Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

What decision is to be made?

As a Data Analyst/Business Analyst in the bank, you are to decide if the customers that apply for loan in the bank are creditworthy.

What data is needed to inform those decisions?

To make this decision, we need to predict that each customer that applies for loan in the bank is credit worthy. To do this, we will need:

1. Data on all past loan applications (Training Dataset)
2. A list of customers that are currently applying for the loan (Test Dataset)

The predictor variables needed to inform this decision are:

- Account Balance
- Payment Status of Previous Credit
- Purpose of the Loan
- Credit Amount
- Value Savings Stocks
- Length of Current Employment
- Instalment per cent
- Most valuable available asset
- Age years
- Type of Apartment
- No of Credits at this Bank

What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

The decisions to be made are whether the customer should be given the loan or not. This is a Yes/No, True/False situation and the model used in this category is the Binary Model. Under this model, we can use either Logistic Regression Models, Decision Tree Models, Random Forest Models or Boosted.

Step 2: Building the Training Set

The data types in the dataset were converted to suit the data type given below:

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

Missing Values

For the training dataset, there were 2 fields with missing values. The Age-years field had only 12 missing data, so, the median of the entire field where imputed where the data was missing in the field.

Whereas, for the “Duration in Current Address” field, there were many missing values (344), so the entire field was removed due to the amount of missing values when compared to the total number of records in the field.

```
df.isnull().sum()
```

```
Credit-Application-Result            0
Account-Balance                      0
Duration-of-Credit-Month             0
Payment-Status-of-Previous-Credit    0
Purpose                              0
Credit-Amount                        0
Value-Savings-Stocks                 0
Length-of-current-employment         0
Instalment-per-cent                  0
Guarantors                           0
Duration-in-Current-address        344
Most-valuable-available-asset        0
Age-years                           12
Concurrent-Credits                   0
Type-of-apartment                    0
No-of-Credits-at-this-Bank           0
Occupation                           0
No-of-dependents                     0
Telephone                            0
Foreign-Worker                       0
```

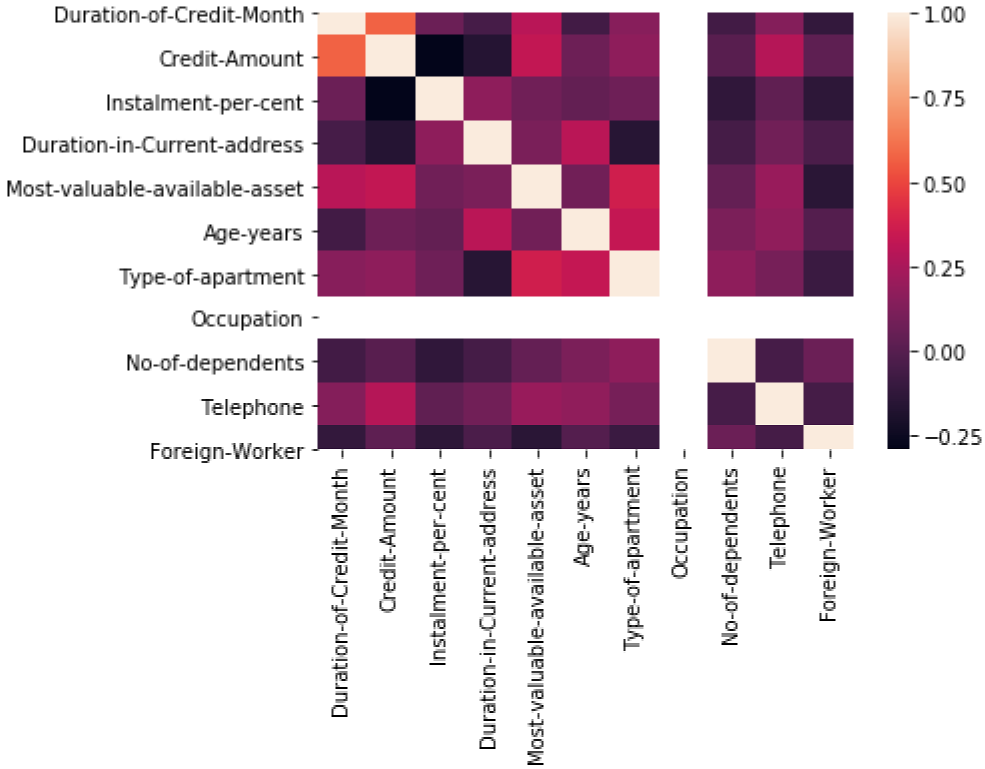
Data Features and the total number of missing values

Correlation

The correlation of each feature was checked numerically and using a heatmap. It was noticed that there was no correlation of up to 0.7 in the dataset. So, there is absence of miscorrelation.

