globsyn

Taking People To The Next Level

globsyn

finishing school

PREDICT CREDIT CARD ACCEPTANCE

Group Members:

Devanjan Chatterjee, DR. SUDHIR CHANDRA SUR DEGREE ENGINEERING COLLEGE, 162550110087

Anutirtha Saha, DR. SUDHIR CHANDRA SUR DEGREE ENGINEERING COLLEGE, 162550110014

Bishal Bhowmik, B.P. PODDAR INSTITUTE OF MANAGEMENT AND TECHNOLOGY, 161150110037

Anshuman Swaroop Das, B.P. PODDAR INSTITUTE OF MANAGEMENT AND TECHNOLOGY, 161150110024

Abhisek Das, B.P. PODDAR INSTITUTE OF MANAGEMENT AND TECHNOLOGY, 161150110003

Table of Contents

- Acknowledgement
- Overview
- Project Objective
- Project Scope
- Project Requirements
- Data Description
- Model Building
- Screenshots
- Future Scope of Improvements
- Conclusion

Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my faculty (Prof. Arnab Chakraborty) for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by him/her time to time shall carry me a long way in the journey of life on which I am about to embark.

I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

(Devanjan Chatterjee, Anutirtha Saha,
Bishal Bhowmik, Abhisek Das,
Anshuman Swaroop Das)

OVERVIEW

Commercial banks receive a lot of applications for credit cards. Many of them get rejected for many reasons, like high loan balances, low income levels, or too many inquiries on an individual's credit report, for example. Manually analyzing these applications is mundane, error-prone, and time-consuming (and time is money). Luckily, this task can be automated with the power of machine learning and pretty much every commercial bank does so nowadays. In this project, we will build an automatic credit card approval predictor using machine learning techniques, just like the real banks do.

Project Objective

In this project we have a small credit card dataset for simple econometric analysis (taken from Kaggle, originally from William Greene's book Econometric Analysis).

Goal: To predict whether a credit card application will be accepted based upon various data about the applicant. For this goal, a multivariable regression analysis will be undertaken including linear regression. We plan to solve this by splitting the dataset into train and test data and create 4 different types of models from the data: Decision Tree, Linear Regression, Naive Bayes and K-NN and also do the performance evaluation of each model.

The steps are as follows:

- First, we will start off by loading and viewing the dataset.
- We will see that the dataset has a mixture of both numerical and non-numerical features, that it contains values from different ranges, plus that it contains a number of missing entries.
- We will have to preprocess the dataset to ensure the machine learning model we choose can make good predictions.
- After our data is in good shape, we will do some exploratory data analysis to build our intuitions.
- Finally, we will build a machine learning model that can predict if an individual's application for a credit card will be accepted.

Project Scope

The broad scope of the Prediction of Credit Card Acceptance project includes:

- In this project we analyzed a dataset of personal and bank account data of a certain customer. The dataset consists of various other factors that influence card availability.

- By this project we can predict whether a credit card application will be accepted or not.

- By analyzing the data, we will build a predictor model by using some well-known pre-processing methods such as imputing missing values, label encoding, scaling the columns values and finally applying the model on training data set and evaluating the model with testing data set.

- We will also see a few additional approaches on how to improve the model performance.

Project Requirements

Hardware requirements:

□ CPU: Dual core 64-bit 2.8 GHz 8.00 GT/s CPUs
□ RAM: 2 GB RAM (recommended 4 GB RAM)
□ Storage: 2 GB for installation of Anaconda Navigator.
□ Internet access to download the files from Anaconda Cloud or a USB drive containing all of the files you need with alternate instructions for air gapped installations.

