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

- What decisions needs to be made?
  We have to decision that which customers to be classified to be eligible for loan through classification modelling.

- What data is needed to inform those decisions?
  - ***Data on all past applications***
  - ***The list of customers that need to be processed in the next few days***

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  As decision has two outcomes, model will be Binary.

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.***

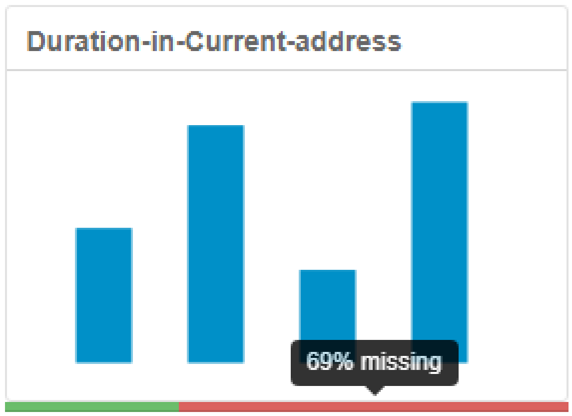
<u>Here are some guidelines to help guide your data cleanup:</u>

- 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, there is not any highly correlated field. I have not observed any correlation above 0.7 in the numeric fields.

Full Correlation Matrix

	Credit.Application.Result.num	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years
Credit.Application.Result.num	1.0000000	-0.2043168	-0.2009899	-0.0653449	-0.1379166	0.0567366
Duration.of.Credit.Month	-0.2043168	1.0000000	0.5704408	0.0795146	0.3047342	-0.0663189
Credit.Amount	-0.2009899	0.5704408	1.0000000	-0.2856309	0.3277621	0.0686430
Instalment.per.cent	-0.0653449	0.0795146	-0.2856309	1.0000000	0.0781104	0.0405397
Most.valuable.available.asset	-0.1379166	0.3047342	0.3277621	0.0781104	1.0000000	0.0854367
Age.years	0.0567366	-0.0663189	0.0686430	0.0405397	0.0854367	1.0000000
Type.of.apartment	-0.0218604	0.1531405	0.1686831	0.0829360	0.3796504	0.3330748
No.of.dependents	-0.0387889	-0.0604413	0.0055003	-0.1164661	0.0507817	0.1177351
Telephone	-0.0273066	0.1475443	0.2920589	0.0255102	0.1909078	0.1764790
Foreign.Worker	0.0056897	-0.1064163	0.0318954	-0.1182555	-0.1405878	-0.0032847

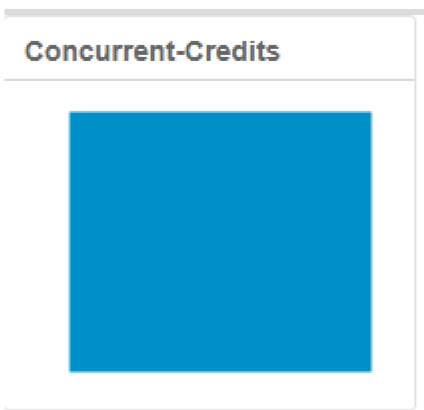
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
  Yes, Duration-in-Current-address has 69% of missing values.



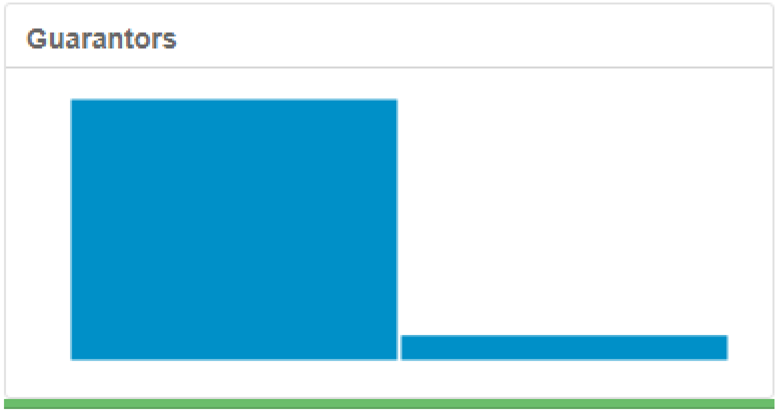
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

  Yes, 6 fields that I removed due to this low variability problem. Occupation and Concurrent-Credits fields have only value, while others removed fields have only 2 values in which one is quite dominant over the other.

