STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Predicting Approval Status of Loan Applicants

Guided By:
Dr. Amir H. Gandomi

Presented By:
Akshay Kirolikar
Ankur Morbale
Gaurav Venkatraman

Date: 11/29/2018

Introduction

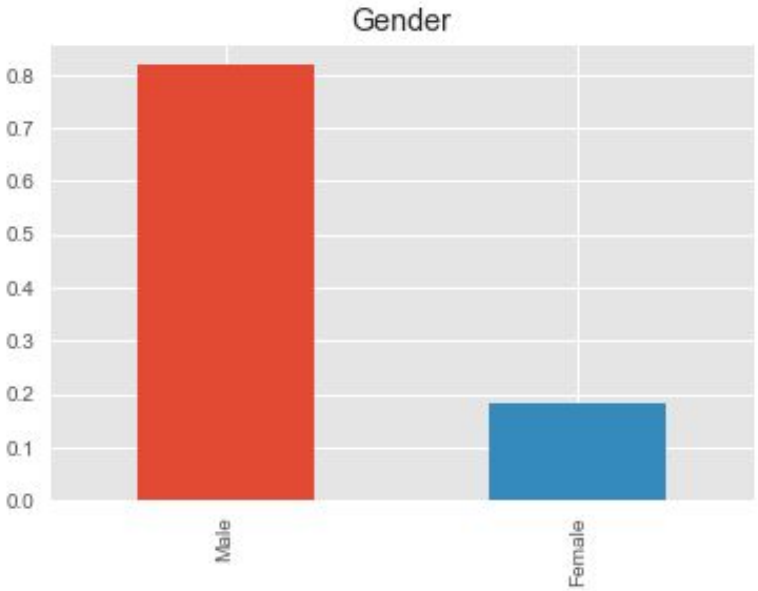
- Data Source : analyticsvidhya.com

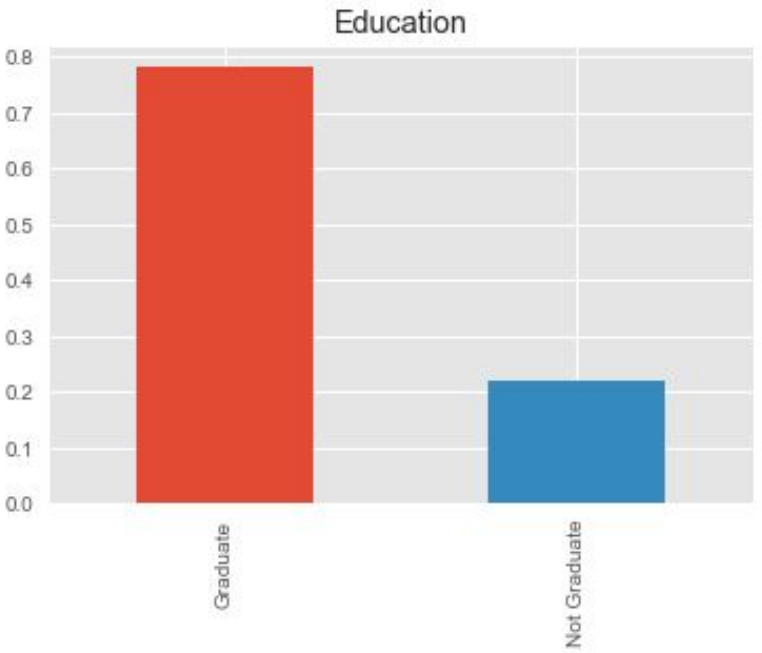
- Understand and clean the dataset
- Identifying significant independent variables
- Principal Component Analysis
- Prediction using classification algorithms
- Conclusion from our learnings
- Tools: Python(pandas, sklearn, matplotlib)