	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worker
Duration-of-Credit-Month	1.000000	0.573980	0.068106	-0.050649	0.299855	-0.066319	0.152516	NaN	-0.065269	0.143176	-0.115916
Credit-Amount	0.573980	1.000000	-0.288852	-0.158069	0.325545	0.068643	0.170071	NaN	0.003986	0.286338	0.025493
Instalment-per-cent	0.068106	-0.288852	1.000000	0.173393	0.081493	0.040540	0.074533	NaN	-0.125894	0.029354	-0.133411
Duration-in-Current-address	-0.050649	-0.158069	0.173393	1.000000	0.109297	0.301966	-0.157550	NaN	-0.056646	0.084925	-0.036587
Most-valuable-available-asset	0.299855	0.325545	0.081493	0.109297	1.000000	0.085437	0.373101	NaN	0.046454	0.203509	-0.146005
Age-years	-0.066319	0.068643	0.040540	0.301966	0.085437	1.000000	0.333075	NaN	0.117735	0.176479	-0.003285
Type-of-apartment	0.152516	0.170071	0.074533	-0.157550	0.373101	0.333075	1.000000	NaN	0.170738	0.101443	-0.089848
Occupation	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No-of-dependents	-0.065269	0.003986	-0.125894	-0.056646	0.046454	0.117735	0.170738	NaN	1.000000	-0.048559	0.065943
Telephone	0.143176	0.286338	0.029354	0.084925	0.203509	0.176479	0.101443	NaN	-0.048559	1.000000	-0.055516
Foreign-Worker	-0.115916	0.025493	-0.133411	-0.036587	-0.146005	-0.003285	-0.089848	NaN	0.065943	-0.055516	1.000000

Numerical Representation of Correlation among features

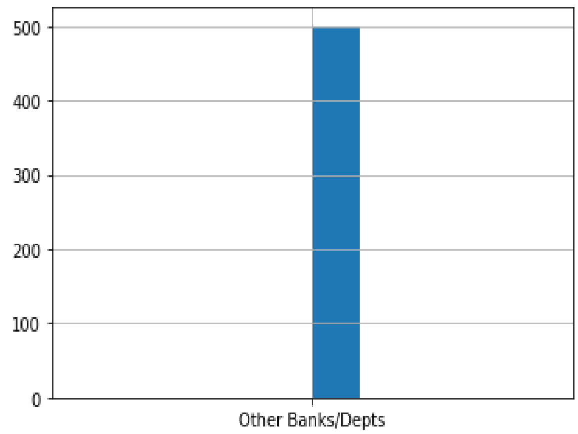


Graphical Representation of Correlation using Heatmap.

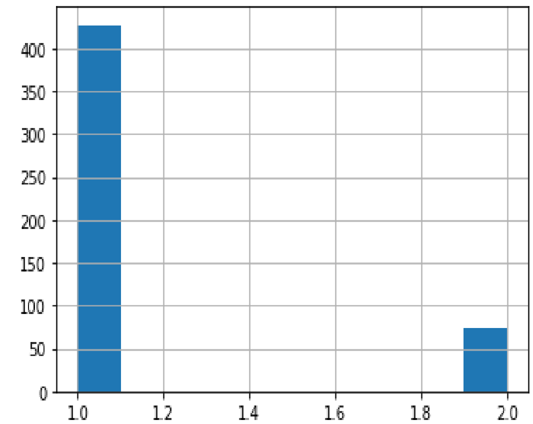
Usefulness to Prediction
The telephone column was removed because it has no relevance to prediction

Low Variability
The columns with low variability are visualized below. There were 5 columns with Low Variability in the dataset and they are: Guarantors, Concurrent Credits, No. of dependents, Occupation and Foreign Worker.