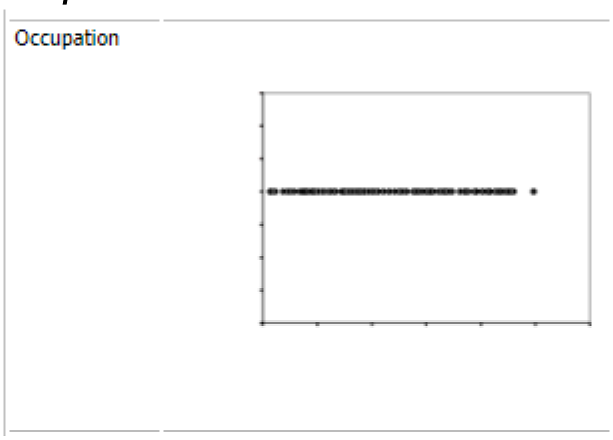
  Concurrent-Credits



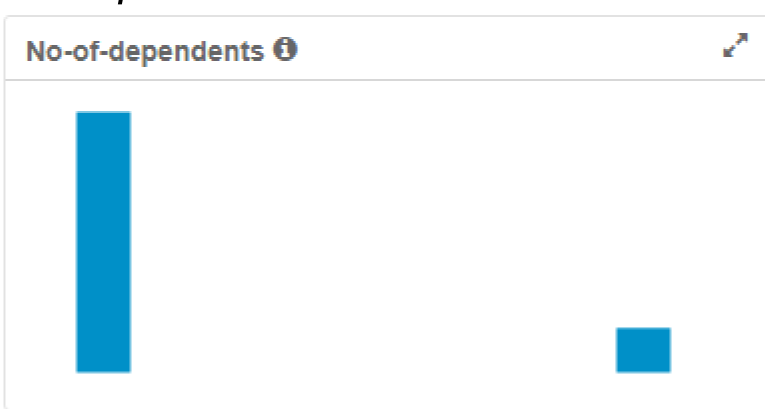
  Guarantors



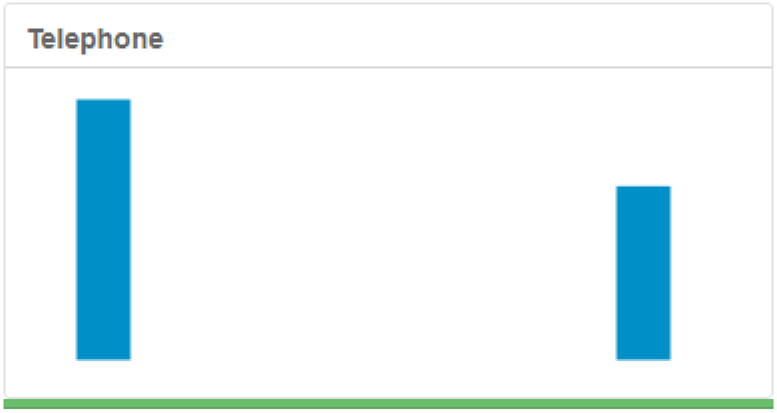
Occupation



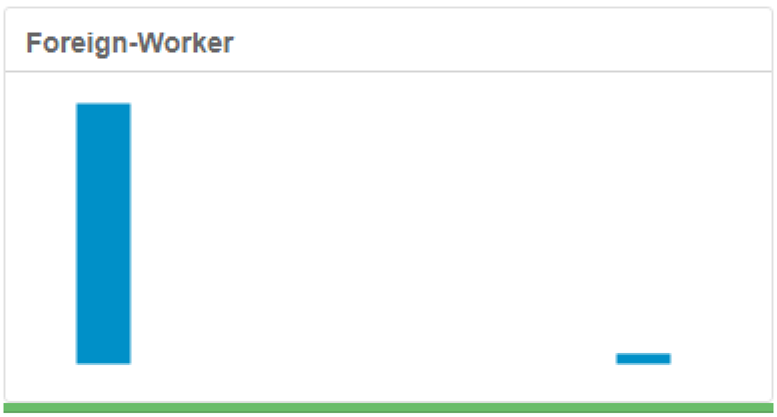
No-of-dependents



Telephone



Foreign-Worker



- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Done

13 of 13 Fields ▾ | Cell Viewer ▾ | 500 records displayed | Search | Data | Metadata | Actions ▾

