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?
  Determine for everyone of the 500 customers if a loan will or won't be approved.

- What data is needed to inform those decisions?
  Historical data for previous credit request with "Creditworthy" and "Non-creditworthy" cases. And data from new customers to evaluate their credit requests. Also for both data sets are required to have the most reliable possible data for next fields:
  - Credit Amount
  - Account Balance
  - Age years
  - Duration of previous Credit Month
  - Most valuable available asset

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary: to classify customer request as "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 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.

Fields removed due to low variability or many null values

  - foreign worker: may cause skew for low variability
  - no-of-dependents: may cause skew for low variability
  - occupation: all same category, may cause skew for low variability
  - telephone: was removed because it is not relevant for the analysis.
  - concurrent credits: all same category, may cause skew for low variability
  - Guarantors: may cause skew for low variability
  - Duration in current Address: it was removed, because this field has 69% of missing values, and that is too high and try to impute it is not an option, because the generated values potentially will cause skew.

The field to be imputed is:

  - Age-years: since this field only have missing a low amount of values, around 2%. It has been imputed with the median value of 33.

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

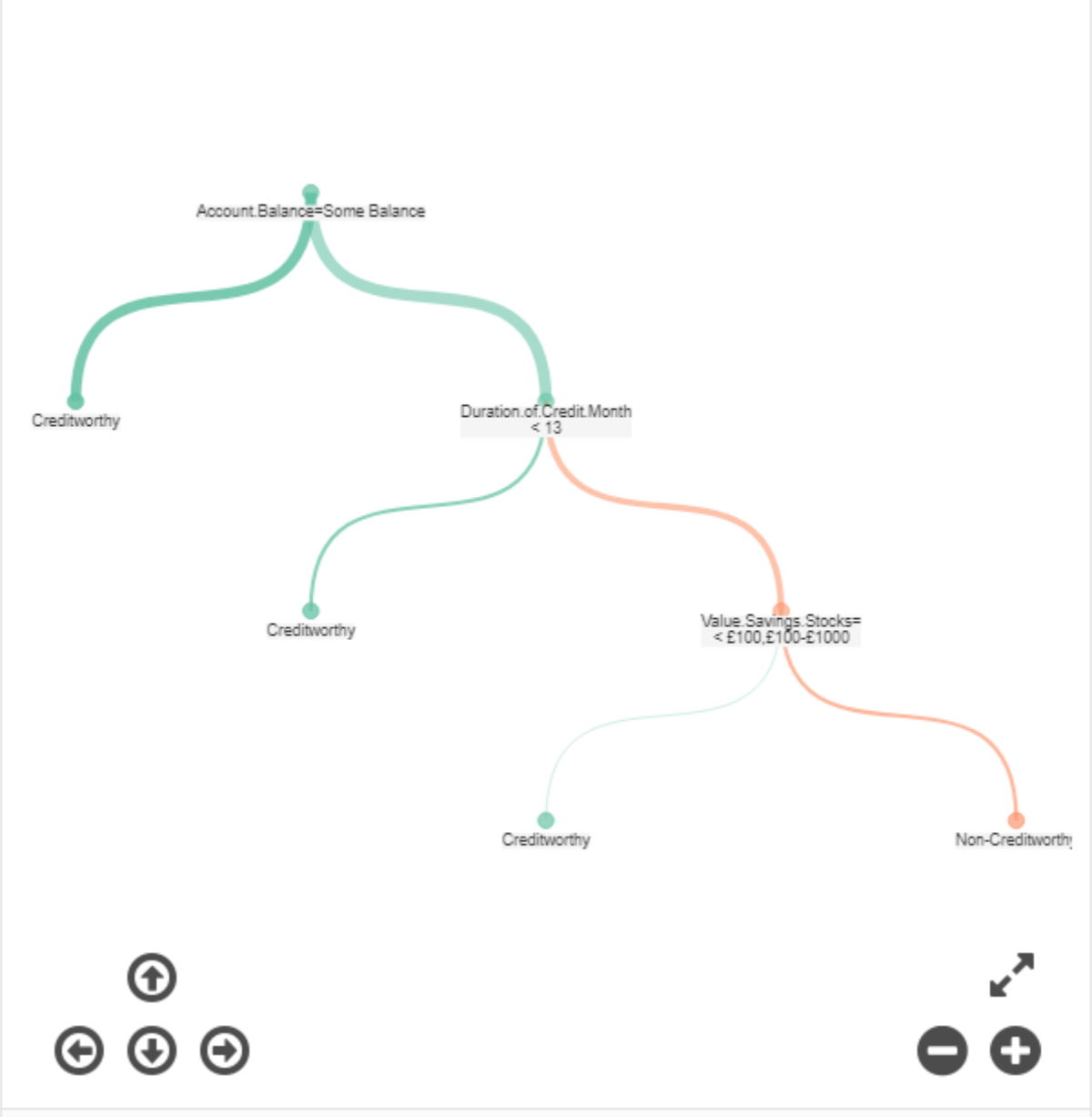
*Answer these questions for **each model** you created:*

