Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made? Normally the company processes 200 loan applications per week manually. Due to recent changes in the business landscape, the loan department needs to process 500 loan applications this week. The business decision that needs to be made is which of the 500 applicants are creditworthy and a loan application can be approved.

- What data is needed to inform those decisions? Data needed to make this decision consist of historical data on previous loan applications like: account balance, previous application results, concurrent credits, and other financial data associated with loan applications.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions? In order for us to determine whether a loan application is creditworthy on non-creditworthy we will need to develop a binary classification model.

Step 2: Building the Training Set

First, I conducted an association analysis looking for highly correlated fields. The results of the analysis are shown below in the correlation matrix. There were no two fields shown to be highly correlated (.70 or greater) with each other.

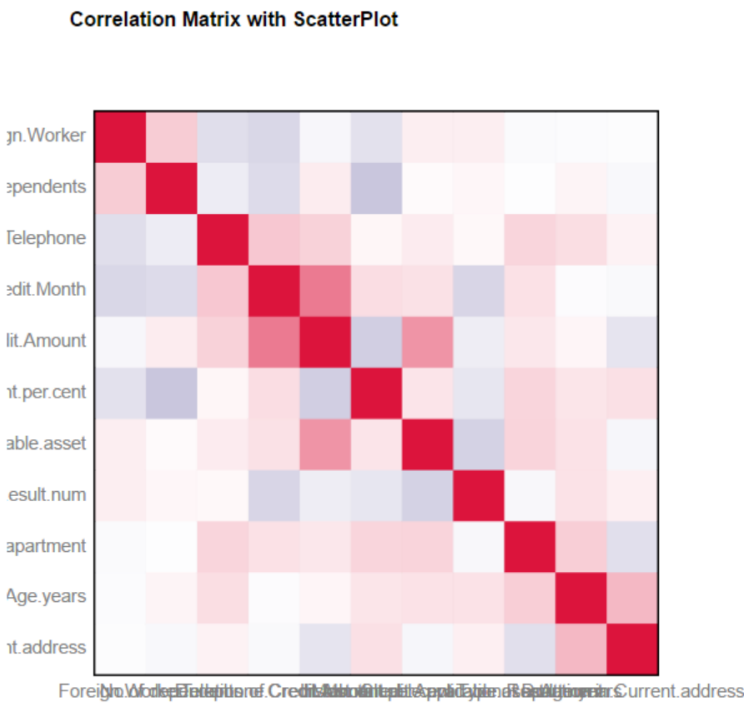


Figure 1: Correlation Matrix

Using the Field Summary Tool, I discovered there were two fields missing data:

- Duration in current address is missing 69% of its data. Given this high volume of missing data I will exclude it from analysis.
- Age years is also missing data. Given that it is only missing 2% of its data I will impute the average of 36 years into the missing fields.

Summarizing the data also shows several fields with either low variability or data that is not uniform due to a single value for the field. Therefore, in order not to skew the analysis these fields will be removed from the analysis.

- Concurrent Credits and Occupation have only one value for the field.
- Guarantors, Foreign Worker and No of Dependents show low variability because a large portion of the data is skewed to one value.

Additionally, telephone will not add any value to the analysis of creditworthiness and will therefore also be removed from the analysis.

