PREDICTION OF APPROVAL OF

CONSUMER

PERSONAL LOANS

A PROJECT REPORT

*Submitted By*

JEEVA.K (ML01006)



SSJ IT Solutions Private Limited

Raebareli – 229001

OCT 2021

DECLARATION

I hereby declare that, The project report on “PREDICTION OF APPROVAL CONSUMER PERSONAL LOANS” has been carried out by own efforts and fact arrived at my observation under the guidance & motivation of project mentor **Mr. Kushagra Srivastava.**

|  |
| --- |
| **Signature** |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **JEEVA.K** |
|  |
|  |
|  |

Place: Getticheviyur - 638110

Date: 08-10-2021

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to **Mr. Kushagra Srivastava** for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards member of SSJ IT Solutions Private Limited for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to **Mr. Kushagra Srivastava** for giving me this great opportunity.

ABSTRACT

In our banking sedulity, banks have several merchandises to vend notwithstanding the most reservoir of pecuniary gain of any bank is on its credit line. in order that they will earn from the interest of these loans that they credit. A bank's profit or loss depends to a hulking extent on loans. daily a hulking variety of integers piece plays for loans, for a range of functions. Not with standing, these expectorants are not sure and everybody can not be approved. Every space, we have a tendency to examine a variety of cases wherever folks do not repay the adultness of the loan quantum to the bank's thanks to that they suffer vast losses. The peril associated with making a decision on loan favor is immense. that the plan of this plan is to assemble loan knowledge from multiple knowledge sources and use multiple machine learning algorithms on this knowledge to uproot vital words. This model is employed by the institutes in creating the proper call to homologate or reject the loan request of the purchasers. during this paper, we have a tendency to examine real bank credit knowledge and conduct multiplex machine learning algorithms on the word that corroborate the credit good of shoppers so as to formulate a bank peril automatic system

|  |  |  |
| --- | --- | --- |
| **S. NO.** | **CONTENTS** | **PAGE NO.** |
|  | **ABSTRACT** |  |
|  | **LIST OF ABBREVIATIONS** |  |
| **1** | **INTRODUCTION** |  |
|  | 1.1 PREFACE OF LOAN APPROVAL | **1** |
|  | 1.2 MACHINE LEARNING | **1** |
|  | 1.3 ABOUT THE PROJECT | **2** |
| **2** | **LITERATURE SURVEY** |  |
|  | 2.1 LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH : MULTIVARIATE STATISTICAL METHODS | **4** |
|  | 2.2 AN APPROACH FOR PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING ALGORITHM | **4** |
|  | 2.3 PREDICTION OF LOAN RISK USING NB AND SUPPORT VECTOR MACHINE | **4** |
|  | 2.4 A NEW CONTRAST PATTERN-BASED CLASSIFIER FOR CLASS IMBALANCE PROBLEMS | **5** |
|  | 2.5 STATISTICAL LEARNING MODEL FOR CLASSIFICATION PREDICTIONS | **6** |
|  | 2.6 CREDIT RISK ANALYSIS AND PREDICTION MODELLING OF BANK LOANS | **6** |
|  | 2.7 CUSTOMER SEGMENTATION BASED ON CLV MODEL | **6** |
|  | 2.8 LOAN DEFAULT PREDICTION WITH MACHINE LEARNING TECHNIQUES. | **7** |
|  | 2.9 DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING | **7** |
|  | 2.10 OBSERVATION AND LOAN PREDICTION PROCESS - ADABOOST | **7** |
|  | 2.11 LOAN PREDICTION BY USING MACHINE LEARNING MODELS AND 80:20 RULE | **8** |
|  | 2.12 LOAN ENABLING PREDICTION SYSTEM | **8** |
| **3** | MATERIALS AND METHODS 3.1 PROBLEM DEFINITION  3.2 PROPOSED METHODOLOGY  3.3 EXPLORATORY DATA ANALYSIS  3.4 OVERALL TASK DIAGRAM | **9** |
|  | **9** |
|  | **9**  **18** |
| **4** | **EXPERIMENTAL SETUP AND PROCEDURE**  4.1 LOAN DATASET  4.2 DETERMINE THE TRAINING AND TESTING DATA  4.3 DATA CLEANING AND PROCESSING  4.4 MODELS IMPLEMENTATIONS  4.5 APPROVAL OR REJECTION | **19**  **19**  **19**  **19**  **20** |
| **5** | **PROPOSED ALGORITHM**  5.1 PROPOSED ALGORITHM  5.2 ACTIVITY DIAGRAM  5.3 SYSTEM FEATURES  5.4 MATHEMATICAL MODEL  5.5 EXTRACTING THE IMPORTANCE FEATURES FOR PREDICTING CREDIT DEFAULTERS  5.6 THE PREDICTIVE MODEL  5.7 SYSTEM ARCHITECTURE | **21**  **22**  **23**  **23**  **23**  **24**  **25** |
| **6** | **EXPERIMENT SETUP**  6.1DEVELOPING A MACHINE LEARNING MODEL  6.2 ACTIVITY DIAGRAM  **CONCLUSION** | **26**  **27**  **30** |
|  | **REFERENCES** | **31** |

**LIST OF ABBREVIATIONS**

ABBREVIATIONS

SVM - Support Vector Machine

NB - Naive Bayes

ML - Machine Learning

NBFC - non- banking financial companies

AUC - average area under the curve

NPC - nonperforming capital

CLV - Customer Lifetime Value

DT - Decision Tree

KNN - K-Nearest Neighbors

RF - Random Forest

CUT - Circuit-under-test

EDA - exploratory data analysis

KMC - K-Means Clustering

ANOVA - ANALYSIS OF VARIANCE

BNB - Bernoullin’s Naive Bayes

DC - Data Cleaning

DP - Data Processing

DV - Data Verification

NBA - Naive Bayes Algorithm

CL - Classification

RD - Report Deliver Success

**CHAPTER 1**

**INTRODUCTION**

**1.1 PREFACE OF LOAN APPROVAL**

Loan Prediction is extremely useful for workers of banks moreover as for the mortal additionally. The aim of this method is to supply a fast, immediate and simple thanks to opt for the worthy candidates. Dream housing non depository financial institution deals all told loans. they need a presence across all urban, semi-urban, and rural areas. Customer-first applies for a loan at that time company or bank validates the customer's eligibility for a loan. the corporate or bank desires to change the loan eligibility method (real-time) supported client details provided whereas filling the applying kind. These details square measure Gender, legal status, Education, range of Dependents, Income, Loan quantity, Credit History, and others. This project has taken the information of previous customers of assorted banks to whom on a group of parameters loans were approved. that the machine learning model is trained on it record to induce correct results. the most objective of this project is to predict the protection of loans. To predict loan safety, the Gradient boosting rule and Bernolins algorithms square measure used. First |the information is clean therefore on avoid the missing values within the data set.