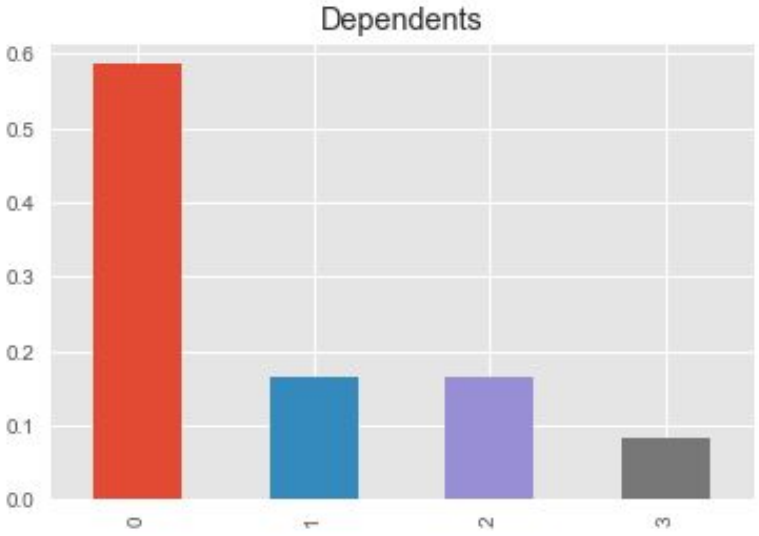
About the Dataset

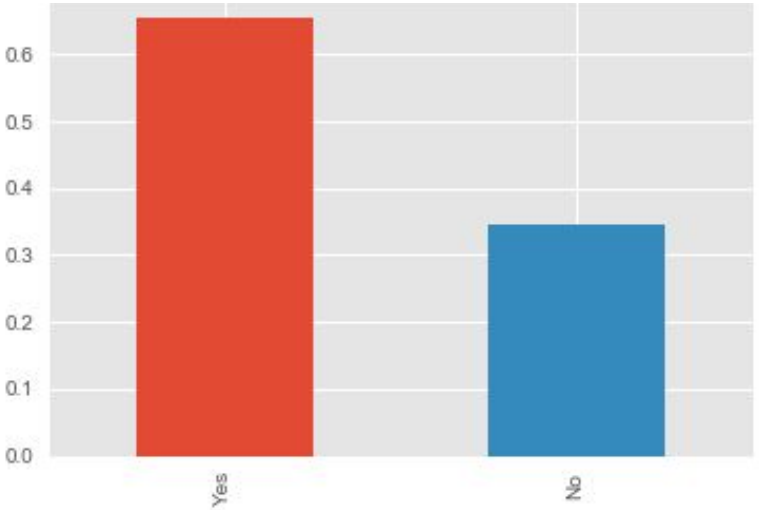
- Dataset : Loan Prediction
- **614** Observations based on **13** attributes
- Target variable **Loan Status** (Yes/No)
- Goal: To predict Loan status(Yes/No) based on Loan Id, Gender, Marital Status, Dependents, Highest qualification, Self-employed or not, income of applicant and the co applicant, loan amount, loan term, credit history, the property area.

