* **Contents**
* Summary
* Introduction
* About data
* Objectives
* Statistical tools and software’s
* Understanding variables, it’s distribution and relation with default status
* Model Building
* Scope and Limitations
* Major findings
* References

**Summary:**

**Objective:** Now a day’s the prediction of defaulting the borrower in future is a challenging task for credit card companies. Therefore the main objective of this project is to develop prediction models for defaulting the borrower in the future by taking advantage of available technological advancement.

**Methods and material:** We used three popular data mining algorithms (artificial neural network, decision tree and Naïve Bayesian classifier) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset (150000 instances). Also we use two sampling methods to develop models (undersampling and oversampling).

**Results:** The results indicated that the Naïve Bayesian classifier is the best predictor with accuracy 92.4% and sensitivity 31.6% on the holdout dataset and with accuracy 79.1% and specificity 81.4% on the undersampling dataset and decision tree (J48) is the best predictor with accuracy 94.9% and specificity 99.7% on oversampling dataset. Performance of logistic regression for given dataset is not well.

1. **Introduction:-**

We know that now a day’s bank play a crucial role in market economy. Before giving a credit loan to borrowers, bank decides who is bad (defaulter) or good (non defaulter) borrower. The prediction of borrower status i.e. in future borrower will be defaulter or non defaulter is a challenging task for bank. The loan defaulter prediction is a binary classification problem. For market and society to function, individuals and companies need access to credit. Credit scoring algorithm, which make a guess of probability of default, are the methods bank use to determine whether or not a loan should be granted. In this project, historical data on 150000 borrowers is analyzed. I implement various classification algorithms namely artificial neural network, logistic regression, decision tree and Naïve Bayesian to predict the probability of borrower will fail in next two year. Also data is imbalanced so I use various sampling technique like undersampling and oversampling (Maciej A. Mazurowski, 2008) and compare the results with regular technique of building the model.

The problem is to classify borrower as defaulter or non defaulter. It is commonly desired for banks to classify borrower accurately so as to manage their loan risk better and increase business. However developing such a model is a very challenging due to growing demand for loans. Data size is very huge and some question arises about relevancy of data. Some banks are using data warehouse technologies to support development and update the models. Mostly banks use the logistic regression to evaluate customer risk. This project will attempt to use some traditional classification models like Artificial Neural Network, Logistic regression, Decision tree and Naïve Bayesian classifier and attempt to create superior model using Sample Explore Modify Model and Assess (SEMMA) methodology to better predict customer default.

1. **About data:-**
   1. **Data source:**

The data set used for this project is obtained from the competition titled “Give Me Some Credit” in kaggle.com. The file contains various parameters such as Monthly Income, Number of Dependents, age, number of open credit lines and loans etc. It is stored in csv format.

* 1. **Data description:**

In thisdatathere are 150000 instances and ten continuous variables and one binary variable coded as 0 or 1 and which is our response variable. Means we classify the new borrowers either a 0 or 1. The list of variable is as follows,

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Variable Name** | **Type** |
| **1** | **Serious Delinquent in 2 years** | **Binary** |
| **2** | **Revolving Utilization Of Unsecured Lines** | **percentage** |
| **3** | **Borrowers Age** | **Integer** |
| **4** | **Debit Ratio** | **percentage** |
| **5** | **Monthly Income** | **Real Number** |
| **6** | **Number Of Open Credit Lines And Loans** | **Integer** |
| **7** | **Number of Real Estate Loans Or Lines** | **Integer** |
| **8** | **Number Of Time 30-59 Days Past Due Not Worse** | **Integer** |
| **9** | **Number Of Time60-89 Days Past Due Not Worse** | **Integer** |
| **10** | **Number Of Times 90 Days Late** | **Integer** |
| **11** | **Number Of Dependents** | **Integer** |

* 1. **Terminology:-**

I describe some term related with this project like credit card, delinquency, credit score, utilization etc.

**i) Credit Card:-**

Credit card is a type of financial account. By using credit cards, customers can offers a bank’s money instead of their own to pay for a product or service today, and over time, they repay the bank. For the benefit of using someone else’s money, customers will often need to pay interest, as expected with other types of loans.