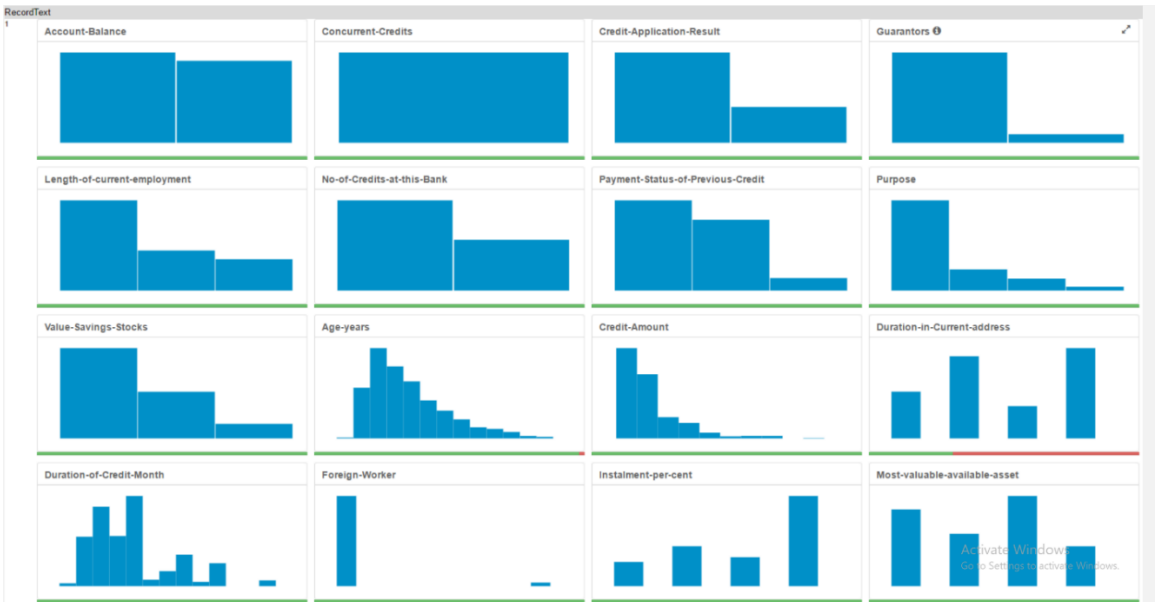


Figure 2: Field Summary of variables

Step 3: Train your Classification Models

Stepwise Logistic Regression:

1. Setting credit application result as the target variable, I ran the regression against the remaining variables and the results in figure 3 were returned. We can see that there are several variables with p-values below the .05 mark. They are: account balance, purpose, credit amount, payment status, length of current employment, and installment per cent.

Report for Logistic Regression Model StepwiseRegression_DefaultRisk

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)

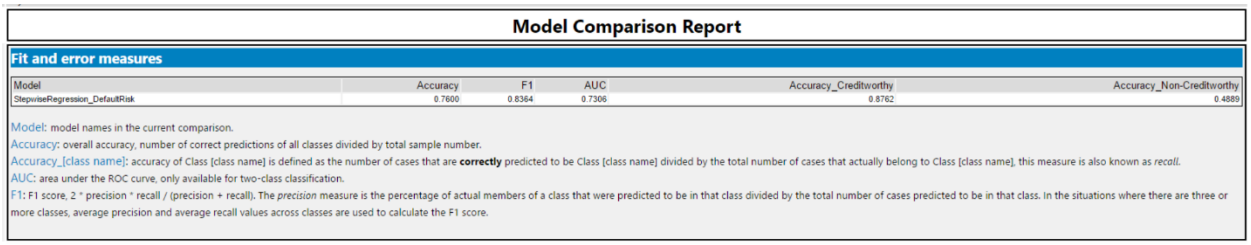
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Figure 3: Logistic Regression Report

2. When we look at how the model performed against the validation set, we see that the model has an accuracy of 76% but is biased towards predicting the applicants as non-creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseRegression_DefaultRisk	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Figure 4: Logistic Regression Model Comparison

Decision Tree:

1. Keeping the same target variable credit application result and running the decision tree model on the remaining variables we get the most important variables. The top three are account balance, duration of credit month, and credit amount.

