TEST Engineering & Management

*May - June 2020*
*ISSN: 0193-4120 Page No. 4450-4453*

# Credit Risk Analysis

**[1]Sudhir KumarPandey, [2]Ashwin Kumar U M**

[1,2]School of Computing and Information Technology, REVA University, Bengaluru, India
[1]ssddpp39@gmail.com, [2]ashwinkumarum@reva.edu.in

***Article Info***
***Volume 83***
***Page Number: 4450-4453***
***Publication Issue:***
***May - June 2020***

***Article History***
***Article Received:****19 November 2019*
***Revised:*** *27 January 2020*
***Accepted:*** *24 February 2020*
***Publication:*** *12 May 2020*

**Abstract**

The several techniques for credit scoring were used to create credit score cards. Due to its desirable features (robustness and transparency) logistic regression model is among them the most widely used in the banking industry.Although some modern techniques (support vector machine) were applied to credit scoring and showed superior predictive accuracy, they have problems with interpretability of the results. Therefore, those specialized methods were not commonly used in practice.Logistic regression with random coefficients is suggested to improve predictive accuracy of logistical regression.The proposed model will boost logistic regression prediction accuracy without sacrificing desirable features.The proposed method of developing the credit scorecard is expected to lead to successful credit risk management in practice.

***Keywords****: logistic regression, Exploratory Data Analysis WOE and information value, DiscreditingPredictors/Binning.*

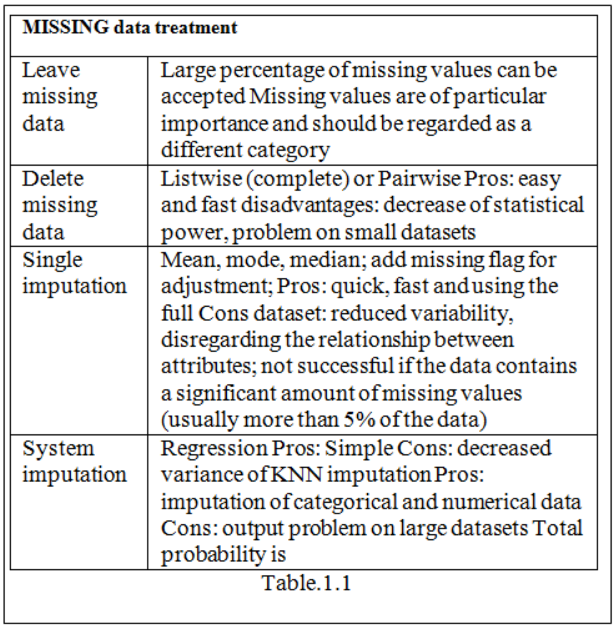
## 1. Introduction

Be mindful of your credit scorecard??Did you have your credit rejected and do not know?. A financial institution or bank makes use of credit ratings to decide Who is applying for a loan number, at what interest rate and loan limits the more a financial institution is comfortable or bank can be of the customer's creditworthiness .but, Bank can be the creditworthiness of the customer.but a credit score may not form part of a daily credit report A mathematical formula is used to transform and data into a three-digit number in the credit report, used by lenders to make credit decisions,but Yet the exact method offices used to determine credit scores remain a mystery. The aimthis project is to use credit rating tools to assess the particular risk client and to create a scorecard model.Credit scoring is the use of a mathematical model to attribute a probability to a credit application,And it is a type of Artificial Intelligence, based on predictive modeling, which assesses the probability that a client who defaults on a credit obligation is delinquent or insolvent.

Over the years, different modeling methods have developed to incorporate credit scoring.Notwithstanding variety, the Credit Scorecard model sticks out and is used by almost 90 per cent of developers of scorecards.As a framework for statistics / machine learning,Its scores may be used explicitly as measures of probability and thus to provide direct feedback for risk-based pricing.First, descriptions of how to use credit scoring to build a consumer credit scorecard will appear as follows. The research will involve an exploratory analysis of data, variable selection, model creation, and scoring.

