CUSTOMER’S ELIGIBILITY TO

GET AN APPROVAL FOR

CREDIT CARD

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 “CUSTOMER’S ELIGIBILITY TO GET AN APPROVAL FOR CREDIT CARD” 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: 6-10-2021

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 the banking sector, each banking cadre contains a colossal dataset for accounts’ MasterCard approbation which needs account identification. account identification suggests that a set of knowledge associated with what accounts would like. It depends on accounts’ beginning data feather of a field of labor, addresses corroboration, credit score, remuneration details, etc. This form primarily concentrates on calling the approbation of credit cards to accounts' victimization machine erudition. Machine Erudition is that the scientific study of algorithms and applied mathematics models that computers use to perform specific tasks with no external directions or clog. within the current trend, this form is attainable victimization several algorithms like “ Decision Tree Classifier”. This helps banks to enjoy high profitableness to satisfy their accounts. Notwithstanding, the presently prevailing system shows AN delicacy proportion of accuracy percentage of about 99%.. The planned system aims at ad-libbing the delicacy quantitative relation whereas victimization solely a number of algorithms.

|  |  |  |
| --- | --- | --- |
| **S. NO.** | **CONTENTS** | **PAGE NO.** |
|  | **ABSTRACT** |  |
|  | **LIST OF ABBREVIATIONS** |  |
| **1** | **INTRODUCTION** |  |
|  | 1.1 PREFACE OF LOAN APPROVAL | **1** |
|  | 1.2 MACHINE LEARNING | **2** |
|  | 1.3 ABOUT THE PROJECT | **4** |
| **2** | **LITERATURE SURVEY** |  |
|  | 2.1 MACHINE LEARNING APPROACH AND HORSE RACE ANALYSIS IN CREDIT CARD FRAUD | **5** |
|  | 2.2 SUPERVISED MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUDULENT TRANSACTION DETECTION | **5** |
|  | 2.3 CUSTOMER SEGMENTATION BASED ON CLV | **5** |
|  | 2.4 DETECTING FRAUD IN ADVERSARIAL ENVIRONMENTS | **6** |
|  | 2.5 CREDIT APPROVAL DATASET | **6** |
|  | 2.6 SEMI-SUPERVISED DEEP FUZZY CMEAN CLUSTERING FOR IMBALANCED MULTI-CLASS CLASSIFICATION | **7** |
|  | 2.7 DATASET FROM CLIENTS LIVE WEBSITE | **7** |
|  | 2.8 QUERY LANGUAGE FOR DATA MINING APPLICATIONS | **7** |
|  | 2.9 IMPROVED K-MEANS CLUSTERING ALGORITHM FOR PREDICTION ANALYSIS USING CLASSIFICATION TECHNIQUE IN DATA MINING | **8** |
|  | 2.10 CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES | **8** |
|  | 2.11 COMPARISON OF SEVERAL DATA MINING METHODS IN CREDIT CARD DEFAULT PREDICTION | **9** |
|  | 2.12 CLASSIFICATION USING NB APPROACH AND KNN APPROACH | **9** |
| **3** | MATERIALS AND METHODS 3.1 APPRROACHING METHOD  3.2 PROPOSED METHODOLOGY  3.3 EXPLORATORY DATA ANALYSIS  3.4 OVERALL TASK DIAGRAM | **10** |
|  | **10** |
|  | **10**  **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 OVERALL TASK 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**  **23**  **24**  **24**  **24**  **25**  **26** |
| **6** | **EXPERIMENT SETUP**  6.1DEVELOPING A MACHINE LEARNING MODEL  6.2 ACTIVITY DIAGRAM  **CONCLUSION** | **27**  **28**  **31** |
|  | **REFERENCES** | **32** |