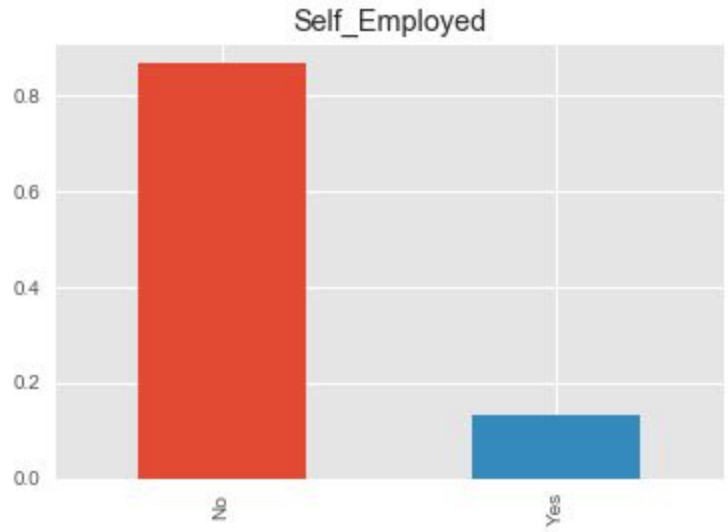
Data Visualisations

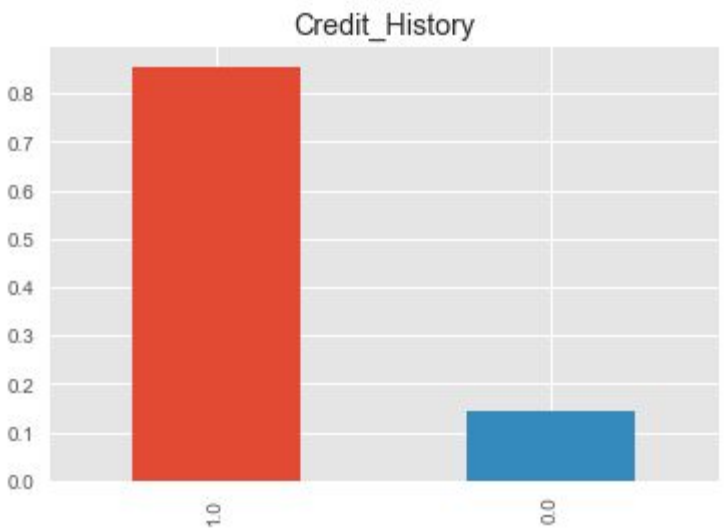


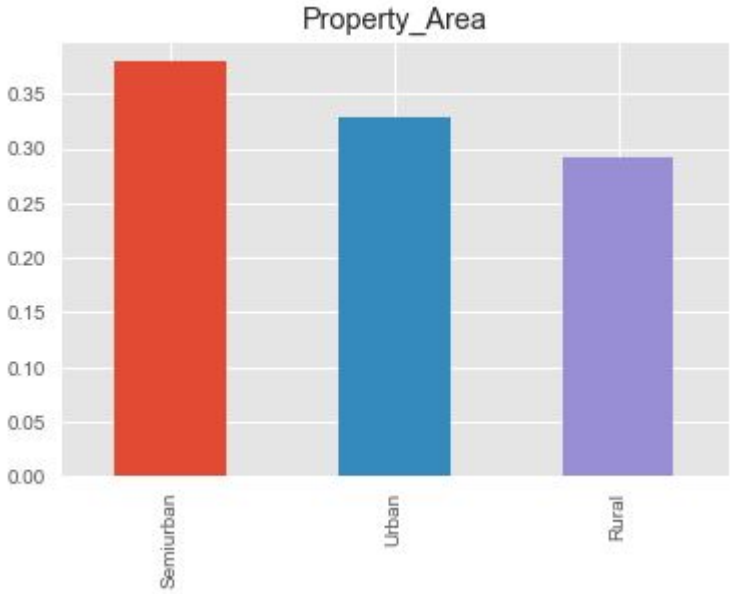







Self_Employed
0.8
0.6
0.4
0.2
0.0
No
Yes


Credit_History
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1.0
0.0


Property_Area
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Semiurban
Urban
Rural


Loan_Status
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Y
N

Data Preprocessing

Missing values

Mode:

- Dependents
- Credit history
- Gender
- Married
- Education
- Self Employed
- Loan Amount Term

Median:

- We subsetted Education based on the 2 classes- 'Graduate and non graduate' and replaced corresponding missing values of Loan Amount using Median of graduate and non graduate classes.

Data Preprocessing

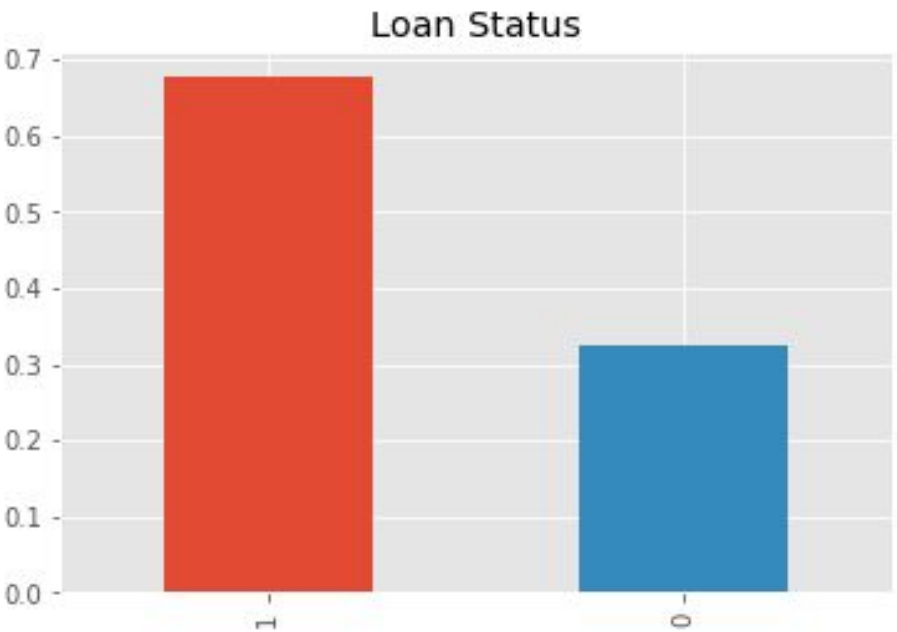
- We combined the Co applicant Income and Applicant Income to form a new column as Total Income
- Categorical variables transformed to dummy variables:

Property Area, Dependents, Gender, Married, Education, Self Employed, Credit History, Loan Status

- Loan Amount, Total Income, Loan Amount Term variables were standardized
- Dataset split into training and test set in 80:20 ratio

Over-Sampling

We performed (SMOTE)Synthetic Minority Over-sampling Technique to balance the data.

