IMPLEMENTATION OF SUMMARY STATISTICS AND DATA PREPROCESSING USING PYTHON DATA SCIENCE PACKAGES

[1]Mr.S.S.Aravinth, [2]Mr.V.Rajagopal, [3]Mr.S.Ramesh, [4]Mr.K.Santhosh,
[5]Ms.M.Nithiyarani, [6]Ms.P.Pavithra,
[1]AP/CSE, Dhirajlal Gandhi College of Technology, Salem, Tamilnadu
[2]I-M.E-CSE, Dhirajlal Gandhi College of Technology, Salem, Tamilnadu
[3,4,5,6]IV-CSE, Dhirajlal Gandhi College of Technology, Salem,Tamilnadu
Mail:aravinth.cse@dgct.ac.in[1]

ABSTRACT

Data analytics is the major technique to draw the insights from data and information which are extracted from anywhere. While preparing the analysis based reports, its very essential to produce the accurate results. Statistical methods are used to work on different types of dataset where the analysis is performed. In this paper, few of summary statistics, descriptive statistics are applied to clean the untidy data. Initially, the load prediction dataset is extracted from various providers to demonstrate the working nature of statistical methods.The process of data exploration is the preparatory stage to do the distribution analysis. Then distribution analysis is carried out to uncover the categorical value analysis. It is a chain of process to identify the deviation on dataset with variables and relations. Based on the credit card history, the loan approval process is carried after removing the missing and erroneous data.

INTRODUCTION:

To track the loan eligibility, computer data analyst has started to help the bank officers to speed up the process of loan approval. Here we take an approach between computer science and loan justice to develop a data analysis paradigm that can help solve approval of loan faster. This proposed work will use clustering based models to help in identification of loan approval patterns. This idea discusses some terminology that is used in loan justice and loan approval departments and compare and contrast them relative to data analysis systems. Automate loan eligibility referred from the person that is believed to have committed the past loan. Customer segmentation is carried out to find the original customers such as retained customers, old customers,and satisfied customers and so on. This process is reducing the complications in finding the real customers who increase the profit of the loan sanction banking bodies.

I.DATA PREPARATION:

The library pandas are used here to load the dataset. And alsopandas are used to explore the data both with descriptive statistics and data visualization.Before the analysis, the simple data analysis such as conversion of data into a corrected data type, recoding the variable into a readable format and selection of relevant variablesare carried out.

II.DATA PREPROCESSING:

The data set I contains several diagrams with plenty of information about the accounts of the bank customers such as loans,loan amount and credit cards. Here, my main purpose is to predict loan status about loan for each account. Thus, the most important diagram here is “loan prediction”. And I also need to use Applicant id and applicant Income to combine them together. Finally, the diagram required is highlighted in the following figure.