**LIST OF ABBREVIATIONS**

ABBREVIATIONS

SVM - Support Vector Machine

NB - Naive Bayes

ML - Machine Learning

AUC - average area under the curve

DFCM - Deep Fuzzy C-Mean

UCI - University of California Irvine

WRFM - weight of Recency, Frequency, and Monetary

GBT - Gradient Boosted Trees

NPC - nonperforming capital

CLV - Customer Lifetime Value

DT - Decision Tree

KNN - K-Nearest Neighbors

RF - Random Forest

EDA - exploratory data analysis

KMC - K-Means Clustering

ANOVA - ANALYSIS OF VARIANCE

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 CREDIT CARD APPROVAL**

The correct assessment of the privy line of credit pitfall is of paramount momentousness for disposition institutions. Credit status could be a universally used style that helps moneybags establishments set the chance for a play to fail the obligation and judge whether or not to grant credit or not. The precise judgment of the trustiness of expectant permits moneybags establishments to extend the quantum of granted credit whereas minimizing attainable losses. The credit trade has exercised tremendous growth within the onetime numerous decades (Crook et al., 2007). The outsize range of possible contenders driven the event of refined recipes that modify the credit favor procedure and supervise the have the health of the Maecenas. the big volume of loan portfolios conjointly implies that modest breakthroughs in situation perfection might lead to important savings for having establishments (West, 2000). The intention of a credit situation model is to kind credit contenders into 2 kinds the “ good credit” class that is prone to repay the obligation and also the “ bad credit” class that ought to be denied credit thanks to the high probability of defaulting on the obligation. The kind is contingent on sociodemographic characteristics of the Maecenas (ditto as age, education reach, occupation, and income), the compensation Performance on former loans, and also the variety of loans. These models are applicable to bitty businesses since there are also supposed to be extensions of a private guest. with in the former numerous decades, varied quantitative ways were planned within the literature to judge guest loans and help the credit situation perfect (for a review, see. g. Crook et al., 2007). These models may be sorted into constant and non-parametric or data processing models. the foremost in style constant models are straightforward discriminant analysis and provision reversion. Linear discriminant analysis was the primary constant tactics hint for credit walk functions (Reichert et al., 1983). This approach has attracted review thanks to the specific nature of the information and also the unarguable fact that the strife matrices of the nice credit and unhealthy credit platoons are normally distinct. The provision reversion (Wiginton, 1980) permits swamping these undersupplies and has wax a common credit walk tool of interpreters in plutocrat establishments. Non-parametric tactics applied to credit scoring embody the k-nearest neighbor (Henley and Hand, 1996), call trees (Frydman et al., 1985; Davis et al.1992), artificial neural networks (Jensen, 1992), hereditary programming (Ong et al., 2005), and support vector machines.(Baesens et al., 2003). a lot of recent, analyses on cross-data processing approaches have shown promising results (Lee et al., 2002; Hsieh, 2005; Lee and boo bracket, 2002). whereas the pursuit of advanced classifiers for credit echelon employments could be a critical attempt, ameliorated accurateness may be simply achieved by aggregating scores anticipated by an associate ensemble of individual classifiers. West et al. (2005) begin that the accurateness of an associate ensemble of neural networks is superior thereto of one neural network in credit echelon and failure auguring employments. This process proposes a credit echelon model of account loans supported by varied coherent models. within the coming section, beginning analysis and information changeover are conferred. this can be followed by a contour of the information sets and a comparison of the prognosticative accurateness of the models. A discussion of the relative alms of the attributes to separate the nice credit and unhealthy credit brackets is further given.

**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**

Credit card approval is a very important process for banking organizations. The system approved or reject the credit cards. Using Machine learning we predict the credit card approval with maximum accuracy.

**CHAPTER 2**

LITERATURE REVIEW

**2.1 MACHINE LEARNING APPROACH AND HORSE RACE ANALYSIS IN CREDIT CARD FRAUD**

There have been important development within the space of credit card fraud within the past through supervised learning, unsupervised learning, and deep learning among others. Supervised learning algorithms commit to notice the perform that may effectively map the input to the corresponding output in labelled coaching knowledge. Gabriel Rushin et al discuss the Associate in a Nursing approach that compares the skills of various supervised classification models like logistical Regression, Gradient Boosted Trees (GBT), and Neural Networks in police investigation fraud

**2.2** **SUPERVISED MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUDULENT TRANSACTION DETECTION**

They as well explore the stamp generation exploitation each sphere experience Associate in Nursing and auto encoder – Associate in Nursing unattended tactics for stamp engineering. Using these two ways that during which besides the first dataset, they . Do six differentia sets on it to look at the classifiers. The results of their study show that the differentia set that was created exploitation discipline experience managed to boost the foreign terrorist college value by 1-4 whereas the autoencoder selections superimposed no Breakthrough. Amongst the supervised classifiers, Neural Networks performed the foremost effective on a adultness of the datasets, followed by GBT. nonetheless GBT showed OK prognosticating powers, Extreme Gradient Boosting Trees ( XGBoost) has shown to be feast of wise within the Mastercard

