Research Article

Loan Forecast by Using Machine Learning

A. Kulothungan[a], Neha[b], Himanshu[c], and Kanak Gupta[d]

[a] Assistant Professor (CSE), SRMIST, DELHI-NCR Campus, Modinagar
[b,c,d] M.Tech Student (CSE), SRMIST, DELHI-NCR Campus, MODINAGAR

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: In the improvement of banking sector many people are applied for bank loans but in bank has its limited benefit for given of limited people only, now searching out of which the loan can be given it will be a safe option for the bank is a classic process. So, in this project we try to reduce the risk factor behind collecting the safe people so to save lots of bank attempt or benefits. On the basis of this data/familiar, by using the machine learning model which give the most precise result. The main purpose of this project to forecast either imputing the loan given that person will be safe or not. In paper, we are forecast loan data by using some machine learning algorithm like Decision Tree, Logical Regression and Classification.

Keywords: Machine Learning, Decision Tree, Prediction, Python

1. Introduction

The main purpose of this paper is to forecast either imputing the loan to a certain person will be safe or not. Implemented this loan prediction problem using Decision Tree algorithm and data cleaning in Python as there are missing values in the dataset. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical attributes and 6 categorical attributes. The commendation of a customer for approve loan rely on multiple parameters, such as credit history, installment etc. [2]

2. Literature Review

Chitra Jalota, et. al. [9] shows that Educational Data Mining (EDM) can be helpful in developing a user action and vision model. Educational Data Mining is a regimentation that uses a data mining process in the educational atmosphere. In this methodology used, A nearby K-neighbor can be a effortless algorithm that keeps total possible affairs and distinguishes new affairs based on the proportionality rate (e.g., distance activities). KNN was used in statistical measurement and pattern esteem ahead in the early 1970s as a non-structural process. It can be a category of algorithms that attempt to accept the basic bond between a data group via procedure that imitate the way the human brain works. In this sense, neural networks question the method of neurons, either biotic or mock in nature.The target algorithm of the vector support machine is to search for a hyper plane in the N-dimensional space (N - number of elements) that clearly separates the data points.

Chiagoziem C. Ukwuoma, et. al. [10] shows that In this paper extant an in-deepness review of the literature on student performance crystal grazing and the process of data mining for crystal grazing student performance. The results showed a firm relationship amid student visits and execution. In this methodology used, Types of classifications predict labels for classified categories; and forecasting models predict continuous value functions. Consolidation is that the function of distributing the population or data points into different clusters such data points within the same clusters are almost just like other data points in the same cluster. In simple terms, the purpose is to distinguish clusters that have the same characteristics and are assigned to collections. Social network analysis differs from traditional collections. It requires the classification of objects based on their links and as their symbols. While a group of traditional art collections include only the same elements, and cannot be used in social network analysis.

Ismail Hmiedi, et. al. [11] shows that the skill of extracting usable acquaintance from big green data is a new period of profitable disquisition. Nowadays, due to the sharply growing corpus data to comprise all eventual regions such as medical, educational, merchandise and so, in the emerging field of data analysis. In this methodology used, Python is a translator, advanced programming language and general purpose. The Python architecture philosophy emphasizes the readability of the code with its remarkable use of large white. Its language-building and object-oriented approach aims to help editors in writing clear, logical code for little and enormous projects. It is integrated educated method for planning, deconstruction and variant activities that work by building a number of decision trees during exercitation and classification which is a class approach or predicting the meanings of each tree. Adding data to data analysis with unusual techniques to extend the quantity of data by adding modified copies of existing data or recently created data generated from existing data. It works as normal and helps to reduce imbalance when training a machine learning model.

Arushi Jain, et. al. [12] shows that Financial Fraud Detection can be a topic that affects many industries such as banking, insurance, government agencies, enforcement. A large number of currency exchanges and transactions occur daily and fraud cases are on the rise. In this methodology used Asset retrieval can also be a mathematical model that in its basic form uses a structural function to simulate binary dynamics, although there are more complex extensions. In multivariate analysis, asset order estimates the parameters of the order model.It specifies joint conditional probability distribution. They are also known as Belief Network, Bayesian Network or Probabilistic Network. It is a flowchart like skeleton in which each internal node represent a “test” on the attribute, each branch represent outgrowth of the test and each leaf node represents a class label.

