Credit Scoring using Machine Learning Techniques

Sunil Bhatia
Computer Science Department
VESIT, Chembur
Mumbai University

Pratik Sharma
Computer Science Department
VESIT, Chembur
Mumbai University

Rohit Burman
Computer Science Department
VESIT, Chembur
Mumbai University

Santosh Hazari
Computer Science Department
VESIT, Chembur
Mumbai University

Rupali Hande
Computer Science Department
VESIT, Chembur
Mumbai University

ABSTRACT
Lenders such as banks and credit card companies while reviewing a client's request for loan use credit scores. Credit scores help measure the creditworthiness of the client using a numerical score. Now it has been found out that the problem can be optimized by using various statistical models. In this study a wide range of statistical methods in machine learning have been applied, though the datasets available to the public is limited due to confidentiality concerns. Problems particular to the context of credit scoring are examined and the statistical methods are reviewed.

Keywords
Data Mining, Credit Scoring, Logistic Regression, LDA, XGBoost, Random Forest.

1. INTRODUCTION
The process of deciding to accept or reject a client's credit by banks is commonly executed via judgmental techniques and/or credit scoring models. Earlier, financial institutions and most banks used the method of judgmental approach that is based on the 5 C's, which are condition, character, collateral, capital and capacity. In this modern computerized world, this process of deciding can be optimized using statistical methods in machine learning.

Thus banks and financial institutions to improve the process of assessing creditworthiness of an applicant during the credit evaluation process develop Credit scoring models. Credit scoring is a system creditors (banks, insurance companies) use to assign credit applicants to either a ''good credit'' group the one that is more likely to repay the debt or a ''bad credit'' group the one who has a high possibility of defaulting on debt or any financial obligation i.e. not paying within the given deadline.

Construction of credit scoring models requires data mining techniques. Using, demographic characteristics, historical data on payments and statistical techniques, these models can help in identifying the important demographic characteristics, which is related to credit risk, and assign a score to each customer. The probability that an applicant will default must be calculated from information about the applicant provided at the time of filing the application, and this estimate will thus serve as the basis for his/her creditworthiness.

In the paper [1] the four machine learning methods reviewed for Credit scoring jare statistical methods, Hybrid Methods, Artificial Intelligence method, and ensemble learning method. Statistical model includes LDA (Linear Discriminant Analysis), MARS, Decision tree. AI methods include ANN, SVM, and K-Nearest method. Paper also discusses about behavioral scoring method. Behavioral scoring makes a decision about management of credit based on the repayment performance of existing customers during a certain predefined period of time. It also includes repayment behavior and payment history of the client. According to this paper ensemble learning has better prediction accuracy and classification ability and is thus widely applied to personal credit evolution.

This paper [2] deals with the design aspects related to financial fraud detection. The aim of feature selection is to improve both the actual and computational performance of the solution, as well as providing a better understanding of the problem. Feature ranking algorithms assign rating to individual features based on certain attributes such as accuracy, content and consistency and choose a suitable subset on the basis of ranking. Performance metrics are used as small increase in performance can lead to large economic benefits. In Classification method accuracy, sensitivity, specificity, precision, false positive rate are the performance measure. In clustering Hopkins statistic is the performance measure. The paper tests various algorithms such as GA1, GA2, DT1, DT2, SVM etc. for determining the best prediction method. It has been found that if misclassification costs are high, techniques with higher sensitivity such as GP1 or neural networks may be suitable choices. If receptiveness to minor chances in dataset is desired then the ant colony optimization or neural networks could be appropriate. Overall the support vector machine could be considered to have the best performance with the highest accuracy.

The study applies the credit scoring techniques using data mining of payment history of members from a recreational club [3]. Classification performance of credit scorecard model, logistic regression model and decision tree model were compared. Classification error rates were 27.9%, 28.8% and 28.1% respectively. The cut off score also known as the threshold can be determined by the value of K-S Test for each bucket of score in the validation sample. The target variable is payment status which is a binary variable with 2 categories: default (consisting of persons who have defaulted) and non-default (consisting of person who have not defaulted) which were coded using numerical values (1 and 0). Out of 2765 members, 35% were found to be defaulters. The majority of the members are male (80%) and more than half (74%) of the customers are from non-government sector. Two main limitations are the availability of data and sample selection issues.

SAS advanced analytical techniques have a proven ability to quickly and accurately forecast the risk of credit losses [4]. This paper uses Data mining in SAS for credit scoring process. Application data, financial data, credit bureau data, character data, performance data, demographic data are different types of data used as input to an application scorecard. As data mining is used, Credit scoring is more of inductive classification where the learning process is supervised by a vector of known outcomes in the training data. Decision trees is a technique used where each branch is a classification question and the leaves of the trees are segments of the data that fall within a specific class. E.g. Good or bad. Splits are made on the basis of a given variable and its specific value. Decision trees can handle missing values without imputation and also can divide the data on each branch without losing any data. It is found that Decision trees and logistic regression are excellent techniques for credit scoring models. Also decision trees has an advantage that it is easy to understand the result as each leaf node can be traced back to the root node.

