# CREDIT RISK PREDICTION ANALYSIS PREPROCESSING USING CREDIT RISK ASSESSMENT MODEL ALGORITHM

G. Arutjothi [1]*, Dr. C. Senthamarai [2]
[1,2]Department of Computer Applications, Govt. Arts College (Autonomous), Salem-7, Tamil Nadu, India.

## I. INTRODUCTION

Today’s real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results. There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. Data transformations, such as normalization, may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements. These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

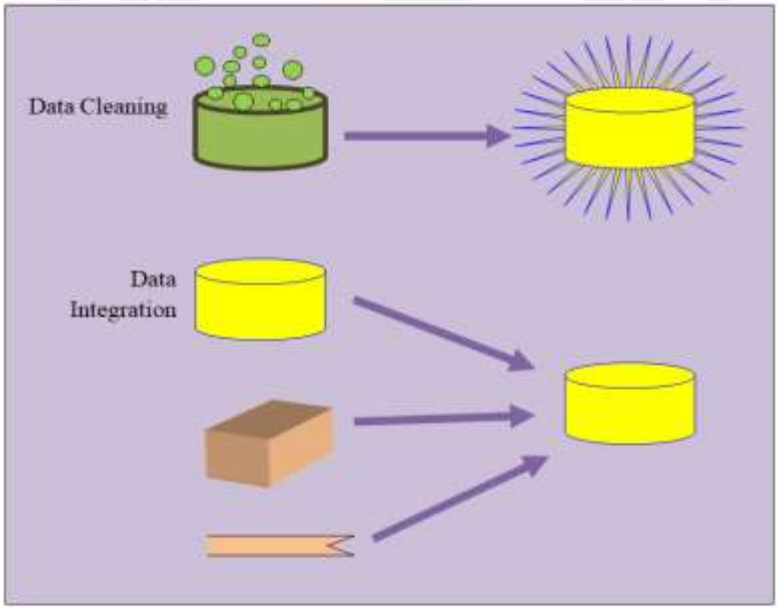

**Figure 1: Forms of Data Preprocessing**

Figure 1 summarizes the data preprocessing steps described here. Note that the above categorization is not mutually exclusive. For example, the removal of redundant data may be seen as a form of data cleaning, as well as data reduction. The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information, e.g., by choosing the default value 'January 1' displayed for birthday. Errors in data transmission can also occur. There may be technology limitations, such as limited buffer size for coordinating synchronized data transfer and consumption.

## II. LITERATURE SURVEY

**Krunal M. Surti , Mr. Ashish Patel** (2017) proposed linear regression with principle based grouping and logistic regression is utilized. The preprocessing is utilized for to perform investigate, dissect and decide the factor that assume pivotal job to discover credit default. A changing model that changes over each passage into vector. Each incentive in the vector is a probability esteem, that is, each element of each credit is changed to a comparing an incentive by measurable procedures or Naive Bayes. They propose a technique to manage URI with no inquiry. It parts URI way string into tokens, so applies Naive Bayes to get their probability esteem. the examination of three algorithms given by scaled conjugate angle back propagation, Levenberg Marquardt and One-step secant back propagation (SCG, LM and OSS). **Simon Lohm¨uller, Fabian Rabe, Andrea Fendt, Bernhard Bauer, Lars Christoph Schmelz[2018]:** Proposed to improve SON Management models with psychological Machine Learning (ML) strategies. In this way, the recreated conduct of three diverse SON Functions is investigated and depicted by a Linear Regression (LR) Model. In a second step, execution information of network cells are broke down for similitudes utilizing k-Means Clustering. The findings of these two steps are then consolidated by fitting the models onto littler bunches of cells. At last, the utility of these models for foreseeing the execution of the network is assessed and the diverse phases of refinement are contrasted and one another. **M.Ozgur Cingiz, Ahmet Unudulmaz, Oya Kalıpsız[2013]:** Proposed prediction of task issue impacts that can cause misfortune in software venture regarding their qualities on risk factors and likewise they need to rank our risk components to see how they can give insight concerning venture issue impacts independently. For these reason five grouping strategies for prediction of issue effect and two channel highlight choice techniques for positioning significance of risk factors are utilized. Software risk the executives expects to find unusual risks before the beginning of software venture. Software risk variables can have diverse effect for progress rate of software ventures. Some risk elements can impact achievement rate of venture essentially however some others can't be. These results are like related works that shows SVM and MLPs achievement rates are higher than different classifiers achievement rates. **Wei Pang, Xiaofang Xie, Pengfei Fan, Jiaqi Liu[2016**]: Proposed a versatile work refinement and collocation focuses strategy the improvement issue into various lattices, utilize the Lagrange interjection technique to estimated the advancement issue in each work, and then figure the discrete errors. In view of the discrete errors, separate the most extreme error of the work into two equivalent networks and increment the collocation focuses for the rest networks, which does not meet the resilience, to improve the precision of the arrangement. At the point when the state factors and the control factors are non-smooth, utilize the worldwide collocation directs insertion polynomial toward take care of the issue, which may cause extraordinarily influence of the exactness.

