A Mini Project Report
On
Loan Prediction

Submitted to
Pune Institute of Computer Technology, Pune

In partial fulfillment for the award of the Degree of

Bachelor of Engineering
in
Information Technology
by

Dyaneshwari Abuj	33301
Soham Ahire	33303
Amod Dhopavkar	33304
Arpit Singh Batra	33306
Aniket Bedare	33310

Under the guidance of → **Prof. S. B. Deshmukh and Dr. A. M. Bagade**

Department Of Information Technology

Pune Institute of Computer Technology, College of Engineering

Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043

2019-2020

CERTIFICATE

This is to certify that the project report entitled

Loan Prediction

Submitted by →

Dyaneshwari Abuj	33301
Soham Ahire	33303
Amod Dhopavkar	33304
Arpit Singh Batra	33306
Aniket Bedare	33310

is a bonafide work carried out by them under the supervision of Prof. S. B. Deshmukh and Dr. A. M. Bagade and it is approved for the partial fulfillment of the requirement of Software Laboratory Course-2015 for the award of the Degree of Bachelor of Engineering (Information Technology)

Prof. S. B. Deshmukh
Internal Guide
Department of Information Technology

Dr. A. M. Bagade
Head of Department
Department of Information Technology

Date : 15/04/2020
Place: Pune

ABSTRACT

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. We try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which gives the most accurate result. The main objective is to predict whether assigning the loan to a particular person will be safe or not.

CONTENTS

List of Tables

List of Figures

1. Introduction:

a. Motivation

1. Why is loan prediction important?
2. What exactly is the reason that despite the use of similar technologies, the NPAs are at their highest ever valuation.

b. Purpose/need/application

1. To find the various factors depending on which a particular loan is sanctioned.
2. To determine the overall percentage of loans sanctioned for different groups/ classes of people.
3. To find definitively whether there is a gender bias observed in the sanctioning of the loans by the banks.

2. Problem Description:

Distribution of the loans is the core business part of almost every bank. The main portion of the bank's assets directly came from the profit earned from the loans distributed by the banks. The prime objective in the banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approve loans after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasizes different weights to each factor but in real life sometimes loans can be approved on the basis of a single strong factor only, which is not possible through this system. Loan Prediction is very helpful for employees of banks as well as for the applicant also. The aim of this Paper is to provide a quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each feature taking part in loan processing and on new test data the same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific applications so that it can be checked on priority basis. This Paper is exclusively for the managing authority of the Bank/finance company, the whole process of prediction is done privately no stakeholders would be able to alter the processing. Results against particular Loan Id can be sent to various departments of banks so that they can take appropriate action on application. This helps all others departments to carry out other formalities.

3. Literature Survey:

3.1. Machine Learning

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data, such as from sensor data or databases. Machine Learning is a scientific discipline that addresses the following question: 'How can we program systems to automatically learn and to improve with experience? 'Learning in this context is not learning by heart but recognizing complex patterns and making intelligent decisions based on data. The difficulty lies in the fact that the set of all possible decisions given all possible inputs is too complex to describe. To tackle this problem the field of Machine Learning develops algorithms that discover knowledge from specific data and experience, based on sound statistical and computational principles.

Six Machine Learning Models are used for prediction. Their details are given below →

1. **Decision Trees (C5.0):** The basic algorithm of decision tree [7] requires all attributes or features should be discretized. Feature selection is based on greatest information gain of features. The knowledge depicted in the decision tree can be represented in the form of IF-THEN rules. This model is an extension of C4.5 classification algorithms described by Quinlan.

2. **Random Forest (RF):** Random forests [8] are a group learning system for characterization (and relapse) that work by building a large number of Decision trees at preparing time and yielding the class that is the mode of the classes yield by individual trees.

3. **Support Vector Machine (SVM):** Support vector machines are administered learning models that use association or learning algorithms which analyze features and identified pattern knowledge, utilized for application classification. SVM can productively perform a regression utilizing the kernel trick, verifiably mapping their inputs into high dimensional feature spaces [9].

4. **Linear Models (LM):** The Linear Model [10] is numerically indistinguishable to a various regression analysis yet burdens its suitability for both different qualitative and numerous quantitative variables.

5. **Neural Network (Nnet):** Neural networks [14] are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

6. **Adaboost (ADB):** Adaboost short for " Adaptive Boosting ". It is delicate to noisy information data and outliers. It is different from neural systems and SVM because Adaboost preparing methodology chooses just those peculiarities known to enhance the divining power of the model, decreasing dimensionality and conceivably enhancing execution time as potentially features don't have to be processed.[14]

3.2 Random Forest

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes the model's prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: "A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models."