## 2. Related Work

### 1. Data Preparation and Exploratory Data Analysis



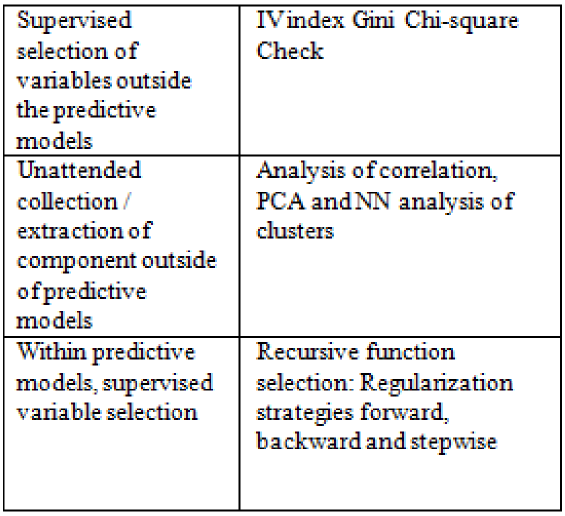
| MISSING data treatment | |
|---|---|
| Leave missing data | Large percentage of missing values can be accepted Missing values are of particular importance and should be regarded as a different category |
| Delete missing data | Listwise (complete) or Pairwise Pros: easy and fast disadvantages: decrease of statistical power, problem on small datasets |
| Single imputation | Mean, mode, median; add missing flag for adjustment; Pros: quick, fast and using the full Cons dataset: reduced variability, disregarding the relationship between attributes; not successful if the data contains a significant amount of missing values (usually more than 5% of the data) |
| System imputation | Regression Pros: Simple Cons: decreased variance of KNN imputation Pros: imputation of categorical and numerical data Cons: output problem on large datasets Total probability is |

Table.1.1

Exploration and clean-ofData are iterative moves to one another.Information research involves. The analysis is both univariate and bivariate, ranging including univariate statistics and the distribution of frequencies to associations, cross-and analysis of characteristics.Before making a decision about how to handle missing values, different curriculums, where every course is attributed to one state. The other steps in this system are combined versions of measure of liner association correlation-analysis-based.

Outliers are another "beast" in our data because their existence will render incorrect, statistical assumptions under which we are constructing a model.For example, outliers may be a valuable source of knowledge when detecting fraud therefore, replacing them with a mean or median and mode value would be a bad idea.

Analyzing the outliers using univariate and multivariate regression. We may use visual tools such as histograms Box plots and statistical approaches like mean and standard deviations for detection.Clustering of distant clusters,Judging what should be considered an outlier is not as easy as finding missing values.



| | |
|---|---|
| Supervised selection of variables outside the predictive models | IVindex Gini Chi-square Check |
| Unattended collection / extraction of component outside of predictive models | Analysis of correlation, PCA and NN analysis of clusters |
| Within predictive models, supervised variable selection | Recursive function selection: Regularization strategies forward, backward and stepwise |

Table 1.2

## 2. Logistic Regression

### Comparison with other models:

### Logistic regression vs SVM:

□ SVM Can handle non-linear solutions but LR can handle only linear solutions.

□ Linear Help Vector Machine treats outliers better, because the solution receives full margin.

□ Loss of hinge in SVM outperforms log loss in regression of logistics.

### Logistic Regression vs Decision Tree:

□ Decision tree is better at handling co linearity than Logistic regression.

□ Decision trees are better than LR for the categorical values.

### Logistic Regression vs Neural Network:

□ NN Can provide support for non-linear solutions where Logistic regression is impossible.

□ LR Have convex loss functions, so it will hang at a minimum locally, while NN can hang.

□ LR Overperforms NN when training data is smaller and features increasing, whereas NN needs detailed training data.

### Logistic Regression vs Naive Bays:

□ Naive Bays algorithm is a generative model while Logistic regression is a model of discrimination.

□ Naïve Bays works well with limited datasets, while Logistic regression + regularization can perform equally well.

