Project: Creditworthiness

Step 1: Business and Data Understanding

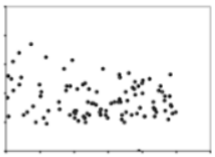
- What decisions needs to be made?
  As a loan officer at a young and small bank that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not.
  I need to use a series of classification models to figure out the best model and provide a list of creditworthy customers in the next two days.

- What data is needed to inform those decisions?
  - Data on all past applications.
  - The list of customers that need to be processed in the next few days.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Use the binary model to get a decision and the possibility of making it by identifying people taking or not.

Step 2: Building the Training Set

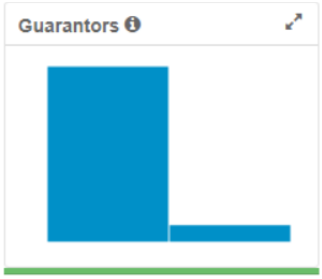
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  Using summary tool and the result is
  – age-years => 2.4% data missing and to solve the problem impute the data using the median.

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	

- Duration-in-Current-address => 68.8% data missing , should be removed.

Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150

- Occupation => has one value should be removed
- Concurrent-Credits => has (other banks/depts) value should be removed
- Guarantors => Most of the data is null should be removed.



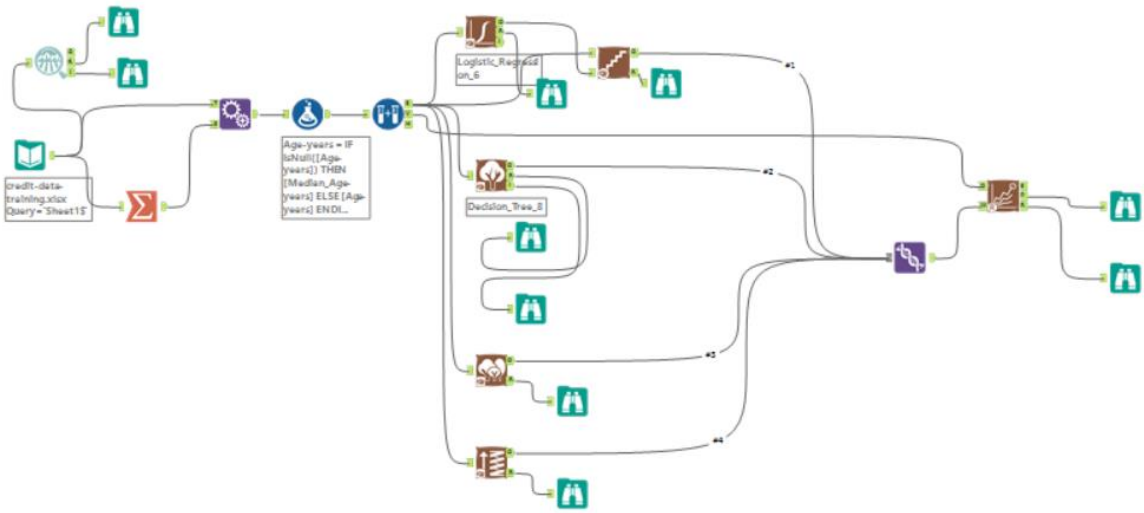
- Foreign-Worker => Most of the data is one value should be removed
- No-of-dependents => Most of the data is one value should be removed
- Telephone => It has to be removed because it doesn't do us any good.

Step 3: Train your Classification Models

- 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?

Used 4 model (Logistic Stepwise, Decision Tree, Forest Model, Boosted Model)
Workflow in alteryx:



- Logistic Stepwise: the significant predictive variables are Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment, and Installment per Cent.

Records 1 to 10

Report

Report for Logistic Regression Model Logistic_stepwise

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

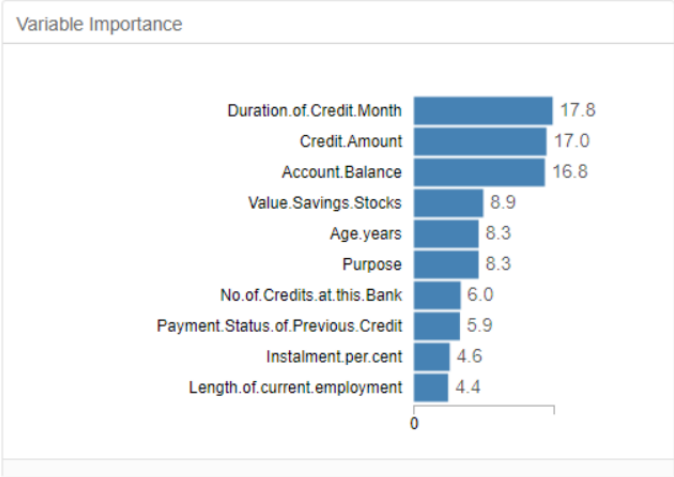
(Dispersion parameter for binomial taken to be 1)

