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?

Answer: Due to the current situation, the bank is confronted with the situation whether 500 potential customers get a loan or not. Each of them has to be proven.

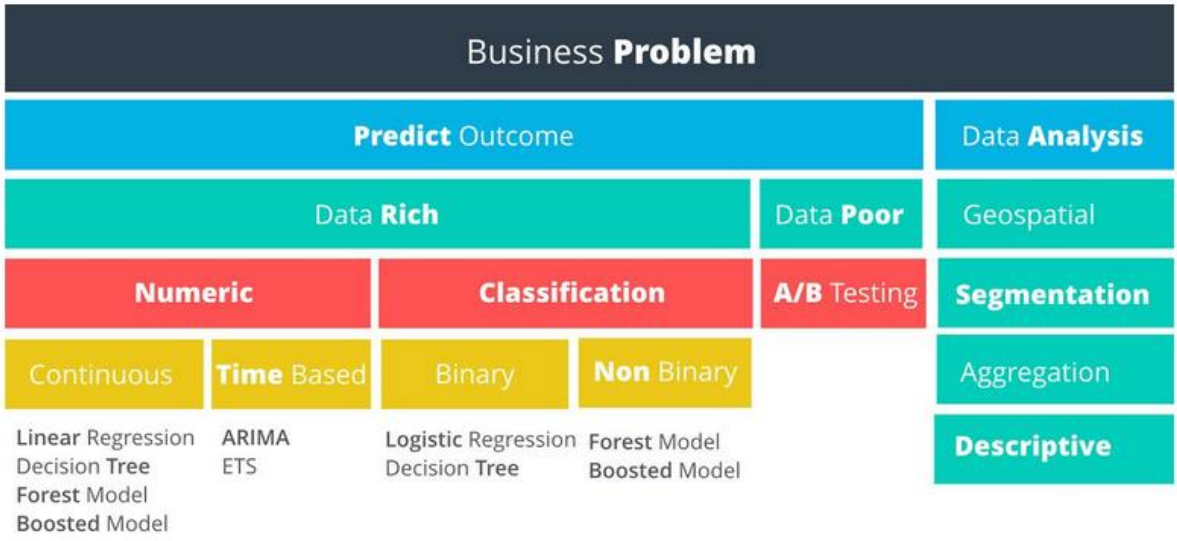
- What data is needed to inform those decisions?

Answer: To make that decision we need the Data from the past applications as well as the list of new customer applications. Useful variables from the data "Credit-data-training" are the following.

- *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*
- *Age-years*
- *Type of apartment*
- No of Credits at this Bank

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

Answer: If we go back to your overview about the analytical Models, we have a business problem -> predict outcome -> Data Rich -> Classification -> Binary/Non Binary. So, I will use the logistic regression, Decision Tree, Forest Model and Boosted 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

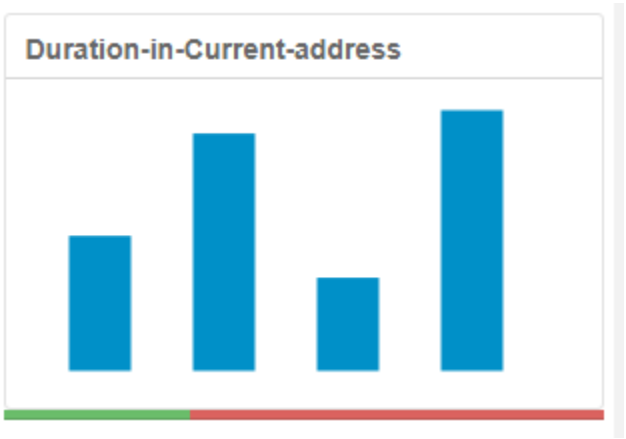
To achieve consistent results reviewers expect.

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.