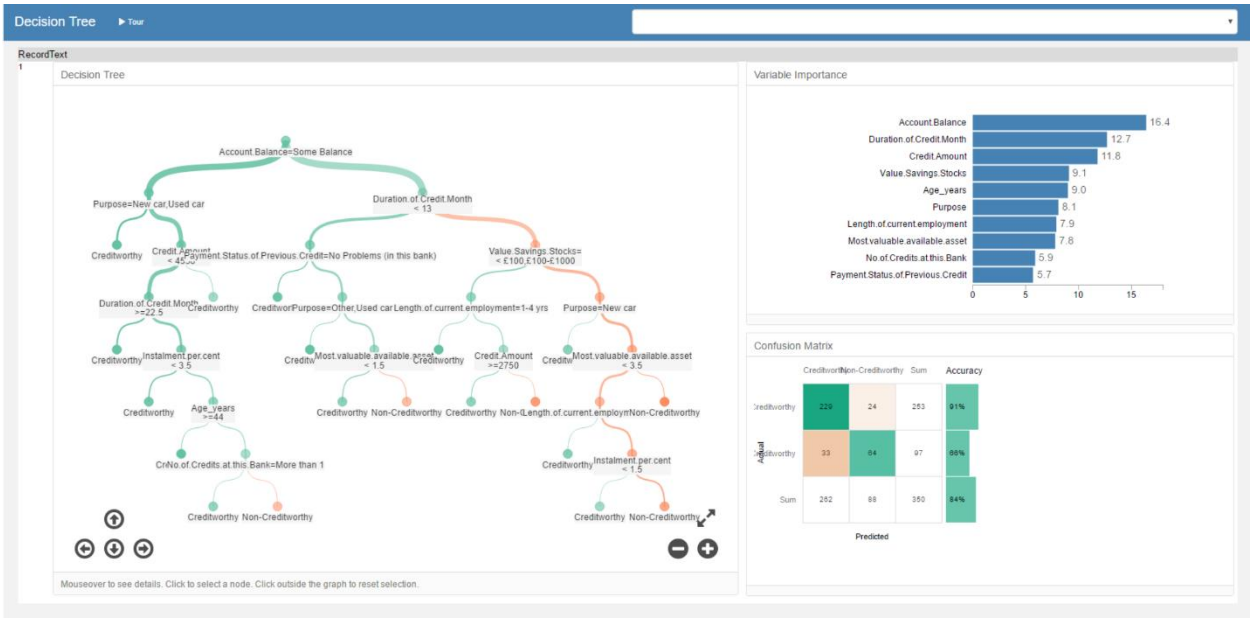


Figure 5: Decision Tree, Confusion Matrix, Variable Importance

2. The model comparison gives an overall accuracy of approx. 67% with creditworthy accuracy coming in at 79% and non-creditworthy approx. 38%. This model is also showing bias towards rating applicants as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree_DefaultRisk	0.6667	0.7685	0.6272	0.7905	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Figure 6: Decision Tree Model Comparison

Forrest Model:

1. In the Forrest Model credit application result was set as the target variable and the model was ran against the remaining variables. This model gives us credit amount, age_years, duration of credit month and account balance as the most important variables.