fraud discipline as seen within the work by Sahil Dhankhad et al

**2.3 CUSTOMER SEGMENTATION BASED ON CLV**

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

**2.4 DETECTING FRAUD IN ADVERSARIAL ENVIRONMENTS**

. K-means is an associate degree unattended technique accustomed to separate unlabelled knowledge into K distinct subgroups supported by a series of input options. The within-cluster variation is that the quantity that observations among a similar cluster dissent from each other according to a nominative comparison metric. A good clustering assignment is one that minimizes the add of the within-cluster variation across all clusters. Euclidean distance is employed because the metric for cluster the accounts. One challenge with K suggests that cluster is determining the best range of clusters to use, especially in the case wherever the user lacks a deep understanding of the problem’s domain or the goal is to spot new subgroups within the info. The add of the within-cluster variations at each level of clusters is aforethought. within the ideal case, the sum of within-cluster variations (in this case, the number of square distances) is drastically reduced up to some range of clusters “K”, whereon it falls insignificantly after. In that case, the quantity of clusters that's acceptable for the dataset is about to “K”

**2.5 CREDIT APPROVAL DATASET**

This section presents the result carried whereas estimating ways with the Australian credit card blessing dataset. This wide used dataset is in public acquirable at the University of California Irvine (UCI) machine erudition depot. As bestowed in Table one, contains fifteen. attributes (A01 − A15) and one family variable (A16). The differentia names and their values were changed to senseless symbols to shield the confidentiality of the information. 9 of those attributes (A01, A04, A05, A06, A07, A09, A10, A12, A13) ar categorical. Amongst the diametric six attributes, three (A11, A14, A15) contain number values and so the rest 3 (A02, A03, A08) hold real mathematics. A unnumbered missing values also are determined for these attributes. The UCI storehouse. mentions this dataset as a noteworthy one thanks to the intermixture of unceasing and nominal attributes with both infinitesimal and enormous mathematics of values. Further, the set variable contains 2 values (i) ( positive) wherever MasterCard usages are approved, and (ii) − ( negative) wherever the usages are rejected. during this dataset, MasterCard usages are approved for44.49 of cases whereas the rest55.51 of uses are rejected. In this work, we've a tendency to compare the performance of the slapdash wheel with indispensable classifiers procurable in the literature. we have trained the model and yanked the effectiveness touchstones victimization 10-foldcross-validation tactics. significantly, the depth and noise scrap of the slapdash wheel is defined as three and zero.5 during this study severally. we have allowed about a hundred trials for dimension issue

**2.6 SEMI-SUPERVISED DEEP FUZZY CMEAN CLUSTERING FOR IMBALANCED MULTI-CLASS CLASSIFICATION**

In 2019, Ali Arshad gave a multi-class classification model for eighteen datasets from the UCI repository. Semi-Supervised Deep Fuzzy C-Mean (DFCM-MC) was utilized in this process for a bunch of semi-supervised knowledge. They introduced a brand new label for the untagged knowledge by fuzzy c-mean. They used the labelled knowledge (supervised knowledge) and untagged data (unsupervised data) with the new label that extracted the discriminatory info that was used for classification. The accuracy rate of DFCM-MC was 82% and therefore the fmeasure was 78.16%.

**2.7 DATASET FROM CLIENTS LIVE WEBSITE**

User behavior is assessed during this scope by computing the chance of the present interval from the previous interval log knowledge. Activities of the user within the money application will be collected from the great user log of the code that has been introduced behind the money application to record completely different forms of user activities throughout his/her navigation over the money application. For this simulation live activities log from a shopper web-based application

**2.8 QUERY LANGUAGE FOR DATA MINING APPLICATIONS**

Simply, data processing is associated automatic discovery process of helpful data from large resources. Imielinski and Virmani (1999) introduced data processing as finding out patterns in massive databases. Even it's potential to call data processing because of the exploration science from knowledge. Hen and restrain (2006) considered data processing as a step to the data discovery process. data discovery may be a rotating method and its steps are provided as follows:

Data Cleaning

Data Transformation

Knowledge Representation

Data Mining

Data Integration

Data Selection

**Fig. 1. Data Mining**

**2.9 IMPROVED K-MEANS CLUSTERING ALGORITHM FOR PREDICTION ANALYSIS USING CLASSIFICATION TECHNIQUE IN DATA MINING**

In 2017, Arpit Bansal conferred a modification in an exceedingly clustering model of the k-means rule. This modification relies on standardization. The man of science to search out the results used the Cancer Dataset. the initial knowledge was extremely dimensional, however solely 5 attributes had been finally thought of supported needs. This paper showed that the accuracy rate for the present rule is capable 57.14% whereas the improved rule recorded 92.86%.