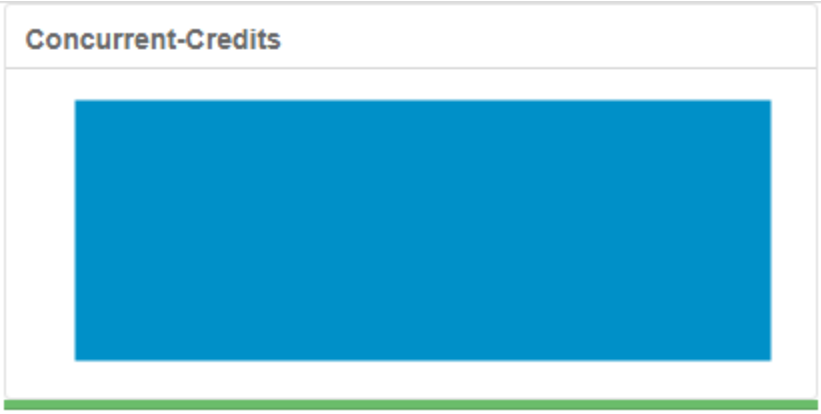
You can answer the question in two ways. Either you analyze the data and recognizes inconsistencies based on the completeness or the structure of the data. Or you look again at the business problem and relates the data to the problem and decides on the basis of common sense whether the data can be useful or not.

1. *Answer: Duration in Current address: Too many null values -> needs to be removed*



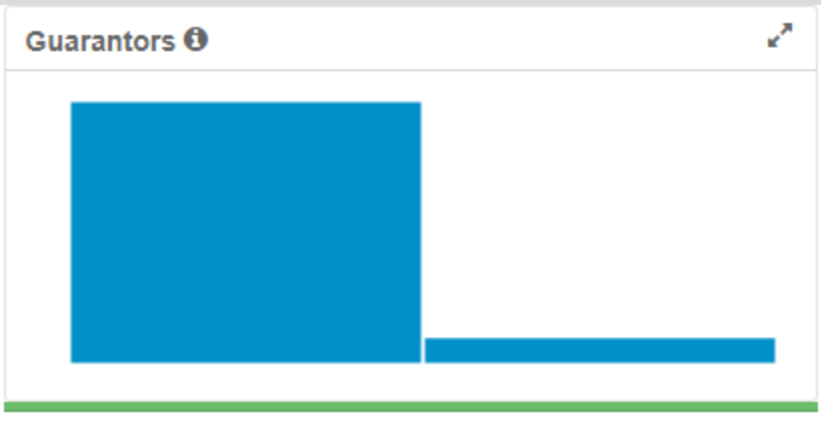
---

2. *Concurrent-Credits: Just one position = no meaningfulness -> needs to be removed*



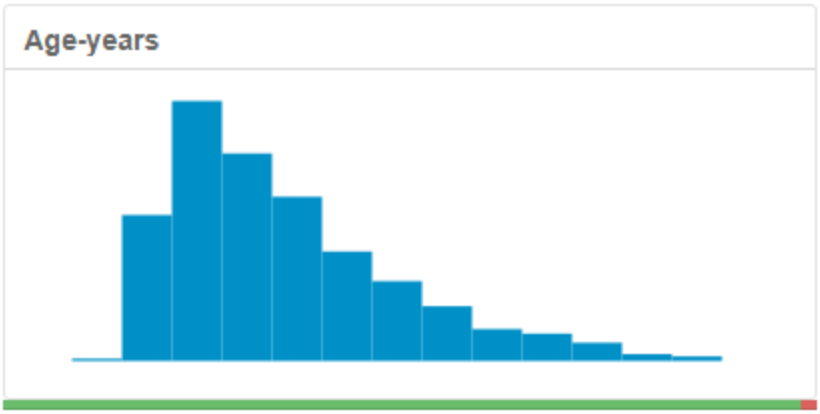
---

3. *Guarantors: One Position is very dominate -> needs to be removed.*



---

4. *Age-years: There are some missing data, which you can see by the little red bar but since it is numeric, I can use the impute function in Alteryx and choose the median to fill up the nulls.*



Besides that, the variables “Foreign-Worker”, “No-of-dependents”, “Telephone” can be removed because there is no logical connection between these variables and the business problem.

Here is a field-overview about all variables and their structure.

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.

- 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)

1. *Logistic Regression with Stepwise*

Answer: I used the "credit-Application-Result" as the Target variable and select all other variables as the predictor variable. Based on the stars on the right side you can see that, "Account Balance", "Purpose" and "Credit amount" are these variables with the strongest significant/ a p-value of less than 0.05.

Report

Bericht über Logistic Repression-Modell Stepwise_Credit

Basis-Übersicht

Aufruf:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Normabweichungs-Residuen:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Koeffizienten:

	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	.

