Udacity Business Analyst Nanodegree Project: Creditworthiness

Jason Grenig

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?
  We need to determine whether or not a loan applicant is creditworthy.

- What data is needed to inform those decisions?
  We need to be able to accurately predict whether an individual will be creditworthy or not. This means using statistically significant fields from our dataset to create and validate a strong prediction model. For this dataset, we'll use logistic-stepwise, decision tree, forest, and boosted models, and then compare them to see which is best for our analysis.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We'll use a non-binary model, since there are more than 2 categorical variables in the dataset.

Step 2: Building the Training Set

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.
  I removed 7 fields from the dataset:
    1. Duration in Current Address - 69% missing values
    2. Concurrent-Credits - only 1 type of data-same thing for every record.
    3. Occupation - all values were "N/A".
    4. Guarantors - low variability
    5. Foreign Workers - low variability

6. No. of dependents - low variability
7. Telephone - not required to determine loan applicant creditworthiness

I also imputed the Age-years field. It was missing 2% of values. I chose to impute the median age to the missing values. Because the Age-years plot is right-skewing, median will be more accurate to use than mean.

Field Summary Report

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-Stepwise Regression Model:

Q1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
A1. The most statistically significant predictor variables are: Account.BalanceSome Balance, PurposeNew car, and Credit.Amount. These are indicated with the smallest p-values (and the greatest number of asterisks) in the below report.

Record | Report

1

Report for Logistic Regression Model log_step_credit

2 *Basic Summary*

3 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)

4 Deviance Residuals:

5

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

6 Coefficients:

7

	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)

8 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

Logistic-Stepwise Regression Model Report

Q2. 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?

A2. This model has an overall accuracy of 76%, with an 88% accuracy for predicting creditworthy individuals, and a 49% accuracy for predicting non-creditworthy individuals. The confusion matrix shows there are roughly 9x more records that are actually creditworthy vs non-creditworthy. This imbalance leads to a bias in the model's ability to predict creditworthy more accurately than non-creditworthy.

Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
log_step_credit	0.7600	0.8364	0.7306	0.8762	0.4889

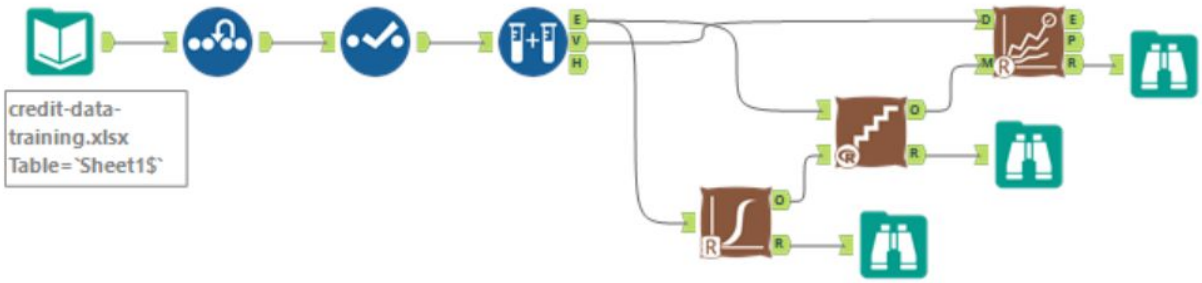
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of log_step_credit

