Project: Creditworthiness

Business and Data Understanding

Ques 1: What decisions needs to be made?
Answer: The decision that needs to be taken here is to classify the customers either as creditworthy or non-creditworthy. And based on that we need to classify the customers into two categories.

Ques 2: What data is needed to inform those decisions?
Answer: The data that we have is: credit-data-training.xlsx which contains the data of the customers to whom bank has provided the loan to and based on that data we can make our predictive model to analyze the customers-to-score.xlsx data set and can categorize the customers into creditworthy and non-creditworthy.
The variables which will be useful in deciding the creditworthiness of the customer will be:
Account Balance, Credit Amount, Payment Status of Previous Credit, PurposeNew car, Value Savings Stocks, Age_Years, Duration of Credit Month

Ques 3: What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
Answer: The model that we need to build for this problem is Binary model as we need to decide whether the customer is creditworthy or not creditworthy.

Building the Training Set

Ques 1: In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
Answer: The data that was removed:

Occupation, Concurrent credit (having only one type of data)
Guarantors, Foreign workers, No. of dependents (low variability)
Duration in current Address (69% missing values)
Telephone (not required to decide the creditworthiness of the customer)

The *'Age'* field is imputed as it has 2% of missing values and instead of imputing if we remove these 2% values it may have an adverse effect on the other attributes thus this value is imputed by replacing the *'null'* values with the *'median'* values. We took median because all the data for age is shifted towards left.

Train your Classification Models

Logistic- Stepwise Regression

Ques1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Answer: From the below chart it can be observer that the most important predictor variables for the logistic stepwise regression model is *Account.BalanceSome Balance, PurposeNew car, Credit.Amount.* The p-values of all these variables can be observed from the below table.

Report for Logistic Regression Model X

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(config$Link), 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, AIC: 352.5

Report for the Logistic Regression

Ques2: 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?

Answer: Below is the model comparison report for the Logistic-Stepwise Regression which shows that this model has an accuracy of 76%. With the accuracy of 80% to predict the creditworthy

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseRegression_Credit	0.7600	0.8364	0.7306	0.8000	0.6286

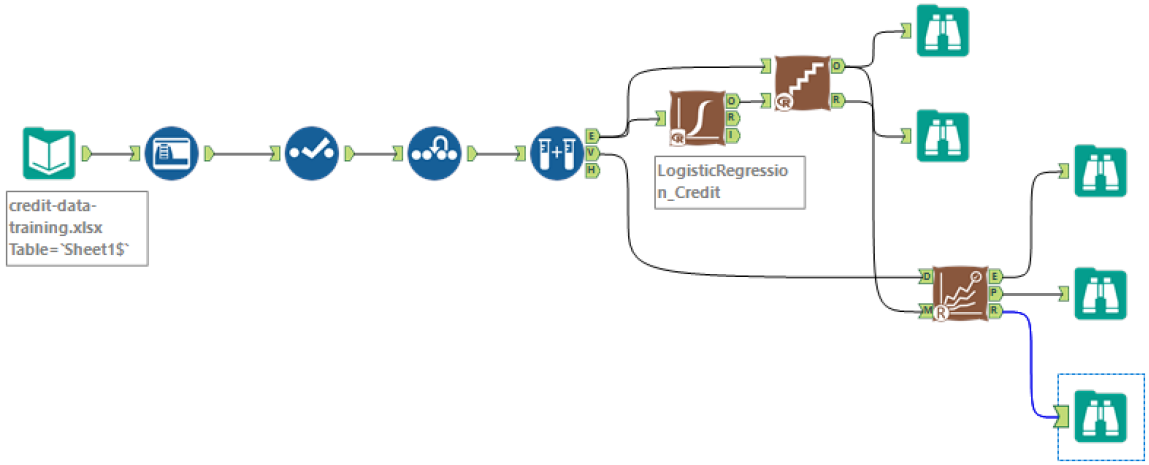
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of 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)