Record	Credit-Application-Result	Account-Balance	Duration-of-Credit-Month	Payment-Status-of-Previous-Credit	Purpose	Credit
1	Creditworthy	Some Balance	4	Paid Up	Other	1,494
2	Creditworthy	Some Balance	4	Paid Up	Home Related	1,494
3	Creditworthy	Some Balance	4	No Problems (in this bank)	Home Related	1,544
4	Creditworthy	Some Balance	4	No Problems (in this bank)	Home Related	3,380
5	Creditworthy	No Account	6	Paid Up	Home Related	343
6	Creditworthy	Some Balance	6	No Problems (in this bank)	Home Related	362
7	Non-Creditworthy	No Account	6	Some Problems	Home Related	433
8	Creditworthy	No Account	6	Paid Up	Home Related	454

1 of 1 Fields ▾ | Cell Viewer ▾ | 1 rec

Record	Avg_Age-years
1	35.574

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:*

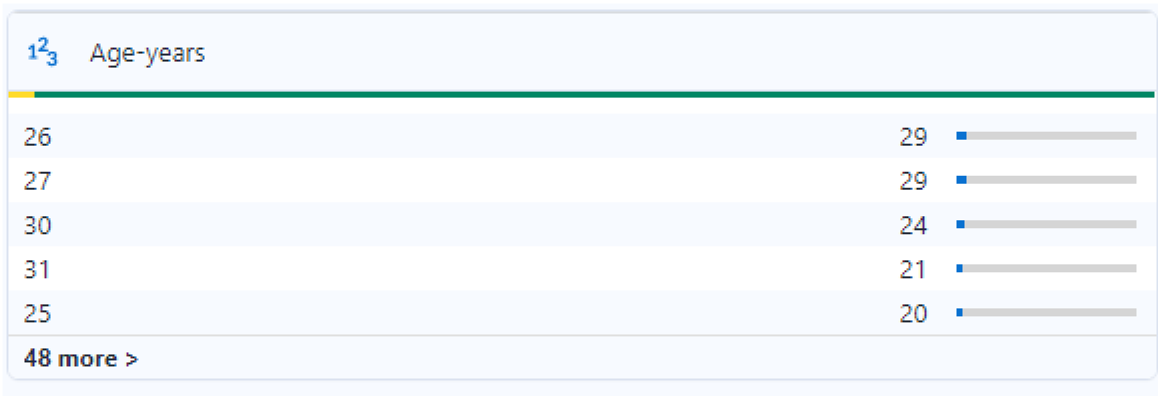
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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 imputed Age field by using the median value. The reason for imputing is because it has NULL values, as observed in the browse tool. The Yellow part of the line represents the presence of NULL values.



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.
  Based on p-values and variable importance charts, I concluded that Account-Balance, Credit-Amount and Duration-of-Credit-Month are the most important predictor variables.