**1.2 MACHINE LEARNING**

**Machine learning (ML)** is that the scientific study of algorithms and maths models that PC systems use to perform a selected task whereas not pattern specific directions, making an attempt forward to patterns and logical thinking instead. it's seen as a bunch of engineering. Machine learning algorithms build a mathematical model supported sample information, named as "training data", so on turn out predictions or choices whereas not being expressly programmed to perform the task. Machine learning algorithms unit of mensuration utilized in AN passing large sort of applications, like email filtering and PC vision, wherever it's exhausting or unworkable to develop a daily rule for effectively collaborating within the task. Machine learning (ML) is that the study of PC algorithms that improve mechanically through expertise. it's seen as a bunch of engineering. Machine learning algorithms build a model supported sample information, named as "training data", so on turn out predictions or choices whereas not being expressly programmed to try to and do this.

A set of machine learning is closely associated with procedure statistics, that focuses on creating predictions pattern computers; however not all machine learning is maths learning. The study of mathematical improvement delivers ways in which, theory, and application domains to the sphere of machine learning. the process could also be a connected field of study, specializing in preliminary information analysis through unattended learning. In its application across business issues, machine learning is besides spoken of as prognostic analytics.

1. **Supervised Learning**

* Supervised learning may be a task of machine learning wherever an operation is being learned by the machine that then maps Associate in Nursing input to Associate in Nursing output supported input-output pairs. This operation is inferred from a tagged and trained information consisting of a group of coaching examples. this instance examines the coaching information Associate in Nursingd produces an inferred operate. This operate may be used for mapping new examples. Associate in the Nursing best state of affairs is needed for the algorithmic program to accurately confirm the category labels within the case of unseen instances that successively needs algorithmic program to generalize from the training information in an exceedingly cheap manner. supervised learning constructs a model which might predict supported proof and proof even within the presence of uncertainty. supervised Learning is split into 2 types:
* **Classification** - This method separates the data and it provides a fixed output which may be a ‘yes’ or a ‘no’ or binary values such as ‘0’ or ‘1’. Example: The working status of a person who can be either employed or not employed.
* **Regression -** This method fits the data and it gives continuous random values. Example: Prediction of a weather change.

1. **Classification**

Classification may be a technique wherever the response worth may be expected by separating the info into categories. this system aims to breed category assignments. so as to extract models that describe necessary knowledge categories or to predict future knowledge trends, knowledge analysis is employed as 2 forms particularly classification and prediction. Classification may be a data processing technique in machine learning wherever cluster membership for knowledge instances may be expected. Examples are:

* Recognition of a type of car in a photo
* Finding whether the mail is a mail spam or a message from a friend.
* Predicting the weather condition.

There are various classification algorithms.

The classification algorithms are:

* Linear Classifiers
* Nearest Neighbor
* Support Vector Machines
* Decision Trees
* Boosted Trees
* Random Forest

**1.3 ABOUT THE PROJECT**

Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

**CHAPTER 2**

LITERATURE REVIEW

**2.1 LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH : MULTIVARIATE STATISTICAL METHODS**

Loan Imprimatur Prognosticating supported Machine Learning Approach " Author-Kumar Arun, Garg Ishan, Kaur Sanmeet Year-2018. the foremost object of this process is to prophesy whether or not or not assignment the loan to express person square measure safe or not. This methodology is resolving into four sections

* Data Collection
* Comparison of machine learning models on collected data
* Training of system on most promising model
* Testing

**2.2** **AN APPROACH FOR PREDICTION OF LOAN APPROVAL USING MACHINE LEARNING ALGORITHM**

The authors M. A. Sheikh, A. K. Goel, and T. Kumar state that supply Regression models are performed and therefore the completely different measures of performances square measure computed. By employing a supply regression approach, the proper customers to be targeted for granting loans may be simply detected by evaluating their probability of default loans. The model concludes that a bank mustn't solely target the made clients for granting loan however it ought to assess the opposite attributes of a customer likewise that play a really vital half in credit granting selections and predicting the loan defaulters.

**2.3 PREDICTION OF LOAN RISK USING NB AND SUPPORT VECTOR MACHINE**

To Explore the Machine Learning (ML) algorithmic program for Predict the Loan sanctionative method ” Authors E. Chandra Blessie, R. Rekha-Year-2019 Extending credits to corporates and objects for the sleek functioning of growing savings like the Republic of India is bound. As a multiplying range of patrons applies for loans within the bank's and non- banking financial companies (NBFC), it's very taxing for banks and NBFCs with restricted capital to contrive a regular resolution and safe procedure to loan moneybags to their borrowers for his or her bones-and-cents demands. additionally, in recent times NBFC pools have suffered a big downfall in terms of stock worth. it's contributed to a contagion that has conjointly unfold to alternative bones-and-cents stocks, unfavorably moving the bar in recent times. during this paper, a trial is created to condense the difficulty concerned in cherry-picking the appropriate one who may repay the loan on time thereby keeping the bank’s nonperforming capital (NPC) on hold. this can be achieved by feeding the history records of the account World Health Organization noninheritable loans from the bank into a trained machine education model that may yield associate degree correct result. The supereminent focus of the paper is to see whether or not or not it'll be safe to apportion the loan to a selected person. Experimental tests begin that the Naïve Thomas Bayes model has higher performance analysis. Experimental tests begin that the Naïve Thomas Bayes model has higher performance than alternative models in terms of loan foretelling

In 2018, Shenghui principle bestowed a classification model for the MasterCard default knowledge set within the bank from Taiwan victimization 5 clump algorithms. 10-fold cross-validation was accustomed to getting average area under the curve(AUC) and therefore the correct rate of the model. light-weight GBM (high-performance Gradient Boosting framework engineered by Microsoft Company) was the best accuracy rate. The model of sunshine GBM achieved associate accuracy magnitude relation by F1-measure equal 89.34%.

