Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Project: Creditworthiness

The Business Problem

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week.

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days.

You have the following information to work with:

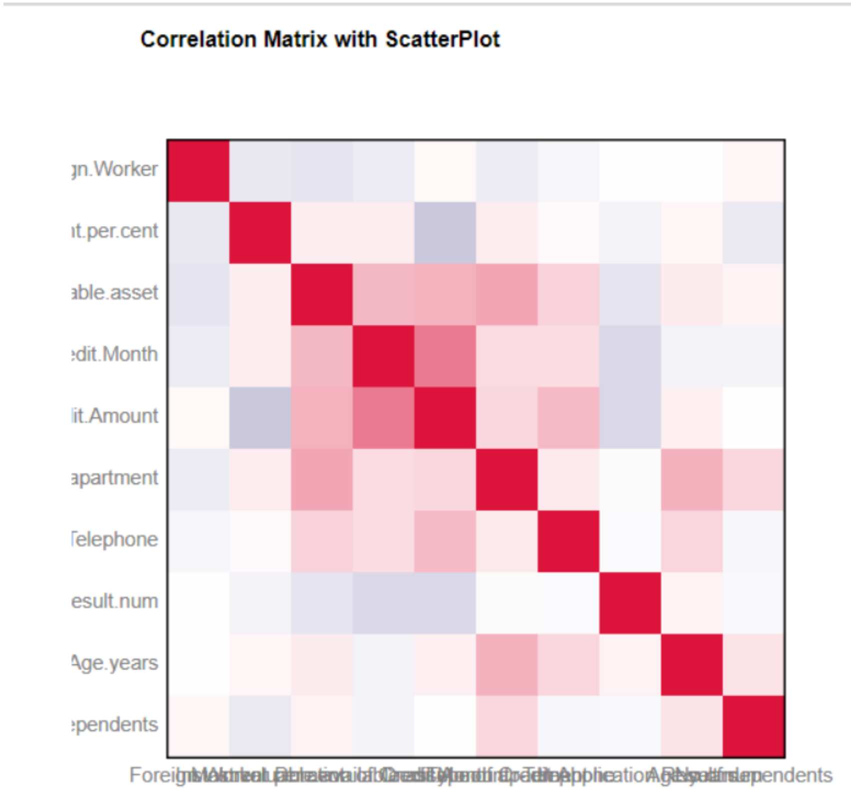
1. Data on all past applications
2. The list of customers that need to be processed in the next few days

Step 1: Business and Data Understanding

The key decision that needs to be made throughout this analysis is to develop an effective model and evaluate the creditworthiness of the new loan applicants. In order to develop a model to estimate whether the new loan applicants are creditworthy or not, we need a historical data of past applications along with basic credit-related information of the past applicants.
For the modeling, we will be conducting 4 kinds of binary models (Logistics, Decision Tree, Forest, and Boosted) and compare all 4 models against each other to determine which model explains the creditworthiness of applicants the best.

Step 2: Building the Training Set

The data has already been cleaned up with proper format. So, we move on to run an association analysis with all numerical variables in Alteryx to find any correlations within the variables. Below is the correlation matrix with scatterplot.



As it can be seen from the correlation matrix plot, none of the variables are highly correlated to each other (correlation all less than 0.7) Therefore, we should not remove any variables yet.
Then, a field summary tool was used to find general distributions of the data for all given variables. Below is the result of the field summary tool in Alteryx.

First thing we notice in the above distribution plots is that variable 'Duration-in-Current-address' have a lot of missing values (69% missing). Because it has too many missing data, we can remove this variable from the dataset for better analysis. However, the variable 'Age-years' has a few missing data, so we keep this variable and substitute median value for all missing values.

Secondly, variables 'Concurrent-Credits', 'Occupations' have a uniform distribution that there are no other variations of the data. Similarly, variables 'Guarantors', 'Foreign-Worker', ''No-of-dependents' have very low variability that the data field is heavily skewed towards one type of data. Therefore, we proceed with removing these variables from the dataset.

