Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here:
https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

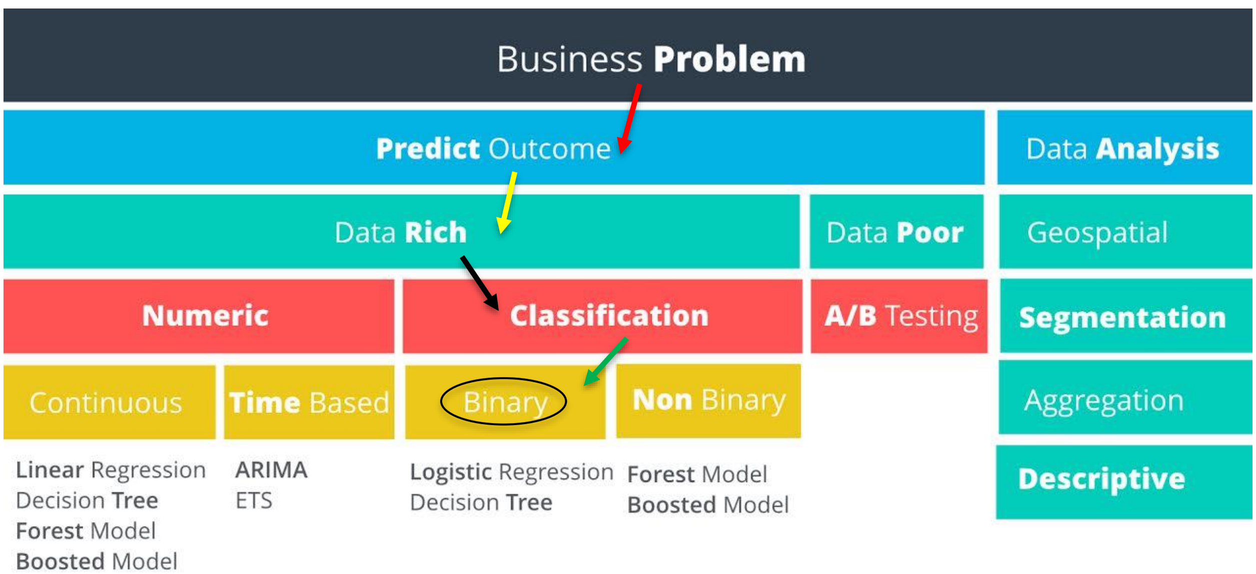
Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

The key business decision that needs to be made is if each new loan applicant is credit-worthy and should be granted a loan. The bank has been approving loan applications by hand but due to a huge influx of customers the loan application assessor should devise an automated way of completing the above task.

The bank has previously collected about twenty different kinds of data relevant to customers such as *Age*, *Account Balance*, *Payment Status of Previous Credit* etc., which will be used to construct the new customer evaluation model (some variables were discarded as non-suitable for the analysis).

Since we are dealing with new customers, we need to predict an outcome and this is a problem for which we do have data. The question is if a customer is creditworthy or not, so two possible outcomes, meaning that we need to construct a *Binary Classification Model*.



Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

The *Association Analysis* tool provides information on correlation between fields. As we can see below there are no highly correlated variables (correlation > 0.70) in our dataset.

