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?**

A training set needs to be made that can distinguish between creditworthy and non-credit worthy individuals from a list of credit applicants. Choosing what variables to use will determine how this decision is made.

- **What data is needed to inform those decisions?**

A total of 13 variables are needed to make this decision which are Credit-Application-Result, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount Value, Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset, Type-of-apartment, No-of-Credits-at-this-Bank, Age_years

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

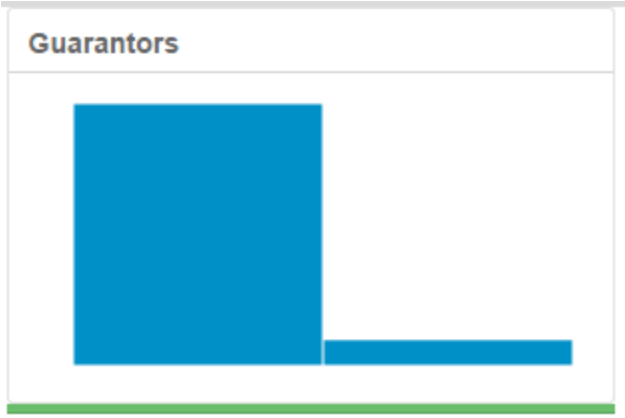
Since we need to determine if an applicant is creditworthy or non-creditworthy this is a binary model.

Step 2: Building the Training Set

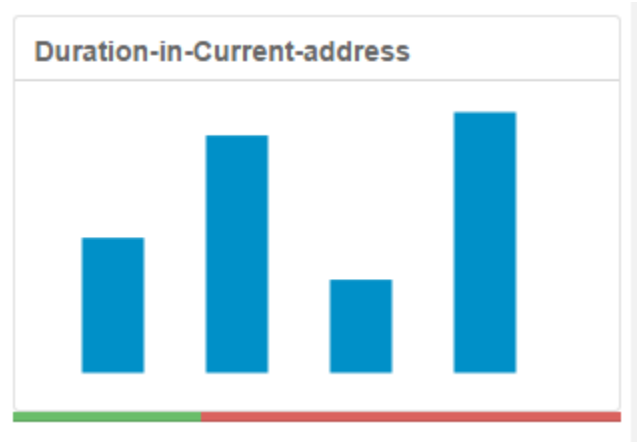
Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Answer this question:

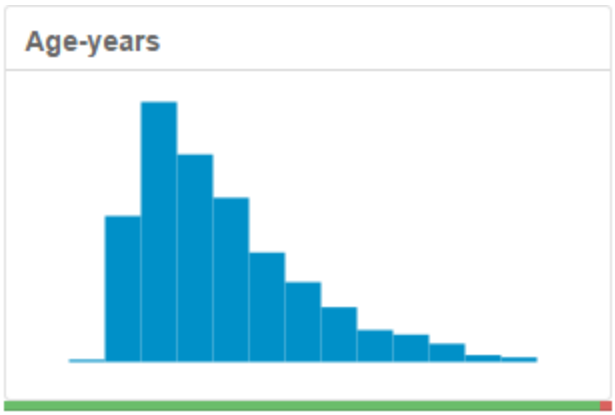
- **In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.**



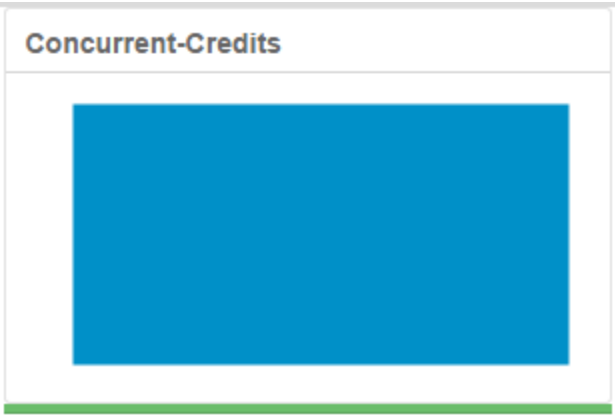
Guarantors had low variability which means the data is skewed to one side. To avoid this field skewing the data I removed it.



Duration in current address was removed due to the field missing 69% of its data. Imputing this field would result in inaccurate data so it was removed.

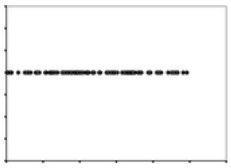


Age was only missing 2% of its data. This field was imputed utilizing a median age of 33 years old. This brought the average age to 36 years old.