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

Logistic-Stepwise Model Comparison Report

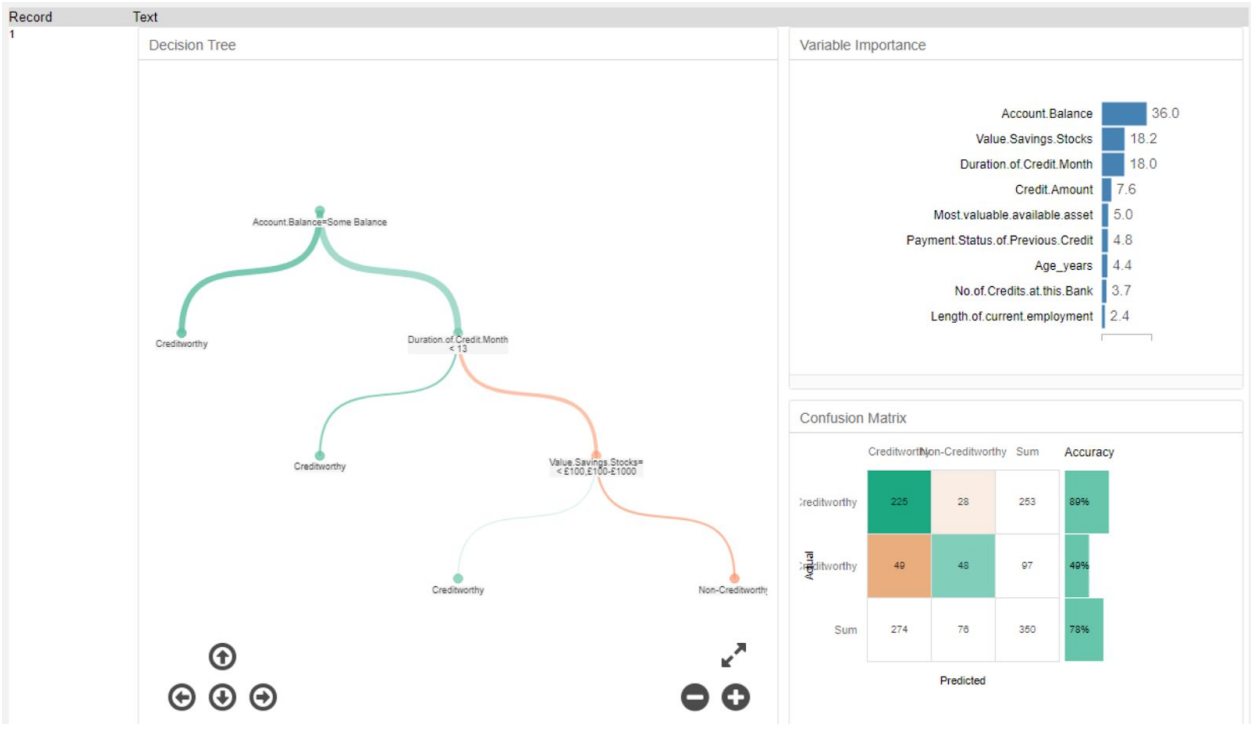


Alteryx Workflow for Logistic-Stepwise Model

2. Decision Tree Model:

Q1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

A1. In the Decison Tree model, the most significant predictor variables are: Account Balance, Value.Savings.Stocks, and Duration.of.Credit.Month.



Decision Tree Interactive Report

Q2. 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?

A2. The overall accuracy is 74%, with an 87% accuracy of predicting creditworthy, and a 46% accuracy of predicting non-creditworthy individuals. Again, the model is biased from a greater number of records for creditworthy individuals. This model performed slightly worse than the Logistic-Stepwise model.

Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667

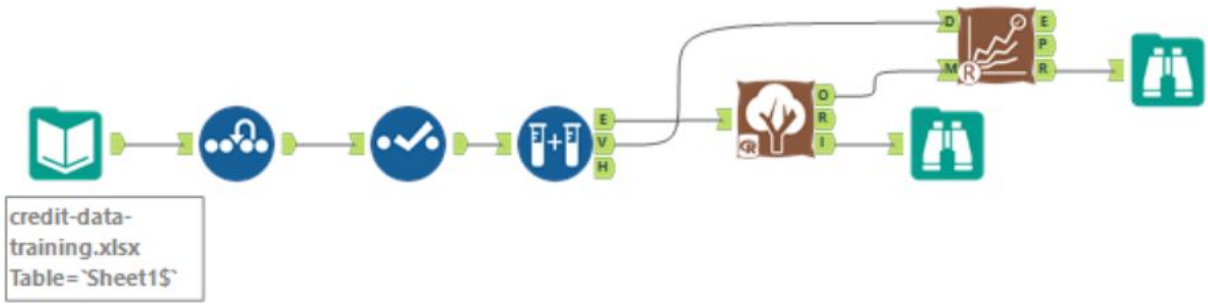
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of DT_Credit

