

Project: Creditworthiness

By Bipin Kumar Sultania

Business and Data Understanding

Ques 1: What decisions needs to be made?

Answer: A small bank, which usually receives 200 loan applications each week and approves by hand, unexpectedly received nearly 500 loan applications due to a financial crisis that struck a large bank last week. As a loan officer, I need to learn how to handle all of these loan applications within a week. On the basis of the classification models that I have recently studied, I need to systematically assess the creditworthiness of these new loan applicants and to provide my manager with a list of creditworthy clients in the next two days.

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. Based on it, we can make the 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: We need to use the Binary model to help make these choices, as we want to classify people who are creditworthy and not creditworthy for loans.

Building the Training Set

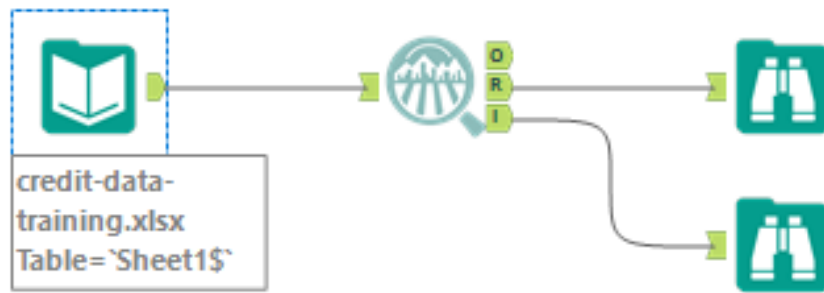
Ques: In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

Answer: In summary

impute Age-years and

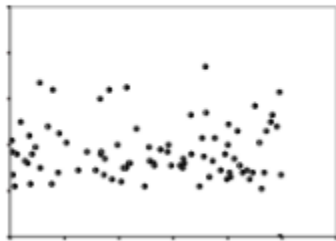
remove Duration-in-Current-address, Occupation, Concurrent-Credits, Guarantors, Foreign-Worker, No-of-dependents, and Telephone.

Alteryx Workflow Used



Imputing:

Age-years space has 2.4 percent missing data. I will measure the data using the median of the entire data field as indicated in the project description.

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502

Removing

Duration-in-Current-address: 68.8% data missing.

Occupation: Only value "1"

Concurrent-Credits: Has only "Other Banks/Depts" (500 instances)

Guarantors: Data is skewed toward "None".

Foreign-Worker: Data is skewed toward "1".

No-of-dependents: Data is skewed toward "1".

Telephone: Cannot tell us anything about the creditworthiness of the applicant.



The 'Age' category is imputed because it has 2 percent of the missing values only. If these values are withdrawn, it could have an detrimental impact on the other attributes, so this value is imputed by replacing the 'null' values with the 'median' values. We took median as all age data is moved to the left.

Train your Classification Models

Logistic Stepwise

Ques 1: 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 predictive variables are Account Balance, Payment Status of Previous Credit, Purpose, Credit Amount, Length of Current Employment, and Instalment per Cent.

Report for Logistic Regression Model Loan_Step

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

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 and 62% for non-credibility.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Loan_Step	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.

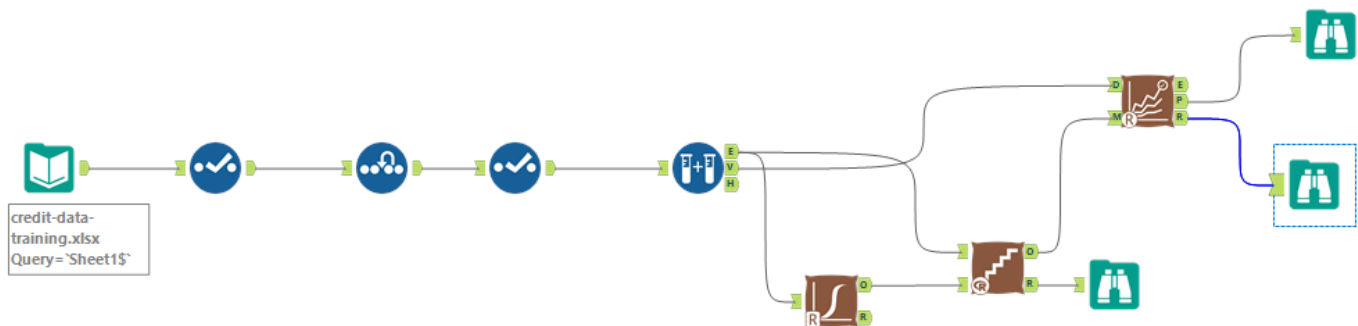
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