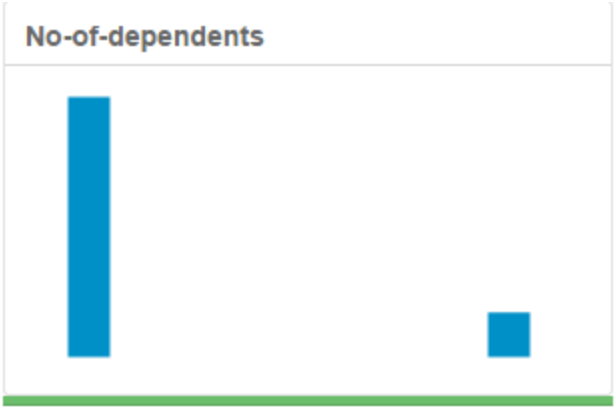


Concurrent credits was removed. This field only has one value and is of no use towards building a model.

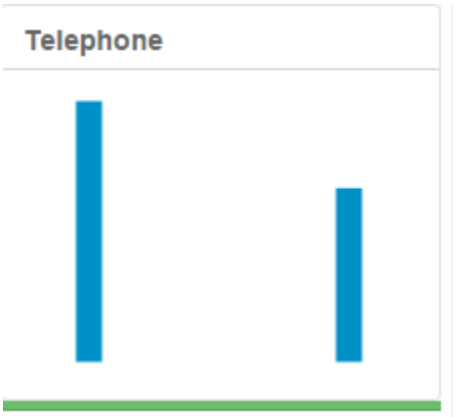
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".



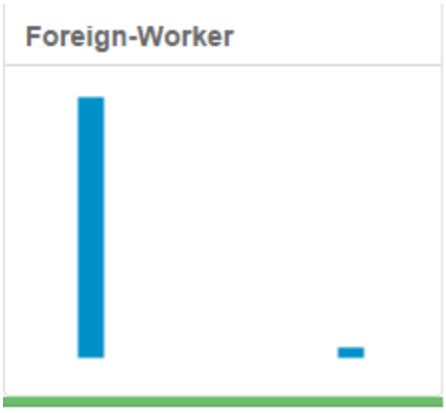
Occupation was removed due to having one value. Without any variability there is no use in using this field to build a model from.



Number of dependents was removed for having low variability, which can skew the results if a model is built from this field.



This field was removed as requested in the supporting material to this project.



Foreign Worker was removed for having low variability in the field which can skew data if used in building a model from.

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.**

  Logistic model

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, AIC: 352.5

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

Significant variables are Account balance some balance, Payment status of previous credit, Purpose new car, credit amount, length of current employment <1yr , and Instalment per cent

Decision Tree Model

```
Call:
rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount +
Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Type.of.apartment +
No.of.Credits.at.this.Bank + Age_years, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

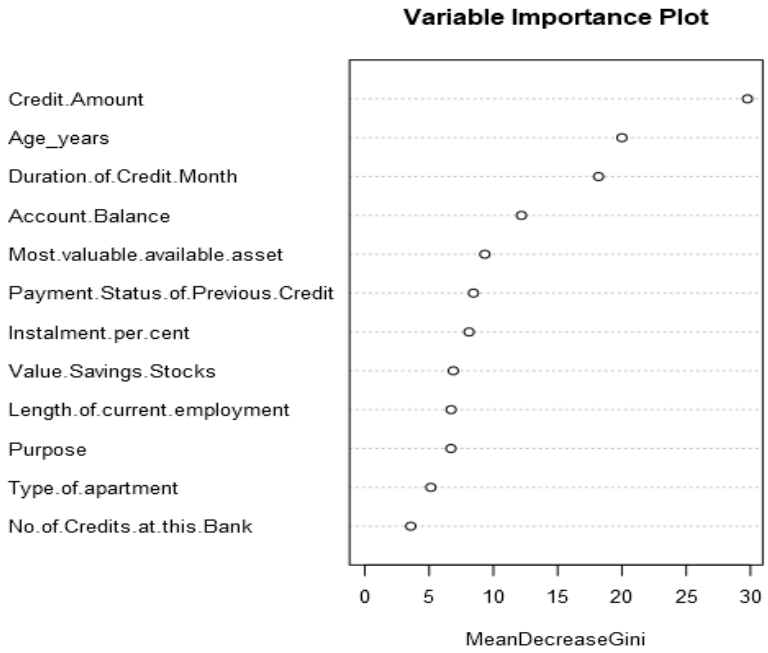
Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