Lastly, variable 'Telephone' does not seem be related to evaluate the creditworthiness of applicants, so we decided to remove these variables as well. Therefore, total of 7 columns are removed from the dataset.

Minseok (Richard) Park
Predictive Analytics for Business Nanodegree

Step 3: Train your Classification Models

First, we created a sample with estimation (70% of the dataset) and validation (30% of the dataset) Then all 4 kinds of binary models were created and compared to see which binary model explains the creditworthiness of applicants. For all 4 models, the target variable is set as 'Credit-Application-Result'

1. Logistic Regression (stepwise)

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

As a result of running logistic regression with stepwise selection, the top 3 most significant variables that have lowest p-values are 'Account-Balance', 'Credit-Amount', and 'Purpose' Then we applied this model in validation samples to see the accuracy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_stepwise	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.

Confusion matrix of logistic_stepwise

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

The overall accuracy of this model is 76%. The model did a fair job in classifying Creditworthy applicant correctly (87.62%), but it did not do a good job in classifying non-

creditworthy applicants with accuracy of 48.89%. Therefore, the model may be biased towards predicting applicants as non-creditworthy.

2. Decision Tree

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.6733	0.7721	0.6296	0.7905	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 situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of DT_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Above are the outputs after conducting a decision tree model, and a model comparison report after the validation. The top 3 most significant variables for the decision tree model are ‘Account-Balance’, ‘Duration-of-Credit-Month’, and ‘Credit-Amount’. As a result of the confusion matrix, overall accuracy of the model is 84%, and the accuracy for classifying the creditworthy applicants is 91%, and non-creditworthy applicants 66%. However, applying the model with the validation samples, the overall accuracy is at 67.33%, accuracy of creditworthy 79.05% and accuracy of non-creditworthy 40%. According to the result, the decision tree model may also be biased towards predicting applicants as non-creditworthy.

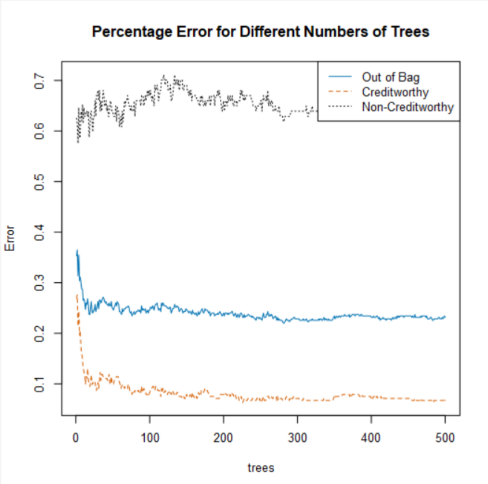
3. Forest Model

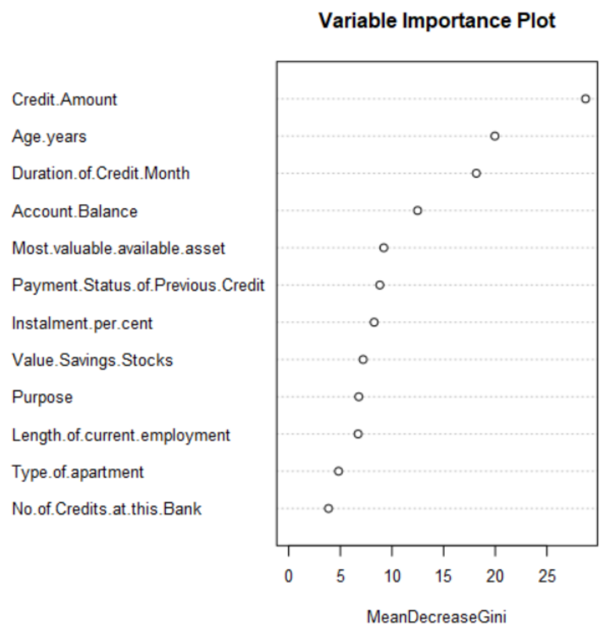
OOB estimate of the error rate: 23.1%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.067	236	17
Non-Creditworthy	0.66	64	33