Biswajit Panja, et. al. [13] shows that Crime is one of the major problems facing many urban areas compared to rural areas. There are various types of law enforcement such as crime, sex crime, theft, violent crime, arson, criminal / drug charges, cybercrime and more.In this methodology used It is a form of machine learning, where you do not need to monitor the model. Supervised reading allows you to collect data or generate data extraction from previous experiences. Unchecked machine reading helps you to discover all kinds of unknown patterns in data.The term KDD stands for Information Access. It refers to a comprehensive data acquisition process and emphasizes the high utilization of certain Data Mining techniques. It is a field of interest for researchers in a variety of fields, including artificial intelligence, machine learning, pattern recognition, data detail, statistics, master system information acquisition and data perception. It is a slice of computer program designed to mimic the way the human brain examine and procedure knowledge. It is the basis of artificial intelligence (AI) and extricates problems that may seem improbable at mathematical levels.

Muqaddas Gull, et. al. [14] shows that people don’t need to go anywhere to shop. The average person spends hours trying to find out what to do online. In this busy world, online shopping is best for working men and women but the consumer concerned is personal, psychological and social and cultural. In this methodology used, It is a procedure of selection and scaling knowledge on a variety of processes based on an established system, which empowers a person to answer relevant questions and analyze results. It is data not processed for use. The difference is occasionally built up between data and knowledge that data is the final product of data processing. Raw data that has been procedure in sometimes called cooked data. It is the process of purifying and converting raw data before processing and analyzing. It is an important step in the process of processing and often involves retrieving data, making data adjustments and integrating data sets to maximize data.

DATA MINING IN BANKING

It is the procedure of analyzing information from various origins and compressing it into contingent knowledge that may be used to help raise revenue and reduce expenditure. Its main intention is to find interactions or patterns between multiple zones in volumetric databases.

It has four major components:

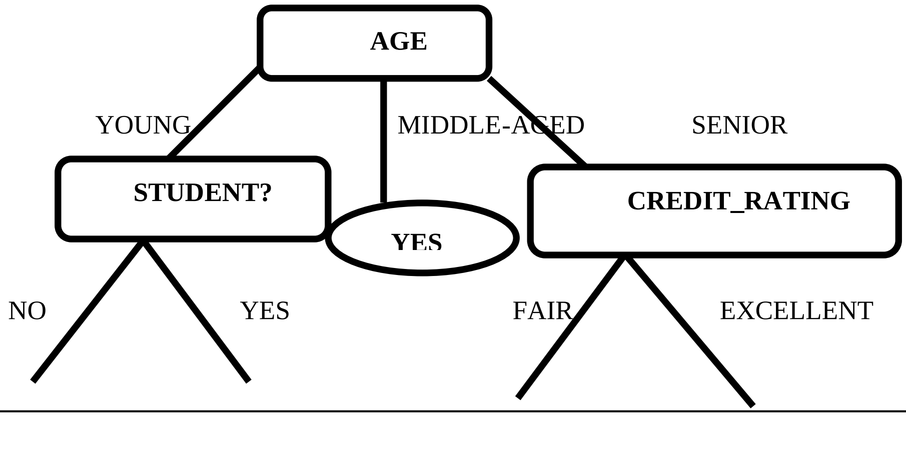
- Extract, modify, and upload transaction data to a database
- Store and manage data in various data systems
- Analyze data on application
- Presents information in a useful format

The data mining process is simple and formation of three stages. The first phase of testing usually strikes up with data processing which comprises Data Cleaning, Data Conversion and selection of record subsets and data sets with multiple variable numbers. Thereafter, appropriate dynamic identification and determination of model difficulties should be made to clarify the critical analysis using a variety of graphical and mathematical methods.

MACHINE LEARNING

Decision Tree

In machine learning method decision tree algorithm performs classification and regression [2]. Decision trees are widely used in the banking business due to high precise and potential to devise a statistical model in simple language.