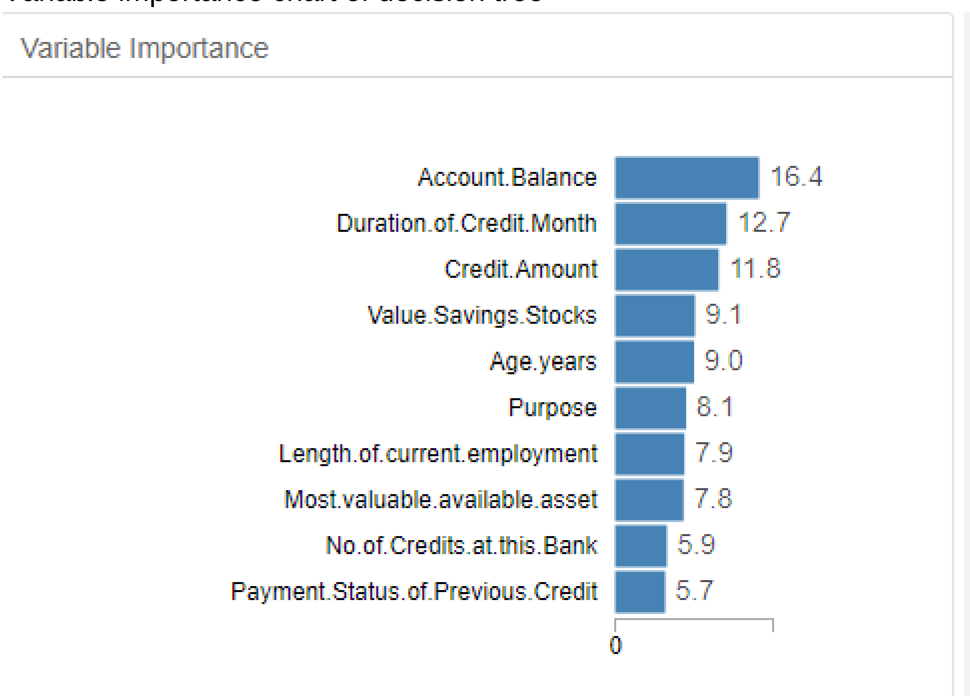
p-values of logistic regression

	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	

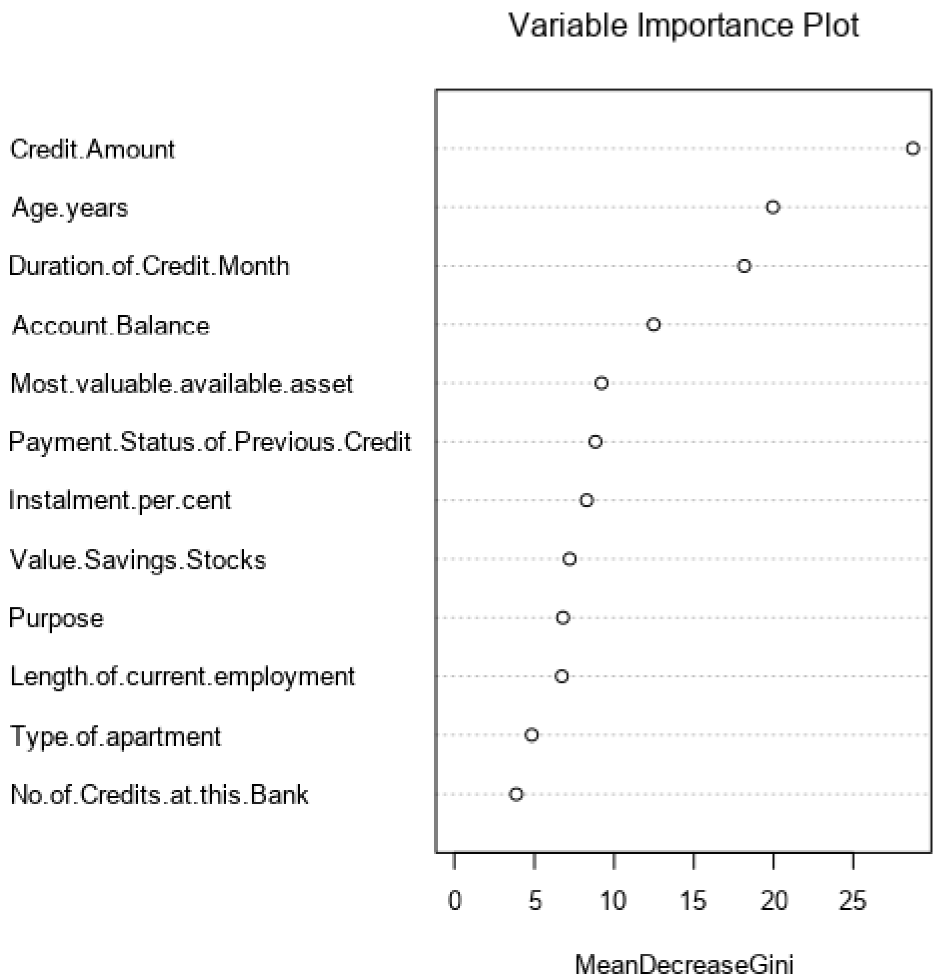
p-values of step-wise

	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	.