**ii) Default:-**

When payments are not made in time and according to the agreement signed by the card holder, the account is said to be in default.

**iii) Delinquency:-**

The term delinquent commonly refers to a situation where a borrower is late or overdue on a credit payment. If borrower is in default, the bank that has issued card is entitled to charge borrower a higher interest rate for that particular billing period as a penalty.

**iv) Credit Score:-**

Credit score is numerical expression based on a level of analysis of a person’s credit files. Simply it is representation of creditworthiness of a borrower.

**v) Utilization:-**

Utilization is a percentage of consumer’s available credit that borrower has used. The credit utilization ratio is key component of borrower credit card.

1. **Objectives:-**

Keeping in view the above discussion, the following objectives has been laid down;

1. To study the relationship between available variables and status of default.
2. To model the relationship between available variables and status of default.
3. **Statistical tools:-**

i) Frequency Table

ii) Logistic Regression

iii) Data mining classifiers: Artificial Neural Network, Decision Tree and Naïve Bayesian classifier

1. **Statistical and data mining software:-**

i) SAS

ii) WEKA

1. **Understanding variables and relation with default status:-**

In this section we give description of each variable. First we check relation of all variables with default status by using frequency table. The frequency distribution of each variable is as follows.

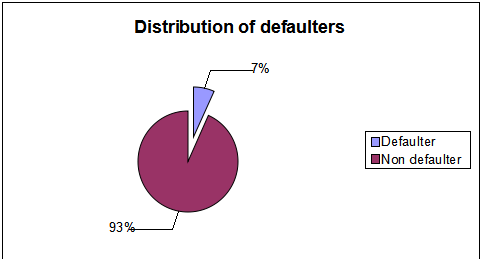
**6.1. Frequency Table:-**

**i) Serious Delinquent in 2 years:**-

**Description**:

This is of binary type. Our algorithm is used to predict this. This depicts whether a person experienced 90 days past due delinquency or worse.

The number of defaulter instances in this variable is 10026 i.e. 6.68% among the all instances.



**ii) Number of revolving utilization of unsecured lines:-**

**Description:**

Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits. This is in percentage.

This variables contain some outliers. Most of the value of this variable lies between 0 and 1 but sometimes it exceed to 1 because some borrowers use more credit than credit limit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | | Defaulter | |
| Revolving utilization | Number | % of cases | Number | % of cases |
| less than 0.25 | 87657 | 58.44 | 1873 | 2.14 |
| 0.25-0.5 | 21055 | 14.04 | 1114 | 5.29 |
| 0.5-0.75 | 13764 | 9.18 | 1394 | 10.13 |
| 0.75-1.0 | 24203 | 16.14 | 4408 | 18.21 |
| **1.0-2.0** | **2950** | **1.97** | **1183** | **40.1** |
| above 2 | 371 | 0.25 | 54 | 14.56 |

From above table it is clear that as a revolving utilization of unsecured lines increases the percentage of defaulters is increase. If revolving utilization of unsecured lines is between 1 to 2 then chances of borrowers will be default is high.

**iii) Borrowers Age:-**

**Description:**

This variablecontains the age of the borrower in years. It is of integer type. This column didn’t contain any missing data. The frequency table and percentage of defaulter of this variable is,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | | Defaulter | |
| Borrower Age | Number | % of cases | Number | % of cases |
| less than 25 | 3028 | 2.02 | 338 | 11.16 |
| 26-35 | 18458 | 12.31 | 2053 | 11.12 |
| 36-45 | 29819 | 19.88 | 2628 | 8.81 |
| 46-55 | 36690 | 24.46 | 2786 | 7.59 |
| 56-65 | 33406 | 22.27 | 1531 | 4.58 |
| above 65 | 28599 | 19.07 | 690 | 2.41 |

From above table we can say that the elder borrowers are most likely to defaulter than older borrowers. Near about 85% borrowers have age greater than 35.

**iv) Debit Ratio:-**

**Description:**

This variable contains data in percentage form. It is obtained by dividing the sum of monthly debt payments, alimony, living costs with monthly gross income.