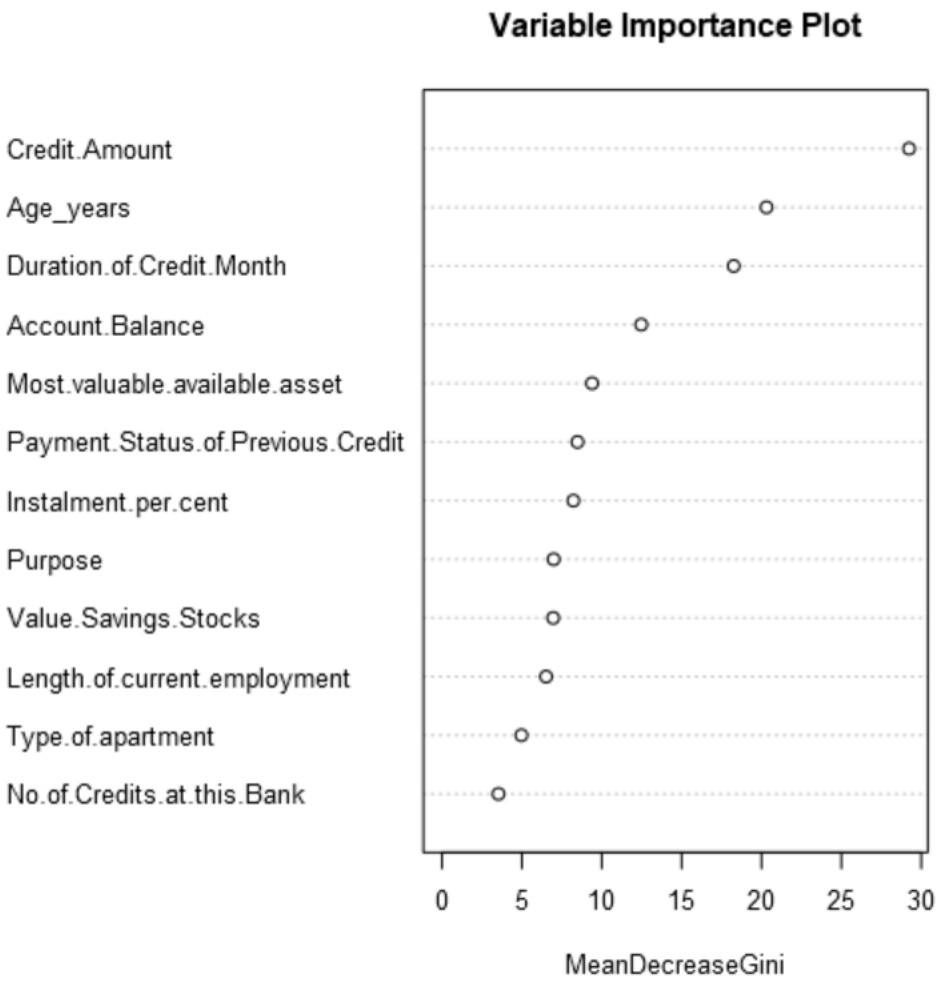


Figure 7: Variable Importance Plot

2. The model comparison gives an overall accuracy of 79% with creditworthy accuracy coming in at 96% and non-creditworthy 40%. This model is also showing bias towards rating applicants as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
ForrestModel_DefaultRisk	0.7933	0.8670	0.7428	0.9619	0.4000

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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Figure 8: Forrest Model Comparison Report

Boosted Model:

1. In the Boosted Model credit application result was set as the target variable and the model was ran against the remaining variables. This model gives us account balance and credit amount as the most important variables.

