ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.6

Analysing Credit Risk using Statistical and Machine Learning Techniques

Amit Chhotelal Vishwakarma[1], Ramesh Solanki[2]
MCA Student[1], Assistant Professor[2]
Department of Master of Computer (MCA) Application
Vivekanand Education Society of Information Technology, Mumbai, Maharashtra, India

Abstract:
Credit risk is associated with the lending institutions, banks, and customers requesting for a loan. It's a calculated risk which lending institutions and banks take while giving a loan to the customer. Lenders such as banks while considering a loan application from customer checks various factors such as customer Credit score, Banks statements, Any previous loan repayment schedule etc. Based on the various factors lenders take the risk of giving credit to the customer. In this paper, we will discuss various factors that are taken into consideration while giving a loan to the customer. Further will discuss a various method which has been introduced over the year which defines the creditworthiness of the customer. In next section will discuss statistical and machine learning algorithm which are used to check creditworthiness. Then will compare the algorithms based on the review from previously done paper as the data of public is difficult to obtain due to confidentiality concern. The main objective of the paper is to study and review some related statistical learning method and machine learning methods in context to credit risk and the process of model creation and deployment.

Keywords: Credit Risk, Data Mining, Logistic Regression, Linear Regression, Linear Discriminant Analysis, ANN, Decision Tree.

I. INTRODUCTION

Credit approval is a tedious task in financial Industries like banks and lending institutions. Identifying the customer which are defaulters before giving loan is an appreciable and troublesome task of the bankers. Bankers deciding whether to accept or reject customer credit request is commonly executed by judging technique, credit scoring and credit risk predictive modelling by machine learning. Earlier, most Banks and lending institute used the method of judging technique that is based on the five characteristics of credit also known as 5C's of credit [1]. The 5C's of credit methodology defines the creditworthiness of potential borrowers. The judgmental techniques analysis five characteristics of the customer and various other condition of a loan and together with this factor it tries to estimate the customer of being a defaulter. The 5 C's of credit methodology that is used for evaluating a customer is based on both qualitative and quantitative measures that look at a customer's credit score, credit reports, income statements and other documents relevant to the customer's financial situation, and also consider information about the loan itself. The 5 C's are [2]

Character:Character defines the first characteristics of credit methodology; It refers to a borrower track record for repaying the debts. All the information regarding customers loan or credit is present on customer credit reports which are generated by the major credit bureaus CIC (credit information company). A credit report contains all the information regarding how much the customer as borrowed in the past and whether he has repaid the credit amount on time and with interest. These reports also contain information regarding the credit instalment was paid in time, judgment by experts, and bankruptcies. Credit scores is another tool that lending institute use to get a quick snapshot of customers credit worthiness before looking at actual credit reports.

Capacity: Capacity evaluates the borrower's ability to repay a loan by comparing income against recurring debts and assessing the borrower's debt-to-income (DTI) ratio. In addition to examining income, lenders look at the period that an applicant has been at his job and stability of the job.

Capital:Lending institution considers the proportionate contribution of capital from the customer with respect to potential credit loan amount. A large contribution by the customer determines fewer chances of customer being defaulter. For example, a customer who has down payment amount for the home they find it easier to get a home loan. Down payment indicates the customer interest and seriousness in home buying which make the lender and banks comfortable in approving loan amount.

Collateral: A collateral is a type of loan secured against the borrower's property (home) through a written note of indebtedness. Collateral is always seen as an extra security for the lender in case the customer(borrower) defaults on the loan. It simply means that the customer is asking credit or loan against its own assets. The lender has the right to seize the property if the borrower fails to repay the loan or defaults in the monthly EMI's. Loans that are secured by collateral typically have a lower rate of interest than unsecured loans.

Condition:Condition plays a crucial part while the lenders are assessing the customer. The condition can be of loan, customer etc. Condition of loan refers to interest rate and the principal amount which influences the lenders to desire to finance the customer. The condition for customer refers to how and what purpose a customer intends to use the money. An example such as, when the customer applies for the mortgage loan or a car loan the lenders or banks are more likely to approve the loan because of their specific purpose and disapprove loan with unspecified purpose. Credit score