Bedeutungscodes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Verteilungsparameter für binomial als 1 angenommen)

Null-Abweichung: 413.16 auf 349 Freiheitsgrad
Verbleibende Abweichung: 328.55 auf 338 Freiheitsgrad
McFadden R-Quadrat: 0.2048, Akaike-Informationskriterium 352.5

In the model comparison you can see that the logistic regression model has an accuracy of 78% while the accuracy of creditworthy is 90.48% and the accuracy for non-creditworthy is 48.89%. Because of the accuracy you can say that the logistic regression model has a bias towards correctly predicting creditworthy individuals.

Model Comparison Report

Fit and error measures

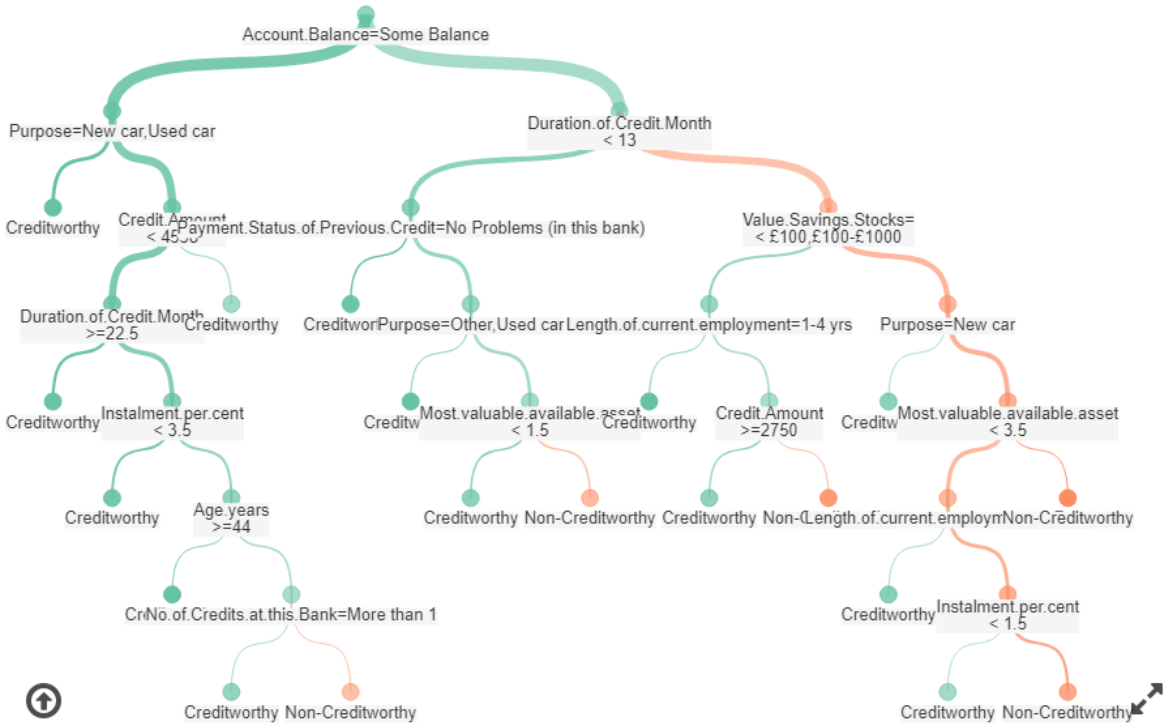
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit	0.6733	0.7721	0.6296	0.7905	0.4000
FM_Credit	0.7933	0.8681	0.7415	0.9714	0.3778
BM_Credit	0.7800	0.8596	0.7446	0.9619	0.3556

