Project: Predicting Default Risk "Creditworthiness"

Step 1: Business and Data Understanding

- What decisions needs to be made?
  A small bank which typically gets 200 loan applications per week and approves them by hand. Due to a financial scandal that hit a competitive bank last week, suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!
  As a loan officer, my manager wants me to figure out how to process all of these loan applications within one week. Based on the classification models that I learned recently, I need to systematically evaluate the creditworthiness of these new loan applicants and provide a list of creditworthy customers to my manager in the next two days.

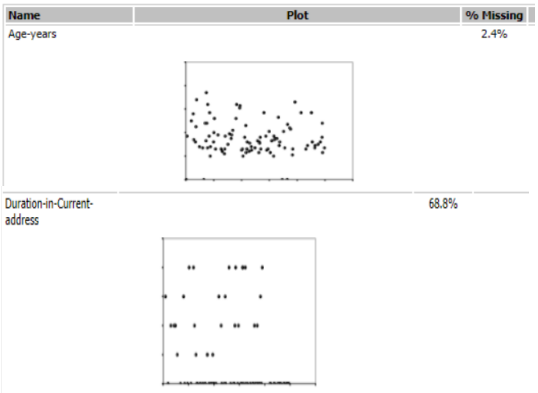
- What data is needed to inform those decisions?
  We need data on all past applications such as Credit Application Result, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Credit Amount, Value Savings Stocks, Length of current employment, and Instalment per cent, and we need new data from the 500 customers.

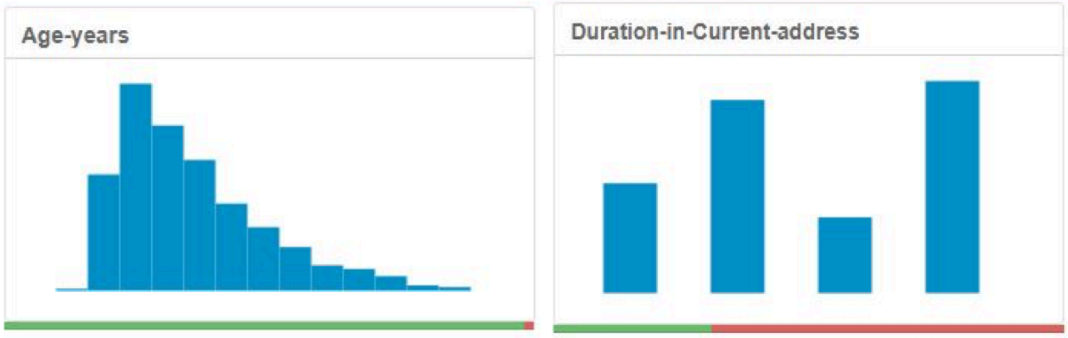
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We need to use the Binary model to help make these decisions, creditworthy and non-creditworthy customers.

Step 2: Building the Training Set

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

The Field Summary tool provided a landscape of all variables.
impute Age-years and remove Occupation, Duration-in-Current-address, Foreign-Worker, Concurrent-Credits, Guarantors, No-of-dependents, and Telephone.
Because the Duration-in-Current-address has a lot of missing values 68.8% and we impute the Age-years variable that has just 2.4% of missing values.





* Due the low-variability, we removed the fields Foreign-Worker, Concurrent-Credits, Guarantors, and No_ of_dependents.

* The Telephone field was removed because included private information and is not relevant to classification. See the graph below

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?

Logistic Regression Model:

- From the below chart, most important predictor varibles for the model are: Account.Balance, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent, Most.valuable.available.asset, respectively With low P-value.
- From the Model Comparison report show this Model has an accuracy of 76% and accuracy of 80% to predict the creditworthy.