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

F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. 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 Loan_Step

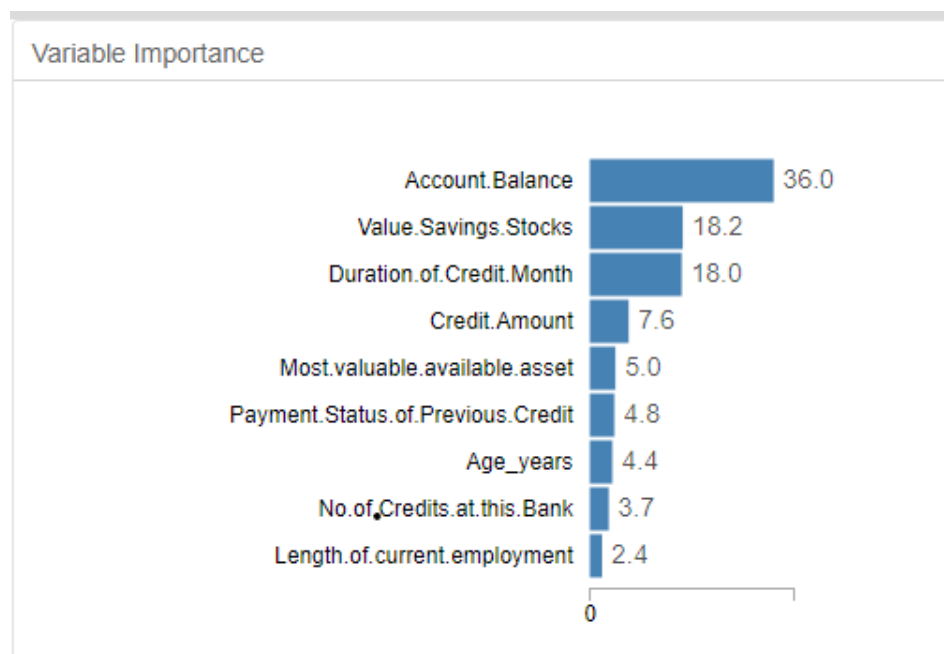
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22



Decision Tree:

Ques 1: 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: Based on the variable importance report, the top 5 predictive variables are Account Balance, Value Savings Stocks, Duration of Credit Month, Credit amount and Most valuable available asset



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 75% with the accuracy of predicting creditworthy customers of approximately 79% and non-creditworthy customers as 60%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Loan_DT	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*.

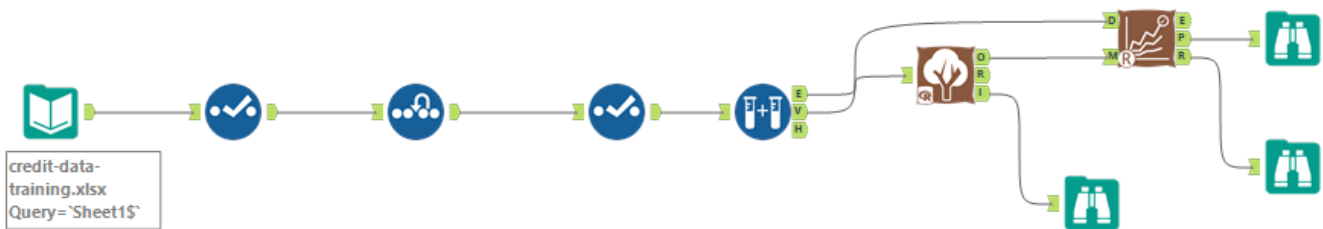
AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. 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 Loan_DT