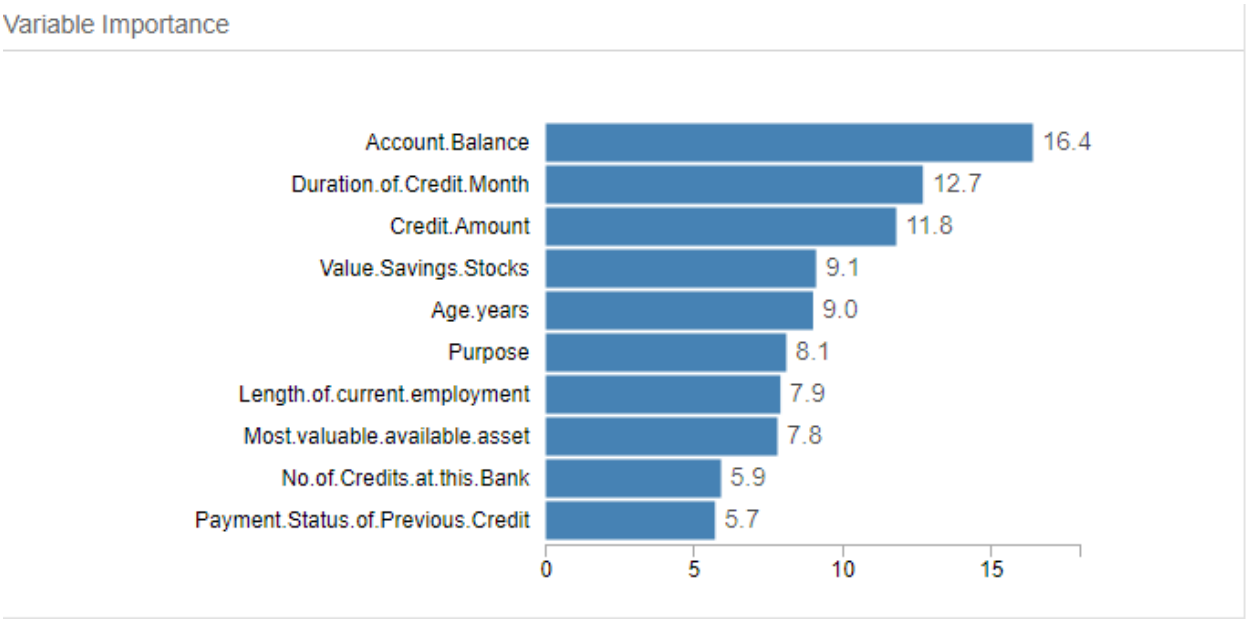
Confusion matrix of LR_Credit

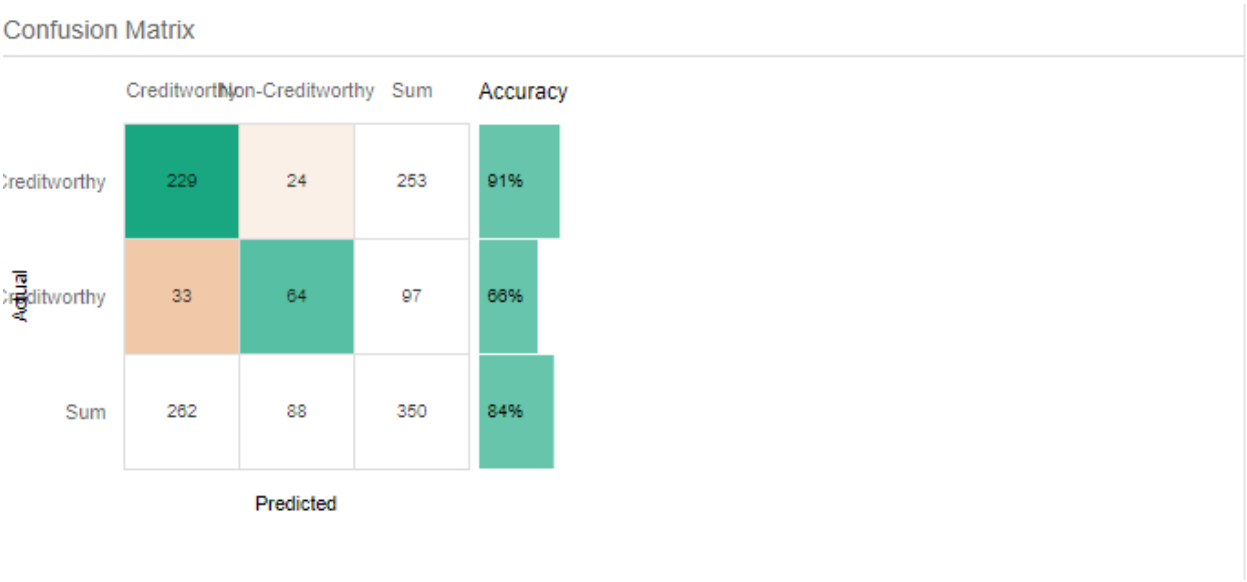
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

2. *Decision Tree*

Answer: I used the "credit-Application-Result" as the Target variable and select all other variables as the predictor variable. Under the section variable importance, you can see the most important variables in descending order. "Account Balance" is the most important variable followed by "duration of credit month" as well as "credit amount".







Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit	0.6733	0.7721	0.6296	0.7905	0.4000
FM_Credit	0.7933	0.8681	0.7415	0.9714	0.3778
BM_Credit	0.7800	0.8596	0.7446	0.9619	0.3556

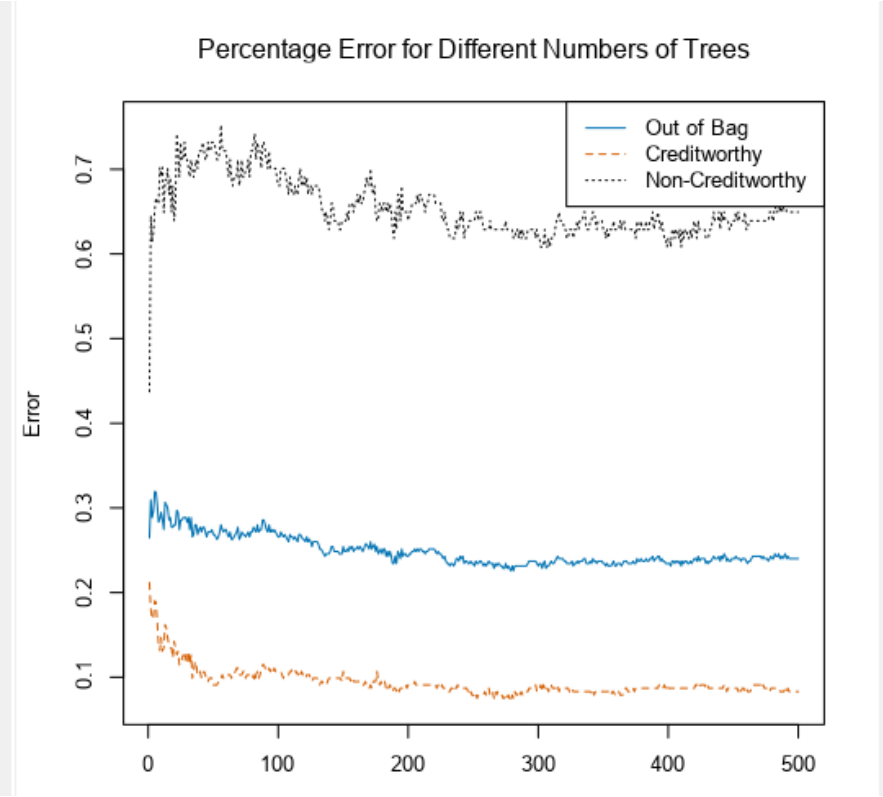
Confusion matrix of DT_Credit

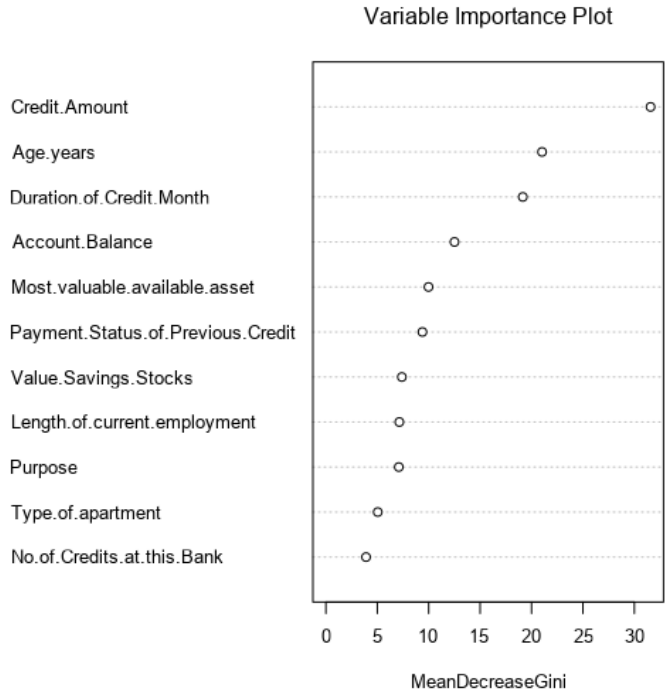
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

The overall accuracy is 67% while the accuracy of creditworthy is 79% and the accuracy for non-creditworthy is 40%. You can say that the decision tree model has a bias towards correctly predicting creditworthy individuals.