Debit ratio is the ratio of **total debit** and **total asset**. Most of the time it is lies between 0 and 1 but sometime it is exceed to 1 because sometime total asset is much less than total debit.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | | Defaulter | |
| Debit ratio | Number | % of cases | Number | % of cases |
| less than 0.25 | 52361 | 34.91 | 3126 | 5.97 |
| 0.25-0.5 | 41347 | 27.56 | 2529 | 6.12 |
| 0.5-0.75 | 15728 | 10.49 | 1484 | 9.44 |
| 0.75-1.0 | 5427 | 3.62 | 596 | 10.98 |
| 1.0-2.0 | 4092 | 2.73 | 539 | 13.17 |
| above 2 | 31045 | 20.7 | 1752 | 5.64 |

As like revolving utilization of unsecured lines if debit ratio is increases the proportion of borrower will be defaults is increase. The percentage of defaulters is high when debit ratio is in between 1 to 2.

**v) Monthly Income:-**

**Description:**

This column contained the information about the monthly income of an individual.

The unit of income is unknown i.e. we don’t know the given income of borrowers is in rupees or in dollars. Also this column contains some missing value. We don’t know the exact value of those observations. Near about 29731(20%) observations are missing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | | Defaulter | |
| Monthly income | Number | % of cases | Number | % of cases |
| below 5000 | 55859 | 46.45 | 4813 | 8.62 |
| 5001-10000 | 46091 | 38.32 | 2752 | 5.97 |
| 10001-15000 | 13035 | 10.84 | 547 | 4.2 |
| above 15000 | 5284 | 4.39 | 245 | 4.64 |

From the above table we can say that nearly 85% borrowers have monthly income less than 10000 Unit. Proportion of defaulters are much high when income of borrower is below than 5000 Unit.

**vi) Number of open credit lines and loans:-**

**Description:**

This column contained information about the number of open loans such as car loans, house loans and lines of credit (ex. Credit card).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of open credit lines and loans | Total | | Defaulter | |
| Number | % of cases | Number | % of cases |
| below 5 | 46590 | 31.06 | 3922 | 8.42 |
| 6 to 10 | 60400 | 40.06 | 3345 | 5.54 |
| 11 to 15 | 29184 | 19.46 | 1804 | 6.18 |
| 16 to 20 | 9846 | 6.56 | 676 | 6.87 |
| 21 to 25 | 2841 | 1.89 | 191 | 6.72 |
| 26 to 30 | 785 | 0.52 | 62 | 7.9 |
| above 30 | 354 | 0.24 | 26 | 7.34 |

Above table shows that nearly 70% borrowers use less than 10 open credit lines and loans. The proportion of defaulters is overall equal.

**vii) Number of real estate and loans:-**

**Description:**

This variable contained information about number of mortgage and real estate loans including home equity lines of credit an individual have taken.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of real estate loans & lines | Total | | Defaulter | |
| Number | % of cases | Number | % of cases |
| below 5 | 149207 | 99.47 | 9884 | 6.62 |
| 6 to 10 | 699 | 0.47 | 121 | 17.31 |
| 11 to 15 | 70 | 0.05 | 16 | 22.86 |
| 16 to 20 | 14 | 0.01 | 3 | 21.43 |
| above 20 | 10 | 0.01 | 2 | 20 |

From above table we can say that almost 99% borrowers have a number of real estate loans is less than five. But proportion of defaulter borrowers is high when number of real estate loans is greater than 5.

**viii) Number of times 30-59 days past due but not worse:-**

**Description:**

This variable contains number of times borrower has been 30-59 days past due but no worse in the last 2 years.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of times 30-59 days late | Total | | Defaulter | |
| Number | % of cases | Number | % of cases |
| 0 | 126018 | 84.01 | 5041 | 4 |
| 1 | 16033 | 10.69 | 2409 | 15.03 |
| 2 | 4598 | 3.07 | 1219 | 26.51 |
| 3 | 1754 | 1.17 | 618 | 35.23 |
| 4 | 747 | 0.5 | 318 | 42.57 |
| 5 | 342 | 0.23 | 154 | 45.03 |
| 6 | 140 | 0.09 | 74 | 52.86 |
| 7 | 54 | 0.04 | 28 | 51.85 |
| above 7 | 314 | 0.21 | 165 | 52.54 |

From the above table it is clear that the percentage of defaulter borrowers is high if they are more times late to pay their credit (30-59 days late).