**2.10 CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES**

In the analysis conducted by Vaishali, K-means cluster is applied to willy-nilly generate knowledge to separate it into teams supported however possible a fraudulent dealing is to occur. To account for the non-numeric attributes, One-Hot-Encoding (among different methods) should be applied. whereas K-means and K-modes clustering solely works with numeric and categorical attributes severally, K-prototypes, an alternate, could be a hybrid technique that applies characteristics of each to the information. This side makes K-prototypes helpful since mastercard knowledge tends to own a mixture of numeric and categorical attributes. The rule, as represented by Zhexue Huang works by distribution every datum to a cluster with the image closest thereto as shown by the similarity live. The method dynamically updates the K-prototypes, in every iteration, to maximize the similarity of the information points at intervals a cluster and the difference of the information points in numerous clusters.

**2.11 COMPARISON OF SEVERAL DATA MINING METHODS IN CREDIT CARD DEFAULT PREDICTION**

In 2018, Shenghui Yang presented a family model for the credit card failure data set in the bank from Taiwan using five clustering algorithms. 10-foldcross-validation was used to get the average area under the arch (AUC) and the correct rate of the model. Light GBM (high- performance Inclination Boosting configuration erected by Microsoft Company) was the top exactitude rate. The model of Light GBM achieved an exactitude proportion by F1- measure equal 89.34.

**2.12 CLASSIFICATION USING NB APPROACH AND KNN APPROACH**

Once acquainted with the connected concepts; represent a client long-term worth. the most tool for client relationship management, to answer elementary questions on attracting, retaining and promoting clients are often customer Lifetime value (CLV) (Kumar & Lemon & Parasuraman, 2006). once it involves talking concerning the client long-run value means that a long-run method, with profit for the company or organization. during this method organizations attract customers and take a look at to retain customers within the following years thus that corporations are attempting to spot their customers and maintain their most useful customers and also the potential profitable one to spice up their profit and to eliminate weaker shoppers to scale back the company's prices thus increase profitability. the worth as a result of customer's long-run relationships ensuing from the attracting method, maintaining and up the client referred to as client long-run value (khajvand, 2010)

**CHAPTER 3**

MATERIALS AND METHODS

**3.1 APPRROACHING METHOD**

Conventional supervised approaches to mastercard fraud build classifiers on group action information across all accounts at intervals a bank’s client dataset. options square measure built at the transaction level regardless of the origin account, and used to predict fraud. This analysis study utilizes a a lot of advanced approach, during which options square measure initial built victimization account level information instead of group action level information. These accounts square measure then clustered into distinct teams supported behavioral patterns, that serves useful in distinguishing unique client teams that were antecedently unknown or not captured. Classifiers square measure then engineered on every cluster of accounts using their individual sets of transactions to coach the classifier

**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, 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.

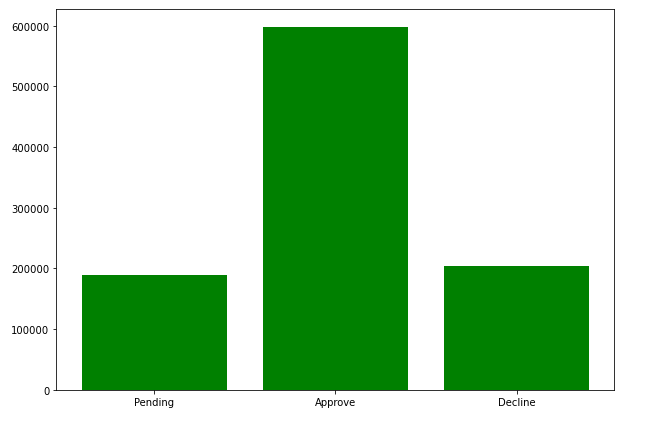
Some patterns will be simply known with univariate analysis area unit Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and variance.

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.