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asset	Age.years	Type.of.apartment	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	1.000000	0.565054	0.145637	-0.032494	0.128814	-0.018171	0.126967	-0.185180	0.238437	-0.207298
Credit.Amount	0.565054	1.000000	-0.253286	-0.136621	0.457147	0.040486	0.100413	0.082721	0.192532	-0.045994
Instalment.per.cent	0.145637	-0.253286	1.000000	0.131231	0.115114	0.111456	0.178926	-0.293380	0.038515	-0.155458
Duration.in.Current.address	-0.032494	-0.136621	0.131231	1.000000	-0.047386	0.301966	-0.163386	-0.036814	0.055112	-0.015787
Most.valuable.available.asset	0.128814	0.457147	0.115114	-0.047386	1.000000	0.123579	0.182744	0.019435	0.083395	0.071932
Age.years	-0.018171	0.040486	0.111456	0.301966	0.123579	1.000000	0.208552	0.046996	0.141103	-0.020939
Type.of.apartment	0.126967	0.100413	0.178926	-0.163386	0.182744	0.208552	1.000000	-0.010189	0.179688	-0.026742
No.of.dependents	-0.185180	0.082721	-0.293380	-0.036814	0.019435	0.046996	-0.010189	1.000000	-0.097632	0.218454
Telephone	0.238437	0.192532	0.038515	0.055112	0.083395	0.141103	0.179688	-0.097632	1.000000	-0.168472
Foreign.Worker	-0.207298	-0.045994	-0.155458	-0.015787	0.071932	-0.020939	-0.026742	0.218454	-0.168472	1.000000

During the cleanup process, I visualized the available data using the *Field Summary* tool on Alteryx. Seven variables were removed from the dataset and I imputed values in one other variable.

Variables that were removed:

1) Occupation: only one distinct value (1) throughout, cannot add value to the model.
2) Concurrent credits: same as occupation, only one distinct value (Other Banks/Depts) throughout.

3) Duration in current address: 69% of the values are missing, cannot impute.
4) Guarantors: Low variability, 91.4% vs 8.6% values distribution.
5) Foreign Worker: Low variability, 96.2% vs 3.8% values distribution.

For the last two variables, I suspected at least *No. of Dependents* to be one of them due to low variability (85.4% vs 14.6% value distribution), however I examined the remaining variables' correlation to *Credit application result* and corresponding values to build a solid case for excluding them.

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value	
Most.valuable.available.asset	-0.232248	0.0050930	**
Duration.of.Credit.Month	-0.215149	0.0096065	**
Instalment.per.cent	-0.130496	0.1190020	
Age.years	0.123088	0.1416213	
Credit.Amount	-0.092205	0.2717004	
Foreign.Worker	0.072525	0.3876717	
Duration.in.Current.address	0.067284	0.4229716	
Type.of.apartment	-0.039360	0.6395134	
No.of.dependents	0.038037	0.6508161	
Telephone	0.030838	0.7136766	

The correlation analysis suggests little association between No. of dependents and creditworthiness, the same goes with the Telephone field, which does not imply any particular relationship with banking status. So:

6) No. of dependents: low variability, low correlation to target variable, high p-values.
7) Telephone: lowest correlation to target variable, highest p-values.

It is also noted that there is a 2% of the *Age-years* data missing. Since this is a numerical field we can impute missing values using the average or the median of the available data. The median is 33 and the average is ca. 36 (rounded up from 35.637 years). In this case, the median appears to be more suitable given the distribution of the field values as seen in the Field Summary visualization above. Most values are between 25 years and 30 years of age, hence the median is closer to that trend. Having imputed the new average age of the customers is ca. 36 again (rounded up from 35.574 years).

Step 3: Train your Classification Models

Having created estimation and validation samples according to the project requirements, four predictive models were trained as explained below.

Logistic Regression Model inc. Stepwise Tool: With the assistance of the *Stepwise* tool on Alteryx and after multiple iterations the predictor variables chosen for this model are: *Installment per cent*, *Credit amount*, *Purpose*, *Account Balance*, *Length of current employment*, *Payment status of previous*. All categorical values with at least one statistically significant value were included as it will most likely affect the outcome of the analysis. P-values of all statistically significant predictor variables are well within the acceptable range.

Although the R-squared value is low (0.1963), the model achieves a 0.7800 accuracy against the validation sample (0.8103 for Creditworthy and 0.6765 for Non=Creditworthy). There is a slight bias of the model toward predicting non-creditworthy statuses for clients that are actually creditworthy.

Visualizations of the findings are included below.