Software requirements:

□ Anaconda Navigator v3.6.4 or higher
□ SciKit learn package for Python
□ Any web browser like Google Chrome

Dependencies:

□ sklearn
□ numpy
□ pandas

Data Description

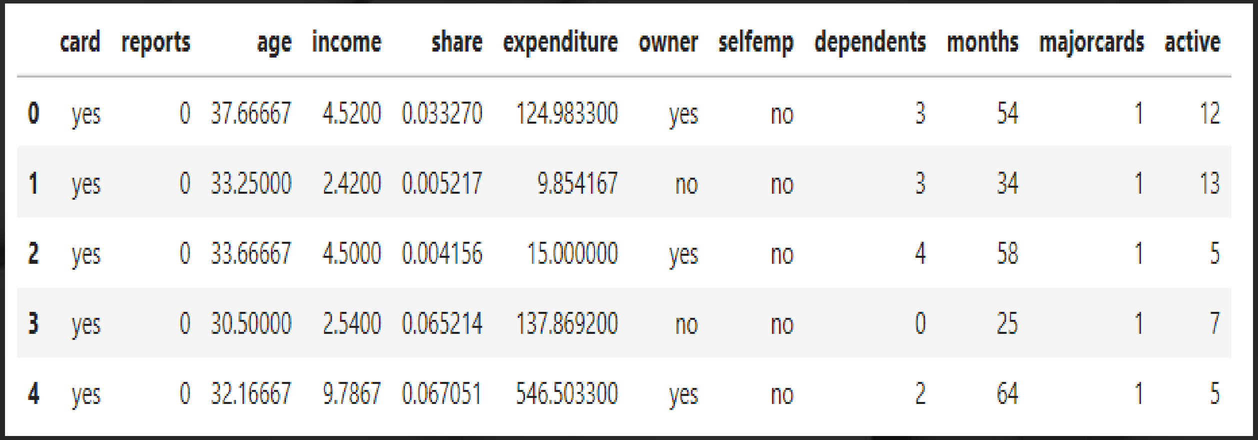
This dataset is taken from Kaggle, originally from William Greene's book Econometric Analysis.
The present exercise will study the Credit Card Account Data. This is a data frame with **1319 observations** on the following variables:

Target Variable:

Card: Dummy variable, 1 if application for credit card is accepted, 0 if not accepted.

The predictors influencing the card are:

- **reports:** Number of major derogatory reports
- **age:** Age n years plus twelfths of a year
- **income:** Yearly income (divided by 10,000)
- **share:** Ratio of monthly credit card expenditure to yearly income
- **expenditure:** Average monthly credit card expenditure
- **owner:** 1 if owns their home, 0 if rent
- **selfempl:** 1 if self employed, 0 if not.
- **dependents:** 1 + number of dependents
- **months:** Months living at current address
- **majorcards:** Number of major credit cards held
- **active:** Number of active credit accounts



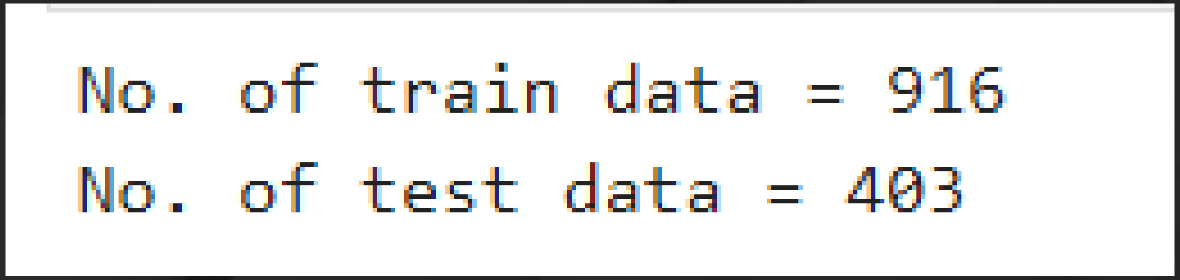
	card	reports	age	income	share	expenditure	owner	selfemp	dependents	months	majorcards	active
0	yes	0	37.66667	4.5200	0.033270	124.983300	yes	no	3	54	1	12
1	yes	0	33.25000	2.4200	0.005217	9.854167	no	no	3	34	1	13
2	yes	0	33.66667	4.5000	0.004156	15.000000	yes	no	4	58	1	5
3	yes	0	30.50000	2.5400	0.065214	137.869200	no	no	0	25	1	7
4	yes	0	32.16667	9.7867	0.067051	546.503300	yes	no	2	64	1	5

❖ First five columns of the data set

Data Description

Train and Test Data –

Dividing the whole data set into Train and Test Data randomly. While we will “train” our models with the help of the train data set, we will test the accuracy of their prediction with the help of the “test data set.

