Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?
  - ✓ I need to evaluate the creditworthiness of the new 500 loan applicants.

- What data is needed to inform those decisions?
  - ✓ I need past loan applicant's information on credit application results and the data used to rate those results like Duration of credit, credit amount, installment, age of the applicant.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  - ✓ The model type will be Binary as I will be predicting an applicant to be either creditworthy or non-creditworthy.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Here are some guidelines to help guide your data cleanup:

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".
  - ✓ No numerical data fields that are highly correlated.

	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	Telephone
Duration-of-Credit-Month	1.00	0.57	0.07	0.30	-0.07	0.14
Credit-Amount	0.57	1.00	-0.29	0.33	0.07	0.29
Instalment-per-cent	0.07	-0.29	1.00	0.08	0.04	0.03
Most-valuable-available-asset	0.30	0.33	0.08	1.00	0.09	0.20
Age-years	-0.07	0.07	0.04	0.09	1.00	0.18
Telephone	0.14	0.29	0.03	0.20	0.18	1.00

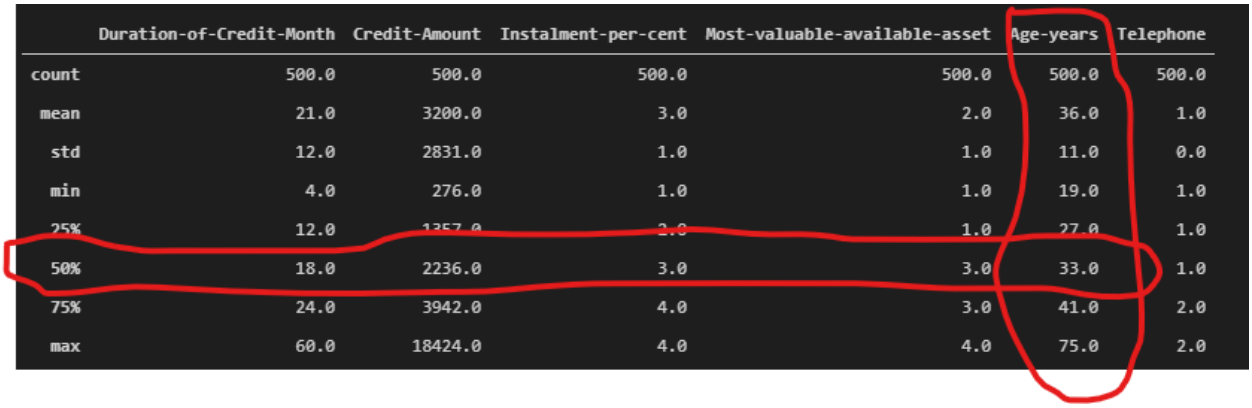
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed.
    - ✓ Yes, Duration in current address and Age have missing values of 344 and 12 respective.
    - ✓ I removed Duration in current address for having a lot of missing data

```
Counting Missing Values

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
dtype: int64
```

- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.
    - ✓ Yes, I removed below fields with low variability.
        - ➢ Concurrent credits
        - ➢ Occupation
        - ➢ Guarantors
        - ➢ Type of apartment
        - ➢ No of dependents
        - ➢ Foreign worker

- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)
    - ✓ The imputed field is Age-years, There 12 applicants with empty age data. I cannot remove these applicants as I will lose 2.4% of the data. I will fill all empty data with an age median of 33



	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	Telephone
count	500.0	500.0	500.0	500.0	500.0	500.0
mean	21.0	3200.0	3.0	2.0	36.0	1.0
std	12.0	2831.0	1.0	1.0	11.0	0.0
min	4.0	276.0	1.0	1.0	19.0	1.0
25%	12.0	1357.0	2.0	1.0	27.0	1.0
50%	18.0	2236.0	3.0	3.0	33.0	1.0
75%	24.0	3942.0	4.0	3.0	41.0	2.0
max	60.0	18424.0	4.0	4.0	75.0	2.0