**2.4 A NEW CONTRAST PATTERN-BASED CLASSIFIER FOR CLASS IMBALANCE PROBLEMS**

. Loan Forecasting using Decision Tree and Random Forest ” Author-Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar Year-2020. In India, the number of people or clubs applying for loans gets increased every epoch. The bank has to put in a lot of work to assay or forecast whether the client can pay back the loan quantity or not (defaulter or non-defaulter) in the given time. The end of this process is to find the nature or background or credibility of the customer that's applying for the loan. We use exploratory data analysis how to deal with the problem of approving or rejecting the loan request or in a short loan forecast. The main focus of this process is to determine whether the loan given to a particular person or a club shall be approved or not.

**2.5 STATISTICAL LEARNING MODEL FOR CLASSIFICATION PREDICTIONS**

The authors Vimala and Sharmili projected a loan prediction model victimization Naive Bayes(NB) and Support Vector Machines(SVM) Algorithms in 2019. Naïve Thomas Bayes, associate freelance speculation approach, encompasses applied math concerning knowledge classification. On the opposite hand, SVM uses an applied math learning model for the classification of predictions. Dataset from the UCI repository with twenty-one attributes was adopted to gauge the projected technique. Experimentations terminate that, instead of individual performances of classifiers(NB and SVM), the mixing of NB-associated SVM resulted in an economical classification of loan prediction.

**2.6 CREDIT RISK ANALYSIS AND PREDICTION MODELLING OF BANK LOANS**

In 2006, Sudhamathy advised a risk analysis technique in enabling a loan for the shopper's victimization of the R package. the varied modules embrace information choice, pre-processing, feature extraction and choice, building the model, prediction followed by the analysis. The dataset used for analysis during this technique was adopted from the UCI repository. To fine-tune the prediction accuracy, the pre-processing operation includes the subsequent sub-processes: detection, ranking, and removal of outliers, removal of imputation, and equalization of the dataset by proportional bifurcation relating to testing and coaching method. Further, the feature choice method improves prediction accuracy. once evaluated, the DT model resulted in 94.3% prediction accuracy.

**2.7 CUSTOMER SEGMENTATION BASED ON CLV MODEL**

In 2016, M. Ayoubi explained a client segmentation model that supported the ballroom dancing algorithmic rule and Kohonen neural network. client segmentation relies on effective factors on Customer Lifetime Value (CLV). The dataset of 56000 customers of the “Taavon bank” was employed in this analysis. Firstly, by exploiting the means that of a ballroom dancing approach, the optimum variety of clusters was determined. Then,” Kohonen neural network" was applied. supported the WRFM (the weight of Recency, Frequency, and Monetary) model, the worth of every cluster was calculated.

**2.8 LOAN DEFAULT PREDICTION WITH MACHINE LEARNING TECHNIQUES**

In the analyzing process named Loan Default Prediction with Machine Learning Techniques, the author L. Lai states that the Loan business is one of all the key financial gain sources for banks. However, loan default issues are a serious issue for loan businesses. With the increase of the large information era and therefore the development of machine learning techniques, these days we have a lot of choices for classifying and predicting loan default, aside from manual process. Models as well as XGBoost, random forest, k nearest neighbors, and multilayer perceptrons. Our result shows the promising application of machine learning techniques within the monetary trade.

**2.9 DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING**

In 2016, Aboobyda Jafar Hamid and Tarig Muhammad Ahmed conferred a loan peril prognosis model that supported the word mining strategies, like call Tree (J48), Naïve Thomas Bayes (NB), and Thomas Bayes web approach. The procedure followed was piloting set drop, piecing the model, Applying the model, and at last. Rating the preciseness. This approach was executed exploitation wood hen Tool and regarded a dataset with eight attributes, namely, gender, job, age, credit quantum, credit history, purpose, hull, and class. Evaluatingthesemodels on the dataset, experimental results ended that J48 predicated generally loan prognosis approach responded in improved preciseness than the contradictory strategies

**2.10 OBSERVATION AND LOAN PREDICTION PROCESS - ADABOOST**

Arun, Ishan and Sanmeet prompt a load prediction procedure victimization millilitre models. The sub-processes embrace information assortment, feature choice, training, testing, and analyzing the performance of the current model. The dataset with ten options was used for the observation and loan prediction method. varied millilitre approaches employed in the current methodology embrace lumen, DT, RF, SVM, NN, and Adaboost ways. Further, the authors prompt few important parameters that play a significant role within the loan prediction method for varied millilitre models, such that, it helps bankers in approval of loans to the purchasers supported their necessities.

**2.11 LOAN PREDICTION BY USING MACHINE LEARNING MODELS AND 80:20 RULE**

Supriya, Pavani, Saisushma, Vimala Kumari, associate degreed Vikas given associate ML-based loan prediction model. The modules at intervals the gift approach were data assortment and pre-processing, applying the metric capacity unit models, coaching job followed by testing the data. throughout the pre-processing stage, the detection and removal of outliers and imputation removal method were assigned. at intervals the gift methodology, SVM, DT, KNN, and gradient boosting models were wont to predict the probabilities of this standing regarding the authorization methodology. the normal 80:20 rule was adopted to separate the dataset into coaching job and testing processes. Experimentation over that DT has significantly higher loan prediction accuracy than the other models

**2.12 LOAN ENABLING PREDICTION SYSTEM**

In 2016, Kacheria, Shivakumar, Sawkar, and Gupta recommended a loan enabling prediction procedure that supported by the NB approach integrated with K-Nearest Neighbor (KNN) and binning algorithms. The seven parameters thought of were financial gain, age, profession, AN existing loan with its tenure, amount, and approval standing. The sub-processes embody Pre-processing (handling the missing values with KNN and knowledge refinement victimization binning algorithm), Classification victimization the NB approach, and change the dataset oftentimes leads to an acceptable improvement within the loan prediction method. Experimentation places forth the conclusion that integration of KNN and binning rule with NB resulted in improved prediction of loan enabling method.

