LOAN PREDICTION USING MACHINE LEARNING

Aman Sharma
M.Sc. in Big Data Analytics, Department of Computer Science
Ramakrishna Mission Vivekananda Educational And Research Institute
Belur Math, Howrah
Pin-711202, West Bengal
June 27, 2020

A **Project Report** SUBMITTED
TO RAMAKRISHNA MISSION VIVEKANANDA EDUCATIONAL AND RESEARCH INSTITUTE
FOR THE COURSE OF **MACHINE LEARNING** IN **M.Sc. in Big Data Analytics**

ACKNOWLEDGEMENT

I take immense pleasure in thanking the great many people who helped in completion of my project associated with it. My deepest thanks to my faculty guide i.e, Dr./Br. Dripta Mj. for guiding and correcting me with attention and care while doing the project Loan Prediction Using Machine Learning. I greatly appreciate the efforts he took to go through my data and make necessary corrections as and when needed. I also want to express my thanks to Dr./Br. Mrinmay Mj., Professor and Swathy Prabhu Mj., H.O.D for extending their support. Words are inadequate in offering my thanks to the various helpful people of RKMVERI for their encouragement and cooperation in carrying out the project work. I would also thank my Institution and my faculty members without whom this project would have been distant reality. Finally, yet importantly I would like to express my thanks to my beloved parents for their blessings my friends/classmates for their help and wishes for the successful completion of this project.

AMAN SHARMA
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, West Bengal
June 27, 2020

Contents

1 PROJECT BACKGROUND **1**

2 INTRODUCTION **1**

3 PROJECT IDEA **2**

4 PROJECT DATASET **2**

5 PROJECT WORK **2**

5.1 Dataset Description . . . 2
5.2 Visualization Of Different Catagories . . . 3
5.2.1 BAR PLOTS FOR DIFFERENT FEATURES . . . 3
5.2.2 SCATTER PLOT FOR FEATURES . . . 7
5.2.3 CORRELATION PLOTS BETWEEN FEATURES-I . . . 7
5.2.4 DISTRIBUTION GRAPH . . . 9
5.2.5 BOX PLOT . . . 9
5.2.6 CORRELATION PLOTS BETWEEN FEATURES-II . . . 11
5.3 Terminologies Used . . . 11
5.3.1 SCATTER DIAGRAM . . . 11
5.3.2 PRINCIPAL COMPONENT ANALYSIS . . . 12
5.3.3 LOGISTIC REGRESSION . . . 12
5.3.4 KNEIGHBORS CLASSIFIER . . . 12
5.3.5 SUPPORT VECTOR CLASSIFIER . . . 13
5.3.6 DECISION TREE CLASSIFIER . . . 13
5.3.7 LOSS FUNCTION . . . 13
5.3.8 ACCURACY . . . 13
5.3.9 PRECISION . . . 13
5.3.10 RECALL . . . 13
5.3.11 F1 SCORE . . . 13
5.3.12 SPECIFICITY . . . 13

6 PROJECT RESULTS **14**

6.1 Results From PCA . . . 14
6.2 Results From Model Analysis . . . 15

7 CONCLUSION **17**

8 REFERENCES **17**

1 PROJECT BACKGROUND

Machine Learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience. Machine Learning focuses on the development of computer programs that can access data and use it to learn for themselves.
The recent significant increase in loan default has generated interest in understanding the key factors predicting the non-performance of these loans.However, despite the large size of the loan market, existing analyses have been limited by data. This project report, **Loan Prediction Using Machine Learning**,is based on model created to predict the factor responsible for payment or non-payment of loans.
Here we have **Loan Datatset** which consists different catagorical and numerical data. This report presents an analysis of data concerning loan delinquency. The analysis is based on observations, each containing specific characteristics of persons who are borrowing the loan.
After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between characteristics and loan default were identified. After exploring the data, **Principal Component Analysis(PCA)** has been done to identify the features that contributing the most in loan payment. Later, different models are created to predict the accuracy of loan giving. The efficiency of the model is increased by identifying the most valuable and contributing features among the given features in the data set.
Thus, this report basically includes the steps followed during the implementation of machine learning model for loan payment prediction and presents the different descriptive statistics and content rich visualization.

2 INTRODUCTION

Loan Prediction is a very important part in our today's life. More often than not, it seems to have been seen that people from different background,different social status,different education level apply to get a loan. But not everyone who applies is approved of loan.
We have applied **PCA** on Dataset and get the idea of features require to describe proeperly.
In this Project, it is worked on how different bakground,social status and education level can effect your loan application and depending on the result we can conclude. Here we have applied **Logistic Regression Model**, **KNeighbors Classifier Model**, **SVC** and **Decision Tree Classifier Model** on our **Loan Dataset** and check which model can predict better.

3 PROJECT IDEA

The idea behind this Proejct is to build models that will classify the loan prediction properly.It is based on the applicants's martial status, education, number of dependents, and employments. We can visualize the dataset and can get some information. We have also applied **PCA** to get the idea of how much components are required to explian the data properly. At last, we define models and calculate **Accuracy** to get the idea of fitted models.

4 PROJECT DATASET

The Dataset that I have used here for my Project is called **loan.csv**. The dataset has a shape of **(614,13)**. Here there are different coloumns indicating different features contributing in the Project.
The coloumn names are **Loan ID, Gender, Married, Dependents, Education, Self Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan Amount Term, Credit History, Property Area, Loan Status** that will appear respectivley in the dataset.

5 PROJECT WORK

In this Project,different kind of Visualizations are done as we have different catagories. Depending on the relationships between features we can also conclude various important aspects of the Train Dataset.

5.1 Dataset Description

Here a summary has been presented in a tabular form to describe the datatset.
The Dataset has both **Catagorical** and **Numerical** Data. We can only summarize the Numerical Data.