## III. TYPES OF DATA SOURCES

First, let us have a closer look at what data to gather. Data can originate from a variety of different sources and provide different types of information that might be useful for the purpose of credit risk modeling, as will be further discussed in this section. We discuss the most prominent data sources and types of information available in a typical organization, but clearly not all possible data sources and types of information. Furthermore, some overlap may exist between the enlisted categories.

### Transactional data

It’s a first important source of data. It consists of structured and detailed information capturing the key characteristics of a customer transaction (e.g., cash transfer, installment payment). It is usually stored in massive online transaction processing (OLTP) relational databases. This data can also be summarized over longer time horizons by aggregating it into averages, absolute or relative trends, maximum or minimum values, and so forth.

### Contractual, subscription, or account data

It may complement transactional data. Contractual data includes information about the type of product (e.g., loan) combined with customer characteristics. Examples of subscription data are the start date of the relationship, characteristics of a subscription such as type of services or products delivered, levels of service, cost of service, product guarantees, and insurances. The moment when a customer subscribes to a service offers a unique opportunity for the organization to get to know the customer unique in the sense that it may be the only time when a direct contact exists between the bank and the customer, either in person, over the phone, or online and as such it offers the opportunity to gather additional information that is nonessential to the contract but may be useful for credit risk modeling. Such information is typically stored in an account management or customer relationship management (CRM) database.

### Sociodemographic Information

Subscription data may also be a source of sociodemographic information, since subscription or registration typically requires identification. Examples of socioeconomic characteristics of a population consisting of customers are age, gender, marital status, income level, education level, occupation, and religion. Although not very advanced or complex measures, sociodemographic information may significantly relate to credit risk behavior. For instance, it appears that both gender and age are very often related to an individual’s likelihood to default: Women and older individuals are less likely to default than men and younger customers. Similar characteristics can also be defined when the basic entities for which default is to be detected do not concern individuals but instead companies or organizations. In such a setting one rather speaks of slow-moving data dimensions, factual data, or static characteristics. Examples include the address, year of foundation, industrial sector, and activity type. These do not change over time at all, or do not change as often as do other characteristics such as turnover, solvency, number of employees, and so on. These latter variables are examples of what we will call behavioral information.

### Data Poolers

In recent times, data poolers have increased in importance in the credit risk modeling industry. Examples are Experian, Equifax, CIFAS, Dun & Bradstreet, Thomson Reuters, and so on. The core business of these companies is to gather data (e.g., sociodemographic information) in particular settings or for particular purposes (e.g., credit risk assessment, fraud detection, and marketing) and sell it to interested customers looking to enrich or extend their data sources. In addition to selling data, these data poolers typically also build predictive models themselves and sell the outputs of these models as risk scores. This is a common practice in credit risk; for instance, in the United States the FICO score is a credit score ranging between 300 and 850 provided by the three most important credit data poolers or credit bureaus: Experian, Equifax, and TransUnion.

### Surveys

Surveys are another source of data, and this information is gathered via offline methods such as mail, or via online modes including telephone, website, and social media interactions (e.g., Facebook, LinkedIn, or Twitter). Surveys may aim at gathering sociodemographic data, but also behavioral information.

### Behavioral information

It concerns any information describing the behavior of an individual or an entity in the particular context under study. Such data is also called fast-moving data or dynamic characteristics. Examples of behavioral variables include information with regard to preferences of customers, usage information, frequencies of events, and trend variables. When dealing with organizations, examples of behavioral characteristics or dynamic characteristics are turnover, solvency, or number of employees. Marketing data results from monitoring the impact of marketing actions on the target population, and concerns a particular type of behavioral information.