3

Confusion matrix of StepwiseRegression_Credit

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

Model Comparison Report for Logistic-Stepwise Regression Model

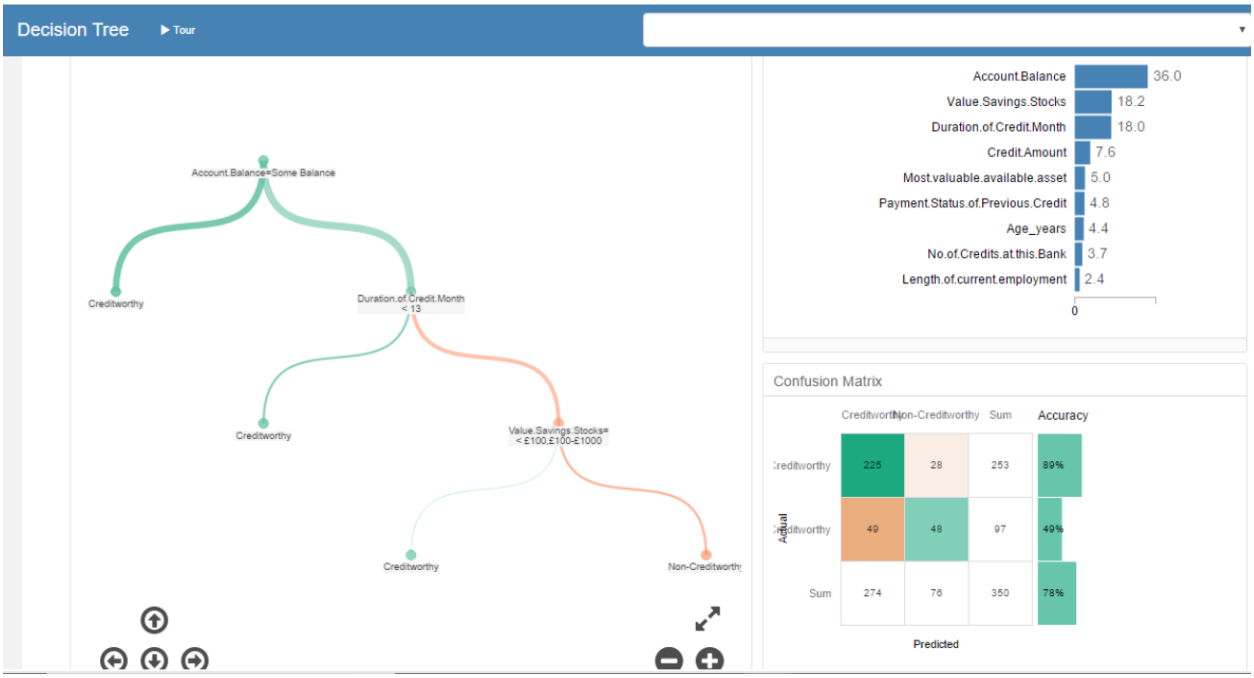


Alteryx Work Flow for the Logistic-Stepwise Regression

Decision Tree Model

Ques1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Answer: The significant predictor variable for Decision Tress Model can be observed from the variable importance graph and these are: *Account Balance, Value Savings Stocks and Duration of Credit Month.*



Interactive Report for the Decision Tree Model

Ques2: 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?

Answer: Below is the model comparison report of the Decision tree and as per the report the accuracy of this model is 74% with the accuracy of predicting creditworthy customers as 79% and non-creditworthy customers as 60%.

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree_Credit	0.7467	0.8273	0.7054	0.7913	0.6000

