Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions :

● What decisions needs to be made?

- The objective is to build a predictive classification model to classify whether customers who applied for loan are creditworthy.

● What data is needed to inform those decisions? The data needed can be summarized in three categories and I listed few for each category:

- Personal information
  - length of employeement
  - duration of current address
  - number of dependent
- Basic account information
  - have account in bank or not and the balance
  - duration of credit month
- Loan information
  - loan purpose
  - credit aount

● What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

- Since the desired outcome is creditworthy or not, it would be a binary classification model. We'll go run logistic regression model, decision tree model etc respectively to compare the model accuracy.

Step 2: Building the Training Set

EDA

Fields Summary

	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean	Layout
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295	Layout - View Browse Tool Report Tab
2	Credit-Amount	Numeric	276	18424	2236.5	2831.386861	0	464	3199.98	Layout - View Browse Tool Report Tab
3	Duration-in-Current-address	Numeric	1	4	2	1.150017	68.8	5	2.660256	Layout - View Browse Tool Report Tab
4	Duration-of-Credit-Month	Numeric	4	60	18	12.30742	0	30	21.434	Layout - View Browse Tool Report Tab
5	Foreign-Worker	Numeric	1	2	1	0.191388	0	2	1.038	Layout - View Browse Tool Report Tab
6	Instalment-per-cent	Numeric	1	4	3	1.113724	0	4	3.01	Layout - View Browse Tool Report Tab
7	Most-valuable-available-asset	Numeric	1	4	3	1.064268	0	4	2.36	Layout - View Browse Tool Report Tab
8	No-of-dependents	Numeric	1	2	1	0.35346	0	2	1.146	Layout - View Browse Tool Report Tab
9	Occupation	Numeric	1	1	1	0	0	1	1	Layout - View Browse Tool Report Tab
10	Telephone	Numeric	1	2	1	0.490389	0	2	1.4	Layout - View Browse Tool Report Tab
11	Type-of-apartment	Numeric	1	3	2	0.539814	0	3	1.928	Layout - View Browse Tool Report Tab
12	Account-Balance	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]	[Null]
13	Concurrent-Credits	String	[Null]	[Null]	[Null]	[Null]	0	1	[Null]	[Null]
14	Credit-Application-Result	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]	[Null]
15	Guarantors	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]	[Null]
16	Length-of-current-employment	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]	[Null]
17	No-of-Credits-at-this-Bank	String	[Null]	[Null]	[Null]	[Null]	0	2	[Null]	[Null]
18	Payment-Status-of-Previous-Credit	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]	[Null]
19	Purpose	String	[Null]	[Null]	[Null]	[Null]	0	4	[Null]	[Null]
20	Value-Savings-Stocks	String	[Null]	[Null]	[Null]	[Null]	0	3	[Null]	[Null]

This dataset contains 20 variables including 11 continuous variables and 0 categorical variables. We can gain more overview from the histogram of each variable.

Before we jump into the model part, we'll clean and manipulate the dataset for model preparation.

1. Missing values

- The `Duration of current address` has 69% missing data so this field should be removed.
- We'll impute the missing value with median data in the `Age-year` field since it only has 2.4% missing data. Since the age is right skewed, so we'll impute the data with median instead of mean.

2. Association Check: Correlation

- We want to make sure the numeric variables are not highly correlated with each other.