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

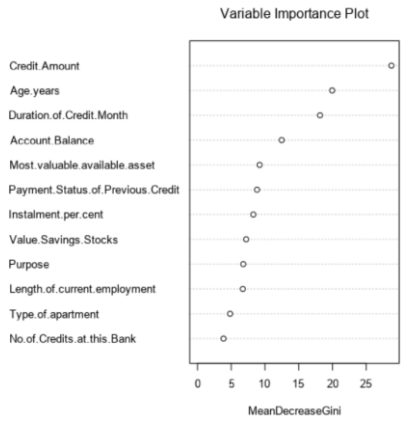
Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

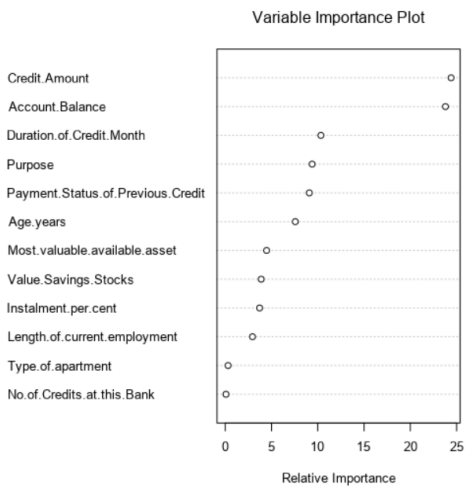
- Decision Tree: the 3 important predictive variables are Duration of Credit Month, Credit Amount, Account Balance.



- Forest Model: the 3 important predictive variables are Credit Amount, Age Years, and Duration of Credit Month.



- Boosted Model: the 3 important predictive variables are Credit Amount, Amount Balance, and Duration of Credit Month.



- then used the model comparison tool of these four models:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.6867	0.7854	0.6524	0.7544	0.4722
Forest_model	0.7933	0.8681	0.7368	0.7846	0.8500
Boosted_model	0.7867	0.8632	0.7524	0.7829	0.8095
Logistic_stepwise	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	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of Decision_Tree

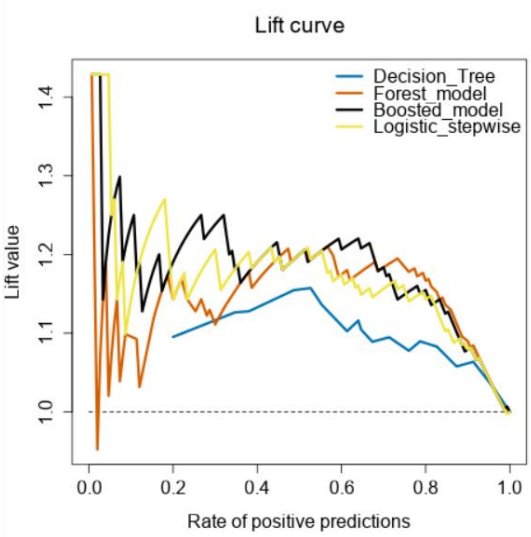
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	28
Predicted_Non-Creditworthy	19	17

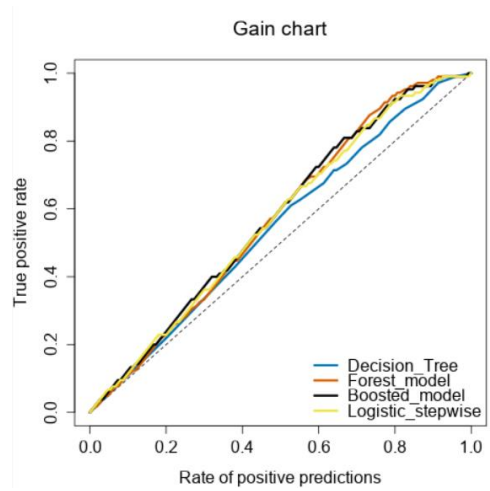
Confusion matrix of Forest_model

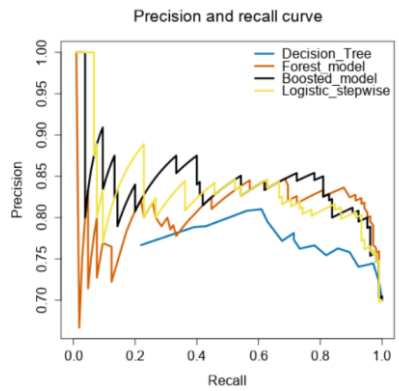
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