1 **Report for Logistic Regression Model Logistic_Step**

2 *Basic Summary*

3 Call:
glm(formula = Credit.Application.Result ~ Instalment.per.cent + Credit.Amount + Purpose + Account.Balance + Length.of.current.employment + Payment.Status.of.Previous.Credit, family = binomial(logit), data = the.data)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.352	-0.731	-0.456	0.769	2.458

6 Coefficients:

7

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.5783608	6.414e-01	-4.0202	6e-05 ***
Instalment.per.cent	0.3426933	1.325e-01	2.5873	0.00967 **
Credit.Amount	0.0002076	5.453e-05	3.8070	0.00014 ***
PurposeNew car	-1.6344313	6.137e-01	-2.6633	0.00774 **
PurposeOther	-0.4435055	8.242e-01	-0.5381	0.59049
PurposeUsed car	-0.7315961	3.976e-01	-1.8400	0.06577 .
Account.BalanceSome Balance	-1.5715598	3.037e-01	-5.1742	2.28e-07 ***
Length.of.current.employment4-7 yrs	0.3678284	4.537e-01	0.8107	0.41752
Length.of.current.employment< 1yr	0.7564408	3.833e-01	1.9733	0.04846 *
Payment.Status.of.Previous.CreditPaid Up	0.2117362	2.952e-01	0.7174	0.47316
Payment.Status.of.Previous.CreditSome Problems	1.3053044	5.089e-01	2.5648	0.01032 *

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: 332.06 on 339 degrees of freedom
McFadden R-Squared: 0.1963, AIC: 354.1

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Step	0.7800	0.8507	0.7352	0.8103	0.6765

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]

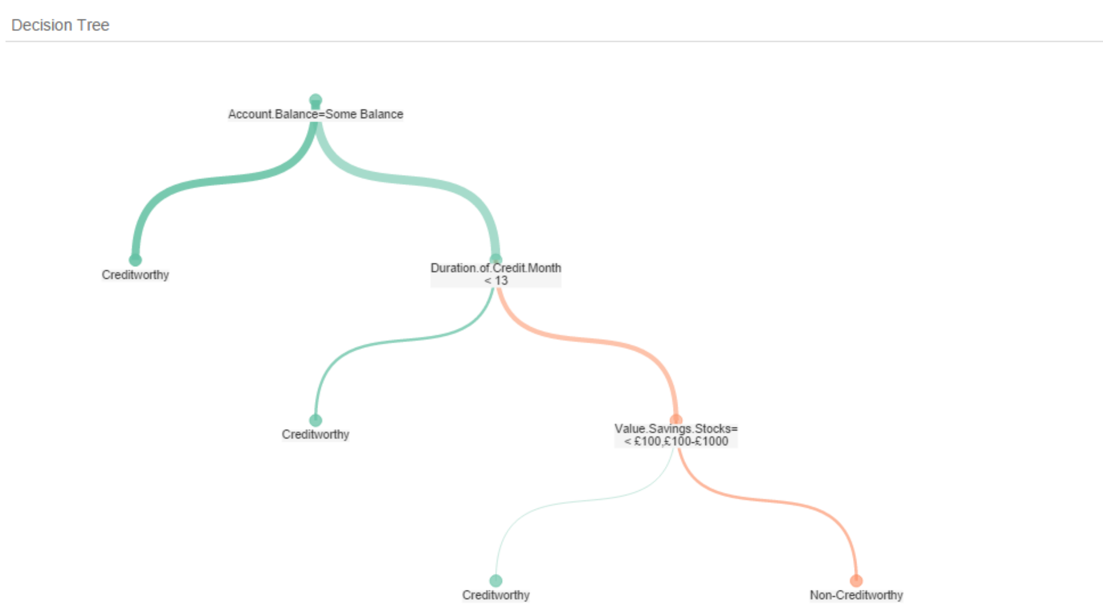
AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

