

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 need to be made?

The decision is whether the new loan applications are creditworthy or not.

- What data is needed to inform those decisions?

The previous applications will inform the decisions. In this project, the “credit-data-training.xlsx” is the material to make the prediction.

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

Since there are only two options for the results as “worthy or not worthy”, it is a binary classification 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**.*

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 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:

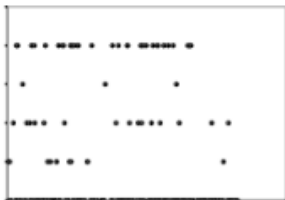
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

Too many variables can make the model overfitting which will give a higher r-square value. Even though it will look good in the training data, it will not reflect good results in the predictions for the new datasets.


I removed "Guarantors", "Duration-in-Current-address", "Concurrent-Credits", "Occupation", "No-of-dependents", "Telephone", and "Foreign-Worker". These fields have either missing data, low variability or not useful for the prediction. The details listed in the table below.

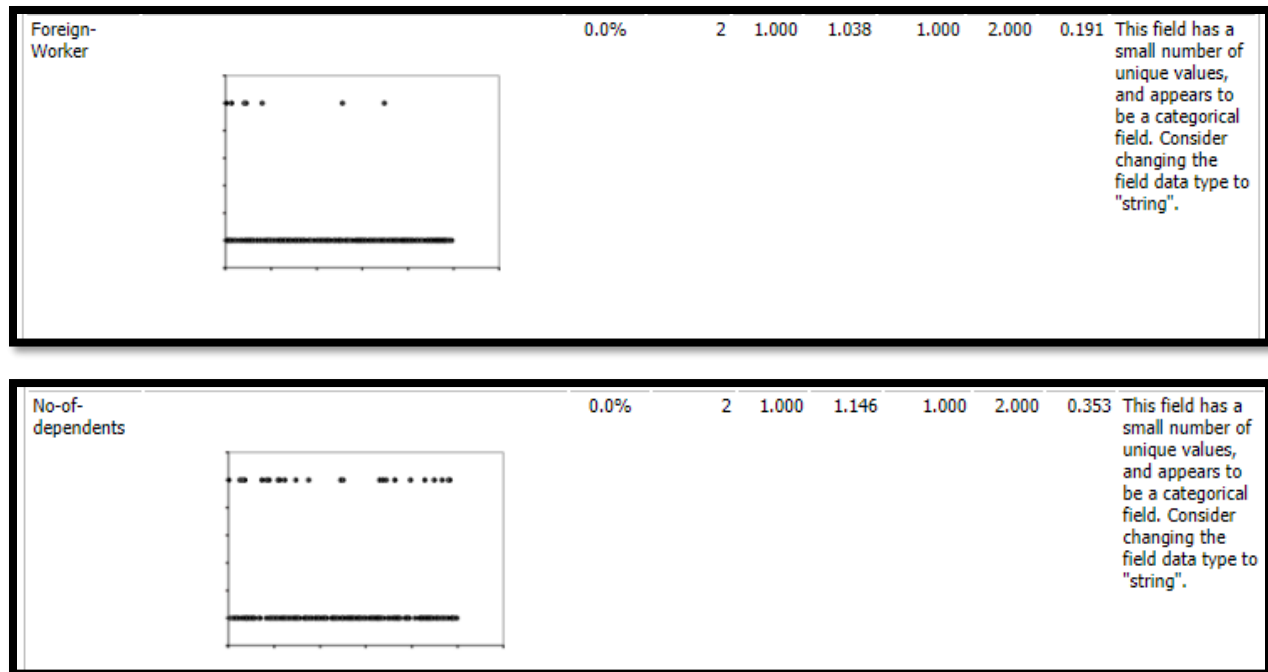
There were some missing values for Age field. The null values are replaced with median 33.

Removed Fields	Missing Data	Low Variability
Guarantors		X
Duration-in-Current-address	X	
Concurrent-Credits		X (1 value)
Occupation		X (1 value)
No-of-dependents		X
Telephone	No useful data	
Foreign-Worker		X

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts	500	500
Guarantors	0.0%	2	Yes	None	43	457

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".
------------	---	------	---	-------	-------	-------	-------	-------	---