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

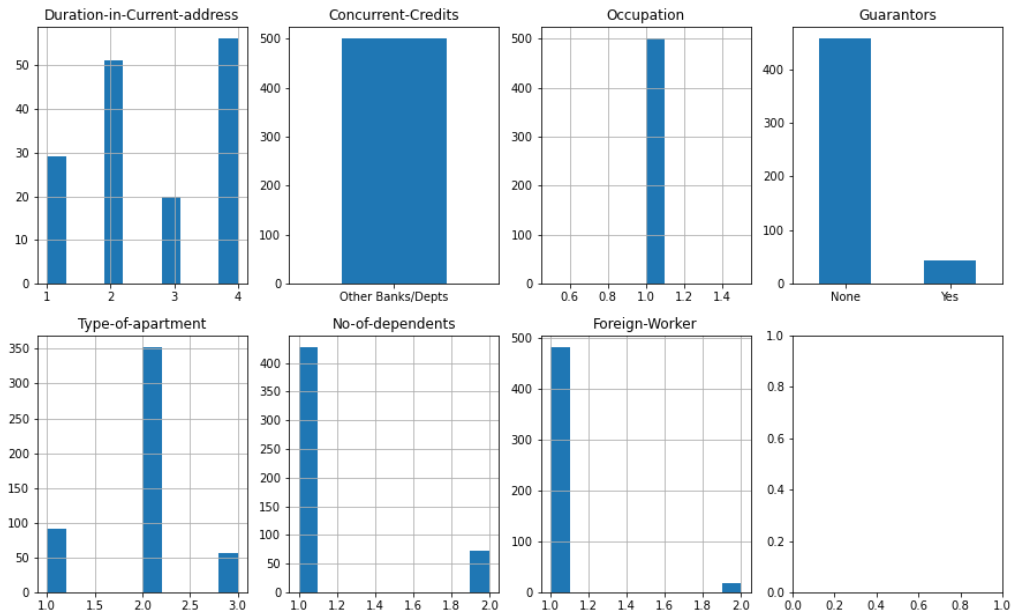
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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

To achieve consistent results reviewers expect.

Answer this question:

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  - ✓ I removed below fields.
    - ➢ Duration in a current address – Due to many missing data
    - ➢ Concurrent credits – Due to low variability
    - ➢ Occupation – Due to low variability
    - ➢ Guarantors – Due to low variability
    - ➢ Type of apartment – Due to low variability
    - ➢ No of dependents – Due to low variability
    - ➢ Foreign worker – Due to low variability



Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

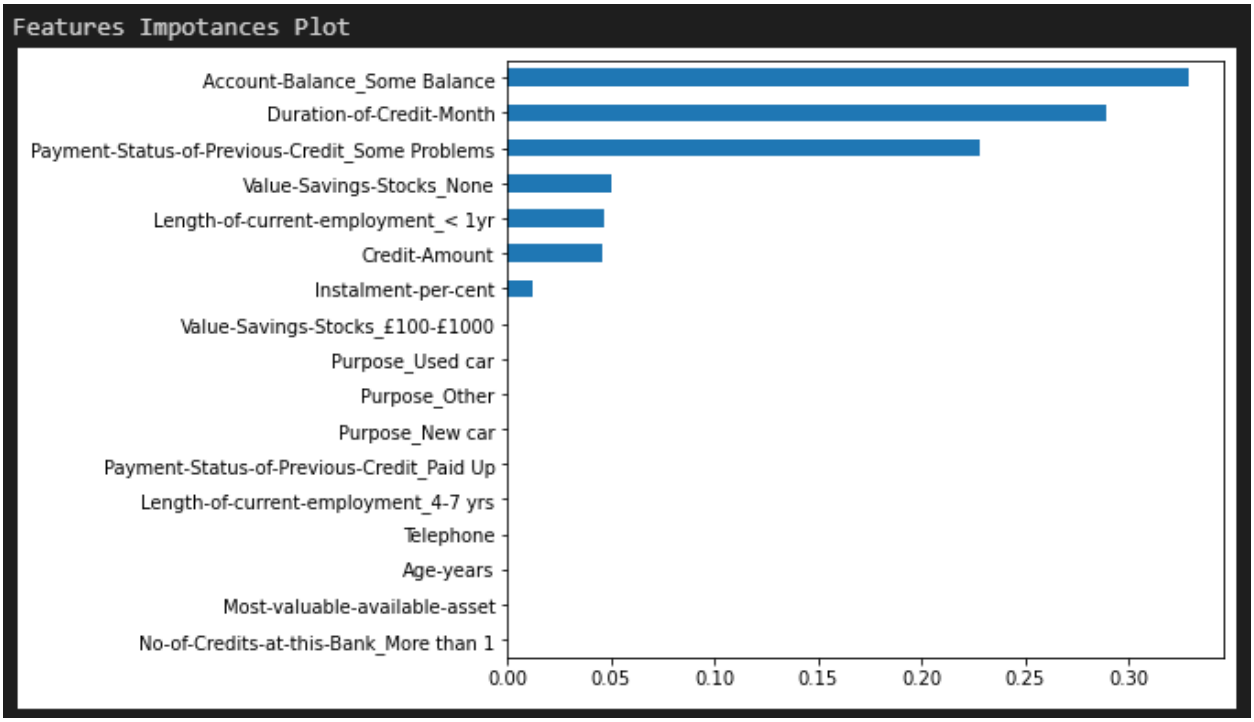
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