3. *Forest Model*

Answer: I used the "credit-Application-Result" as the Target variable and select all other variables as the predictor variable. Besides that, I selected default 500 trees. Under the section "variable importance Plot" you can see the most important variables. The top 3's is "Credit Amount", "Age years" and "Duration of Credit Month"





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit	0.6733	0.7721	0.6296	0.7905	0.4000
FM_Credit	0.7933	0.8681	0.7415	0.9714	0.3778
BM_Credit	0.7800	0.8596	0.7446	0.9619	0.3556

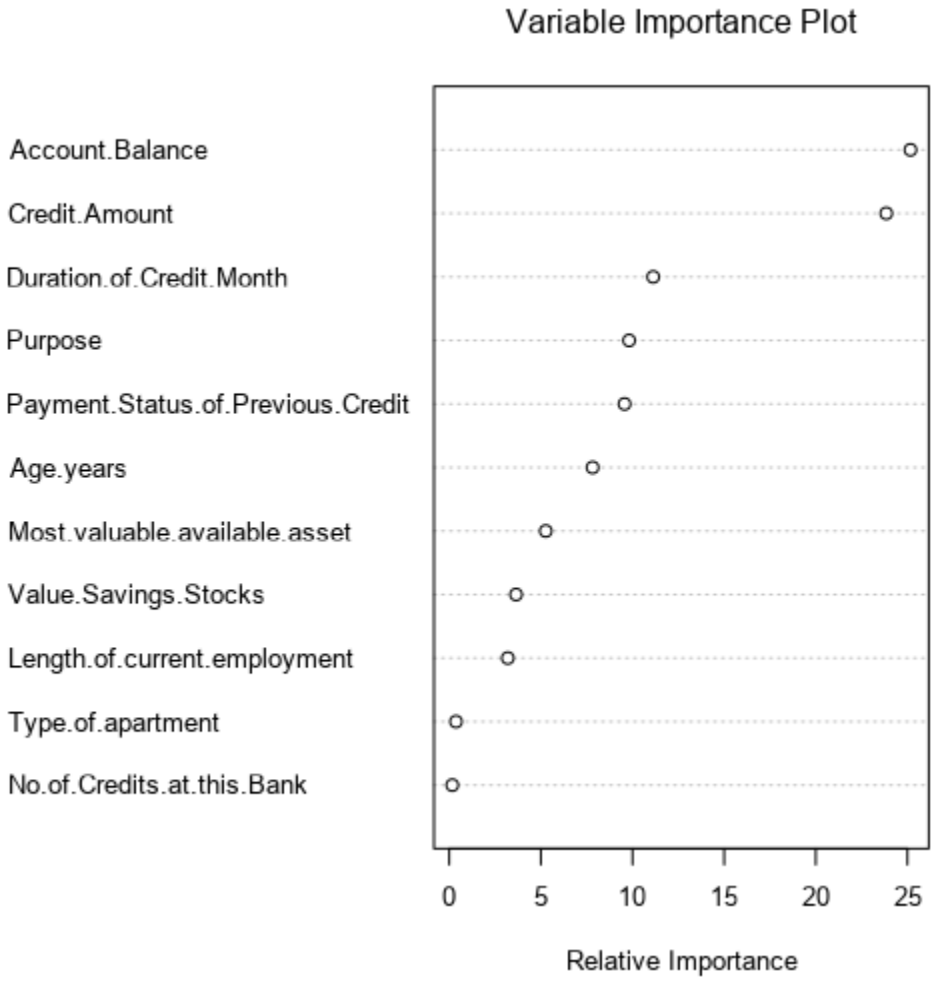
Confusion matrix of FM_Credit

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

The overall accuracy is 79% while the accuracy of creditworthy is 97.14% and the accuracy for non-creditworthy is 37.78%. You can say that the forest model has a bias towards correctly predicting creditworthy individuals even stronger than the decision tree, the logistic regression model and the boost model.

4. *Boosted Model*

Answer: I used the "credit-Application-Result" as the Target variable and select all other variables as the predictor variable. Under the section "variable importance Plot" you can see the most important variables. The top 3's is "Account Balance", "Credit Amount" and "Duration of Credit Month"



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit	0.6733	0.7721	0.6296	0.7905	0.4000
FM_Credit	0.7933	0.8681	0.7415	0.9714	0.3778
BM_Credit	0.7800	0.8596	0.7446	0.9619	0.3556

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	29
Predicted_Non-Creditworthy	4	16

The overall accuracy is 78% while the accuracy of creditworthy is 96.19% and the accuracy for non-creditworthy is 35.56%. You can say that the forest model has a bias towards correctly predicting creditworthy individuals even stronger than the decision tree and the logistic regression model.

