneficial to have an idea about the relationship between the characteristics of people that are already accepted (clients) and whether they are paying back their loans on time or not i.e. payment performance. To some extent, the second is kind of testing whether the decision of accepting/rejecting (or the decision criteria) the applicants is the right one or not. Therefore, investigating the effects of some characteristics of credit clients on clients' payment performance becomes crucial. Second, by ranking customers according to predicted default probabilities, a bank will have a chance to minimize the expected default or misclassification rate subject to some exogenous acceptance rule (Carling et al., 1998).

As a reaction to an increasing competition and bankruptcies, banks all over the world are trying hard to improve the process of loan origination in corporate banking.

Practitioners estimate that improvements in risk management can decrease credit losses by 20 to 40%. Third, no research has been done on characteristics consumer credit applicants and/or clients of any Turkish financial institution in order to develop credit-scoring criteria for the banking sector in.

As our final statistical analysis, a logistic binary regression 3 is used to determine the effect of financial and demographic variables on the payment performance of clients. The estimated change in the log of P (Late Payment) /P (On Time) 4 computed for every value of every factor variable except reference values. First values of factor variables are kept constant when the others' coefficients are calculated. The results of the logistic binary regression is presented in the following table:

Independent Variable	Coefficents (p-values)
Constant	-2.2865 (.001)
Home Loan*	-.4455 (.472)
Individual Support Loan*	-.0837 (.795)
Interest Rate	-.0837 (.000)
Sex (male is constant)	-.1645 (.432)
Age	.00677 (.556)
Married**	-.2718 (.285)
Divorced**	-.0796 (.893)
Widowed**	.9578 (.335)
Income	-2.234E-12 (.984)
Principal	3.9159E-11 (.243)
Number of Payment	.0274 (.047)
Owner (with Credit Debt)***	-1.086 (.321)
Renter***	.3878 (.105)
Family's Home***	.4631 (.091)
Company's House***	.778 (.591)
Occupation	-.01806 (.080)

In conclusion, despite the results being localized only to Turkey, the same model can be used at a global level by increasing the sample size and number of codes for occupation can be decreased to have a better chance to interpret the results between the client's payback reliability and occupation properly. Also, payment performance can also be measured as a continuous variable instead of a binary variable for more accurate results.

III. DATASET

Our dataset was taken from Kaggle and it has extensive details about the borrower.The dataset has a good mix of attributes(Real,Integer,Categorical).

It had 73 columns and 855969 rows initially.

The few important attributes selected for visualization were :

Loanamt : Amount(asked) in dollars.

Term: duration of the loan in months.

Intrate: Interest rate for the loan.

Installment: Amount to be paid back in installments.

Grade: grade of the loan sanctioned.

Emptitle: borrower's job details.

home ownership: whether the borrower has any home ownership.

annualincome: Borrower's annual income

Issueddate: Loan issued date

Purpose: Loan's purpose

defaultind(target variable):a binary number indicating whether the credit has been defaulted or not

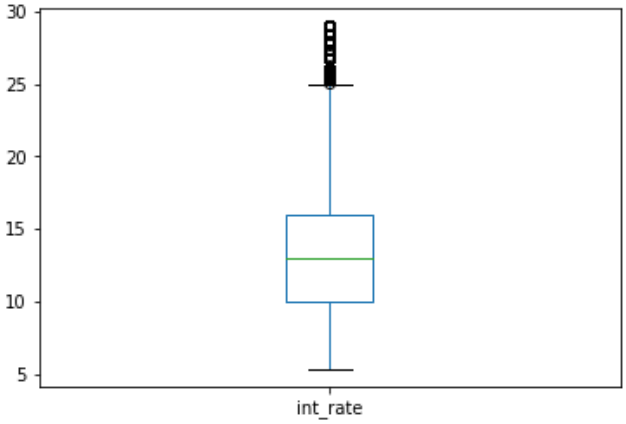
There were columns with missing values and we dropped columns with high miss rate and unwanted /irrelevant columns were dropped.Missing values were handled carefully,few were replaced by "missing" and quantitative variables were imputed by mean of the column and mode replacement was done for categorical variables. After handling missing data, 51

columns remained in our dataset.
We then identified outliers and few attributes had extreme outliers which gives us meaningful insights. Dataset was checked was incomplete,inconsistent,duplicate and incorrect entries and data cleaning was done appropriately. Preprocessing techniques like dimensionality reduction, range transformations and standardisation.