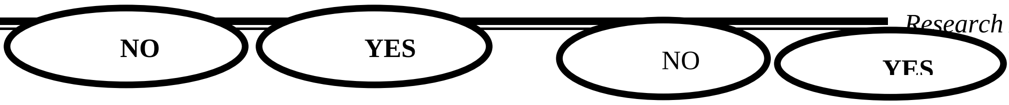


Fig. 1: Decision tree

Logistic Regression

Logistic Regression is a statistical model, it is the basic form uses a logistic function to model a binary dependent variable, although many more complex extension exist. In regression analysis, logistic regression is estimating the parameter of a logistic model.

PROPOSED MODEL

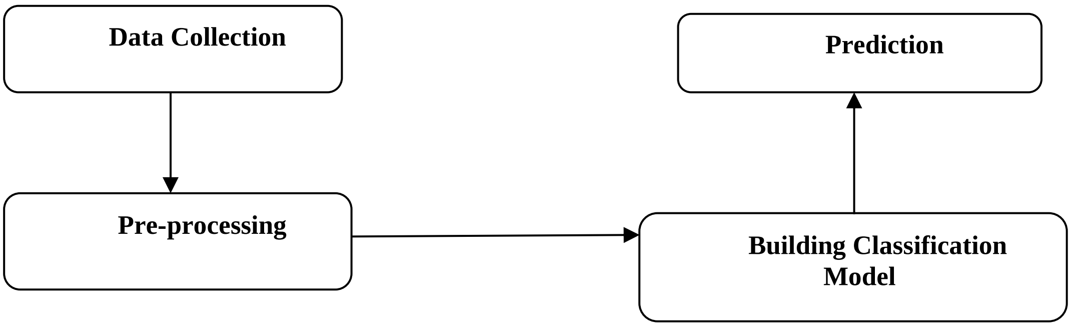
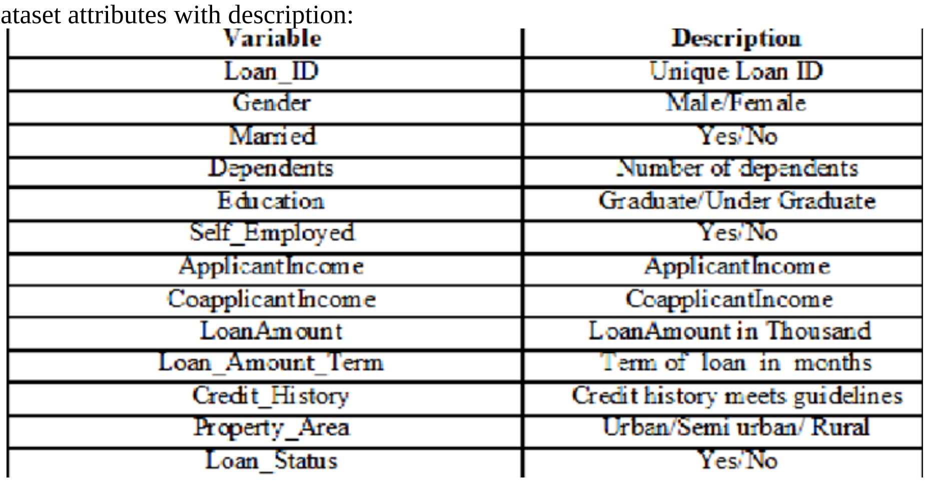


Fig. 2: Proposed Model

3. Data Collection

Data Collection means to collect the data to measure and analyze research for standard validated approaches. Below is the dataset attributes with description:



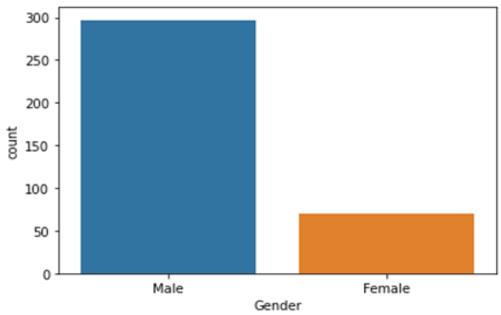
Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/Female
Married	Yes/No
Dependents	Number of dependents
Education	Graduate/Under Graduate
Self_Employed	Yes/No
ApplicantIncome	ApplicantIncome
CoapplicantIncome	CoapplicantIncome
LoanAmount	LoanAmount in Thousand
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/Semi urban/ Rural
Loan_Status	Yes/No