Table 1: DATASET DESCRIPTION FOR NUMERICAL FEATURES

CATAGORY	ApplicantIncome	CoapplicantIncome	LoanAmount	LoanAmountTerm	CreditHistory
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	403.459283	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
0.25	2877.500000	0.000000	100.000000	360.00000	1.000000
0.50	3812.500000	1188.500000	128.000000	360.00000	1.000000
0.75	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

count = No of observations in each catagory.
mean = Average no of observations in each catagory. On an Average what is the no denoting the numerical features.
sd = The Deviation from mean of each observation under each catagory. It deonets how much devaited the datatset features are.
0.25 = It is the 25th percentile of the data. How much observations lie in the 25th percentile of the dataset.
0.5 = It is the 50th percentile of the data. How many observations lie in the 50th percentile of the datatset. It is also called **MEDIAN**.
0.75 = It is the 75th percentile of the data. How many observations lie in the 75th percentile of the dataset.
max = Maximum of each Numerical features in the dataset.

5.2 Visualization Of Different Catagories

5.2.1 BAR PLOTS FOR DIFFERENT FEATURES

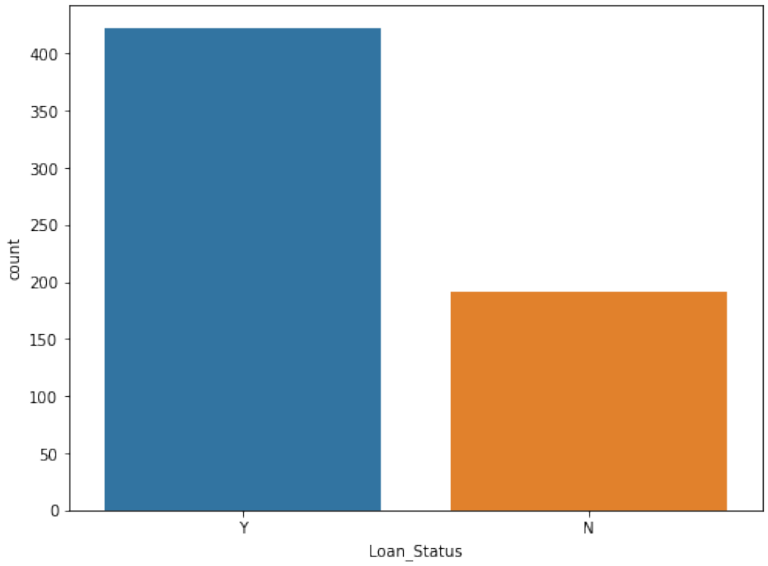

Figure 1: Percentage Count of Loan Status

The percentage of Y class : **0.69**
The percentage of N class : **0.31**

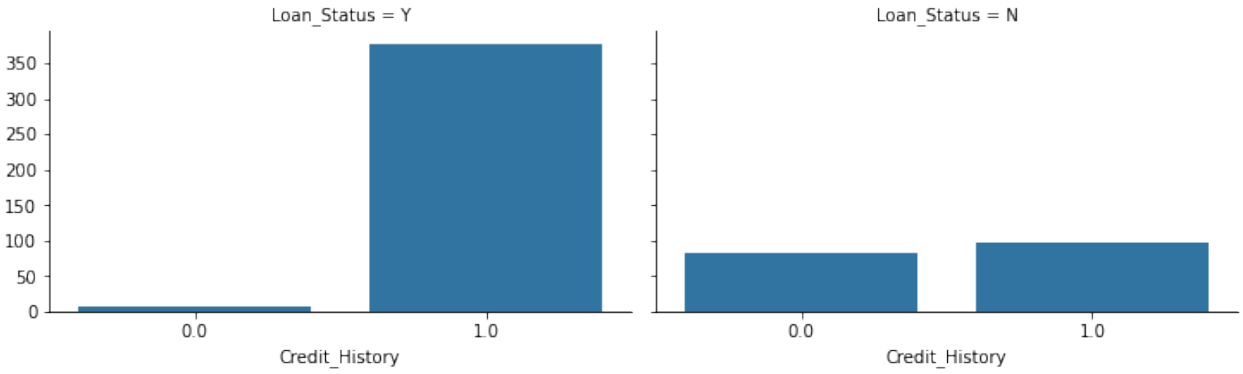

Figure 2: Loan Status Association With Credit History

So, we can say that for **Loan Status = Y**
we didn't give a loan for most people who got Credit History = 0
but **we did give a loan for most of people who got Credit History = 1**
So if you got Credit History = 1 , **you will have better chance to get a loan**.

For **Loan Status = N**, we can say that
it is **almost equal** for the cases, whatever your credit history maybe.
So it is not a good way to measure the association between **Credit Histrory and Loan Status**.

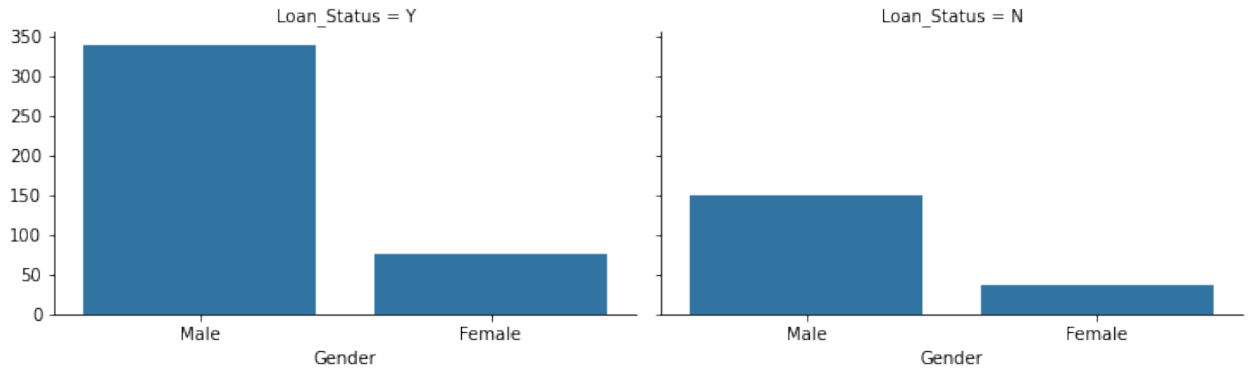

Figure 3: Loan Status Association With Gender

Here it is seen that,in **Loan Status= Y**
Most of the **Males** get loan as well as **Females**, though the no is less.
So regarding gthe loan Status, Gender is not a deciding factor.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever your Gender maybe.
So it is not a good way to measure the association between **Gender and Loan Status**.