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

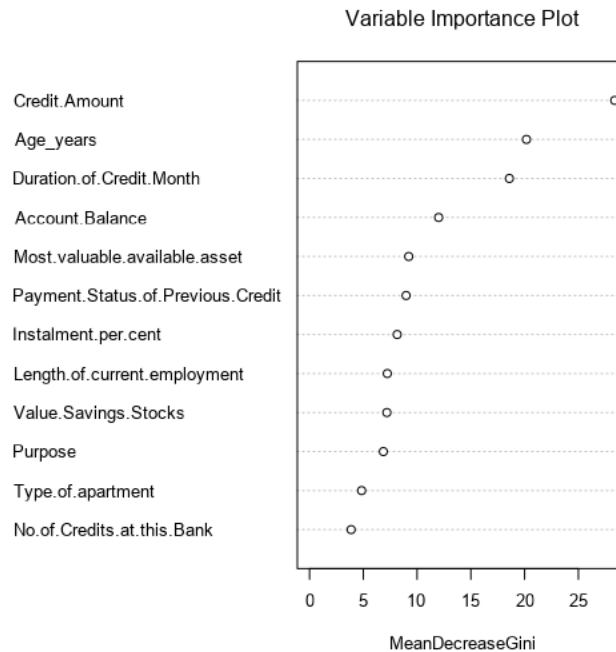
Alteryx Workflow



Forest Model:

Ques 1: 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: Based on the variable importance plot, the top 3 predictive variables are Credit Amount, Age Years, Duration of Credit Month, Account Balance and Most valuable available asset



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?

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Loan_FM	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, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. 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 Loan_FM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

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.

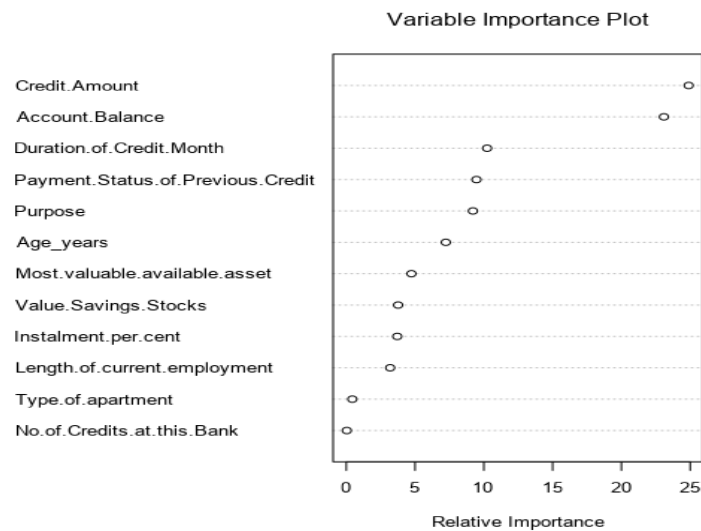
F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Loan_FM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Alteryx Workflow

Boosted Model:

Answer: Based on the variable importance plot, the top 5 important predictive variables are Credit Amount, Amount Balance, Duration of Credit Month, Payment Status of Previous Credit and Purpose



Ques 2: 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?

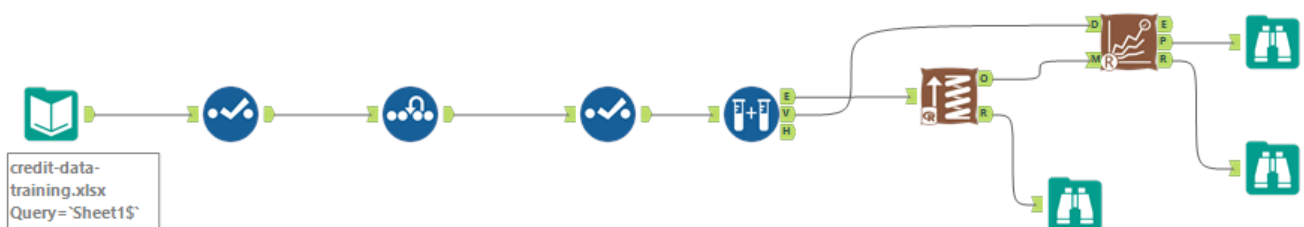
Answer: The model comparison report shows that the accuracy of this model approx. 80% with accuracy of predicting creditworthy customers to be 79% and non-creditworthy customers to be 82%

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Loan_BM	0.7933	0.8670	0.7509	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, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. 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 Loan_BM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Alteryx Workflow



After all the predictive analysis I used the model comparison tool to validate and compare the accuracy of these four models:

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Loan_DT	0.7467	0.8273	0.7054	0.8667	0.4667
Loan_FM	0.8000	0.8707	0.7361	0.9619	0.4222
Loan_BM	0.7933	0.8670	0.7509	0.9619	0.4000
Loan_Step	0.7600	0.8364	0.7306	0.8762	0.4889

Based on the model comparison study above, the overall accuracy of the Logistic Stepwise model is 0.7600; the overall accuracy of the Decision Tree is 0.7467; the overall accuracy of the Forest Model is 0.8000; and the overall accuracy of the Boosted Model is 0.7867.

