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 to create model.
  - ✓ I need new 500 loan applicants to predict who is creditworthy from the model created.

- 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	Type-of-apartment
Duration-of-Credit-Month	1.00	0.57	0.07	0.30	-0.06	0.15
Credit-Amount	0.57	1.00	-0.29	0.33	0.07	0.17
Instalment-per-cent	0.07	-0.29	1.00	0.08	0.04	0.07
Most-valuable-available-asset	0.30	0.33	0.08	1.00	0.09	0.37
Age-years	-0.06	0.07	0.04	0.09	1.00	0.33
Type-of-apartment	0.15	0.17	0.07	0.37	0.33	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
        - ➢ Telephone
        - ➢ 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

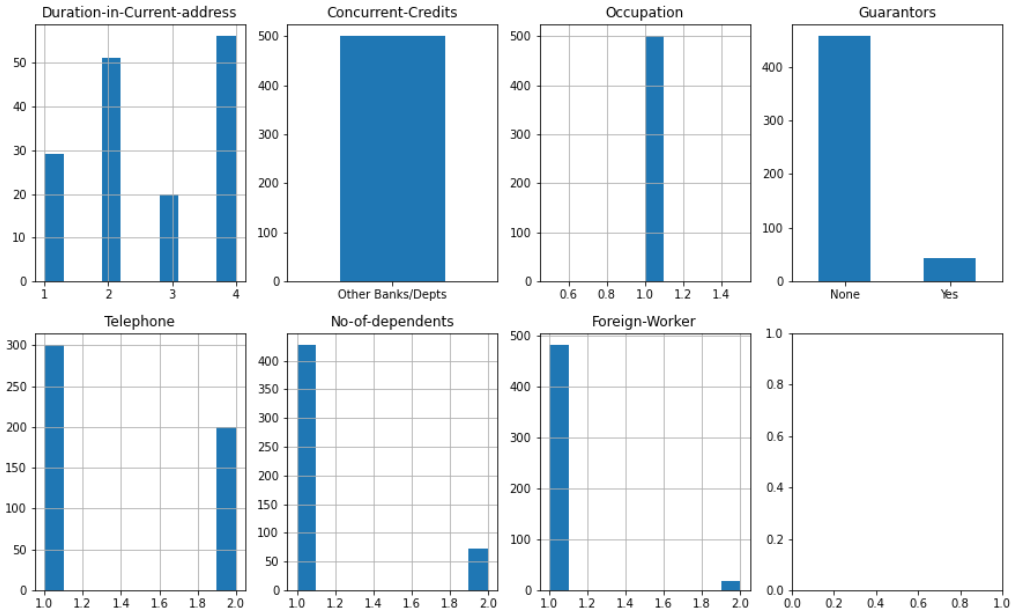
	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	Type-of-apartment
count	500.0	500.0	500.0	500.0	500.0	500.0
mean	21.0	3200.0	3.0	2.0	36.0	2.0
std	12.0	2831.0	1.0	1.0	11.0	1.0
min	4.0	276.0	1.0	1.0	19.0	1.0
25%	12.0	1357.0	2.0	1.0	27.0	2.0
50%	18.0	2236.0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

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)*

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 the 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
    - ➢ Telephone– 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:*