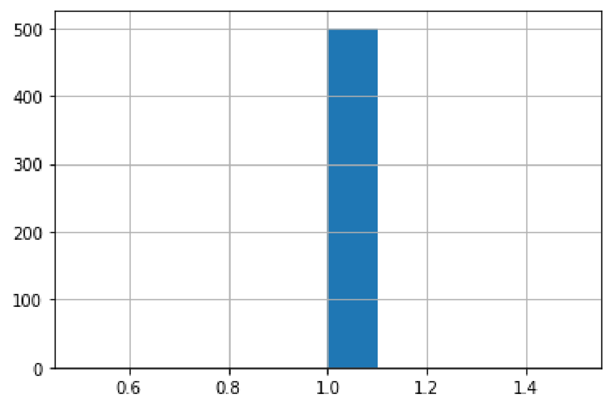
Guarantors



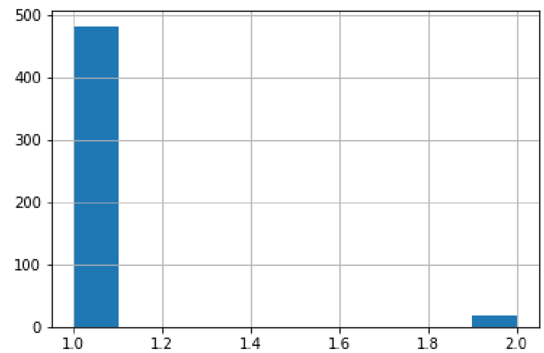
Concurrent Credits



No. of Dependents



Occupation



Foreign Worker

Deductions

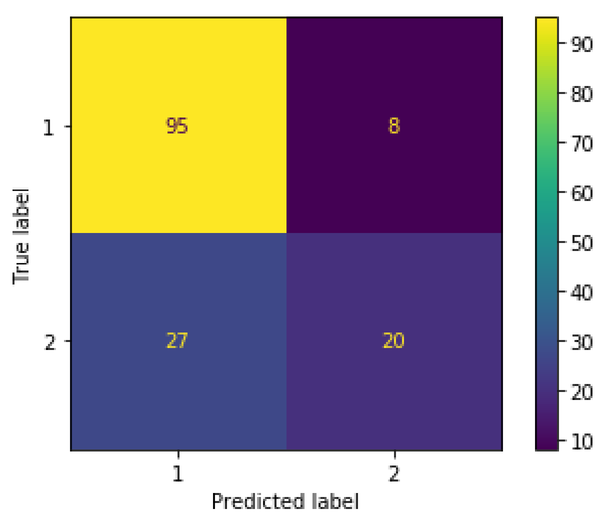
After cleaning and preparation of the dataset, it was discovered that there are 13 columns in the dataset and the Mean of the Age-years column is 35.574 which is approximately 36.
The Mean of the Age-Years was found to be 35.574 and approximately 36.

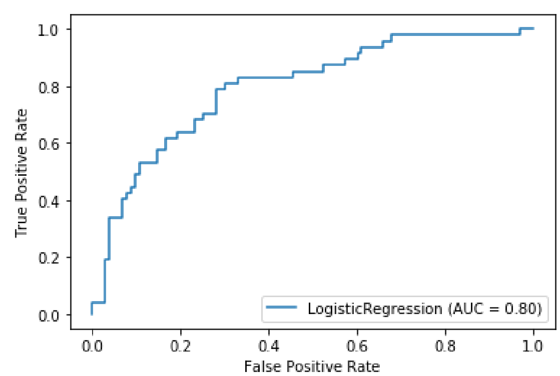
Step 3: Train your Classification Models

The data was divided into 2 parts, the Estimation sample or Train Set, which covers 70% of the dataset and the Validation Set/Test set which account for 30% of the dataset.