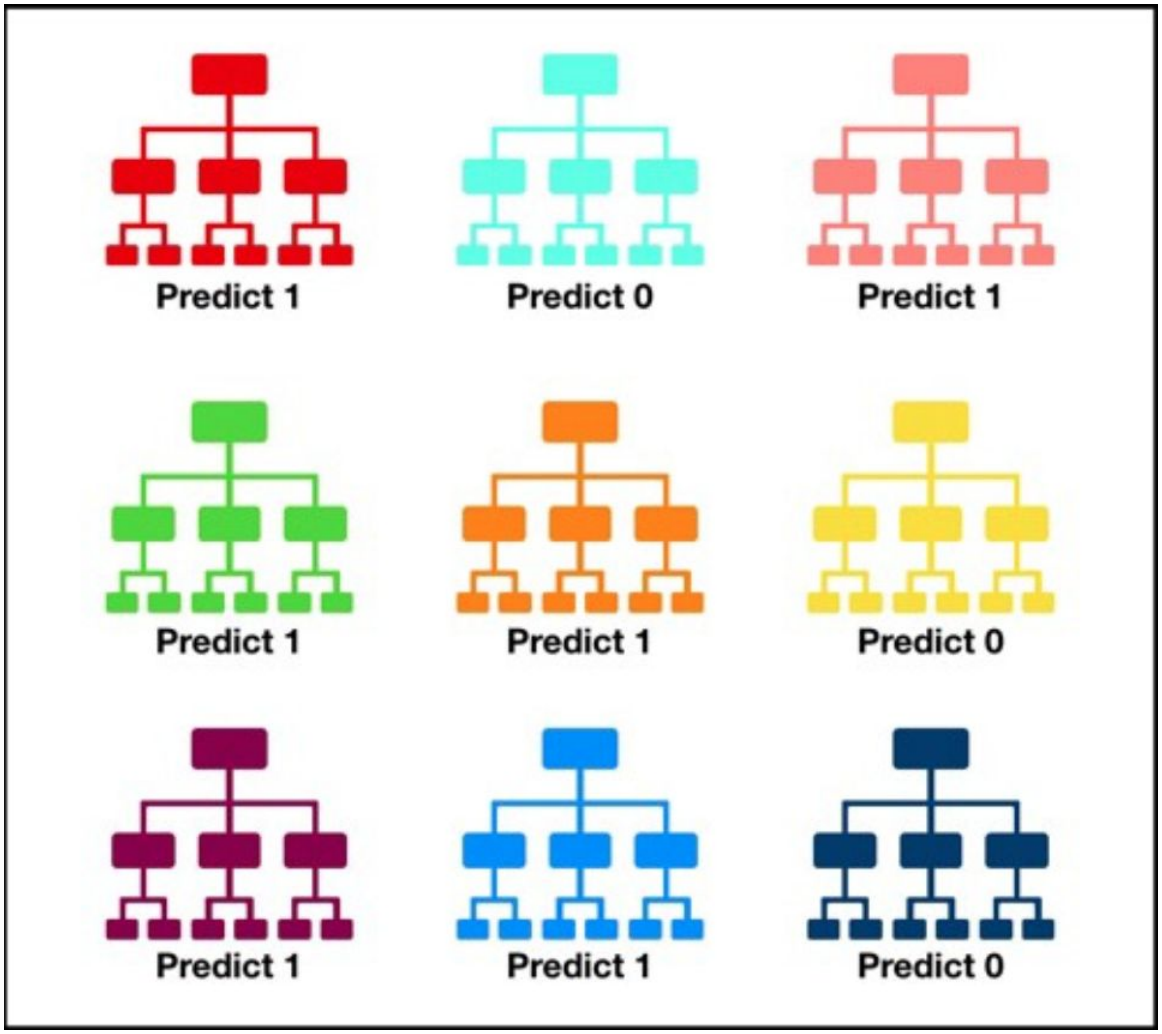


Fig 1. Random Forest Explanation

Tally → Six 1s and 3 0s. Therefore, final prediction 1

The low correlation between models is the key. Just like how investments with low correlations (like stocks and bonds) come together to form a portfolio that is greater than the sum of its parts, uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

The prerequisites for random forest to perform well are:

1. There needs to be some actual signal in our features so that models built using those features do better than random guessing.

2. The predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

In Random Forest, our chances of making the correct prediction increases with increase in the number of unrelated trees. That is, higher the number of trees, more accurate will be our model.