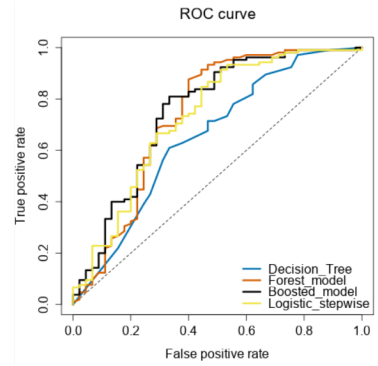
Confusion matrix of Logistic_stepwise

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









Step 4: Writeup

- 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

use Forest model considering overall accuracy, PPV, NPV and the F1 score.

model comparison tool (Overall Accuracy+ “Creditworthy” and “Non-Creditworthy” + ROC graph):

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.6867	0.7854	0.6524	0.7544	0.4722
Forest_model	0.7933	0.8681	0.7368	0.7846	0.8500
Boosted_model	0.7867	0.8632	0.7524	0.7829	0.8095
Logistic_stepwise	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	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of Decision_Tree

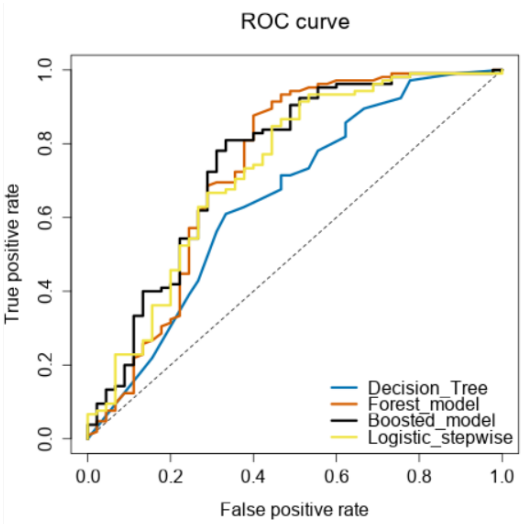
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	28
Predicted_Non-Creditworthy	19	17

Confusion matrix of Forest_model

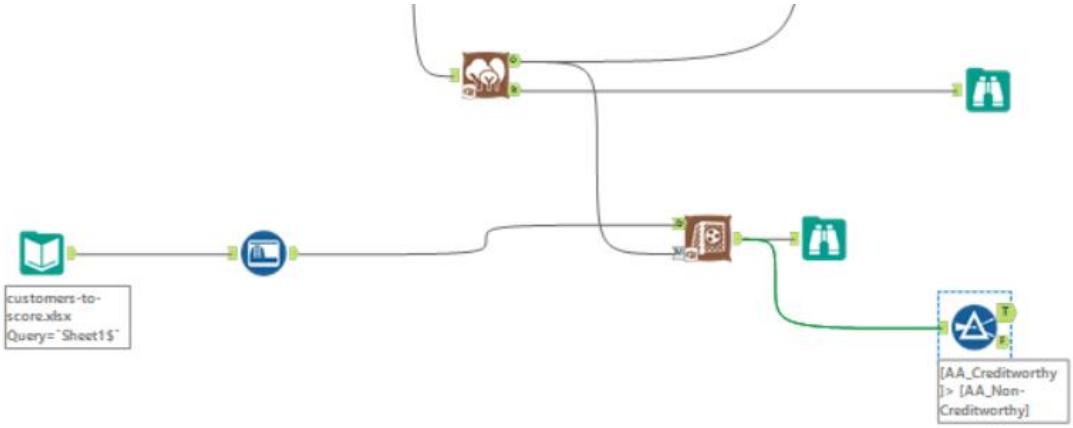
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of Logistic_stepwise

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



- How many individuals are creditworthy?



408 individuals are creditworthy.

Workflow :