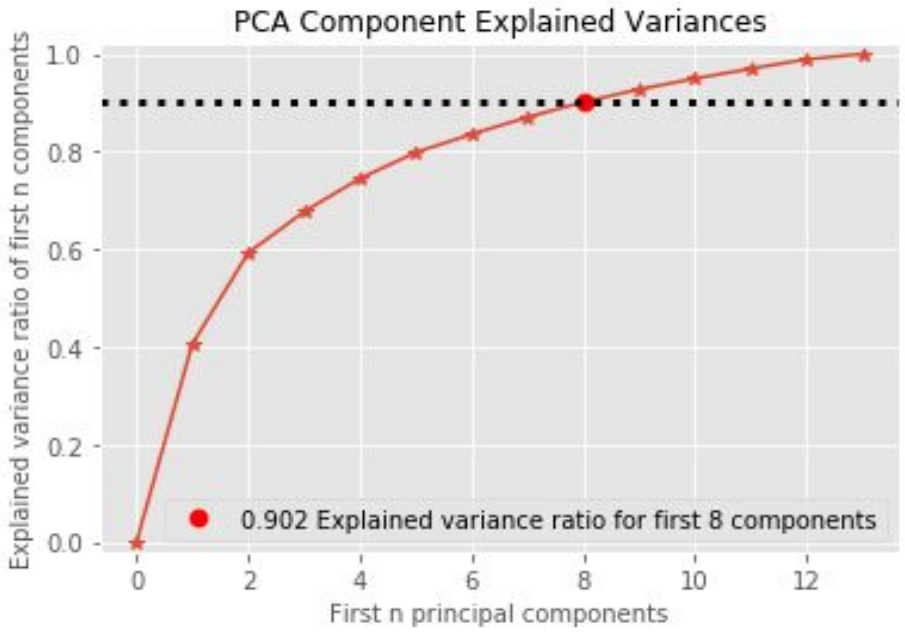


Value Counts:
1 => 332
0 => 159

Value Counts:
1 => 332
0 => 332

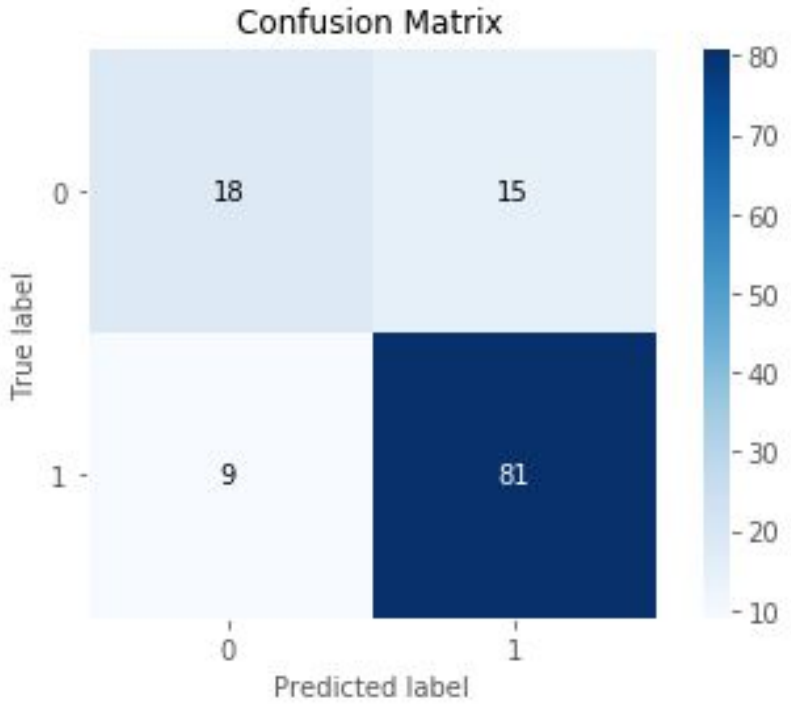
Principal Component Analysis

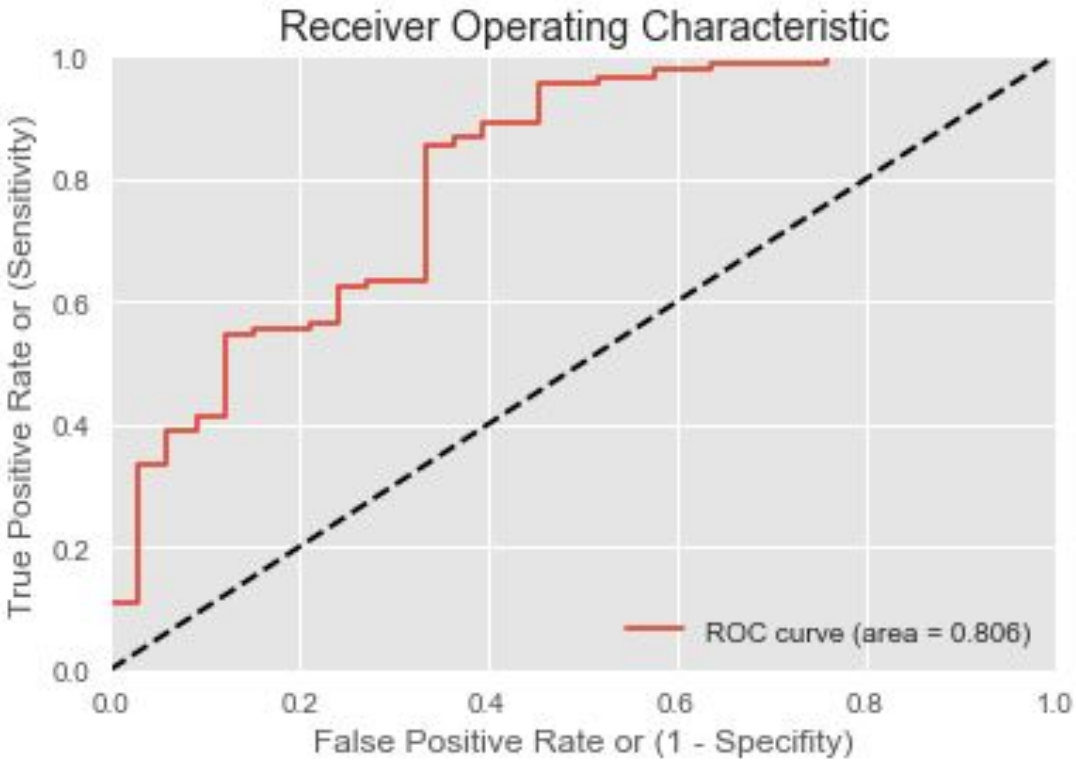