Logistic Regression

	Feature_Importance	Features
11	0.455159	Payment-Status-of-Previous-Credit_Some Problems
7	0.369986	Account-Balance_No Account
13	0.292450	Value-Savings-Stocks_None
17	0.234103	Length-of-current-employment_< 1yr
4	0.129206	Most-valuable-available-asset
3	0.014430	Instalment-per-cent
0	0.012577	Duration-of-Credit-Month
2	0.000103	Credit-Amount
5	-0.020098	Age-years
16	-0.127113	Length-of-current-employment_4-7 yrs
19	-0.131531	No-of-Credits-at-this-Bank_More than 1
1	-0.145652	Purpose
6	-0.166917	Type-of-apartment
18	-0.188796	No-of-Credits-at-this-Bank_1
12	-0.209471	Value-Savings-Stocks_< £100
10	-0.295703	Payment-Status-of-Previous-Credit_Paid Up
14	-0.403305	Value-Savings-Stocks_£100-£1000
15	-0.427316	Length-of-current-employment_1-4 yrs
9	-0.479782	Payment-Status-of-Previous-Credit_No Problems ...
8	-0.690313	Account-Balance_Some Balance

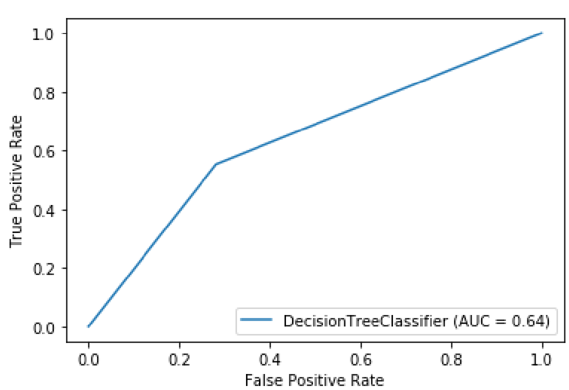


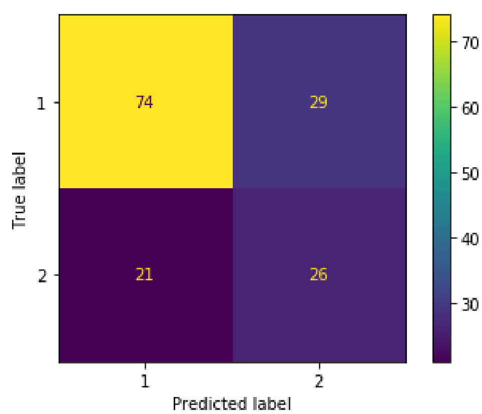


The Overall Accuracy is 0.7667

Decision Trees

	Features	Feature_Importance
2	Credit-Amount	0.276888
0	Duration-of-Credit-Month	0.163133
5	Age-years	0.146341
3	Instalment-per-cent	0.066642
8	Account-Balance_Some Balance	0.065773
4	Most-valuable-available-asset	0.055565
11	Payment-Status-of-Previous-Credit_Some Problems	0.041246
16	Length-of-current-employment_4-7 yrs	0.039390
9	Payment-Status-of-Previous-Credit_No Problems ...	0.027268
1	Purpose	0.025374
17	Length-of-current-employment_< 1yr	0.022313
10	Payment-Status-of-Previous-Credit_Paid Up	0.019483
19	No-of-Credits-at-this-Bank_More than 1	0.019264
14	Value-Savings-Stocks_£100-£1000	0.013485
12	Value-Savings-Stocks_< £100	0.009532
6	Type-of-apartment	0.008304
13	Value-Savings-Stocks_None	0.000000
15	Length-of-current-employment_1-4 yrs	0.000000
7	Account-Balance_No Account	0.000000
18	No-of-Credits-at-this-Bank_1	0.000000