### Logistic Regression vs KNN:

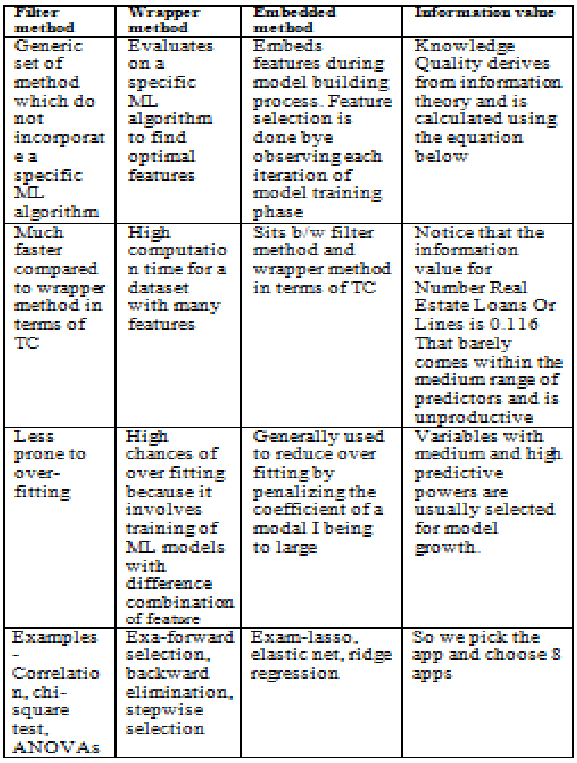
□ K-nearest neighbors areA model which is not parametric, where Logistic regression Is a model parametric.

□ K-nearest neighbors are comparatively slower than Logistic Regression.

□ K-nearest neighbors supportNon-linear solutions for which LR only supports linear solutions.

### Feature selection

Filter vs. Wrapper vs. Embedded vs. Information Value methods:



| Filter method | Wrapper method | Embedded method | Information value |
|---|---|---|---|
| Generic set of method which do not incorporate a specific ML algorithm | Evaluates on a specific ML algorithm to find optimal features | Embeds features during model building process. Feature selection is done bye observing each iteration of model training phase | Knowledge Quality derives from information theory and is calculated using the equation below |
| Much faster compared to wrapper method in terms of TC | High computation time for a dataset with many features | Sits b/w filter method and wrapper method in terms of TC | Notice that the information value for Number Real Estate Loans Or Lines is 0.116 That barely comes within the medium range of predictors and is unproductive |
| Less prone to over-fitting | High chances of over fitting because it involves training of ML models with difference combination of feature | Generally used to reduce over fitting by penalizing the coefficient of a modal I being to large | Variables with medium and high predictive powers are usually selected for model growth. |
| Examples - Correlation, chi-square test, ANOVAs | Exa-forward selection, backward elimination, stepwise selection | Exam-lasso, elastic net, ridge regression | So we pick the app and choose 8 apps |

Table.1.3

## 3. Review of Literature

In This article, the actual project is based Whose main aim was to provide credit risk management To the foundations of finance.

Financial institutions face multiple dilemmas concerning the approval process for loans And apart from taking the risk of granting a loan to a consumer who may become default They even lack the chance to make profit By refusing loan to a customer who can pay his obligations This is why many financial institutions have recently embraced the Credit scoring models capable of identifying hidden trends in very large databases with the goal of classifying clients as default or non-default. Credit scoring model was developed with Oracle Data Miner software kit used for classification using Generalized Linear Model.

The model was created with great predictive confidence and precision. But also provided accurate results regarding the selection of features The microfinance institution therefore agreed to implement this model as a decision-making aid[1]

To improve the solidity and precision of the credit evaluation model We research individual credit risk, pick a Logistic Regression Method And a non-statistical DP-algorithm neural network system Methods which are most widely used by domestic and foreign banks.Additionally,The BP neural network should first be created,And then the performance tests of the model neural network Can be used as a standalone feature,Together with certain characteristic variables, such as the logistic regression model input variables.Ultimately, Logistic regression model to determine customer's personal credit The empirical study shows the following benefits of this approach Predictability is higher than using the Logistic regression model merely;This model's robustness is greater than just using the neural network model BP.The personal credit evaluation model for personal credit rating is therefore of practical significance in the BP-Logistic mixed strategy[2]