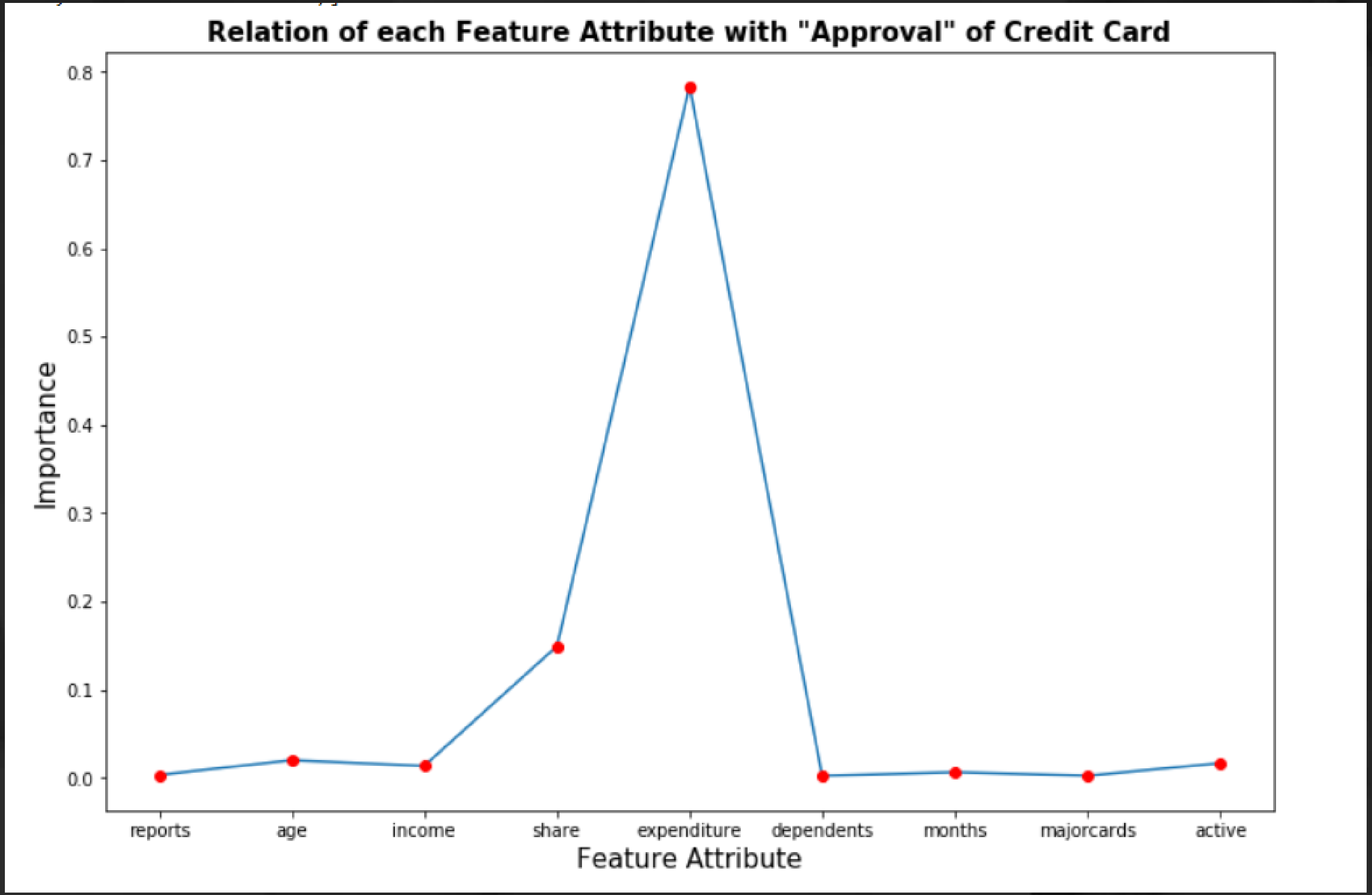


```
No. of train data = 916
No. of test data = 403
```