3.3 Spearman's Rank Correlation

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

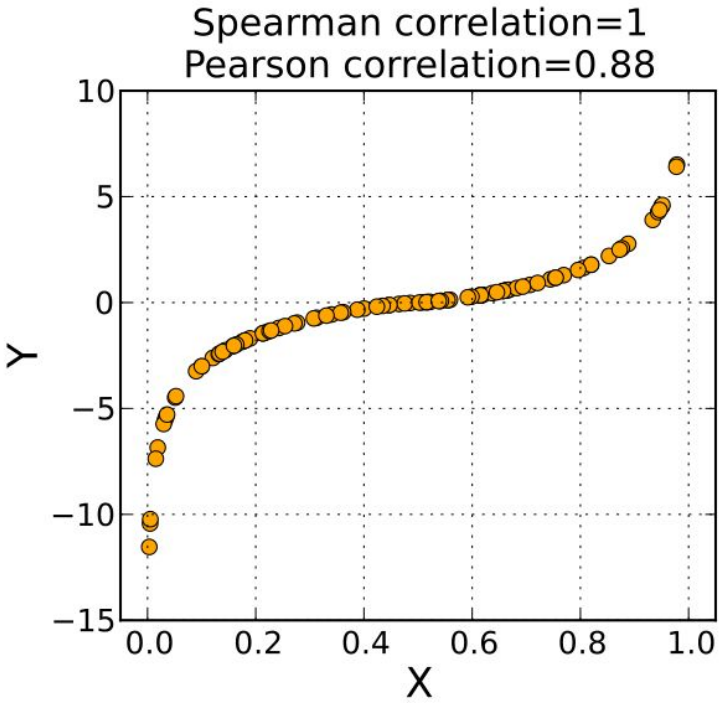


Fig 2. Spearman Correlation

3.4 Data Set

The training data set is now supplied to the machine learning model, on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, models predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it concludes on the basis of the training data sets.

Sr. No.	Variable Name	Description	Type
1.	Loan_ID	Unique Loan ID	Integer
2.	Gender	Male/ Female	Character
3.	Maritial_Status	Applicant married (Y/N)	Character
4.	Dependants	Number of dependents	Integer
5.	Education_Qualification	Graduate/ Undergraduate	String
6.	Self_Employed	Self Employed (Y/N)	Character
7.	Applicant_Income	Applicant income	Integer
8.	Co_Applicant_Income	Co-Applicant income	Integer
9.	Loan_Amount	Loan amount in thousands	Integer
10.	Loan_Amount_Term	Term of loan in months	Integer
11.	Credit_History	Credit history meets guidelines	String
12.	Property_Area	Urban/ Semi Urban/ Rural	Integer
13.	Loan_Status	Loan Approved(Y/N)	Character

Tab 1. Data Set Description

3.5 Loan Prediction Methodology

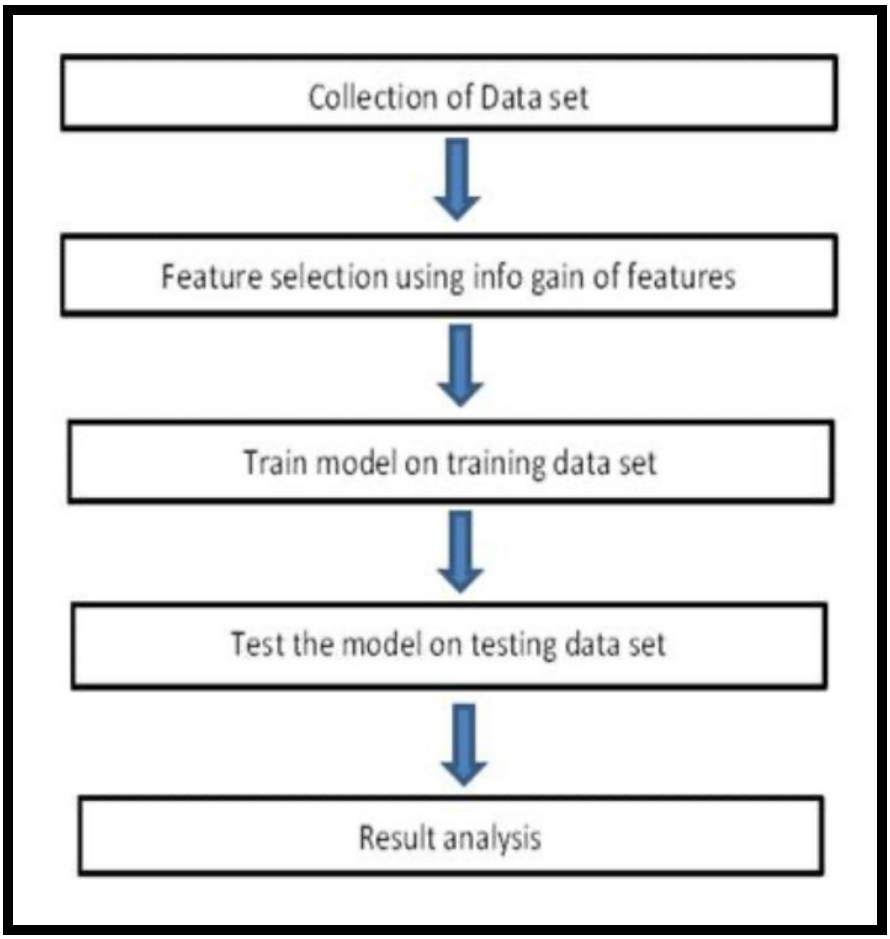


Fig 3. Loan Prediction Methodology

4. Design and Implementation:

Irrespective of the process model employed, software development comprises several tasks. Requirements acquisitions, conceptual modelling, risk analysis, database design, coding, testing and software maintenance are some of the tasks involved. Further, each task entails a specific set of skills for its accomplishment.