Plots





Model Comparison Report

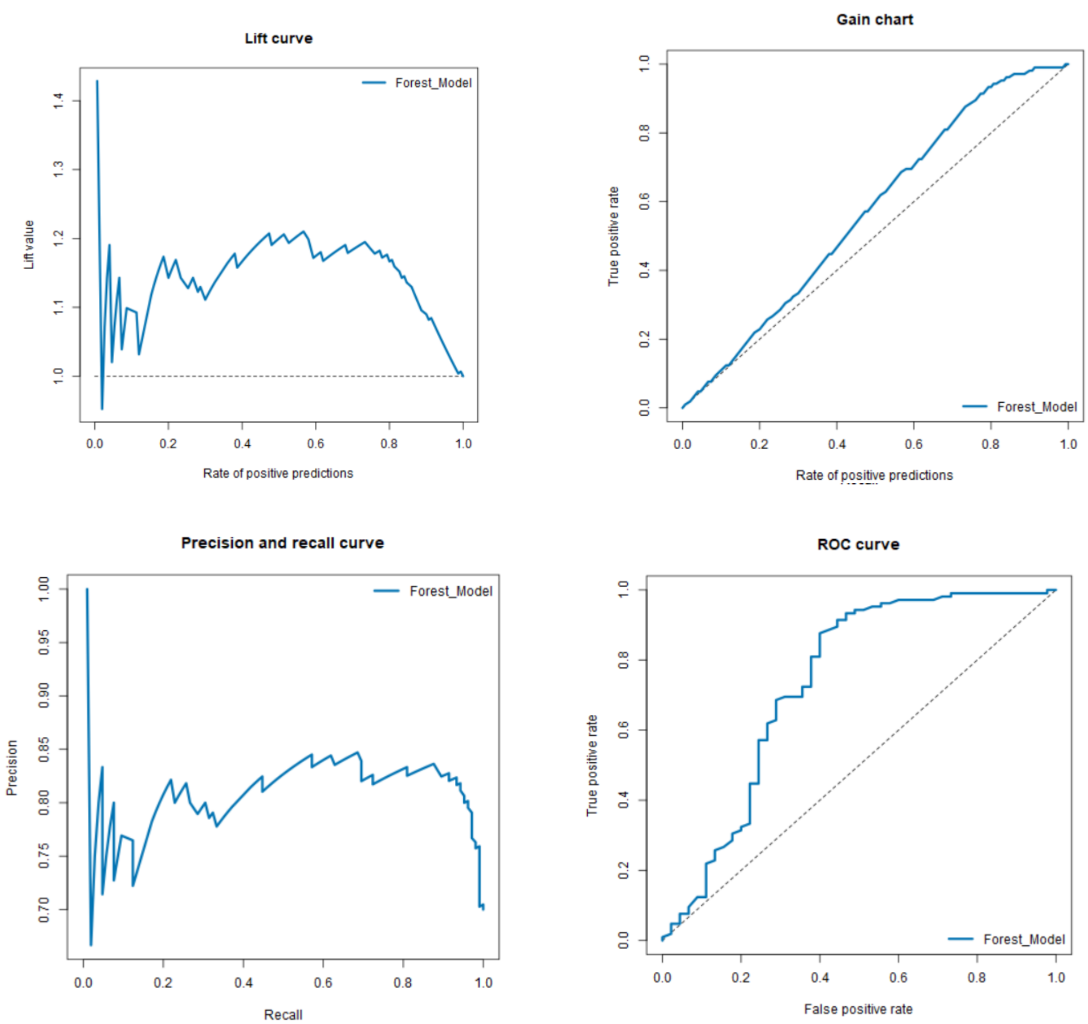
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.7933	0.8681	0.7368	0.9714	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.