The study's principal aim is to test the assumption That network model enhances predictive efficiency Of the classical algorithms for credit scoring. To this end we propose how to build network-based models.Centered on the correlations pair wise between the time observed.Set of financial indicator referring to the different borrowing companies.We extract centrality indicators from these models and use them to Logistic regression augmentation and tree models.Our statement of study is confirmed by empirical analysis.That shows how classical scoring algorithms include network parameters,Such as functional regression and CART, predictive accuracy does also increase[3]

In this paper we demonstrated the value of using the real example-dependent financial cost Linked to credit company when selecting a credit score model In addition, our analyses verified that each ex cost includes Ample and cost-sensitive, example-dependent process As with the Bayes low risk classifier, better results are obtained In the context of greater efficiency, regardless of the algorithm used to estimate the probabilities.Various real-world classification problems are cost-sensitive in nature, depending on example,Where the costs of misclassification differ from one example to another A common example of a cost-sensitive classification is credit scoring.Usually, however, it is handled using approaches that do not take into account the actual financial costs of the lending sector[4]

## 4. Methodology

### A- Data Preparation and Exploratory DataAnalysis

Data research involvesThe analysis is both Univariate and bivariate, From univariate statistics and frequency distributions through correlations, cross-tabulation and study of characteristics.

### B-Discrediting Predictors/Binning

Discrediting is the method of turning a numerical function into a categoric alone, as well as grouping and consolidating categorical features. Example of grouping ‘age’ or same bank feature is shown below

### C-Scorecard — Model Building

Two additional steps are needed before constructing scorecard model.

When a continuous variable is split into few groups, or grouping a discrete variable

Component in a few classes or grouping a singleWe may calculate the Weight of Proof (WOE) value in a few categories for each function Details of WOE calculation are given in the following section

### D-Weight of Evidence (WOE)

WOE tests The intensity of a characteristic function in the distinction between good and poor accounts, which is based on ratio of good and bad applicants in each category. Negative values indicate that a given category isolates a higher ratio of Bad applicants than successful applicants. The difference is calculated in each variable between the ratio of the better and poor.

For example:

**WOE**=In (Distribution of Goods/Distribution of Bad)

## 5. Regression Equation

$$Y = e^{(b0 + b1*x1\ldots bn*xn)} / (1 + e^{(b0 + b1*x1\ldots bn*xn)})$$

a. Y is normal chance

b. b.The explanatory element xi is I

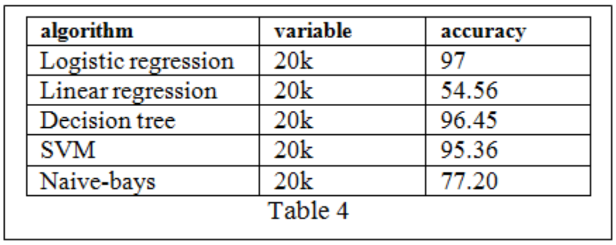
c. bi is the explanatory factor I d regression coefficient.

d. n is the number of variables explanatory to

e. The reasons why logistic regression is ideally suited for analysis of credit risk are

f. LR model is possible, because its values The two ends of the numeral line vary.

## 6. Result and Discussion

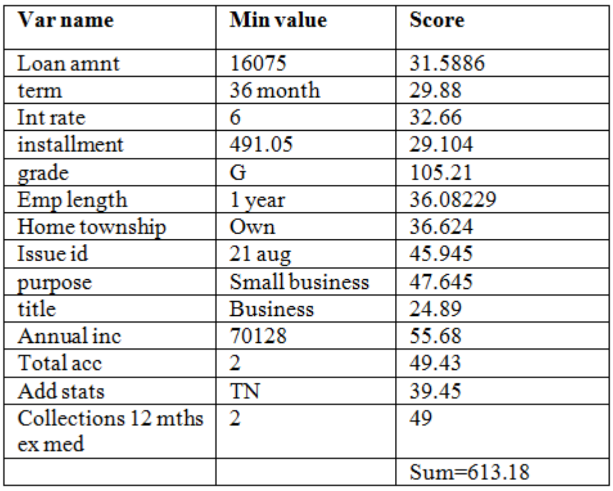


| algorithm | variable | accuracy |
|---|---|---|
| Logistic regression | 20k | 97 |
| Linear regression | 20k | 54.56 |
| Decision tree | 20k | 96.45 |
| SVM | 20k | 95.36 |
| Naive-bays | 20k | 77.20 |