Credit scoring is used to describe formal statistical methods for classifying applicants for credit into good and bad risk classes [5]. Predictor variables are used to yield estimates of the probabilities of defaulting. Factors, which need to be considered, are the cost of collecting and analyzing information, expected returns on good and bad loans that have been accepted, the fact that loans may be profitable even if the borrower defaults, the attrition rate etc. Population drift can also be a problem in credit scoring applications. A widespread practice used in credit scoring is reject inference. It describes the practice of attempting to infer the likely true class of the rejected applicants and then using this information to yield a new scorecard that is superior to one built on only those accepted for credit thus making the dataset more reliable. Risk based pricing, loan servicing, and review function and fraud scorecard are other issues which are addressed in this paper.

2. Machine Learning Techniques

2.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is derived from Fisher's linear discriminant, a classification method used in machine learning to discover a linear combination of features that characterizes two or more classes of objects. The resulting combination of the most desired features is used as a dimensionality reduction before later classification or as a linear classifier.

If one wants to preserve the difference between the classes as well while reducing the dimensions, LDA is the option available that determines the discriminant dimension in the response pattern space, on which the ratio of between-class over within-class variance of the available data is maximized.

Linear Discriminant Analysis method is thus easy and predicts models whose accuracy is as good as other methods that are complex.

In Credit Scoring model, we determine worthiness to receive credit. This is done by determining the probability that the person will default in future or not. Thus, we can use LDA to predict that according to the person's details he/she falls under default or not-default category and accordingly grant credit.

The score function is calculated as follows:

$Z=\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d$

$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta}$: Score function

$\downarrow$

$S(\beta) = \frac{\underline{Z_1} - \underline{Z_2}}{Variance\ \ of\ Z\ within\ groups}$

On a given score function, the problem is to estimate the linear coefficients of variables that maximize the score that can be found as follows:

$\beta = C^{-1}(\mu_1 - \mu_2)$: Model coefficients

$C = \frac{1}{n_1+n_2}(n_1 C_1 + n_2 C_2)$: Pooled covariance matrix

Where

β : Linear model coefficients

C_1, C_2 : Covariance matrices

μ_1, μ_2 : Mean vectors

A way of assessing the effectiveness of the discrimination thus created is by calculating the Mahalanobis distance between two groups formed by LDA. A distance of more than 3 means suggests that the two averages differ by more than 3 standard deviations which thus is an indicator of good separability between the two groups.

$$\Delta^2 = \beta^T(\mu_1 - \mu_2)$$

$$\Delta : Mahalanobis\ distance\ between\ two\ groups$$

Finally, a new point (i.e. A person's profile) is classified into default (C1) or not-default (C2) by projecting it onto the maximally separating direction and classifying it in the class C1 (default) if:

$\beta^T(x - (\frac{\mu_1+\mu_2}{2})) \geq log\frac{p(C_1)}{p(C_2)}$

β^T : Coefficients vector

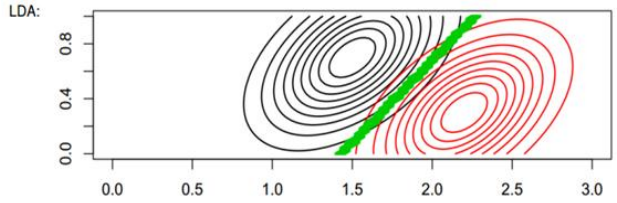
x : Data vector

$\frac{\mu_1+\mu_2}{2}$: Mean vector

$\frac{p(C_1)}{p(C_2)}$: Class probability

A sample graph in LDA is as follow:

of the M. The best split on these m is used to split t



2.2 B. Random Forest Algorithm

Decision Trees-

Tree based learning algorithms like Decision Trees are considered to be one of the best and mostly used in the category of supervised learning methods. Tree based methods encourage predictive models with stability, high accuracy and easy of exploration. Tree based methods map the non-linear relationships with a good accuracy. This method breaks down the dataset into smaller and smaller subsets of data while in the same period an associated decision tree is developed in an incremental manner. We thus get a tree with decision nodes and tree nodes. Decision trees can handle both categorical and numerical data. A sample Decision Tree is as follows:

After the decision tree is ready for our database, we will create a set of rules that defines the major objective of our project i.e. knowing if the credit risk will be good or bad.