Confusion matrix of Forest_Model

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



The Forest Model have OOB estimate of the error rate with 23.1%, and classification error rate for creditworthy is 6.7%, and for Non-creditworthy 66%. Top 3 most significant variables for the forest model are 'Credit-Amount', 'Age-years', and 'Duration-of-Credit-Month'

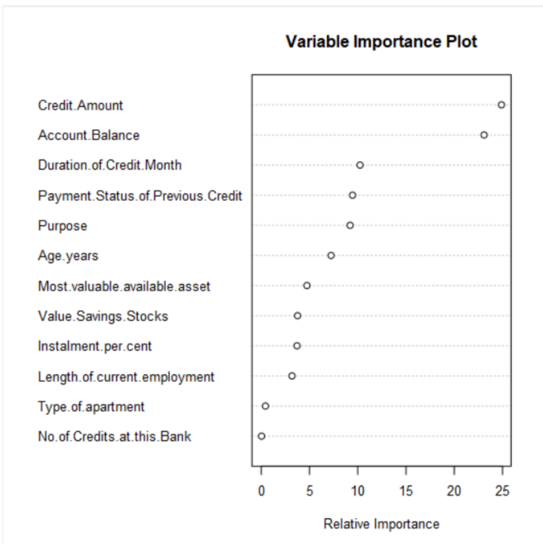
Applying this forest model with our validation samples, the overall accuracy is 79.33%: 97.14% for classifying applicants as creditworthy, and 37.78% accuracy for classifying applicants as non-creditworthy.

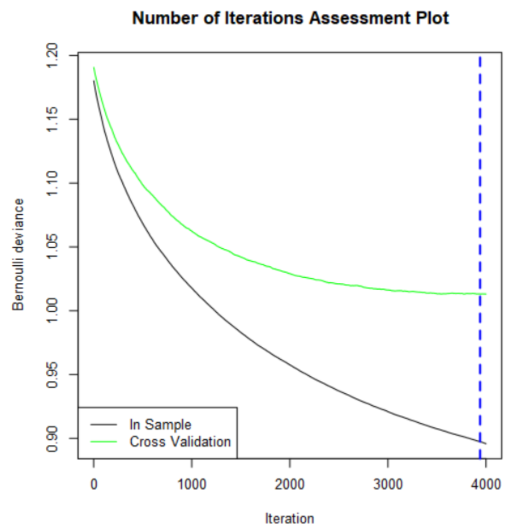
4. Booted Model

Report for Boosted Model Boosted_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3940