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

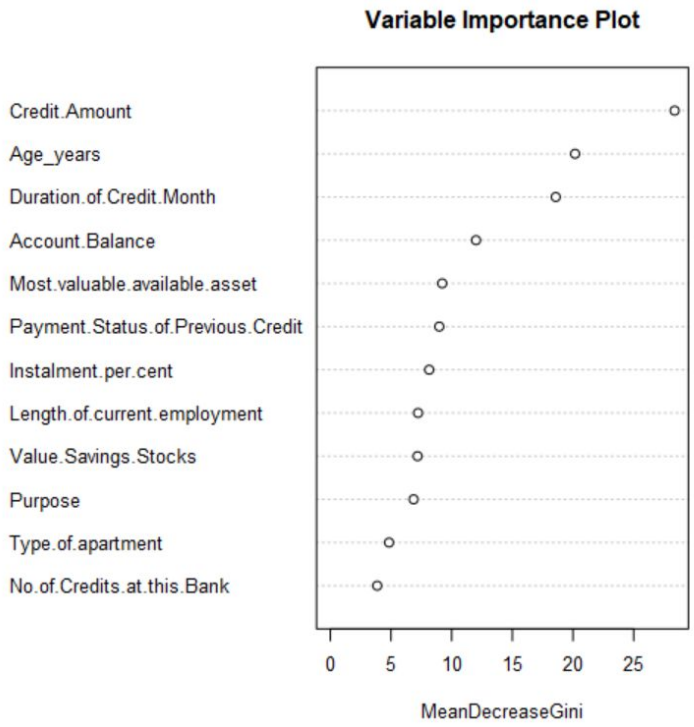
Decision Tree Model Comparision Report



Alteryx Workflow for Decision Tree Model

3. Forest Model:

Q1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
A1. According to the Forest model, Credit.Amount, Age_years, and Duration.of.Credit.Month are the most significant predictor variables. Interestingly, Account.Balance was the 4th most significant predictor variable, after being the most important in the Decision Tree and Logistic-Stepwise models.



Forest Model Variable Importance Plot

Record	Report
1	*Basic Summary*
2	Call: randomForest(formula = 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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years, data = the.data, ntree = 500, replace = TRUE)
3	Type of forest: classification Number of trees: 500 Number of variables tried at each split: 3
4	OOB estimate of the error rate: 24%
5	Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.639	62	35

Forest Model Confusion Matrix

Q2. 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?

A2. The Forest model has an overall accuracy of 80%, with a 96% accuracy in predicting creditworthiness, and a 42% accuracy in predicting non-creditworthiness. The ratio of creditworthy to non-creditworthy records is a bit higher in this model than the Decision Tree or Logistic-Stepwise, which skews the ability of the model to predict the target variable. This model has the highest overall accuracy so far though.

Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_credit	0.8000	0.8707	0.7361	0.9619	0.4222

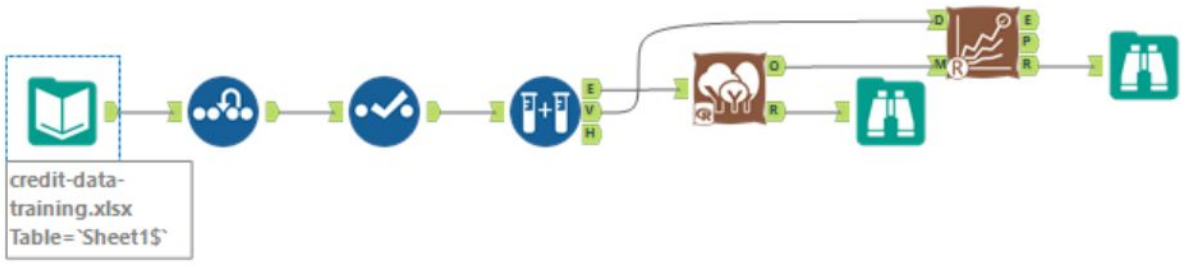
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Forest_credit