In this section, we are going to discuss the modular high level architecture describing various modules and interaction among these modules. Architecture of our development cycle can be mainly divided into three stages namely:-

4.1 Dataset Collection

This phase is one of the most important phases of the project development cycle as it is the only dataset which defines the level of accuracy and stability that will be achieved and reflected in results. Like every other project this one also starts with collection of database which comprises of various songs name along with their ranks based on yearly basis

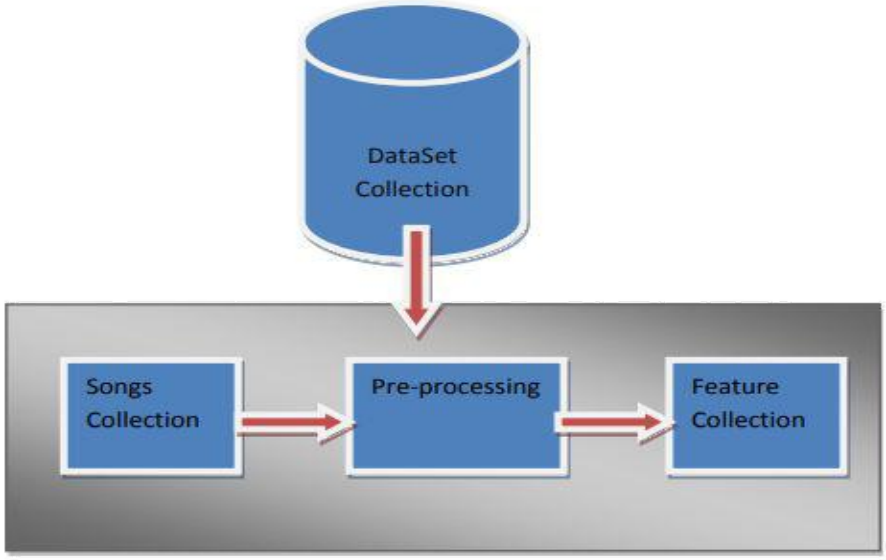


Fig 3.1 Composition of phase 'Data Collection'

Fig 4. Dataset Collection

4.2 Data Preprocessing

1. **Data cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

2. **Data integration:** Using multiple databases, data cubes, or files. Integration of Various files.

3. **Data transformation:** Normalization and Aggregation.

4.2.1 Data Cleaning

The dataset used here contains NaN values and outliers. These not applicable values and the values which are not correct must be either removed or replaced as they may alter the results. Removing such values results in reduction of the number of records, which again impacts the accuracy of our model.

Thus, the NaN values and outliers are replaced with the **median** value.