### Unstructured Data

Unstructed Data embedded in text documents (e.g., e-mails, web pages, claim forms) or multimedia content can be interesting to analyze. However, these sources typically require extensive preprocessing before they can be successfully included in a credit risk modeling exercise. Analyzing textual data is the goal of a particular branch of analytics (i.e., text analytics). Given the high level of specialization involved, this book does not provide an extensive discussion of text mining techniques. For more information on this topic, you could consult academic textbooks on the subject.

### Contextual or Network Information

A second type of unstructured information is contextual or network information, meaning the context of a particular entity. An example of such contextual information concerns relations of a particular type that exist between an entity and other entities of the same or a different type. An example in credit risk modeling could be liquidity dependencies between corporate counterparts. Taking into account these complex network relationships allows us to model system risk whereby the default of one company may create a knock-on effect in the network of interconnected companies.

### Qualitative Expert Based Data

Another important source of data is qualitative, expert-based data. An expert is a person with a substantial amount of subject matter expertise within a particular setting (e.g., credit portfolio manager, brand manager). The expertise stems from both common sense and business experience, and it is important to elicit this knowledge as much as possible before the credit risk model building exercise commences. It will allow for steering the modeling in the right direction and interpreting the analytical results from the right perspective. A popular example of applying expert-based validation is checking the univariate signs of a regression model. For instance, an example already discussed relates to the observation that a higher age often results in a lower credit risk. Consequently, a negative sign is expected when including age in a default risk model yielding the probability of an individual being a defaulter. If this turns out not to be the case (e.g., due to bad data quality or multi-collinearity), the expert or business user will not be tempted to use the credit risk model at all, since it contradicts prior expectations.

## IV. PROPOSED ALGORITHM

### Credit Risk Assessment Model Algorithm

As indicated by issues of information class imbalanced and information repetition in credit hazard evaluation, we propose a credit chance appraisal technique utilizing DNN's with grouping and combining examining calculation as appeared in Figure 2. The technique comprises of two sections: Data Equalization and DNN's characterization. In the initial segment, the information is bunched into various subgroups to accomplish contrast between info information by grouping and combining examining calculation. In the second part, various DNN's (base classifiers) are prepared to accomplish decent variety of classifiers utilizing distinctive preparing sets that are numerous bunched subgroups. The combining technique depends on lion's share casting a ballot strategy, which takes larger part casting a ballot expectation class as forecast result. The point by point calculation of the model is appeared underneath.