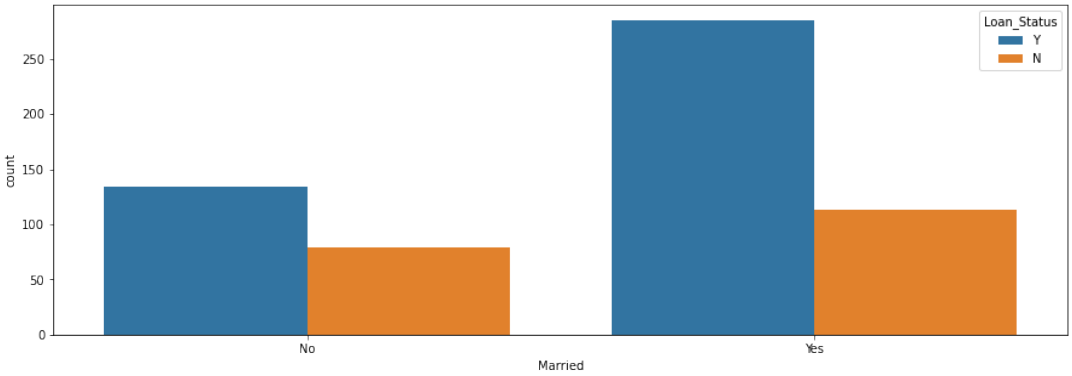

Figure 4: Loan Status Association With Marriage

Here it is seen that,in **Loan Status= Y**
Most of the person who are **Married** get loan as well as **Unmarried Person**, though the no is less.
So regarding gthe loan Status, married person has better chance in getting loan.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever your martial status maybe.
So it is not a good way to measure the association between **Loan Status and Martial Status**.

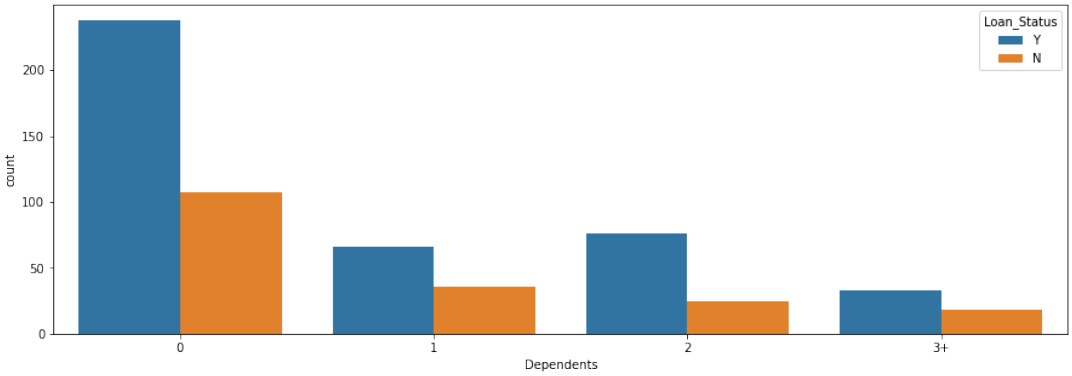

Figure 5: Loan Status Association With Dependents

Here it is seen that,in **Loan Status= Y**
Those who have **0 Dependents** have most chance of getting loan.
It decreases **gradually** regarding the dependent.
So, accordingly **Dependents with 1,2 and 3+** also have chance of getting loan but of less no.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever Dependents maybe.
For **Dependents 0** it is also most and **decreasing gradually**.

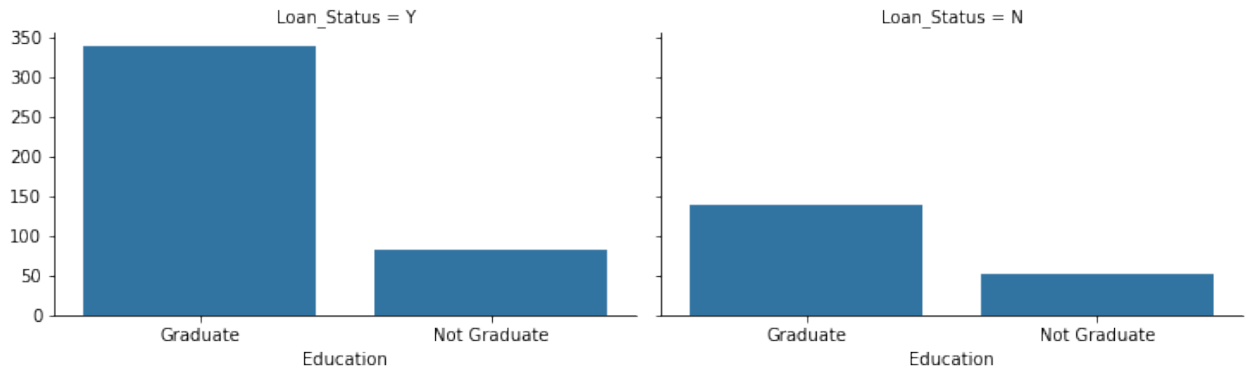

Figure 6: Loan Status Association With Education

Here it is seen that,in **Loan Status= Y**
Most of the person who are **Graduated** get loan as well as **Ungraduated**, though the no is less.
So regarding gthe loan Status, Education is not a deciding factor.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever your Eduaction status maybe.
So it is not a good way to measure the association between **Loan Status and Education Status**.