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

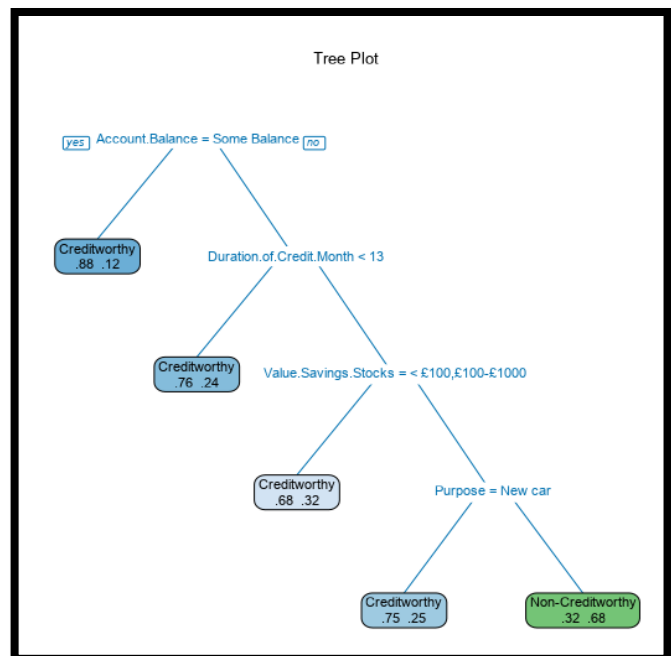
Answer these questions for **each model** you created:

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

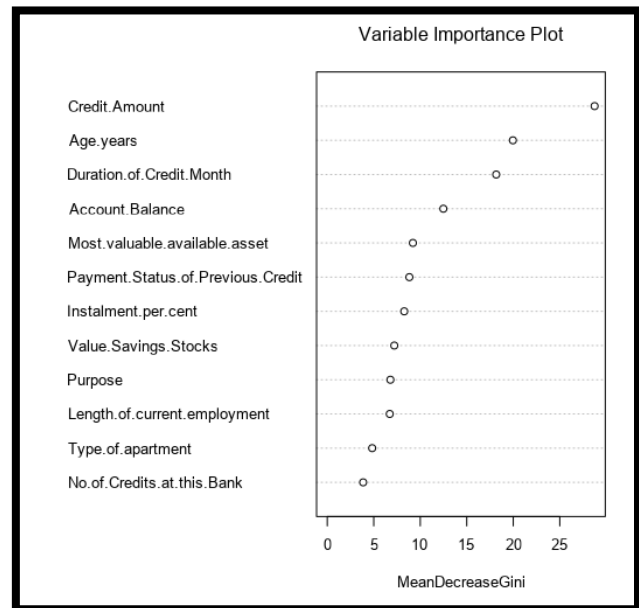
Logistic Regression: “Some Balance” value of Account-Balance is the most significant variable. “New car” value and “Used car” value of Purpose field are the next most significant variables. Please see the report below for the details.

	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	

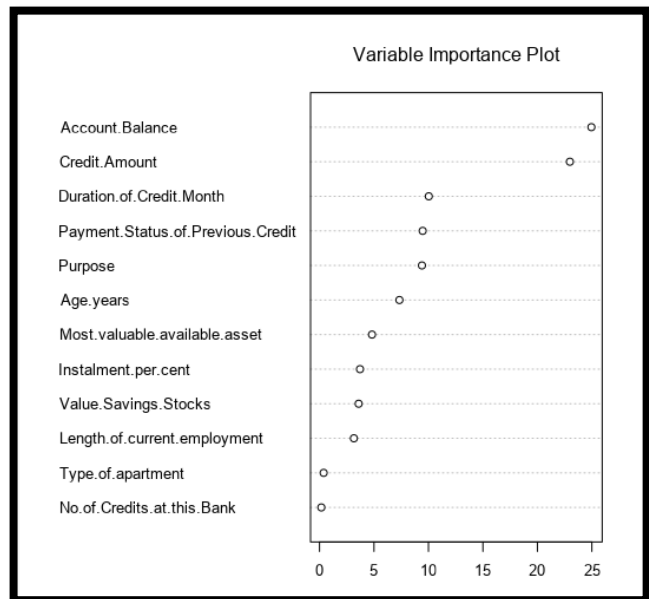
Decision Tree: “Some Balance” value of Account-Balance is the most significant variable.



Forest Model: Credit-Amount is the most significant predictor value.



Boosted Model: The most important variable is Account-Balance predictor. However, Credit-Amount is very close too. See the figure below.



- 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/Stepwise:

The overall percent accuracy is 76%. While the creditworthy accuracy is significantly high which is 88%, non-creditworthy is %49. This model bias towards correctly predicting creditworthy results.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

The confusion matrix is shown below. The model predicted 92 correct creditworthy and 23 times incorrect creditworthy. On the other hand, non-creditworthy is predicted 13 times correctly and 22 times incorrectly.