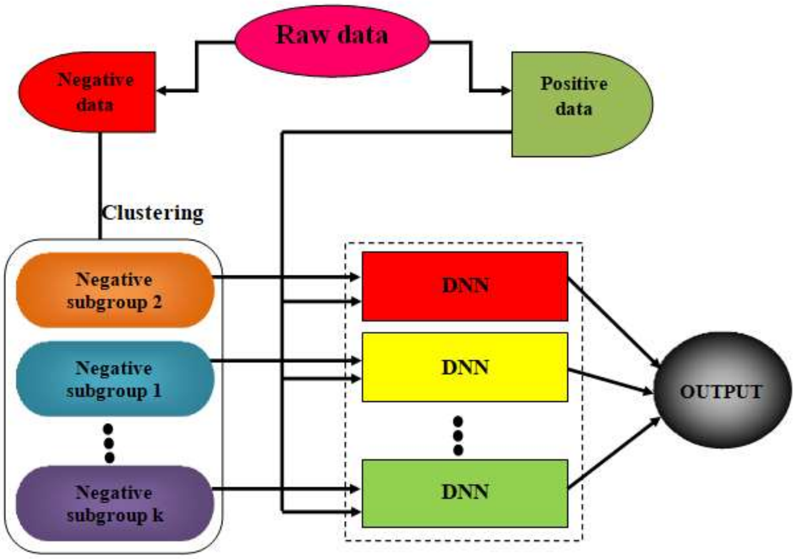

**Figure 2: Credit Risk Assessment Model Structure**

**Algorithm:** CREDIT RISK ASSESSMENT MODEL ALGORITHM

Input : Raw dataset *D*, number of Clustering center *k*

Output : Integration of DNN learning algorithm *L*

1. Raw dataset *D* is divided into training set $Z= \{x_i,y_i\}, i=1,2,.....n$ and test set $T=\{x* i,y* i\}, i=1,2,....m,$ and *Z* is divided into positive class samples *Q* and negative class samples *P*.
2. *Q* is clustered into k subgroups $A_i(i^{\epsilon}k)$ via k-means algorithm
3. *for (i=0; i<=k; i++)* {
4. Balanced subgroups $R* i$ is formed by the merger of $A_i$ subgroups and positive class samples *P*.
5. Balanced subgroups $R* i$ is classified using the DNN model $C_i$
6. }
7. The integrated algorithm *L* is composed of all base DNN $\{C_1, C_2, .... C_k\}$and *T* is classified by majority voting method:

$$y *= arg \max_{y} \sum_{i,\epsilon C} C_i(x^*, y)$$

## 4.4 EXPERIMENTAL RESULTS

### Credit Utilization

| Payment Month | Payment Amount |
|---|---|
| **10** | **2500** |
| **20** | **3900** |
| **30** | **3180** |
| **40** | **4632** |
| **50** | **4890** |

**Table 4.1: Table of Credit Utilization**

The table of credit utilization describes payment month and payment amount. 10th amount is 2500, 20th amount is 3900, 30th month amount is 3180, 4th month amount is 4632 and 5th month amount is 4890.

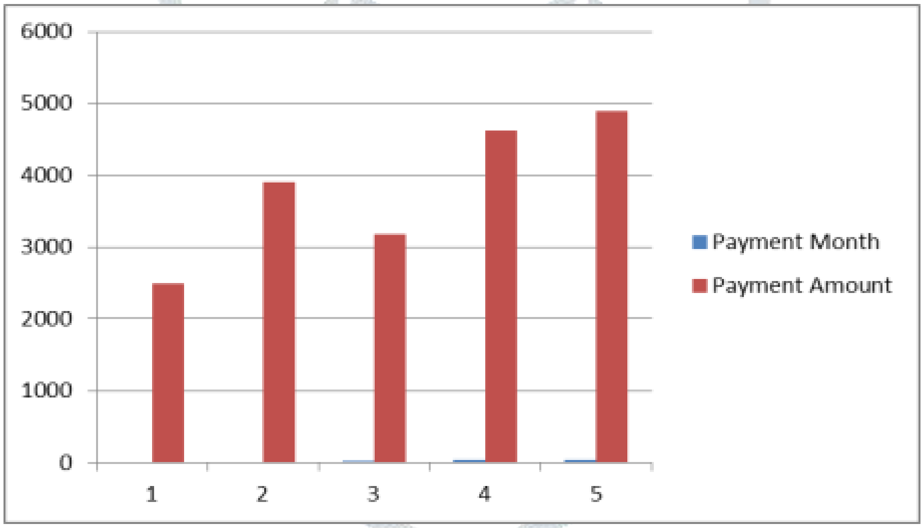

**Figure 4.3: Comparison chart of Credit Utilization**

The chart of credit utilization explain payment month and payment amount. Payment month in X axis and payment amount in Y axis.10th amount is 2500, 20th amount is 3900, 30th month amount is 3180, 4th month amount is 4632 and 5th month amount is 4890.

## Loan Growth

| No of months | Loan Amount | Payment Amount |
|---|---|---|
| 12 | **17500** | **16900** |
| 24 | **16300** | **15800** |
| 36 | **15000** | **12350** |
| 48 | **15900** | **14670** |
| 60 | **14210** | **11001** |

**Table 4.2: Table of Loan Growth**

The table of loan growth describes No of month, loan amount and payment amount. 12th month loan amount is 17500 and payment amount is 16900, 24th month loan amount is 16300 and payment amount is 15800, 36th month loan amount is 15000 and payment amount is 12300, 48th month loan amount is 15900 and payment amount is 14670, 60th month loan amount is 14210 and payment amount is 11001.