Pearson Correlation Analysis

Full Correlation Matrix

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

	Type.of.apartment	Occupation	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	0.126967	NaN	-0.185180	0.238437	-0.207298
Credit.Amount	0.100413	NaN	0.082721	0.192532	-0.045994
Instalment.per.cent	0.178926	NaN	-0.293380	0.038515	-0.155458
Duration.in.Current.address	-0.163386	NaN	-0.036814	0.055112	-0.015787
Most.valuable.available.asset	0.182744	NaN	0.019435	0.083395	0.071932
Age.years	0.208552	NaN	0.046996	0.141103	-0.020939
Type.of.apartment	1.000000	NaN	-0.010189	0.179688	-0.026742
Occupation	NaN	1.000000	NaN	NaN	NaN
No.of.dependents	-0.010189	NaN	1.000000	-0.097632	0.218454
Telephone	0.179688	NaN	-0.097632	1.000000	-0.168472
Foreign.Worker	-0.026742	NaN	0.218454	-0.168472	1.000000

- An assoication analysis is performed on the numeric variables and there are no variables which are highly correlated with each other (the `abs(correlation) is > 0.7` and the p-value is also not significant).

3. Varibility Check

- We also want to remove data with low variability. Referring to the fields summary plots above, `Guarantors` , `Foreign Worker`, `No of Dependents` show low varibility where more than 80% of the data skewed towards to one value. These three fields should be removed in order not to skwe our model results.

4. Irrelevancy Check

- `Telephone` field should be removed since its irrelevancy to the creditworthy.

Summary

Category	Field	Process
Missing Value	Duration of current address	Removed
	Age-Years	Impute missing with median
Low Varibility	Guarantors	Removed
	Foreign Worker	Removed
	Occupation	Removed
	Concurrent Credits	Removed
	No of Dependents	Removed
Irrelevancy	Telephone	Removed

Step 3: Train Classification Models

- First, I created Estimation and Validation samples where 70% of the dataset should go to Estimation and 30% of entire dataset should be reserved for Validation. Set the Random Seed to 1.

- Then I'll create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model.

- The target variable for all models is `credit application result`.

1. Logistic Regression (Stepwise)

- summary of the model

Report for Logistic Regression Model LogisticModel_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	.

- From this model we can tell that `Account-Some balance` ,`payment status CreditSomeProblems` , `Purpose`, and `Credit Amount` are the significant predictor variables with significant p-value.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticModel_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

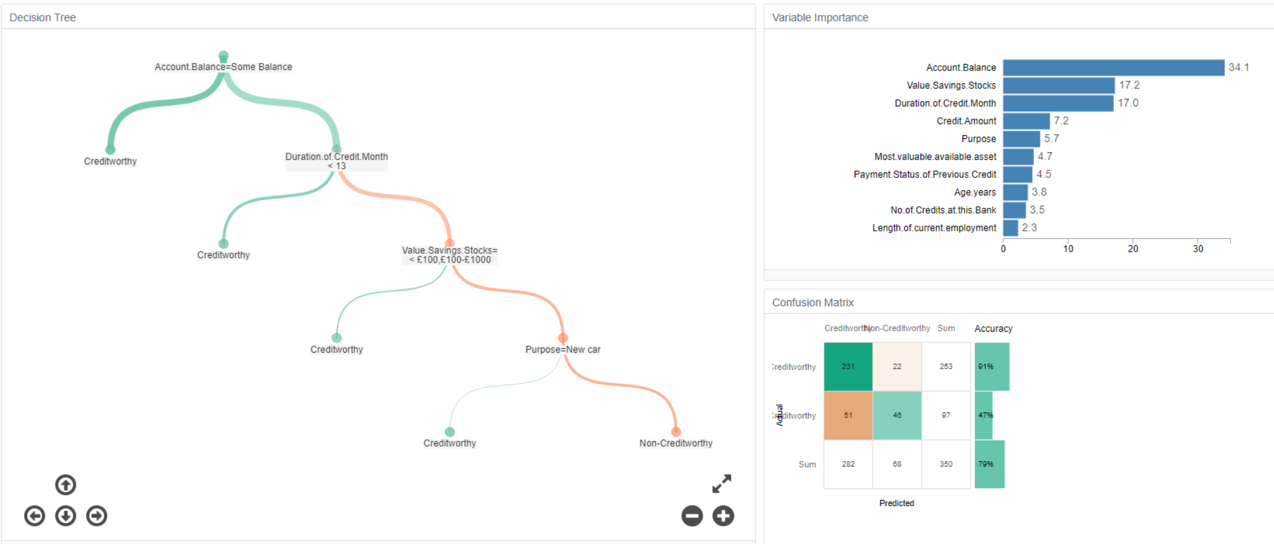
Confusion matrix of LogisticModel_Stepwise

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

While this stepwise model has 76% accuracy and its Non-creditworthy group is 48%.