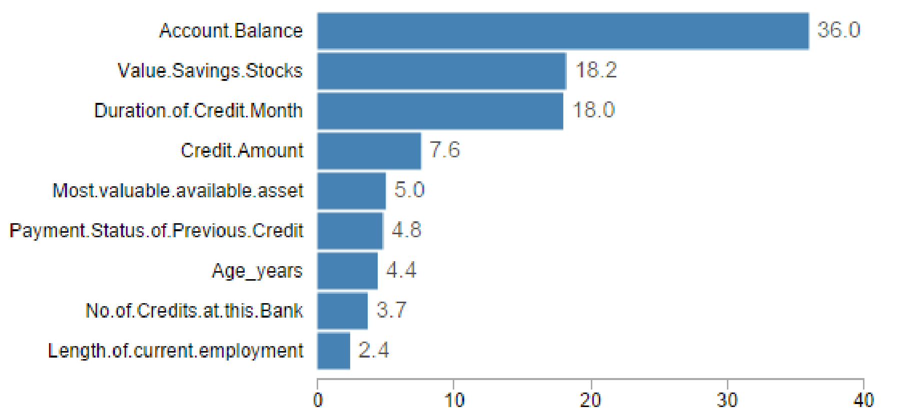
Confusion matrix of Logistic_Step

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

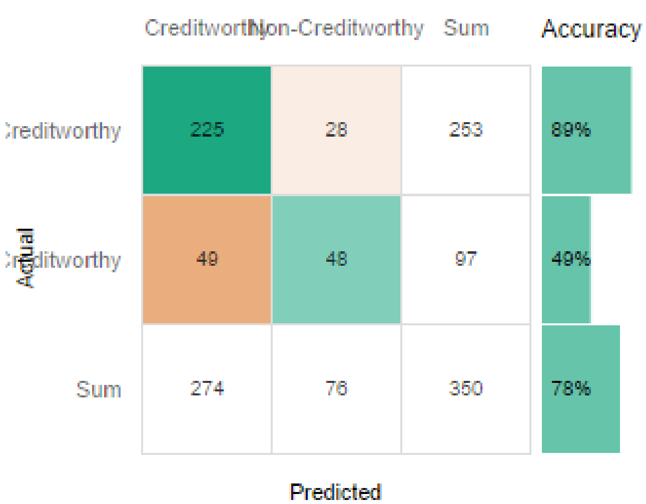
Decision Tree Model: The most important variables for the forest model analysis are *Account Balance (36.0)*, *Value saving stocks (18.2)* and *Duration of credit month (18.0)*, followed by another six variables that can have some effect on the model (see Variable importance graph). Accuracies achieved on estimation sample are shown within the confusion matrix.



Variable Importance



Confusion Matrix



The decision tree model achieves a 0.7467 accuracy against the validation sample (0.7913 for Creditworthy and 0.6000 for Non=Creditworthy). Again, there is a tendency for falsely predicting non-creditworthy statuses for creditworthy applicants, i.e. 49 applicants out of 97 were falsely predicted as non-creditworthy (orange box).

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Model	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)

Confusion matrix of Decision_Tree_Model

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

Forest Model: Forest models create multiple different estimation and validation samples and the follow a similar analysis path as the decision trees model, in our case 500 different combinations were examined. Running the model:

1 *Basic Summary*

2 Call:
randomForest(formula = Credit.Application.Result ~ Type.of.apartment + Most.valuable.available.asset + Instalment.per.cent + Credit.Amount + Duration.of.Credit.Month + Age_years + Purpose + Account.Balance + Value.Savings.Stocks + Length.of.current.employment + Payment.Status.of.Previous.Credit + No.of.Credits.at.this.Bank, data = the.data, ntree = 500)