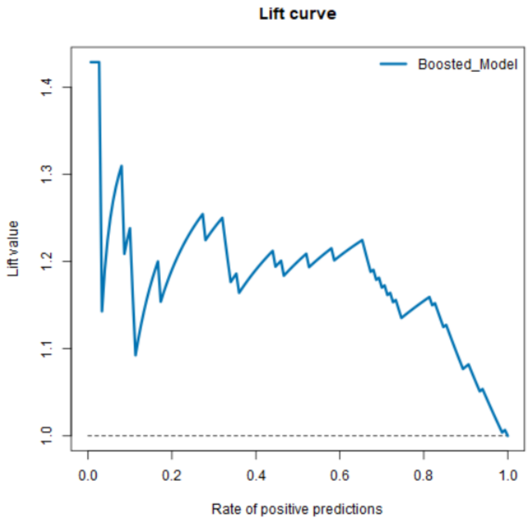
Model Comparison Report

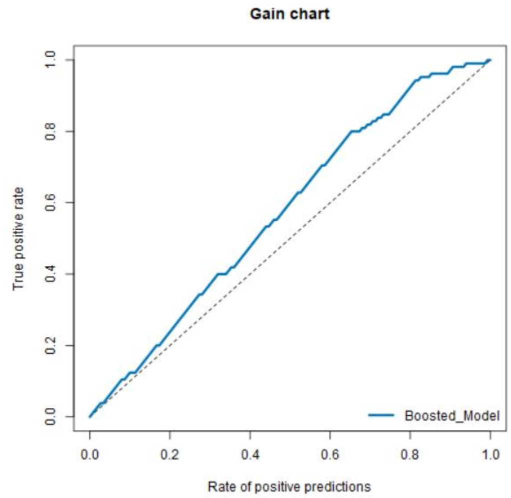
Fit and error measures

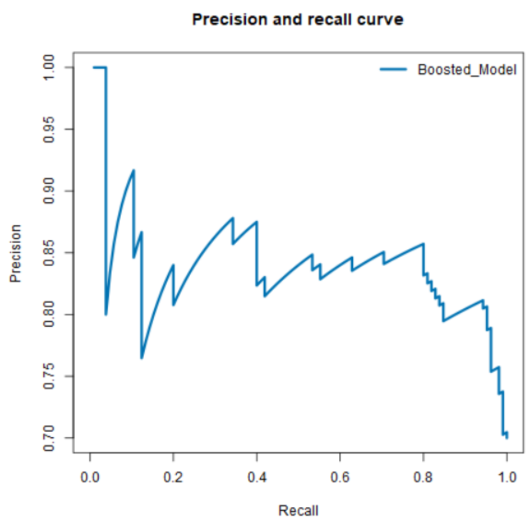
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7933	0.8670	0.7509	0.9619	0.4000

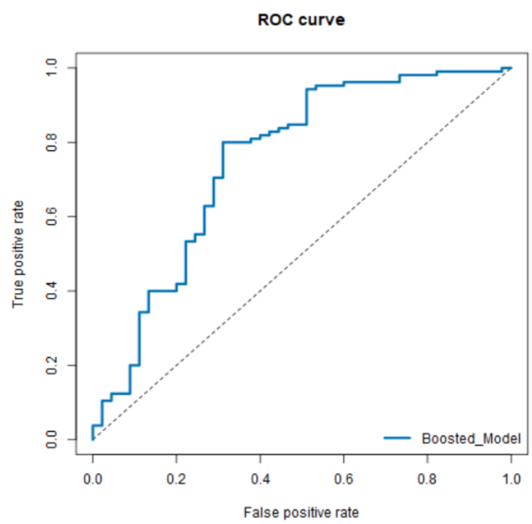
Confusion matrix of Boosted_Model

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









For the boosted model, variables that seem to be the most significant are 'Credit-Amount' and 'Credit-Balance'. Applying the boosted model with validation samples, the overall accuracy is 79.33%. The accuracy of classifying applicants with creditworthy is 96.19%, and for non-creditworthy 40%.

Step 4: Writeup

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7933	0.8670	0.7509	0.9619	0.4000
logistic_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted_Model

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

Confusion matrix of DT_Model

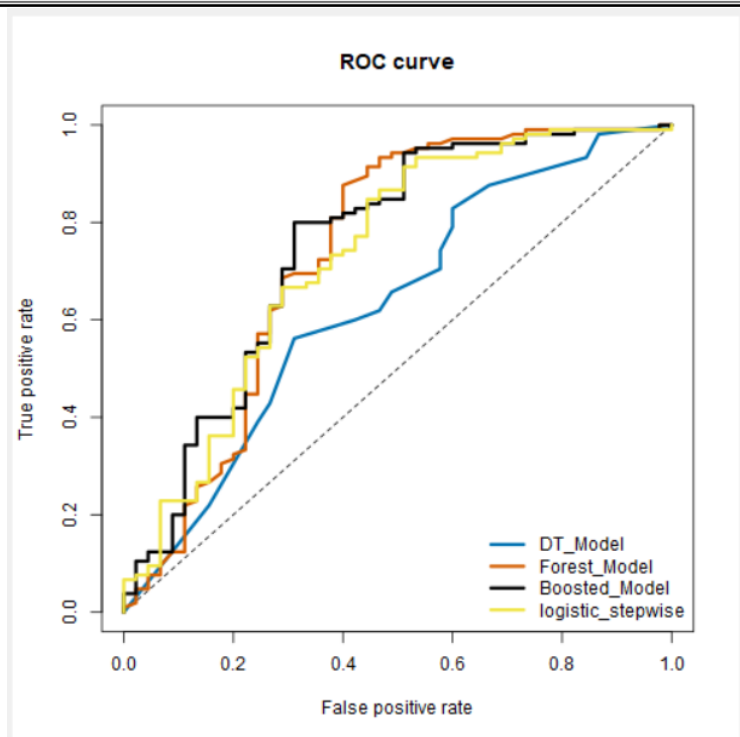
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