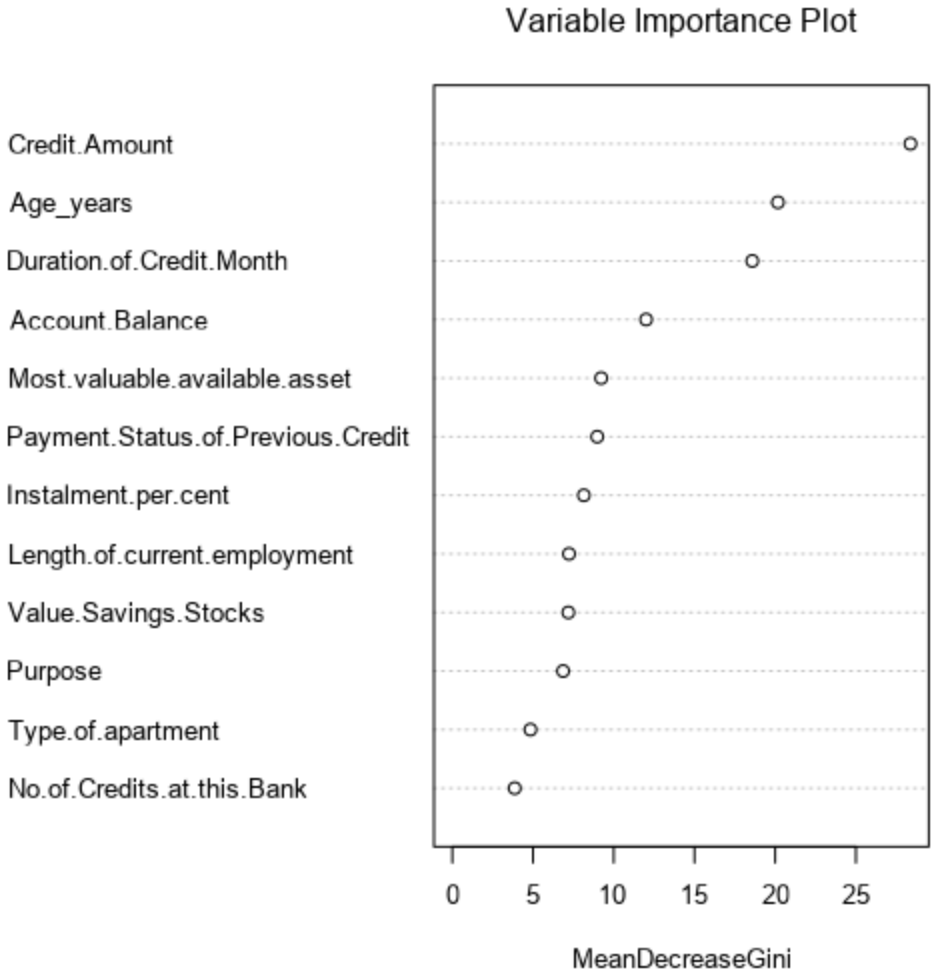
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