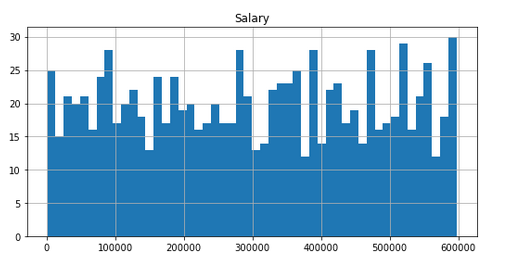
**Bar Chart**

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. 2. Bar Chart – Prop\_Area**

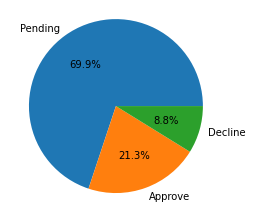
**Histogram**

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. 3. Histogram – Salary**

**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. 4. Pie Chart – tdecision**

**Frequency Polygons**

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

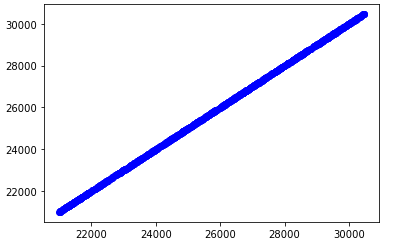
1. **BIVARIATE ANALYSIS**

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**

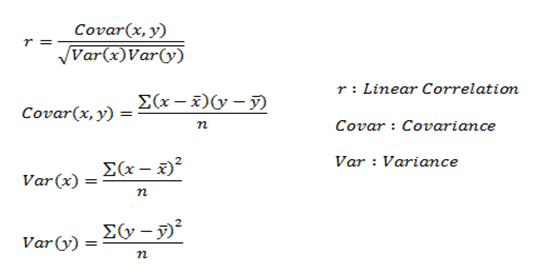
A scatter plot represents individual things of information victimization dots. These plots build it easier to determine if two variables are related to each other. the following pattern indicates the kind (linear or non-linear) and strength of the affiliation between two variables.



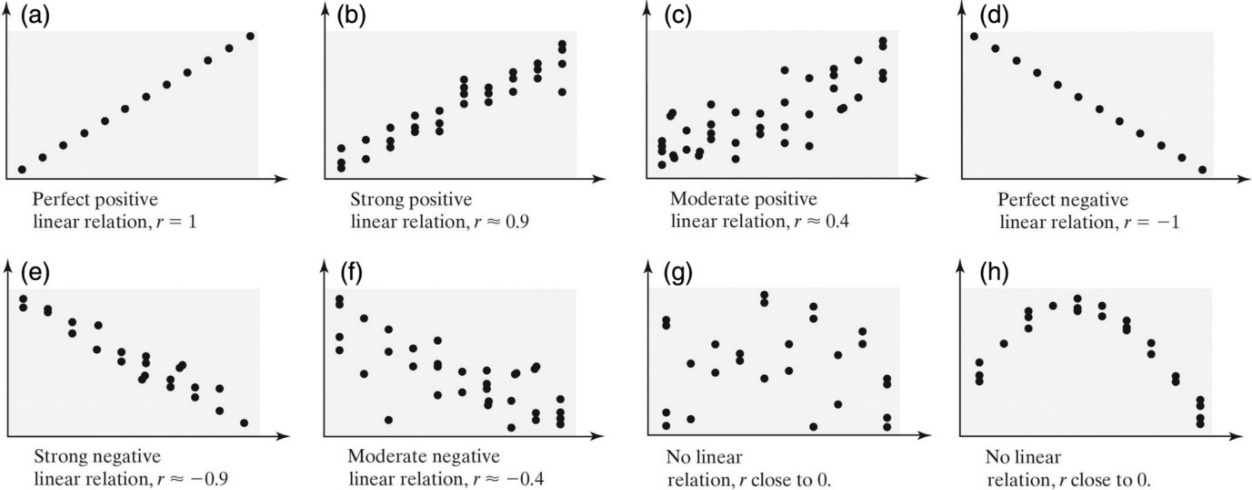
**Fig. 5. Scatter Plot – ExDebt (Liability)**

**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.*

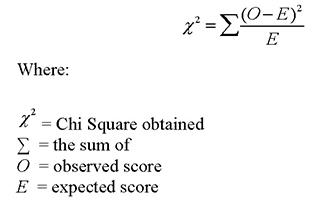


**Fig. 6. Entire Process of Linear Correlation**

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

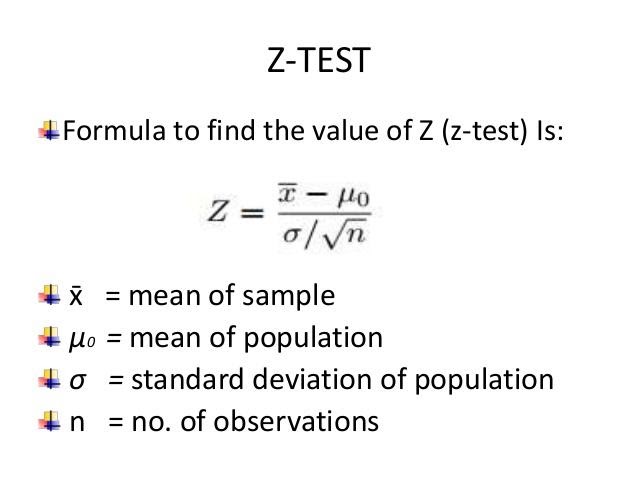
**Chi-square Test**

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.