**CHAPTER 3**

MATERIALS AND METHODS

**3.1 PROBLEM DEFINITION**

Banks, Housing Finance Companies, and a few NBFC deal in various sorts of loans like housing loans, personal loans, business loans,s, etc everywhere in the part of countries. These companies have existence in Rural, Semi-Urban, and concrete areas. After applying for a loan by a customer these companies validate the eligibility of consumers to urge the loan or not. This project provides an answer to automate this process by employing a machine learning algorithm. therefore the customer will fill a web application form. this type consists of details like Sex, legal status, Qualification, Details of Dependents, Annual Income, Amount of Loan, Credit History of the Applicant, and others.

**3.2 PROPOSED METHODOLOGY**

The major objective of this project is to derive patterns from the datasets that area unit used for the loan enabling method and make a model that supported the patterns derived within the previous step. Classification data processing algorithms area units accustomed strain the probable loan defaulters from the list. For analysis functions, essential inputs like gender, age, legal status, residential standing, job, income, loan expectation, existing consumer, account balance, total debt, etc., area unit collected and accustomed realize the acceptable attributes.

**3.3 EXPLORATORY DATA ANALYSIS**

In statistics, start word associate exploratory data analysis (EDA) is an Associate in Nursing approach to dissecting word sets to synopsize their main characteristics, generally with visual ways. A reckonings model is normally used or not, but primarily EDA is for seeing what the knowledge can tell the U.S.A. on the way aspect the formal modelling or thesis testing task.

EDA tackle specific tasks such as:

* + Spotting miscalculations and missing data;
  + Mapping out the beginning structure of the data;
  + Connecting the foremost necessary variables;
  + Listing anomalies and outliers;
  + Testing a propositions/ checking hypotheticals associated with a chose model;
  + Establishing a chintzy model (one which will be accustomed to make a case for the information with bottommost predictor variables);
  + Estimating parameters and deciding the associated confidence intervals or borderlines of error. Specific applied mathematics functions and approaches you will be capable to perform with these tools include
  + Clustering and dimension reduction approaches, that abet you to form graphical displays of high dimensional knowledge containing several variables;
  + Univariate visualisation of every field within the raw dataset, with contour statistics;
  + Bivariate visualizations and contour statistics that permit you to assess the connection between every variable within the dataset and also the target variable you ’re trying at;
  + Multivariate visualizations, for mapping and understanding commerces between fully different fields within the data;
  + K-Means clump (creating “ centers” for every cluster, supported the closest mean); prognosticative models, e.g. reversion toward the mean.

1. **UNIVARIATE ANALYSIS**

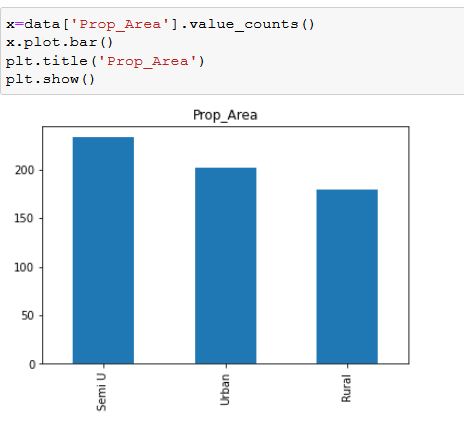
Uni suggests that one and random variable suggests that variable, therefore in univariate analysis, there's just one dependable variable. the target of the univariate analysis is to derive the info, outline and summarize it, and analyze the pattern gift in it. in an exceeding dataset, it explores every variable singly. it's potential for two varieties of variables- Categorical and Numerical.

Some patterns will be simply known with univariate analysis area unit Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and variance.

1. **FREQUENCY DISTRIBUTION TABLES**

The statistical distribution table reflects however typically a happening has taken place within the information. It offers a short plan of the information and makes it easier to seek out patterns.

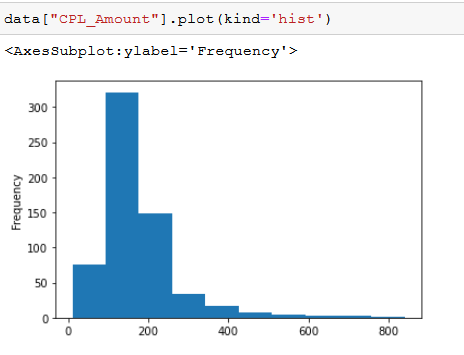
**Bar Chart**

The chart is incredibly convenient whereas examination classes of information or completely different teams of information. It helps to trace changes over time. it's best for visualizing separate knowledge

**Fig. 1. Bar Chart – Prop\_Area**

**Histogram**

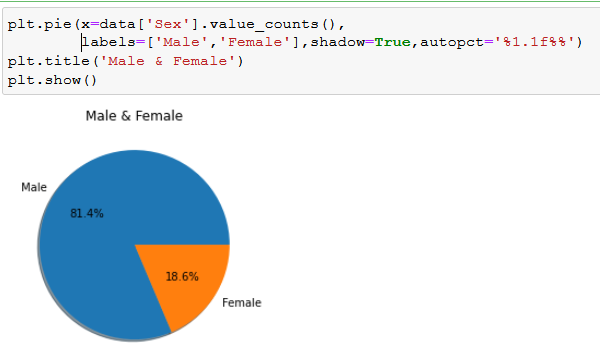
Histograms are the same as bar charts and show constant categorical variables against the class of knowledge. Histograms show these classes as bins that indicate the number of knowledge points during a vary. it's best for visualizing continuous knowledge.



**Fig. 2. Histogram – CPL\_Amount**

**Pie Charts**

Pie charts square measure in the main wont to comprehend however a bunch is softened into smaller items. the total pie represents 100%, and also the slices denote the relative size of that specific class.



**Fig. 3. Pie Chart – Sex**

**Frequency Polygons**

Similar to histograms, a frequency polygonal shape is employed for comparison datasets or displaying the additive statistical distribution.