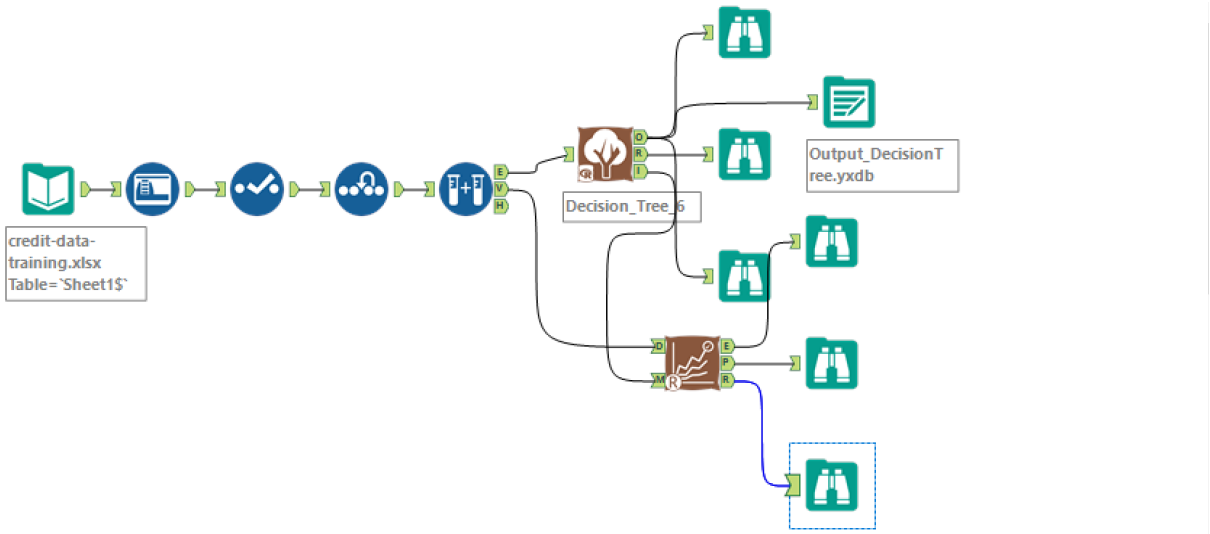
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of 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)

3

Confusion matrix of DecisionTree_Credit

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

Model Comparison Report for Decision Tree-Model



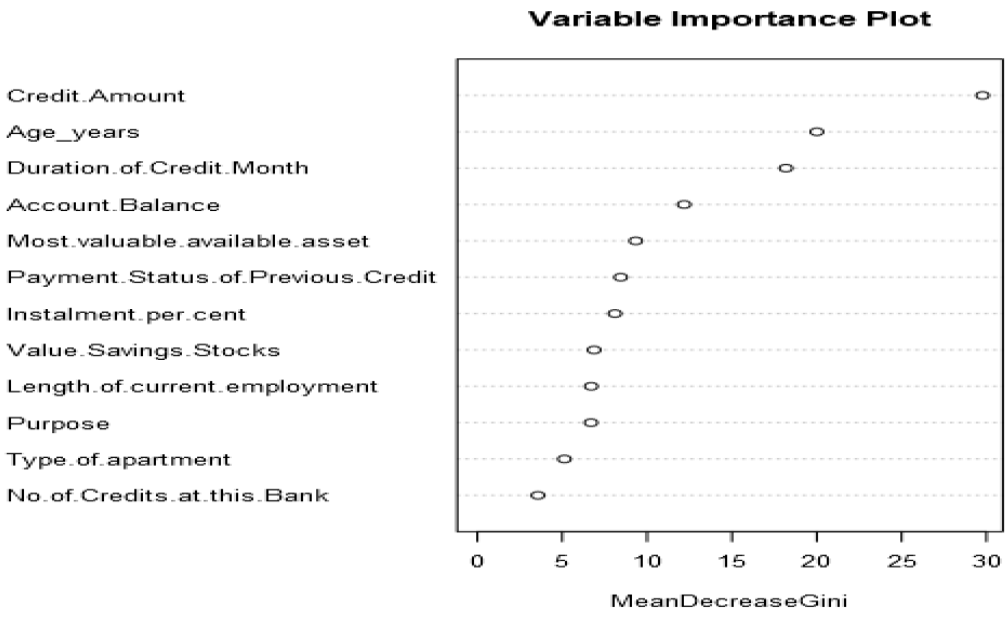
Alteryx Work Flow for the Decision Tree Model

Forest Model

Ques1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Answer: From the variable importance plot of the forest model we can infer that the most important predictor variables for this model are:

Credit Amount, Age_Years, Duration of Credit Month



Variable Importance Graph for the Forest Model

Ques2: 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?

Answer: The model comparison report shows that the accuracy of this model is 80% with accuracy of predicting creditworthy customers to be 79% and non-creditworthy customers to be 82%.

