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 essential to financial institutions which provide loans to individuals and businesses.Credit loans have risk of being defaulted if they exceed a certain limit. To understand levels of risks of credit users,credit providers usually collect large amount of information about borrowers. This paper gives insights about the credit data and some predictive analytic techniques that can be used to analyze and determine credit risk.The model is tested on a data set for different customers to predict if it is approved or denied. The four different algorithms used in this model are: K-Nearest Neighbors Algorithm(KNN), Support Vector Machine(SVM) and Logistic Regression(LG) and Random Forest .Various performance metrics like accuracy,F1-score, precision and recall of these models are tabulated and results are compared.**

***Index Terms*—KNN,SVM,LG,Random Forest**

I. Introduction

Credit risk is the possibility of a risk of default on a debt by a client or failure to satisfy contractual obligations. The importance of mitigating credit risk has increased in developing countries for both borrowers and lenders. As a result, banks around the world have started revising their lending policies and are trying to identify which of their customers are most likely to default on their loans. The challenges to successful credit risk management are:

- Inefficient data management
- Insufficient risk tools
- Inadequate reporting
- Poor data visualization capabilities

The aim of this paper is to examine the relationship between the borrower's credit payment performance and some demographic variables (such as marital status, age, sex, occupation, country of residence) and some financial variables (such as income, loan size, last payment date, next payment date) so that banks can make a more informed choice as to which clients are more likely to not default on their loans resulting in higher revenue for the bank as well.

II. Review Of Literature

Previous papers have focussed on the relationship between the lenders' decisions and the details of their application rather than comparing against demographic variables and their characteristics. It is of utmost importance not only to draw a relationship between applicants who have borrowed a loan but it is also equally important to draw a relationship between accepted clients and whether they are paying back their loans on time (this is where the last payment date and next payment date play a very important role). Therefore, investigating some of the characteristics of clients becomes extremely crucial and banks can make a more informed choice as to which applicants are more likely to default on their loans. This paper aims to draw relationships from a variety of variables (52 to be exact) and make a more informed choice on the kind of clients banks usually receive giving banks a chance to improve their risk management system, something which is very important in developing countries as their financial system is not only unstable but also identifies the concept of credit in a real sense. The data set chosen for this paper has a good mix of categorical and nominal variables, allowing us to make an informed choice as to whether a client will default on their loan or not and will benefit banks greatly.

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. Methodology

Evaluation of the models is done based on the performance metrics like accuracy,precision,specificity and ROC score.Confusion matrix is also taken into consideration to evaluate different models based on their performance metric.

A. K-Nearest Neighbors Test

This is a lazy learning model where the learner does no training at all when training data is provided. In the training phase, it stores the entire data set but does not perform any calculations.Only when test data is provided, it computes the distance between the query point and all the other points. It is a supervised algorithm which can be used for both classification and regression. In our case, it classifies the test instances based on top k neighbours and it takes a majority voting between the top-k neighbour classes and assigns the most similar class to the test instance. This method can be very slow for large datasets as it has to perform the distance calculation every time a new instance is given. The different distance measures used are:

Euclidean distance:$\sqrt{\sum_{i=1}^{k}(Xi-Yi)^2}$

L1 distance : $\sum_{i=1}^{k}|Xi-Yi|$

Euclidean metric is chosen as a distance metric. To find the k nearest neighbors, finding the best k value can be done by analysing the given data. Usually, a larger value of k is preferred as it avoids overfitting and reduces the noise. Other methods include:

- picking k between 3 and 10.
- k= n+1
- k=$\sqrt{n}$ where n is the number of classes
- Elbow method where the k value with minimum error is chosen.

We choose k=n+1 and also test the model for k=5 and k=10.