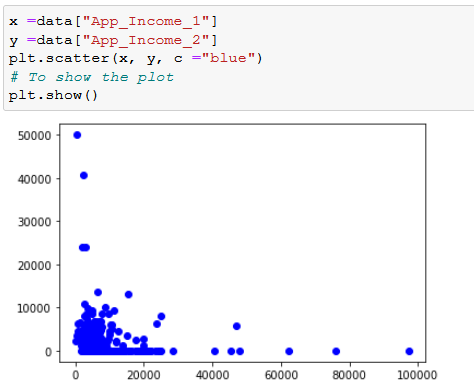
1. **BIVARIATE ANALYSIS**

Bi means that two and variable quantity means that variable, thus here there are two variables. The analysis is expounded on the cause and also the relationship between the two variables. There are three kinds of quantity analysis

1. **Bivariate Analysis of two Numerical Variables (Numerical-Numerical)**

**Scatter Plot**

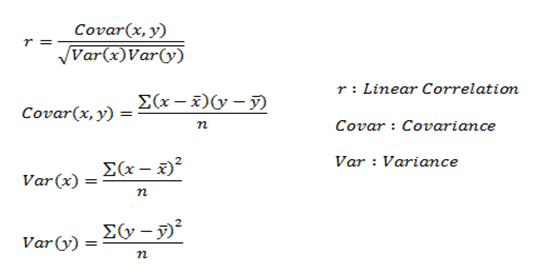
A scatter plot represents individual things of information victimization dots. These plots build it easier to determine if two variables are related to each other. the following pattern indicates the kind (linear or non-linear) and strength of the affiliation between two variables.



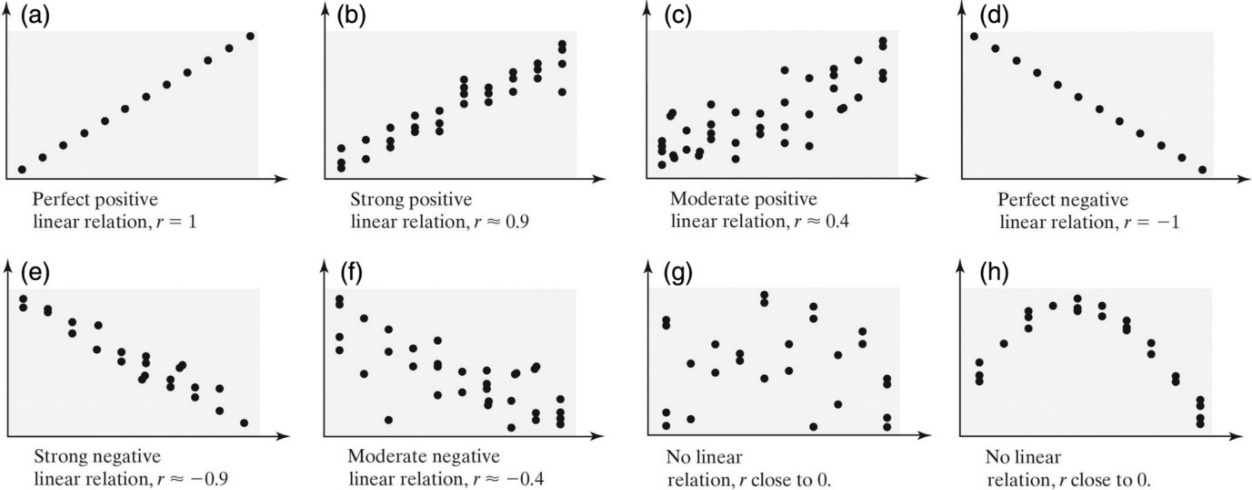
**Fig. 4. Scatter Plot – App\_Income\_1 & 2**

**Linear Correlation**

Linear Correlation represents the strength of a straightforward relationship between two numerical variables. However, there is no tendency to alter alongside the values of the equal volume, If there is no correlation between the two variables.



*Here, r measures the strength of a linear relationship and is always between -1 and 1 where -1 denotes perfect negative linear correlation and +1 denotes perfect positive linear correlation and zero denotes no linear correlation.*

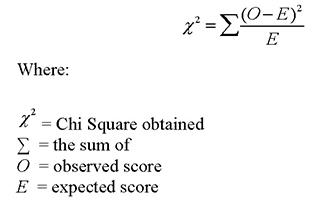


**Fig. 5. Entire Process of Linear Correlation**

1. **Bivariate Analysis of two categorical Variables (Categorical-Categorical)**

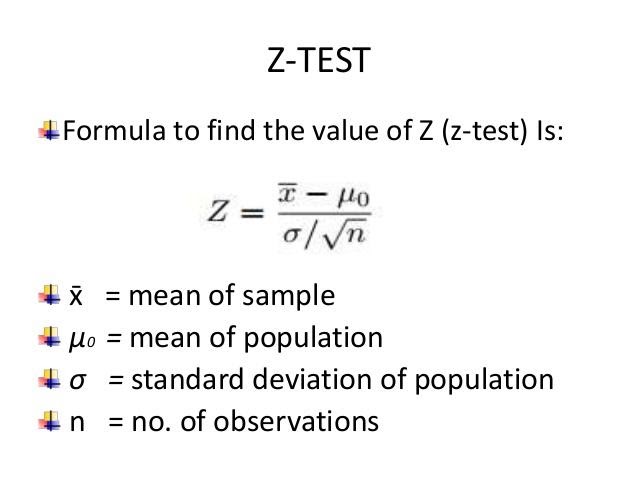
**Chi-square Test**

The chi-square check is employed to determine the association between categorical variables. Its calculation supported the distinction between expected frequencies and also the discovered frequencies in one or additional classes of the frequency table. A likelihood of zero indicates a whole dependency between two categorical variables and a likelihood of 1 indicates that 2 categorical variables are utterly freelance.

Here, subscript c indicates the degrees of freedom, O indicates discovered price, and E indicates expectation.

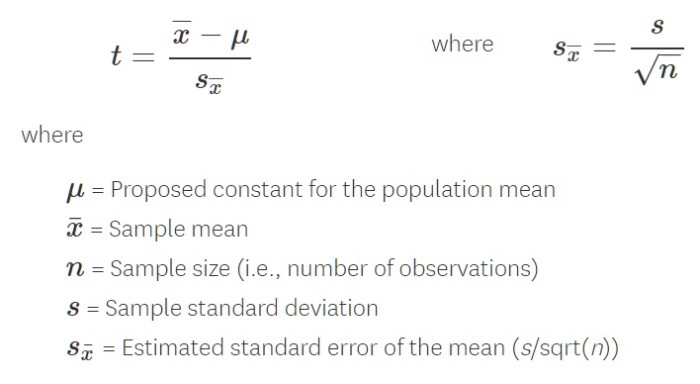
1. **Bivariate Analysis of one numerical and one categorical variable (Numerical-Categorical)**

**Z-test**

Z and T-tests are important to calculate if the difference between a sample and a population is substantial.