Data Description

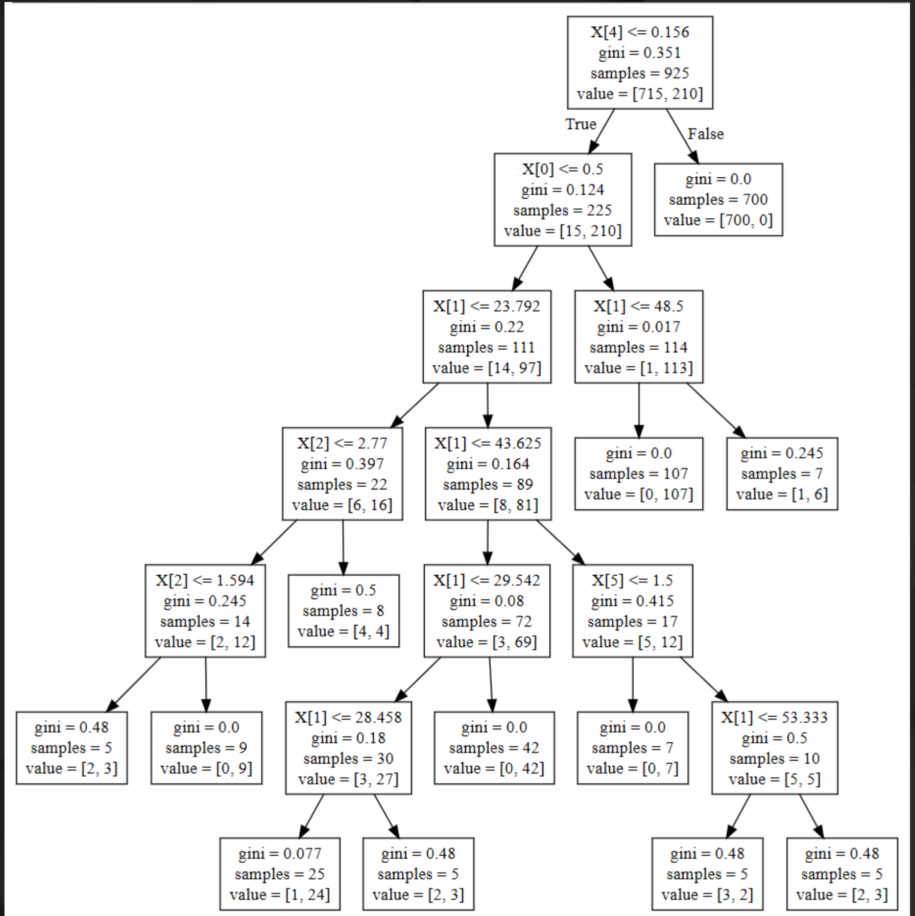
Determining Principal Attributes –

With the help of this line plot we can come to a conclusion that “share” and “expenditure” are the principal feature attributes affecting the outcome of the approval. While “reports”, “age”, “income” and “active” affect the outcome very minimally. The other 3 attributes, viz, “dependents”, “months” and “majorcards” can be neglected as feature attributes affecting the outcome, according to this plot



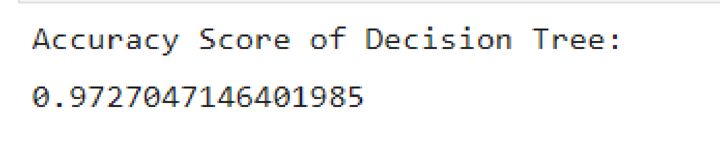
Model Building

- **Decision Tree**



Accuracy Score –

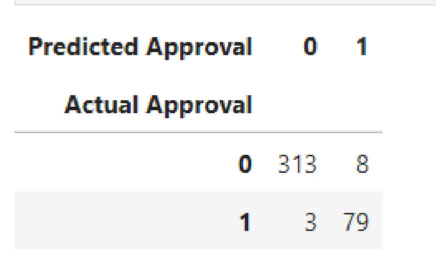
The decision tree has a 97.27% accuracy rate while predicting the outcomes of the Test data



```
Accuracy Score of Decision Tree:

0.9727047146401985
```