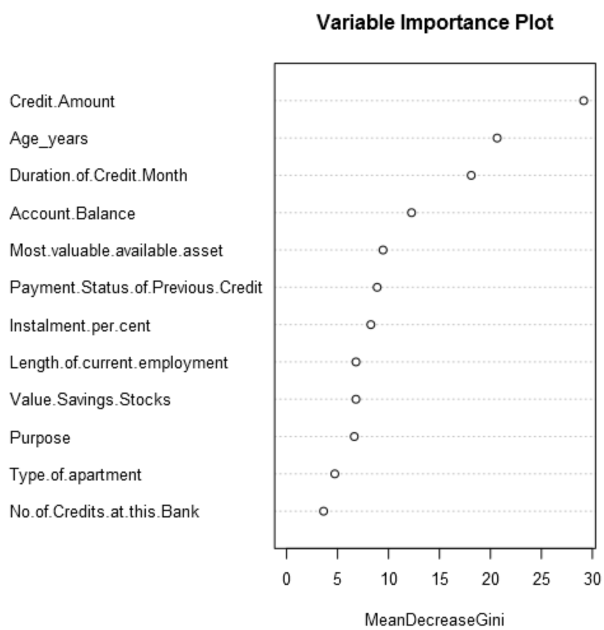
3 Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

4 OOB estimate of the error rate: 36.6%

5 Confusion Matrix:

6

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.083	232	21
Non-Creditworthy	0.649	63	34



The most important variables appear to be *Credit amount*, *Age_years*, *Duration of credit month* and *Account balance*, followed by another eight variables of reducing significance as shown above.

The forest model's performance against the validation sample is the highest so far, i.e. 0.8067 (0.7969 for Creditworthy and 0.8636 for Non-Creditworthy). The model is predicting slightly better actual Non-Creditworthy applicants, as shown in the confusion matrix below. However, it is not a bias per se as it performs very well for actual Creditworthy cases.

Model Comparison Report

Fit and error measures

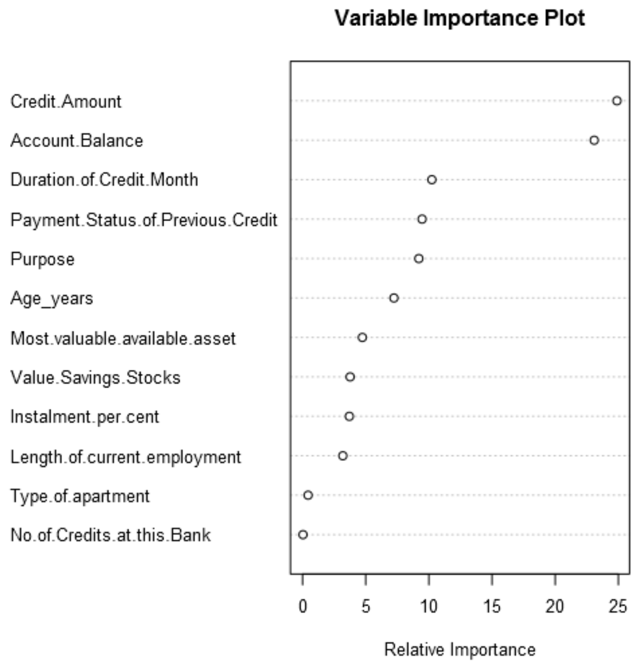
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8067	0.8755	0.7456	0.7969	0.8636

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 Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Boosted model: a boosted model performs multiple iterations of error reducing analysis of decision trees. Our model indicates that the most important variables are *Credit amount* and *Account balance* followed by another ten variables of reducing impact.



