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?
  Answer: The business decision that needs to be made here is to determine amongst a new set of customers those whom it will be profitable for the business to approve loans to.

- What data is needed to inform those decisions?
  Answer: Data like age, income, loan amount, loan duration of existing customers who have applied for loans in the past, their creditworthiness status that can be used to build a model and data on the new customers who are applying for loan.

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

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".
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
  the Median. I picked the median as the imputation value because the data is skewed to the right, so the mean would not have an accurate representation of the data.
- 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.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

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:*

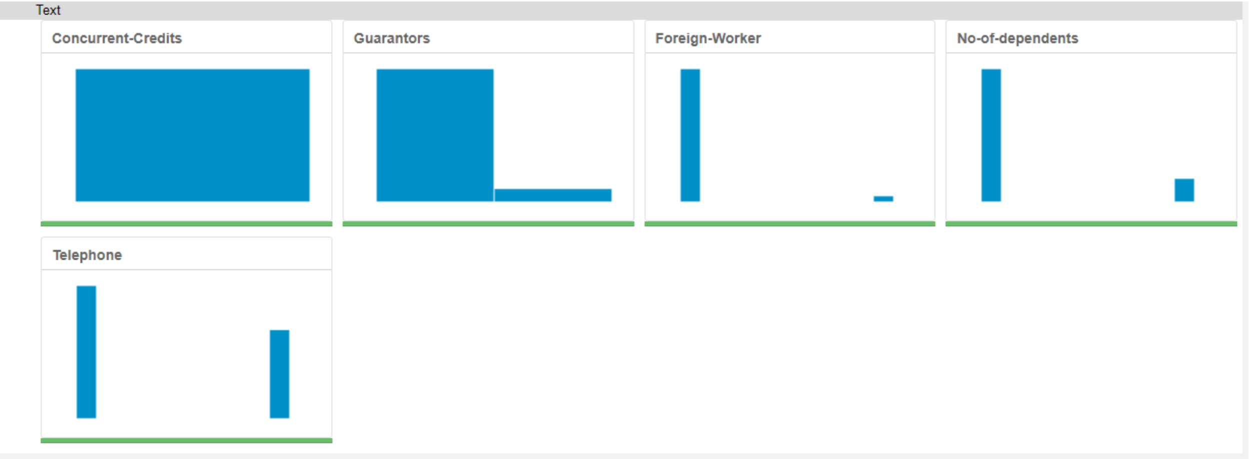
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address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

Answer: I dropped 7 fields during data cleanup. There are two fields with incomplete data in the dataset- The duration in current address and the age fields. During the cleanup process, I dropped the duration in current address field because it has more than more half of its data missing (69%) and data imputation will not be accurate. Age has 2% data missing so I imputed with the Median as specified. I dropped the other five fields due to low variation in the data namely foreign workers, guarantors, no of dependents, occupation and telephone as shown in their frequency charts below.



Step 3: Train your Classification Models

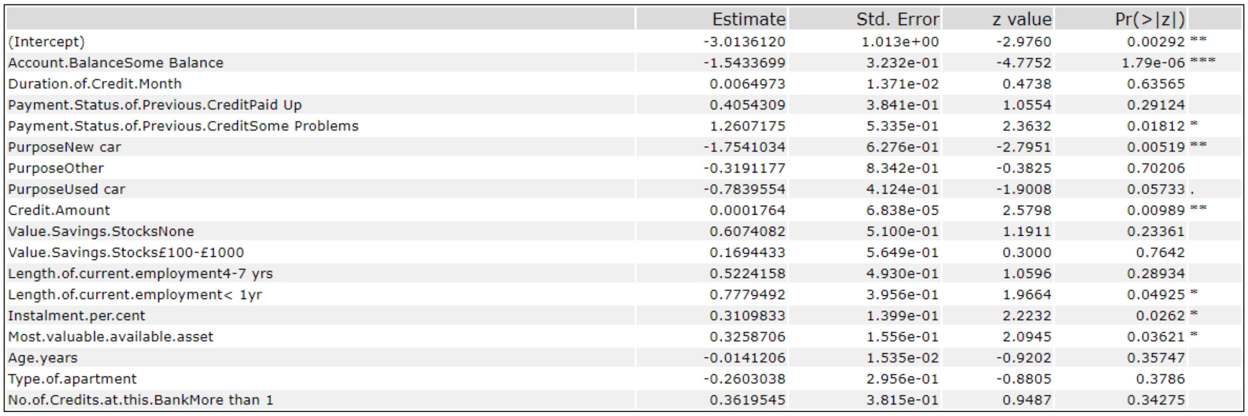
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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

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

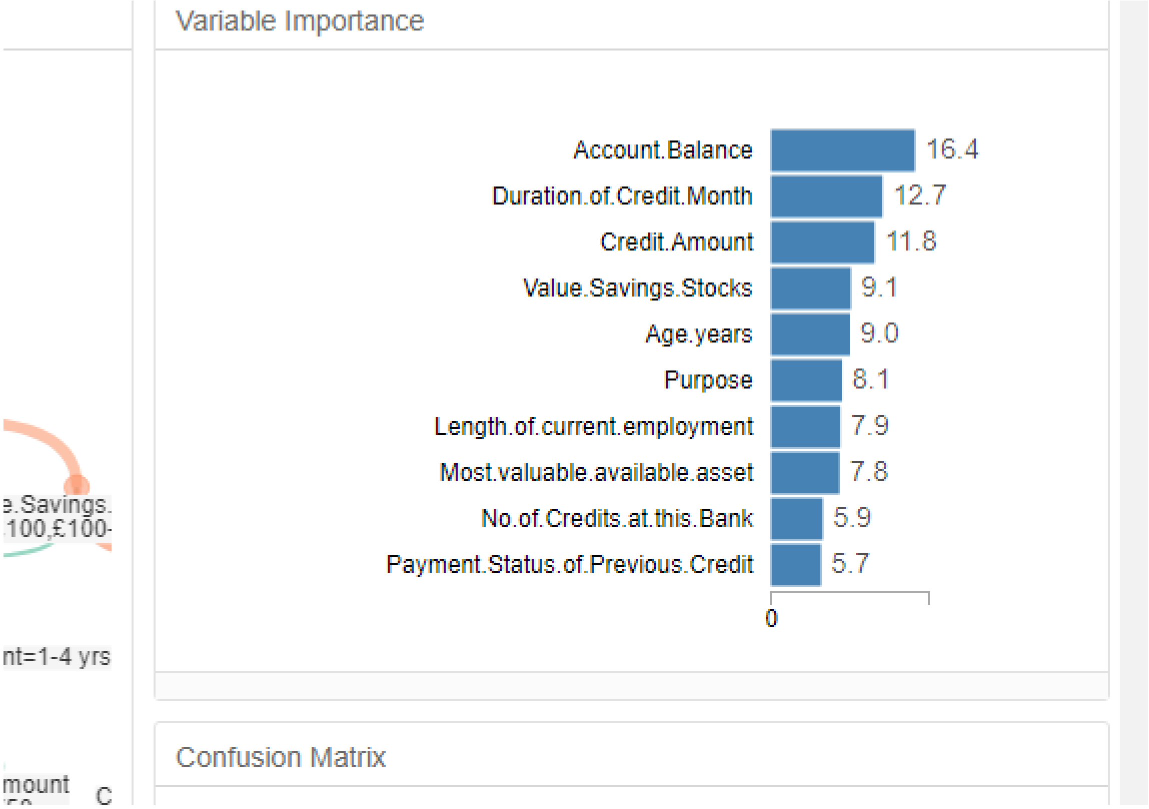
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  Answer:

Logistic Regression:The statistically significant features in the model are shown below. The features with the asterisks* are statistically significant. The higher the number of asterisks a feature has, the more statistically significant it is, which implies that Account.BalanceSome Balance is the most statistically important followed by PurposeNew car while features like Most.valuable.available.asset are not statistically significant.

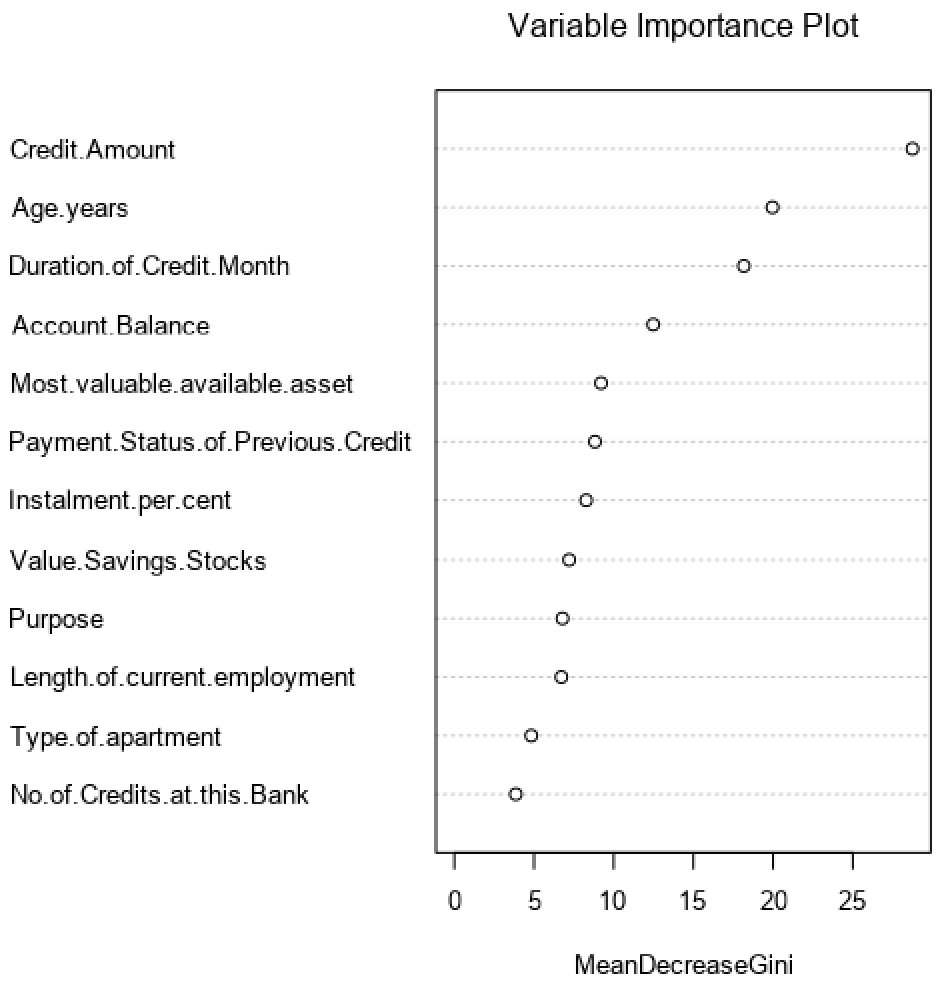


	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	

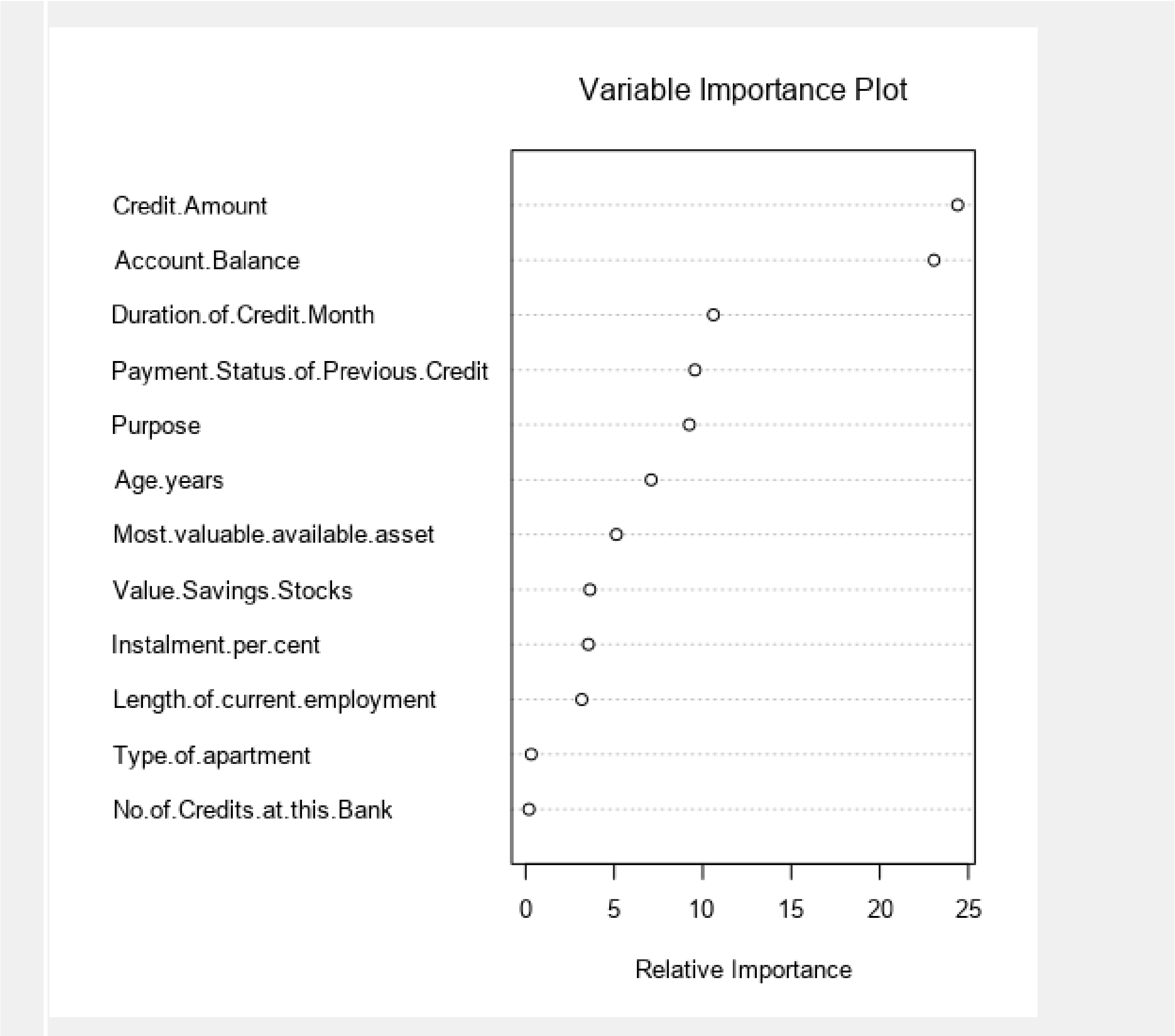
Decision tree: The variable importance chart below lists the features in order of importance to the model, the most important is Account balance, while the least important is Payment status of Previous Credit while some features like installment per cent were not used by the model at all.


Variable Importance
Account.Balance 16.4
Duration.of.Credit.Month 12.7
Credit.Amount 11.8
Value.Savings.Stocks 9.1
Age.years 9.0
Purpose 8.1
Length.of.current.employment 7.9
Most.valuable.available.asset 7.8
No.of.Credits.at.this.Bank 5.9
Payment.Status.of.Previous.Credit 5.7
0
e.Savings.
100,£100-
nt=1-4 yrs
mount
C
Confusion Matrix

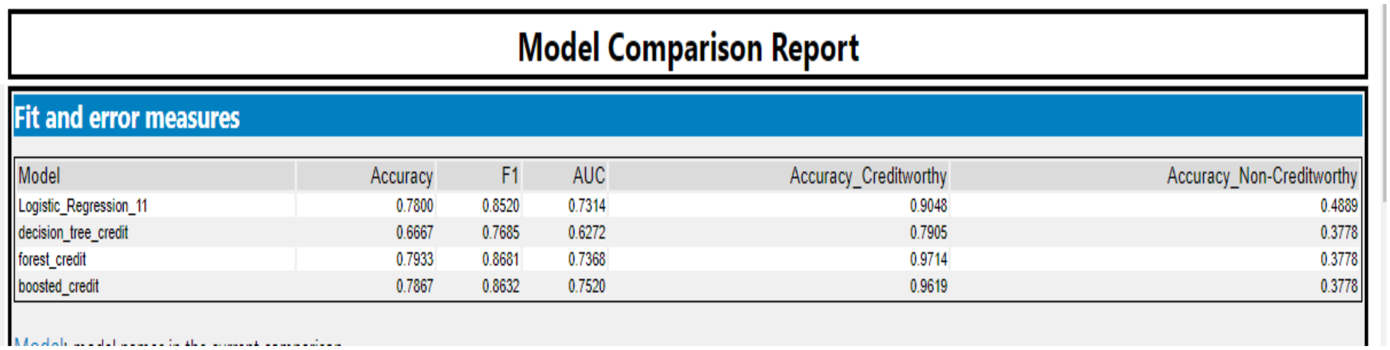
Forest Model: The variable importance plot for the forest model is shown below. The most important feature to the model is credit Amount, the least important field is occupation.



Boosted Model: The variable importance chart for the boosted model is shown below. The most important features to model are credit.Amount, account balance while the least importance are no of credits at this bank with the feature occupation not important to the model at all.



- 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?
  Logistic Regression: The logistic Regression has an accuracy at 78%, which is strong.
  PPV= true positives \ (true positives + false positives) = 95 / (95+23) =.81
  NPV= true negatives\ (true negatives + false negatives) =22/ (22+10) = .69



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_11	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree_credit	0.6667	0.7685	0.6272	0.7905	0.3778
forest_credit	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_credit	0.7867	0.8632	0.7520	0.9619	0.3778

Model: model names in the current comparison

Decision Tree model: The decision tree has the lowest accuracy at ~67%, not quite strong. The model is also biased at predicting creditworthiness because the difference between the PPV and NPV is quite high.
PPV= 92 / (92+23) =.75
NPV=22/ (22+17) = .44

Forest Model:This model also has a high accuracy at ~79%, does very well at predicting credithworthiness(the highest accuracy score), it also has the highest f1 score but not so good at predicting non credithworthiness. The evaluation of the confusion matrix below shows the model is not biased.
PPV= 102 / (102+28) =.79
NPV=17/ (17 + 3) = .85

Boosted Model: The boosted model has an accuracy of ~79%, a high precision rate at ~79% it has the least bias though as shown below.
PPV= 101 / (101+28) =.78
NPV=17/ (17+4) = .81

(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"

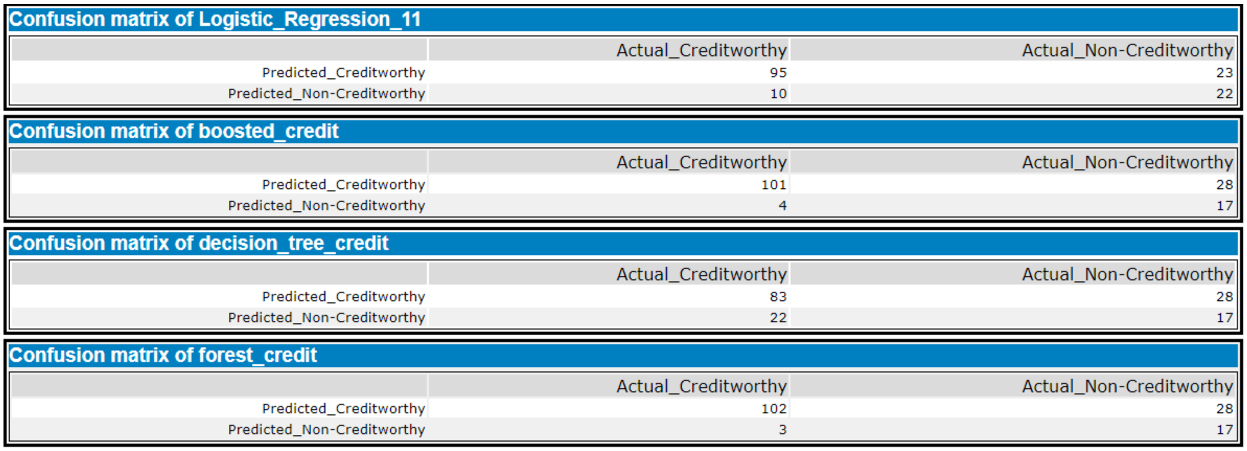
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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:
    - ○ Overall Accuracy against your Validation set
    - ○ Accuracies within "Creditworthy" and "Non-Creditworthy" segments

- ROC graph
- Bias in the Confusion Matrices

Answer: I chose the forest model because it has the highest accuracy of all the models. Although the model has a very high accuracy in its overall predictions, it does not have as high accuracy in its predictions of non-creditworthiness. With regards to the confusion matrix, as I have shown in my analysis above, there is not much difference between the NPV and PPV values which indicated that the model is not biased. In terms of the ROC graph (shown below), the model is arguably the highest and the area under the curve while not the highest (it is the second highest) is quite high. Because we are interested only in the accuracy for this problem, the forest model is the best fit.



Confusion matrix of Logistic_Regression_11

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Confusion matrix of boosted_credit

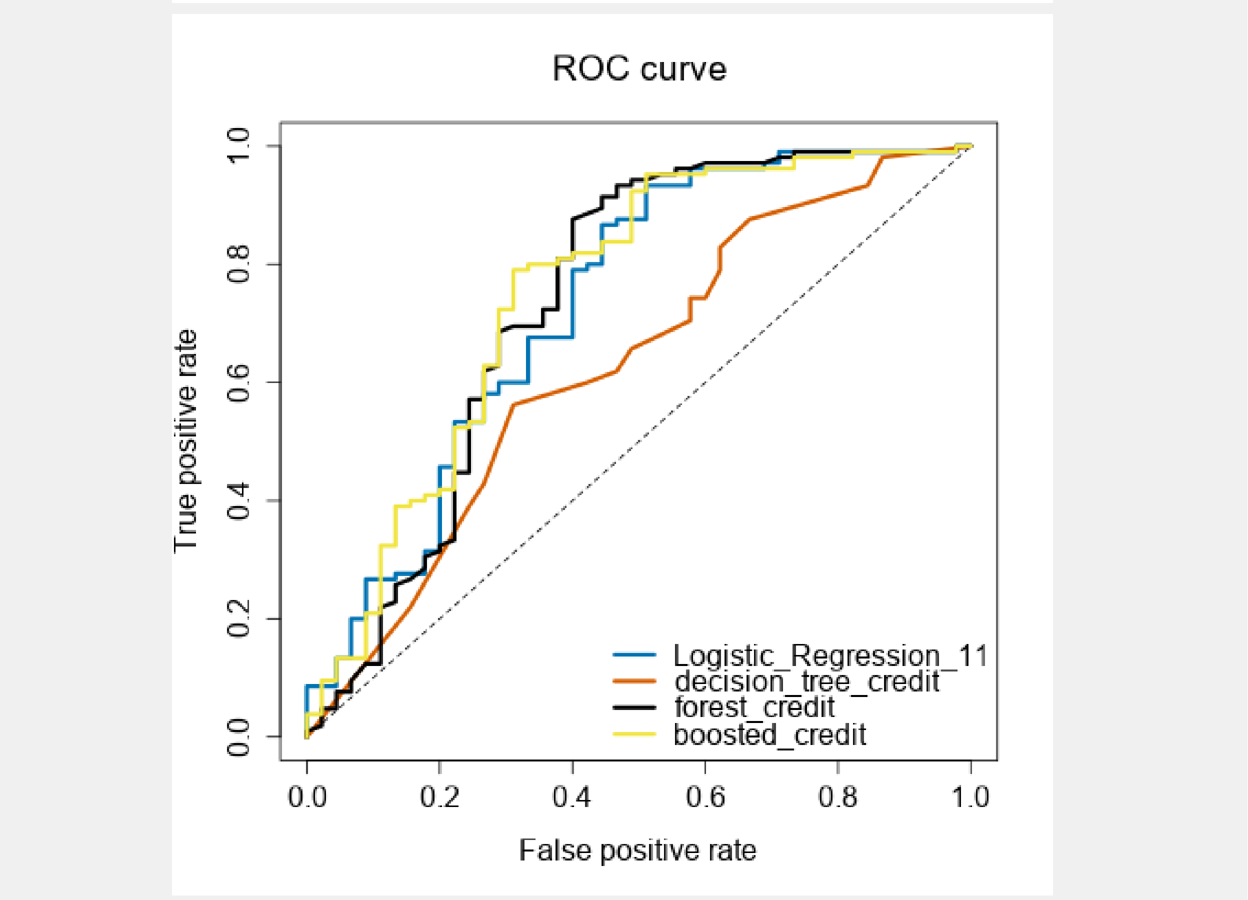
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of decision_tree_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of forest_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17



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

- How many individuals are creditworthy?
  Answer: 408 people

<u>Before you Submit</u>

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.