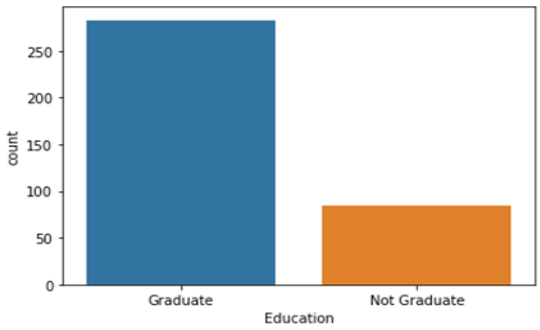
Table 1: Data Collection

IMPLEMENTATION

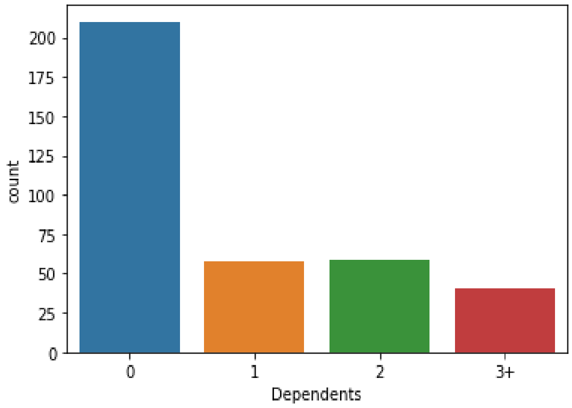
In this work, I have predicted loan status will be approved or not, it will show in Yes/No. This is some of terms like Gender for Male/Female, Married for Yes/No. Here various attributes I will deal with various attributes fill the missing values and do some analysis to create some new attributes like I will see many assets in the project. Firstly, Import the module that I needed in the project and then load the dataset. After the loading dataset, find the missing values in numerical term in mean operation (LoanAmount, Loan_Amount_Term, Credit_History) and then find the missing values in categorical term with the help of mode operation (Gender, Married, Dependents, Self_Employed) mode operation means most frequently occurring values.

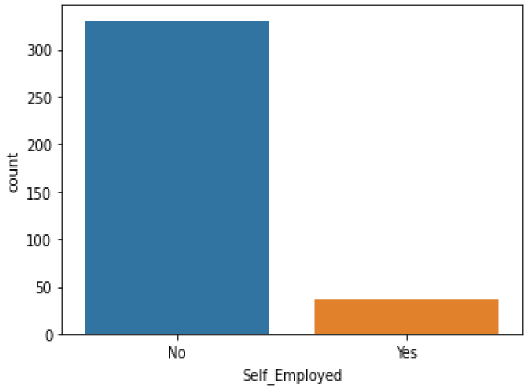
After that Expletory Data Analysis, in this analysis the data for each attributes like Gender, Education, Dependents, Self_Employed. From this analysis I got intuition to building the model for other categorical attributes.





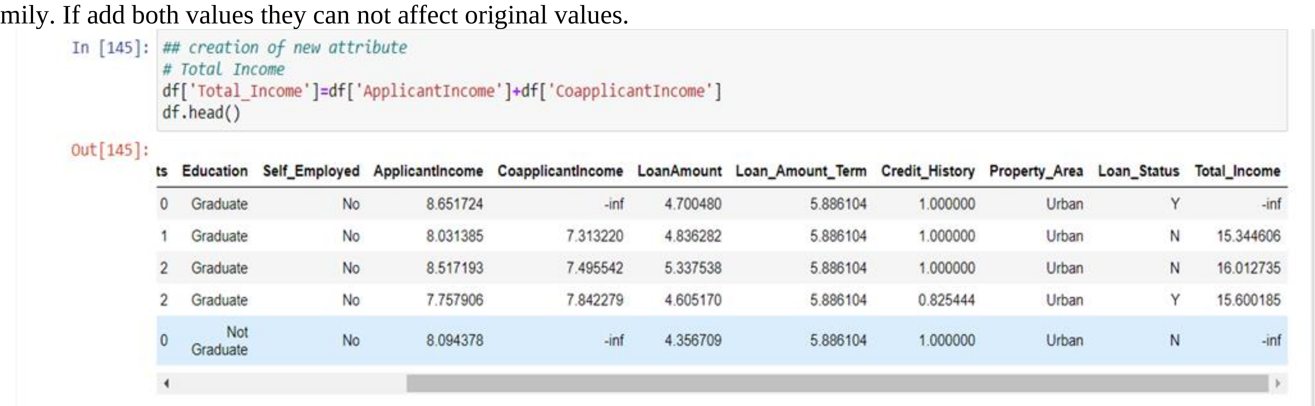
- In Gender, Majority of applicants are Male only small portion of applicants are Female.
- In Education, Majority of Graduation is more as compare to not graduate applicants.