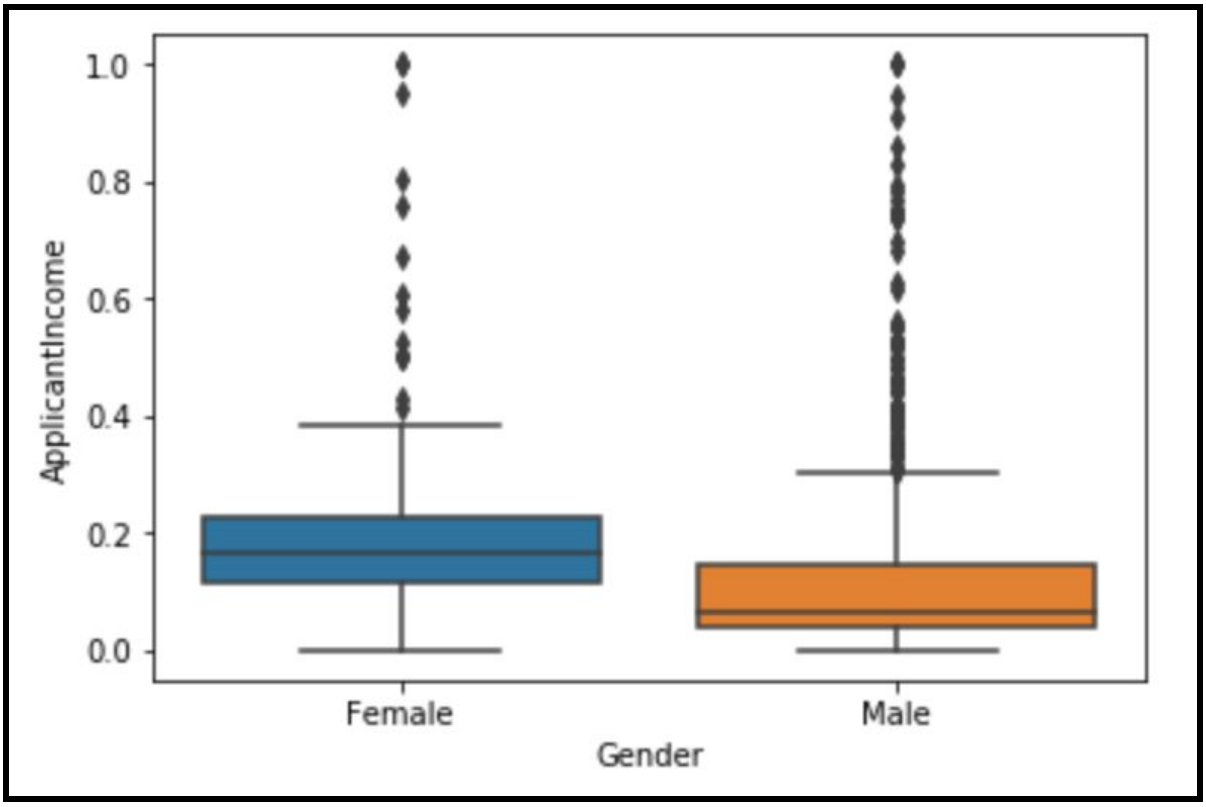


Fig 5. Outliers in the Dataset

4.3 Importance of Columns

The importance of columns when considering a particular loan request is to be sanctioned is determined. This gives the primary categories depending upon which, the decision of whether the request is to be accepted is made.

	cols	imp
10	Credit_History	0.664756
9	Loan_Amount_Term	0.073993
6	ApplicantIncome	0.072821
7	CoapplicantIncome	0.056520
11	Property_Area	0.037865
8	LoanAmount	0.025991
12	LoanAmount_na	0.021695
3	Dependents	0.017915
0	Loan_ID	0.011977
2	Married	0.009735
14	Credit_History_na	0.006335
4	Education	0.000397
1	Gender	0.000000
5	Self_Employed	0.000000
13	Loan_Amount_Term_na	0.000000

Fig 6. Importance of columns in the Dataset

4.4 Spearman Correlation

The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation. Spearman's correlation coefficient, (ρ, also signified by rs) measures the strength and direction of association between two ranked variables.

Here, the Spearman Correlation is used to visualise the connection between two related columns

```
# Correlation Function
def correlate(df2):
    corr = np.round(scipy.stats.spearmanr(df2).correlation, 4)
    corr_condensed = hc.distance.squareform(1-corr)
    z = hc.linkage(corr_condensed, method='average')
    fig = plt.figure(figsize=(16,10))
    dendrogram = hc.dendrogram(z, labels=df2.columns, orientation='left', leaf_font_size=16)
    plt.show()
```