    I. Logistic Regression
    The most important predictor variables are credit amount, account balance, payment status of previous credit, and purpose.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Duration-of-Credit-Month	-0.2125	0.1644	-1.2923	0.1963	-0.5347	0.1098
Credit-Amount	-0.4479	0.1908	-2.3482	0.0189	-0.8218	-0.0741
Instalment-per-cent	-0.2489	0.1574	-1.5816	0.1137	-0.5573	0.0595
Most-valuable-available-asset	-0.1162	0.1517	-0.7658	0.4438	-0.4134	0.1811
Age-years	0.0586	0.1536	0.3815	0.7029	-0.2425	0.3597
Telephone	0.0526	0.1490	0.3533	0.7239	-0.2393	0.3446
Account-Balance_Some Balance	1.0963	0.2838	3.8628	0.0001	0.5400	1.6525
Payment-Status-of-Previous-Credit_Paid Up	0.2997	0.3214	0.9327	0.3510	-0.3301	0.9296
Payment-Status-of-Previous-Credit_Some Problems	-1.8092	0.6626	-2.7304	0.0063	-3.1078	-0.5105
Purpose_New car	1.2993	0.5696	2.2811	0.0225	0.1829	2.4156
Purpose_Other	1.1346	1.1089	1.0232	0.3062	-1.0387	3.3080
Purpose_Used car	0.5665	0.3941	1.4375	0.1506	-0.2059	1.3388
Value-Savings-Stocks_None	-0.0454	0.3952	-0.1148	0.9086	-0.8199	0.7291
Value-Savings-Stocks_£100-£1000	0.4645	0.4615	1.0065	0.3142	-0.4401	1.3691
Length-of-current-employment_4-7 yrs	0.3513	0.4274	0.8220	0.4111	-0.4864	1.1890
Length-of-current-employment_< 1yr	-0.2864	0.3247	-0.8820	0.3778	-0.9228	0.3500
No-of-Credits-at-this-Bank_More than 1	0.5265	0.3464	1.5197	0.1286	-0.1525	1.2055