**ix) Number of times 60-89 days past due but not worse:-**

**Description:**

This variable contains the information about the number of times borrower has been 60-89 days past due but no worse in the last 2 years.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of times 60-89 days late | Total | | Defaulter | |
| Number | % of cases | Number | % of cases |
| 0 | 142396 | 94.93 | 7256 | 5.1 |
| 1 | 5731 | 3.82 | 1777 | 31.01 |
| 2 | 1118 | 0.75 | 561 | 50.18 |
| 3 | 318 | 0.21 | 180 | 56.6 |
| 4 | 105 | 0.07 | 65 | 61.9 |
| 5 | 34 | 0.02 | 21 | 61.76 |
| Above 6 | 298 | 0.2 | 166 | 55.7 |

As like previous table this table also shows that defaulting rate is increases as number of times delinquency of borrowers in 60-89 days increases.

**x) Number of times 90 days late:-**

**Description:**

This variable had information about the number of times an individual was late by 90 days or more in paying their bills.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of times 90 days late | Total | | Defaulter | |
| Number | % of cases | Number | % of cases |
| 0 | 141662 | 94.44 | 6554 | 4.63 |
| 1 | 5243 | 3.5 | 1765 | 33.66 |
| 2 | 1555 | 1.04 | 776 | 49.9 |
| 3 | 667 | 0.44 | 385 | 57.72 |
| 4 | 291 | 0.19 | 195 | 67.01 |
| 5 | 131 | 0.09 | 83 | 63.36 |
| 6 | 80 | 0.05 | 48 | 60 |
| above 6 | 371 | 0.25 | 220 | 59.3 |

The proportion of defaulters is high when the more times borrower has late to pay their credit bill.

**xi) Number of dependent:-**

**Description:**

This variable contained information about the number of dependents in the family excluding themselves. Near about 3924 values of this variable are missing. We replace all missing values of this column by mode of that attribute and mode is ‘0’.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total | | Defaulter | |
| Dependent | Number | % of cases | Number | % of cases |
| 0 | 90826 | 60.55 | 5095 | 5.61 |
| 1 | 26316 | 17.54 | 1935 | 7.35 |
| 2 | 19522 | 13.01 | 1584 | 8.11 |
| 3 | 9483 | 6.32 | 837 | 8.83 |
| 4 | 2862 | 1.91 | 297 | 10.38 |
| 5 and above | 991 | 0.66 | 99 | 9.98 |

From above table we can say that most of the individuals i.e. 61% are single or no anyone is dependent on his. But the distribution of defaulting borrowers is same for all values except for those individuals are single.

* 1. **Descriptive Analysis:-**

Following table shows descriptive statistic of all variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Mean | SD | Q1 | median | Q3 |
| utilization | 6.05 | 249.76 | 0.0299 | 0.1542 | 0.5590 |
| age | 52.30 | 14.77 | 41 | 52 | 63 |
| debitratio | 353.01 | 2037.82 | 0.1751 | 0.3665 | 0.8683 |
| income | 6670.22 | 14384.67 | 3400 | 5400 | 8249 |
| credit\_lines | 8.45 | 5.15 | 5 | 8 | 11 |
| real\_estate | 1.02 | 1.13 | 0 | 1 | 2 |
| del\_30\_59 | 0.42 | 4.19 | 0 | 0 | 0 |
| del\_60\_89 | 0.24 | 4.16 | 0 | 0 | 0 |
| del\_90 | 0.27 | 4.17 | 0 | 0 | 0 |
| dependent | 0.76 | 1.12 | 0 | 0 | 1 |

\* SD: Standard Deviation Q1: First quartile Q3:Third quartile

**Conclusion:-**

i) Average of revolving utilization of unsecured lines of all borrowers is **6.05** with standard deviation **249.76**. But **75%** borrowers have a revolving utilization is less than **0.56**.

ii) On a average age of borrowers is nearly equal to **52** and also **50%** borrowers have a age below **52** or above.

iii) Mean of debit ratio is **353.01** and **75%** borrowers have debit ratio less than **0.87** means very few observations have debit ratio is high and which is affected on average of debit ratio.

iv) On a average borrowers have income is **6670.22** unit.