Fig 7. Spearman Correlation Function Code

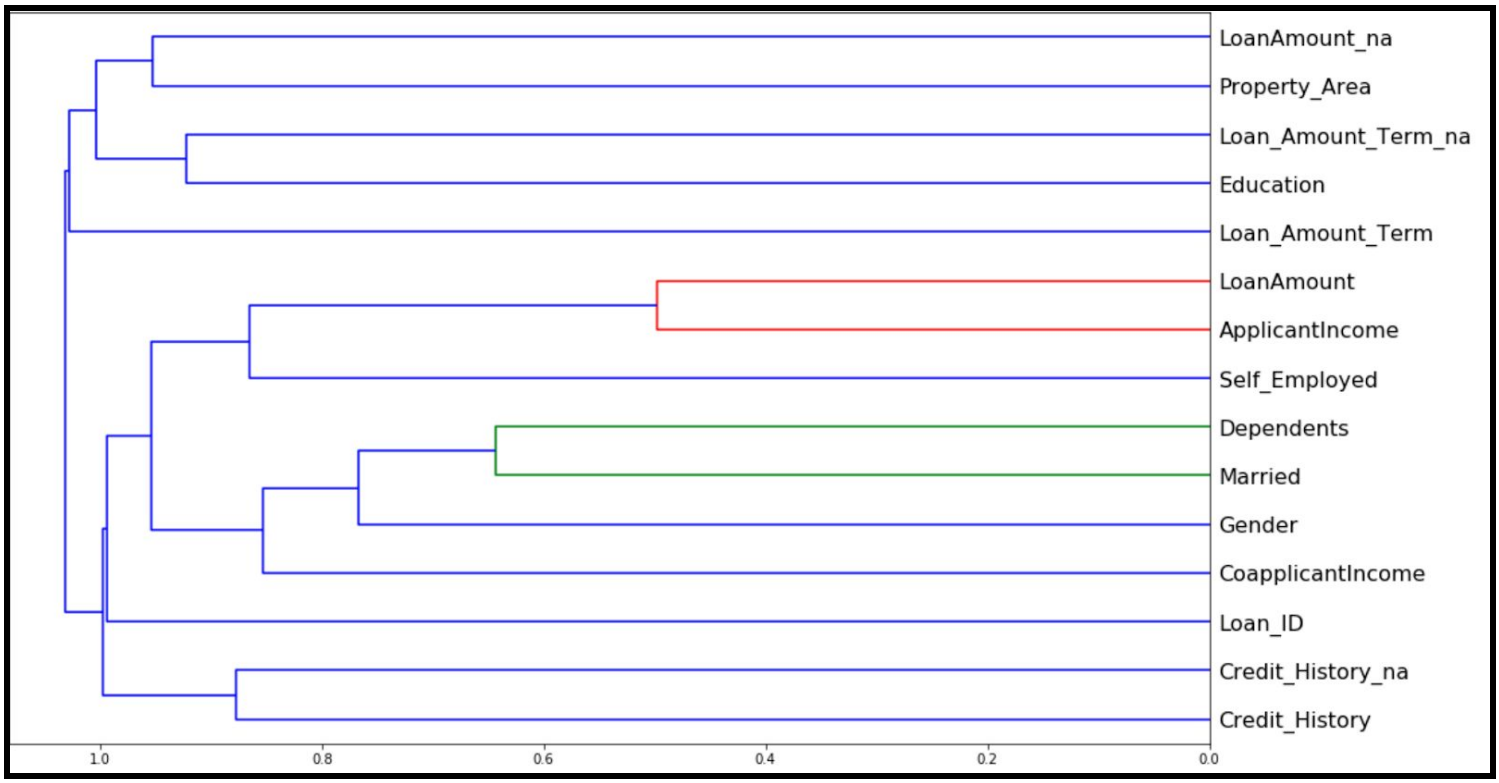


Fig 8. Spearman Correlation Function Graph

4.5 Random Forest Classifier

Random Forest Classifier is the most accurate currently available technique of classification. In this method, a collection of unrelated decision trees work together to predict the outcome. Random Forest works well only with unrelated decision trees. More the number of trees in the forest, higher will be the accuracy of the model.

For the given task of loan prediction, a Random Forest Classifier with →
Number of Trees = 40
Max. Depth = 2
Min. leaves = 2

```
def RF(df,Xtest3,ytest3):
    predictions3 = clf1.predict(X_test3)
    m=metrics.accuracy_score(y_test3, predictions3)
    return m
```

```
clf1=RandomForestClassifier(n_estimators=40, min_samples_leaf=2, max_features=0.5,max_depth=2)
X_train3, X_test3, y_train3, y_test3 = train_test_split(df2,y,test_size=0.33,shuffle=True,random_state=42)
clf1.fit(X_train3,y_train3)
print("Accuracy:",RF(df2,X_test3,y_test3))
```