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

Forest Model Comparison Report

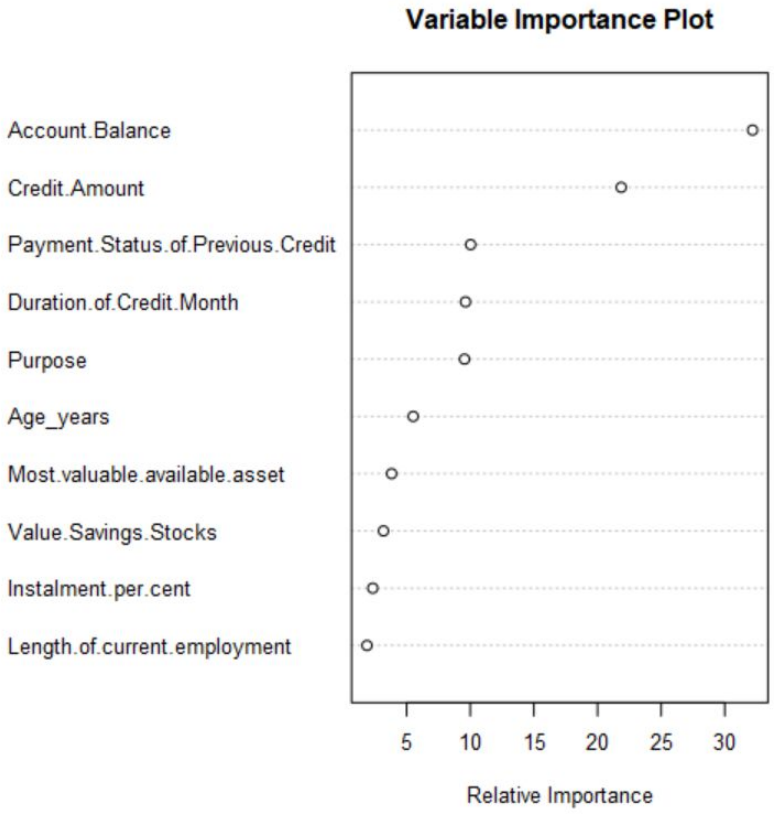


Alteryx Workflow for Forest Model

4. Boosted Model:

Q1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

A1. In the Boosted model, Account.Balance and Credit.Amount are the most important predictor variables. There is a bit of drop off before the next tier of variables on the Relative Importance axis.



Boosted Model Variable Importance Plot

Q2. 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?

A2. The overall accuracy for the Boosted model is 79%, slightly worse than the Forest model. Similarly to the Forest model, the Boosted model accurately predicted 101 individuals to be creditworthy. The Boosted model made 2 less predictions on non-creditworthy individuals, which gave it just a 38% accuracy vs 42% accuracy for the Forest model. Again the number of records for creditworthy vs non-creditworthy contributed to the model's bias in more accurately predicting creditworthy individuals.

Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_credit	0.7867	0.8632	0.7524	0.9619	0.3778

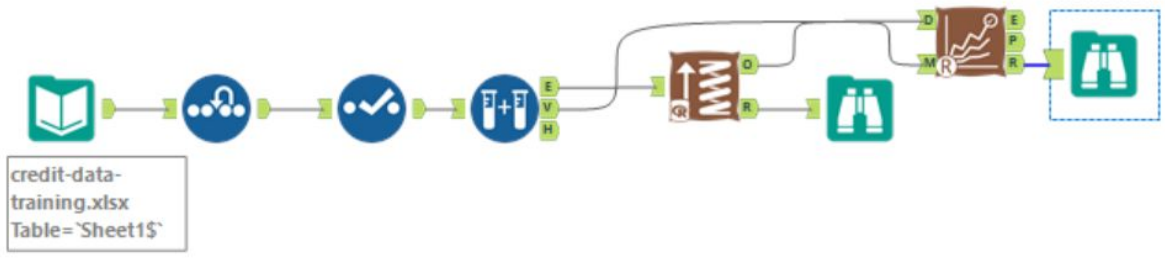
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Boosted_credit

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

Boosted Model Comparison Report



Alteryx Workflow for Boosted 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"

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

I chose the Forest model. The Forest model has the highest overall accuracy at 80%, slightly better than the Boosted model which had 79%. The Forest model was able to accurately predict 2 more non-creditworthy individuals than the Boosted model. Both the Decision Tree and Logistic-Stepwise models had higher prediction accuracy for non-creditworthy loan applicants. However these models were only able to predict creditworthy applicants at 87% and 88% accuracy respectively. This is much less than the Forest model which had 96% accuracy in predicting creditworthy applicants.

All models faced the same bias - having many more records for creditworthy than non-creditworthy individuals. Therefore no model was more biased than another.

Based on the ROC curve (below), the Forest model reaches the top the fastest, slightly faster than Boosted. This solidifies my choice to use the Forest model in production.

Record | Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_credit	0.8000	0.8767	0.7361	0.9619	0.4222
Boosted_credit	0.7867	0.8632	0.7524	0.9619	0.3778
log_step_credit	0.7600	0.8364	0.7306	0.8762	0.4889