2. Decision Tree

- Decision Tree



- From this model we can tell that `Account-Some balance`, `Value Saving Stock`, `Duration of Month` are the significant predictor variables with high variable importance.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222

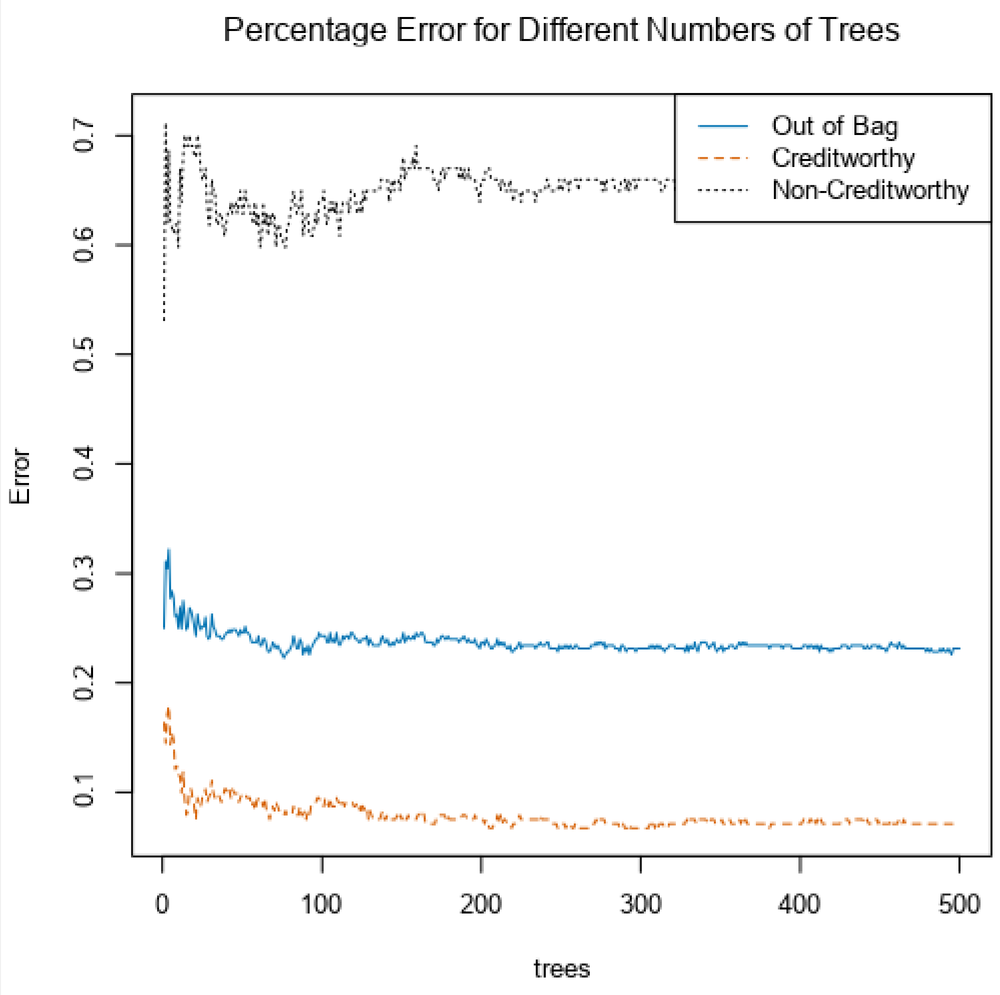
Confusion matrix of DT

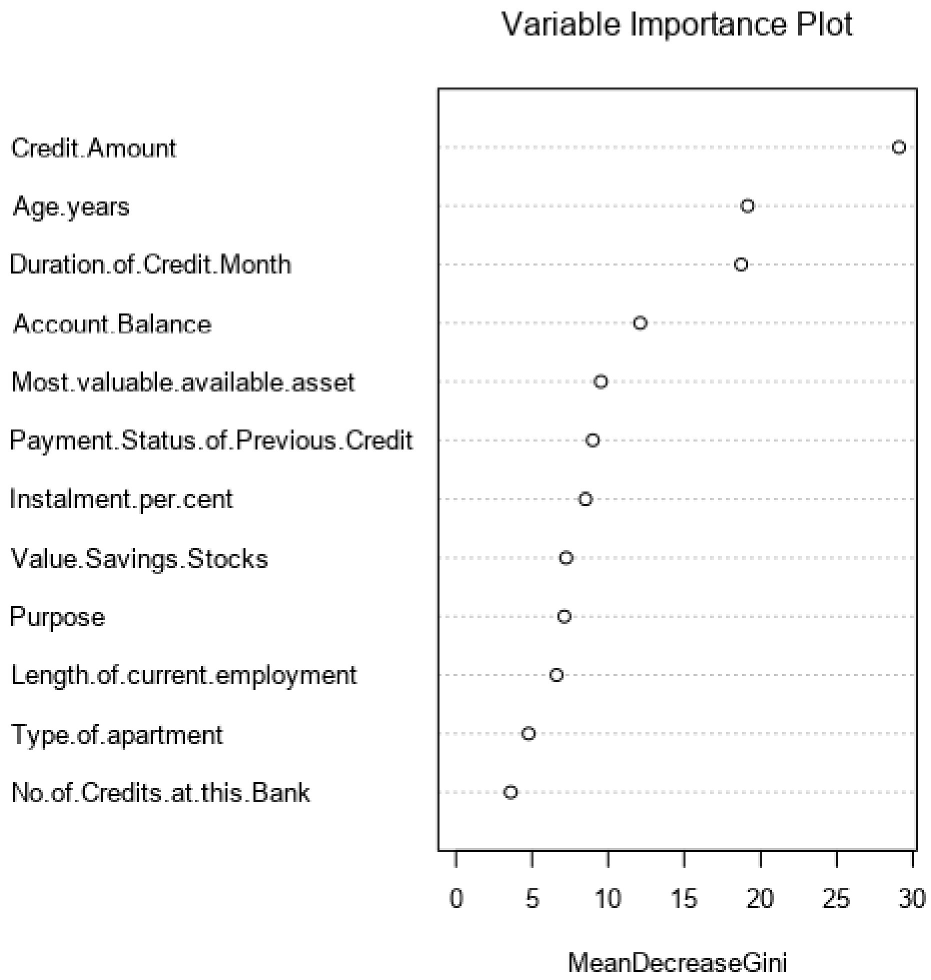
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

- The overall Model Accuracy is 79% while the accuracy for Non-creditworthy group is 42%.

3. Forest Model

- Random Forest Model





- From this **Variable Importance Plot** we can tell that `Credit Amount` ,`Age.years` , `Duration of credit month` are the significant predictor variables with high variable importance.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RandomForest	0.7933	0.8670	0.7403	0.9619	0.4000

Confusion matrix of RandomForest

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

- The overall acuracy is 79.33% and the Non-creditworthy accuracy is 40%.