**6.3.T-Test of each variable with response variable:-**

Now I check the relationship between defaulter variable and all other variables. This will be checking by using test of equality of means i.e. T test with unequal variance. Here we take two groups as defaulter (1) and non defaulter (0). Here we use Welch’s t-test for testing equality of two means. By using SAS command PROC TTEST following tables are obtained,

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean of each variable | | |
|  | Serious Delinquent | |  |
| Variables | 0 (mean) | 1(mean) | P-value |
| Revolving utilization of unsecured lines | 6.1689 | 4.3673 | 0.2248 |
| Borrowers Age | 52.7514 | 45.9266 | 0.0001 |
| Number of times 30-59 days past due | 0.2801 | 2.3885 | 0.0001 |
| Debit ratio | 357.2 | 295.1 | 0.0001 |
| monthly income | 6747.8 | 5630.8 | 0.0001 |
| number of open credit lines and loans | 8.4936 | 7.8823 | 0.0001 |
| number of times 90 days late | 0.1352 | 2.0914 | 0.0001 |
| number real estate loans or lines | 1.0204 | 0.9885 | 0.0001 |
| number of time 60-89 days past due not worse | 0.1267 | 1.828 | 0.0001 |
| number of dependent | 0.7434 | 0.9482 | 0.0001 |

\* 0 : Non Defaulter 1 : Defaulter

**7. Model Building:-**

First I delete those observations that contain at least one missing observation. After deleting missing observations our data contains 120269 observations and 11 variables. Among these observations defaulting customers are 8357 (i.e. 6.95%).

As we seen in above section some variables contains unlikely values like age variable has value ‘0’. So the following changes made in data,

i) ‘0’ value replaced by median of corresponding variable in age variable.

ii) ‘98’ and ‘96’ value replaced by 13 in number of delinquent in 30-59 days variable.

iii) ‘98’ and ‘96’ value replaced by 11 in number of delinquent in 60-89 days variable.

iv) ‘98’ and ‘96’ value replaced by 17 in 90 days late for paying credit bill.

Where **96 represent non response** and **98 represent other**.

After changes in data we use artificial neural network, logistic regression, decision trees and naïve Bayesian classifier algorithms. This models have been selected for this study because of their popularity in the recent literature. I first give a short description of these classification models.

**7.1.1. Artificial Neural Network:-**

Artificial neural network(ANN) are commonly known as biological inspired, sophisticated analytical technique, capable of modeling extremely complex non-linear function. We used a popular ANN architecture called multilayer perceptron (MLP) with back-propagation. The MLP is known to be a powerful function approximator for prediction and classification problems. It is the most commonly used and well studied ANN architecture. The MLP is essentially the collection of nonlinear neurons organized and connected to each other in a feedforward multilayer structure.

**7.1.2. Logistic Regression model:-**

Logistic regression is the generalization of linear regression. It is used primarily for predicting binary or multiclass dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predicting point estimate of the event itself, it builds the model to predict the odds of its occurrence. In two class problem odds greater than 50% would mean that the case is assigned to the class designed as “1” and “0” otherwise. While logistic regression is a powerful modeling tool, it assumes that the response variable is linear in the coefficients of the predictor variable.

**7.1.3. Decision Tree:-**

Decision tree are powerful classification algorithm that are becoming increasing more popular with the growth of data mining in the field of information systems. In the literature there are so many popular decision tree algorithms like ID3, C4.5, C5 etc. As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, they use mathematical algorithms like information gain, Gini index to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is conducted. The objective of the splitting pair that maximizes the homogeneity of the resulting two or more subgroup samples. The most commonly used mathematical algorithm for splitting includes Entropy based information gain used in ID3, C4.5 and C5. Gini index used in CART. In this study I choose to use J48 algorithm as our decision tree method which is a simple C4.5 decision tree for classification.