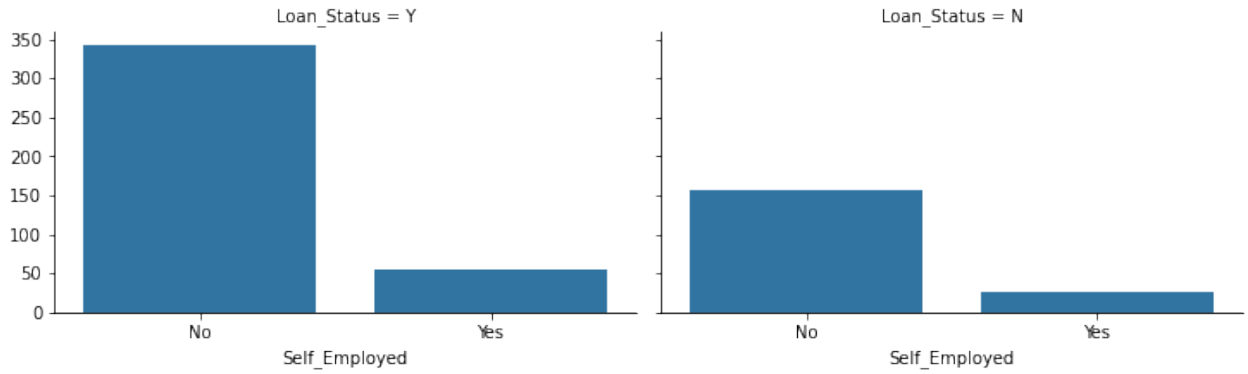

Figure 7: Loan Status Association With Self-Employment

Here it is seen that,in **Loan Status= Y**
Most of the person who are **Not Self-Employed** get loan as well as **Self-Employed**, though the no is less.
So regarding the loan Status, Self- Employment is not a deciding factor.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever your Self-Employment status maybe.
So it is not a good way to measure the association between **Loan Status and Self-Employment Status**.

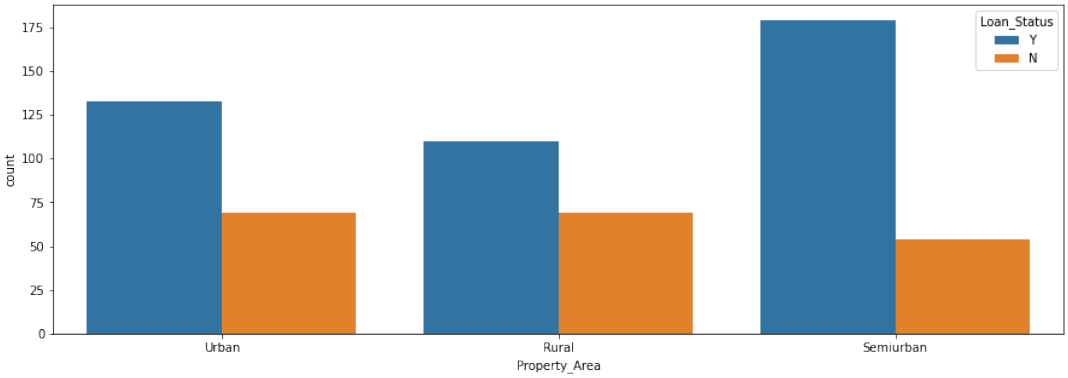

Figure 8: Loan Status Association With Property Area

Here it is seen that,in **Loan Status= Y**
Most of the person who are **From SemiUrban Area** get loan as well as **From Urban And Rural Area**.
So regarding the loan Status, SemiUrban Area People has got more than 0.50 chance to get a loan.

For **Loan Status = N**, it is clear that
it is **almost same** for the cases, whatever your Property Area maybe.
So it is not a good way to measure the association between **Loan Status and Property Area**.

5.2.2 SCATTER PLOT FOR FEATURES

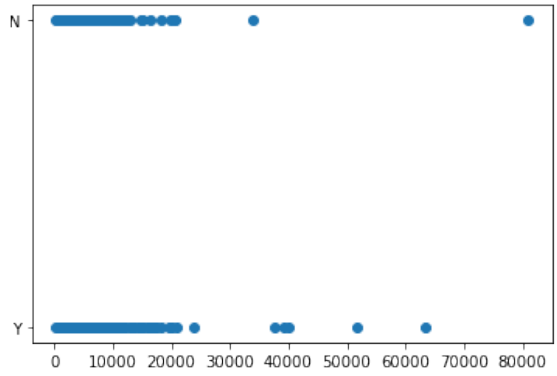

Figure 9: Loan Status Association With ApplicantIncome

So from the **Scatter Plot**, it is clear that there is **No Relationship** between Loan Status With ApplicantIncome.