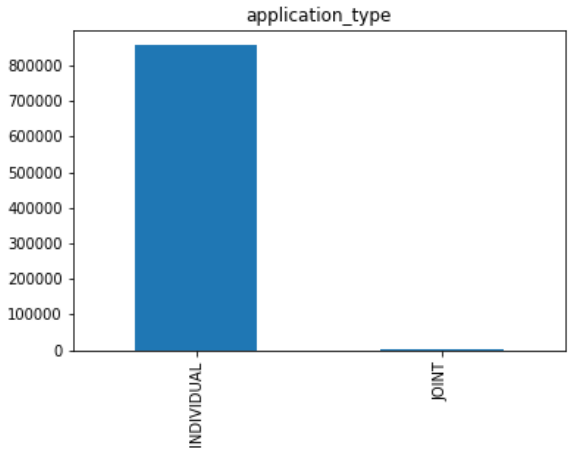
IV. Initial Insights

We plotted histograms and line plots for data visualization and boxplot for outlier detection. We did not remove outliers as they give us meaningful insights. Below are some of the important graphs and plots for which the corresponding insights have been explained.

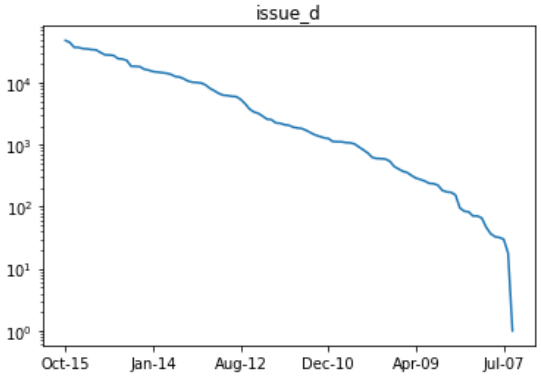
Outliers in intrate:
We can see some outlier values in intrate column which needs to analysed when we do univariate analysis



Individual and joint account:
It can be inferred that more borrowers have taken individual loans compared to joint accounts .



Issued:
From the below image we can infer that the issued loans has been increased exponentially from jul-07 to oct-15



V. PROBLEM STATEMENT AND APPROACH

The aim of our project is to predict whether the borrower will default or not. In this huge dataset with multiple records and attributes various inferences can be drawn and many model could be built and experimented around such as decision trees,logistic regression,KNN,Gradient Boosting Machine(GBM).However our minds are revolving around addressing one question that is given the data about a borrower can we predict whether the borrower will default the loan or not.
Bringing a solid answer to the above problem statement we help the credit issuers estimate and eliminate the credit default risks.
We plan to use Gradient Boosting Machine(GBM),complex regression model,KNN or SVD. The task is little bit challenging due to the complexity however it should be accomplished with ease and we hope to bring out excellent results from the modelling and analysis.

Acknowledgment

We would like to express out profound gratitude to Dr. Gowri Srinivasa and the entire Data Analytics team, for encouraging and providing us with this opportunity to get hands-on experience in the field, and guiding us along the way. We would also like to thank the Computer Science and Engineering department at PES University, for always inspiring us to conduct frequent research and inculcating a problem-solving discipline in us.

References

[1] Subashini, B., and K. Chitra. "Enhanced System for Revealing Fraudulence in Credit Card Approval." International Journal of Engineering Research and Technology. Vol. 2. No. 8 (August-2013). ESRSA Publications, 2013.
[2] Al-Zoubi, Ala Rodan, Ali Alazzam, Azmi. (2018). "Classification Model for Credit Data".The Fifth HCT INFORMATION TECHNOLOGY TRENDS (ITT 2018), Dubai, UAE, Nov., 28 - 29, 2018
[3] Xinyu Gao, Yu Xiong, Zehao Xiong and Hailing Xiong. "Credit Default Risk Prediction Based On Deep Learning." Research square
[4] Ozdemir, Ozlem and Boran, Levent. (2004). An Empirical Investigation on Consumer Credit Default Risk. Turkish Economic Association, Working Papers.
[5] handani, Amir E., Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms." Journal of Banking and Finance 34.11 (2010): 2767-2787.