Table 4

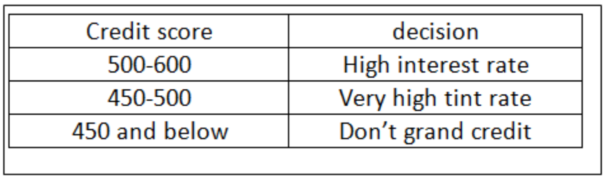
### Decision Making from Scorecard

- Choosing a score cut off that is appropriate for the Loan Product.
- For example: Secured Loans may be approved for a lower score than unsecured loans.
- Scores not only are used for assessment but also for predictive purposes.
- The Model scores combined with business considerations are used for final decision making.



| Var name | Min value | Score |
|---|---|---|
| Loan amnt | 16075 | 31.5886 |
| term | 36 month | 29.88 |
| Int rate | 6 | 32.66 |
| installment | 491.05 | 29.104 |
| grade | G | 105.21 |
| Emp length | 1 year | 36.08229 |
| Home township | Own | 36.624 |
| Issue id | 21 aug | 45.945 |
| purpose | Small business | 47.645 |
| title | Business | 24.89 |
| Annual inc | 70128 | 55.68 |
| Total acc | 2 | 49.43 |
| Add stats | TN | 39.45 |
| Collections 12 mths ex med | 2 | 49 |
|  |  | Sum=613.18 |

Let cut off=600 or more then 600 grand credit



| Credit score | decision |
|---|---|
| 500-600 | High interest rate |
| 450-500 | Very high tint rate |
| 450 and below | Don't grand credit |

Let cut off=600 or more then 600 grand credit

## 7. Conclusion

This paper is based on the actual project whose principal objectivewas to provide finance institutions with the management of credit risk.Financial institutions face lotsof dilemmas With regard to the loan approval process, as apart from taking the risk of authorizing a loan to the customer who can become a borrower defaulter,they still have little hope of profiting from declining loans to a borrower who is capable of paying his obligations. This is whymany financial institutions Lately, credit scoring models are being implemented which are capable of identifying hidden trends in very large databases toClassify customers as default or non-default Credit Scoring System showed surprisingly high accuracy and recognize a 100% of default clients.This model was confirmed by the institution's members who would be used to help the decision making process When authorizing a loan application and thereby predicting default customers

## References

[1] Jasmina Nalić, "Using Data Mining Approaches to Build CreditScoring Model"17th International Symposium INFOTEH-JAHORINA, 21-23 March 2018

[2] Alejandro Correa Bahnsen, Djamila Aouada and Bjorn Ottersten, "Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring, 2014 13th International Conference on Machine Learning and Applications

[3] Huang Weidongl, Zhu Xiangwei2, Su Qinglingl;"Research on Application of Personal Credit Scoring based on BP- Logistic Hybrid Algorithm" 20IO International Conference on Computer Application and System Modeling (ICCASM 2010)

[4] D. Zhang, H. Huang, Q. Chen and Y. Jiang. "Comparison of credit scoring models Third International Conference of Natural Computation, 2007

[5] Shalev-Shwartz, S. and S. Ben-David. Understanding Machine Learning. Cambridge University Press, Cambridge UK. 2014.

[6] Hand, D., H. Mannila, and P. Smyth. Principles of Data Mining. MIT Press,Cambridge,MA.2002.

[7] Shan Liang, Qiao Yang. Dataing Risk Control-Credit Score Modeling Course, Electronic Industry Press.2018

[8] Zhou Zhihua. Machine learningTsinghua University Press. 2015.

[9] J. Sobehart and S. Keenan, "Measuring Default Accurately," Credit Risk SpecialReport, Risk Magazine, Mar. 2001, pp. S31–S33.

[10] A. Kraus, "Recent Methods from Statistics and Machine Learning for Credit Scoring," Doctoral Dissertation, Faculty of Mathematics,Computer Science and Statistics, Ludwig Maximilian University of Munich, Munich (Germany), 2014.