The boosted model achieves a 0.7933 overall accuracy (0.7891 for Creditworthy and 0.8182 for Non-Creditworthy) with no bias at all as it can predict equally well both outcomes. The confusion matrix to support this view is shown below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7933	0.8670	0.7509	0.7891	0.8182

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	27
Predicted_Non-Creditworthy	4	18

Step 4: Writeup

To decide on which model most successfully predicts the creditworthiness of a loan applicant a model comparison is essential. Using the *Union* tool on Alteryx and running all models against the estimation and validation samples simultaneously, we can visualize what was explained in the previous section in one model comparison table. Please see below:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Step	0.7800	0.8507	0.7352	0.8103	0.6765
Decision_Tree_Model	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8067	0.8755	0.7456	0.7969	0.8636
Boosted_Model	0.7933	0.8670	0.7509	0.7891	0.8182

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	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of Decision_Tree_Model

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

Confusion matrix of Forest_Model

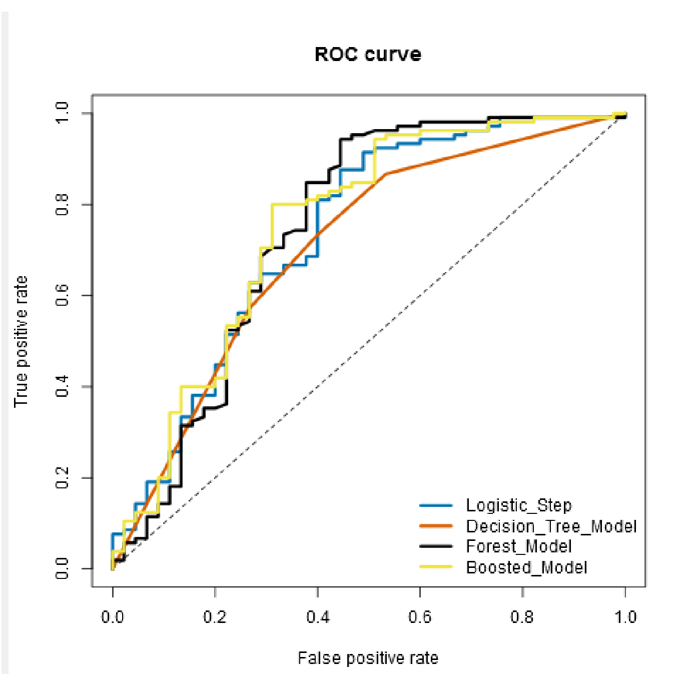
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of Logistic_Step

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

It appears that the **Forest Model** has the highest accuracy (0.8067) followed by the *Boosted Model (0.7933)*, the *Logistic_Step Model (0.7800)* and the *Decision Tree Model (0.7467)*. When considering the results in detail, we can see that all models perform almost the same (+-0.03) at the Actual_Creditworthy category, the top two performers being the *Logistic_Step* and *Forest models*. This makes our judgement more difficult and thus we also need to look at actual Non-Creditworthy occurrences. Immediately it is apparent that the *Logistic_Step* and *Decision Models* are not trustworthy enough as they perform poorly in this category (0.6000 – 0.6765). The highest accuracy is achieved by the *Forest Model* (0.8636) and then by the Boosted Model (0.8182) – there also is no obvious bias within those two as shown by the confusion matrices.

ROC curves are a common way to visualize the performance of a binary classifier. Curves enclosing more area under them are a sign of a better performing classifier.



The ROC curve graph illustrates a close race between the Boosted (yellow curve) and Forest (black curve) models which can be quantified by looking at the area under the curve (AUC) metrics (please see model comparison report above). The Boosted model has enclosed 0.7509 of the chart area under its curve, whereas the Forest model's AUC is 0.7456.

Taking into consideration all aspects of the analysis, it is considered that the ***Forest Model*** performs best in overall (highest overall accuracy which is what the bank manager is interested in).

The *Forest* model object was extracted and then input to a new Alteryx workflow to reduce analysis load. Using the *Score* tool, it is predicted that 406 of the applicants are Creditworthy and 94 applicants are Non-Creditworthy.

Record #	Credit Application Result	Count
1	Creditworthy	406
2	Non-Creditworthy	94

Appendix – Alteryx Workflows