Below please find the confusion matrix for all these models:

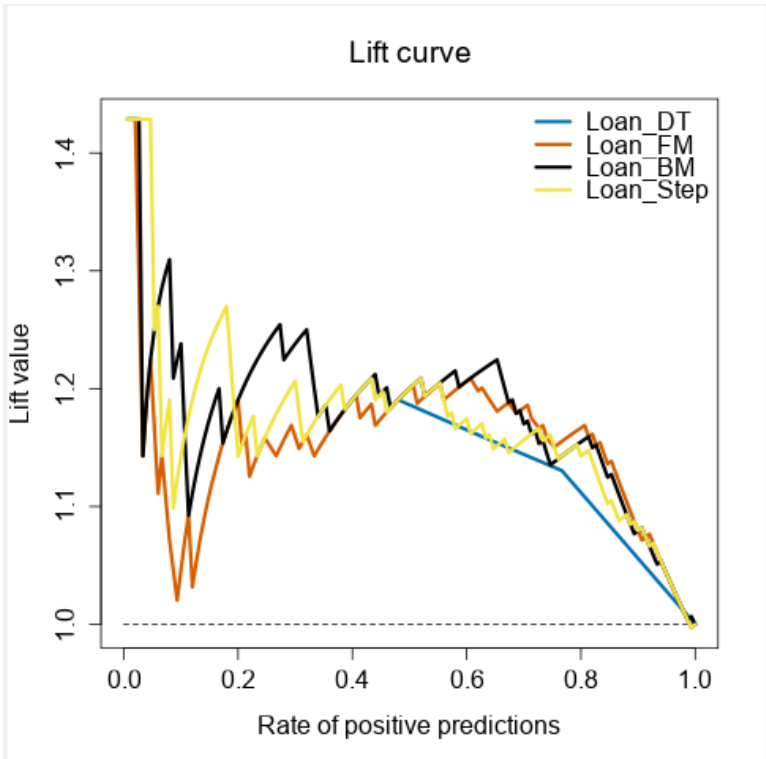
Confusion matrix of Loan_BM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

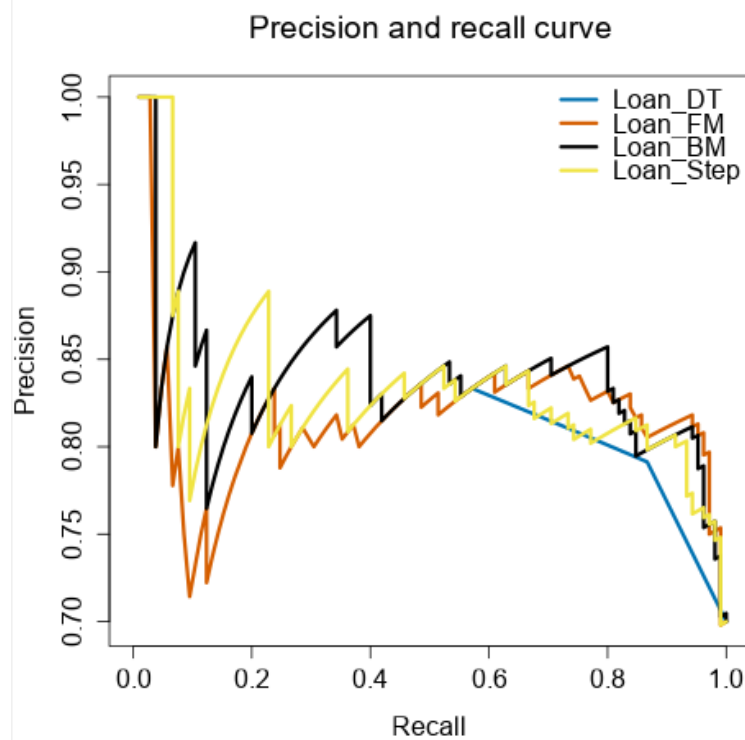
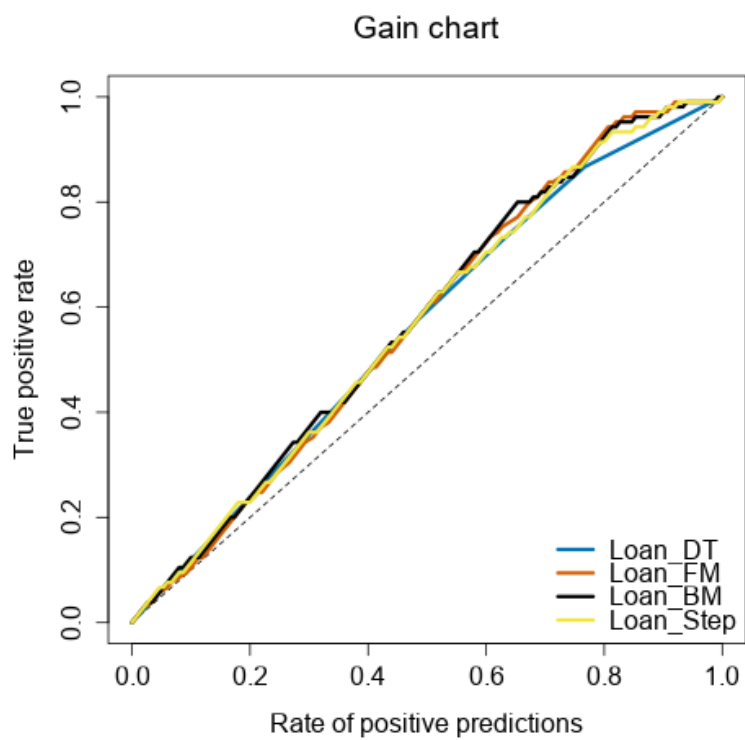
Confusion matrix of Loan_DT		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Loan_FM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Loan_Step		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

