Project: Creditworthiness **BERNS**

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

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?
  Based on a bunch of data from customers, we want to identify and predict which are creditworthy or non-creditworthy.

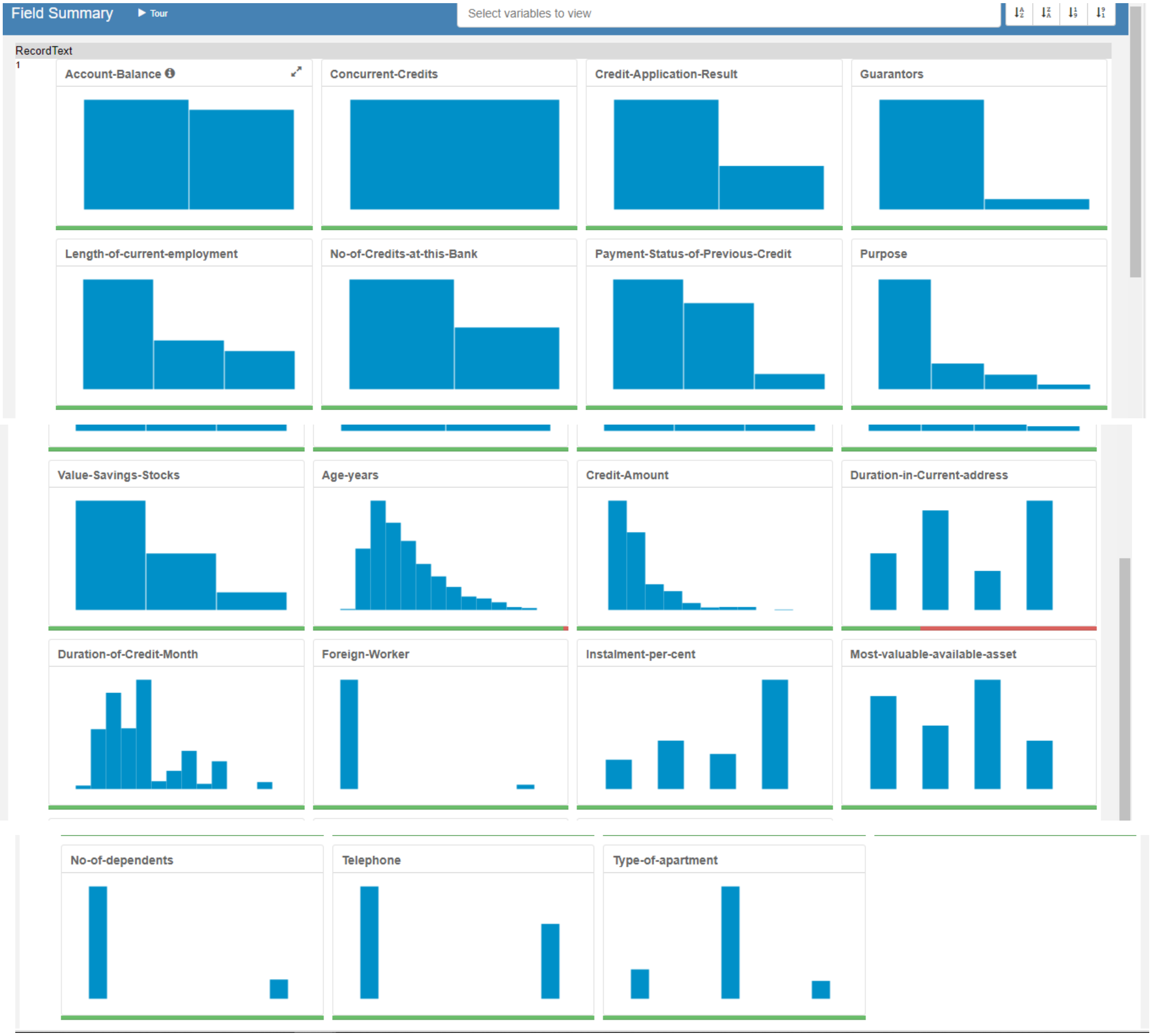
- What data is needed to inform those decisions?
  Using the dataset of customers/past applications who already had a loan from the bank, we will make a predictive model. Using this model we can then score other customers to see if there are creditworthy or not. Some of the important data we can use to build our model includes: account balances, credit amount, Purpose, value savings stocks, age, duration of credit month, and payment status of previous credit.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Since the answer to this business problem is one of two choices, it is a binary classification problem.