If the probability of Z is small, the difference between the two averages is more significant.

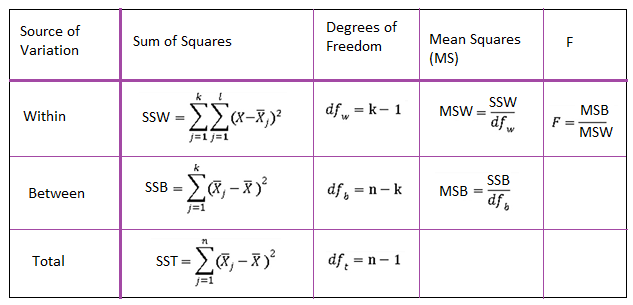
**T-Test**



If the sample size is large enough, then we use a Z-test, and for small sample size, we use a T-test.

1. **ANALYSIS OF VARIANCE (ANOVA)**

The analysis of variance take a look at is employed to see whether or not there's a big distinction among the averages of quite two teams that are statistically completely different from one another. This analysis is acceptable for comparison the averages of a numerical variable for quite two classes of a categorical variable.

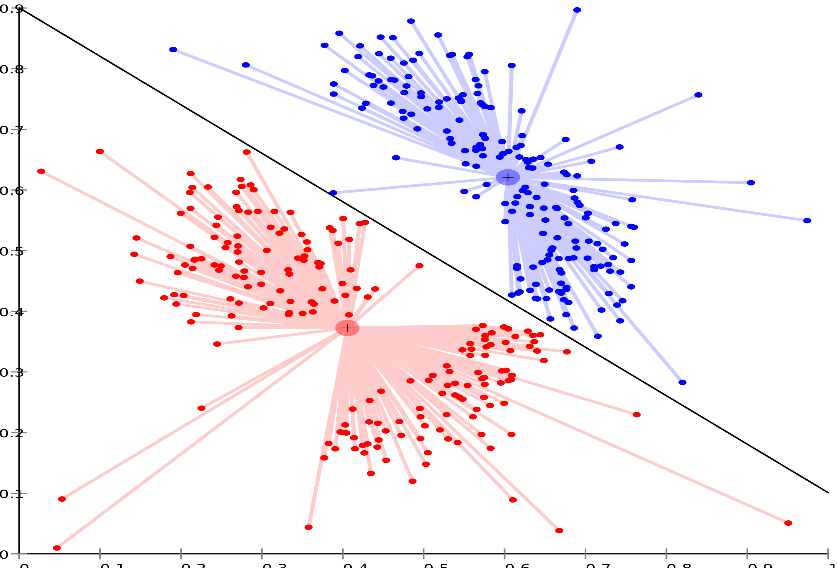


**Fig. 6. Formulation Table**

1. **MULTIVARIATE ANALYSIS**

Multivariate analysis is needed once over two variables ought to be analyzed at the same time. it's a staggeringly laborious task for the human brain to check a relationship among four variables in an exceeding graph and therefore the statistical method is employed to check additional advanced sets of information. forms of statistical method embrace Cluster Analysis, correlational analysis, multiple correlation Analysis, Principal element Analysis, etc. over twenty alternative ways to perform statistical method exist and that one to settle on depends upon the kind of information and therefore the finish goal to realize. the foremost common ways that are:

**A. Cluster Analysis**

Cluster Analysis classifies totally different objects into clusters in a very manner that the similarity between two objects from an equivalent cluster is most and negligible otherwise. it's used once rows and columns of the info table represent an equivalent unit and therefore the life represents distance or similarity.

**Fig. 7. Cluster Analysis**

**3.4 OVER ALL TASK FLOW DIAGRAM**

Input Data Output Result

**CHAPTER 4**

**IMPLEMENTATION DETAILS (MODULES):**

**4.1 LOAN DATASET**

Loan Dataset is very useful in our system for the prediction of more accurate results. Using the loan Dataset the system will automatically predict which customer’s loan it should approve and which to reject. The system will accept the loan application form as an input. A justified format of the application form should be given as an input to get processed.

**4.2 DETERMINE THE TRAINING AND TESTING DATA**

Typically, Here the system separates a dataset into a coaching set and testing set, most of the information used for coaching, and smaller parts of the knowledge area unit used for testing. once a system has been processed by mistreatment the coaching set, it makes the prediction against the check set

**4.3 DATA CLEANING AND PROCESSING:**

In Data cleaning the system detect and correct corrupt or inaccurate records from the database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative

**4.4 MODELS IMPLEMENTATIONS:**

**A. Gradient Boosting Algorithm:**

Gradient boosting is a type of machine literacy boosting. It relies on the premonition that the workaday possible succeeding model, when combined with preceding models, minimizes the overall prognostication error. Notwithstanding, either the succeeding target resultant of the case is zero If a small change in the prognostication for a case causes no change in error.

**B. Bernoulli Naive Bayes Classifier:**

Naive Bayes is a kind of classifier that uses the Bayes Theorem. It predicts count chances for each class analogous to the probability that a given record or data point belongs to a particular class. The class with the loftiest probability is considered the most likely class.

**C. Logistic Regression**

Logistic regression is basically **a supervised classification algorithm**. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X. Contrary to popular belief, logistic regression is a regression model

1. Loanapp\_ID

2. Sex

3. Marital\_Status

4. first\_name

5. last\_name

6. email

7. address

8. Dependents

9. Qual\_var

10. SE

11. App\_Income\_1

12. App\_Income\_2

13. CPL\_Amount

14. CPL\_Term

15. Credit\_His

16. Prop\_Area

17. INT\_ID

18. Prev\_ID

19. AGT\_ID

20. CPL\_Status

**4.5 APPROVAL OR REJECTION**

verify the documents and forward the details to the loan evaluator for approval or rejection. The system approves the loan if documents are cleared and reject the loan if documents are not cleared Report is delivered to the applicant according to their status.