5.2.3 CORRELATION PLOTS BETWEEN FEATURES-I

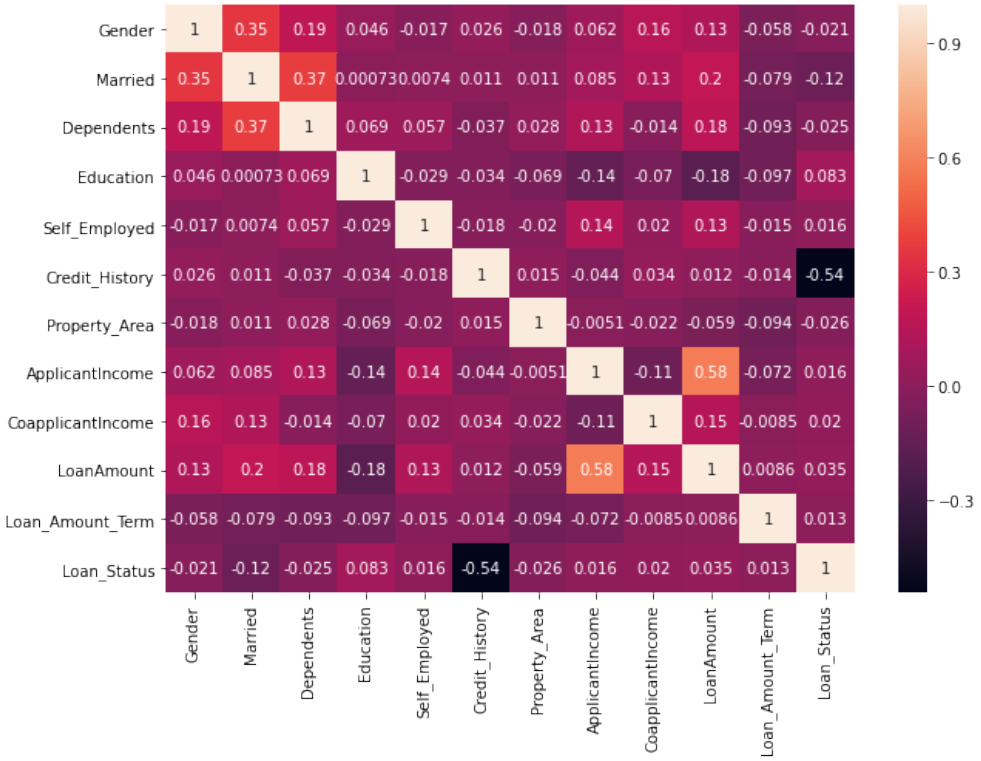

Figure 10: Correlation Plot

So we can see the relationship between each and every features from **Correlation Plot**.
We can conclude that **Credit Histroy and Married Status are good features, actually Credit**

Histroy is the best .
Here we got 0.58 similarity between **LoanAmount and ApplicantIncome**.

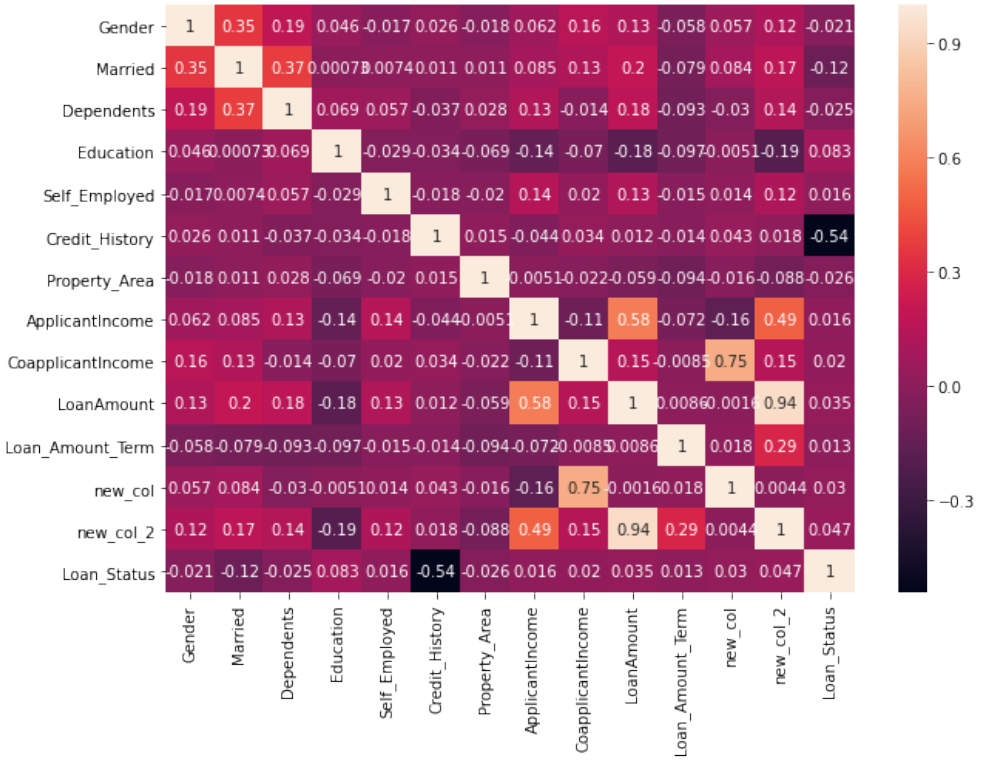

Figure 11: Modified Correlation Plot

Here, I have two new coloumns:
new col = CoapplicantIncome/ ApplicantIncome.
new col2 = LoanAmount *Loan Amount Term.

5.2.4 DISTRIBUTION GRAPH

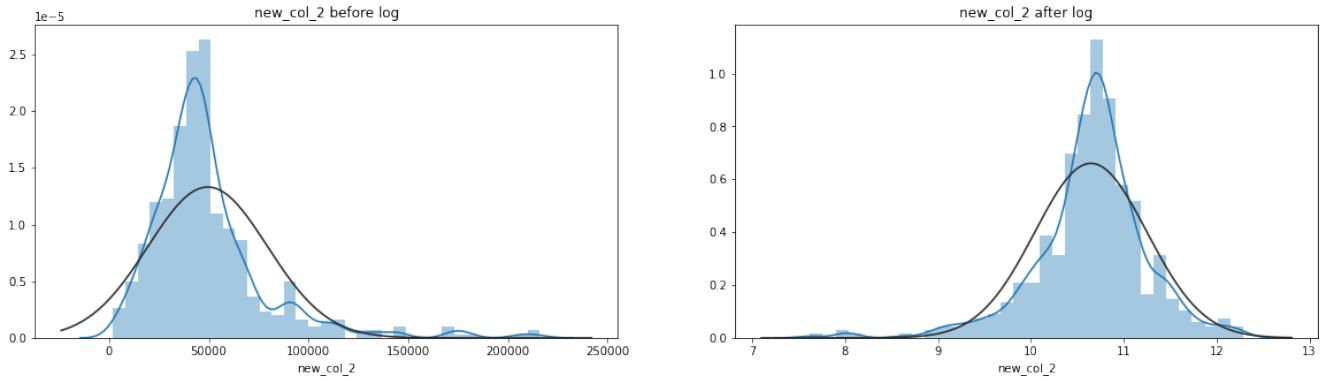

Figure 12: Distribution Plot

For new col2 we can see we got **Right Skewed Distribution**.
We can solve this problem with very simple statistical technique, by taking the logarithm of all the values because the data is **Normally Distributed** that will help improving our model.