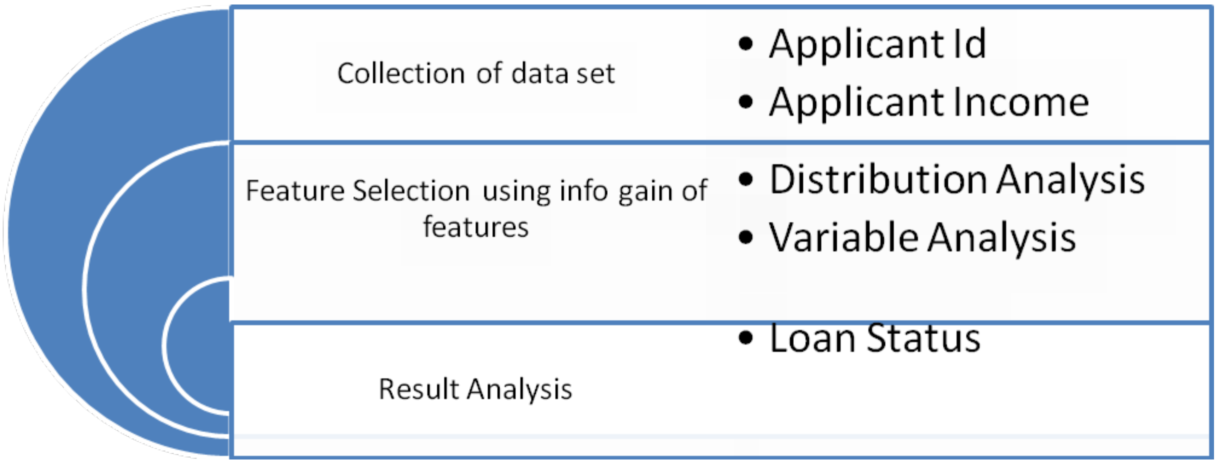


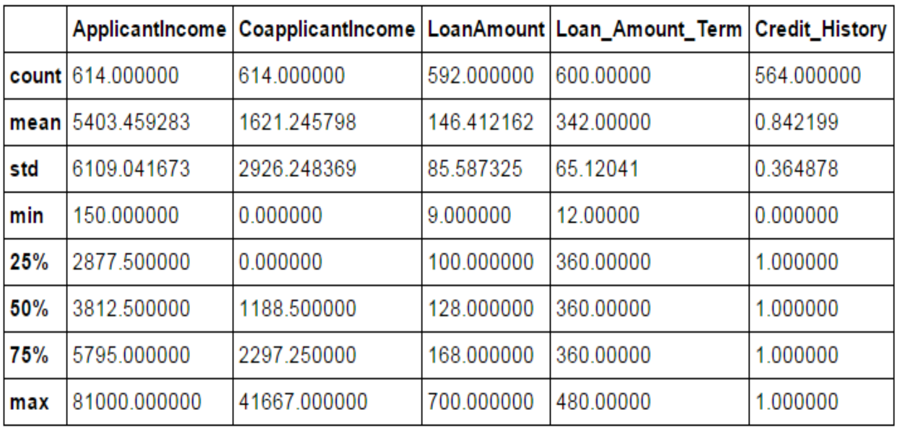
Fig 1 Workflow of loan approval process

II.i) DETECTION OF MISSING VALUES AND REPLACEMENT

In this dataset, describe function is used to understand the summary and description of data. The missed values or error parameters are such as loan amount,loan amount term, credit history, applicant income and co applicant income by finding the sum , average, mean , mode and median values. The frequency distribution value with respective analysis is carried out to provide the insights. Identified missing values are replaced with actual original values. The box blot is used to categorize the education levels of the loan seekers such as graduated and non-graduated.From this the crux is that, more number of non-graduate people is applying for loan compare to the graduates candidates.

III.SUMMARISING DATA

This session summarizes the loan prediction process, which is the most significant feature for predicting the loan approval. The count,mean, median, min, max, 25% 50% ,75% for applicant income,coapplicantincome,loanamount,loan _amount_term,credit_history are as follows:



	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

Table 1.Summaring Variables

IV.DISTRIBUTION ANALYSIS

This section examines the distribution of various numerical variables such as Applicantid and ApplicantIncome. The histograms are plotted based on the those variables as per the Figure . According to this analysis, the applicant average income is very minium to the amount of 10000. If the salary of the emplpyess increase gradually, the loan willingness is not expressed.

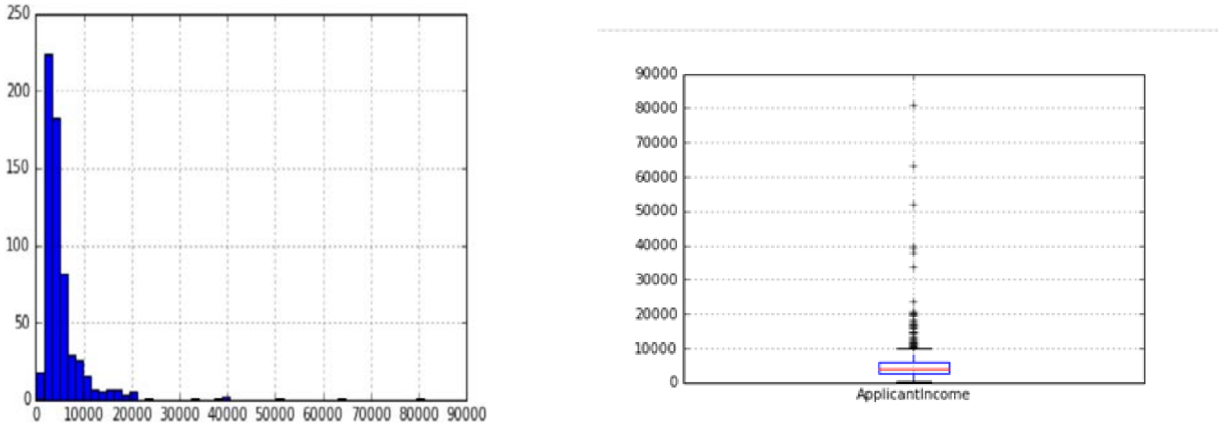
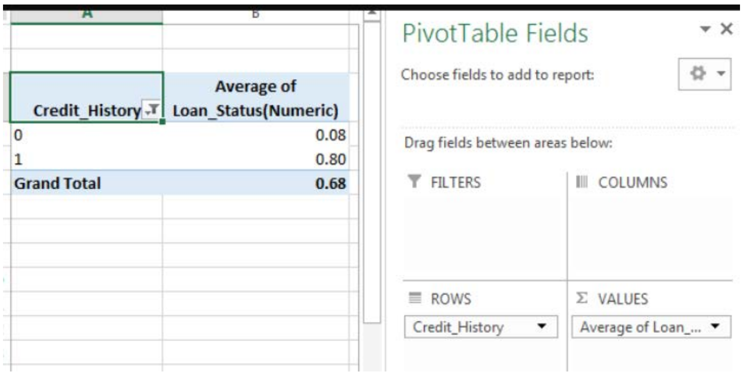