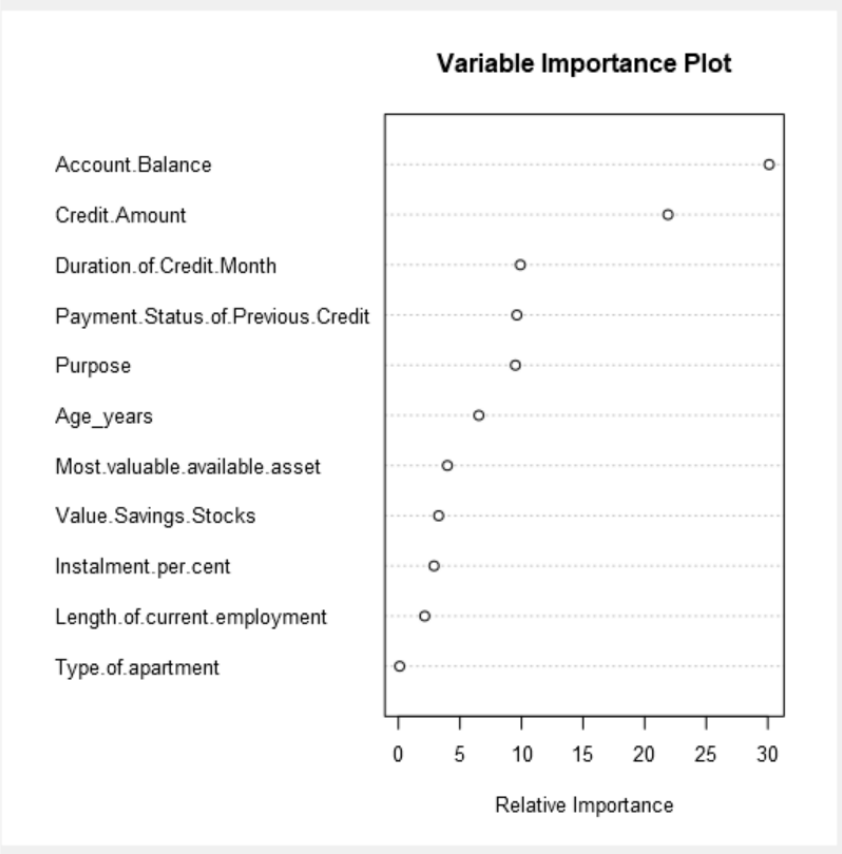
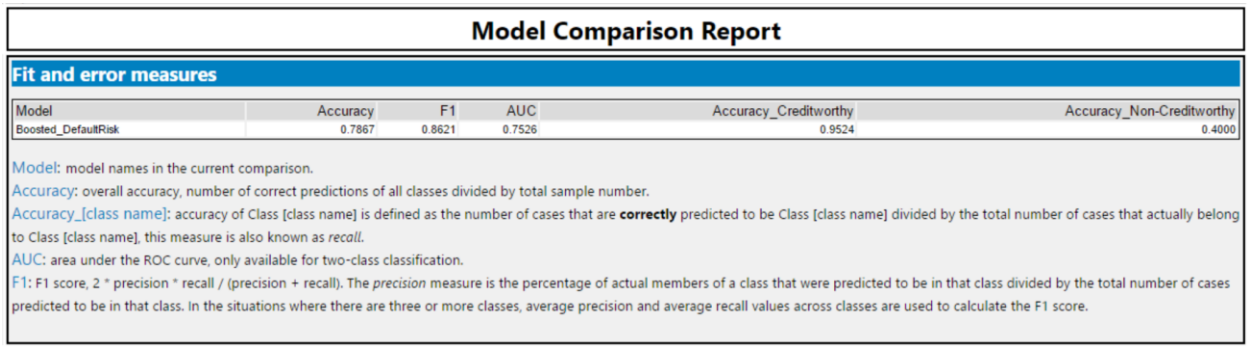


Figure 9: Boosted Model Important Variables

2. The model comparison gives an overall accuracy of just under 79% at 78.67% with creditworthy accuracy coming in at 95.25% and non-creditworthy 40%. This model is also showing bias towards rating applicants as non-creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_DefaultRisk	0.7867	0.8621	0.7526	0.9524	0.4000

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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Figure 10: Boosted Model Comparison Plot

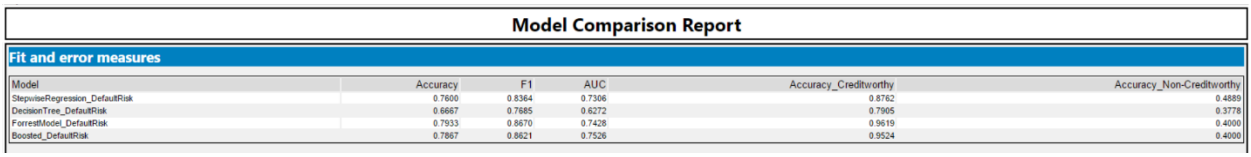
Step 4: Writeup

The Forrest Model had the highest accuracy at 79.33% when used on the validation set. It also had the highest accuracy in predicting creditworthy applications and second highest accuracy predicting non-creditworthy.

Although the Forrest Model showed bias towards predicting applicants as non-creditworthy, all models showed some bias in that direction. This could be looked at two ways, a model with bias towards non-creditworthy applicants could help in mitigating risk to the bank, but at the same time it could run the risk of not approving an application for an otherwise good customer.

Looking at the ROC it is evident that the Forrest Model reaches the highest true positive rate before the other models.