**7.1.4. Naïve Bayesian Classifier:-**

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naïve yet the algorithm tends to perform well and learn rapidly in various supervised classification problems. Naïve Bayes classifier is based on Bayes theorem and the theorem of total probability.

In this classifier we compute the conditional probability P(Cj/X) and assign X to those class Ci having large probability i.e. Xϵ Cj if P(Cj/X)>P(Ci/X) for all i≠j=1,2,….,m.

**7.2. Measures for performance evaluation:-**

In this study I used three performance measures: accuracy, sensitivity, specificity.

i) Accuracy = It is a rate of true classified instances.

ii) Sensitivity= It is a rate of true positive classified instances.

iii) Specificity=It is a ratio of true negative instances and total observed negative instances.

To develop a model we partition the data into two part training dataset and testing dataset. First I build the model on training dataset and then check the performance of model on testing dataset. The methodology and performance of every model is as follows,

**a) Artificial Neural Network:-**

I used **sigmoid activation function** and six hidden unites in hidden layers. By using WEKA software I build the model and corresponding result are as follows,

**Weight matrix:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Input layer to hidden layer | | | | | |
| variable | node 1 | node 2 | node 3 | node 4 | node 5 | node 6 |
| Threshold | -21.03 | -11.01 | -11.69 | -7.48 | -9.77 | -10.55 |
| utilization | 18.78 | 8.84 | 10.08 | 6.84 | 9.64 | 9.09 |
| Age | 0.53 | -2.15 | 1.84 | 4.49 | 0.61 | 5.26 |
| del\_30\_59 | -20.63 | -20.04 | -2.10 | -1.50 | -14.76 | 0.42 |
| debitratio | 20.36 | 11.21 | 8.99 | 6.93 | 7.83 | 10.68 |
| income | 29.45 | 16.67 | 19.14 | 12.62 | 12.43 | 16.24 |
| credit\_lines | 4.63 | 17.51 | -2.05 | -1.62 | 21.07 | -15.80 |
| del\_90 | -38.25 | -32.67 | -28.83 | -24.41 | -24.28 | -23.29 |
| real\_estate | -10.25 | 5.11 | -3.04 | -0.04 | -7.67 | -15.18 |
| del\_60\_89 | -22.72 | -12.46 | -13.49 | -11.23 | -8.35 | -10.05 |
| dependent | -2.13 | -4.53 | -2.87 | 4.95 | -10.50 | 6.98 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | hidden layer to output layer | | | | | |
| Hidden node | Intercept | node 1 | node 2 | node 3 | node 4 | node 5 |
| defaulter | 1.40 | 1.63 | 0.93 | 0.46 | 1.02 | 0.25 |
| non defaulter | -1.40 | -1.63 | -0.93 | -0.46 | -1.02 | -0.25 |

The cutoff point is 0.5 used for the above model and the confusion matrix of testing data is,

|  |  |  |  |
| --- | --- | --- | --- |
| obs.\pred. | Non def. | Defaulter | Total |
| Non def. | 33330 | 264 | 33594 |
| Defaulter | 2091 | 396 | 2487 |
| Total | 35421 | 660 | 36081 |

**b) Logistic Regression model:-**

I develop a model on training dataset by using SAS software. Stepwise variable selection procedure is used for variable selection and result is as follows,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Summary of Stepwise Selection | | | | | |
| Step | Effect | DF | ScoreChi-Square | WaldChi-Square | Pr > ChiSq |
| 1 | del\_30\_59 | 1 | 5764.7015 | 1407.508 | <.0001 |
| 2 | del\_90 | 1 | 1794.1215 | 756.6357 | <.0001 |
| 3 | age | 1 | 580.5198 | 448.2811 | <.0001 |
| 4 | del\_60\_89 | 1 | 230.6534 | 207.9419 | <.0001 |
| 5 | dependent | 1 | 22.8346 | 34.2155 | <.0001 |
| 6 | income | 1 | 15.4233 | 83.9022 | <.0001 |
| 7 | real\_estate | 1 | 44.5306 | 44.5684 | <.0001 |
| 8 | debitratio | 1 | 4.6635 | 4.9928 | 0.0308 |