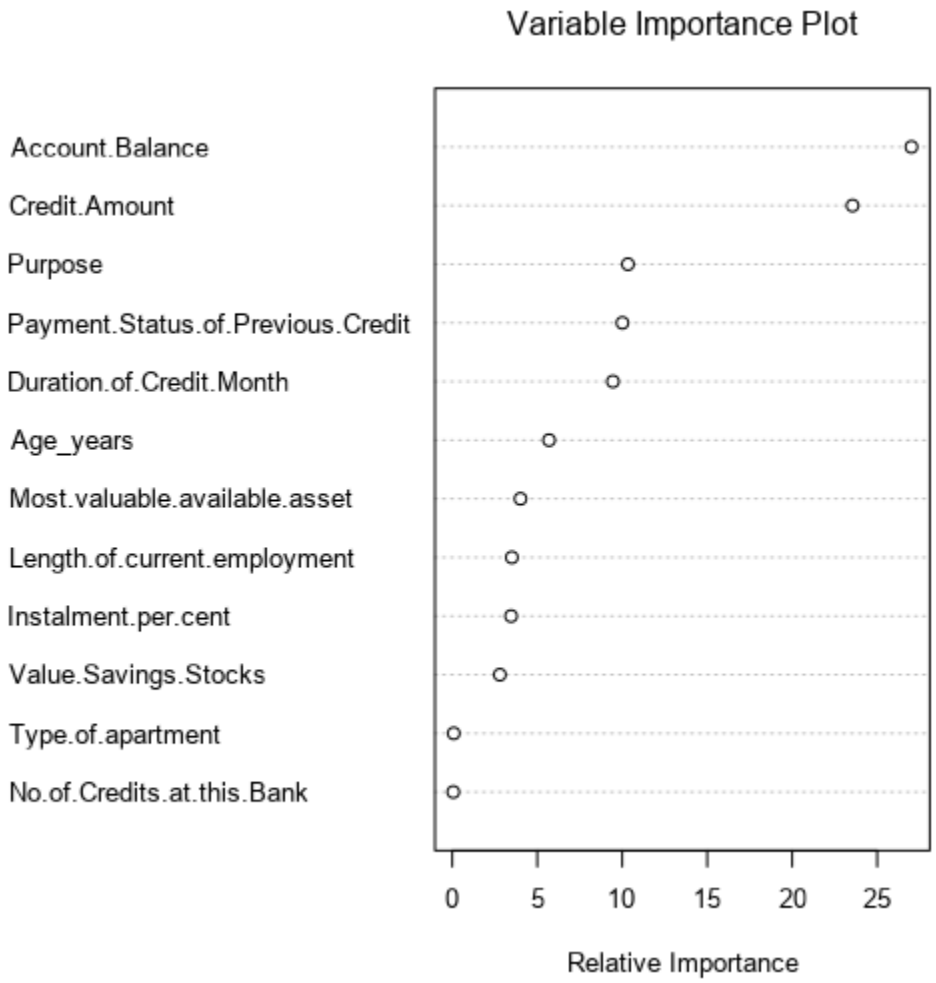
The next predictor variables has been identified as more important by the Tree, Random Forest and Boosted models.

- Credit Amount (0.009)
- Account Balance (1.79e-06)
- Age_years (0.3574)
- Duration.of.Credit.Month (0.6356)
- Most.valuable.available.asset (0.0362)

The selected model has been Random Forest Model, which has generated the next importance for predictor variable importance respectively.




Variable Importance Plot
Credit.Amount
Age_years
Duration.of.Credit.Month
Account.Balance
Most.valuable.available.asset
Payment.Status.of.Previous.Credit
Instalment.per.cent
Length.of.current.employment
Value.Savings.Stocks
Purpose
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
MeanDecreaseGini



The p values generated by the Logistic Model are next:

	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	*
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	
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747	

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

- 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?

The confusion Matrix generated by the Tree model has an accuracy of 78 %. It is important to highlight that the Creditworthy has a higher accuracy (89%) than the Noncreditworthy (49%). That low level of accurary for Noncredithworthy can be considered as a potential bias.

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

The selected model has been Random Forest Model (RF_Model), because in the model comparison table it has the highest accuracy (80%), and for classification of Creditworthy (96%), that is the purpose of the models classification for this project, considering that the classification for Non-creditworthy is (42 %), and in the third place for predicting it.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_MODEL	0.7467	0.8273	0.7054	0.8667	0.4667
RF_MODEL	0.8000	0.8707	0.7361	0.9619	0.4222
LR_MODEL	0.7800	0.8520	0.7314	0.9048	0.4889
BO_MODEL	0.7667	0.8523	0.7397	0.9619	0.3111

Analyzing the Confusion matrix, the RF Model (random forest model) performed better that the others for classifying Creditworthy cases.

Confusion matrix of BO_MODEL

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	31
Predicted_Non-Creditworthy	4	14