**CHAPTER 5**

**PROPOSING METHODS**

**5.1 PROPOSED ALGORITHM**

The following shows the pseudo code for the proposed loan prediction method

1). Load the data

2). Determine the training and testing data

3). Data cleaning and pre-processing.

a) Fill the missing values with mean values regarding numerical values.

b) Fill the missing values with mode values regarding categorical variables.

c) Outlier treatment.

4). Apply the modelling for prediction

a) Removing the load identifier

b) Create the target variable (based on the requirement). In this approach, target variable is loan-status

c) Create a dummy variable for categorical variable (if required) and split the training and testing data for validation.

d) Apply the model: Gradient Boosting Algorithm, Bernoullin’s Naive Bayes, & Logistic Regression

5). Determine the accuracy followed by confusion Matrix.

**5.2 ACTIVITY DIAGRAM**

Trained Dataset

User

System

**5.3 SYSTEM FEATURES**

* + Data collection.
  + Data cleaning and preprocessing
  + Model selection
  + Data verification
  + Classification.
  + Report deliver.

**5.4 MATHEMATICAL MODEL**

Consider any decision problem, where for a given number of inputs, the decision-oriented solution is available so our project is NP-complete but some cases like not proper input format provided or if dataset not trained properly it’s NP-hard. Let s be System: S=I, P, O

S: is a System

I=I1, I2

P= DC, DP, DV, NBA,

CL O=RD

I1: Loan Dataset

I2: Trained Dataset.

Condition : Proper features trained Dataset will give proper output

Failure Condition No Trained Dataset.

**5.5 EXTRACTING THE IMPORTANCE FEATURES FOR PREDICTING CREDIT DEFAULTERS**

The total variety of options inside the bank credit Defaulters dataset. However, not all have a major influence in determining the flexibility of a given client in paying his/her loan or not. The designed system is checked with a test set and therefore the performance is assured. Evolution analysis refers to the outline and model regularities or trends for objects whose behavior changes over time. Common metrics calculated from the confusion matrix square measure Precision; Accuracy

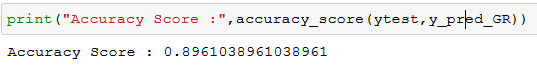
**5.6 THE PREDICTIVE MODEL**

The most important feature of these features is to develop a predictive model using an ordinary linear regression model. This can serve as a tool in determining the creditworthiness of bank clients because these are among the main features taken into consideration by most banks in advancing loans to customers.

**Accuracy:**

It measures however usually the classifier is correct for each true positives and true negative case. Mathematically, it's outlined as:

Accuracy = (True Positive + True Negative)/Total Predictions



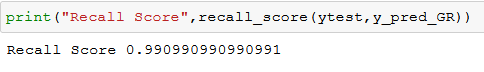
**Sensitivity or Recall:**

measures what number of times did the classifier get facticity yeas correct. Mathematically, it's outlined as

Recall = True Positive/ (True Positive + False Negative) three Specificity

It measures what number of times did the classifier get facticity negatives correct. Mathematically, it's outlined as

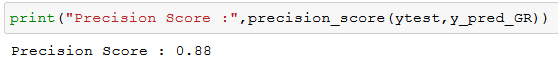
Specificity = (True Negative) / (True Negative + False Positive)



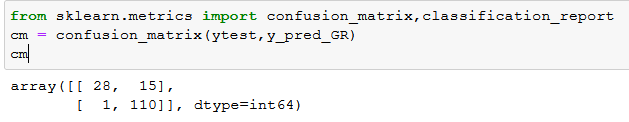
**Precision:**

Precision measures of the full expected to be positive what percentage were truly positive. Mathematically, it's outlined as:

Precision = (True Positive)/(True Positive + False Positive)



**Confusion Matrix:**



**5.7 SYSTEM ARCHITECTURE**

* Gradient Boosting Algorithm
* Bernoulli Naive Bayes Classifier
* Logistic Regression

Data

Collection

Data

Cleaning

Exploratory

Data analysis

Data Split Train & Test Data

Machine Learning Models

Training using various classifier

Predict Best Models

Input

Output Result

**Fig. 8. System Architecture**

**CHAPTER 6**

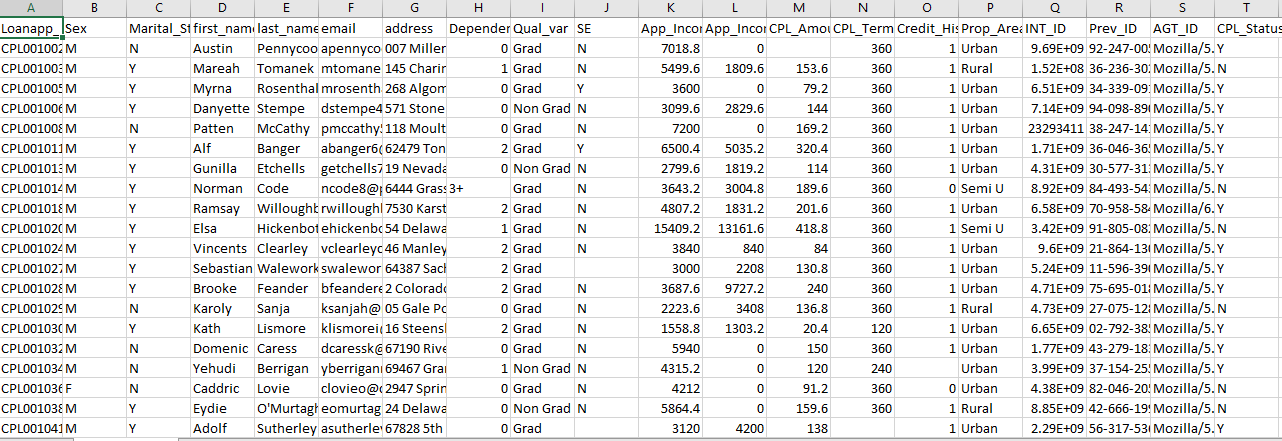
**EXPERIMENT SETUP**

**6.1 DEVELOPING A MACHINE LEARNING MODEL**

The major steps we have a tendency to used in developing the machine learning tasks/algorithms area unit are more mentioned below

**Step 1:** Collect the data:

**Step 2:** Prepare the input data: and therefore the composition of the dataset is shown in Table I.



**Table 1 – Credit Dataset**

**Step 3:** Analyze the input data: perceive the link among completely different options. A plot of the core options and therefore the entire dataset. The dataset is more split into 2/3 for coaching and 1/3 for testing the algorithms. moreover, so as to get a stratified sample, every category within the full dataset is painted in regarding the correct proportion in each the coaching and testing datasets. the varied proportions of the coaching and testing datasets employed in the paper square measure shown in Table I.