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
ForestModel_Credit	0.8000	0.8707	0.7419	0.7953	0.8261

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)

3

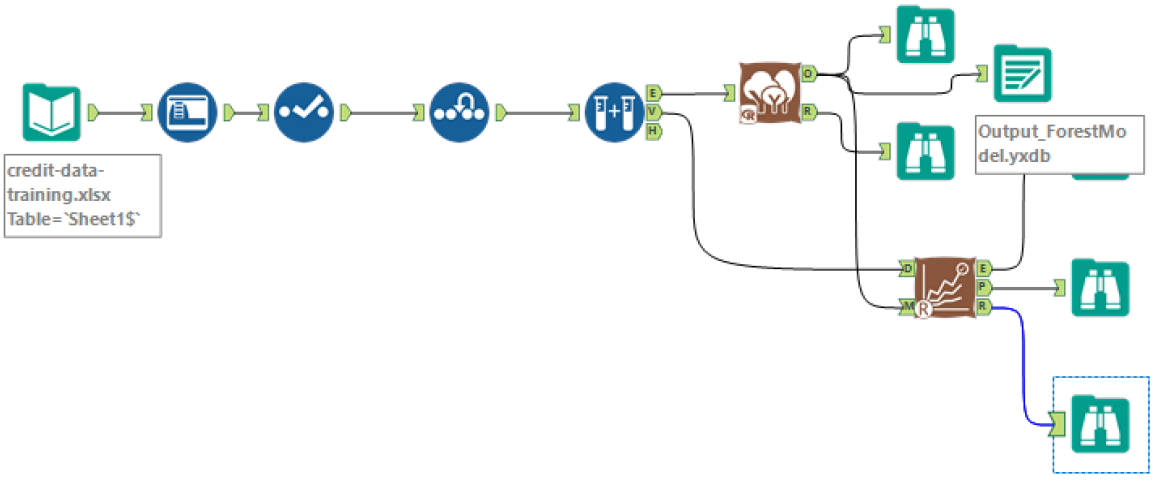
Confusion matrix of ForestModel_Credit

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

4

Performance Diagnostic Plots

Model Comparison Report for Forest Model

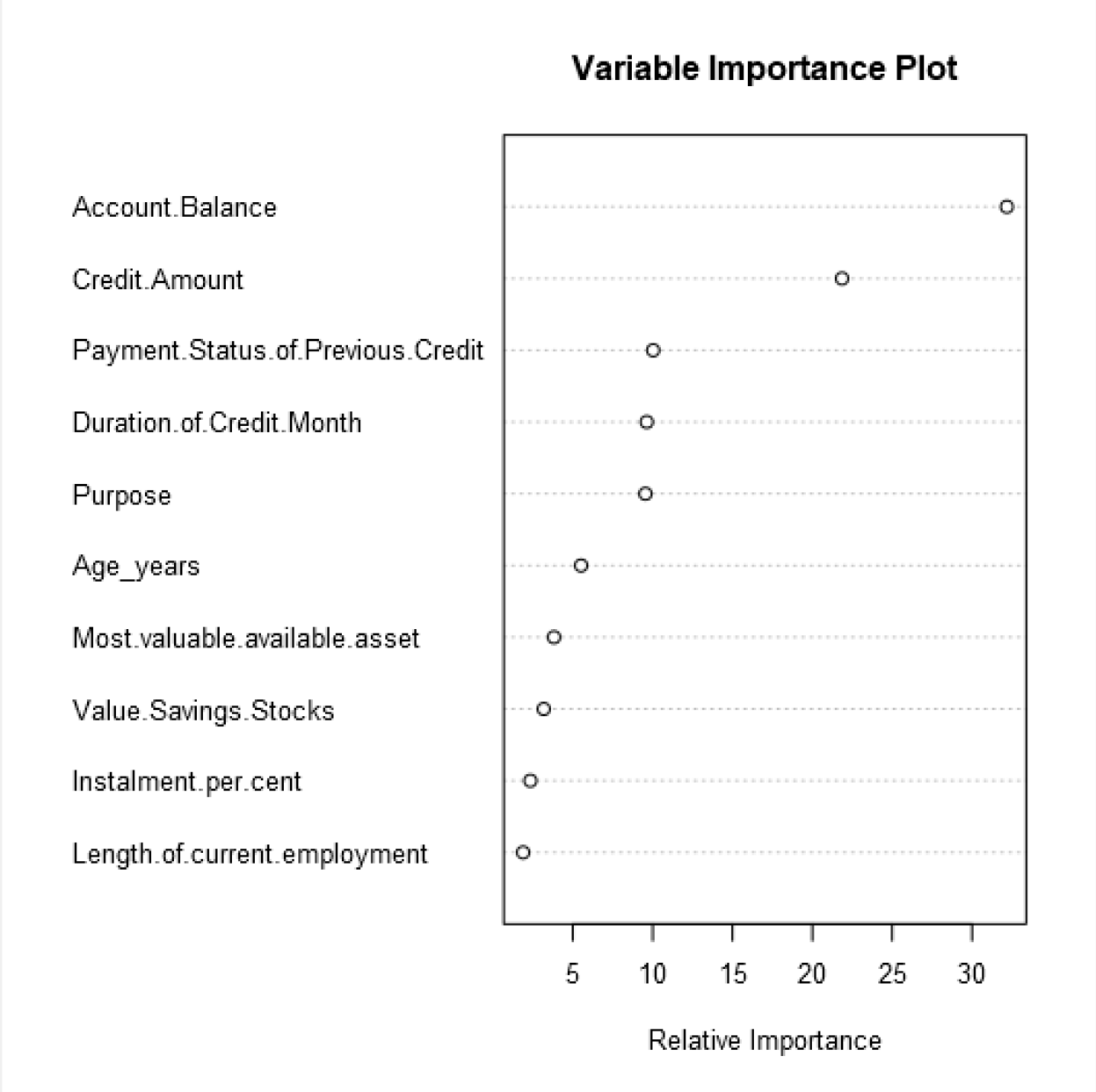


Alteryx Work Flow for the Forest Model

Boosted Model

Ques1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Answer: From the below variable importance plot the three most important predictor variables for Boosted Model is *Account Balance, Credit Amount, Payment Status of Previous Credit*



Variable Importance Graph for the Boosted Model

Ques2: 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?

Answer: The model comparison report shows that the accuracy of this model is 78% with accuracy of predicting creditworthy customers to be 78% and non-creditworthy customers to be 80%.

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BoostedModel_Credit	0.7867	0.8632	0.7524	0.7829	0.8095

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)