Confusion matrix of DT_MODEL

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

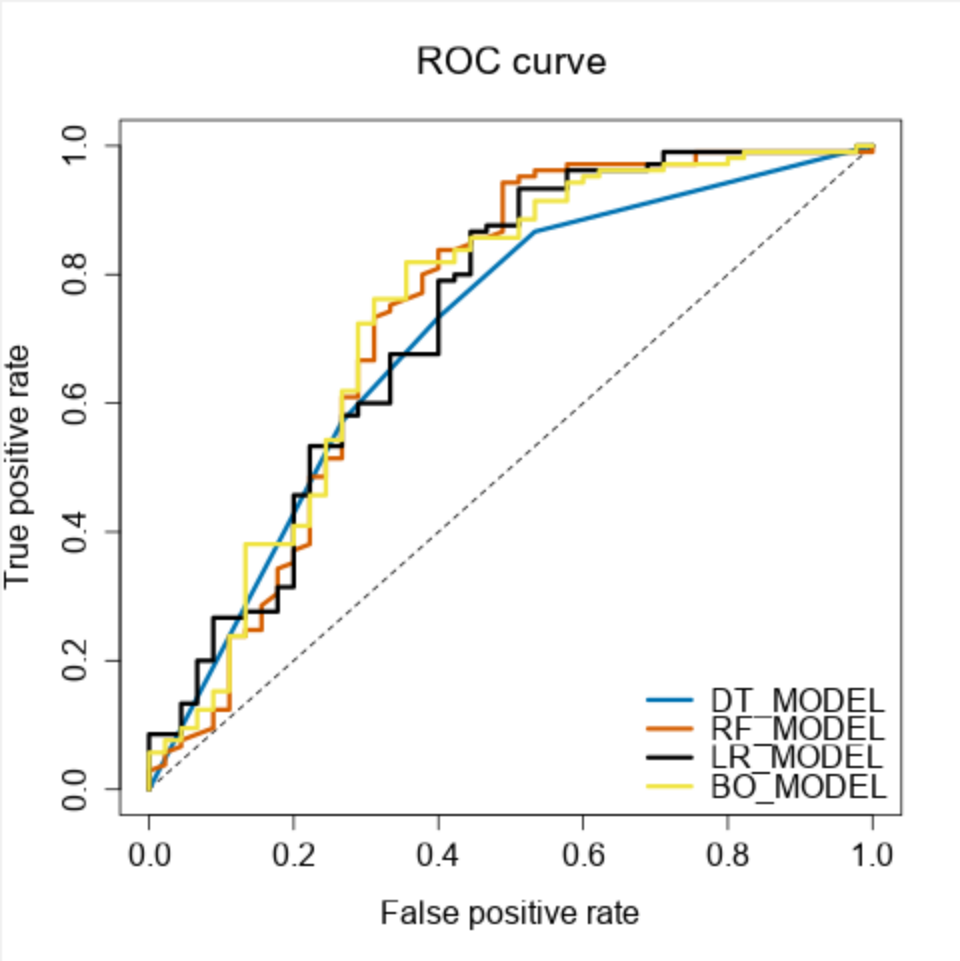
Confusion matrix of LR_MODEL

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Confusion matrix of RF_MODEL

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

Also in the ROC curve below is observed that the RF_MODEL is performing slightly better than the others to classify Creditworthy.

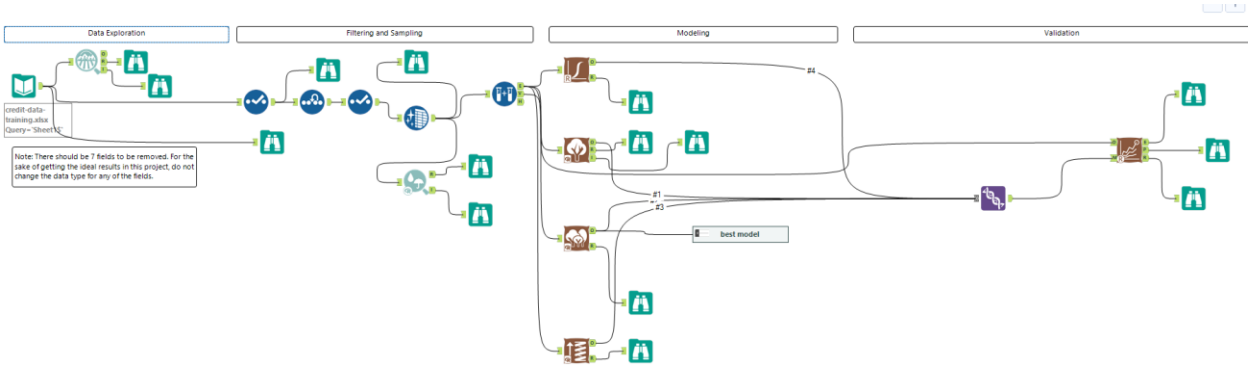


Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

There are 406 customers that has been classified as creditworthy.

Alteryx Model, to evaluate best model



Alteryx workflow applying the Random Forest Model for classification of Creditworthy and Non-creditworthy cases.