5.2.5 BOX PLOT

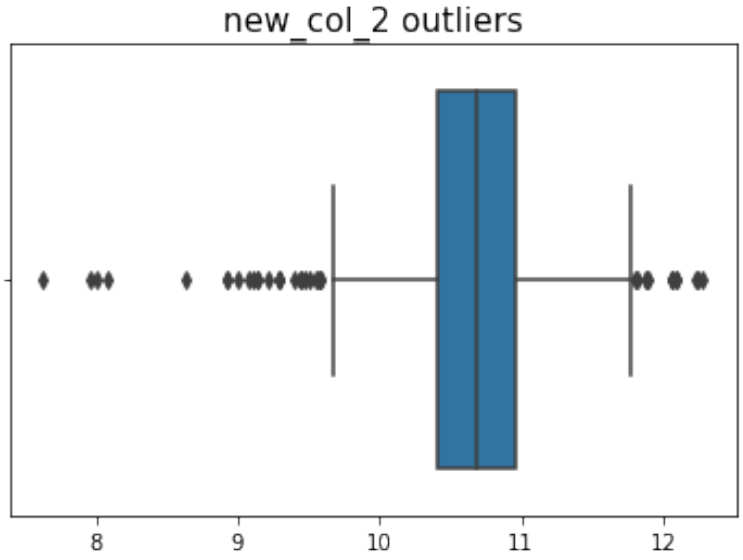

Figure 13: Box Plot WIth Outliers

RESULTS:
Quartile 25: 10.416008100285975 , **Quartile 75**: 10.961277846683982
iqr: 0.5452697463980076
Cut Off: 0.05452697463980077
Lower: 10.361481125646174
Upper: 11.015804821323783

Nubers of Outliers: 218

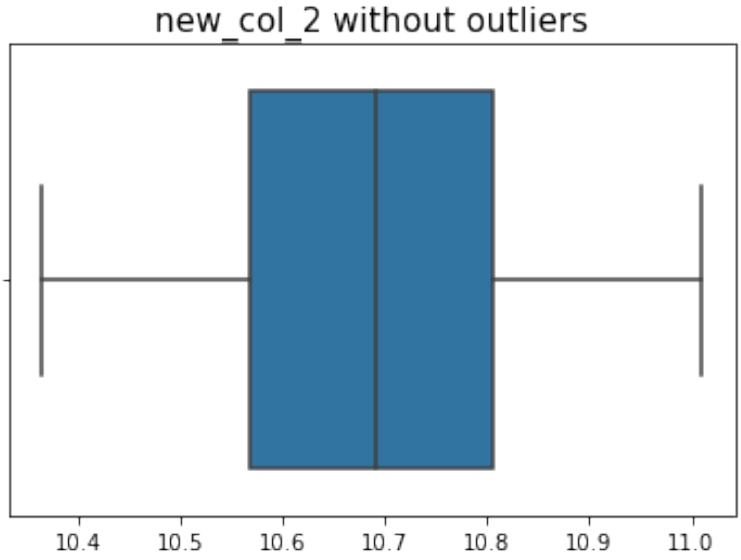

Figure 14: Box Plot WIthout Outliers

5.2.6 CORRELATION PLOTS BETWEEN FEATURES-II

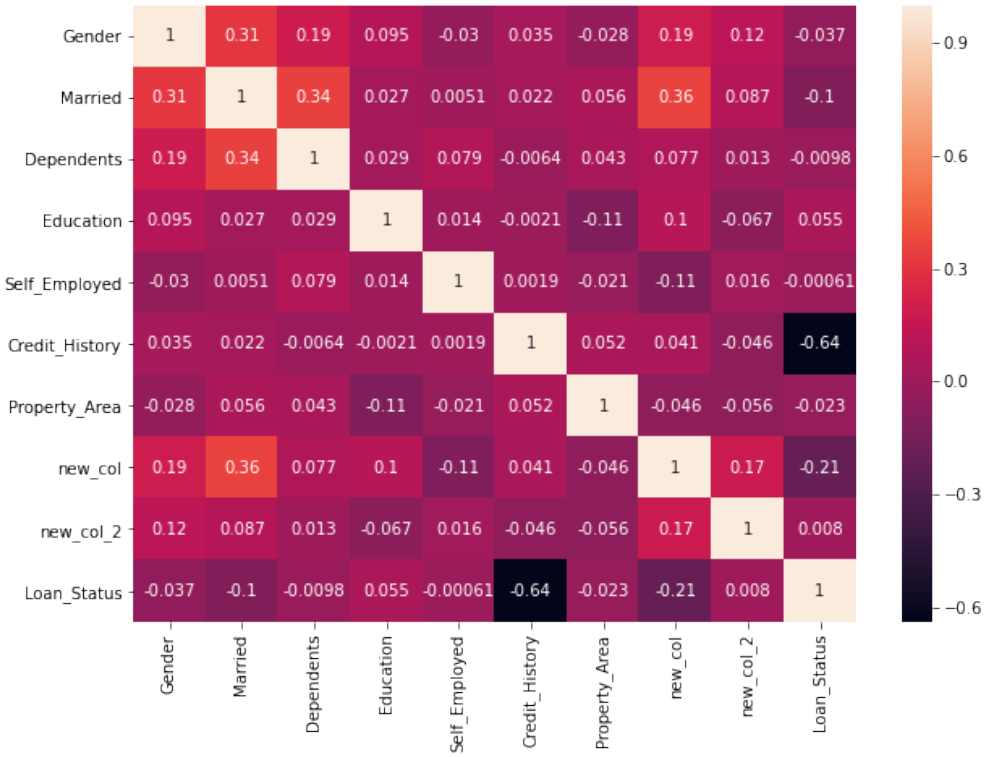

Figure 15: Modified Correlation Plot

We have seen that Self Employed got really **bad corr (-0.00061)**.

5.3 Terminologies Used

5.3.1 SCATTER DIAGRAM

A **Scatter Plot** is a type of plot or Mathematical Diagram using Cartesian Coordinates to display values for typically two varaibles for a set of dataset.
It Generally lies between **+1 to -1**. Where the Extreme values shows perfect correation, **Positive Or Negative** may be. Value close to 0 shows **Zero Correlation**.

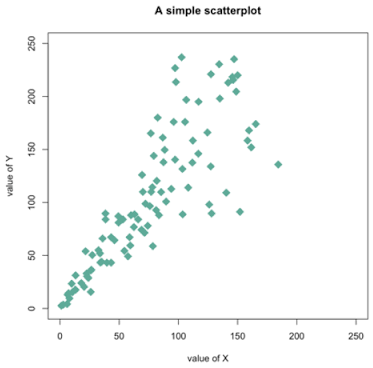

Figure 16: Scatter Plot

5.3.2 PRINCIPAL COMPONENT ANALYSIS

An important **Machine Learning** method for dimensionality reduction is called **Principal Component Analysis(PCA)**.
It is a method that uses simple matrix operations from Linear Algebra and Statistics to calculate a projection of the **Original data into the same number or Fewer Number**.

5.3.3 LOGISTIC REGRESSION

In Statistics, the **Logistic Model** is used to model the Probability of a certain class or even existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat or a dog or classifing the classes properly.

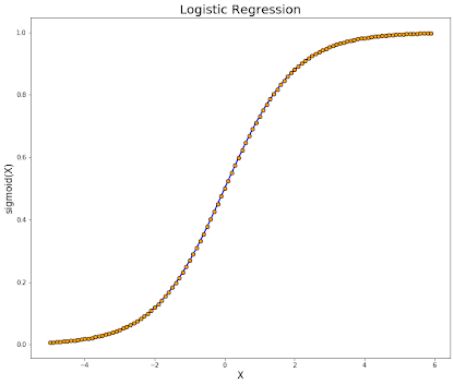

Figure 17: Logistic Regression

5.3.4 KNEIGHBORS CLASSIFIER

The **K-Nearest Neighbors** (KNN) algorithm is a type of supervised machine learning algorithms. KNN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. **KNN works** by finding the distances between a query and all the examples in the data,selecting the specified number examples (K) closest to the query.

5.3.5 SUPPORT VECTOR CLASSIFIER

The objective of a **Linear SVC** is to fit the data provided,returning the **Best Fit hyperplane** that divides or catagorizes data.

5.3.6 DECISION TREE CLASSIFIER

Decision Tree Learning is one of the predictive modelling approaches used in **Statistics, Data Mining and Machine Learning**. It uses a decsion tree to go from observations about an item to conclusions about the item's target value.

5.3.7 LOSS FUNCTION

Machine learns by **means of a Loss Function**. It is a method of evaluating how well specific algorithm models the given data.If predictions deviates too much from actual results, loss function would cough up a very large number.
The Lower the Loss Function, the Better the Predicted Model.

5.3.8 ACCURACY

Measurement of being correct . **Accuracy** gives the measure of how good a model is in terms of performance. It is the quality or state of being correct. It refers to the closeness of a measured value to a standard of known value.

5.3.9 PRECISION

Measurement of consistency,minimizing **false positives** .

5.3.10 RECALL

Measurement of completeness, also known as **" TRUE POSITIVE RATE/SENSITIVITY"**, least **false negatives**.

5.3.11 F1 SCORE

Indicates the balance between Precision and Recall.

5.3.12 SPECIFICITY

Proportion of true negatives that are correctly predicted, also known as**" TRUE NEGATIVE RATE"**.

6 PROJECT RESULTS

The Results that I have got after performing the experiment in **Python** is depicted here below. With the help of these results, it can be said that which classifier model performs best.

6.1 Results From PCA

Filtering the Dataset, I have done the PCA analysis and got the following Result :

If n components is not set, all components are kept (11 in this case).

Total Explained Variance : 11.025700934579454
Cumulative Explained Variance : array([17.3988806 , 30.80902058, 41.63884139, 51.26280458, 60.51732821, 69.14419256, 77.04206011, 84.58157718, 91.27831426, 96.47690067, 100.])
Components Required To Explaine Variance : (cum var exp 0.99) = 11, (cum var exp 0.95) = 10, (cum var exp 0.90) = 9, (cum var exp 0.85) =9