- We had total of 13 variables after addition of dummy variables
- Applied PCA and selected 8 PCs explaining 90.02% of variance



Logistic Regression

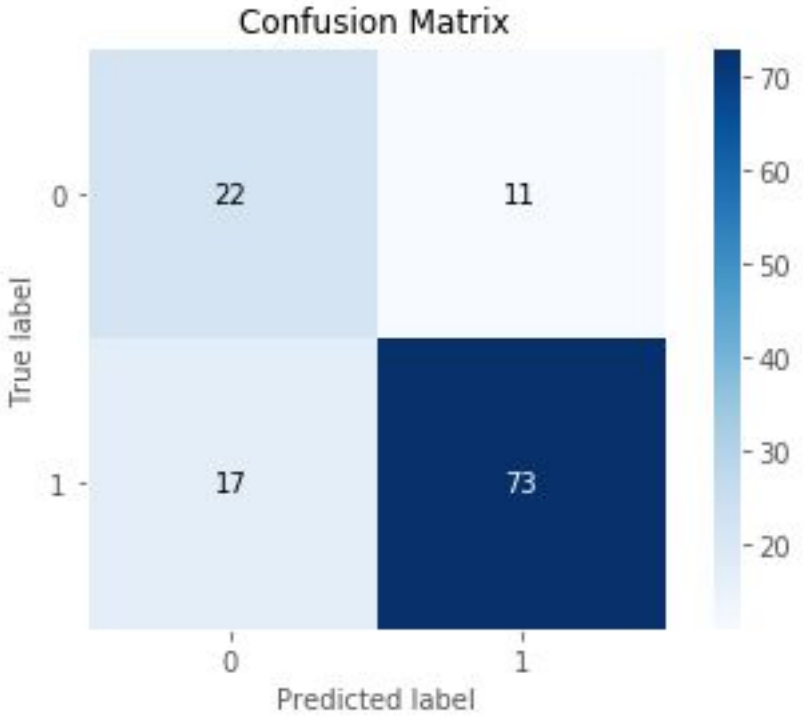
- The accuracy percent achieved is 80.48%
- The confusion matrix is as follows:

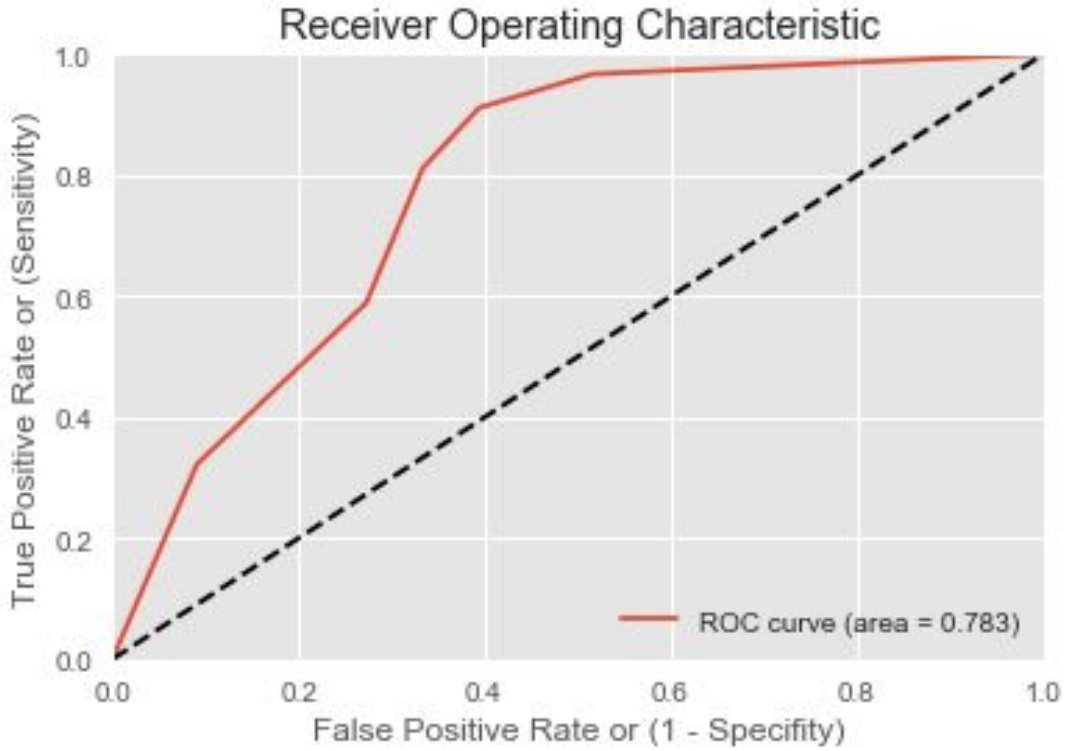




K-Nearest Neighbors

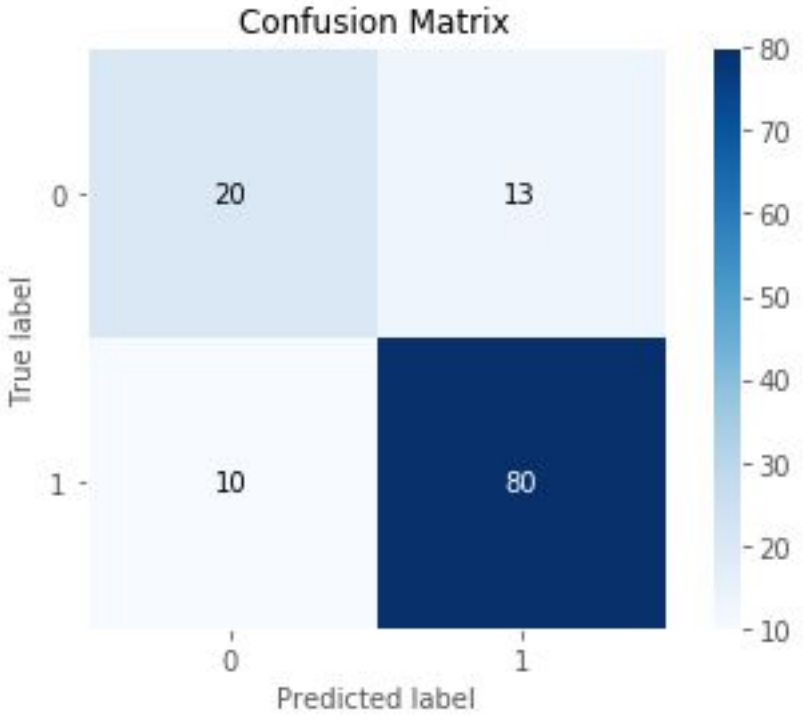
- The accuracy percent achieved is 77.23%
- The confusion matrix is as follows: K=5