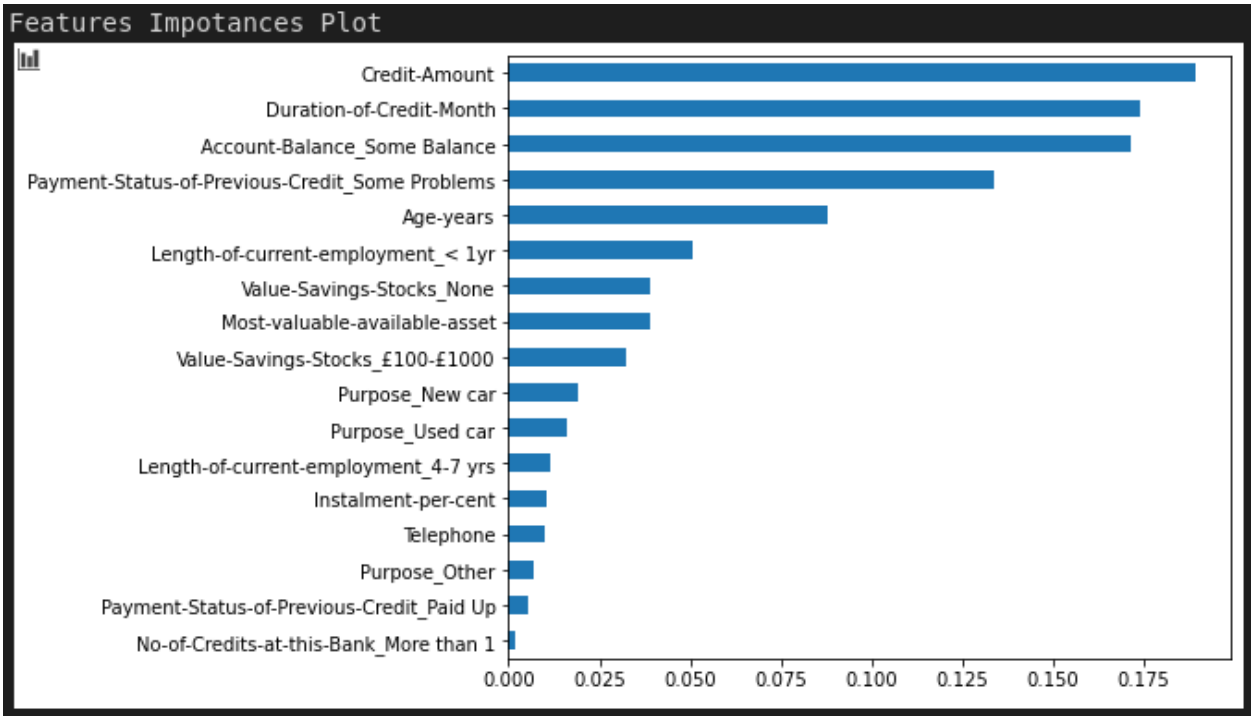
II. Decision Tree

The most important predictor variables are account balance, duration of credit, payment status of previous credit, value savings stocks, length of current employment, credit amount and instalment percent.