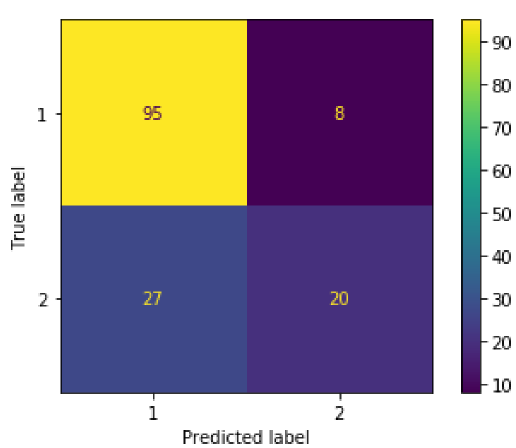


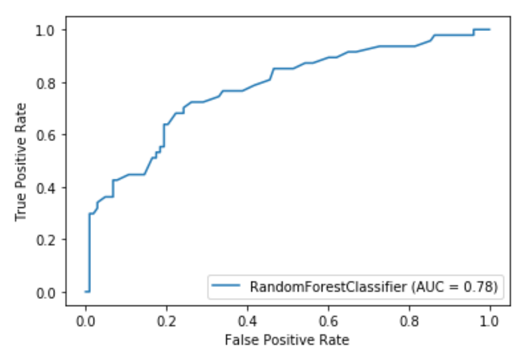


The Overall Accuracy is 0.6666

Random Forest

	Features	Feature_Importance
2	Credit-Amount	0.205994
5	Age-years	0.143945
0	Duration-of-Credit-Month	0.139751
4	Most-valuable-available-asset	0.065392
3	Instalment-per-cent	0.056352
11	Payment-Status-of-Previous-Credit_Some Problems	0.037854
1	Purpose	0.037320
6	Type-of-apartment	0.036297
7	Account-Balance_No Account	0.035520
8	Account-Balance_Some Balance	0.032021
17	Length-of-current-employment_< 1yr	0.030167
13	Value-Savings-Stocks_None	0.029579
14	Value-Savings-Stocks_£100-£1000	0.021528
10	Payment-Status-of-Previous-Credit_Paid Up	0.021498
18	No-of-Credits-at-this-Bank_1	0.020652
9	Payment-Status-of-Previous-Credit_No Problems ...	0.020614
19	No-of-Credits-at-this-Bank_More than 1	0.018615
15	Length-of-current-employment_1-4 yrs	0.018068
16	Length-of-current-employment_4-7 yrs	0.017179
12	Value-Savings-Stocks_< £100	0.011654