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

3

Confusion matrix of Boosted_credit

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

4

Confusion matrix of DT_Credit

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

5

Confusion matrix of Forest_credit

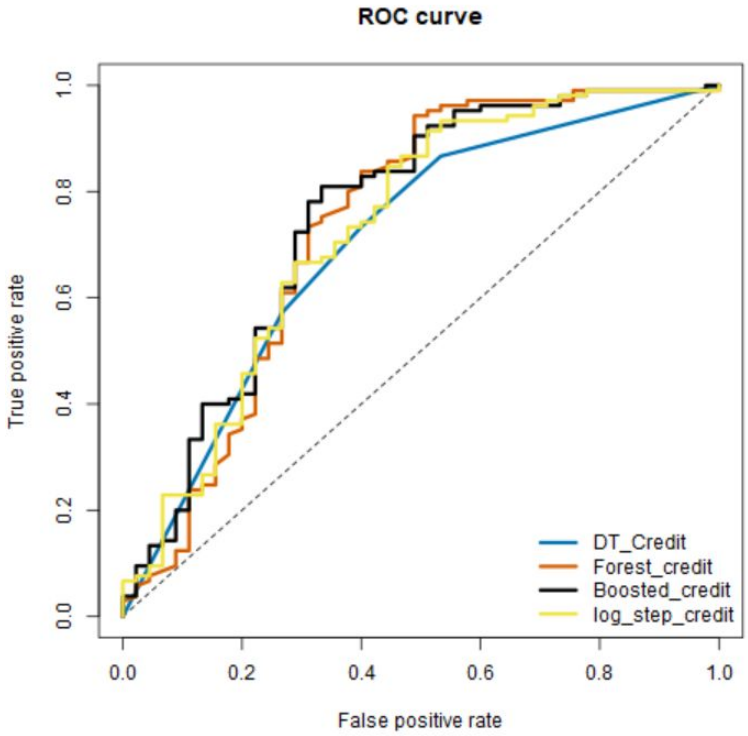
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

6

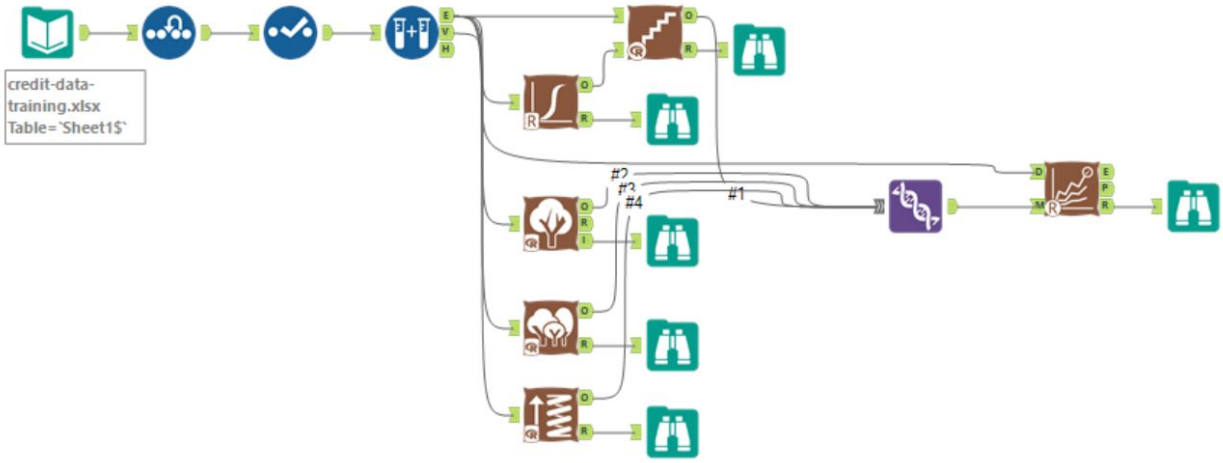
Confusion matrix of log_step_credit

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

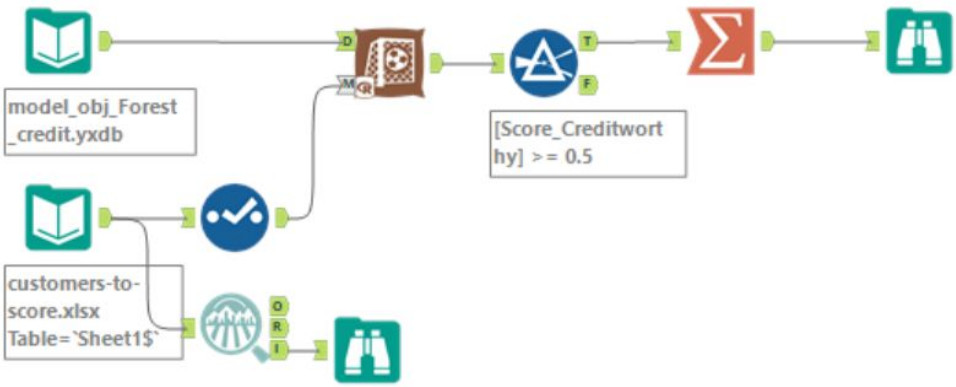
Model Comparison Report for All 4 Models



ROC Curve for All 4 Models



Alteryx Workflow for Model Comparison of All 4 Models



Alteryx Workflow for Scoring the Forest Model

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

- How many individuals are creditworthy?
  The Forest model predicts that 373 of 500 individuals are creditworthy (74.6% of total).