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



Lastly, all four kinds of binary models are compared to determine the best model for predicting creditworthiness of the new applicants. The overall accuracies for all 4 models are fairly low, ranging from 67% ~ 79%. In general, the accuracies of classifying applicants as creditworthy for the 4 models were moderately high, but the accuracies of classifying applicants as non-creditworthy were not (all below 50%).

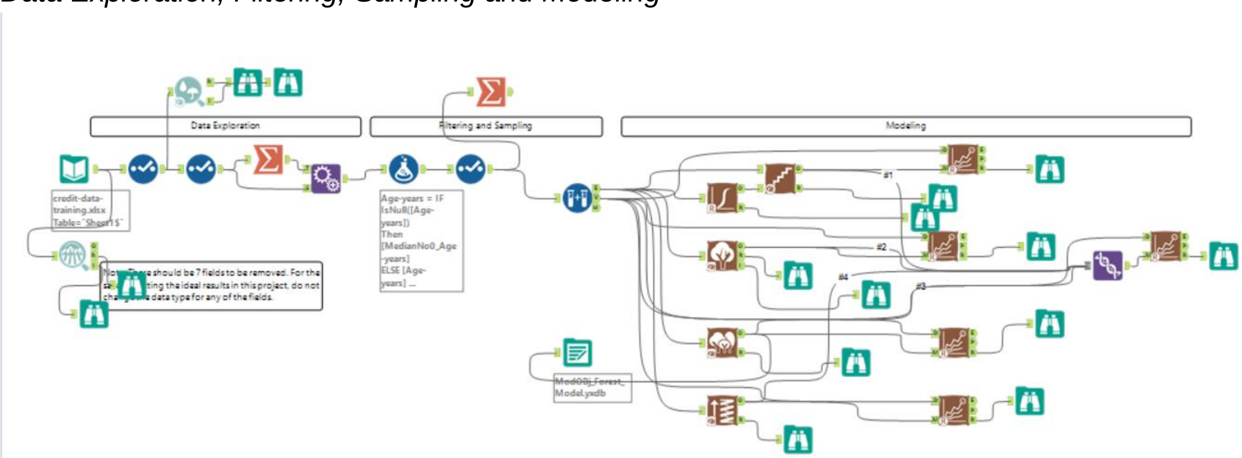
Considering the overall accuracy and the accuracy of predicting applicants as creditworthy, it would be legitimate to say that the forest model seems to be the best model out of the four binary models we tested. boosted model also has same overall accuracy with forest model (79.33%), but the accuracy of the creditworthy for the forest model is slightly higher than that of boosted model.

Then we exported the forest model and applied on the data of the new applicants to predict the probability of an applicant being classified into either creditworthy or non-creditworthy, using a Score tool in Alteryx. Next, we counted all the lists where the probability of being classified as creditworthy is greater than that of non-creditworthy. As a result, 408 applicants from the 500 new loan applicants would be considered creditworthy.

As a final comment, we must keep in mind that the forest model that we used had low accuracy (37.78%) of classifying applicants as non-creditworthy. Therefore, the 98 applicants who were classified as non-creditworthy may be biased that within these people some people could be creditworthy. With this being the issue, I would recommend to carefully review again the applicants who were classified as non-creditworthy and had probability of non-creditworthy close to 50%.

Alteryx Workflow

Data Exploration, Filtering, Sampling and Modeling



Scoring new applicants data with chosen model