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.***

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.



When I summarized the data fields I removed fields and imputed one field. I removed Occupation and Concurrent credits, due to them both having only one type of data. I then removed Guarantors, # Of Dependents, and Foreign Worker due to low variability (where more than 80% of the data looks skewed to one side). I also removed the Duration in current address due to 69% of the values missing, and finally removed the telephone

column because telephone numbers are irrelevant for the nature of what we are trying to figure out.
I kept Age-Years even though it has only 2% of the data missing. I imputed the missing data with the median age instead of the mean because in the graph above, the data in the Age-Years column looks skewed to the left. The median should fix that.

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

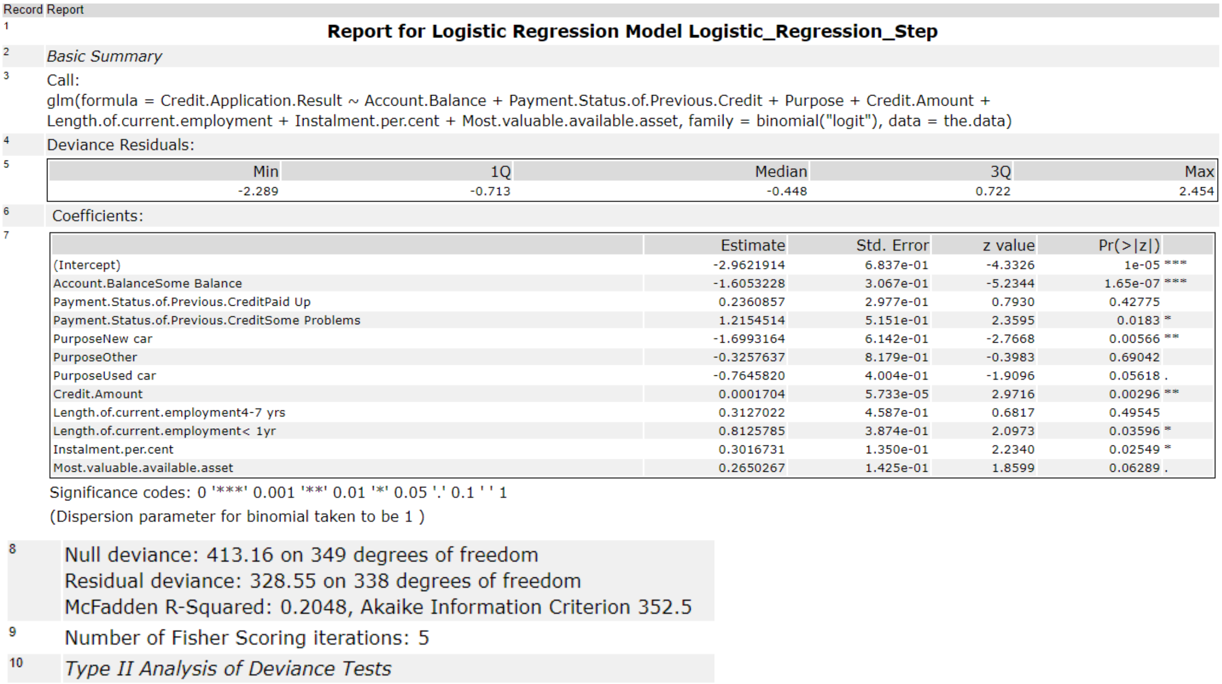
*Answer these questions for **each model** you created:*

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

- 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?

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

LOGISTIC/STEPWISE REGRESSION



Record Report

1 **Report for Logistic Regression Model Logistic_Regression_Step**

2 *Basic Summary*

3 Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

6 Coefficients:

7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183 *
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566 **
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618 .
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596 *
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549 *
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .

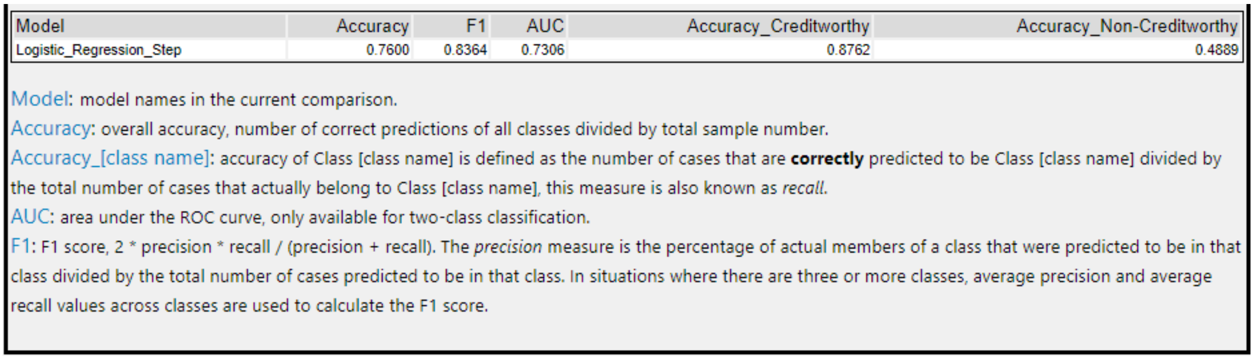
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

I used Credit Application Result as my target variable, then when I performed the regression, the most significant variables with p value less than .05 were Account Balance, PurposeNewCar, and Credit Amount.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_Step	0.7600	0.8364	0.7306	0.8762	0.4889

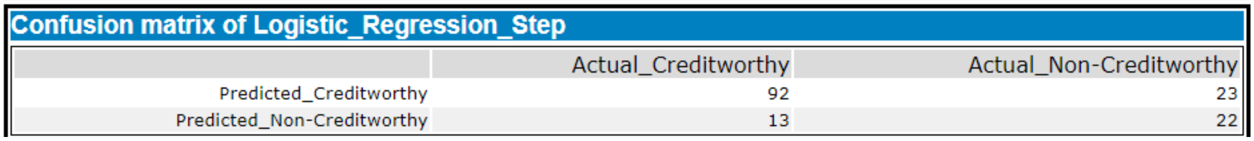
Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

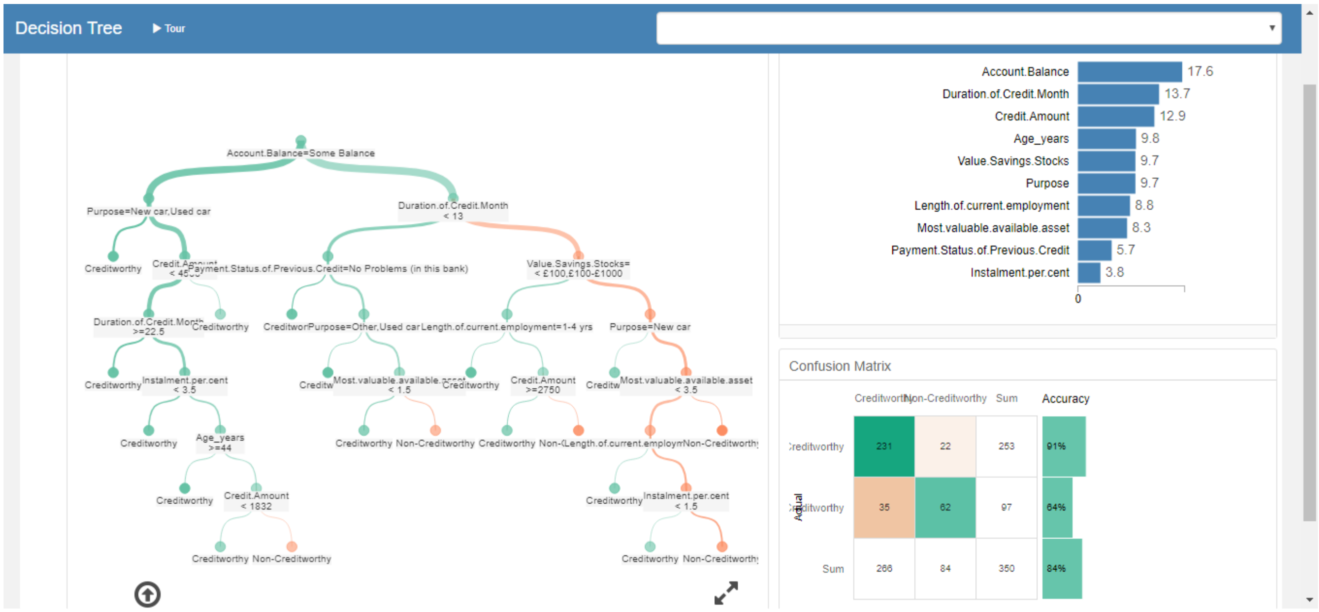


Confusion matrix of Logistic_Regression_Step

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Using the model comparison report, the overall accuracy was 76%. The accuracy for predicting Creditworthy at 87% is higher than Non Creditworthy at 48%. But looking at the confusion matrix, it seems as though this model shows bias or predicting someone non creditworthy.

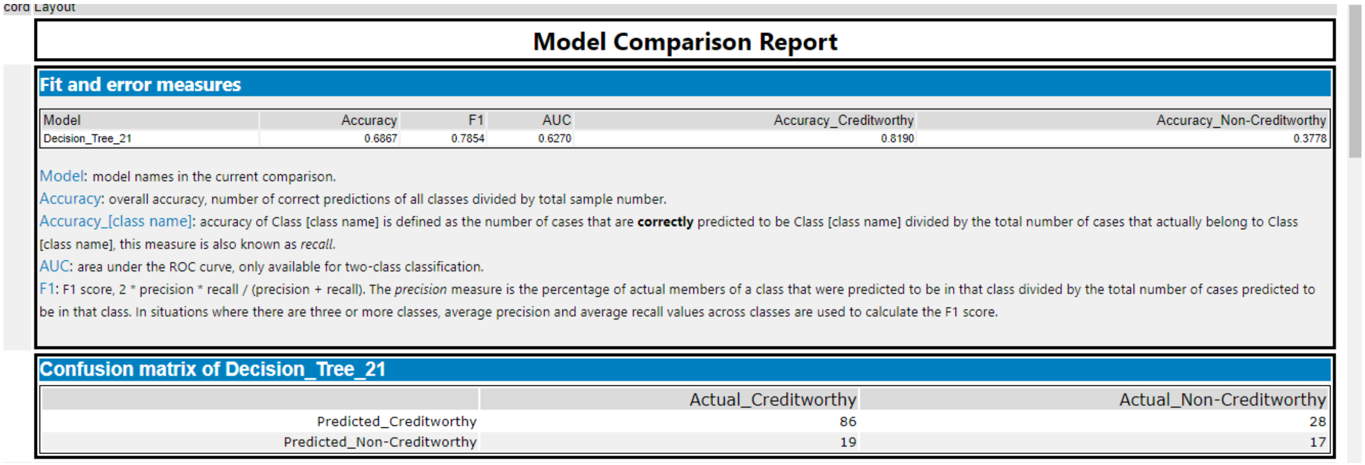
DECISION TREE MODEL



Confusion Matrix

Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	231	22	253	91%
editworthy	35	62	97	64%
Sum	266	84	350	84%

From the variable importance graph, we can see the significant predictor variables which are: Account Balance, Duration of Credit Month, and Credit Amount.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_21	0.6867	0.7854	0.6270	0.8190	0.3778

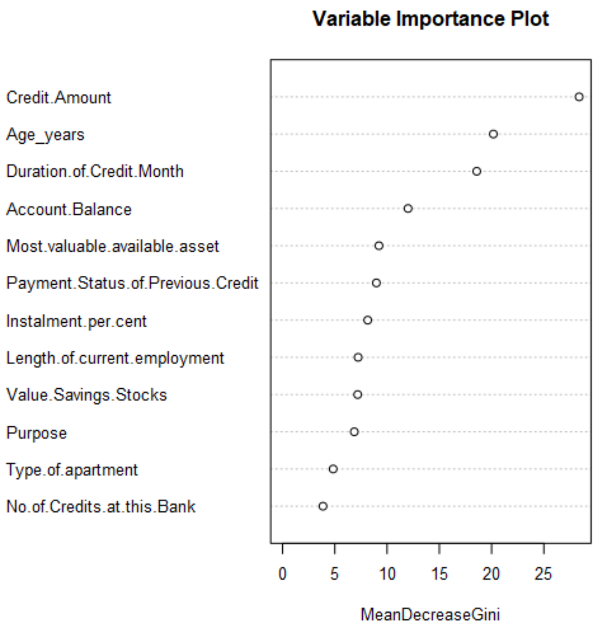
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Decision_Tree_21

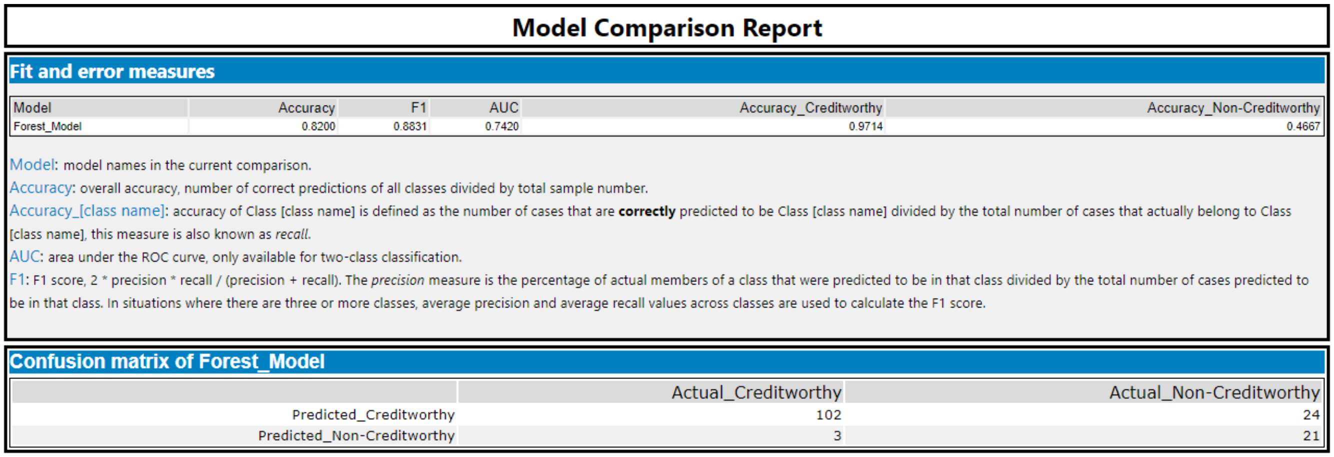
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	28
Predicted_Non-Creditworthy	19	17

From the model comparison report, we can see that this model has an overall accuracy of 68.67%. It has a 81.9% accuracy of predicting someone as Creditworthy and an accuracy of 37.78% of predicting someone as Noncreditworthy. As with the Logistic Regression, it seems as though when looking at the confusion matrix it still has bias of predicting people as Non Creditworthy.

FOREST MODEL



Using the variable importance plot of the Forest Model, we can see that the top 3 important predictor variables are Credit Amount, Age-Years, and Duration Of Credit Month.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8200	0.8831	0.7420	0.9714	0.4667

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

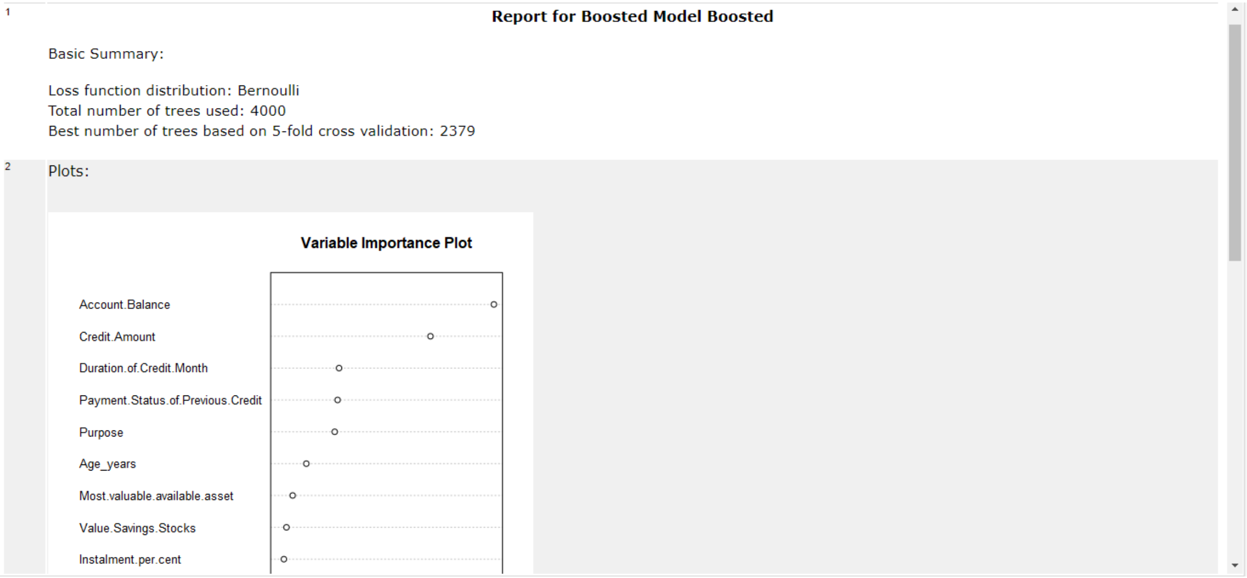
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Using the model comparison report, we can see that this has an overall accuracy of 82%. It has a accuracy of Creditworthy of 97%, and an accuracy of 46% of Non Creditworthy. Looking at the confusion matrix, it seems to be more biased of predicting people as non credit worthy.

BOOSTED MODEL



1

Report for Boosted Model Boosted

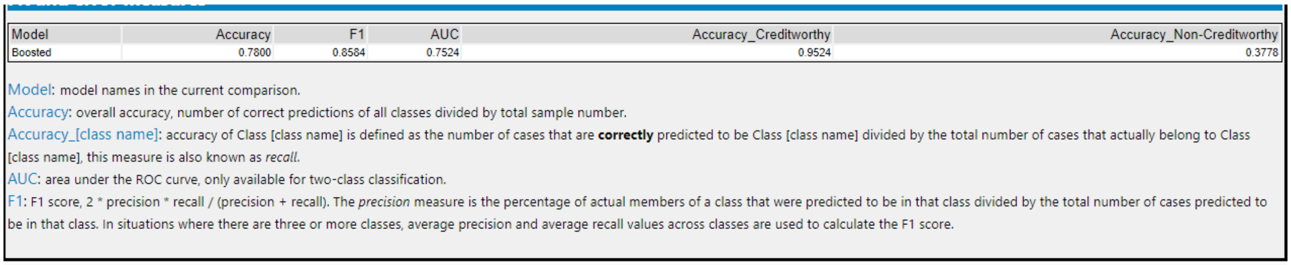
Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2379

2

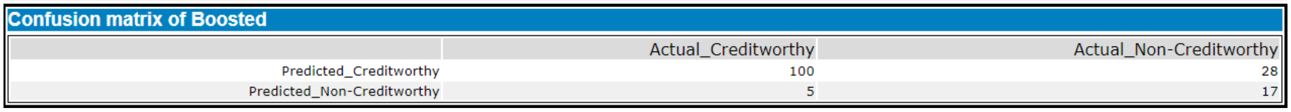
Plots:

Using the variable importance plot of the Forest Model, we can see that the top 3 important predictor variables are Account Balance, Credit Amount, and Duration Of Credit Month.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted	0.7800	0.8584	0.7524	0.9524	0.3778

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.



Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Using the model comparison report, we can see that this has an overall accuracy of 78%. It has an accuracy of Creditworthy of 95%, and an accuracy of 37% of Non Creditworthy. Once again, this model shows bias of predicting people as Non Creditworthy.

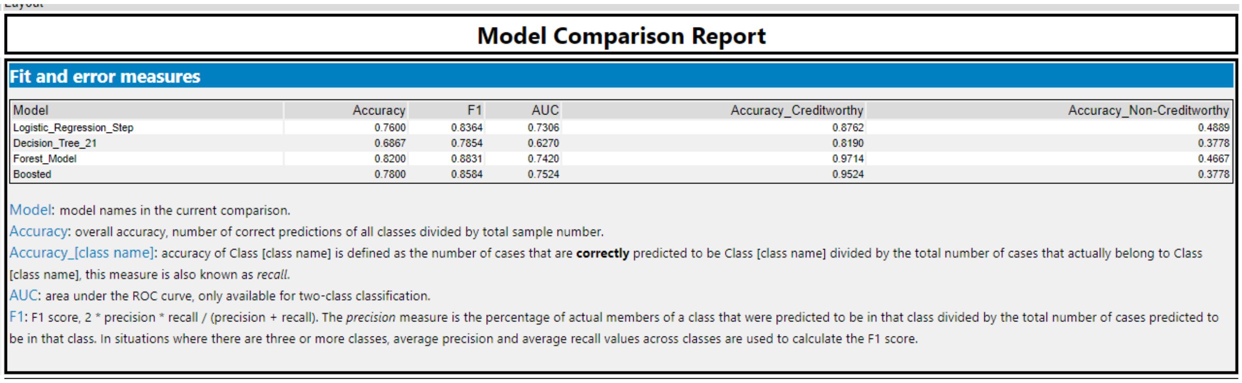
Step 4: Writeup

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



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_Step	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree_21	0.6867	0.7854	0.6270	0.8190	0.3778
Forest_Model	0.8200	0.8831	0.7420	0.9714	0.4667
Boosted	0.7800	0.8584	0.7524	0.9524	0.3778

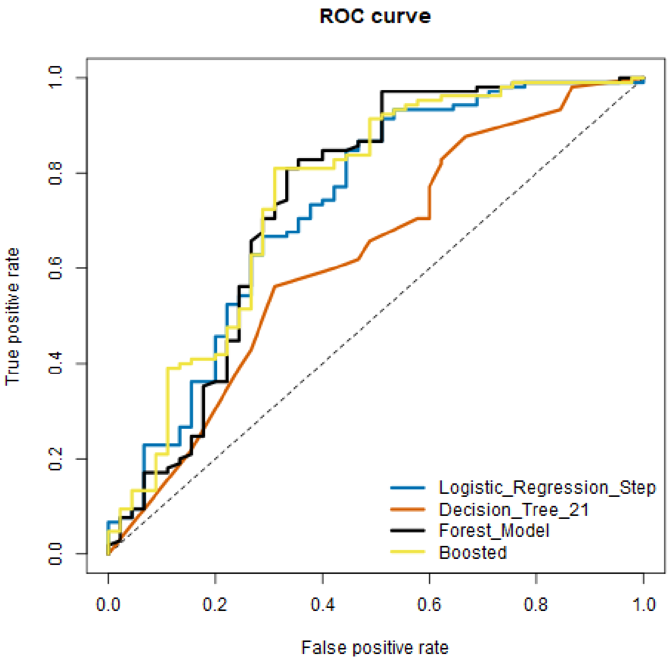
Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

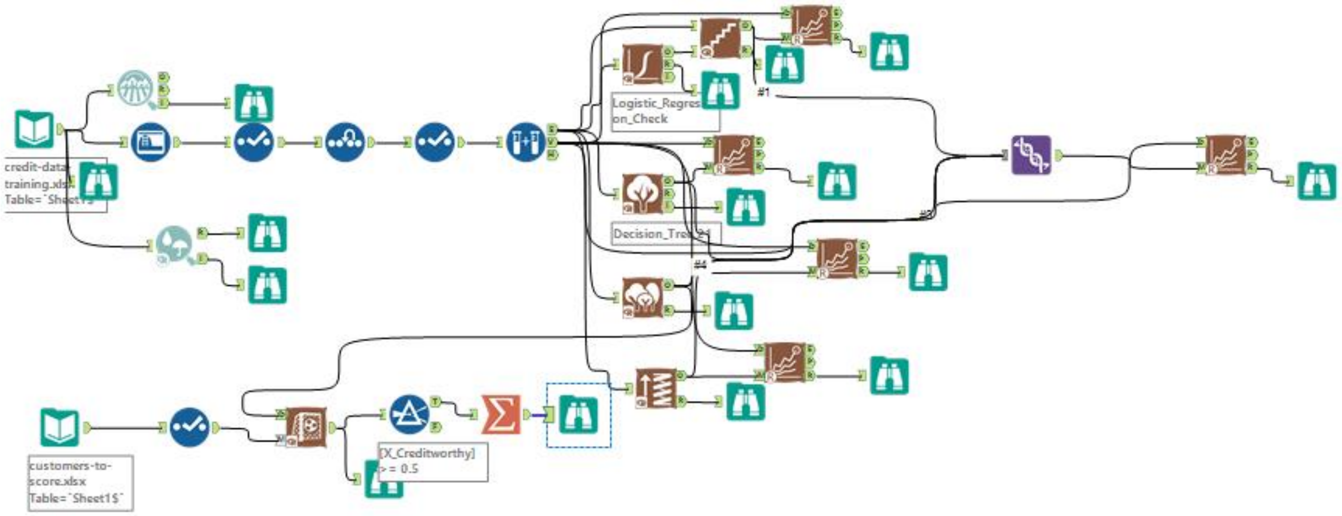


I chose the Forest Model, as it had the highest accuracy of any model at 82%. I will also choose this for my prediction analysis. This model also had the least discrepancies between creditworthy and non-creditworthy.
We can also see from the ROC curve, the Forest Model reached the positive rate the quickest compared to the other models.

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

- How many individuals are creditworthy?

From my results from the Forest Model, I predicted that 408 customers would be credit worthy.


credit-data-training.xlsx
Table=`Sheet1$`
Logistic_Regres on_Check
#1
Decision_Tree#1
#2
#4
customers-to-score.xlsx
Table=`Sheet1$`
[X_Creditworthy] >= 0.5