B. Support Vector Machine

Support Vector Machine is a supervised algorithm which uses the concept of hyperplanes or lines to separate data into classes.This is called a decision boundary which classifies the data points on one side of it as one class and any other point as another class.To do this we find out the points that are closest to both the classes called support vectors. The main goal of SVM is to maximize the margin to find the optimal hyperplane which ensures that the gutter width is maximised which avoids misclassification. Svm uses a kernel that transforms an input n-dimensional space into a higher dimensional space. This is called kernel trick and different kernel used are:

Linear kernel : $K(\overrightarrow{u},\overrightarrow{v})=\overrightarrow{u}.\overrightarrow{v}$

Polynomial kernel : $K(\overrightarrow{u},\overrightarrow{v})=(\overrightarrow{u}.\overrightarrow{v}+b)^n$ n >1

RBF kernel : =$K(\overrightarrow{u},\overrightarrow{v})=e^{(-(|\overrightarrow{u}-\overrightarrow{v}|)^2)/2\sigma^2}$

We used the Standard Scaler pipeline approach as a normalisation technique for scaling the data and RBF kernel for mapping data to a higher dimension.

C. Logistic Regression

Logistic regression is an extension to Simple linear regression which is used to predict the relation between the independent variables and a dichotomous dependent variable. Instead of using a straight line, it uses a LR curve for the prediction. The log odds ratio is given by

$\ln(p/1-p)=\beta 0+\beta 1.x$

The input to this function can be in the range (-inf,+inf) and it outputs a probability p which can be used for classification. The logistic function is in the sigmoid form as shown below:

$p=1/(1+e^{-(\beta 0+\beta 1.x)})$

D. Random Forest

Random Forest classifier is an ensemble model.it consists of multiple decision trees ,there are two types randomness built into the trees each tree is built on a random sample from the original data and at each tree node ,a subset of features are randomly selected to generate the test-split. Random forest uses bagging method wherein the results of various decision trees on different subsets of the dataset are combined and a simple voting score is used to make a final classification.The advantages of using random forest is it provides higher accuracy through cross validation,all the individual models in a random forest can run parallely thus reducing the time to train the model.

V. Results

A. K-Nearest Neighbors Test

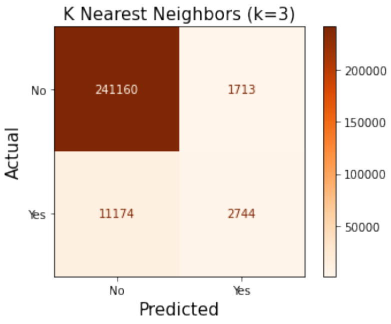
The below table shows the performance of KNN model in terms of accuracy using Euclidean distance measure.

KNN PERFORMANCE RESULTS

K Value	KNN Distance Metric	Accuracy
3	Euclidean	0.9498
5	Euclidean	0.9513
10	Euclidean	0.9497

We observe that k=3 gives the best accuracy. other accuracy results are mentioned in the above table

Below is the confusion matrix for KNN model which shows the high value of True positives.



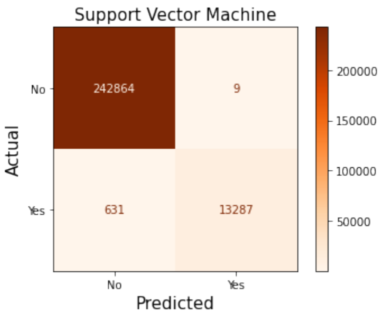
B. Support Vector Machine

The below table shows the performance of SVM using RBF kernel and Standard Scaler for scaling the data to follow Standard Normal distribution with 0 mean and constant variance.

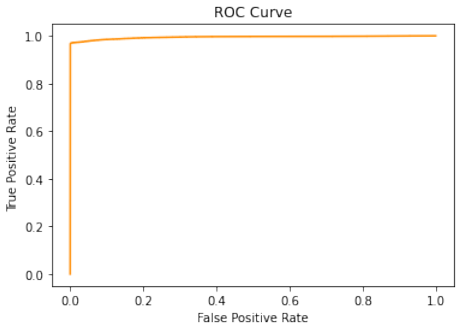
SVM PERFORMANCE RESULTS

Performance Metric	Value
Accuracy	0.9971
Precision	0.9999
F1-Score	0.9726
Recall	0.9467

Below is the confusion matrix for the same in which the values have not significantly changed as compared to KNN model



Below is the ROC curve for SVM model with an ROC-score of 0.9950



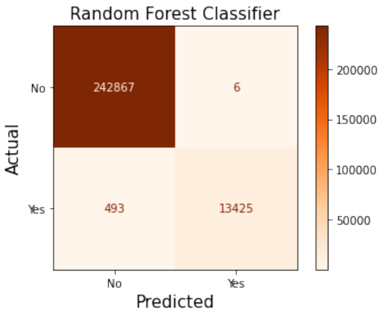
C. Random Forest

Performance metrics for Random Forest model are tabulated below and it performs better than SVM and KNN