- In Dependents, Most of the Loan applicants don't have no dependents, around 50-60 people 1&2 dependents.
- In Self_Employed, Assume some people are doing business and freelancing.

Creation of new attribute: Now let's see the data, do some statistic operation here the ApplicantIncome and CoapplicantIncome added both of them and make a new attribute is called TotalIncome because they are same family. If add both values they can not affect original values.



```
In [145]: ## creation of new attribute
          # Total Income
          df['Total_Income']=df['ApplicantIncome']+df['CoapplicantIncome']
          df.head()
```

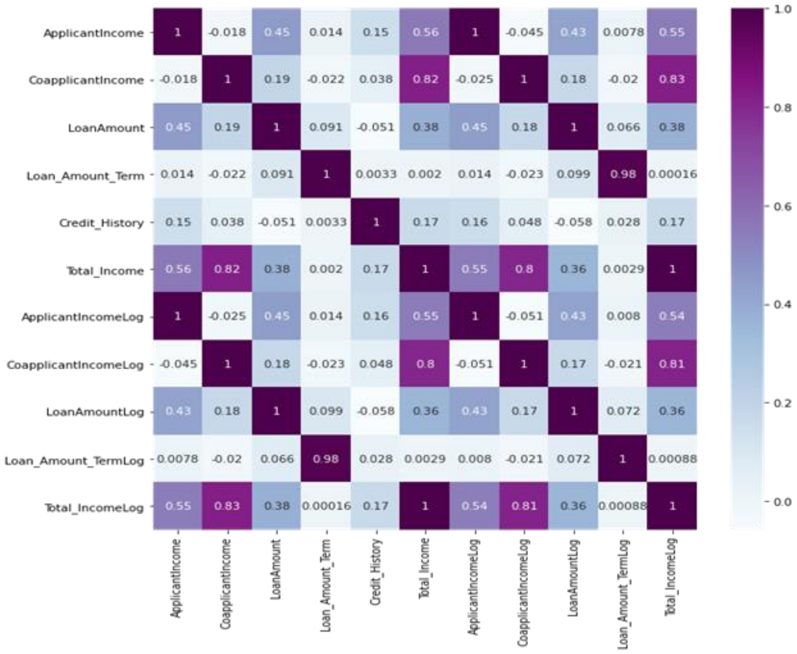
Out[145]:

ts	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status	Total_Income
0	Graduate	No	8.651724	-inf	4.700480	5.886104	1.000000	Urban	Y	-inf
1	Graduate	No	8.031385	7.313220	4.836282	5.886104	1.000000	Urban	N	15.344606
2	Graduate	No	8.517193	7.495542	5.337538	5.886104	1.000000	Urban	N	16.012735
2	Graduate	No	7.757906	7.842279	4.605170	5.886104	0.825444	Urban	Y	15.600185
0	Not Graduate	No	8.094378	-inf	4.356709	5.886104	1.000000	Urban	N	-inf

Fig 3: Creation of new attribute

Correlation Matrix:

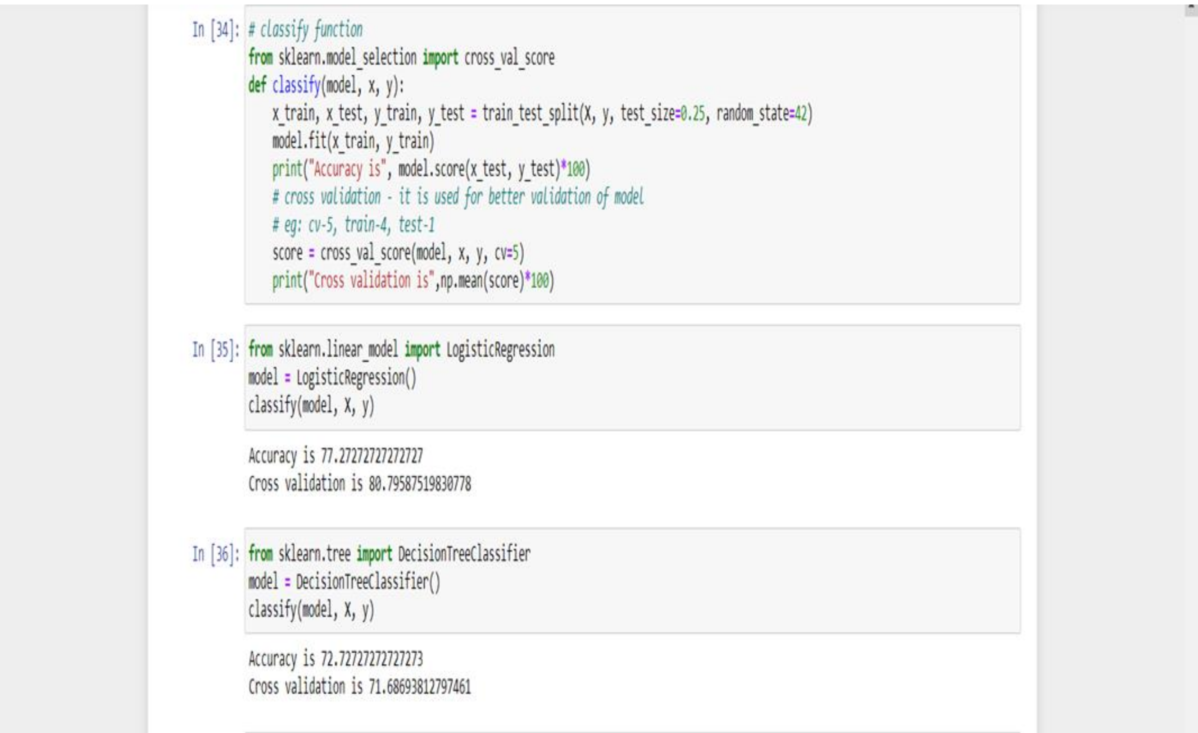
Correlation Matrix is simply a table which which display the correlation. In this matrix, high density plotted with dark color and low density plotted with light color.



	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Total_Income	ApplicantIncomeLog	CoapplicantIncomeLog	LoanAmountLog	Loan_Amount_TermLog	Total_IncomeLog
ApplicantIncome	1	-0.018	0.45	0.014	0.15	0.56	1	-0.045	0.43	0.0078	0.55
CoapplicantIncome	-0.018	1	0.19	-0.022	0.038	0.82	-0.025	1	0.18	-0.02	0.83
LoanAmount	0.45	0.19	1	0.091	-0.051	0.38	0.45	0.18	1	0.066	0.38
Loan_Amount_Term	0.014	-0.022	0.091	1	0.0033	0.002	0.014	-0.023	0.099	0.98	0.00016
Credit_History	0.15	0.038	-0.051	0.0033	1	0.17	0.16	0.048	-0.058	0.028	0.17
Total_Income	0.56	0.82	0.38	0.002	0.17	1	0.55	0.8	0.36	0.0029	1
ApplicantIncomeLog	1	-0.025	0.45	0.014	0.16	0.55	1	-0.051	0.43	0.008	0.54
CoapplicantIncomeLog	-0.045	1	0.18	-0.023	0.048	0.8	-0.051	1	0.17	-0.021	0.81
LoanAmountLog	0.43	0.18	1	0.099	-0.058	0.36	0.43	0.17	1	0.072	0.36
Loan_Amount_TermLog	0.0078	-0.02	0.066	0.98	0.028	0.0029	0.008	-0.021	0.072	1	0.00088
Total_IncomeLog	0.55	0.83	0.38	0.00016	0.17	1	0.54	0.81	0.36	0.00088	1