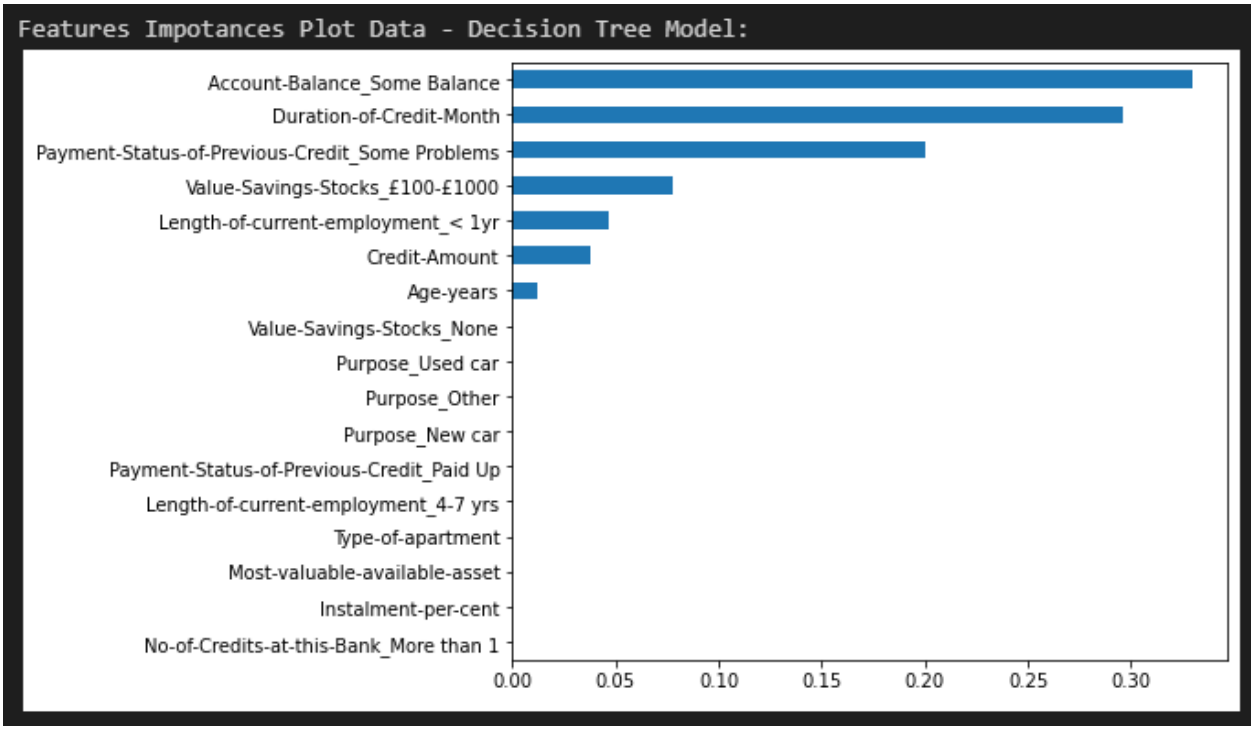
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

    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	-0.2163	0.1638	-1.3199	0.1869	-0.5374	0.1049
Credit-Amount	-0.4279	0.1831	-2.3371	0.0194	-0.7867	-0.0691
Instalment-per-cent	-0.2406	0.1568	-1.5346	0.1249	-0.5479	0.0667
Most-valuable-available-asset	-0.0994	0.1583	-0.6278	0.5301	-0.4097	0.2109
Age-years	0.0774	0.1611	0.4806	0.6308	-0.2383	0.3931
Type-of-apartment	-0.0305	0.1591	-0.1920	0.8477	-0.3424	0.2813
Account-Balance_Some Balance	1.1046	0.2845	3.8830	0.0001	0.5471	1.6622
Payment-Status-of-Previous-Credit_Paid Up	0.2940	0.3223	0.9124	0.3616	-0.3376	0.9256
Payment-Status-of-Previous-Credit_Some Problems	-1.8177	0.6634	-2.7399	0.0061	-3.1179	-0.5174
Purpose_New car	1.3016	0.5689	2.2877	0.0222	0.1865	2.4167
Purpose_Other	1.1096	1.1077	1.0018	0.3165	-1.0614	3.2807
Purpose_Used car	0.5637	0.3945	1.4289	0.1530	-0.2095	1.3369
Value-Savings-Stocks_None	-0.0365	0.3970	-0.0919	0.9268	-0.8147	0.7417
Value-Savings-Stocks_£100-£1000	0.4740	0.4615	1.0271	0.3044	-0.4305	1.3785
Length-of-current-employment_4-7 yrs	0.3373	0.4260	0.7918	0.4285	-0.4977	1.1724
Length-of-current-employment_< 1yr	-0.2885	0.3247	-0.8885	0.3743	-0.9249	0.3479
No-of-Credits-at-this-Bank_More than 1	0.5238	0.3467	1.5109	0.1308	-0.1557	1.2033

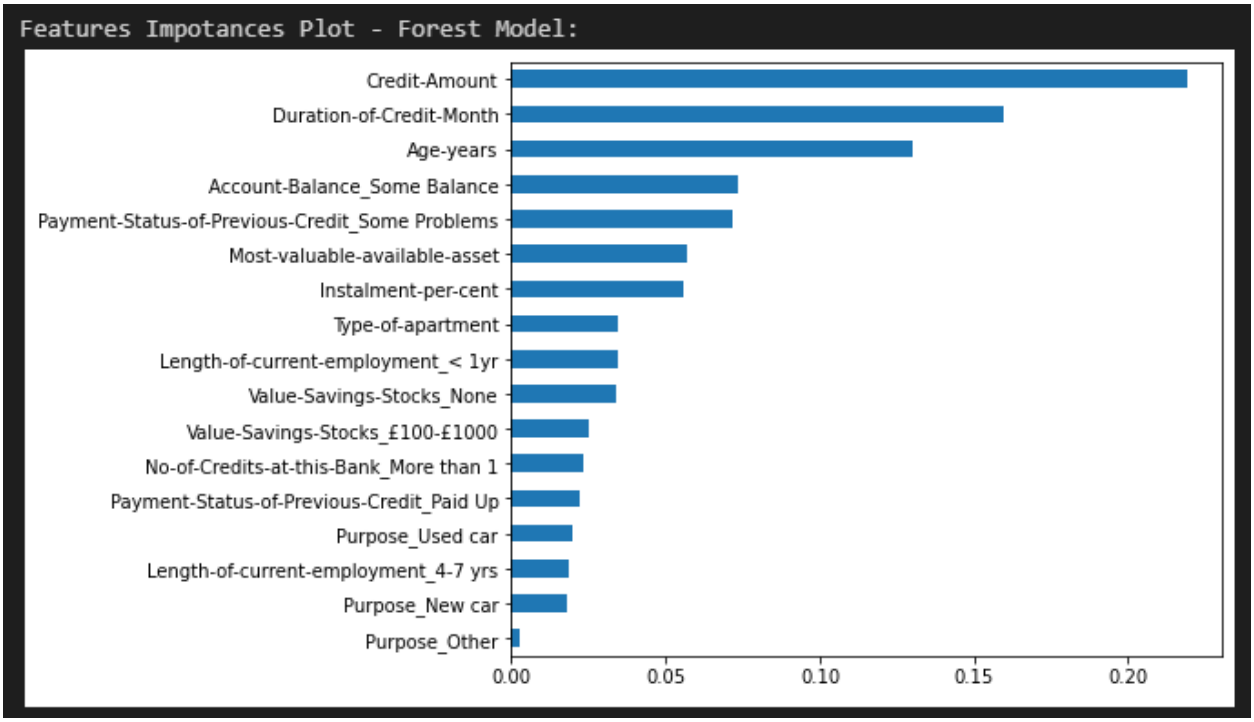
II. Decision Tree

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 age.



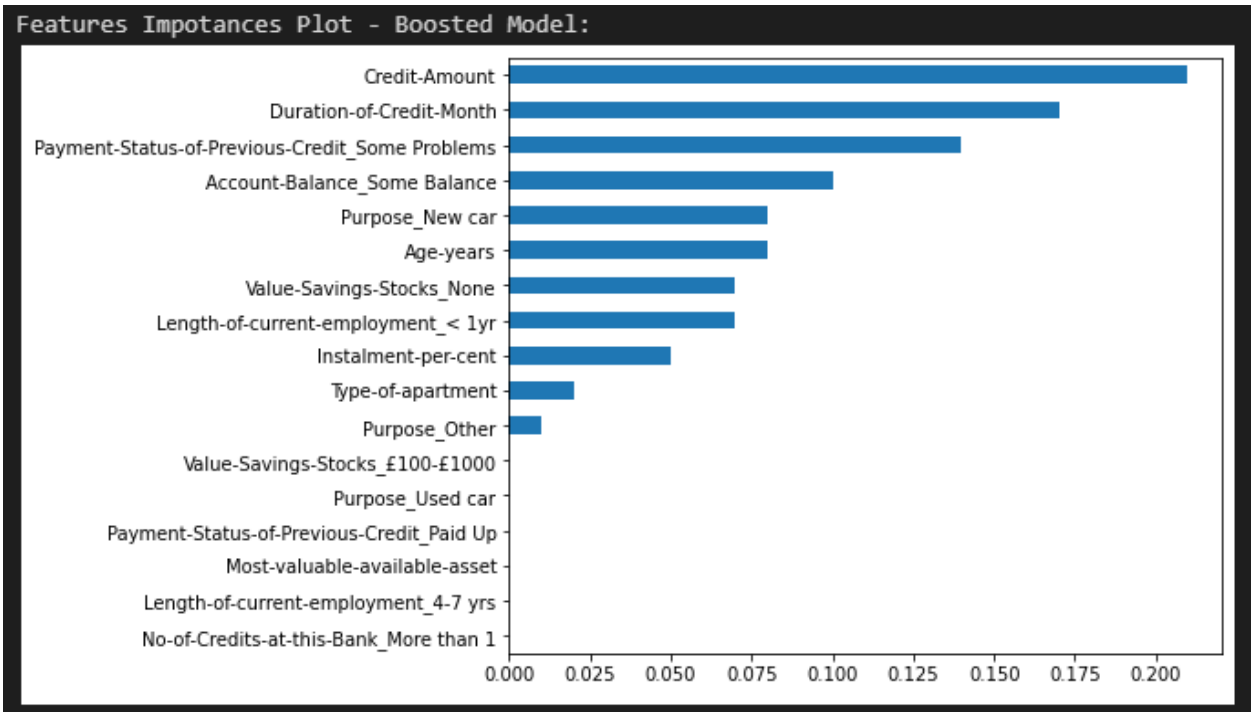
III. Forest Model

All predictor variables are importance



IV. Boosted Model

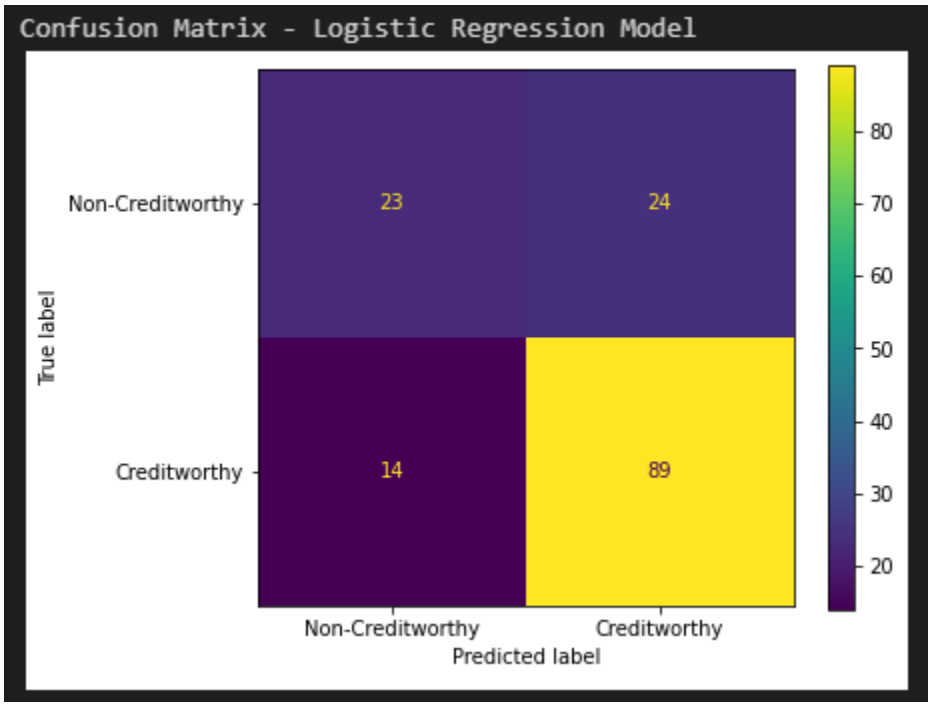
The most important predictor variables are credit amount, duration of credit, payment status of previous credit, account balance, purpose, age, value savings stocks, length of current employment, instalment percent, and type of apartment.



- 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?

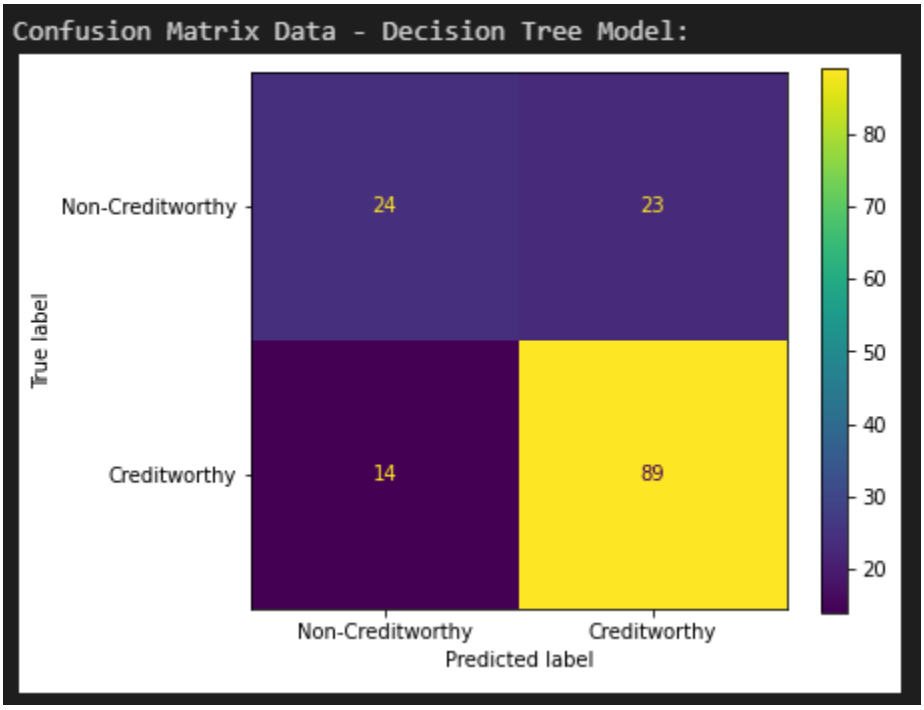
    I. Logistic Regression

    The logistic regression model has an accuracy of 75% in validation data.
    PPV= true positives \ (true positives + false positives) = 89 / (89+24) =.79
    NPV= true negatives\ (true negatives + false negatives) =23/ (23+14) = .62
    After checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.



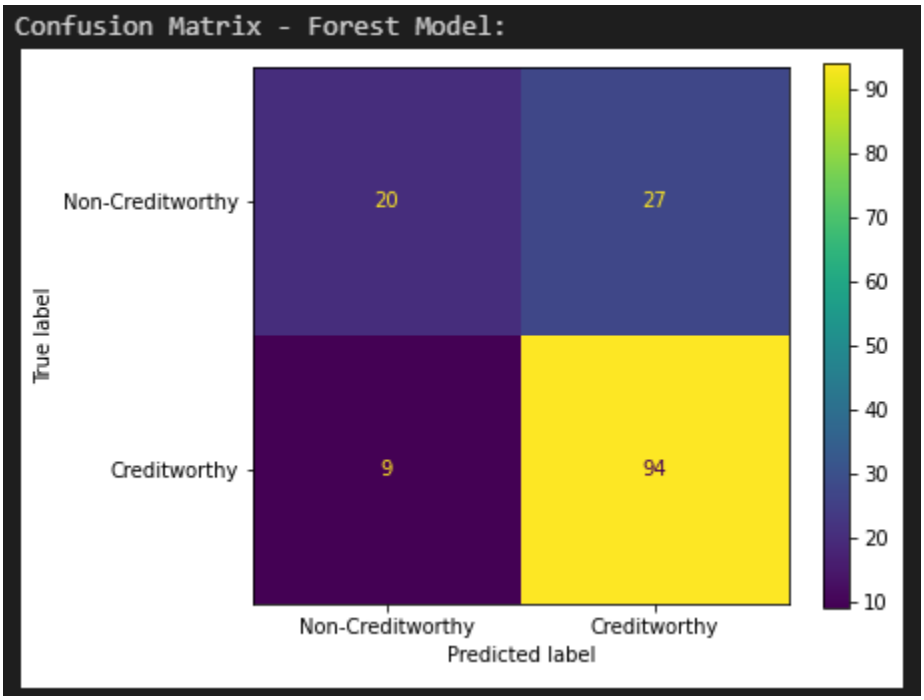
II. Decision Tree

The decision tree model has an accuracy of 75% in validation data.
PPV= true positives \ (true positives + false positives) = 89 / (89+23) =.79
NPV= true negatives\ (true negatives + false negatives) =24/ (24+14) = .63
After checking the confusion matrix there is bias seen in the model's prediction to Creditworthy.



III. Forest Model

The forest model has an accuracy of 76% in validation data.
PPV= true positives \ (true positives + false positives) = 94 / (94+27) =.78
NPV= true negatives\ (true negatives + false negatives) =20/ (20+9) = .68
After checking the confusion matrix there is no bias seen in the model's prediction.



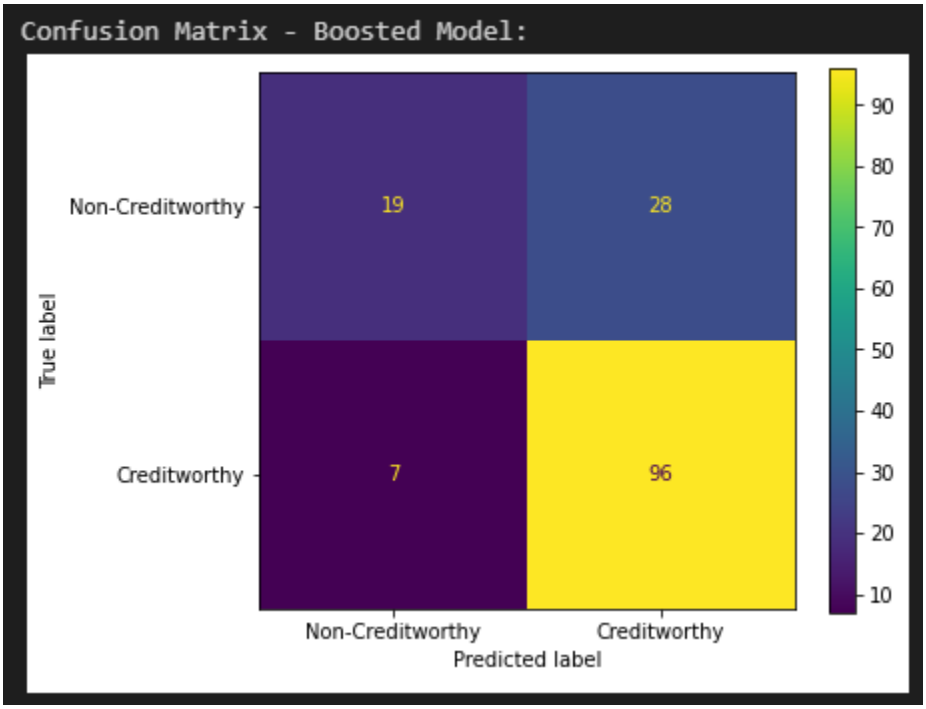
IV. Boosted Model

The boosted model has an accuracy of 77% in validation data.

PPV= true positives \ (true positives + false positives) = 96 / (96+28) =.77

NPV= true negatives\ (true negatives + false negatives) = 19/ (19+7) = .73

After checking the confusion matrix there is no bias seen in the model's prediction.



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

Step 4: Writeup

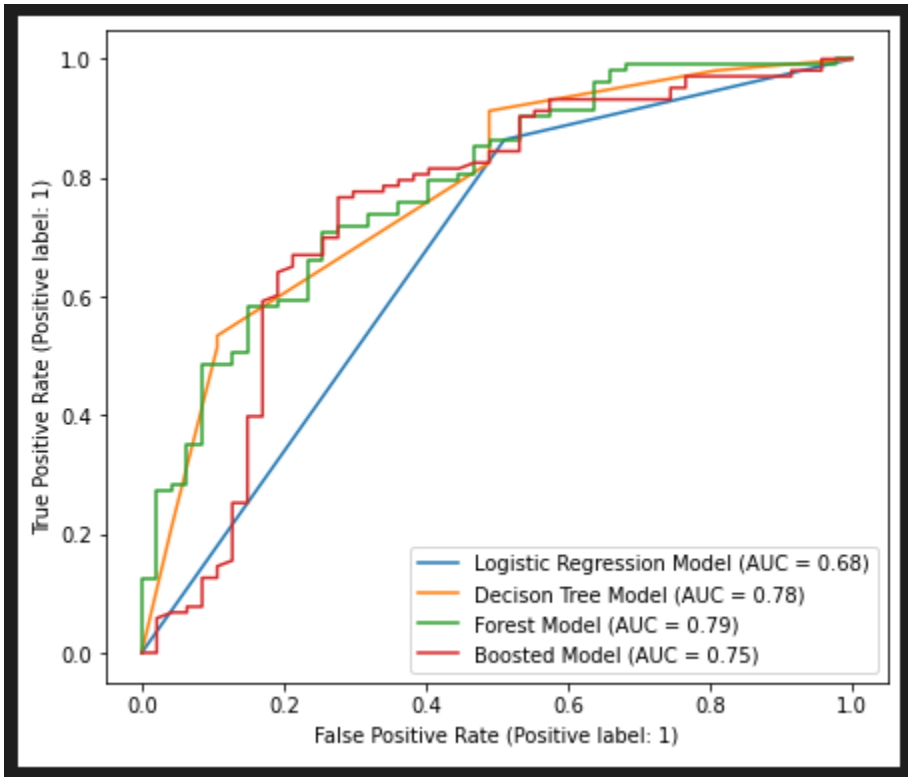
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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 forest model due to high AUC value of 0.79.

  - ○ ROC graph



  - ○ Bias in the Confusion Matrices
    PPV= true positives \ (true positives + false positives) = 94 / (94+27) =.78
    NPV= true negatives\ (true negatives + false negatives) =20/ (20+9) = .68
    After checking the confusion matrix there is no bias seen in the model's prediction.

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 433