Report

Report for Logistic Regression Model Logistic_Regression_6

Basic Summary

Call:
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)

Deviance Residuals:

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

Coefficients:

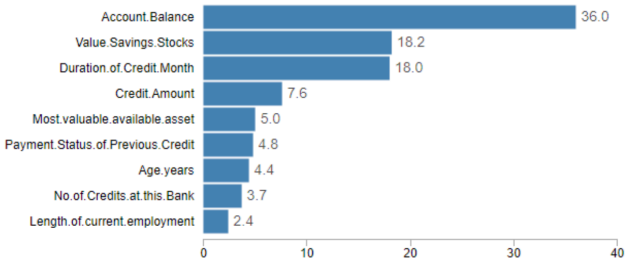
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Decision Tree Model:

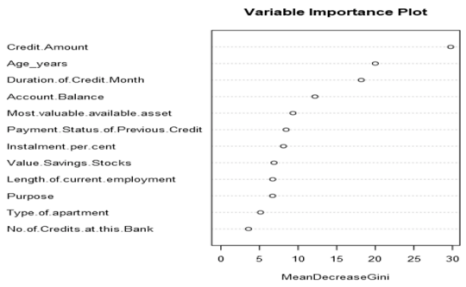
- The significant predictor variable for this model based on the variable importance report below, the top 3 predictive variables are Account Balance, Value Savings Stocks, and Duration of Credit Month.



- From the Model Comparison report show this Model has an accuracy of 74%. With accuracy of 79% to predict the creditworthy customers and non-creditworthy customers as 60%.

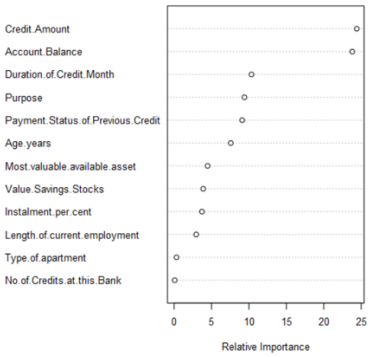
Forest Model:

- based on the variable importance plot, the top 3 important predictive variables are Credit Amount, Age years, and Duration of Credit Month.
- From the Model Comparison report show this Model has an accuracy 80%. With accuracy of predicting the creditworthy customer to be 79% and non-creditworthy customers as 82%.



Forest Model:

- based on the variable importance plot, the top 3 important predictive variables are Credit Amount, Amount Balance, and Duration of Credit Month.
- From the Model Comparison report show this Model has an accuracy 78%. With accuracy of predicting the creditworthy customer to be 78% and non-creditworthy customers as 80%.



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"

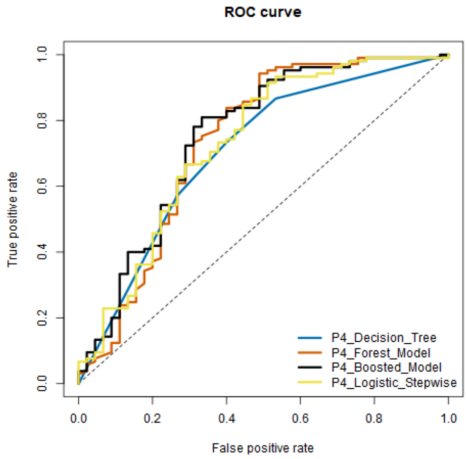
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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.

Based on the Model comparison report, the four models are biased to Creditworthy, therefore we used the overall Accuracy value,PPV, NPV and F1 score to select the highest value, it is the ***Forest Model***. With accuracy is 0.80 and highest F1 score.



Accuracy Creditworthy rate= TP/ actual yes ⇒ 101/105= 0.9619. Being a test with high True positives rate and high sensitivity, it means rarely fail diagnosis. ***Forest Model*** has little to no bias in its predictions as you'll see in the Model Comparison report for all the 4 models since it has one of the least differences in Positive Predictive Value (PPV) and Negative Predictive Value (NPV).

Also from the ROC curve the ***Forest Model*** has a positive rate. The creditworthy and non-creditworthy prediction also almost same hence this model is not biased.

Model Comparison Report for Accuracy:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
P4_Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
P4_Forest_Model	0.8000	0.8707	0.7361	0.9619	0.4222
P4_Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778
P4_Logistic_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Accuracies within “Creditworthy” and “Non-Creditworthy” segments:

Confusion matrix of P4_Boosted_Model

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

Confusion matrix of P4_Decision_Tree

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

Confusion matrix of P4_Forest_Model

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

Confusion matrix of P4_Logistic_Stepwise

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

- How many individuals are creditworthy?
  406 individuals who are creditworthy.

- **Alteryx Workflow**