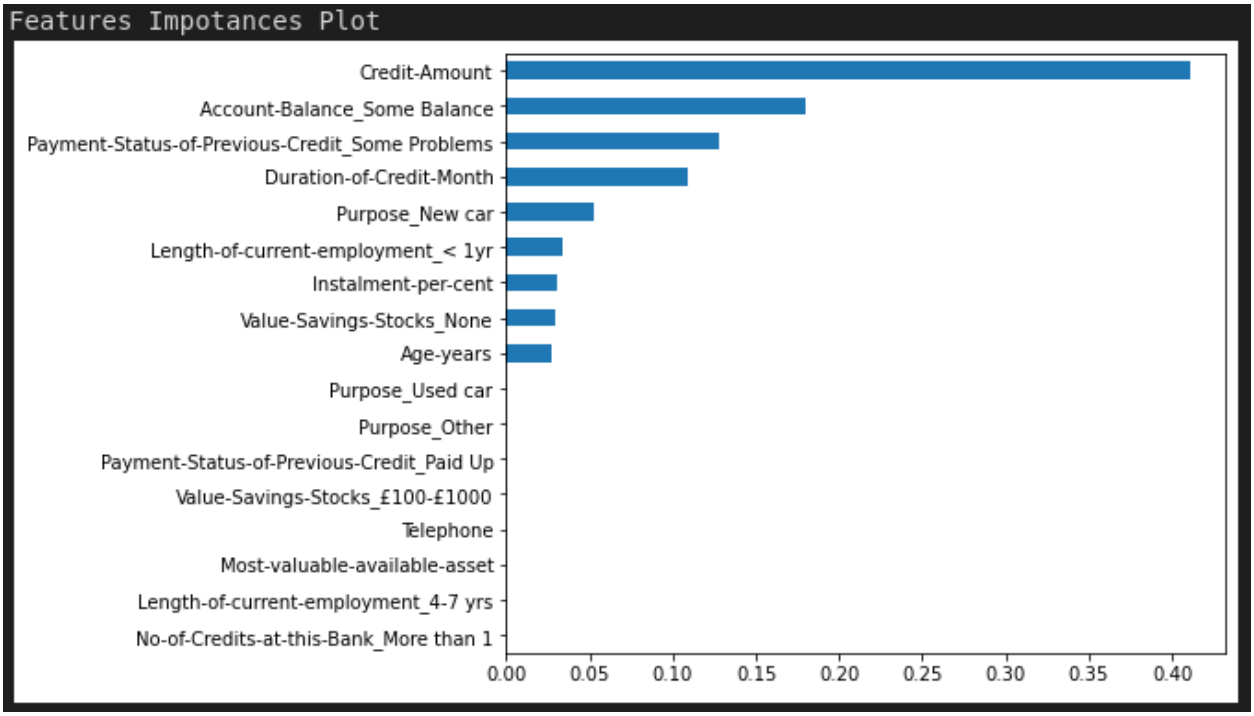
III. Forest Model

All predictor variables are importance



IV. Boosted Model

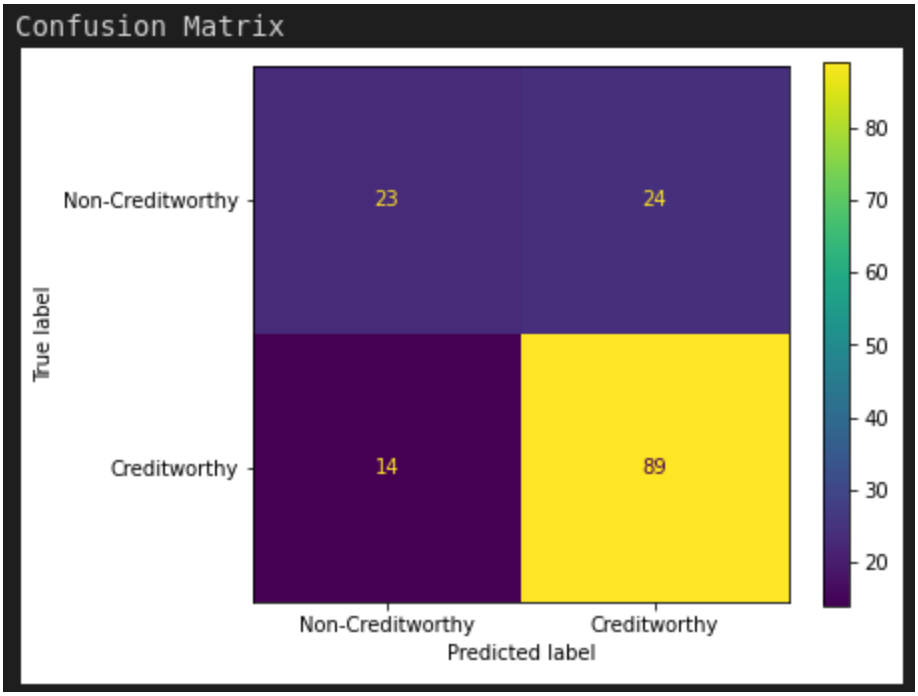
The most important predictor variables are credit amount, account balance, payment status of previous credit, duration of credit, purpose, length of current employment, instalment percent, value savings stocks and age.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

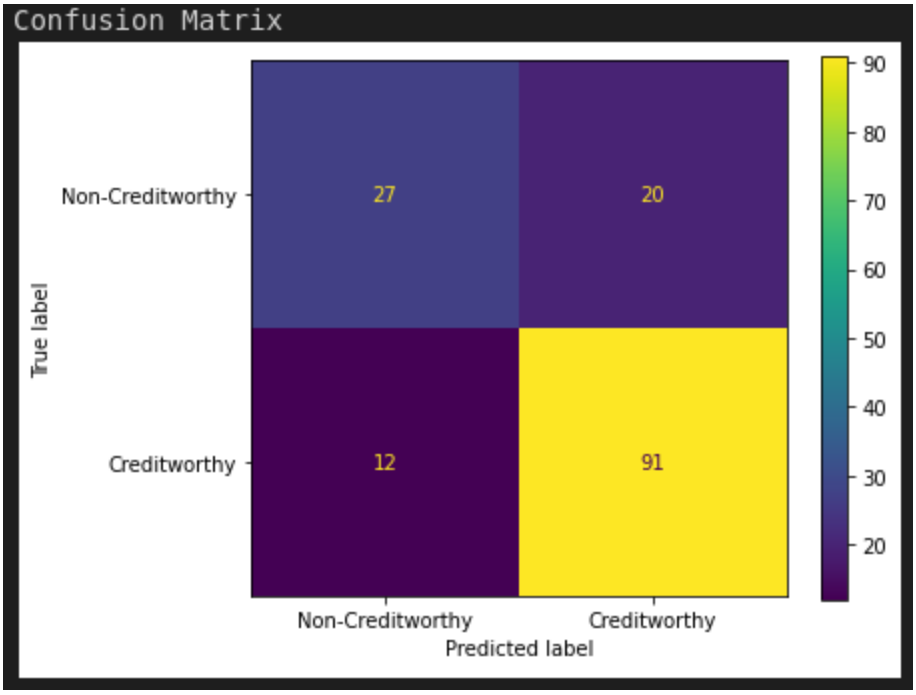
    I. Logistic Regression

    The logistic regression model has an accuracy of 75% in validation data. There bias due to less data on non-creditworthy applicants.



II. Decision Tree

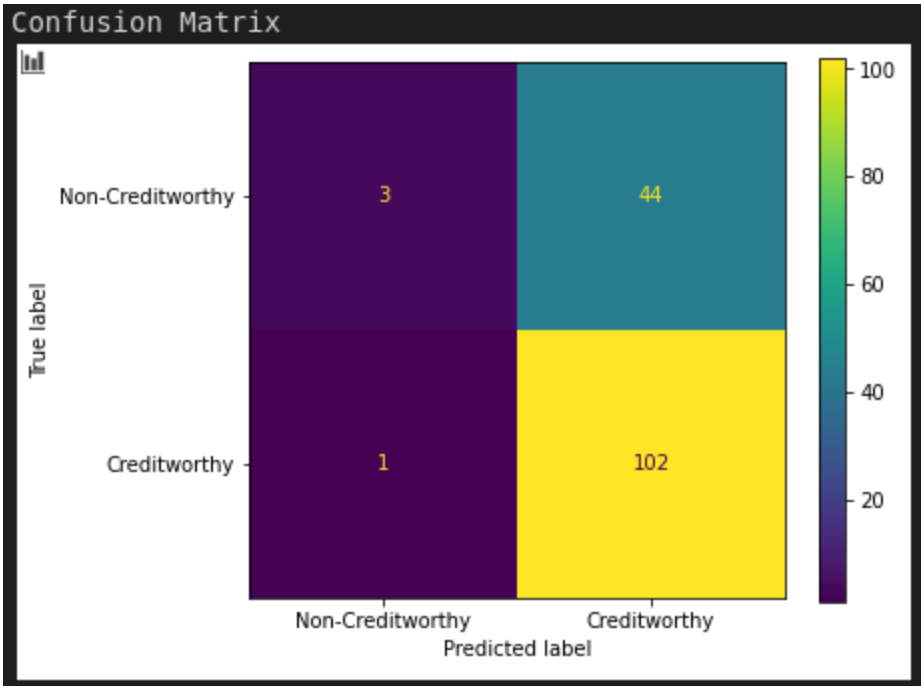
The decision tree model has an accuracy of 69% in validation data. There bias due to less data on non-creditworthy applicants.



III. Forest Model

The forest model has an accuracy of 70% in validation data.

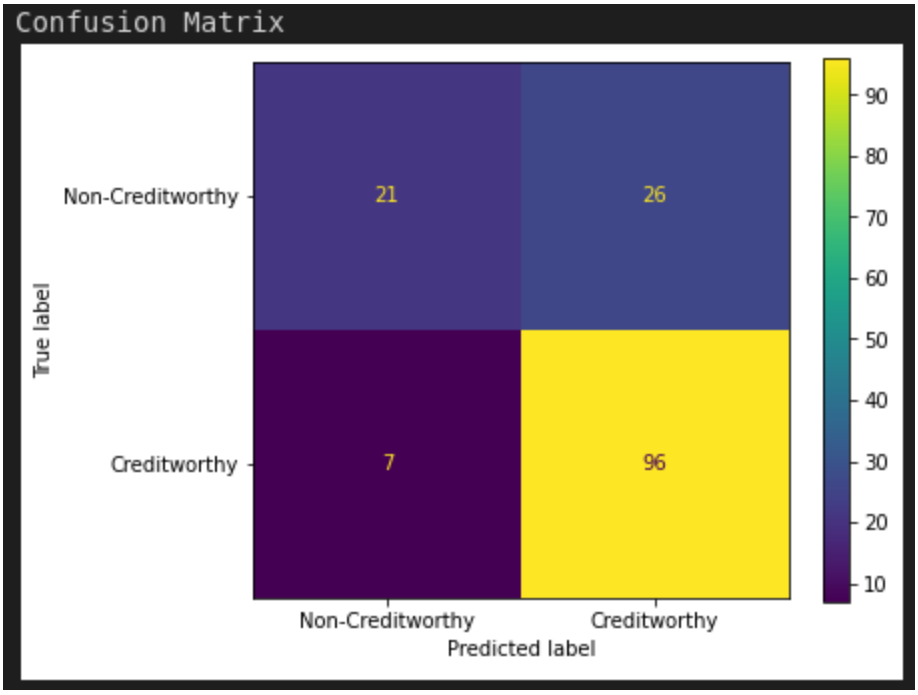
There bias due to less data on non-creditworthy applicants.



IV. Boosted Model

The boosted model has an accuracy of 78% in validation data.

There bias due to less data on non-creditworthy applicants.



You should have four sets of questions answered. (500 word limit)

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

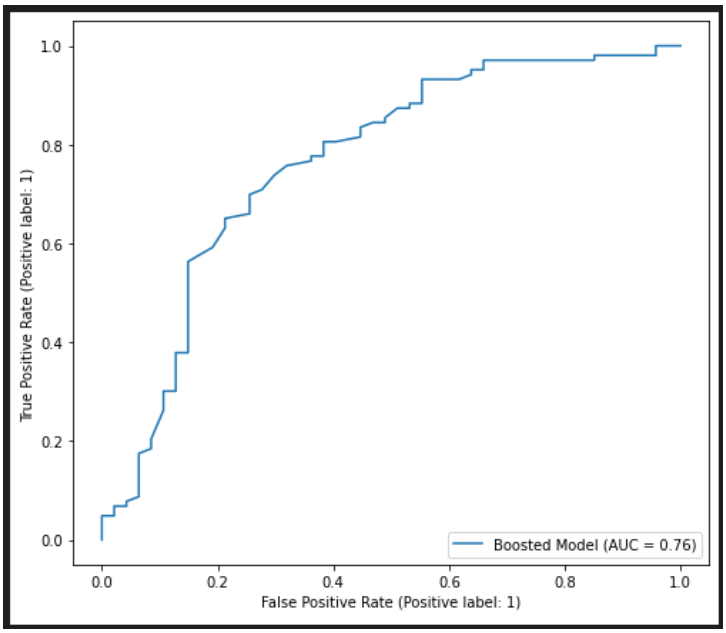
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
  - ✓ I choose boosted model due to high accuracy.
  - ○ Overall Accuracy against your Validation set
    - ✓ The boosted model has 78% accuracy against validation set
  - ○ Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ✓ The creditworthy has precision of 79% and non-creditworthy has precision of 75%
    - ✓ The creditworthy has recall of 93% and non-creditworthy has recall of 45%

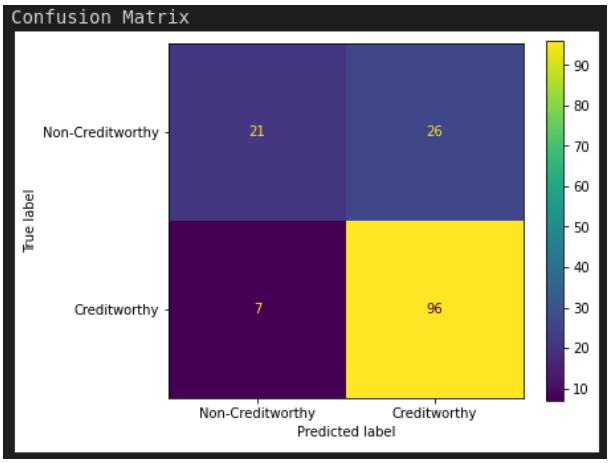
	precision	recall	f1-score	support
Non-Creditworthy	0.75	0.45	0.56	47
Creditworthy	0.79	0.93	0.85	103
accuracy			0.78	150
macro avg	0.77	0.69	0.71	150
weighted avg	0.78	0.78	0.76	150

- ROC graph



- Bias in the Confusion Matrices

There bias due to less data on non-creditworthy applicants.



Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?
  - The individuals predicted to be creditworthy are 441