```
Leaf Summary
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
   2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
   3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
     6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
     7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
      14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
      15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

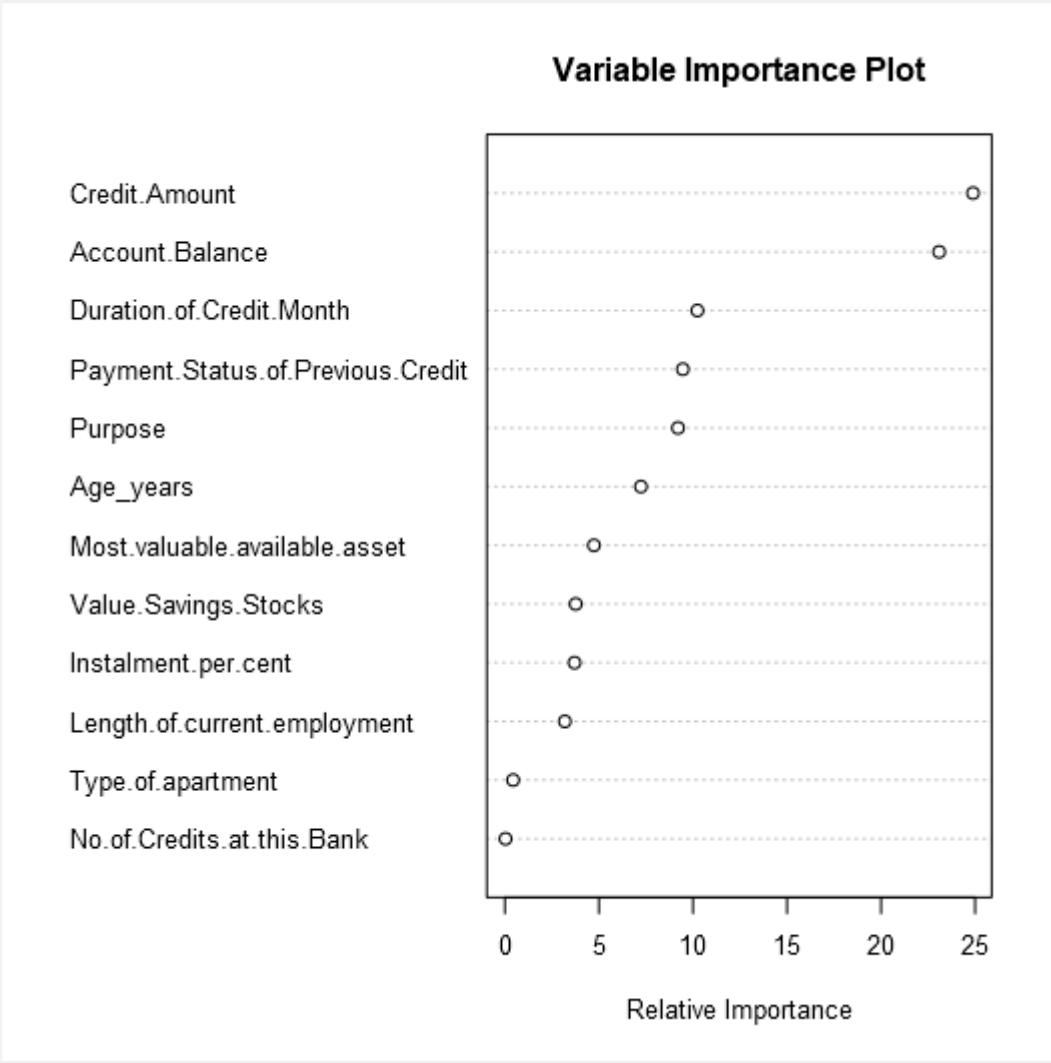
Significant variables are account balance some balance, duration of credit, value savings stocks< $100,100-1000, and value savings=none

Decision Forest Model



Significant variables are Credit amount, age years, duration of credit month, and account balance

Boosted Model



Significant variables are account balance, and credit amount

- **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?**

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decisiontree_model	0.7467	0.8273	0.7054	0.7913	0.6000
DecisionForest_model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7933	0.8670	0.7490	0.7891	0.8182
StepwiseLog_model	0.7600	0.8364	0.7306	0.8000	0.6286

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

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

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of DecisionForest_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Decisiontree_model