After choosing the Forrest Model to score the 500 new applicants, the model returned **412** applicants that **should be** approved for a loan.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepwiseRegression_DefaultRisk	0.7600	0.8364	0.7306	0.8762	0.4889
DecisionTree_DefaultRisk	0.6667	0.7685	0.6272	0.7905	0.3778
ForrestModel_DefaultRisk	0.7933	0.8670	0.7428	0.9619	0.4000
Boosted_DefaultRisk	0.7867	0.8621	0.7526	0.9524	0.4000

Figure 11: Model Comparison for 4 Models

Confusion matrix of Boosted_DefaultRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DecisionTree_DefaultRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of ForrestModel_DefaultRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of StepwiseRegression_DefaultRisk

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

Figure 12: Confusion Matrix for 4 Models

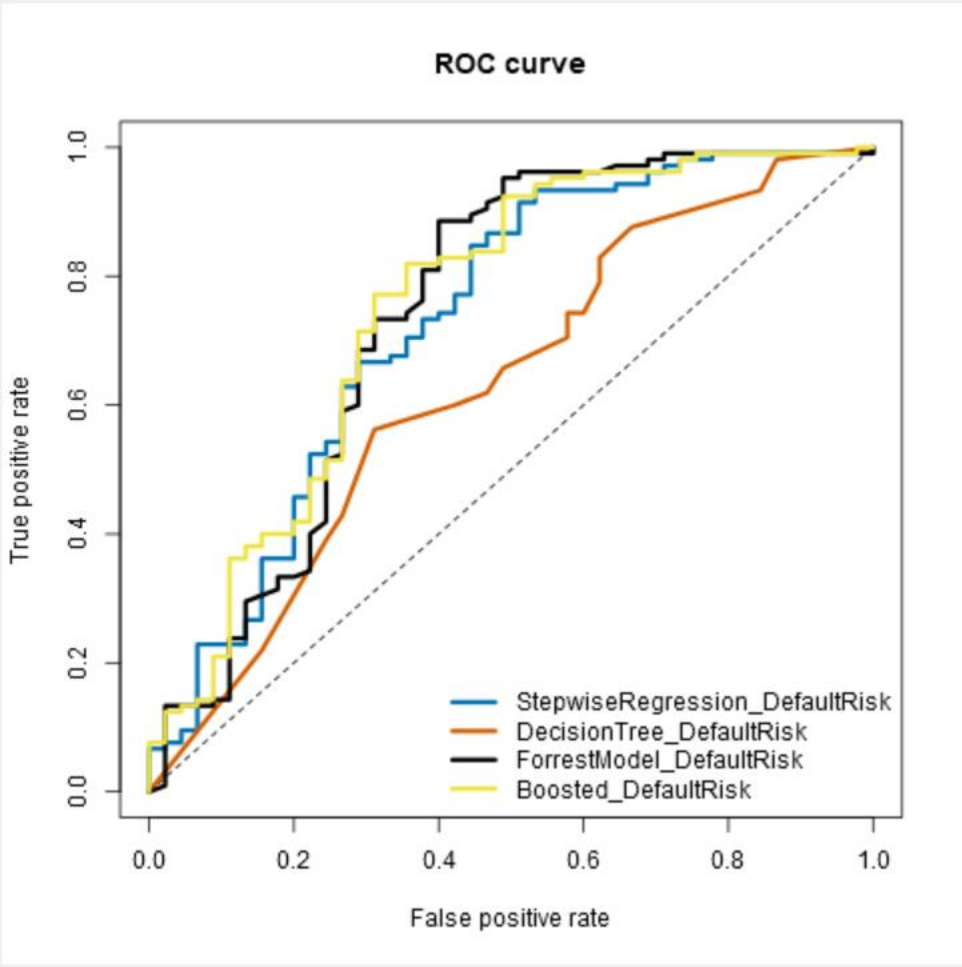


Figure 13: ROC Curve

Alteryx Workflows:

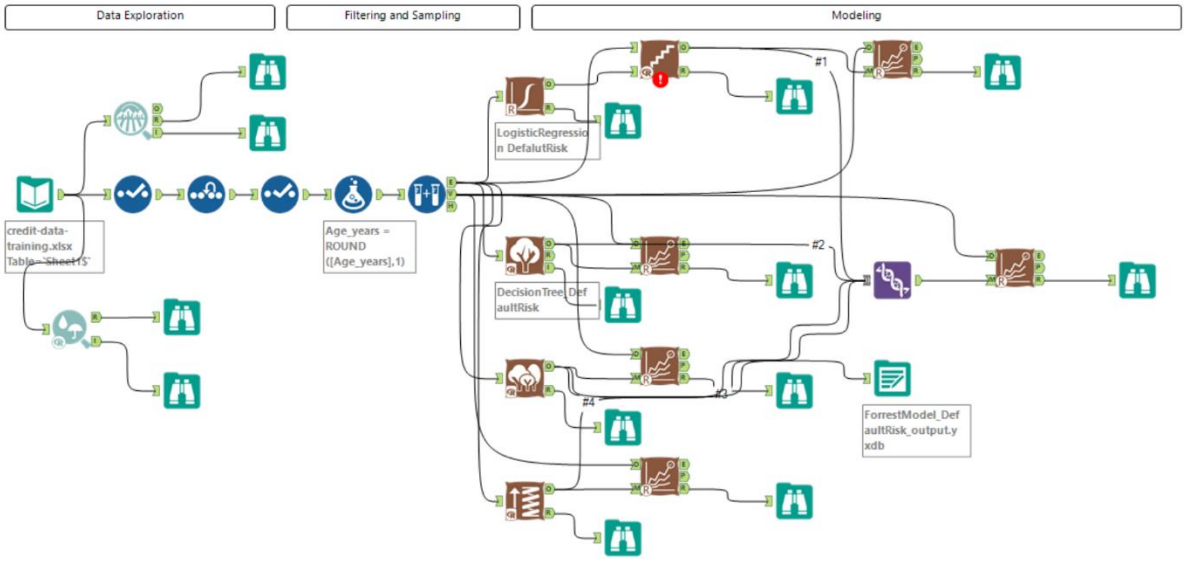


Figure 14: Predicting Default Risk Workflow

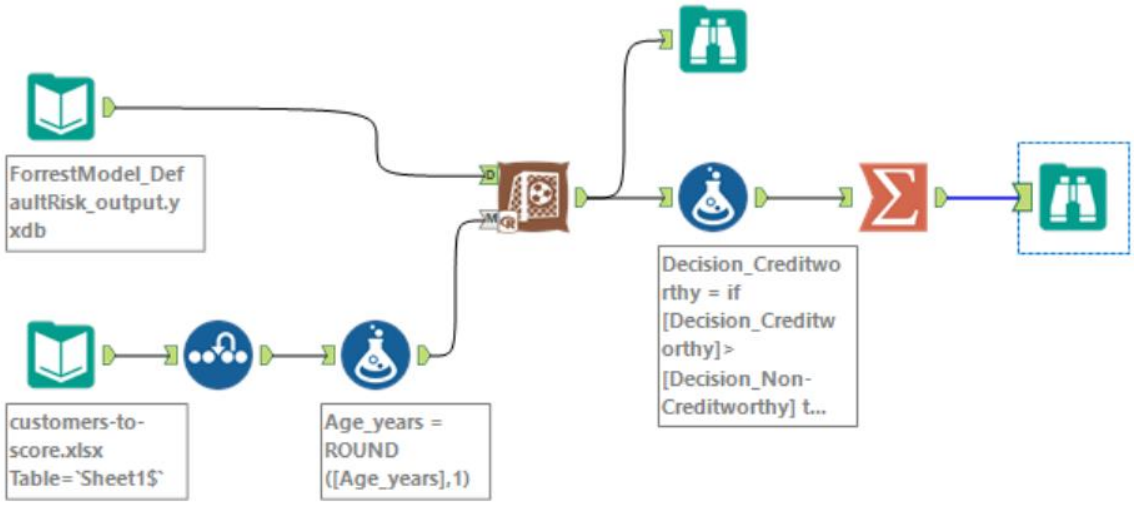


Figure 15: Scoring Customers Workflow