**Step 4:**Train the algorithm: the varied classification algorithms square measure trained employing a completely different set of information. The coaching dataset is shown in Table I

**Step 5:** check the formula: the varied algorithms square measure accustomed predict the effectiveness of the algorithm on the check dataset. In evaluating the performance of the classification algorithms, It includes accuracy, precision, recall, specificity, and F-measure (F1-measure). These values square measure calculated exploitation the Python scikit- learn tool with input values because the entities of the confusion matrix. The formula for the varied evaluating metrics is shown in III, with their definitions. during this paper, a positive instance refers to no(signifying there'll not be a default within the payment of the loan) whereas the negative instance refers to affirmative (signifying there'll be a default within the payment of the loan ).

**6.2** **ACTIVITY DIAGRAM:**

Loan Analysis and Data Visualization

Loan Dataset

Exploratory Data Analysis

Univariate

Bivariate

Data Cleaning

Removal unwanted

observations

Fixing structural errors

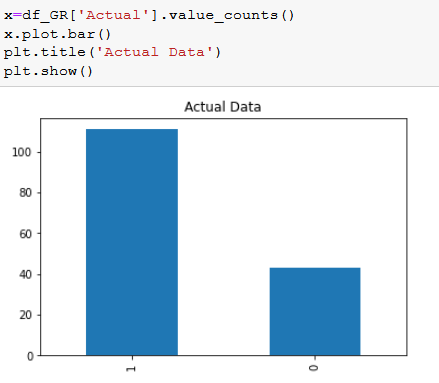
Handling missing data

Managing unwanted outliers

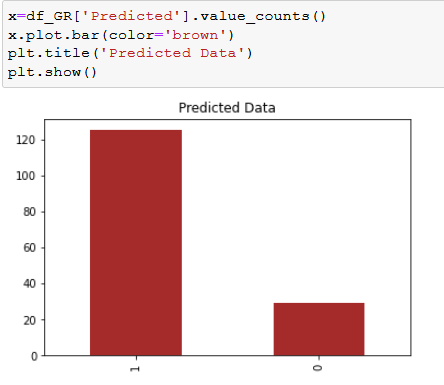
Model Building & Visualization

|  |  |  |
| --- | --- | --- |
| **Sex­­­­** | address | **CPL\_Status** |
| **M** | **xxx** | **Y** |
| **F** | **yyy** | **N** |

**OUTPUT**

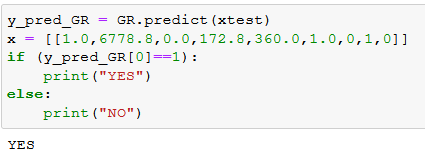


**Fig. 10. Actual data – Bar Chart**

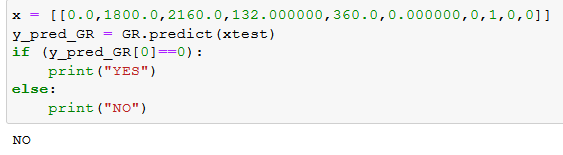


**Fig. 11. Predict data – Bar Chart**

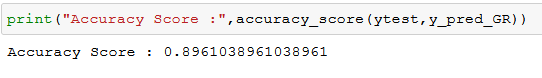
*TO CHECK THE CPL\_STATUS WHETHER YES OR NO*



*:*



***PREDICTION OF THE ACCURACY***



**CONCLUSION**

So here, it will be overconfident that the Gradient Boostting model is extraordinarily economical and offers an improved result in comparison to alternative models. It works properly and fulfills all necessities of bankers. this method properly and accurately calculates the result. It predicts the loan is approved or reject to the loan someone or client terribly accurately.

**REFERENCES**

1. Kumar Arun, Garg Ishan, Kaur Sanmeet, ―Loan Approval Prediction based on Machine Learning Approach‖, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016). B. F. Manly and J. A. N. Alberto, Multivariate statistical methods: a primer. CRC Press, 2016.
2. M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614
3. S. Vimala, K.C. Sharmili, ―Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018
4. O. Loyola-Gonzalez, M. A. Medina-P ´ erez, J. F. Mart ´ ´ınez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. Garc´ıa-Borroto, “Pbc4cip: A new contrast pattern-based classifier for class imbalance problems,” Knowledge-Based Systems, vol. 115, pp. 100–109, 2017.
5. S. Vimala, K.C. Sharmili, ―Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018
6. Sudhamathy, Credit Risk Analysis and Prediction Modelling of Bank Loans using R‖, International Journal of Engineering and Technology (IJET), Vol. 8, No. 5, pp. 1954-1966, Oct-Nov 2016.
7. M. Ayoubi, "Customer segmentation based on CLV model and neural network." International Journal of Computer Science Issues (IJCSI) 13.2 (2016): 31.
8. L. Lai, "Loan Default Prediction with Machine Learning Techniques," 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020, pp. 5-9, doi: 10.1109/CCNS50731.2020.00009.
9. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed, ―Developing Prediction Model of Loan Risk in Banks using Data Mining‖, Machine Learning andApplications: An International Journal (MLAIJ), Vol.3, No.1, pp. 1-9, March 2016.
10. Kumar Arun, Garg Ishan, Kaur Sanmeet, ―Loan Approval Prediction based on Machine Learning Approach‖, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016)
11. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma, Namburi Vimala Kumari, K. Vikas, ―Loan Prediction by using Machine Learning Models‖, International Journal of Engineering and Techniques, Vol. 5, Issue 2, pp. 144-148, Mar-Apr 2019.
12. Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta,Loan Sanctioning Prediction System, International Journal of Soft Computing and Engineering (IJSCE), vol. 6, no. 4, pp. 50-53, 2016