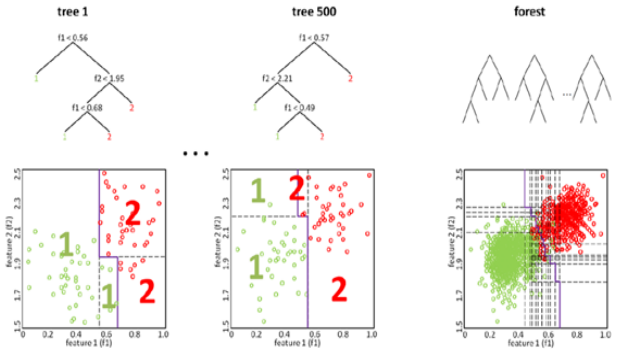
At the end, result of both the algorithms is compared to know if regression or decision tree would prove efficient as compared to the other.

Random Forest Algorithm-

In our project, under Decision Trees we apply Random Forest algorithm. Random Forest is a multifaceted machine learning technique which is capable of performing both classification and regression tasks. It also treats missing values, undertakes dimensional reduction methods, outlier the values and other essential steps at the data exploration stage. It is an ensemble learning technique, where a group of weak models is combined to form a powerful model. To classify a new object based on attributes, each tree assigns a classification i.e. in other words it "votes" for that particular class. The forest (i.e. the algorithm) chooses the classification having the most votes (over all the other trees in the forest) and in case of regression, the algorithm takes average of different outputs.

The algorithm is as follows:

1. Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but with replacement. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out he node. The value of m is held constant while we grow the forest.
3. Each tree is grown to the largest extent possible and no pruning is done.
4. Predict new data by aggregating the predictions of the n tree trees (i.e., majority votes for classification, average for regression).



2.3 Logistic Regression

Logistic regression, developed by David Cox, is one of the most frequently used statistical model used in credit scoring.

Logistic regression can be seen as a special case of the linear model and analogous to linear regression. Logistic and linear regression model differ in their outcome. The outcome of logistic regression is discrete rather than continuous.

Logistic regression models the relationship between independent variables and one or more independent variables. The constraints the estimated probabilities to lie between 0 and 1. Depending on the values of attributes (independent variables), we will find the probability that the dependent variable takes value 0 (default probability). It is commonly used for prediction and forecasting. Profit, sales, diseases, probability of failure of a given process, could all be predicted using regression techniques.

The conditional probability $Pr(Y = 1| X = x)$ is modeled as a function of x and $Y$is a binary output variable. It uses the form

$$P(x) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}}$$

We can also re-express the above equation as

$$log(\frac{P(x)}{1 - P(x)}) = \beta_0 + x \cdot \beta$$

The coefficients β are estimated using the maximum likelihood method. The solution of β0 + x · β = 0 gives us the decision boundary separating the two predicted classes. The logistic function is a nonlinear function and this means that the probability that $Y = 1\ or\ Y = 0$is not constant with constant changes in the predictor variable X.

2.4 XGBoost

XGBoost is an open source implementation of Gradient Boosting Machines. It is a scalable and high performance machine learning system for tree boosting and claims to run 10 times faster than existing solutions. It provides parallel tree boosting that solve many data science problems in a fast and accurate way. It also provides cache access patterns, data compression and sharing for tree boosting.

XGBoost is used for supervised learning problems. The model for XGBoost is tree ensembles which consist of classification and regression trees. Tree ensembles are also a model for random forests. The difference between the two is how we train them. Consider an objective function

$$obj = \sum_{i=1}^{n} l\,(yi, \hat{yi}^{(t)}) + \sum_{i=1}^{t} \Omega\,(fi)$$

Here l is a differentiable convex loss function and Ω is the regularization term. This model is trained in an additive manner. We note the prediction value at step t by $\hat{yi}^{(t)}$, so we have

$$\hat{yi}^{(t)} = \sum_{k=1}^{t} f_k\,(x_i) = \hat{yi}^{(t-1)} f_t(x_i)$$

$$obj^{(t)} = \sum_{i=1}^{n} l\,(yi, \hat{yi}^{(t)}) + \sum_{i=1}^{t} \Omega\,(fi)$$

$$= \sum_{i=1}^{n} l\,(yi, \hat{yi}^{(t-1)} + f_t(x_i)) + \Omega\,(f_t) + c$$

Here $c$is constant. We take the Taylor expansion of the loss function up to the second order.

$$obj^{(t)} = \sum_{i=1}^{n} l\,(yi, \hat{yi}^{(t-1)} + g_i\, f_t(x_i)) + \frac{1}{2} h_i\, f_t^{\,2}(x_i)\,] + \Omega\,(f_t) + c$$