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

Confusion matrix of StepwiseLog_model

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

In ascending order of accuracy the Decision tree had an accuracy of 74.67%, Logistic model 76%, Boosted model 79.33%, and Forest model 80%. The Forest model had the highest overall accuracy. Looking at the confusion matrix all models had variability in accuracy for predicting creditworthy or non-credit worthy. The Forest model had near equal accuracy for both credit worthy & non-credit worthy. Except predicting non-credit worthy (82.61%) was slightly higher than credit worthy (79.53%). The decision tree model predicted higher credit worthy (79.13%) than non-credit worthy (60%) accuracy. The boost model predicted higher non-credit worthy applicants (81.82%) than credit worthy (78.91%), and Logistic model higher credit worthy (80%) than non-creditworthy (62.86%)

Step 4: Writeup

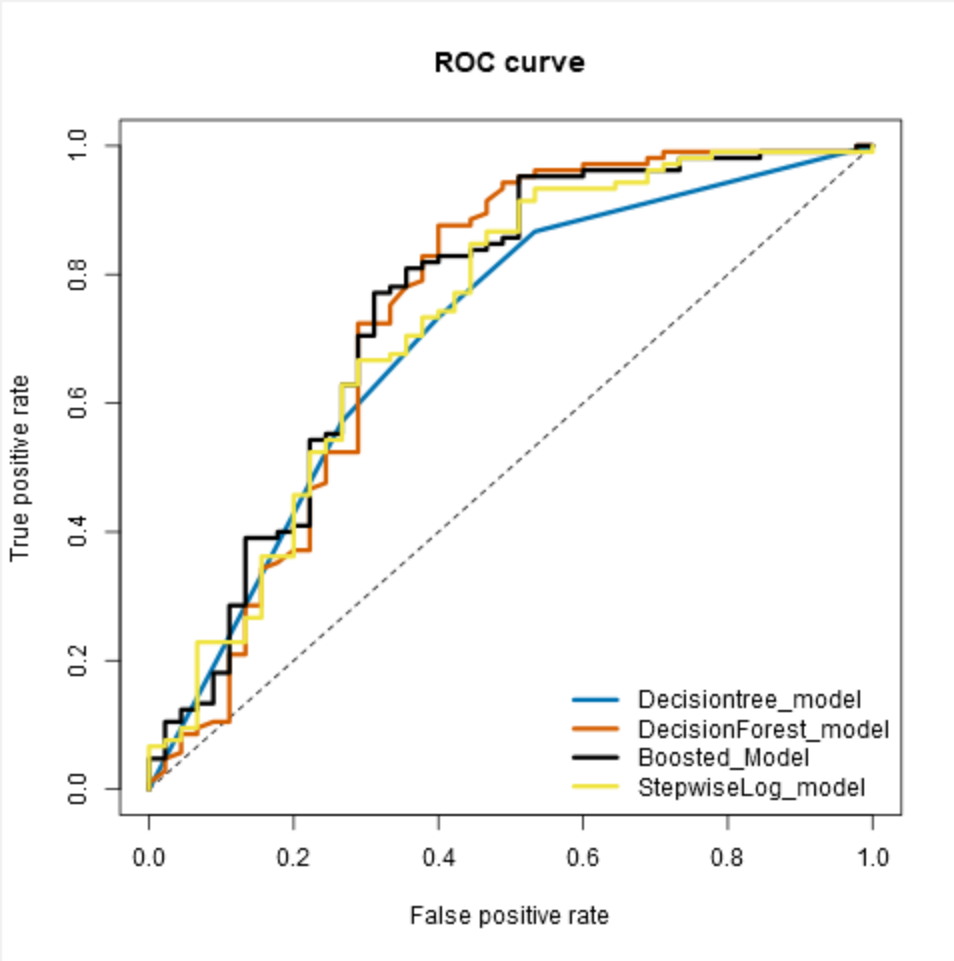
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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**

I choose the Decision Forest Model. The Decision Forest had the highest overall accuracy against the other models at 80%. In regards to the accuracy within creditworthy and non-creditworthy segments the Decision Forest was second place to the Boosted model, however the Decision Forest model had the highest accuracy of non-credit worthy segments at 82.61%

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decisiontree_model	0.7467	0.8273	0.7054	0.7913	0.6000
DecisionForest_model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7933	0.8670	0.7490	0.7891	0.8182
StepwiseLog_model	0.7600	0.8364	0.7306	0.8000	0.6286



The decision tree was the lowest for false positives out of the other models as the iterations began and became the highest for true positive values as iterations went on. This reflects the high rate of accuracy that was achieved with this model.

- **How many individuals are creditworthy?**

407 individuals are 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.