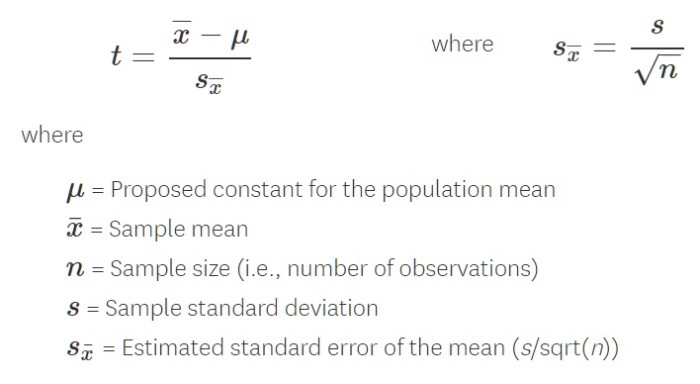
1. **Bivariate Analysis of one numerical and one categorical variable (Numerical-Categorical)**

**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.

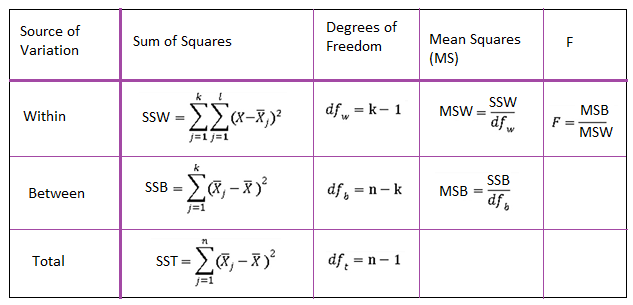
**T-Test**



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.



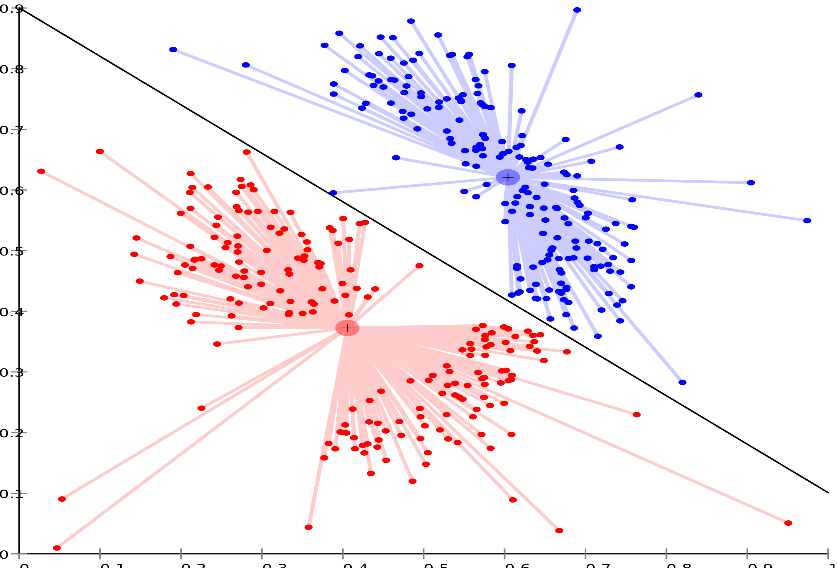
**Fig. 7. Formulation Table**

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:

1. **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. 8. Cluster Analysis**

**3.4 OVER ALL TASK FLOW DIAGRAM**

Input Data Output Result

**CHAPTER 4**

**IMPLEMENTATION DETAILS (MODULES):**

**4.1 CREDIT DATASET**

Credit Dataset is very useful in our system for the prediction of more accurate results. Using the credit Dataset the system will automatically predict which customer’s behaviour it should approve and which to reject. The system will accept the credit score application form as an input. A justified format of the credit score 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. Application id