Fig 4: Correlation Matrix

4. Result:



```
In [34]: # classify function
         from sklearn.model_selection import cross_val_score
         def classify(model, x, y):
             x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
             model.fit(x_train, y_train)
             print("Accuracy is", model.score(x_test, y_test)*100)
             # cross validation - it is used for better validation of model
             # eg: cv-5, train-4, test-1
             score = cross_val_score(model, x, y, cv=5)
             print("Cross validation is",np.mean(score)*100)
```

```
In [35]: from sklearn.linear_model import LogisticRegression
         model = LogisticRegression()
         classify(model, X, y)
```

```
Accuracy is 77.27272727272727
Cross validation is 80.79587519830778
```

```
In [36]: from sklearn.tree import DecisionTreeClassifier
         model = DecisionTreeClassifier()
         classify(model, X, y)
```

```
Accuracy is 72.72727272727273
Cross validation is 71.68693812797461
```

Confusion Matrix:

- It’s also known as Heat map.
- It’s used to reduce the errors

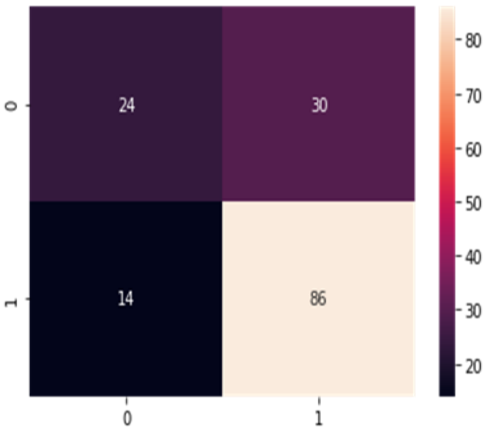


Fig 5: Confusion Matrix

- In the left side, actual values are present and the bottom side, predicted values are present.
- Left diagonal part correctly predicted and right diagonal part not predicted correctly.

5. Conclusion

The systematical procedure begin from data cleaning and processing, Missing value assigning with mice package, then exploratory analysis and finally model building and evaluation. Accuracy is 77.2727 and Cross Validation is 80.7958. A study on sorts of data mining methods, machine learning and many more learning techniques and some algorithm also. Overall, this paper gives the knowledge of how to collect the data for classification of bank data set with the use of different forms of algorithm, methods and techniques.

References

1. Cowell,R.G.,A.P.,Lauritez,S.L.,and Spiegelhalter,D.J.(1999).Graphical models and Expert Systems. Berlin: Springer. This is a good introduction to probabilistic graphical models.
2. Kumar Arun, Garg Ishan, Kaur Sanmeet, May- Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)
3. Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision- Making, Ministry of Education, Hefei University of Technology, Hefei 23009, China
4. Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.
5. Research on bank credit default prediction based on data mining algorithm, The International Journal of Social Sciences and Humanities Invention 5(06): 4820-4823, 2018.
6. Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C. Sealand on july 20, 2018.
7. http://www.ijetjournal.org/volume5/issue2/IJET-V5I2P28.pdf
8. http://sersc.org/journals/index.php/IJAST/article/view/460
9. C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 243-247, doi: 10.1109/COMITCon.2019.8862214.
10. C. C. Ukwuoma, C. Bo, I. A. Chikwendu and E. Bondzie-Selby, "Performance Analysis of Students Based on Data Mining Techniques: A Literature Review," 2019 4th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand, 2019, pp. 1-5, doi: 10.1109/TIMES-iCON47539.2019.9024396.
11. Hmiedi, H. Najadat, Z. Halloush and I. Jalabneh, "Semi Supervised Prediction Model in Educational Data Mining," 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, 2019, pp. 27-31, doi: 10.1109/ACIT47987.2019.8991048.
12. Jain and S. Shinde, "A Comprehensive Study of Data Mining-based Financial Fraud Detection Research," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp.1-4, doi: 10.1109/I2CT45611.2019.9033767.

13. Panja, P. Meharia and K. Mannem, "Crime Analysis Mapping, Intrusion Detection - Using Data Mining," 2020 IEEE Technology & Engineering Management Conference (TEMSCON), Novi, MI, USA, 2020, pp. 1-5, doi: 10.1109/TEMSCON47658.2020.9140074.
14. M. Gull and A. Pervaiz, "Customer Behavior Analysis Towards Online Shopping using Data Mining," 2018 5th International Multi-Topic ICT Conference (IMTIC),Jamshoro, 2018, pp. 1-5, doi: 10.1109/IMTIC.2018.8467262