RANDOM FOREST PERFORMANCE RESULTS

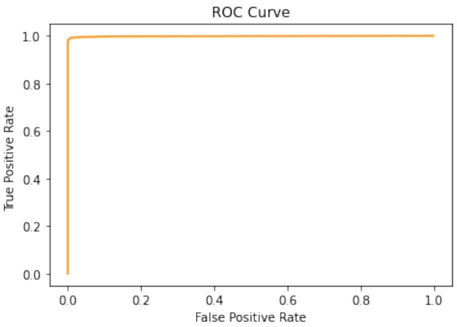
Performance Metric	Value
Accuracy	0.9980
Precision	0.9990
F1-Score	0.9817
Recall	0.9640

below is the confusion matrix for the same in which the values have not significantly changed



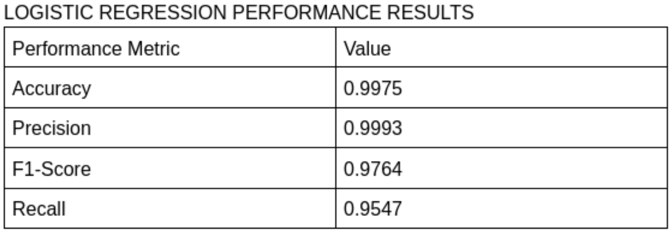
Below is the ROC curve for Random Forest model with an

ROC-score of 0.9984



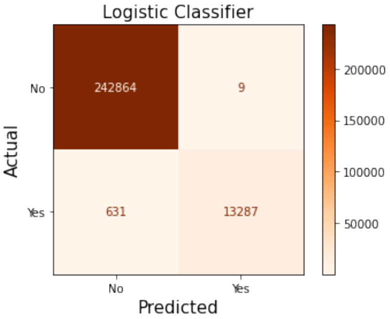
D. Logistic Regression

This model gives accuracy almost similar to previous models but higher F1-score


LOGISTIC REGRESSION PERFORMANCE RESULTS

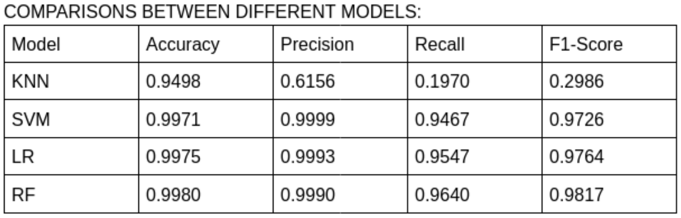
Performance Metric	Value
Accuracy	0.9975
Precision	0.9993
F1-Score	0.9764
Recall	0.9547

Below is the confusion matrix for our final model



Comparision between different models:

We finally compare all the models against various performance metrics and observe that Random Forest performs the best with an F1-score of 98.17 % which gives us a combined result considering both precision and recall. Also, it gives the highest ROC score and hence performs best.


COMPARISONS BETWEEN DIFFERENT MODELS:

Model	Accuracy	Precision	Recall	F1-Score
KNN	0.9498	0.6156	0.1970	0.2986
SVM	0.9971	0.9999	0.9467	0.9726
LR	0.9975	0.9993	0.9547	0.9764
RF	0.9980	0.9990	0.9640	0.9817

VI. Conclusion

The classification models implemented for this dataset can have a very significant impact in profiling credit risk and help reduce the credit default risk for lenders.We have implemented four classification models for this dataset that is K-nearest neighbors,Support Vector,Random Forest and Logistic regression which gave promising results to consider for the classification. The Random Forest model gave us the highest accuracy and F1-score and other models also performed similarly but to compare the models the final metric used was ROC score and Random Forest gives the highest ROC score. It was noticed that Logistic Regression as a higher sensitivity and any change in the parameters would affect the results. To build upon the models, we hope to try other classification models such as Decision trees and Naive Bayes for better and accurate results to improve the risk approval credit system and the predictions.

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.