3

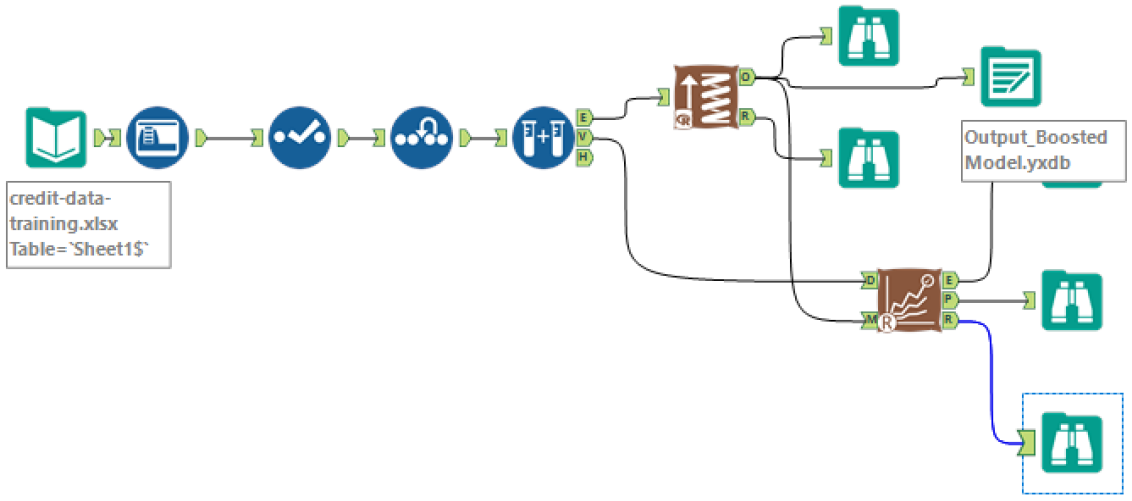
Confusion matrix of BoostedModel_Credit

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

4

Performance Diagnostic Plots

Model Comparison Report for Boosted Model



Alteryx Work Flow for the Boosted Model

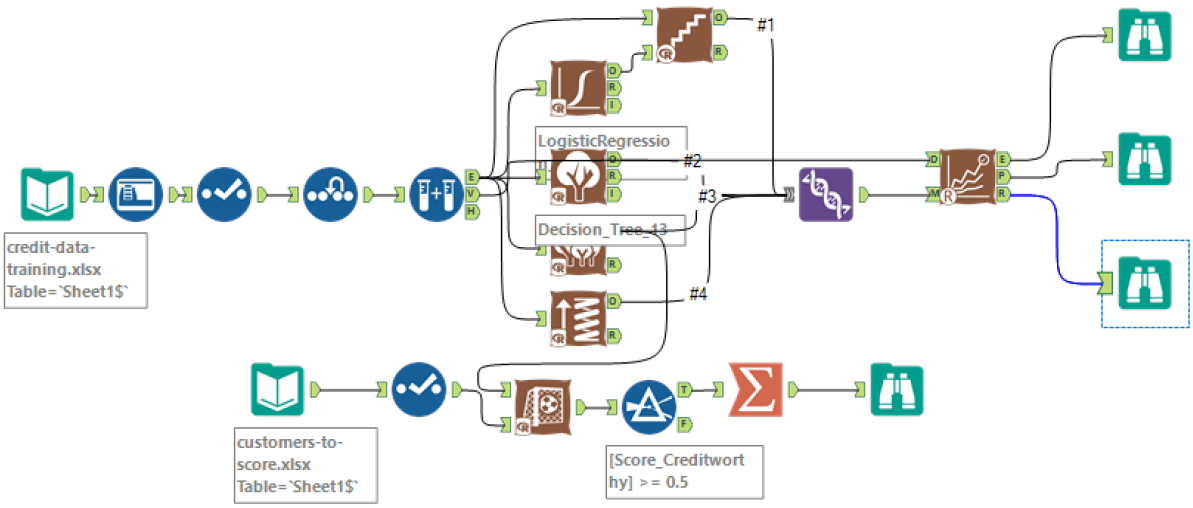
Writeup

Ques1: 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

Answer:
From the below model comparison Report we can observe that FM_Credit i.e. Forest Model is the model with the highest accuracy hence we will choice the same for the prediction analysis. The bias between the creditworthy and non-creditworthy prediction is also almost same hence this model is not biased.
Also, from the ROC curve we can deduce that Forest model reached the positive rate fastest hence this also gives the good reason to select it.

Ques2: How many individuals are creditworthy?
Answer: At the last we score the forest model with our customer_to_Score dataset and finds out that there are 416 customers which are creditworthy from the given dataset.



Alteryx Work Flow for the Comparison of all Models

Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseRegression_Credit	0.7600	0.8364	0.7306	0.8000	0.6286
DecisionTree_Credit	0.7467	0.8273	0.7054	0.7913	0.6000
ForestModel_Credit	0.8000	0.8707	0.7419	0.7953	0.8261
BoostedModel_Credit	0.7867	0.8632	0.7524	0.7829	0.8095

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)

3

Confusion matrix of BoostedModel_Credit

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

4

Confusion matrix of DecisionTree_Credit

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

5

Confusion matrix of ForestModel_Credit

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

6

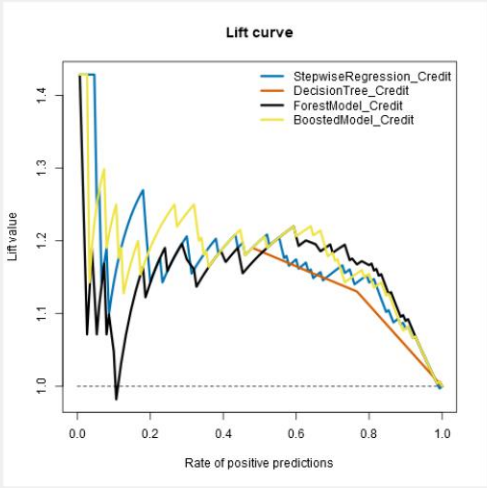
Confusion matrix of StepwiseRegression_Credit

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

7

Performance Diagnostic Plots

8



Model Comparison Report for All Models