2. first\_name

3. last\_name

4. email

5. gender

6. address

7. age

8. tdecision

9. empstaus

10. ExCus (Customer in Past)

11. Source

12. Salary

13. ExDebt (Liability)

14. Booking

15. INT\_ID

16. Prev\_ID

17. AGT\_ID

18. Booking\_Amt

**4.5 APPROVAL OR REJECTION**

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

**CHAPTER 5**

**PROPOSING METHODS**

**5.1 PROPOSED ALGORITHM**

The planned work aims at rising this technology victimization rules just like the call tree rule and also the Decision tree algorithm. The planned system is concerning the upkeep and analysis of client profiles to approve credit cards. This method happens in an automatic method. An automated method means that it involves the help of machine learning. By manually collecting massive sets of information from numerous customers World Health Organization area unit operating in numerous fields and already having credit cards, this information area unit is taken into consideration and trained and tested within the machine. The trained and tested information area unit at the start separated every which way through algorithms. it's then enforced employing a python programing language with the assistance of the boa tool. The text editor is employed in Jupyter notebook. on balance the coaching and testings of information, the machine can predict no matter information it is being given at that point, supported customers’ credit score, it'll predict whether or not credit cards are often approved or not. These predictions happen through two algorithms namely:

* + 1. Decision Tree algorithm
    2. Bagging Classifier

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: Decision Tree Algorithm, Bagging Classifier

5). Determine the accuracy followed by confusion Matrix.

**5.2 OVERALL TASK 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

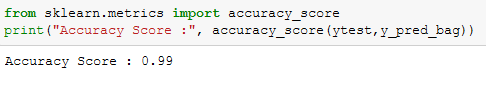
**5.6 THE PREDICTIVE MODEL**

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.