which is represented by the 3-digit number is computed by the external credit bureaus and rating agencies through the information collected from a credit report. Credit report consist of all the information of customer credit loan payment pattern, any late payment of loan taken, defaulter, financial history and current financial situation. Based on the credit score it is decided whether the person is eligible for the loan or not but there are no exact specifications on what constitutes a "good" score from a "bad" score. For an example, CIBIL uses a three-digit credit score ranging from 300 to 900, more the credit score better chances for the loan approval. However, it does not tell you the level of risk for the lending you may be considering it just gives a brief knowledge about the financial status of a person. Normally credit scoring model is used as a supportive element along with many other factors. [3] Credit risk analysis (Financial risk analysis or Loan default risk analysis) and Credit risk management are important aspects of lending institutions (mortgage industry or banks) which provide a loan to customers (individual or businesses). Credit loan has a risk of being defaulted so it is important to understand the customers and credit risk associated with them. For a better understanding of the risk level of the customers taking credit loan, lending institution normally collects a large amount of customer data. This large amount of data collected is utilized in understanding the customer credit risk. In today computerized system this process of deciding can be optimized using statistical and machine learning techniques. Thus, lending institution to improve the process of accessing creditworthiness of customers develops various credit analysis model based on various statistical and machine learning algorithms. Credit analysis model helps the lending institution to categories customers either as good creditors (a group of customers that are more likely to repay the loan) and bad creditors (a group of customers who are having a high possibility of defaulting on a loan or any other financial obligation).Derivation of credit risk analyzing model requires large data and data mining techniques. These models along with demographic data, statistical techniques and payments data can help in identifying the important characteristics, which are related to credit risk, and assigning score or status to customers. The probability of a customer being default must be calculated from the information collected at the time of registering for loan and the information collected using statistical models and machine learning algorithms. Both this information collected will serve as the base for customer creditworthiness. In this paper, the statistical methods which were reviewed for credit risk analysis are a Linear regression, Logistic regression, Linear discriminant analysis, and SVM. And Machine learning or Artificial Intelligence methods, reviewed are ANN, Deep learning, Random Forest and Gradient Boosting Machine. This paper also discusses behavioral scoring method (e.g. Bayesian Behavior Model). This scoring model makes a decision about the customer based on repayment performance of existing customers during the certain predefined period of time. Ensemble modelling is also discussed in this paper. Ensemble modelling is a way to improve your prediction model. Ensemble learning technique provides better prediction accuracy and also classification ability, therefore, is widely applied to credit evolution.This paper also describes the various algorithm and model that can be used for the improvisation of financial fraud detection. The main reasons for selecting this feature are to improve computationally and the actual solution to bring out best performance, as well as providing a better understanding of the problem. An algorithm such as Feature ranking can be used to assign a rating to individual features on the basis of certain attributes such as accuracy, content and consistency and choose a suitable subset of features based on ranking. Performance metrics are used measures a small increase in performance which can lead to large economic benefits. Classification method such as accuracy, sensitivity, specificity, precision, the false positive rate is the method used for the performance measure. This paper studied various algorithms such as Genetic algorithm, Decision tree, Support vector machine etc. for specifying the best prediction method. It has been found that misclassification of any features or attributes costs is high, techniques with a higher sensitivity such as GP (Genetic programming) or neural networks may be suitable choices. If receptiveness to minor changes in the dataset is desired then the neural networks could be appropriate. Overall the support vector machine could be considered to have the best performance with the highest accuracy As it is very difficult to obtain data of user's personal information from banks due to confidentiality concern. In this paper, we use the data provided by the mobile store which sells mobile in instalments. The study applies the credit scoring techniques using the data mining of payment history of the customers of the mobile store. Credit scorecard model, Logistic regression, and Decision tree are compared on the basis of classification performance. The result of the classification performance error rates was 26.7%, 28.2%, and 27.1% respectively. The threshold which is known as the cut off score can be determined by the value of K-S Test for each bucket of the score in the validation sample. The target variable is payment status which is a binary variable with two categories: default (comprise of the customer which have defaulted the payment) and nondefault (comprise of the customer which not defaulted any payments) which were represented by numerical values (1 and 0). Out of 2550 customers, 35% were found to be defaulters. The majority of the members were male (80%) and more than half (68%) of the customer is from non-government sector. The main issues related to the performing credit scoring are the availability of data and selecting the sample from the available data. Credit scoring model is a classification technique in which the learning process is supervised by the vector value which is the outcome of known training data. A decision tree is a machine learning algorithm in which each branch represents the classification question and the leaves represent the specific data which a possible outcome of a single or multiple classification questions can be. The outcome of the classification question is based on the variable given to the question with a specific parameter. A decision tree can handle missing values without any difficulties and can easily divide the data on each branch without losing any of the data. From the study, it is known that the decision tree and the linear regression techniques are the best suitable for credit scoring models and the decision tree have an advantage over linear regression as it is easy to understand and to track the leaf node outcomes back to the root node (parent node). While a customer applies for a loan the lending institution or banks evaluate the customer creditability to repay the loan amount with the interest amount. Lending institute normally measures the profitability and leverage to access credit risk. A profitable lending firm generates enough income to cover the expense of giving the loan and collecting back the loan amount with interest. However, this is not always the case that just checking the parameter which is defined for the classifying the good or bad customers that the customers are liable to get a loan. There are various other factors that are taken into consideration so that the lending institution should make a profit even when the customer defaults. These factors include

other financial information such as liquidity ratio, behavioral information, cost of collection information, excepting interest on the load which have been approved etc. Another most important thing is to update the database timely because some good customer can turn bad and bad customer can turn good. So, for not missing out any reliable customer the database should be updated always. In this paper, Traditional statistical learning and Machine learning method differences are discussed. The common objective of this tool is to learn from data. Both the approaches aim to investigate the underlying relationship of a various factor that defines customers goodwill by using a training dataset. But the difference is that the statistical learning methods assume formal relationships between variables which can be represented in the form of mathematical equations, while on the other hand machine learning methods can learn from data without requiring any rules-based programming. Based on comparison machine learning techniques is more flexible and can better fir the patterns in data. [4]