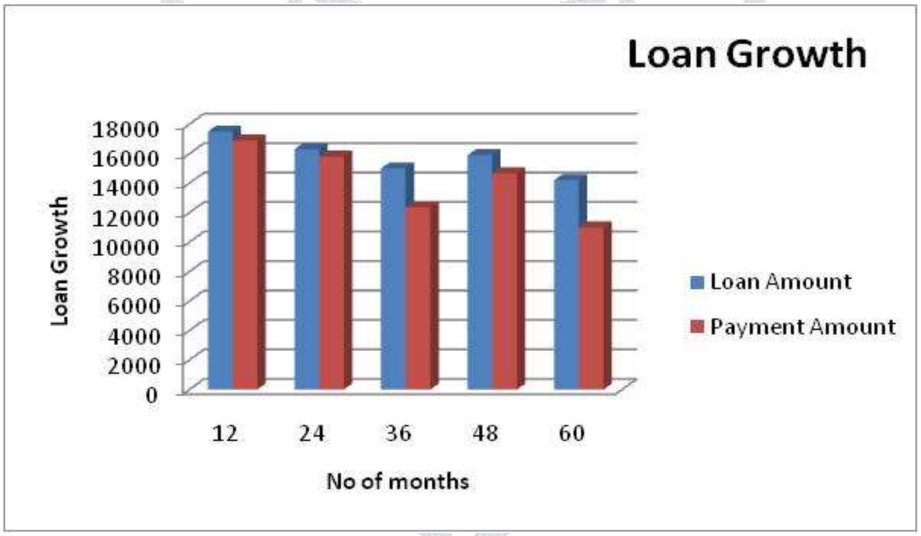

**Figure 4.4: Comparison chart of Loan Growth**

The chart of loan growth explain payment No of months, loan amount and payment amount. No of months in X axis and loan growth in Y axis. 12th month loan amount is 17500 and payment amount is 16900, 24th month loan amount is 16300 and payment amount is 15800, 36th month loan amount is 15000 and payment amount is 12300, 48th month loan amount is 15900 and payment amount is 14670, 60th month loan amount is 14210 and payment amount is 11001.

## Expenses Ratio

| No of months | Credit Amount | Expenses Amount |
|---|---|---|
| **10** | **50000** | **39000** |
| **20** | **150000** | **132000** |
| **30** | **100000** | **78000** |
| **40** | **200000** | **186000** |
| **50** | **250000** | **230000** |

**Table 4.3: Table of Expenses Ratio**

The table of expenses ratio describes No of month, credit amount and expenses amount. 10th month credit amount is 50000 and expenses amount is 39000, 20th month credit amount is 15000 and expenses amount is 132000, 30th month credit amount is 100000 and expenses amount is 78000, 40th month credit amount is 200000 and expenses amount is 186000, 50th month credit amount is 250000 and expenses amount is 230000.

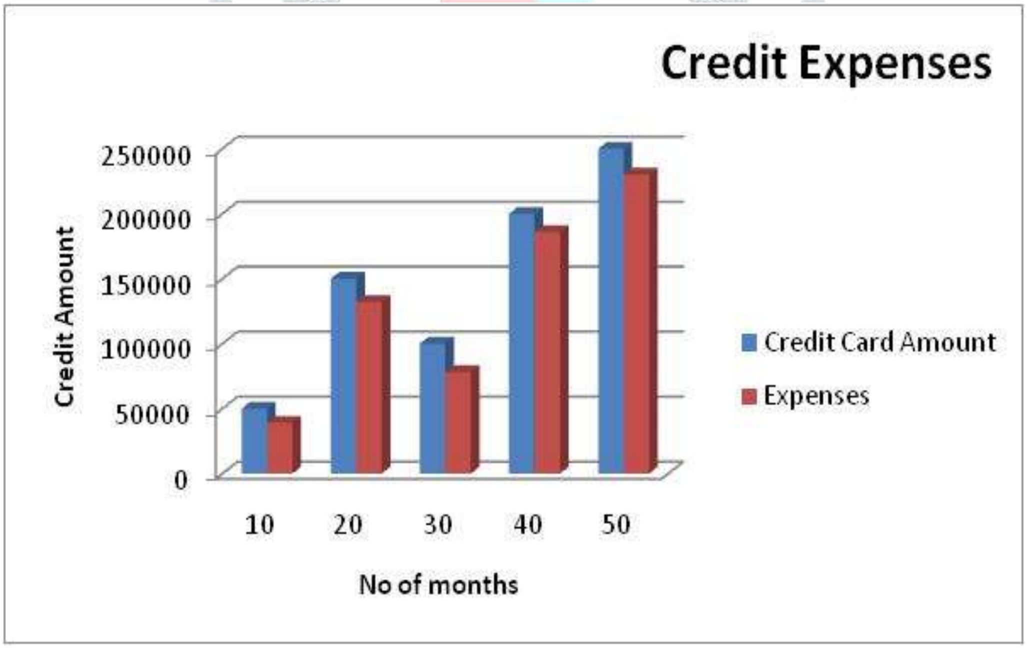

**Figure 4.5: Comparison chart of Credit Expenses**

The chart of credit expenses explain payment No of months, loan amount and payment amount. No of months in X axis and credit amount in Y axis. 10th month credit amount is 50000 and expenses amount is 39000, 20th month credit amount is 15000 and expenses amount is 132000, 30th month credit amount is 100000 and expenses amount is 78000, 40th month credit amount is 200000 and expenses amount is 186000, 50th month credit amount is 250000 and expenses amount is 230000.

**Limit Bal Vs Default Payment**

| month | Male | Female |
|---|---|---|
| **1** | **14500** | **16000** |
| **2** | **10300** | **11900** |
| **3** | **13450** | **18330** |
| **4** | **10450** | **15119** |
| **5** | **11540** | **16890** |

**Table 4.4: Table of Limit Bal Vs Default payment**

Table of Limit Bal Vs Default payment describes the month, male and female. 1st month male payment 14500 and female payment 16000.2nd month male payment 10300 and female payment 11900.3rd month male payment 1340 and 18330.4th month male payment 10450 and female payment 15119.5th month male payment 11540 and female payment 16890.

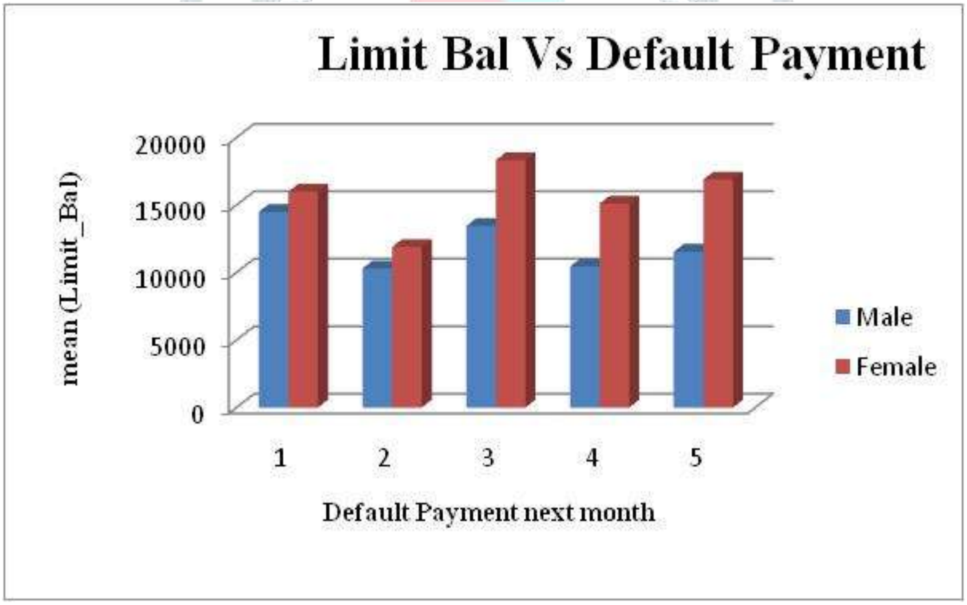

**Figure 4.6: Chart of Limit Bal Vs Default payment**

The chart of limit bal Vs default payment explain the month, male and female. Default payment next month of months in X axis and mean(Limit_Bal) in Y axis. . 1st month male payment 14500 and female payment 16000.2nd month male payment 10300 and female payment 11900.3rd month male payment 1340 and 18330.4th month male payment 10450 and female payment 15119.5th month male payment 11540 and female payment 16890.

# V. CHAPTER SUMMARY

Credit risk assessment is important for financial institutions, which helps them to decide whether or not to accept loan applications from customers. DNN’s are used to establish multiple layer network structure evaluation model in credit risk assessment field, in which can directly obtain feature information to improve accuracy of classification from a large number of customer credit data. Since the general evaluation methods are limited to class imbalance problem of classifiers, clustering and merging sampling algorithm is proposed for generating balanced data and maintaining data diversity. From the experimental results, we can see that the proposed evaluation model has higher classification accuracy and better evaluation performance of imbalanced credit data.

**References**

1. Sudhakar M , Dr. C. V. K Reddy, “Two Step Credit Risk Assesment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 3, March 2016.

2. Anand Motwani, Goldi Bajaj, Sushila Mohane, “Predictive Modelling for Credit Risk Detection using Ensemble Method”, International Journal of Computer Sciences and Engineering Vol.6(6), Jun 2018, E-ISSN: 2347-2693.

3. Sivasree M S, Rekha Sunny T, “Loan Credibility Prediction System Based on Decision Tree Algorithm”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 4 Issue 09, September-2015.

4.Manjeet Kumar, Vishesh Goel, Tarun Jain, Sahil Singhal, Dr. Lalit Mohan Goel,“Neural Network Approach To Loan Default Prediction”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 04 | Apr-2018.

5. Sharjeel Imtiaz, Professor Allan J. Brimicombe, “A Better Comparison Summary of Credit Scoring Classification”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.7, 2017.

6. Krunal M. Surti , Mr. Ashish Patel, “Effective Credit Default Scoring using Anomaly Detection”, IJSTE - International Journal of Science Technology & Engineering | Volume 3 | Issue 11 | May 2017 ISSN (online): 2349-784X.

7. Simon Lohm¨uller, Fabian Rabe, Andrea Fendt, Bernhard Bauer, Lars Christoph Schmelz**,** “SON Function Performance Prediction in a Cognitive SON Management System”, 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW): 7th International Workshop on Self-Organizing Networks (IWSON).

8. M.Ozgur Cingiz, Ahmet Unudulmaz, Oya Kalıpsız, "Prediction of Project Problem Effects on Software Risk Factors", 12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques.

9. Wei Pang, Xiaofang Xie, Pengfei Fan, Jiaqi Liu**,** "An Adaptive Collocation Method and Mesh Refinement for Solving Non-smooth Trajectory Optimization problems", 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics.

10. Irit Nowik**,** "The Game on the Risk in Deviating from Nash Equilibrium", Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management.

11. Joshua T. Cook, Laura E. Ray and James H. Lever, "Dynamics Model for Mobility Optimization and Control of Off-Road Tractor Convoys", 2016 American Control Conference (ACC) Boston Marriott Copley Place.

12. Yahya Choua, Laurent Santandréa, Yann Le Bihan, and Claude Marchand, "Mesh Refinement in Eddy Current Testing With Separated T-R Probes", IEEE TRANSACTIONS ON MAGNETICS, VOL. 46, NO. 8, AUGUST 2010.

13. Jihoon Seo, Juyul Lee, and Keunyoung Kim "Decoding of Polar Code by Using Deep Feed-Forward Neural Networks" IEEE2018,2018 workshop on computing,Networking and Communications.

14. Ghulam Mohi Ud Din, Angelos K. Marnerides**,** "Short Term Power Load Forecasting Using Deep Neural Networks"IEEE2017, 2017.International Conference on Computing, Networking and Communications (ICNC): Green Computing, Networking, and Communications.

15. Carlos Bentes, Domenico Velotto, Susanne Lehner**,** "TARGET CLASSIFICATION IN OCEANOGRAPHIC SAR IMAGES WITH DEEP NEURAL NETWORKS: ARCHITECTURE AND INITIAL RESULTS"IEEE2015.

16. Nimish Shah, Paragkumar Chaudhari, and Kuruvilla Varghese, " Runtime Programmable and Memory Bandwidth Optimized FPGA-Based Coprocessor for Deep Convolutional Neural Network"IEEE2018. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

17. Nima Amini, Seyyed Ali Seyyedsalehi, " Manipulation of attractors in feed-forward autoassociative neural networks for robust learning"IEEE2017.

18. Paul M. Baggenstoss " On the Duality Between Belief Networks and Feed-Forward Neural Networks"IEEE2018.

19. Yiming Tian, Xitai Wang "SVM ensemble method based on improved iteration process of Adaboost algorithm"IEEE2017.

20 Mücahid Barstugan, Rahime CEYLAN "A Discriminative Dictionary Learning-AdaBoostSVM Classification Method on Imbalanced Datasets"IEEE2017.