Confusion matrix of stepwise		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Decision-Tree:

The overall percent accuracy of the Decision-Tree model is 75%. While the creditworthy accuracy is significantly high which is 89%, non-creditworthy is %42. The result is very close to logistic regression model result. This model bias towards correctly predicting creditworthy results.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
tree	0.7467	0.8304	0.7035	0.8857	0.4222

The confusion matrix of the decision-tree model is shown below. The model predicted 93 correct creditworthy and 26 times incorrect creditworthy. On the other hand, non-creditworthy is predicted 12 times correctly and 29 times incorrectly.

Confusion matrix of tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Forest Model:

The forest model overall accuracy is 79%. This makes the forest model the most accurate model. The creditworthy result is 97% and the non-creditworthy is 38%. This model bias towards correctly predicting creditworthy results.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
forest	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix shows that forest model was able to predict 102 correct creditworthy applications. However, 28 creditworthy predictions were incorrect. Also, it was able to predict 17 not worthy applications and failed to predict 3 applications.

Confusion matrix of forest		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Boosted Model:

Boosted model has the second highest overall accuracy value between four models. It has 79% accuracy. The creditworthy result is 96% and the non-creditworthy is 38%. This model bias towards correctly predicting creditworthy results.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
boost	0.7867	0.8632	0.7515	0.9619	0.3778

Confusion matrix shows that boosted model was able to predict 101 correct creditworthy applications. However, 28 creditworthy predictions were incorrect. Also, it was able to predict 17 not worthy applications and failed to predict 4 applications. It is very similar to forest model results.

Confusion matrix of boost		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

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:

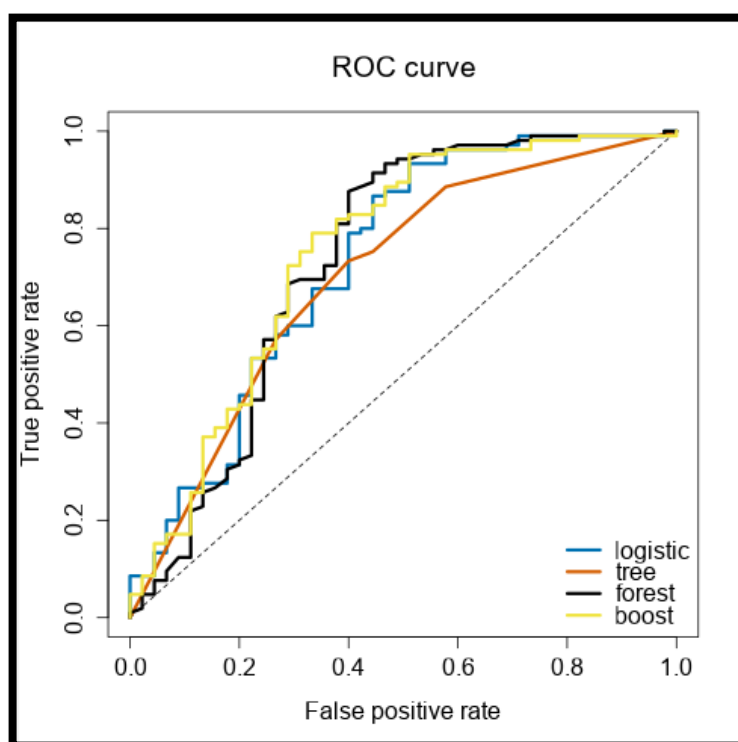
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
 - Overall Accuracy against your Validation set
 - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
 - ROC graph
 - Bias in the Confusion Matrices

Based on the overall accuracy against the validation set, the forest model is the most accurate model with 79.3% accuracy.

Even though there is a slight difference between boosted model and forest model in predicting creditworthy applications, the forest model has better results in the accuracies within “Creditworthy” and “Non-Creditworthy” segments.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic	0.7800	0.8520	0.7314	0.9048	0.4889
tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest	0.7933	0.8681	0.7368	0.9714	0.3778
boost	0.7867	0.8632	0.7515	0.9619	0.3778

If we look at the ROC curve, the boosted model and the forest model are reaching high true positive-false positive ratio. However, the forest model is reaching the highest true positive rate earlier than the boosted model.



Confusion matrix below shows that the forest model still has the best results compared to the other models.

Confusion matrix of boost		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of forest		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of logistic		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Confusion matrix of tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

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

- How many individuals are creditworthy?
408 individuals are creditworthy and 92 individuals are 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.