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)

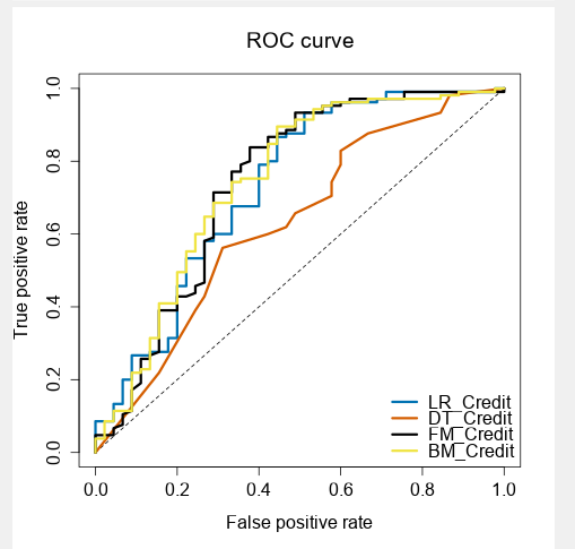
Answer these questions:

- 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

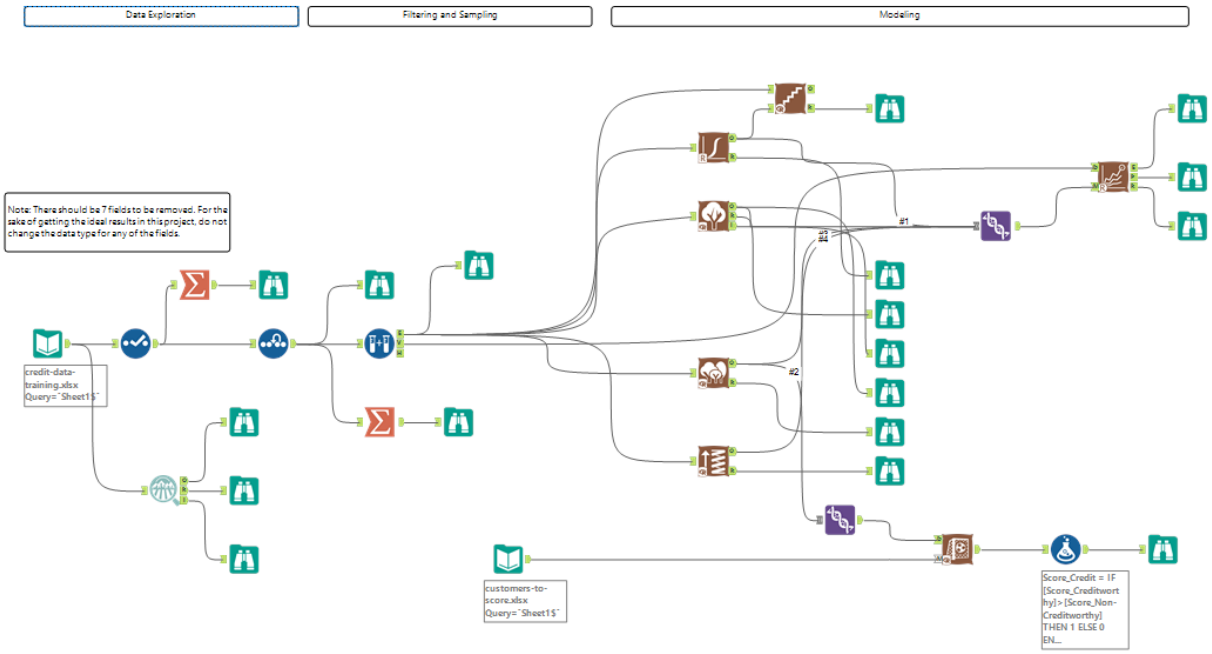
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

Answer: I have chosen the Forest Model since it has the highest accuracy with 79.33. It also has the highest accuracy regarding "Creditworthy". But unfortunately, the third lowest accuracy regarding "non creditworthy". This leads also to a bias towards correctly predicting creditworthy individuals. Regarding the "ROC curve" the Forest Model reaches the true rate at the fastest. Applying the Forest Model to the total 500 potential customers, according to this logic, 405 are creditworthy.



Alteryx Workflow:



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.