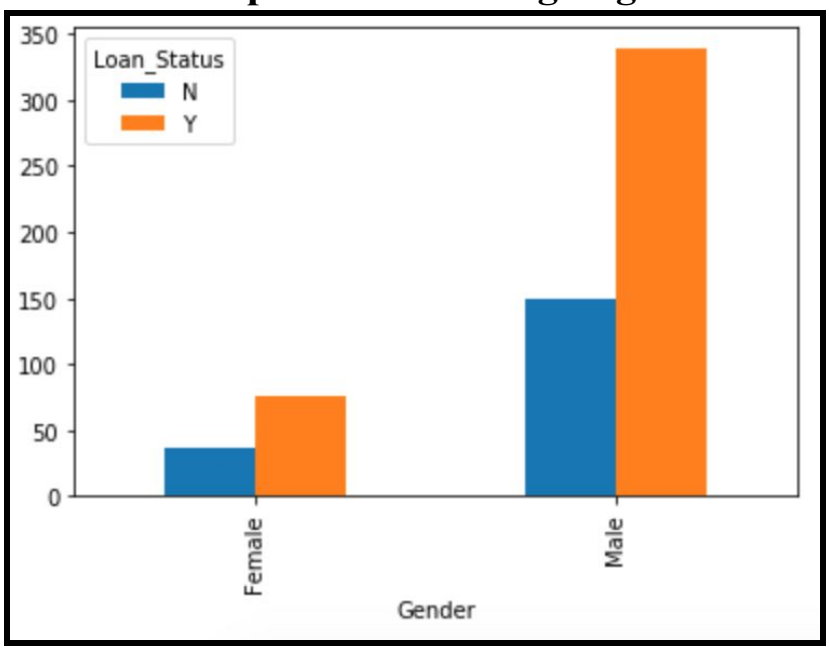
```
Accuracy: 0.7980295566502463
```

Fig 9. Random Forest Classifier Code

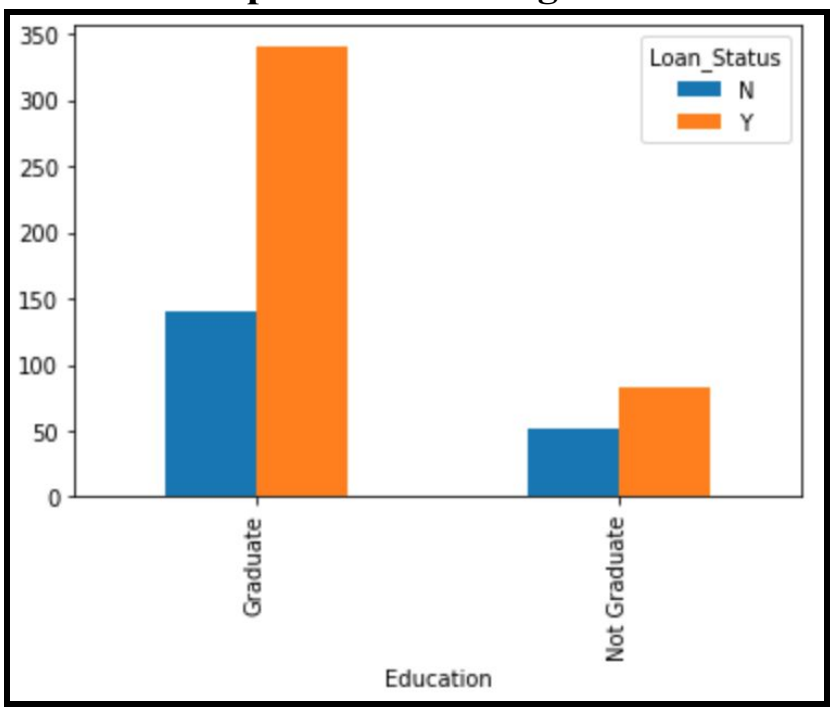
Accuracy = 0.7980295566502463
∴ Accuracy ≈ 79.80%

5. Evaluation:

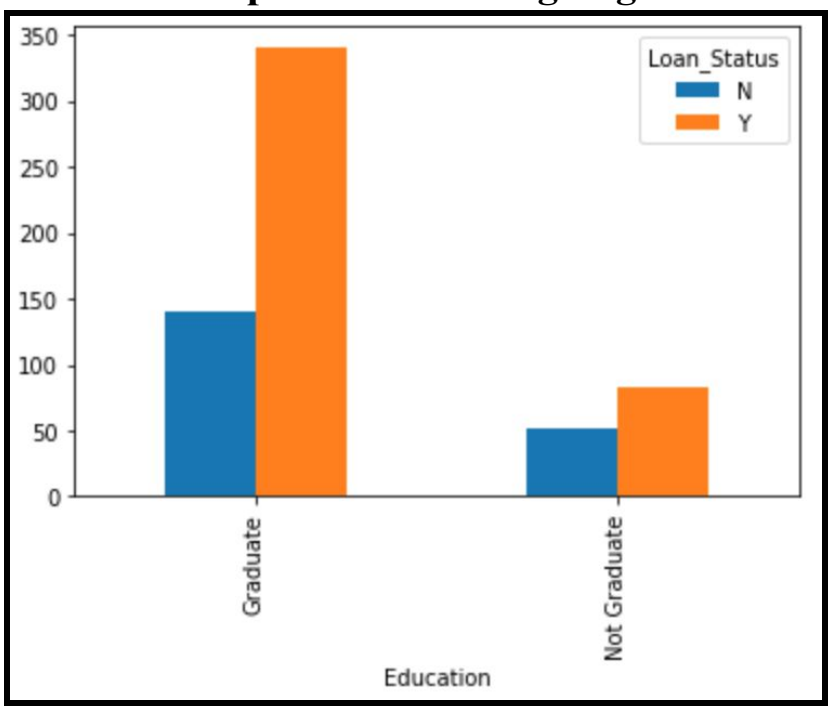
5.1 Loan acceptance according to gender



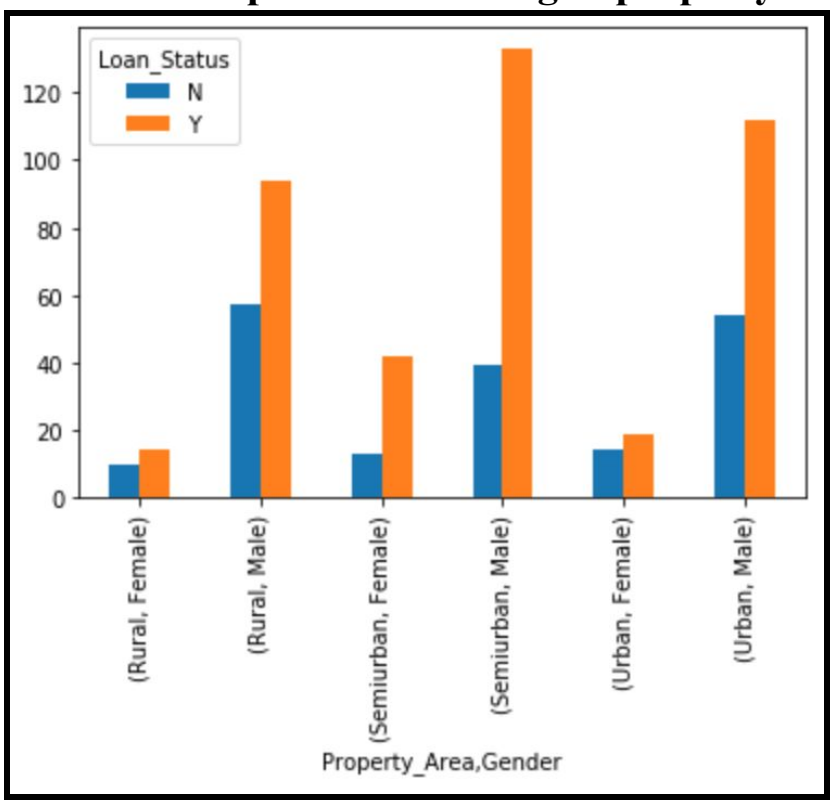
5.2 Loan acceptance according to education