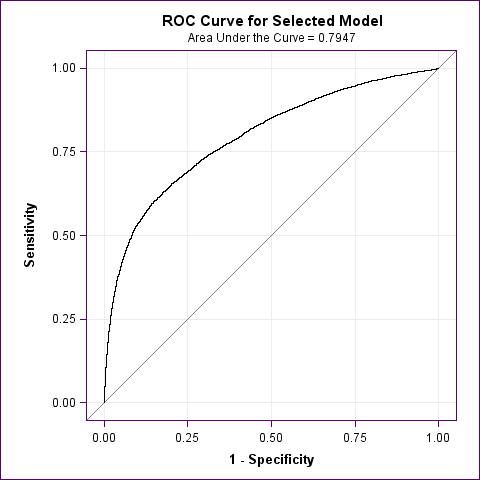
Utilization and number of credit lines variables removed in stepwise procedure.

Estimates of coefficients is,

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| coefficients | intercept | age | del\_30\_90 | debitratio | income | del\_90 | real\_estate | del\_60\_89 | dependent |
| estimate | -1.76 | -0.0241 | 0.5165 | -0.00012 | -0.00004 | 0.5488 | 0.0833 | 0.4193 | 0.0706 |

**ROC Curve of logistic model:-**

ROC curve for logistic regression is as follows,



Since area under the curve is 0.7947, hence our model is appropriate model for given data. From above ROC curve cut off point is 0.5. The confusion matrix of testing data is,

|  |  |  |  |
| --- | --- | --- | --- |
| obs.\pred. | Non def. | Defaulter | Total |
| Non def. | 33372 | 152 | 33524 |
| Defaulter | 2303 | 254 | 2557 |
| Total | 35675 | 406 | 36081 |

**c) Naïve Bayesian Classifier:-**

I use WEKA software to build Naïve Bayesian classifier. The confusion matrix by using testing data is,

|  |  |  |  |
| --- | --- | --- | --- |
| obs.\pred. | Non def. | Defaulter | Total |
| Non def. | 32579 | 1015 | 33594 |
| Defaulter | 1700 | 787 | 2487 |
| Total | 34279 | 1802 | 36081 |

**d) Decision Tree:-**

I use WEKA software to build tree. In WEKA software Gini index is used to determine splitting variable.

**Splitting variable = 90 days late for paying credit bill.**

The confusion matrix by using testing data is as follows,

|  |  |  |  |
| --- | --- | --- | --- |
| obs.\pred. | Non def. | Defaulter | Total |
| Non def. | 33284 | 310 | 33594 |
| Defaulter | 2079 | 408 | 2487 |
| Total | 35363 | 718 | 36081 |

The performance of above all classifiers is compared in following table,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance measure / Classifiers | Accuracy | Sensitivity | Specificity | Area Under Curve |
| ANN | 0.934 | 0.159 | 0.992 | 0.815 |
| Logistic regression | 0.934 | 0.095 | 0.995 | 0.794 |
| Naive Bayesian | 0.924 | **0.316** | 0.969 | 0.795 |
| Decision Tree | 0.933 | 0.164 | 0.991 | 0.685 |

From above table it is clear that accuracy of all classifier is nearly same but sensitivity of Naïve Bayesian classifier is greater than all other classifiers. Naïve Bayesian classifier achieves accuracy 92.475 with sensitivity 31.6 and specificity 96.97

To account for the class imbalance, two standard methods were evaluated, namely undersampling and oversampling (Maciej A. Mazurowski, 2008). In undersampling, instances from the overrepresented class are removed and make a data balance. This method is useful for classification but drawback is that potentially useful instances are discarded from training. In oversampling, instances from the underrepresented class are copied several times such that make a balanced dataset.