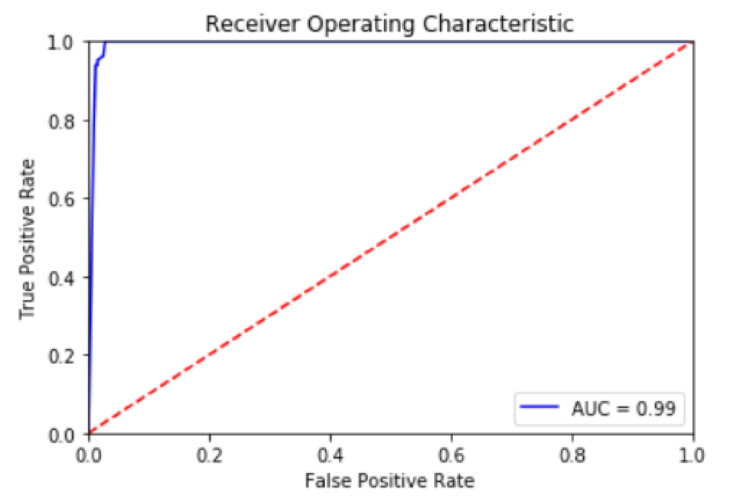
Confusion Matrix -



Predicted Approval	0	1
Actual Approval		
0	313	8
1	3	79

ROC Curve –

The AUC of the ROC curve is 0.99 (close to 1) which shows that the model predicts 99% of the actual positive approval as “yes” while testing

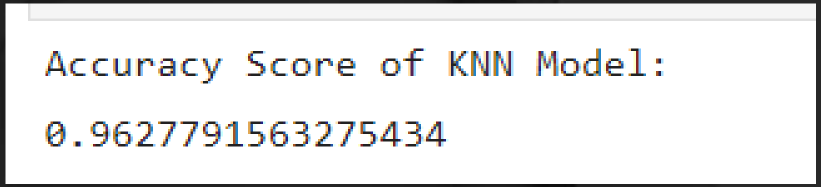


Model Building

- **K-NN**

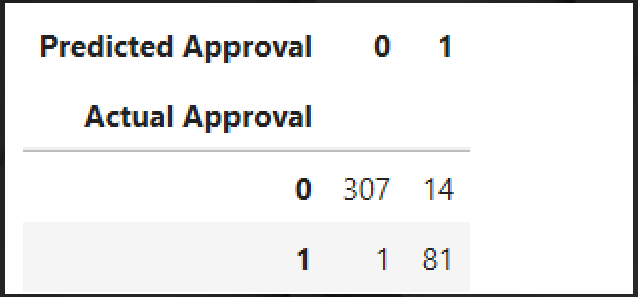
Accuracy Score –

The K-NN model has a 96.27% accuracy rate while predicting the outcomes of the Test data



```
Accuracy Score of KNN Model:
0.9627791563275434
```

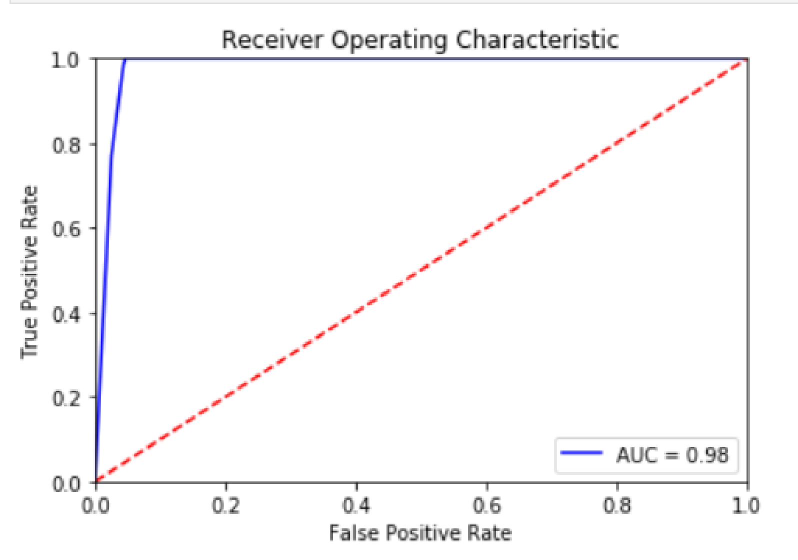
Confusion Matrix -



Predicted Approval	0	1
Actual Approval		
0	307	14
1	1	81

ROC Curve –

The AUC of the ROC curve is 0.98 (close to 1) which shows that the model predicts 98% of the actual positive approval as "yes" while testing

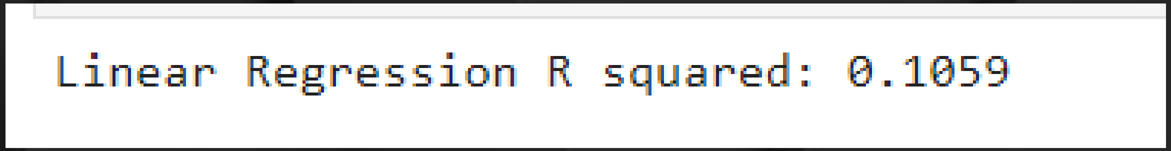


Model Building

- **Linear Regression**

Accuracy Score –

The Linear Regression model has a 10.59% accuracy rate while predicting the outcomes of the Test data



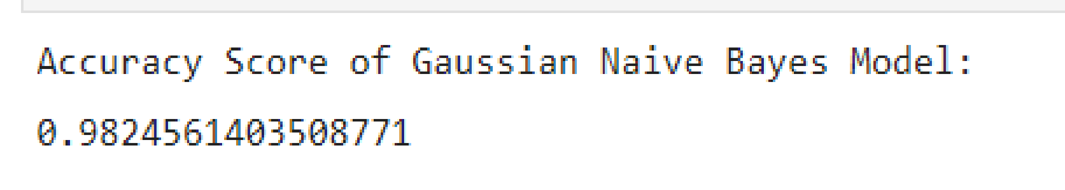
```
Linear Regression R squared: 0.1059
```

Model Building

- **Gaussian Naïve Bayes**

Accuracy Score –

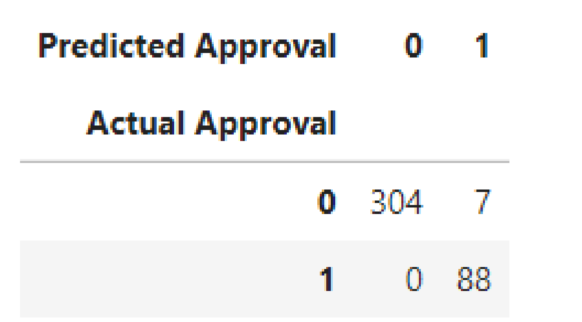
The Gaussian Naïve Bayes model has a 98.24% accuracy rate while predicting the outcomes of the Test data



```
Accuracy Score of Gaussian Naive Bayes Model:

0.9824561403508771
```

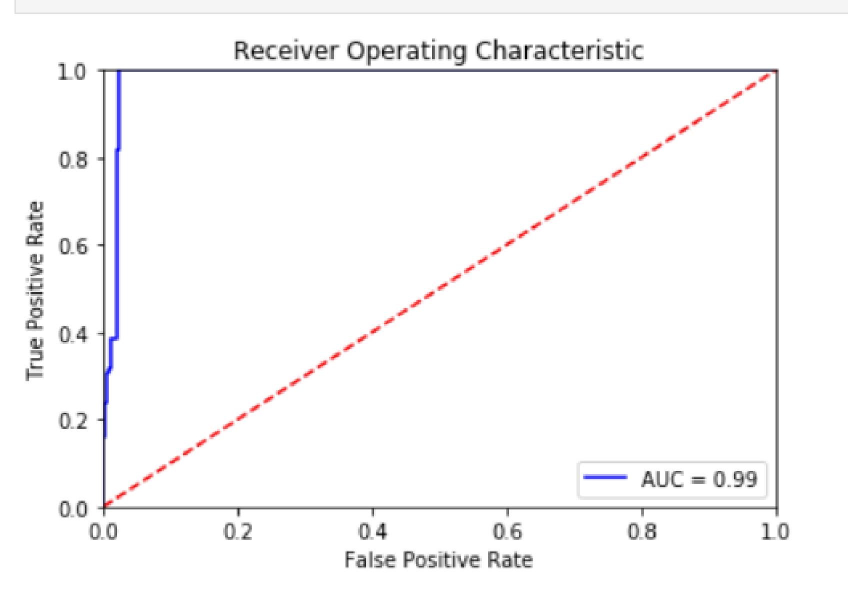
Confusion Matrix -



Predicted Approval	0	1
Actual Approval		
0	304	7
1	0	88

ROC Curve –

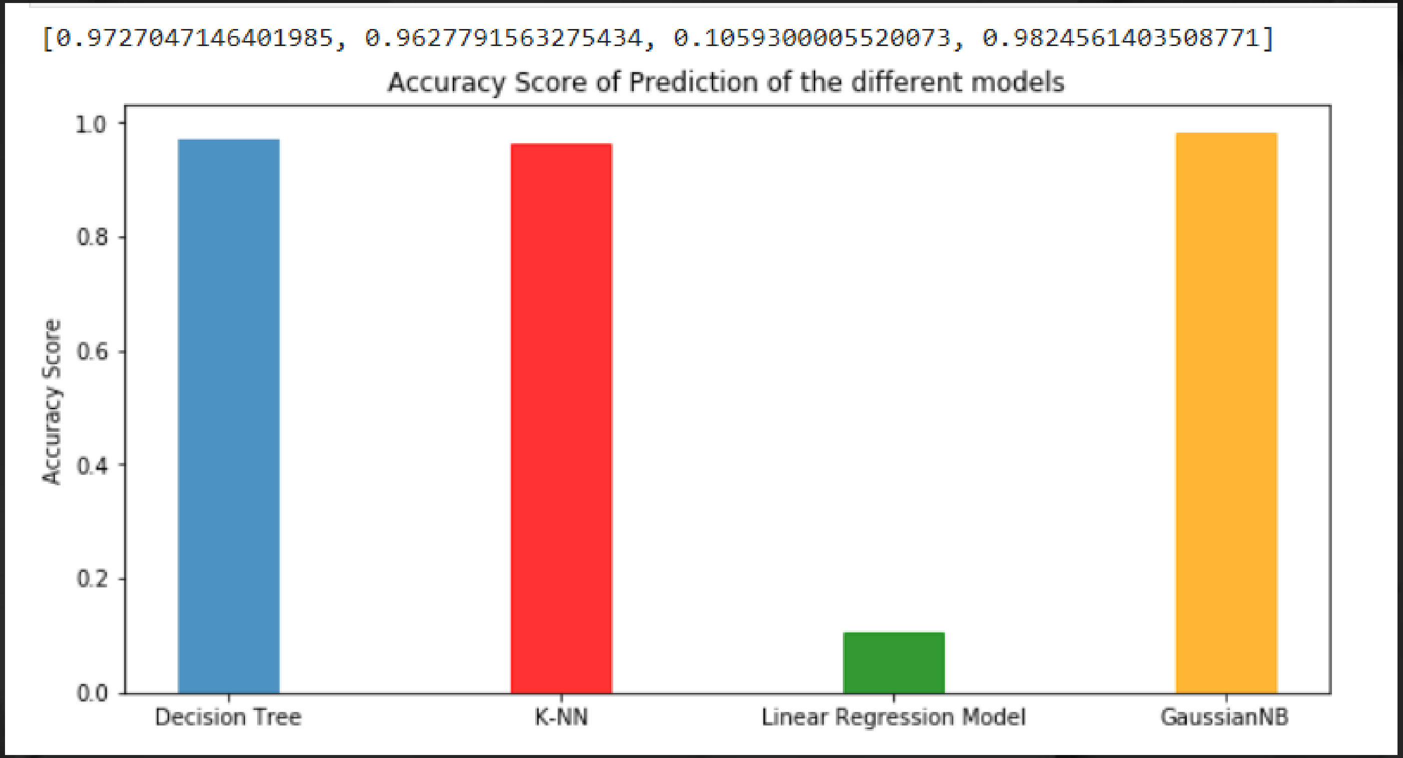
The AUC of the ROC curve is 0.99 (close to 1) which shows that the model predicts 99% of the actual positive approval as "yes" while testing