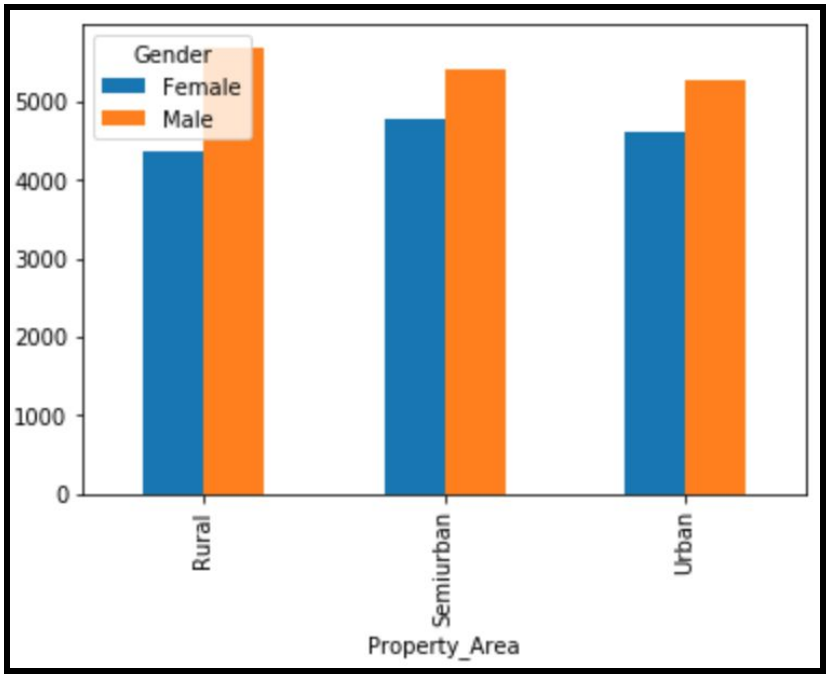
5.3 Loan acceptance according to gender and education



5.4 Loan acceptance according to property area and gender



5.5 Property area owned according to gender



6. Results:

The Random Forest Classifier is the most accurate classifier that is currently available. Using Random Forest Model with the number of trees set to 40, and max depth as 2 we were able to achieve an accuracy of →

Approach Used	Accuracy (in %)
Random Forest Classifier	79.80

Tab 2. Accuracy of Random Forest Classifier

Gender	Ratio of Acceptance
Male	0.6932515337423313
Female	0.6696428571428571

Tab 3. Ratio of Acceptance

```
cnt10,cnt11=0,0
for i,j in enumerate(df["Gender"]):
    if df["Loan_Status"][i]=="Y" and df["Gender"][i]=="Male":
        cnt10+=1
    if df["Loan_Status"][i]=="Y" and df["Gender"][i]=="Female":
        cnt11+=1
print("Number of males accepted",cnt10,"Ratio",cnt10/489)
print("Number of females accepted",cnt11,"Ratio",cnt11/112)
```

```
Number of males accepted 339 Ratio 0.6932515337423313
Number of females accepted 75 Ratio 0.6696428571428571
```

Fig 10. Final acceptance ratio according to genders

7. Conclusion:

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system, So in the near future the so –called software could be made more secure, reliable and dynamic weight adjustment .In near future this module of prediction can be integrate with the module of automated processing system. The system is trained on old training dataset in future software that can be made such that new testing date should also take part in training data after some fix time.

8. Tools Used:

The various tools used for this project are as follows →

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit Learn
5. Jupyter Notebook
6. Train_cats
7. Anaconda
8. Python 3
9. macOS Catalina 10.15.4

9. References:

[1]. Rattle data mining tool: available from http://rattle.togaware.com/rattle-download.html

[2]. Aafer Y, Du W & Yin H 2013, DroidAPIMiner: 'Mining API-Level Features for Robust Malware Detection in Android', in: Security and privacy in Communication Networks Springer, pp 86-103 .

[3]. Ekta Gandotra, Divya Bansal, Sanjeev Sofat 2014, 'Malware Analysis and Classification: A Survey Available from http:// www.scirp.org/journal/jis

[4]. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: International Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).

[5]. J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[6]. Mean Decrease Accuracy https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.hml

[7]. J.R. Quinlan. Induction of decision trees. MachinelearningSpringer, 1(1):81–106, 1086.

[8]. Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News(http://CRAN.R-project.org/doc/Rnews/), 2(3):9–22, 2002.

[9]. S.S. Keerthi and E.G. Gilbert. Convergence of a generalizeSMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.

[10]. J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.