**Undersampling Method:-**

In previous section we build the model on imbalance data i.e 93/7 data. In this section I removed the non defaulter instances randomly to make dataset balanced. We know that 8357 borrowers are defaulters and remaining are non defaulters, so we take 7% sample without replacement from non defaulters and combine data. Again we partition the data into training set and testing set. Develop the model on training dataset and calculate performance measures. Repeat this procedure into 10 times and take average of performance measures. The performance of each model is shown in following table,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Performance Measure | | | |
| Statistical Models |  | Accuracy | Sensitivity | Specificity | AUC |
| ANN | Mean | 0.736 | 0.659 | 0.818 | 0.815 |
|  | SD | 0.008 | 0.083 | 0.085 | 0.005 |
| Logistic regression | Mean | 0.735 | 0.626 | 0.855 | 0.808 |
|  | SD | 0.006 | 0.016 | 0.011 | 0.007 |
| Naïve Bayesian | Mean | **0.791** | **0.814** | 0.766 | **0.86** |
|  | SD | 0.01 | 0.016 | 0.025 | 0.011 |
| Decision Tree | Mean | 0.713 | 0.551 | 0.888 | 0.796 |
|  | SD | 0.018 | 0.071 | 0.045 | 0.007 |

Hence Naïve Bayesian classifier achieves 79.1% accuracy with 81.4% sensitivity and 76.6% specificity. Also area under curve (AUC) is 0.86.

**Oversampling Results:-**

In this section I copied the defaulter instances several times such that data make balance. We know that among 120269 instances 8357 instances are defaulters and remaining are non defaulters. So I copied defaulter instances 13 times and make dataset balance. Now our data contains 220553 instances. The performance of various classifiers is given in following table,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Sensitivity | Specificity | AUC |
| ANN | 0.754 | 0.681 | 0.826 | 0.82 |
| Logistic regression | 0.738 | 0.601 | 0.871 | 0.808 |
| Naïve Bayesian | 0.712 | 0.493 | 0.922 | 0.799 |
| Decision Tree | **0.949** | **0.997** | 0.901 | **0.956** |

From above table it is clear that accuracy of decision tree is larger than all other classifiers. Accuracy achieved by decision tree is 94.96% with sensitivity 99.7% and specificity 90.1% and AUC is 0.956.

**8. Scope and Limitations:-**

**Scope:**

As studied, analysis of this project is concerned with credit risk or bank loan defaults. Also our analysis gives better results. Similarly this analysis is useful for other problems related with medical studies like predicting cancer survivability. Also this project is useful for dealing with imbalance dataset.

**Limitation**:

For this project we have data on 10 variables listed as above. Some variable like occupation, education, living status etc of borrowers are not consider in these study. We develop model only with available variables but if add other important variables then we expect that our models gives better results.

**9. Major Findings:-**

In accordance with the SEEMA methodology, after obtaining the sample data we started by understanding variables and relation with defaulting status. Also skewness of the data has significant effect on analysis. Also dealing with missing values is an important step of the data cleaning. We further modeled our data using various classification techniques artificial neural network, logistic regression, decision tree and Naïve Bayesian classifier. Then when assessing the various techniques, we used ROC curve, Accuracy, Sensitivity, Specificity etc. to compare the results. ROC values were most similar for most of the techniques but results were not great in terms of the sensitivity. This was because of the 93:7 Proportion of the events. So by changing the sampling method ( undersampling and oversampling) we get better results of about 79% sensitivity using undersampling method and 99.7% sensitivity using oversampling method.

**10. Conclusion:-**

In conclusion, we can show that it is possible to use a superior analytics algorithm through the use of different sampling method to correctly classify customers according to their probability of default. We believe that Naïve Bayesian classifier classifies customer as close to true model when data is imbalance but decision tree give better results when we make data balance by using oversampling method.

**11. Reference:-**

i) Mazurowski M.A., Habas P.A., Zurada J.M., Lo J.Y., Baker J.A., Tourassi G.D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced dataset on classification performance. Neural Network, 21, 427-436.

ii) Dursun Delen, Glenn Walker, Amit Kadam (2004). Predicting breast cancer survivability: a comparison of thee data mining methods. Artificial Intelligence in Medicine, 34, 113-127.

iii) Jiawei Han, Micheline Kamber, Jian Pei (2011). Data Mining: Concepts and Techniques, Third Edition.