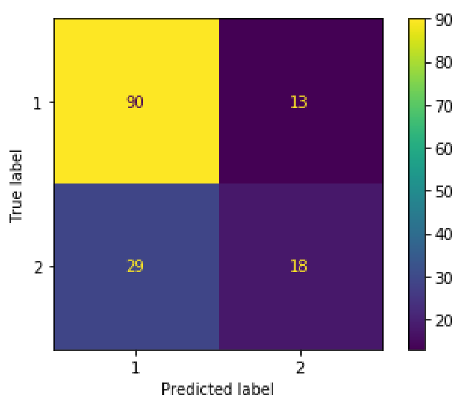


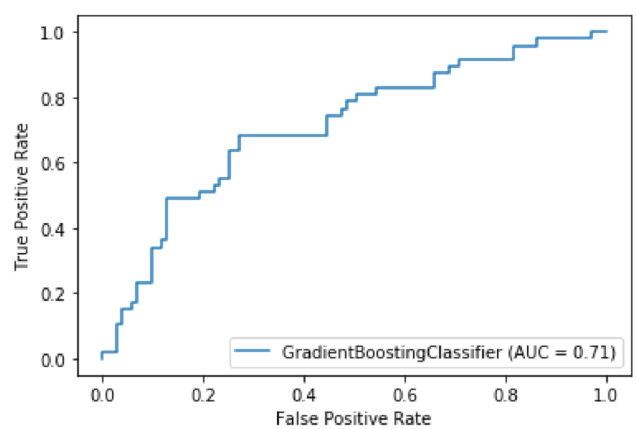


The Overall Accuracy is 0.7666

Gradient Boosting Classifier

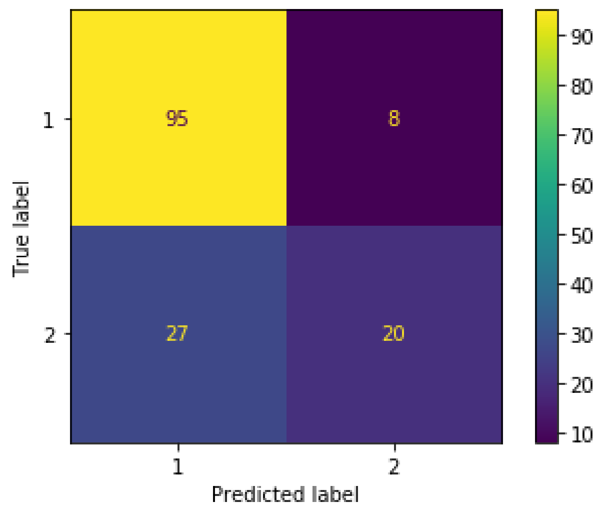
	Features	Feature_Importance
2	Credit-Amount	0.294664
0	Duration-of-Credit-Month	0.193132
5	Age-years	0.141032
11	Payment-Status-of-Previous-Credit_Some Problems	0.076308
8	Account-Balance_Some Balance	0.046309
3	Instalment-per-cent	0.038006
13	Value-Savings-Stocks_None	0.035535
7	Account-Balance_No Account	0.034940
17	Length-of-current-employment_< 1yr	0.033146
4	Most-valuable-available-asset	0.032603
14	Value-Savings-Stocks_£100-£1000	0.018399
1	Purpose	0.018261
6	Type-of-apartment	0.011522
9	Payment-Status-of-Previous-Credit_No Problems ...	0.007636
15	Length-of-current-employment_1-4 yrs	0.005686
18	No-of-Credits-at-this-Bank_1	0.005159
19	No-of-Credits-at-this-Bank_More than 1	0.003048
12	Value-Savings-Stocks_< £100	0.003029
10	Payment-Status-of-Previous-Credit_Paid Up	0.001423
16	Length-of-current-employment_4-7 yrs	0.000163



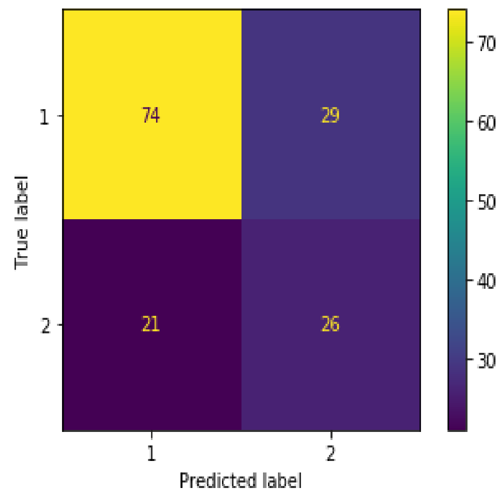


Overall Accuracy: 0.72

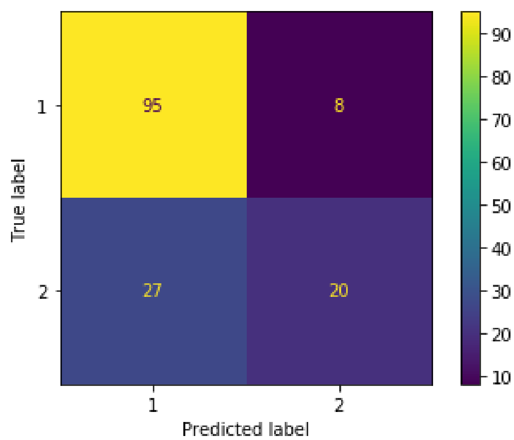
Classification Matrices for all Models



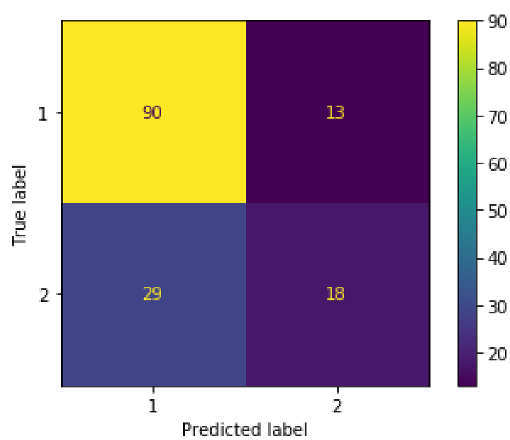
Logistic Regression



Decision Trees



Random Forest Boosted



Model Classification

Deductions:

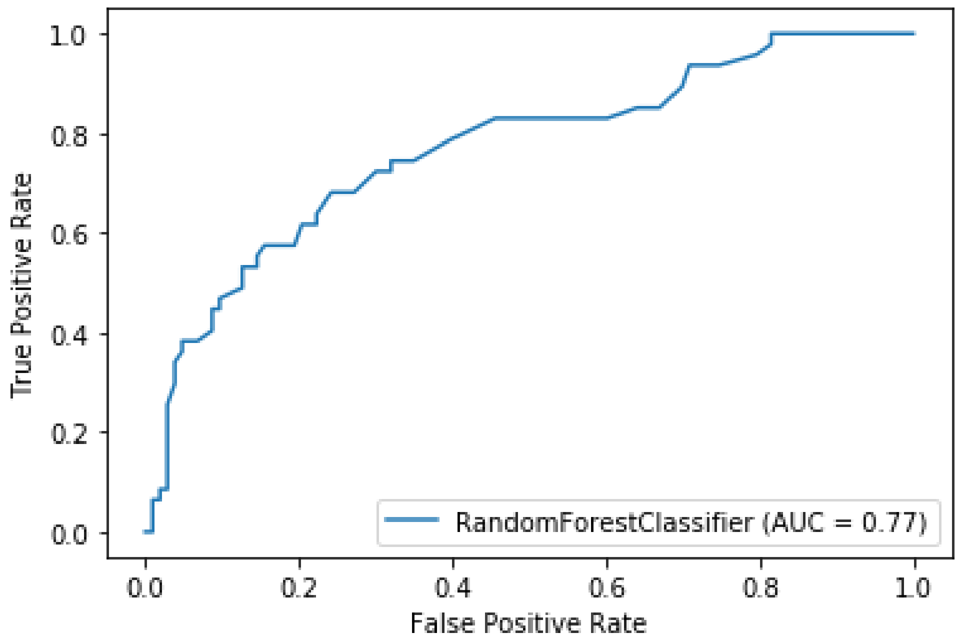
Model	**Sensitivity**	**Specificity**
Logistic Regression	0.7787	0.7143
Decision Tree	0.7789	0.4727
Random Forest Classifier	0.7797	0.7143
Gradient Boosted Classifier	0.7627	0.5806

The Models, Logistic Regression and Random Forest are said to have low bias because their sensitivity is relatively close to specificity, whereas Decision Tree and Random Forest Classifier have bias in prediction.

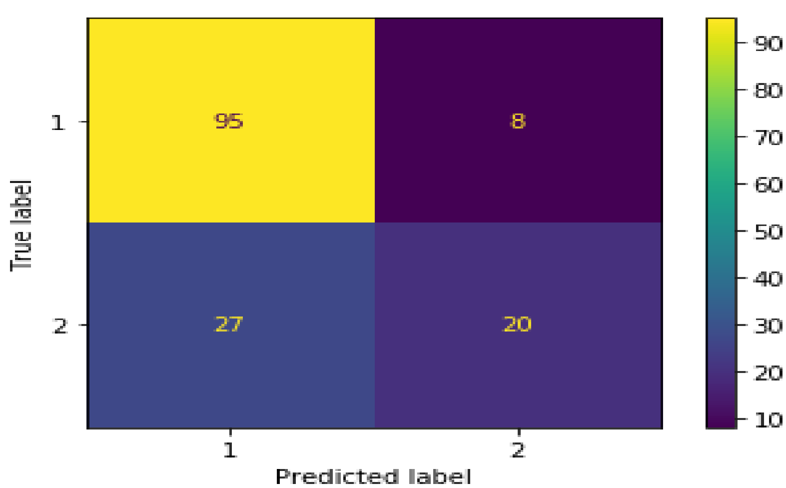
Step 4: Writeup

The models with low bias are Logistic Regression and the Random Forest Models. These two models also have the highest AUC Score. The best predictive model in this case would be the Random Forest Classifier because it has the lowest bias. Additionally, it has a Sensitivity of 77.97 % and a Specificity of 71.43%,an overall accuracy of 76.67 and an AUC Score from the ROC Graph of 78%.

Using the Random Forest Model
The Number of Credit Worthy Individuals are 420
The Number of Non Credit Worthy Individuals are 80



The Random Forest Classifier ROC Curve



Random Forest Classification Matrix