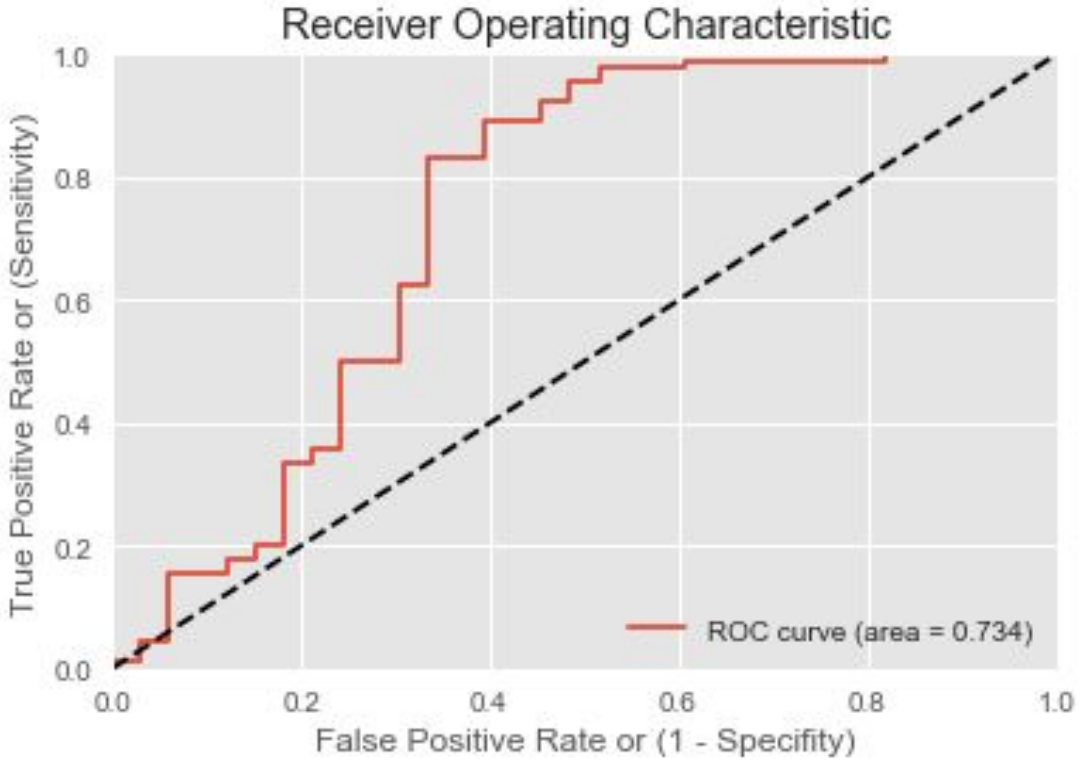




Naïve Bayes

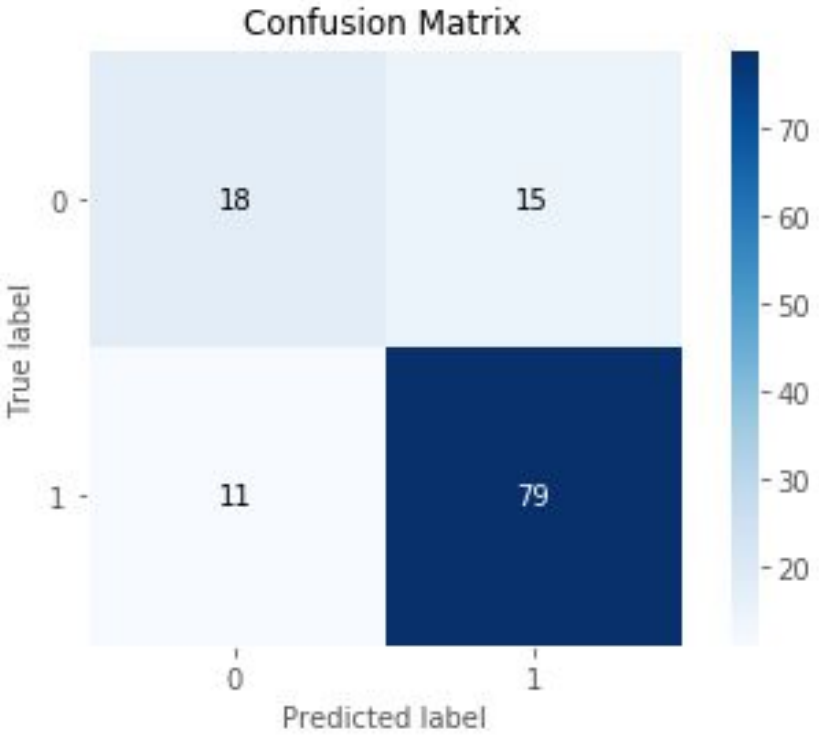
- The accuracy percent achieved is 81.30%
- The confusion matrix is as follows:

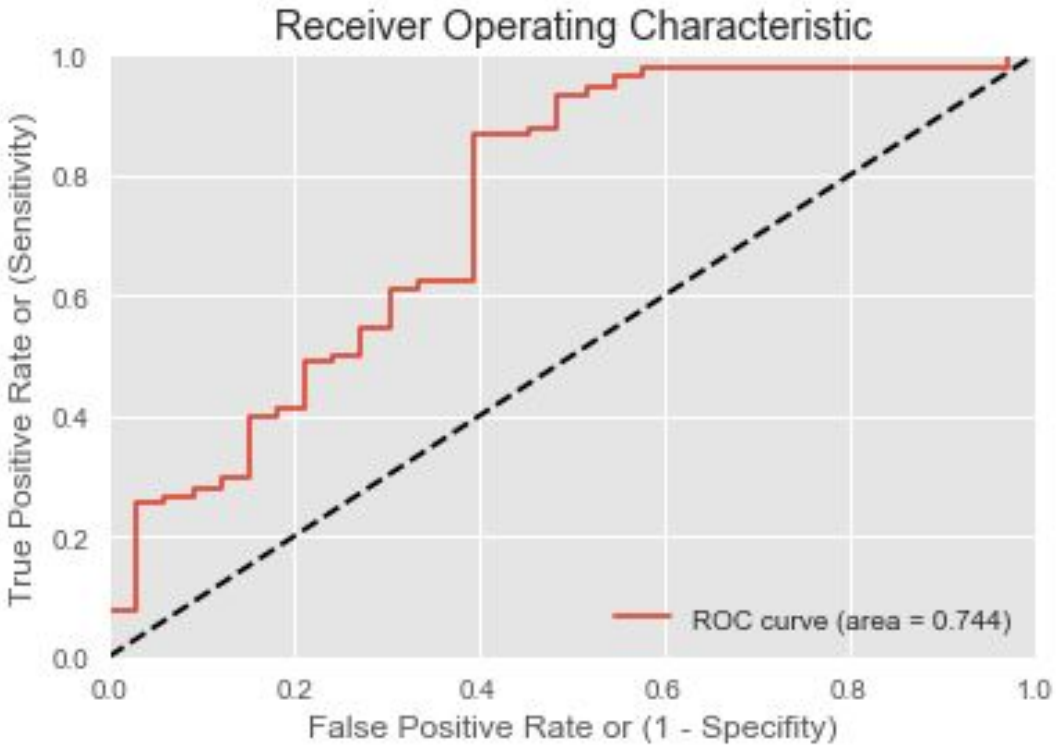




Support Vector Machine

- The accuracy percent achieved is 78.86%
- The confusion matrix is as follows:





Conclusion

1) While performing Logistic Regression, false positives(FP) are 15, in naive bayes, FPs are 13 and in kNNs, FPs are 11. From this, we can infer that although naive bayes displays higher accuracy, kNNs display a lower FP rate and our primary goal is to have reduced FP rates so as to prevent the number of undeserved loans being sanctioned and hence from the banks perspective, kNNs proves to be the optimum solution here.

2) The accuracy showcased by all the model lie in the similar bracket, i.e., 77-82%. According to our problem statement, our key goal is to reduced the FP rates so as to reduce the risk incurred by the bank.

3)This model can help the decision makers to reinforce their decision making through data driven insights.

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu