Project: Creditworthiness

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

1. What decisions needs to be made?

   A customer's credit worthiness must be predicted so we can process loan applications to meet the demand of 500 applications per week. We'd like to use classification modeling to systematically evaluate new loan applications.

2. What data is needed to inform those decisions?

   In order to accurately predict the credit worthiness of a new customer, we need past data for loan customers where the credit worthiness has been determined. We also need various attributes about a customer collected in their loan application and their time as loan customers. For example, these could be their assets, income, employment status and previous payment history. We will use tools like correlation matrix, p-values, stepwise regression, or importance plot to find the predictors with statistical significance.

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

   A binary model would help predict the best outcome because we are trying to provide a Boolean answer of 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".*

- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*
- *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 average 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:*

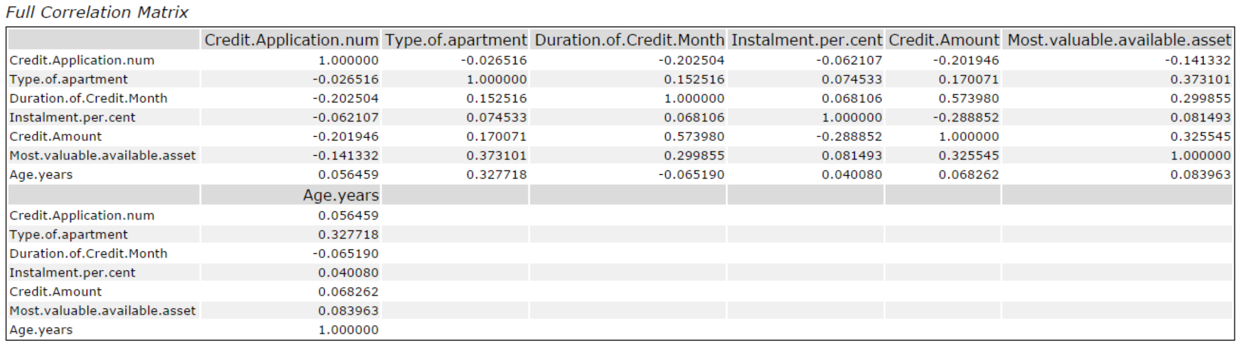
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address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

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

No predictors were removed as a result of multicollinearity. The predictors did not show a high correlation with each other.



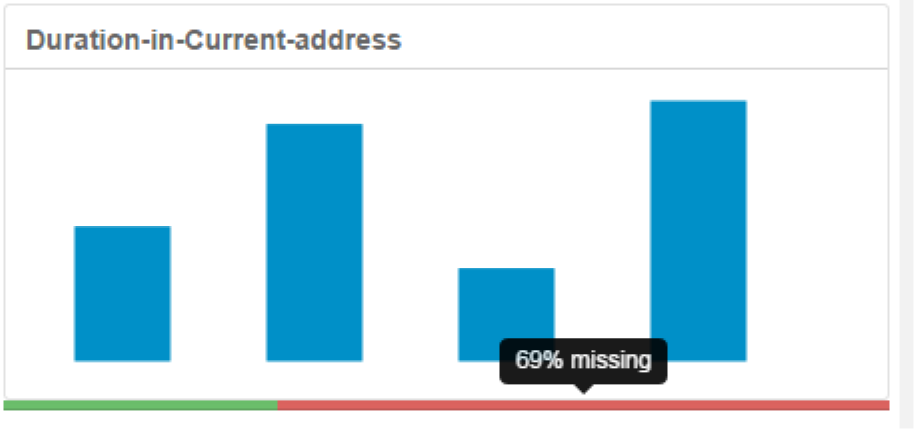
Full Correlation Matrix

	Credit.Application.num	Type.of.apartment	Duration.of.Credit.Month	Instalment.per.cent	Credit.Amount	Most.valuable.available.asset	Age.years
Credit.Application.num	1.000000	-0.026516	-0.202504	-0.062107	-0.201946	-0.141332	0.056459
Type.of.apartment	-0.026516	1.000000	0.152516	0.074533	0.170071	0.373101	0.327718
Duration.of.Credit.Month	-0.202504	0.152516	1.000000	0.068106	0.573980	0.299855	-0.065190
Instalment.per.cent	-0.062107	0.074533	0.068106	1.000000	-0.288852	0.081493	0.040080
Credit.Amount	-0.201946	0.170071	0.573980	-0.288852	1.000000	0.325545	0.068262
Most.valuable.available.asset	-0.141332	0.373101	0.299855	0.081493	0.325545	1.000000	0.083963
Age.years	0.056459	0.327718	-0.065190	0.040080	0.068262	0.083963	1.000000

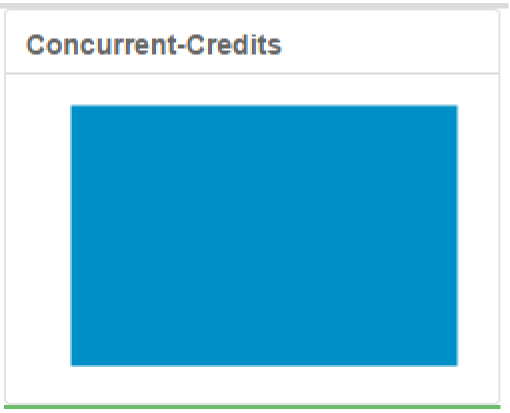
Most of the columns that were removed because they only had a single value. Duration in current address was removed because it had a large number of nulls.

Age was imputed with 36 in cases where it was null.

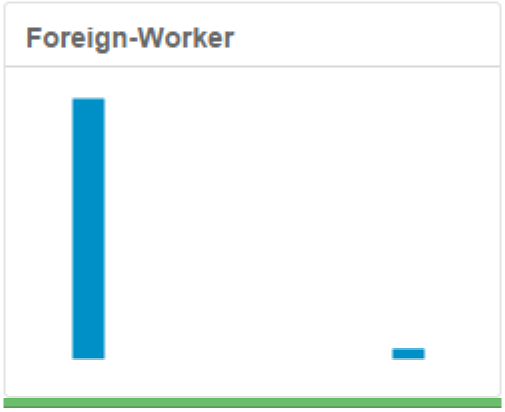
Duration in current address was removed. 69% of the values were null.

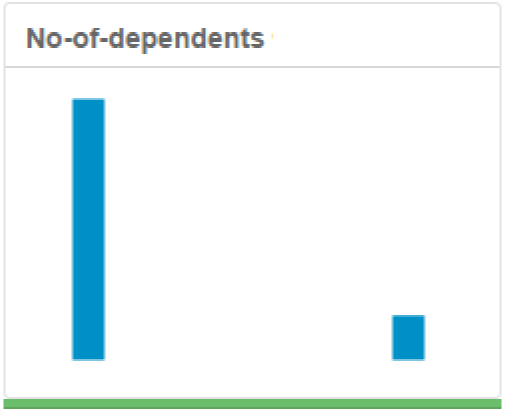


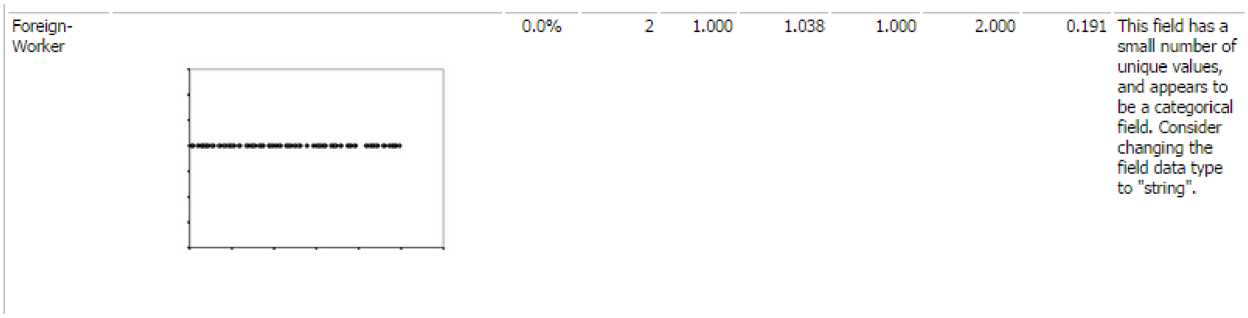
Concurrent-Credits was removed because it only had one value.



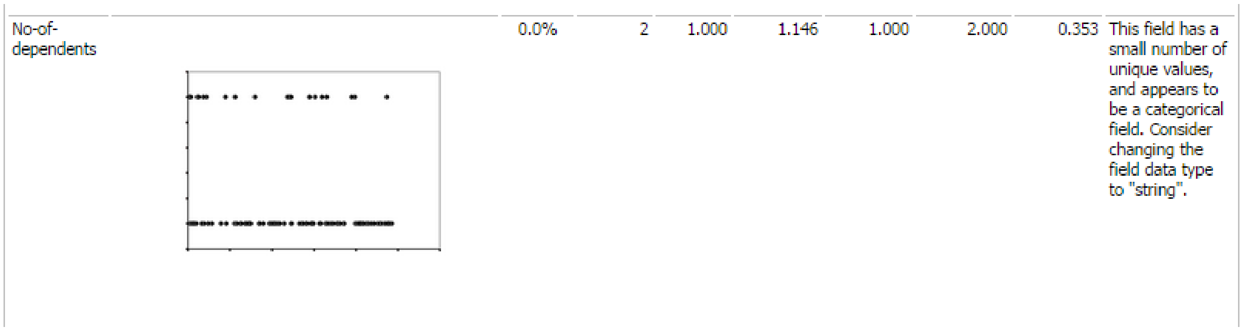
Foreign-Worker and No-of-dependents was removed because a majority of the values were the same.





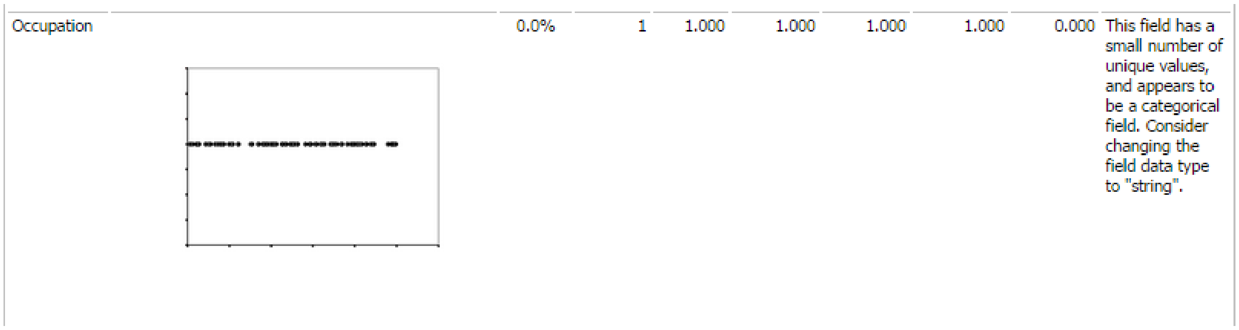


Foreign-Worker		0.0%	2	1.000	1.038	1.000	2.000	0.191	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".



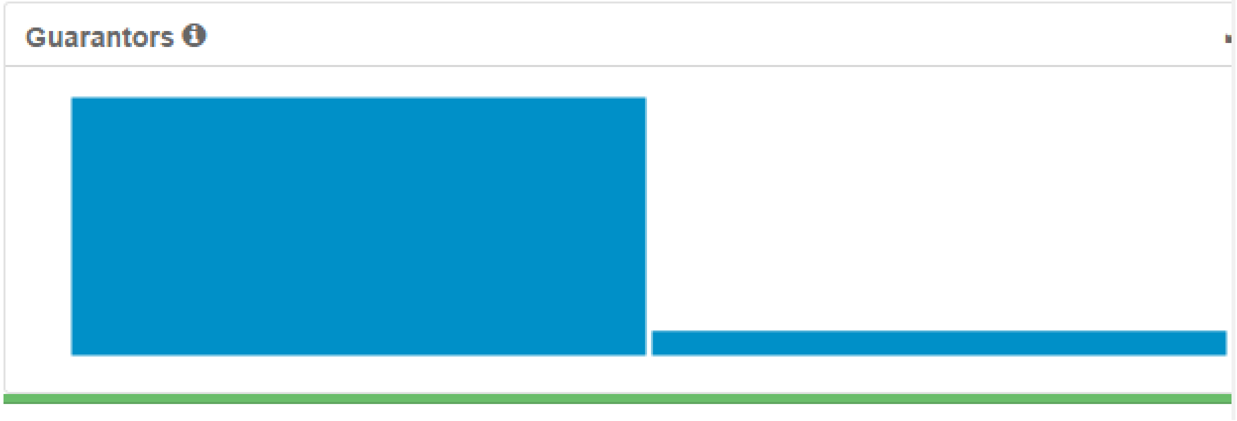
No-of-dependents		0.0%	2	1.000	1.146	1.000	2.000	0.353	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Occupation was removed because all loans applicants had a job.

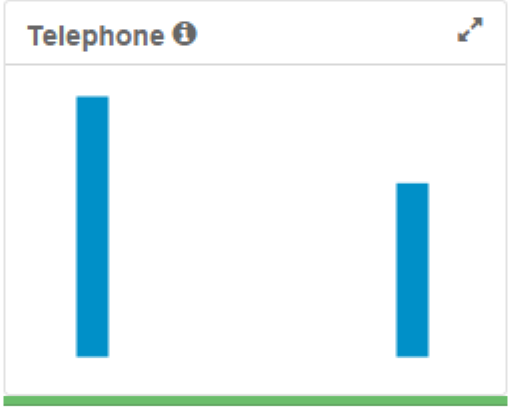


Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Guarantors was removed because there were only two unique values and most of these were none.



Telephone was removed because it only had two values and didn't make sense to use as a predictor.



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:*

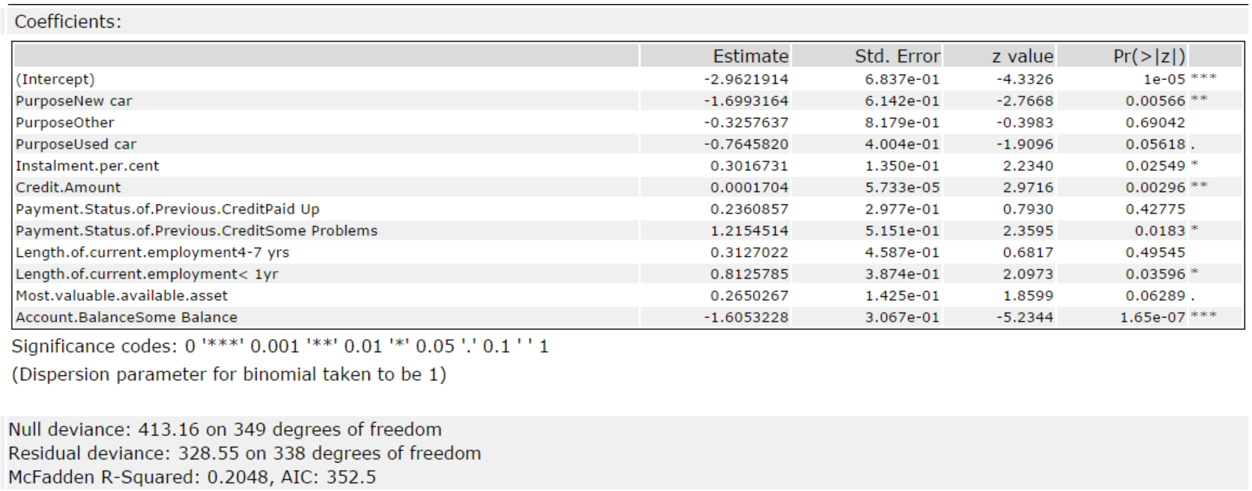
1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

2. 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 Regression

Using the stepwise tool with logistic regression, we have an overall accuracy of 76.00%. It was observed that the r^2 value was .2048. The most valuable predictor was having an account balance of some balance



Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
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 .
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549 *
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
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 *
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 *
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***

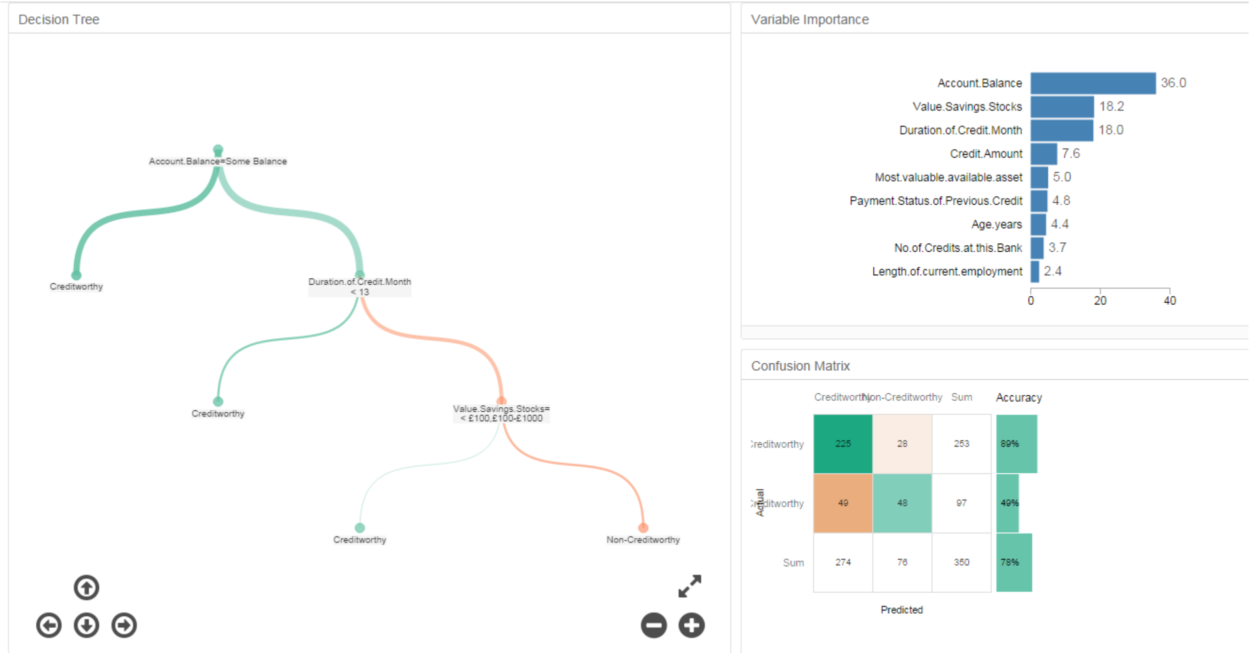
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

Decision Tree

In the decision tree model, we have an overall accuracy of 74.67%. The most significant predictors for that model are account balance, value savings stocks, and duration of credit month.



Decision Tree
Account.Balance=Some Balance
Creditworthy
Duration.of.Credit.Month < 13
Creditworthy
Value.Savings.Stocks= < £100,£100-£1000
Creditworthy
Non-Creditworthy
Variable Importance
Account.Balance 36.0
Value.Savings.Stocks 18.2
Duration.of.Credit.Month 18.0
Credit.Amount 7.6
Most.valuable.available.asset 5.0
Payment.Status.of.Previous.Credit 4.8
Age.years 4.4
No.of.Credits.at.this.Bank 3.7
Length.of.current.employment 2.4
0
20
40

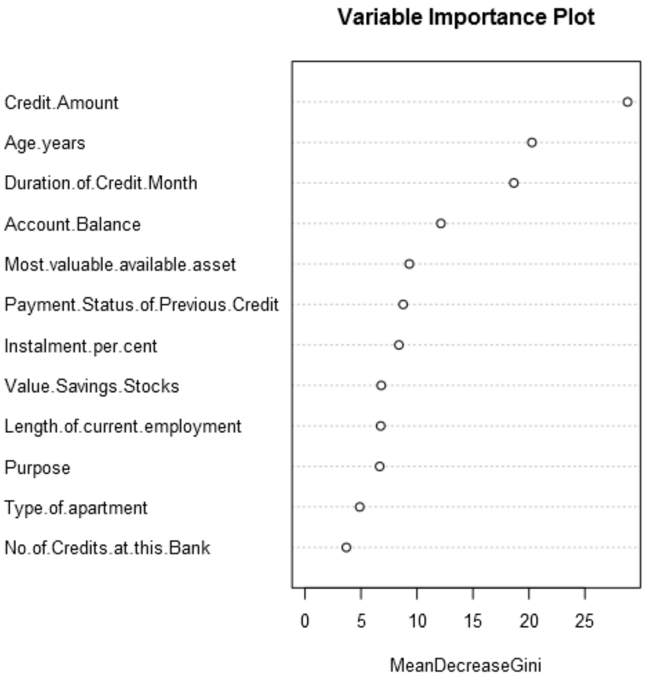
Confusion Matrix

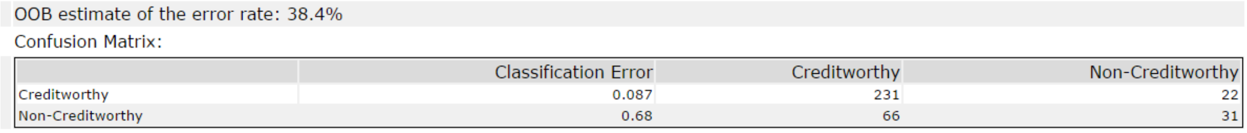
Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	225	28	253	89%
reditworthy	49	48	97	49%
Sum	274	76	350	78%

Predicted

Forest Model

In the forest model, the important variables are credit amount, age years, and duration of credit month. The accuracy was 81.33%.



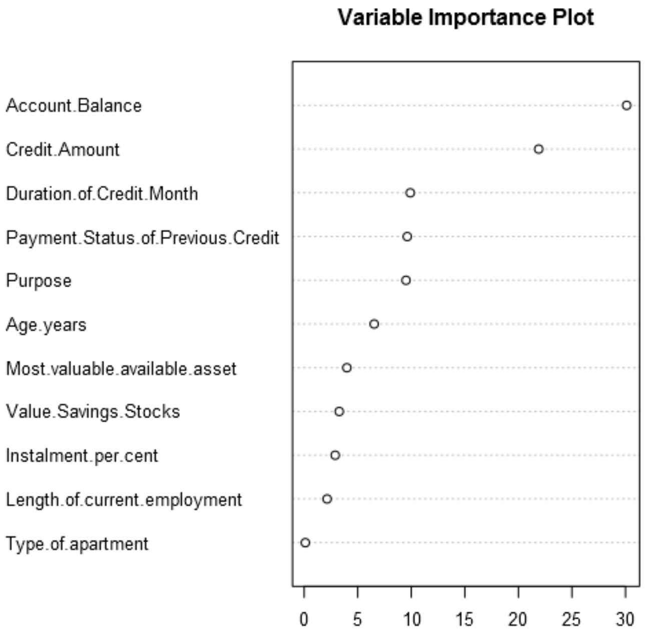

OOB estimate of the error rate: 38.4%

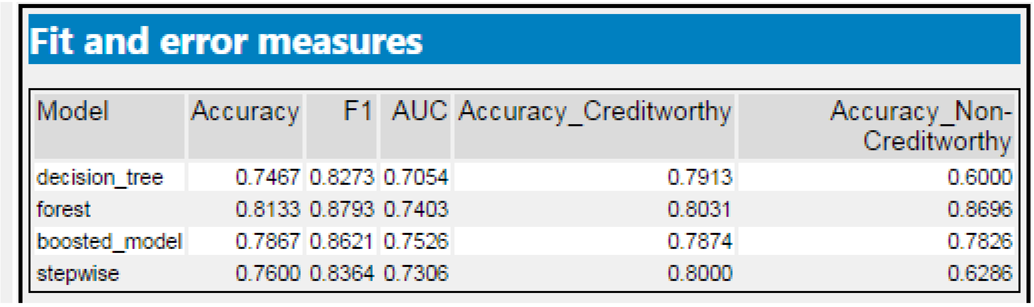
Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.68	66	31

Boosted Model

The boosted model showed account balance, credit amount, and duration of credit month as the most significant variables. It had a 78.67% accuracy





Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
forest	0.8133	0.8793	0.7403	0.8031	0.8696
boosted_model	0.7867	0.8621	0.7526	0.7874	0.7826
stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Step 4: Writeup

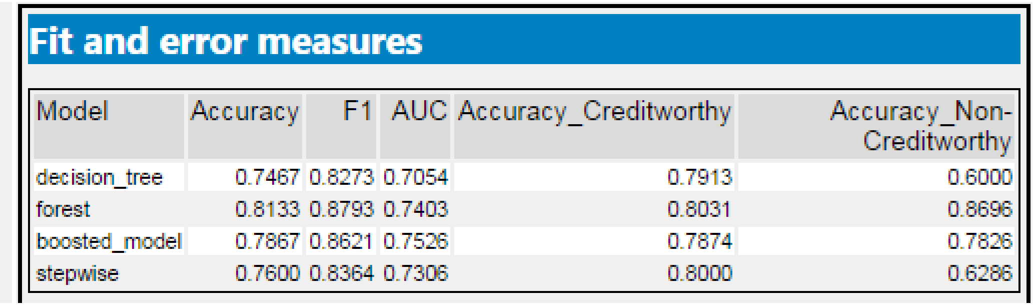
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as “Creditworthy”

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:

1. Which model did you choose to use? Please justify your decision using only the following techniques:

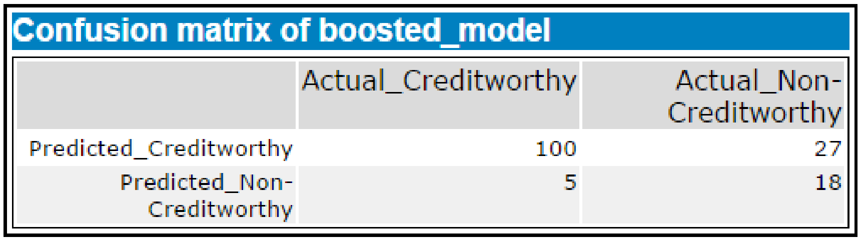
The forest model performed better overall with an overall accuracy of 81.33%. The accuracy of each prediction creditworthy vs non-creditworthy is also better with the forest model.



Fit and error measures

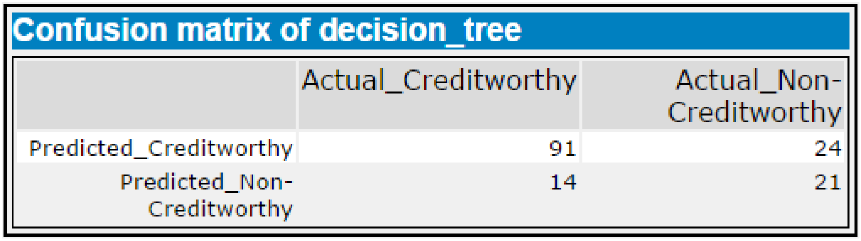
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
forest	0.8133	0.8793	0.7403	0.8031	0.8696
boosted_model	0.7867	0.8621	0.7526	0.7874	0.7826
stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Each model has a similar accuracy when predicting creditworthiness correctly, but overall the predictions for non-creditworthy were not nearly as accurate. This means that if the other models were used, non-creditworthy applicants would be given loans in error.



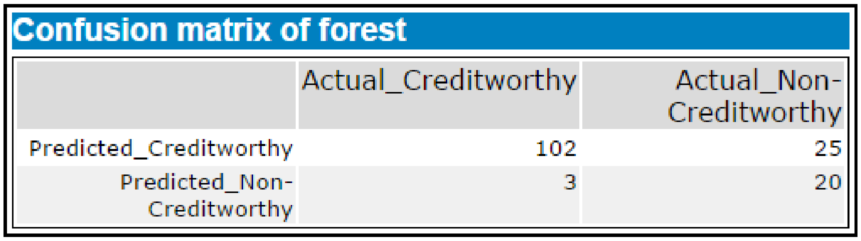
Confusion matrix of boosted_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18



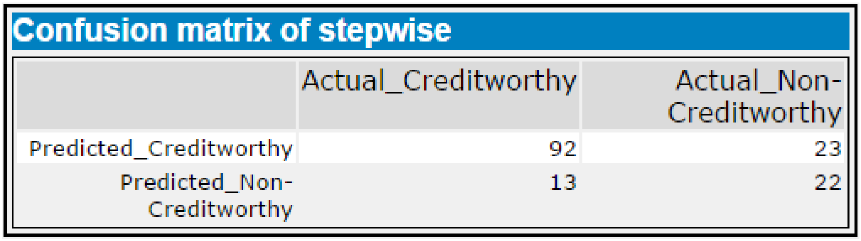
Confusion matrix of decision_tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21



Confusion matrix of forest

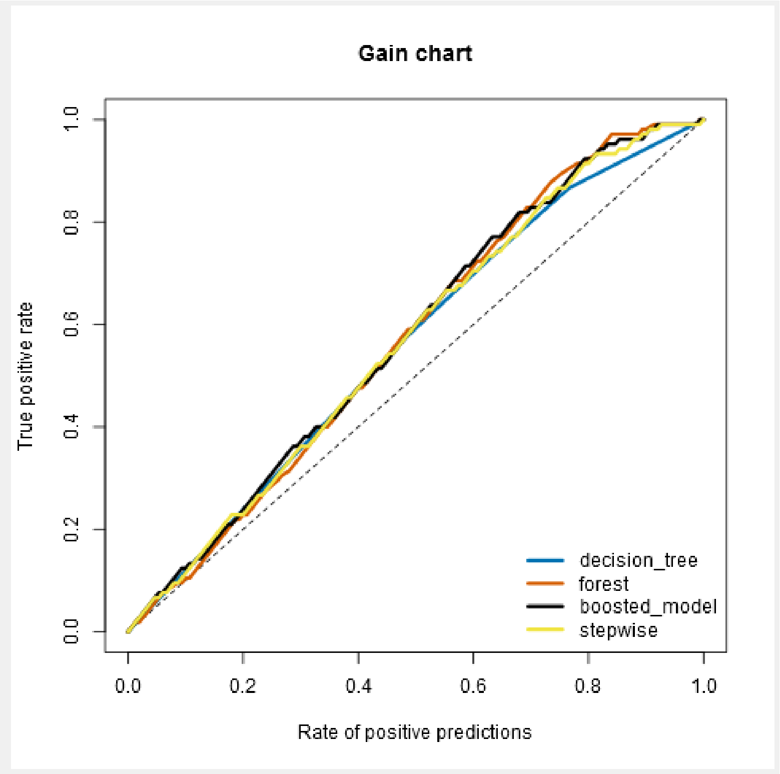
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

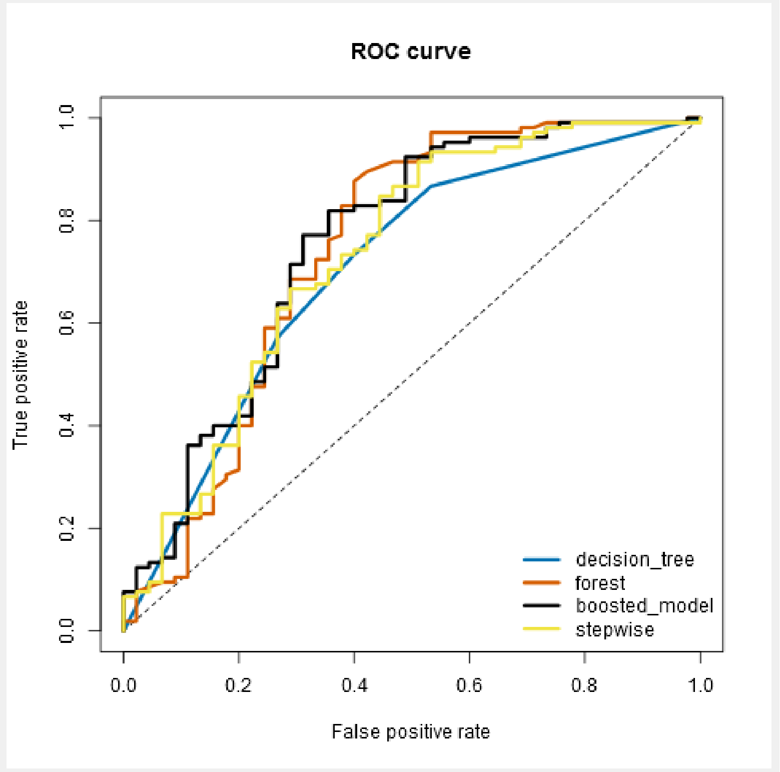


Confusion matrix of stepwise

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

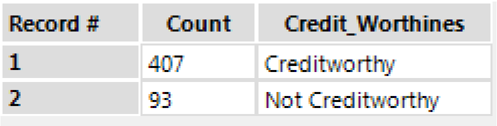
In the gain chart and roc curve, forest model shows the best results by having a better true positive rate than the other models.





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

2. How many individuals are creditworthy?
407 users are creditworthy



Record #	Count	Credit_Worthines
1	407	Creditworthy
2	93	Not Creditworthy

Before you Submit

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