The no of creditworthy is more than the no of non-creditworthy.





We may assume that the models appear to predict individuals who are creditworthy, as they do not predict individuals who are not creditworthy at exactly the same level as those who are.

Writeup

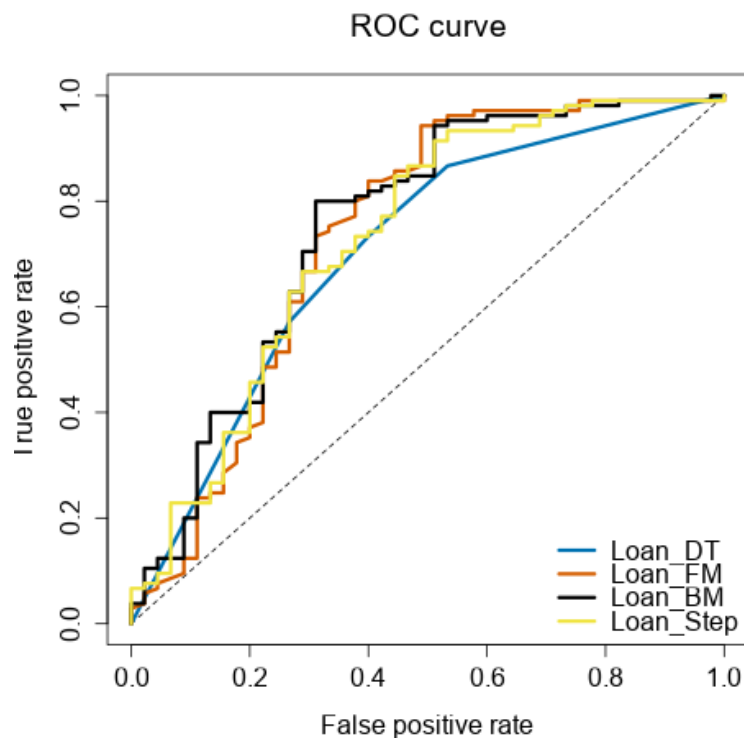
Ques1: Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:

- Overall Accuracy against your Validation set
- Accuracies within “Creditworthy” and “Non-Creditworthy” segments
- ROC graph
- Bias in the Confusion Matrices

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

Answer: From the model comparison report below, we can see the Loan_FM i.e. Forest Model is the most accurate model with 80%, so we'll use the same for prediction analysis. The difference between credit-worthy and non-credit-worthy predictions is also almost the same, so this model is not biased.

Also, we can deduce from the ROC curve that the Forest model achieved the fastest flat curve.



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 406 customers which has more than score of 0.5 in creditworthy from the given dataset. Also, there are 2 customers which has exactly a score of 0.5 which mean they are may or maybe worthy.

Alteryx Workflow for Modelling