Model Building

Comparison of Accuracy Scores of all 4 models :

Decision Tree	97.27%
K-NN	96.27%
Linear Regeression	10.59%
Gaussian Naïve Bayes	98.24%



```
[0.9727047146401985, 0.9627791563275434, 0.1059300005520073, 0.9824561403508771]
```

Hence, we can conclude that the Gaussian Naïve Bayes Model is the model best-suited to make the most accurate prediction of the given data set on “Prediction of Credit Card Approval”

Future Scope of Improvements

From this initial analysis, we are able to conclude that the most significant factors in determining the outcome of a credit application are Expenditure, Income and Shares.

Based on these insights, we can work on building some predictive models. They can be used by analysts in financial sector and be incorporated to automate the credit approval process. These results can also serve as a source of information for the consumers.

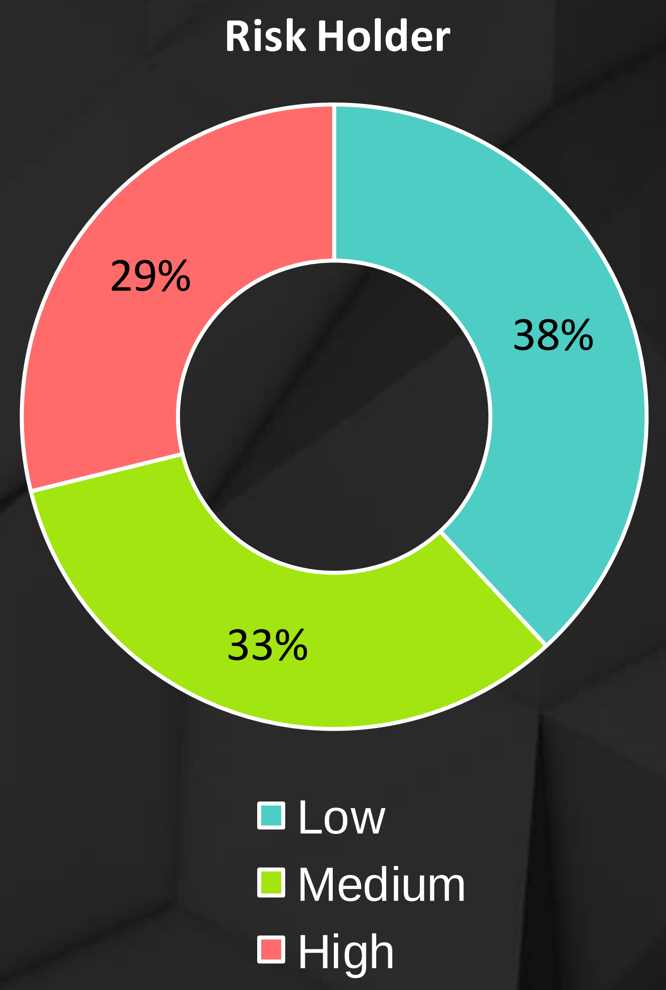
Modern credit analyses employ many additional variables like the criminal records of applicants, their health information, net balance between monthly income and expenses. A dataset with these variables could be acquired. It's also possible to add complementary variables to the dataset. This will make the credit simulations more, similar to what is done by the banks before a credit is approved.

Conclusion

On applying the logistic model, we have achieved the accuracy of 86% which is significantly high. Further, we have also categorized all the applicant into three different categories of risk. To do so, we have calculated the risk probability based on logistic regression.

Users with more than 90% of score considered as the low-risk holder; potentially the good users. Bank can offer good deals or increase in credit card balance for these applicants whereas below 20% considered as high-risk profile, which should get the approval only after proper document check and remaining are medium risk holders.

We also applied the random forest model and achieved the same accuracy. The random forest was especially useful to measure the attribute importance for the given data set.



Thank You!