II. STATISTICAL LEARNING TECHNIQUESS

A. Linear Regression

Linear regression model is used in various application. One of the vital space in which linear regression is employed is credit risk analysis. In statistical learning technique, Linear regression is a method which is used to predict a target variable by fitting the best linear relationship between dependent and independent variable by fitting the best line. The best fit line is known as regression line and is done by making sure that the sum of all the distance between the shape and actual observation at each point of contact is same as possible. Linear regression is of two types: Simple Linear regression (which is characterized by one independent variable) and Multiple linear regression (which is characterized by more than 1 independent variable). [5]

Linear regression method is represented by linear equation y=a * X + b.

This equation represents as:

- y = Dependent variable
- a = Slope
- X = Independent variable
- b = Intercept

Coefficients a and b are derived by minimizing the sum of the squared difference of distance between data points and regression line. The example below shows how weight of the person is calculated using height

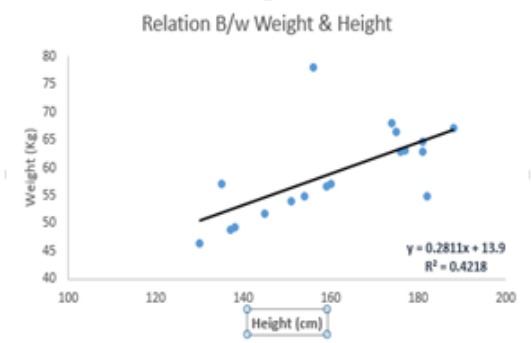

Figure 1

The best fit line has been identified using linear equation y = 0.2811x + 13.9. Now using this equation, we can find the weight with reference to the know height. Logistic regression such as binary and dichotomous, the response variable y∈ {0,1} follows a Bernoulli distribution. The response variable reflects the creditworthy of the customer:

$$Y = \begin{cases} 1, & if\ a\ customer\ goes\ to\ collection \\ 0, & if\ a\ customer\ coontinues\ to\ make\ payemnents \end{cases}$$

The observations $y = (y_1,...y_n)^T$ should be independent, where n indicates the number of the observations. The above assumption is better for the dataset of banks because potential customers and their characteristics are totally independent from each other because only one person per household is able to receive a loan.

where β is a $(m + 1) \times 1$ regression coefficient vector and X is a $n \times (m + 1)$ model matrix containing n observations of m explanatory variables and a constant term.

Let $p_i = P(y_i = 1)$ be the probability of $y_i = 1$

$\text{logit}(p_i) = \log(p_i / (1 - p_i))$ be the logit link function for the observation which is *i.*

The logistic regression model has a linear form for the logit:

$$(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \quad (1)[1]$$

In the above equation x_i is a $m+1$ vector, which contains 1 and m categorical or continuous explanatory variables.

Using the above equation (1.1) probability of $p_i$can be derived by using the exponential function:

$$p_i = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} \quad (2)\ [1]$$

B. Logistic Regression

Logistic regression technique was developed by David Cox. It is one of the most frequently used Statistical learning technique. Logistic regression is 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. Don't get confused by the name of the technique, it is a classification technique rather regression algorithm. This technique is used to estimate discrete values such as 0/1, true/false or yes/no based on given set of independent variables(s). In simple terms, the techniques predict the probability of occurrence of an event by utilizing the data into a logit function. Hence, it is also known as logit regression and its output values lie between 0 and 1 as it predicts the probability [6]. Now, let us try and understand this through a simple example, let's say you have a puzzle to solve and there are possibly only 2 outcomes – either you solve it or you don't. Now imagine, that you are being given a set of puzzles or quizzes in an attempt to understand that in which subjects you are better. The outcome of this study would be somewhat like this – if you are given a science-based tenth-grade problem, you are 60% likely to solve it. On the other hand, if it is grade eight history question, the probability of getting an answer is only 40%. So, Logistic Regression provides you with the most likelihood of any question.

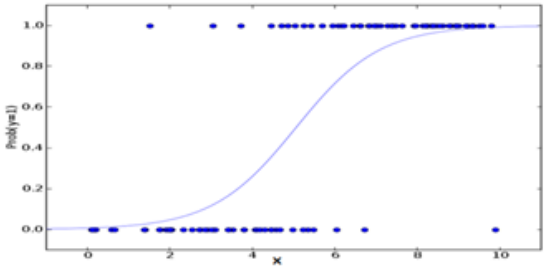

Figure 2

As credit risk analyzingis a binary problem we wish to reduce this outcome to 0 or 1. Logistic regression achieves this by applying a logistic transformation, which restricts the output from [-∞, +∞] to a probability between 0 and 1. In credit risk when there are only two outcome groups (i.e. good and bad) binary logistic regression is used. Multinomial logistic regression refers to cases where more than 2 outcome groups are used (i.e. good, indeterminate, bad). Binary logistic regression equation is represented as below

$$g(x) = \ln\left(\frac{p_g}{1-p_g}\right) = b_0 + b_1x_1 + \ldots + b_nx_n \quad (1)\ [2]$$

$p^g$where $p_g$is the probability of belonging to the good class and is called the *odds* $1 - p_g$*ratio* and $g(x)$ is the logit transform of p_g. The logit transform is a link function used to relate the probabilities of group membership to a linear function of the input features. Many of the Linear regression desirable properties can be seen in Logit such as it is linear in its parameters; may be continuous; and may range from [-∞, +∞] depending on the range of x the logit transform is not the only link function available, for example *pro-bit* and *to-bit* have been used in credit risk analyzing However, logit is the easiest to interpret and generally there is little dissimilarity between it and the performance of the pro-bit and to-bit link [7] The regression coefficients ($b0$ to b_n) are derived using the maximum likelihood estimation (MLE) method. TheMLE is an iterative andintensive calculation approach. It initially begins with guessing the coefficients values and iteratively changes these values to attain the maximize likelihood The logistic model produced in above equation refer to (2.1) can be manipulated to estimate the probabilities of class membership ($p_g$and p_b). The first step is to express the probabilities of class membership in terms of the input features directly:

$$p_g = \frac{exp(b_0 + b_1x_1 + \ldots + b_nx_n)}{(1 + exp(b_0 + b_1x_1 + \ldots + b_nx_n))} \quad (2)\ [2]$$

And the probability of belonging to the bad class:

$$p_b = \frac{1}{(1 + exp(b_0 + b_1x_1 + \ldots + b_nx_n))} \quad (3)\ [2]$$

In the next step, the constant $b0$ and the regression coefficients $b1$ to b_n are used to define a classification model. According to the following rules an instance can be defined as belonging to $p_g$if:

$$b0 + b1x1 + ... + b_nx_n > 0 \quad (4)\ [2]$$

And, similarly, an instance can be defined as belonging to p_b if:

$$b0 + b1x1 + ... + b_nx_n < 0 \quad (5)\ [2]$$

If $p_g = p_b$ then an instance has equal probability of belonging to both classes. These rules are based on a probability cut-off of 0.5. Using a different cut-off value, p_c, the following rules apply:

$$b_0 + b_1x_1 + \ldots + b_nx_n > ln\left(\frac{p_c}{1-p_c}\right) \quad (6)\ [2]$$

and, similarly, an instance can be defined as belonging to p_b if:

$$b_0 + b_1x_1 + \ldots + b_nx_n < ln\left(\frac{p_c}{1-p_c}\right) \quad (7)\ [2]$$

Previously, a disadvantage of logistic regression was the computational intensity required during MLE. however, improvements in computer hardware have made this less of an issue. The most important attraction of logistic regression is that the input can be either continuous or discrete, or it can be combination of both type and also, they do not have all the time normal distributions

C. Support Vector Machine (SVM)

Support vector machine is a kind of kernel Machine and is one of the most widely used for classification tasks. The mathematical formulation can be quite aggressive and involved, but from a geometric point of view, it can be quite simple to understand, explain and visualize. When there are two linearly separable classes x's and o', in two-dimensional space, there can be many different solutions that have good accuracy or minimum errors on the training data as sown in the figure 3.1. [8] The SVM solution matches what an intelligent human agent would naturally choose as the decision boundary, i.e. the one that is the farthest away from either cluster (shown in purple). This makes the decision made bySVM more robust against errors (both random and systematic). For example, some financial ratios such as the debt-to-equity ratio rely on subjective valuations of intangible assets by accountants, which are error-prone.

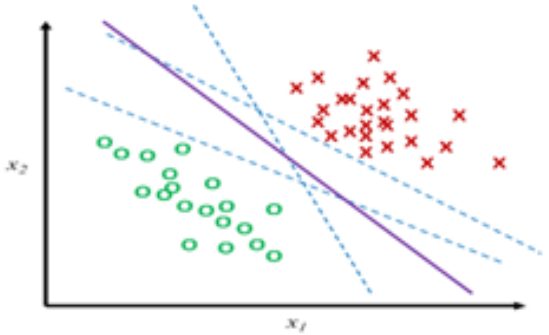

Figure 3

In the above figure (3), the binary separation of x an o is done by using SVM in 2D. The blue dashed line in the figure indicates potential decision boundaries with zero training errors that no labels are located on the wrong side of the line. And the Decision line of SVM is indicated by dark blue line. The concept of choosing the decision boundary with the largest margin is one of the distinguishing characteristics of SVM. Compared to logistic regression, which is more concerned with maximizing the probability of the two classes, SVM concentrates on maximizing the separation between the support vectors and therefore the classification accuracy (i.e. minimizing the generalization error). Since only the support vectors (a subset of training points closest to the boundary) have a significant impact on the decision boundary, the solution can be considered sparse and full knowledge of the posteriori/class probabilities is not necessary; this improves the efficiency of the algorithm. For SVM to learn from the training data, the primal problem given is often converted to its equivalent dual problem, provided in form of equation. In this case, the labels $\vec{y}$ are {-1, 1} instead of {0, 1} as in the logistic regression formulation. Although the dual problem appears to be more complicated than the primal problem, there are many advantages that reduce the complexity of this formulation. When the dimension of the input feature space (D) is much greater than the number of training points (N), the dual problem only needs to estimate N number of α instead of D number of w. Even though this is not usually the case (i.e. N is typically greater than D), only the support vectors will have a non-zero α, therefore it can still be more efficient. Furthermore, $\Phi(x_i)^T\Phi(x_j)$ can be kernelized into $K(x_i, x_j)$ and calculated directly, which bypasses the need to explicitly solve for $\Phi(x)$. Finally, because this formulation of the objective function is convex, a globally optimal solution can always be determined (i.e. the algorithm would not converge to a local optimum). Note that C in the equations is a hyperparameter that sets the degree of regularization. This allows some "slack"

in the model in cases where there are outliers and the two classes are not perfectly separable. When C is large, a hard and narrow margin is obtained between the two classes, while a small C returns a soft and wide margin.

$$\min_{w} \underset{w^T}{\rightarrow} \underset{w}{\rightarrow} + C \sum_i^N max\left(0,1 - y_i \left(\underset{w^T}{\rightarrow} \underset{\emptyset}{\rightarrow} (x_i) + b\right)\right) \quad (1)[3]$$

$$\max_{a_i \geq 0} \sum_i a_i - \frac{1}{2} \sum_{jk} a_j a_k y_j y_k \emptyset\left(x_j\right)^T \emptyset\left(x_k\right) \quad (2)\ [3]$$

$$\text{s.t} \quad 0 \leq a_i \leq C \ \forall i$$

$$\sum_i a_i y_i = 0$$

D. Linear discriminant analysis

Linear discriminant analysis is a classification method used in machine learning. It is derived from Fisher's Linear Discriminant and used to discover a linear combination of features that can be characterized into two or more objects. After the result, the combination with most desired features is used as a dimensionality reduction before later classification or as a linear classifier. If somebody wants to note down the difference between the classes along with reducing dimensions, LDA is the option available that determines the discriminant dimension in the response pattern category, based on the discriminant dimension the ratio of between-class over within-class variance of the available data is maximized. Thus, Linear Discriminant Analysis method is easy predictive model whose accuracy is as good as other methods that are complex. [4] In Credit risk analysis model, we determine the creditworthiness of the customer who wants the credit loan and this can have done by determining the probability that the customer will default in future or not. Thus, we can use Linear discriminant algorithm to predict that according to the customer's details that he/she falls under default or not-default category and accordingly grant credit loan [4].

Let generalize linear discriminant algorithm

Let $y = w1x1 + ... + w_ix_i + ... + w_nx_n$ be any linear combination of the characteristics $x = (x1,...,x_i,...,x_n)$. Adjusting the components of the weight vector $w = (w1,...,w_i,...,w_n)$, results in a projection onto one dimension represented as $y = w^Tx + w0$. Classification is achieved by placing a threshold on y, i.e. $w0$, which is the mid-point of the distance between the means. The objective is to select a projection that best separates two groups.

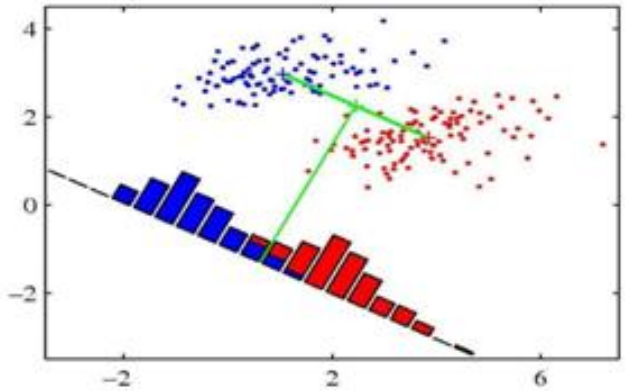

Figure.4. Two classes red and blue sample along with histogram resulting from projection onto the line based on thr fisher's linear discriminant analysis [7].

The simplest measure of separation of the classes is the separation of the class means [7]. The weights, w_n, are selected in order to maximize the distance between the means, and w is P, constrained to have unit length so that n $wn = 1$. Assuming that both the group have a common sample variance then, according to Fisher (1936) suggested a sensible measure of separation as

$$M = \frac{distance\ \ between\ \ sample\ \ means\ \ of\ \ two\ \ group}{(sample\ \ vaariance\ \ \ of\ each\ group\ )^{\frac{1}{2}}} \quad (1)\ [4]$$

where the measure M is the separating distance. This gives a large separation between the class means while also giving a small variance within each class, thereby

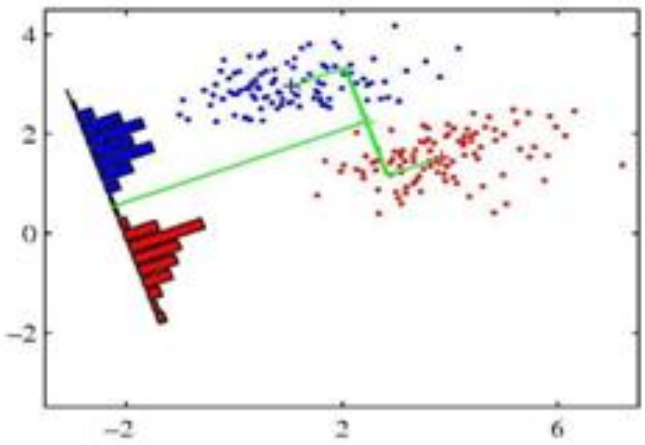

Figure .5. Two classes red and blue sample along with histogram resulting from projection onto the line based on thr fisher's linear discriminant analysis [7].

An attractive feature of LDA is the fast and simple approach to determine the optimal linear separation, merely requiring simple matrix manipulation such as addition, multiplication, and eigenvalue decomposition (Loog&Duin, 2002). LDA makes an assumption that the input features are measured on an interval scale or ratio scale. This enables the ranking of objects and a comparison of size differences between them. Unlike other forms of linear discriminant analysis, Fisher's LDA does not require that the input features are independently and randomly sampled from a population having a multivariate normal distribution. LDA assumes that the different groups have equal variance-covariance matrices[2]

III. MACHINE LEARNING TECHNIQUES

A. Artificial Neural Network

Artificial neural network is a mathematical simulation of a biological neural network. Basically, it is inspired by human brain. Individual neurons in our brain do simple decision making and yet together they control complex human function, cognitive ability etc. Likewise, the individual neurons can be mathematically represented or approximated by logical regression, and therefore artificial neural network can be thought of as multiple layers of connected logistic regression classifier. The figure below illustrates the typical structure of two-layered neural network known as multilayer perceptron [9]. An ANN with this architecture can already be considered a universal approximator, having the capability/flexibility to model any continuous functions. With modern day computers, the trend is to have more hidden units and hidden layers, giving it a deep network architecture. Although an interconnected network with many edges and nodes can be daunting at first, when broken down into its fundamental building blocks an ANN is just calculating the weighted sum of several input features. If this value is larger than a threshold, the activation function "fires" (i.e. returns a value of one instead of zero) [10].

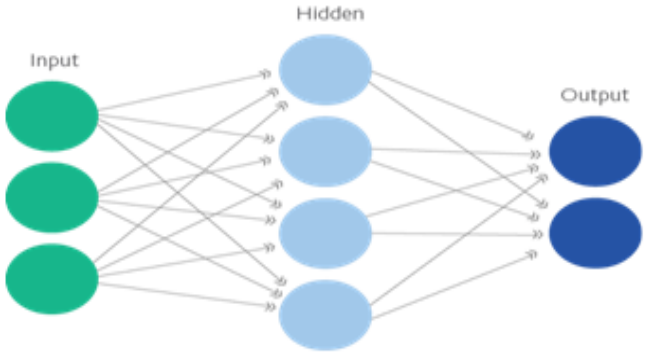

Figure. 6. Structure of a fully connected two-layer artificial neural network

ANN share similarities with a lot of the other machine learning algorithms already discussed. For example, it combines many simple decisions into a complex nonlinear decision boundary. However, unlike SVM (which is convex), ANN is sensitive to the initial parameters and can often get trapped in a local minimum. But one noteworthy benefit of ANN is that it supports sequential learning; this means the ANN can continuously update itself as new data becomes available over time without having to re-train the entire model from scratch like SVM and K-D tree. This attribute of ANN can be critical in stock market forecasting. Also, ANN can be faster than SVM because the model is generally more compact.

B. Gradient Boosting Machine – XGBoost

Gradient Boosting Machine algorithm has a high prediction power and used when there is a lot of data to deal with and make a prediction. Gradient Boosting is Machine is a kind of an ensemble learning algorithm whichbasically combines the output (prediction) of several base predictors in order to improve robustness and accuracy over a single predictor. It basically implements the method of combining multiple predictors which can be weak or average to build strong predictor [11].

XGBoost: It is an implementation of Gradient Boosting Machine algorithm. The predictive power of XGBoost is comparatively higher which makes it one of the important choice for accuracy in the event as it possesses both linear model and tree learning algorithm, which makes the algorithm almost 10 times faster than existing gradient booster techniques. One of the most interesting thing about the XGBoost is that it is called a regularized boosting technique. This helps to reduce overfitting the model. The support includes various objective function, including regression, classification and ranking. It offers tree boosting processing which works in parallel and thathelps to solve many data science problems in a fast and accurate way. It additionally provides cache access pattern, compression of data and sharing for tree boosting [12].XGBoost is machine learning technique which is used for supervised learning problems. TheXGBoost model is tree boosting ensembles machine learning technique 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

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

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

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

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

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

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

Where gi and $hi$are defined as

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

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

C. Random Forest – Decision Tree

Tree-based learning algorithms like Decision Trees are considered to be one of the best and most used in the category of supervised learning methods. Tree-based learning method helps in developing predictive model with high degree of stability, accuracy and ease of exploration. Tree-based methods map the non-linear relationships with a good accuracy. Tree-based methods break down the dataset into smaller and unit of data known as subsets of data and while this process is going on an associated decision tree is developed in an incremental pattern. We thus get an output as a tree with decision nodes and tree nodes which is known as Decision Tree. The main thing about Decision trees is it can handle both categorical and numerical data. A sample Decision Tree is shown in figure 7After the decision tree is ready for the database, a set of rules are created that defines the major objective of the project i.e. knowing if the credit risk will be good or bad [13]. In this paper, we have used Random Forest Algorithm under Decision Trees. Random Forest Algorithm is a multi-task performing algorithm which can perform both classification and regression tasks. It also capable of treating the missing values, it also undertakes dimensional reduction methods, outlier the values and other essential steps at the data exploration stage. Random Forest is an ensemble learning technique, in which group of weak models is combined to form a powerful one. To categories a new object based on an attribute, each tree is assigned a classification in the forest that is in other words, it votes for a particular tree. The algorithm i: e Random forest chooses the classification having the most votes (over all the other trees in the forest) and in case of regression, the algorithm takes an average of different outputs.

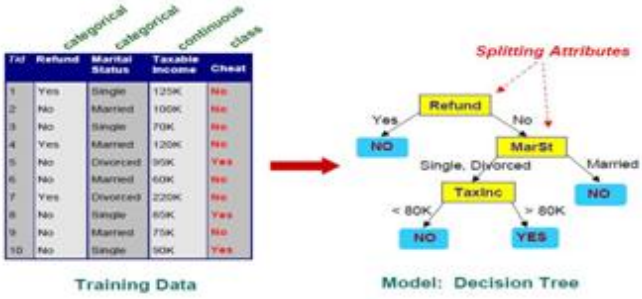

Figure 7

The algorithm is as follows:

1. Assuming that the number of cases in the training set to be N.
2. A sample from the cases are taken at random but with replacing the cases and these sample will become the training set for the growing tree.

3. Let the input variables be M and a number m<M is specified at each node.
4. m, variables are selected at random out of the node and the value of m is kept constant while the forest grows.
5. Each tree in the forest in grown to the largest possible extent and along with it, no pruning is done.
6. And in final stage the prediction of new data is done by aggregating the prediction of the n trees in the forest (i.e., majority votes for classification and the average for the regression technique)

IV. GENERALIZATION OF MODEL DEVLOPMENT

Previously for identifying the credit risk associated with the credit loan of customers many data analysis techniques were used. Lending Institution was not able to rely on this technique due to the changing technical environment and also, they cannot have an impact on the credibility of the stakeholders. Every model required to be more versatile to resist in this changing technical exploration going all around the globe. There are many techniques methods like Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning and statistics available all around but they commonly fall around two categories: Statistical Techniques and Machine Learning Techniques (Artificial Intelligence). This paper visualises the model which is divided into two parts one is used to find the patterns using clustering algorithm and then pass the data through Machine Learning techniques (ANN) network that will self-learn and will be robust enough to detect any types of risk. This model will help us not only to detect the new risk but also provide a means to analyse data. The model will use data mining techniques to mine the data and find different risk associated pattern. This data mining technique will be used to frequently mine the data set so all the patterns are always up to date. After the data mining, the patterns that are formed will be divided into two groups defaulter pattern and non-defaulter patterns. They are nothing but the two clusters. There will be a matching algorithm whose main task is to match the pattern with the real-time data that is feed to the system. The main task of this matching algorithm is to also detect the new patterns as well as to find any discrepancies. When the pattern generated from the data which is fed to the system is matched with the existing pattern and if the customer (borrower) is true then it is labelled as legitimate customer and that data will be further passed as an input to Machine learning techniques (ANN) that will analyses the customers data and thus will increase the prediction accuracy of the credit risk detection system. On the other hand, if the

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

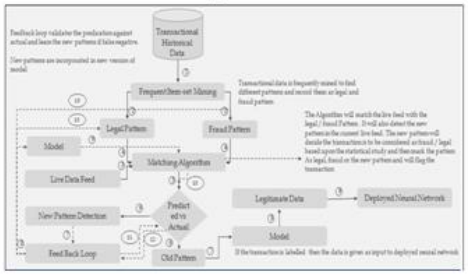

Figure. 8. Pattern matching model with clustering algorithm.

customer is detected as defaulter or fraud to pay loan then it is labelled as a defaulter and the data is then stored as defaulter pattern and the customer is denied for the credit loan. And if any new pattern is detected in the feed data then that data is stored by calling a feedback loop.

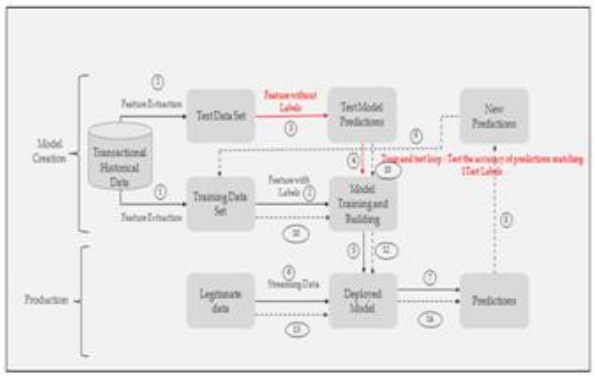

Figure.8. Prediction model using ann.

Which will add the data to the new pattern storage and the customer data stored will be mined again for data pattern and then it will be given as an input to matching algorithm and then it will check for the acceptance level and if still the pattern cannot be decided then there can be annual intervention triggered or the application can be marked as fraud and then provide it as an input to neural network that will more accurately will be able to term the data as defaulter or non-defaulter by adjusting the weights as inputs and make the system more robust The input from the pattern recognition model is given to the Machine learning techniques like Artificial Neural Network that aims to prevent risk associated and identify the default customer to get a loan. The Artificial Neural network model consists of input nodes who provide their respective output to the intermediate node and finally the output from the intermediate node is given to the output node where the actual output is compared with the expected output with certain accuracy set. While the model is tuned to get the output of exact accuracy level that is usually done using backpropagation technique by adjusting the weight at each node. The model is being built using tensor flow library.

V. DEPLOYMENT ARCHITECHTURE

When the model is built the first question arises always is how the model is going to be deployed and can be brought to use. The below-given model explains how exactly the developed model can be deployed. When the customer applies for the loan through a various medium which can be online or by submitting details through form the details of the customer is entered into the system database. All other relevant information is collected which is necessary for the credit risk evaluation. After the insertion of information of data into the database, the processing should be done according to the model discussed and the result should be generated as accepted or rejected quickly as there is multiple customer application to be processed. As the world is growing rapidly towards internet and want result quickly so the processing should have done quickly as possible and it has to be done accurately otherwise it will halter the creditability of the system and also be losing a customer. The parameters of each application are pipelined using an Apache Kafka. The Apache Kafka streamlines the process into batches So that all the transaction can process into batch instead of one after the other. The output of the Kafka is given to the actual Apache engine which uses MLlib where actually our model is deployed. MLlib consists of our two models One for pattern matching and the second model for deriving outcome as accepted or rejected. The outcome of the model needs to be visualized and analyses so that can be done using different

data visualization and Analytical tool. This provides us with the graphical representation and visualization techniques.

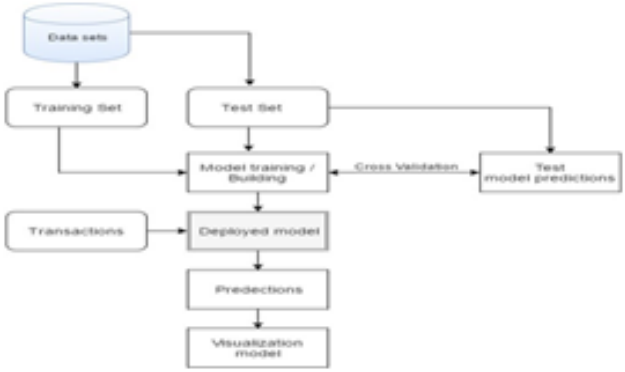

Figure .9. Deployment architecture.

VI. ACKNOWLEDGMENTS

The success and final outcome of this paper required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my research paper. All that I have done is only due to such supervision and assistance and I would not forget to thank them. I owe my deep gratitude to my external project guide Mr. Ujjwal Pathak, who took keen interest on our project work and guided us all along, till the completion of our paper work by providing all the necessary information for developing a good system. I would not forget to remember Mr. Lalit Menghani, of Home Capital for their encouragement and more over for their timely support and guidance till the completion of my paper work. I heartily thank our internal project guide, Professor Ramesh Solanki, Assistant Professor, VES Institute of Technology, Department of MCA for his guidance and suggestions during this project work.

VII. REFERENCES

[1]. Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance 34: 2767–87

[2]. https://www.investopedia.com/terms/f/five-c-credit.asp

[3]. Seetharaman, A, Vikas Kumar Sahu, A. S. Saravanan, John Rudolph Raj, and Indu Niranjan. 2017. The impact of risk management in credit rating agencies. Risks 5: 52.

[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.

[5]. Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33: 1

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

[7]. Bishop, C. (2006). Pattern Recognition and Machine Learning. New York, United States: Springer.

[8]. Trustorff, J., Konrad, P., &Leker, J. (2011). Credit risk prediction using support vector machines. Review of Quantitative Finance and Accounting, 36(4), 565-581.

[9]. Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. Science 313: 504–7.

[10]. Schmidhuber, Jurgen. 2014. Deep Learning in Neural Networks: An Overview. Technical Report IDSIA-03-14. Lugano: University of Lugano & SUPSI.

[11].Friedman, Jerome H. 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29: 1189–232.

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

[13]. Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau. 2008. Random Forests: Some Methodological Insights. Research Report RR-6729; Paris: INRIA.