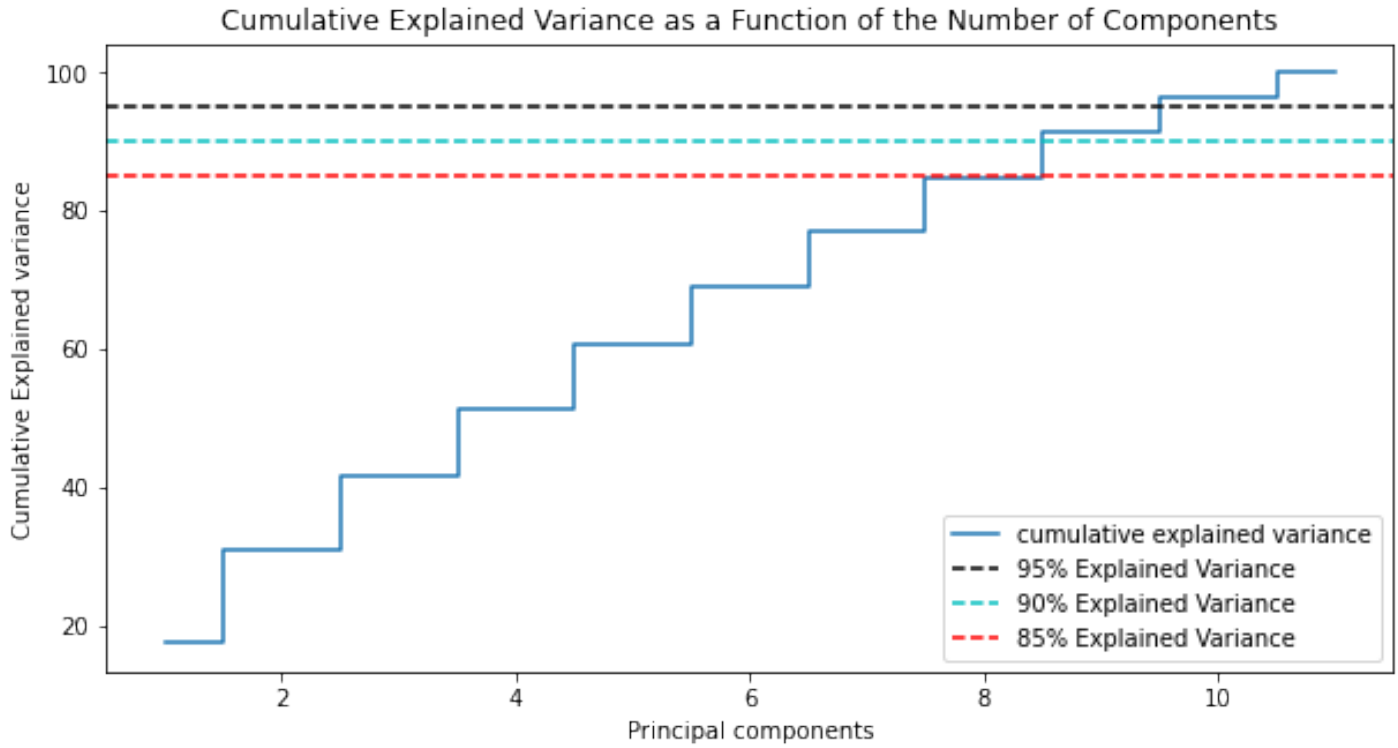

Figure 18: PCA ANALYSIS

6.2 Results From Model Analysis

At the very beginnig performing different models on dataset, the result that we got is depicted below :

LogisticRegression :
pre: 0.930
rec: 0.429
f1: 0.587
loss: 6.542
acc: 0.811
KNeighborsClassifier :
pre: 0.667
rec: 0.364
f1: 0.471
loss: 8.863
acc: 0.743
SVC :
pre: 1.000
rec: 0.013
f1: 0.026
loss: 10.692
acc: 0.690
DecisionTreeClassifier :
pre: 0.929
rec: 0.422
f1: 0.580
loss: 6.612
acc: 0.809

We can see that best model is **Logistic Regression** at least for now, **SVC** is just memorizing the data so it is overfitting .

After some modification we got the result below :

LogisticRegression :
pre 0.894048
rec 0.422500
f1 0.562733
loss 6.752695
acc 0.804490
KNeighborsClassifier :
pre 0.379834
rec 0.207500
f1 0.259954
loss 12.381887
acc 0.641510
SVC :
pre 0.000000
rec 0.000000
f1 0.000000
loss 11.043955
acc 0.680245

DecisionTreeClassifier :
pre 0.919048
rec 0.422500
f1 0.565740
loss 6.611717
acc 0.808571

As it can be seen **SVC** is just memorizing the data, and one can see that here **Decision Tree Classifier is better than Logistic Regression**.

Now we look at the value counts of every label :

For **Gender :**
1 398
0 93
For **Married :**
1 315
0 176
For **Dependents :**
0 292
2 85
1 78
3 36
For **Education :**
0 382
1 109
For **Self Employed :**
0 428
1 63
For **Credit History :**
1 421
0 70
For **Property Area :**
1 179
2 170
0 142
For **new col :**
0.000000 222
0.414374 1
0.912892 1
1.258120 1
0.504299 1
...
0.330420 1
2.332134 1
0.844471 1
0.564642 1
0.824769 1
For **new col2 :**
43200.0 18
39600.0 13

36000.0 11
57600.0 11
46080.0 9
..
3000.0 1
12000.0 1
9072.0 1
25920.0 1
7560.0 1

7 CONCLUSION

In this project, we explore different aspects of incorporating neural, statistical and external features to deep neural networks on the task of **loan prediction**. We also presented in-depth analysis of several state-of-the-art recurrent and convolution architectures. The presented idea leverages features extracted using difderent models.
From the used models, it is seen that **Logistic Regression** performs the best and **Decision Tree Classifier** also performs the same level as the Previous one. But Among all **SVC** performs the worst, it generally overfitts the data and gives very bad result.
Though there is still scope for further improvement in the topic and hope to be explored in future.

8 REFERENCES

1. **Google** 'www.google.com'
2. **Machine Learning Wikipedia** 'en.wikipedia.org/wiki/Machine learning'
3. **NLP for Machine Learning** 'towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d
4. **Loan Prediction as Python Project** 'data-flair.training/blogs/advanced-python-project-detecting-loan-prediction'
5. **Building Machine Learning Systems with Python** by Willi Richert,Luis Pedro Coelho (PACKT PUBLISHING;BIRMINGHAM - MUMBAI)