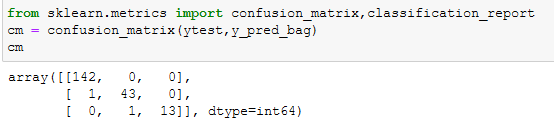
**Accuracy:**

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



**Confusion Matrix:**



**5.7 SYSTEM ARCHITECTURE**

* Decision Tree
* Bagging 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

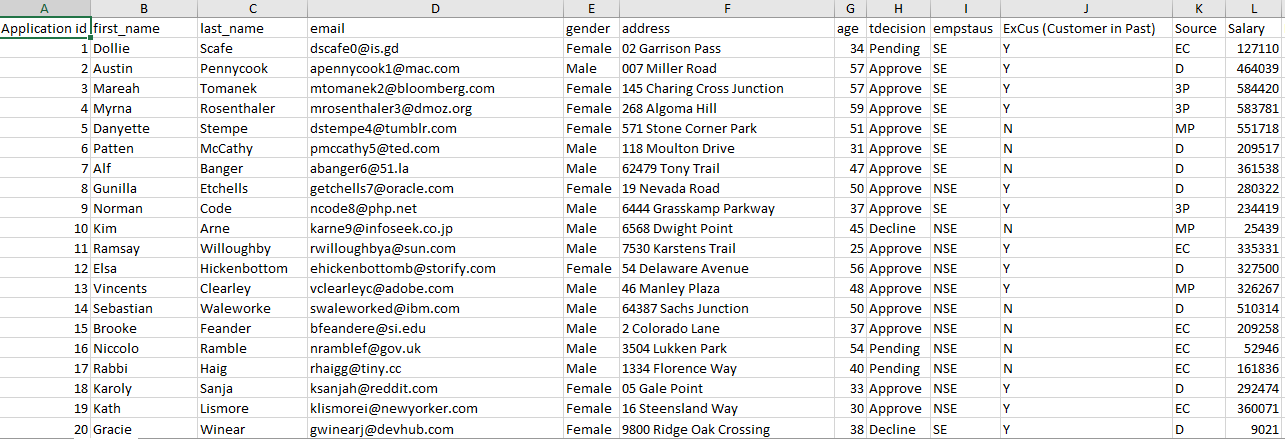
**Fig. 9. System Architecture**

**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

**Table 1. Credit Dataset**

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

**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, with their definitions.

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

Loan Analysis and Data Visualization

Loan Dataset

Exploratory Data Analysis

Univariate

Bivariate

Data Cleaning

Removal unwanted

observations

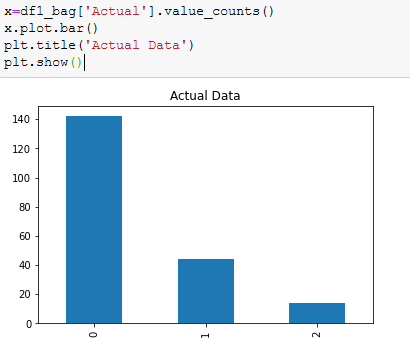
Fixing structural errors

Handling missing data

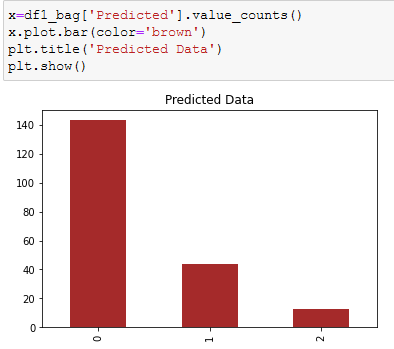
Managing unwanted outliers

Model Building & Visualization

|  |  |  |
| --- | --- | --- |
| **Sex­­­­** | address | **tdecision** |
| **M** | **xxx** | **Approved** |
| **M** | **yyy** | **Rejected** |

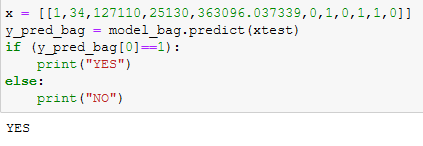
**OUTPUT:** 

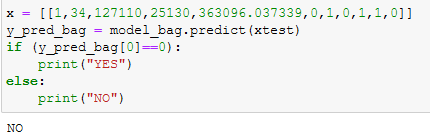
**Fig. 9. Actual data – Bar Chart**



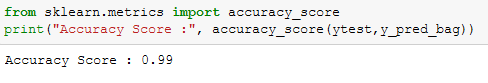
**Fig. 10. Predict data – Bar Chart**

*TO CHECK THE TDECISION WHETHER YES OR NO*

 *:*



**PREDICTION OF THE ACCURACY**



**CONCLUSION**

The datasets of consumers are utterly collected, analyzed and trained. These trained datasets function a serving to consider predicting the approval of credit cards for purchasers. each Decision Tree and Bagging Classifier algorithms would offer good results when nonstop coaching of various sets of collected knowledge. In the period of time once a lot of datasets is trained and tested and when the variables for the ultimate call are inflated, both the decision and Bagging algorithms can show a modification in their coaching and testing accuracies.The accuracy proportion improved compared to the existing system.

**REFERENCES**

1. Rushin, Gabriel and Stancil, Cody et al. April 2017, “Horse Race Analysis in Credit Card Fraud—Deep Learning, Logistic Regression, and Gradient Boosted Tree Gabriel.” 2017 Systems and Information Engineering Design Symposium (SIEDS)
2. Dhankhad, Sahil and Mohammed, Emad et al. 2018, “Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study”, IEEE International Conference on Information Reuse and Integration
3. M. Ayoubi, "Customer segmentation based on CLV model and neural network." International Journal of Computer Science Issues (IJCSI) 13.2 (2016): 31
4. Mead, Adrian and Lewris, Tyler et al. 2018, “Detecting Fraud in Adversarial Environments: A Reinforcement Learning Approach.” 2018 Systems and Information Engineering Design Symposium (SIEDS)
5. Credit Approval Dataset https://archive.ics.uci.edu/ml/datasets/Credit+Approval (2021).
6. A. Arshad, S. Riaz, and L. Jiao. "Semi-Supervised Deep Fuzzy CMean Clustering for Imbalanced Multi-class Classification." IEEE Access (2019)
7. Dataset from clients live website Available: <http://reporting4results.com/actiontracker/login.php>
8. Imielinski, T., and Virmani, A. (1999). Msql – Query Language for Data Mining Applications. Datamining and Knowledge Discovery Journal
9. A. Bansal, M. Sharma, and S. Goel. "Improved k-means clustering algorithm for prediction analysis using classification technique in data mining." International Journal of Computer Applications 157.6 (2017): 0975-8887.
10. Huang, Zhexue, 1997, “Clustering large data sets with mixed numeric and categorical values.” Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference, pp. 21-34
11. S. Yang and H. Zhang. "Comparison of Several Data Mining Methods in Credit Card Default Prediction." Intelligent Information Management 10.05 (2018): 115. 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
12. Khajvand,Mahbubeh, (2010), Customer Segmentation Based On CLV Model , M.B Thesis, Khaje Nasir Uni (In Persian)