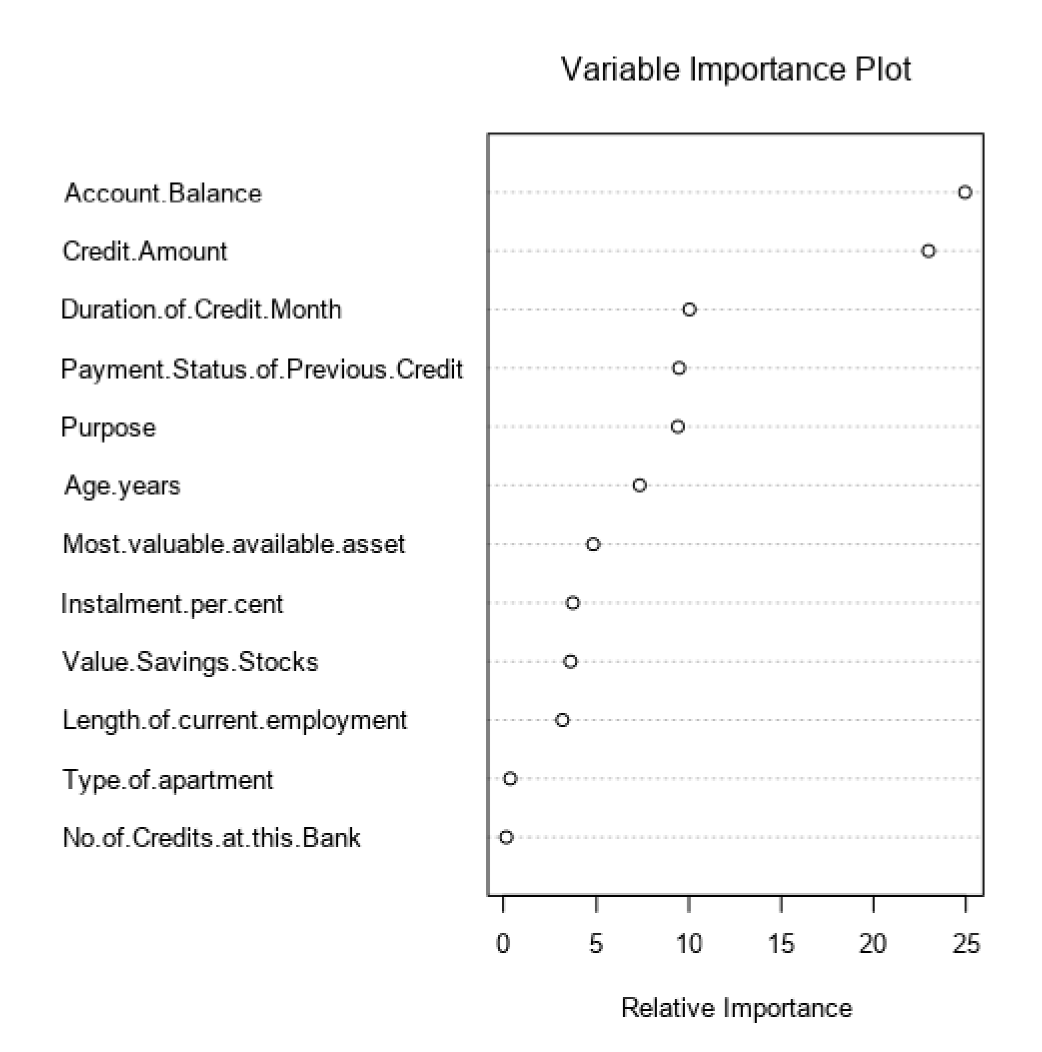
Variable importance chart of decision tree



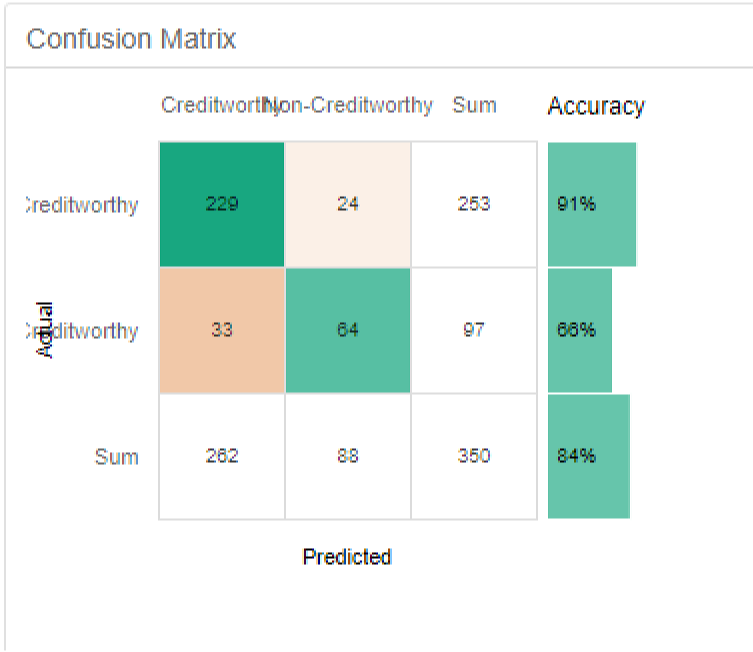
Variable importance chart of Forest model



Variable importance chart of Boosted model



- 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?
Overall internal Accuracy of decision Tree is 84% but in when validated it came out to be only 66.67%. Decision Tree model overfitted with high bias in accuracies.



Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree	0.6667	0.7685	0.6272	0.7905	0.3778

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

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:
    - Overall Accuracy against your Validation set

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments
- ROC graph
- Bias in the Confusion Matrices

From ROC curve, we can observe that forest, boosted and the step-wise were overlapping with each other with slight variations. Forest model and boosted models performed better in determining correct Creditworthy as they were leading vertically at some points. While, step-wise and decision tree models performed better in determining correct non-Credit worthy as they were leading horizontally at some points.
Their confusion matrix results also confirm the ROC results.
Based on the model comparison tool accuracy, I chose Forest model. It predicted the results with slightly higher accuracy on validation data set than other models.
There is a high Bias in all models as we can observe below in results that accuracies of Creditworthy are much higher than the accuracies of non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepWise	0.7600	0.8364	0.7306	0.8762	0.4889
Forest	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted	0.7867	0.8632	0.7515	0.9619	0.3778
DecisionTree	0.6667	0.7685	0.6272	0.7905	0.3778

Confusion matrix of Boosted

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

Confusion matrix of DecisionTree

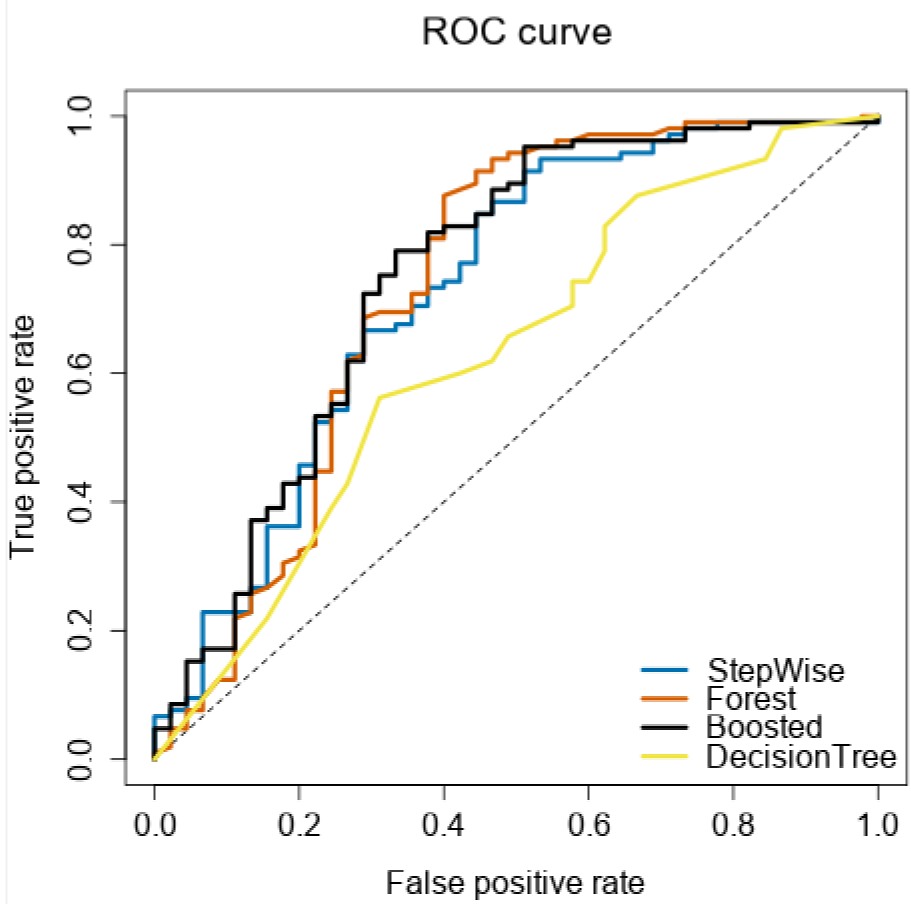
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of Forest

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

Confusion matrix of StepWise

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



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

- How many individuals are creditworthy?
  410 customers

<u>Before you Submit</u>

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