Fig2.Distribution Analysis

V. VARIABLE ANALYSIS

This section examines to understand the variables and its analysis . Excel style pivot table and cross-tabulation are used here. For instance, let us look at the chances of getting a loan based on credit history. This can be achieved in MS Excel using a pivot table as:



Credit_History	Average of Loan_Status(Numeric)
0	0.08
1	0.80
Grand Total	0.68

A
B
PivotTable Fields
Choose fields to add to report:
Drag fields between areas below:
FILTERS
COLUMNS
ROWS
VALUES

ROWS: Credit_History

VALUES: Average of Loan_...

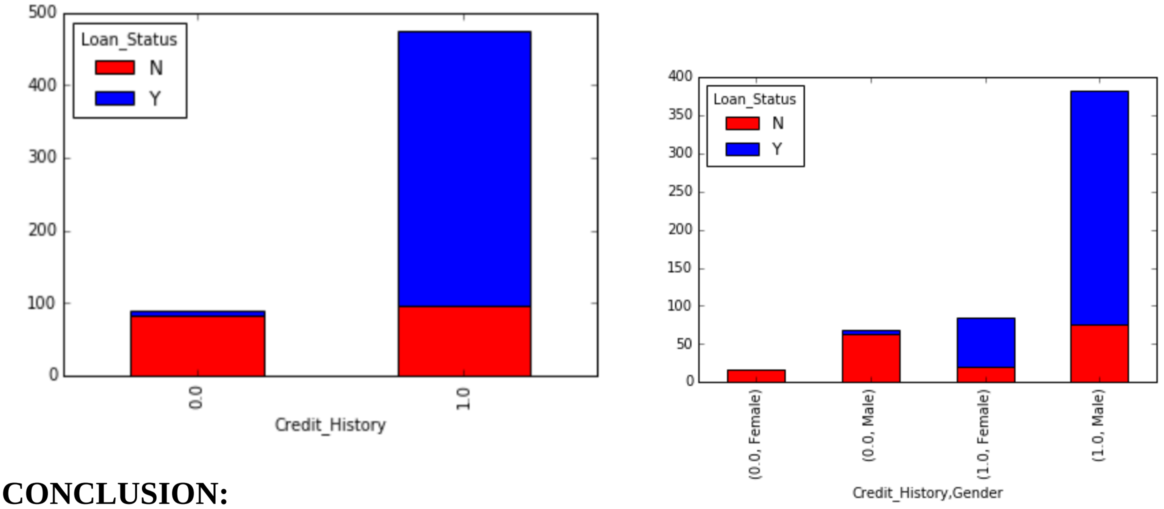
Fig 3.Variable Analysis

Here loan status has been calculated as 1 for Yes and 0 for No. So the mean represents the probability of getting loan.

VI. CREDIT HISTORY ANALYSIS

Looking at the loan_status graph, we note that Credit_history have more records and more than half of the applicants' loan have been approved. There are less applicants loan not approved but still more than half of their applications have been approved. We look at the charts with the same eye to evaluate how each category performed in regards to the approval of the loan.

Fig 4 Gender Analysis



CONCLUSION:

This is just a simple demonstration of gaining the insight of the data and Predicting the loan approval process in respective year. This Project focuses on loan prediction analysis by implementing descriptive statistics on loan dataset using Python. In future, the data munging techniques and various advanced analytics algorithms are used to predict the results. Various predictive models will be proposed to do the data analytics techniques such as regression, classification and clustering. The following table 2 gives a clear picture of numerical values of loan approval candidates and analysis.

S.No	Category	Count	Count - ID	Approval
1.	Males	614-213= 401	M01	55 Y 346 N
2.	Females	614-401=213	F02	Y 153 N 60
3.	Graduated	155	G01	Y 155
4.	Non Graduated	459	NG01	N 59 Y 450

Table 2 output summary

REFERENCES:

[1] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, HomaAtabakhsh, Crime Data Mining: An Overview and Case Studies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003, available at: http://ai.bpa.arizona.edu/

[2] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, HomaAtabakhsh, "Crime Data Mining: A General Frameworkand Some Examples", IEEE Computer Society April 2004.

[3] Road traffic accident data analysis and visualization "jamshidsodikov Tashkent Highway Engineering Institute, Published: May 05,2018".

[4] Analyzing traffic patterns on street segments based on GPS data "Emilian Necula Faculty of Computer Science, University Al, Published:July 2015.