Where gi and hi are defined as,

$$g_i = \delta_{\hat{y}\,(t-1)}\, l(yi,\ \hat{yi}^{(t-1)})$$

$$h_i = \delta^2{}_{\hat{y}\,(t-1)}\, l(yi, \hat{yi}^{(t-1)})$$

3. FIGURES/CAPTIONS

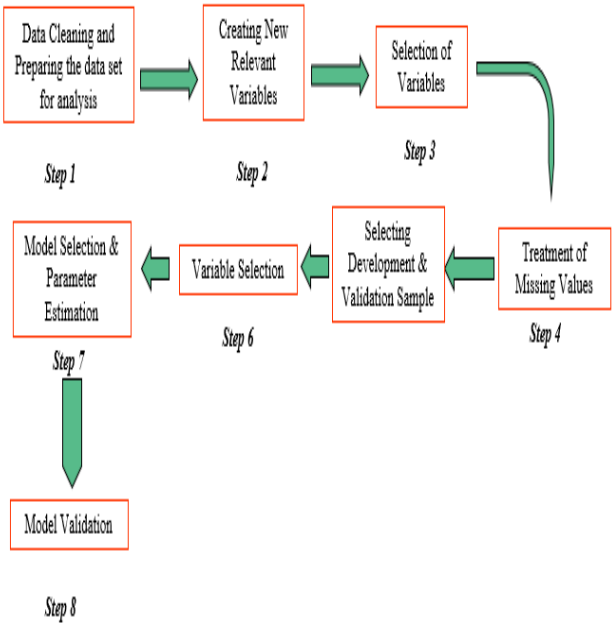


Fig 1 Process Flow for Developing Scoring Model:

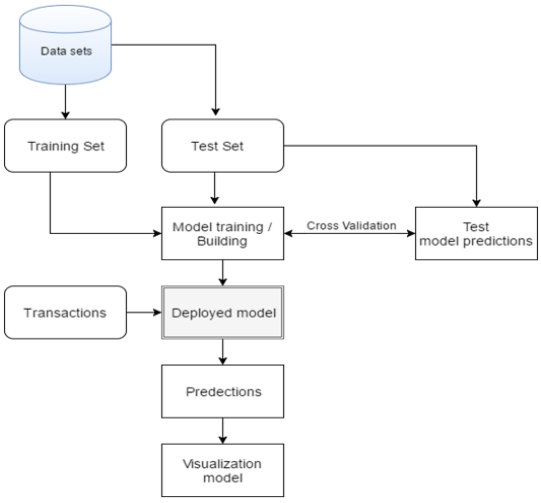


Fig 2 System Architecture:

4. CONCLUSION

The In this work, we have surveyed various machine learning algorithms with respect to credit scoring along with various ensemble techniques. In general, we have decided to implement the model with four different machine-learning models for creating credit risk scorecards. Different analysis techniques will be used to find the accuracies of models.

5. REFERENCES

[1][Li, Xiao-Lin, and Yu Zhong. "An overview of personal credit scoring: techniques and future work." (2012).

[2]West, Jarrod, and Maumita Bhattacharya. "Some Experimental Issues in Financial Fraud Mining." Procedia Computer Science 80 (2016): 1734-1744.

[3]Yap, Bee Wah, Seng Huat Ong, and Nor Huselina Mohamed Husain. "Using data mining to improve assessment of creditworthiness via credit scoring models." Expert Systems with Applications 38.10 (2011): 13274-13283.

[4]Jayagopal, B. "Applying Data Mining Techniques to Credit [5]Hand, David J., and William E. Henley. "Statistical classification methods in consumer credit scoring: a review." Journal of the Royal Statistical Society: Series A (Statistics in Society) 160.3 (1997): 523-541.

[6]James, Gareth, et al. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.

[7]Kennedy, Kenneth. "Credit Scoring Using Machine Learning." (2013).

[8]Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms." *Journal of Banking & Finance*34.11 (2010): 2767-2787.

[9]Yap, Bee Wah, Seng Huat Ong, and Nor Huselina Mohamed Husain. "Using data mining to improve assessment of creditworthiness via credit scoring models." *Expert Systems with Applications* 38.10 (2011): 13274-13283.

[10]Siddiqi, Naeem. *Credit risk scorecards: developing and implementing intelligent credit scoring*. Vol. 3. John Wiley & Sons, 2012.

[11]Hand, David J., and William E. Henley. "Statistical classification methods in consumer credit scoring: a review." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160.3 (1997): 523-541.

[12]Chen, Tianqi, and Tong He. "xgboost: eXtreme Gradient Boosting." *R package version 0.4-2* (2015).

[13]Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

[14]Izenman, Alan Julian. "Linear discriminant analysis." *Modern Multivariate Statistical Techniques*. Springer New York, 2013. 237-280.

[15]Hosmer, David W., and Stanley Lemeshow. *Multiple logistic regression*. John Wiley & Sons, Inc., 2000.

[16]Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System."(2016).