4. Boosted Model

- Boosted Model

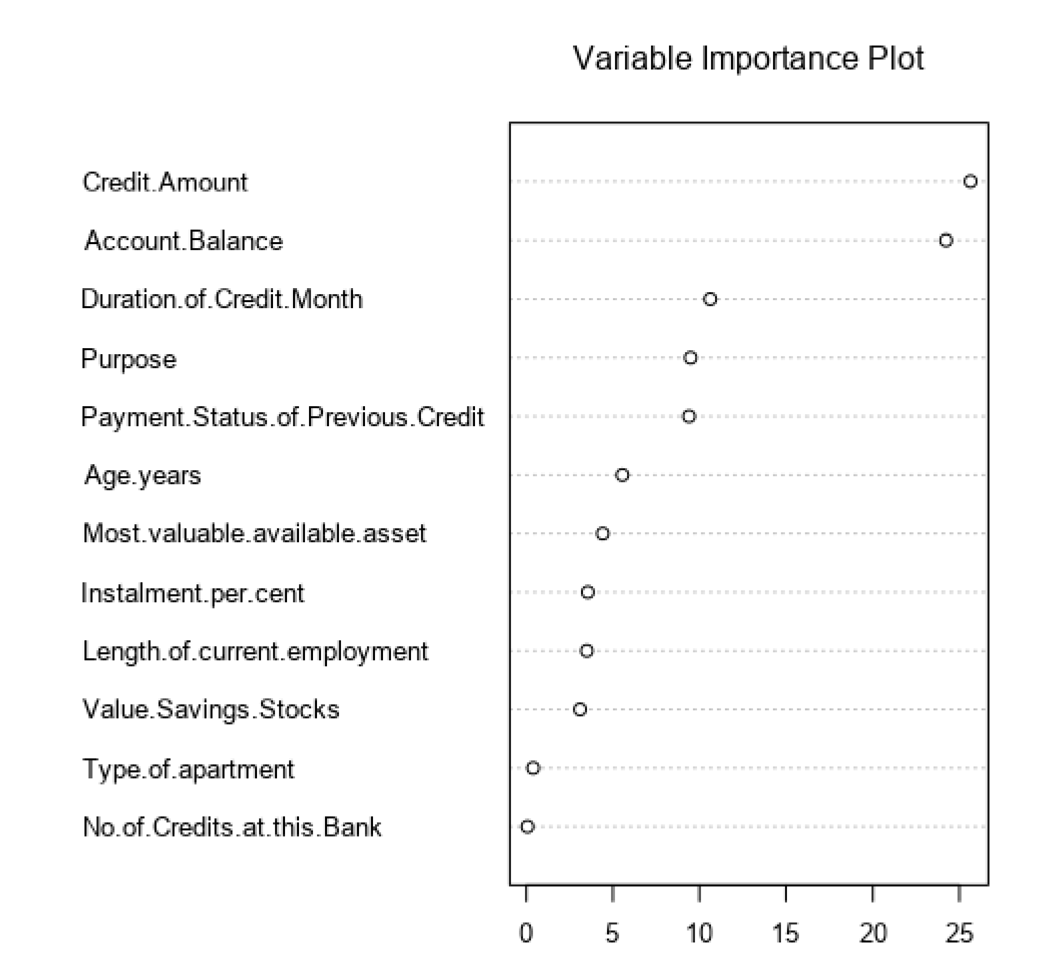
Report

Report for Boosted Model BoostedModel

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1988

Plots:



- From this **Variable Importance Plot** we can tell that `Credit Amount` ,`account balance` , `Duration of credit month` are the significant predictor variables with high variable importance.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BoostedModel	0.7933	0.8670	0.7469	0.9619	0.4000

Confusion matrix of BoostedModel

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

- The overall acuracy is 79.33% and the Non-creditworthy accuracy is 40%.

Step 4: Write-up

Final Model Compare

Here is the final four model comparison

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222
RandomForest	0.7933	0.8670	0.7403	0.9619	0.4000
LogisticModel_Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
BoostedModel	0.7933	0.8670	0.7469	0.9619	0.4000

Confusion matrix of BoostedModel

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

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of LogisticModel_Stepwise

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

Confusion matrix of RandomForest

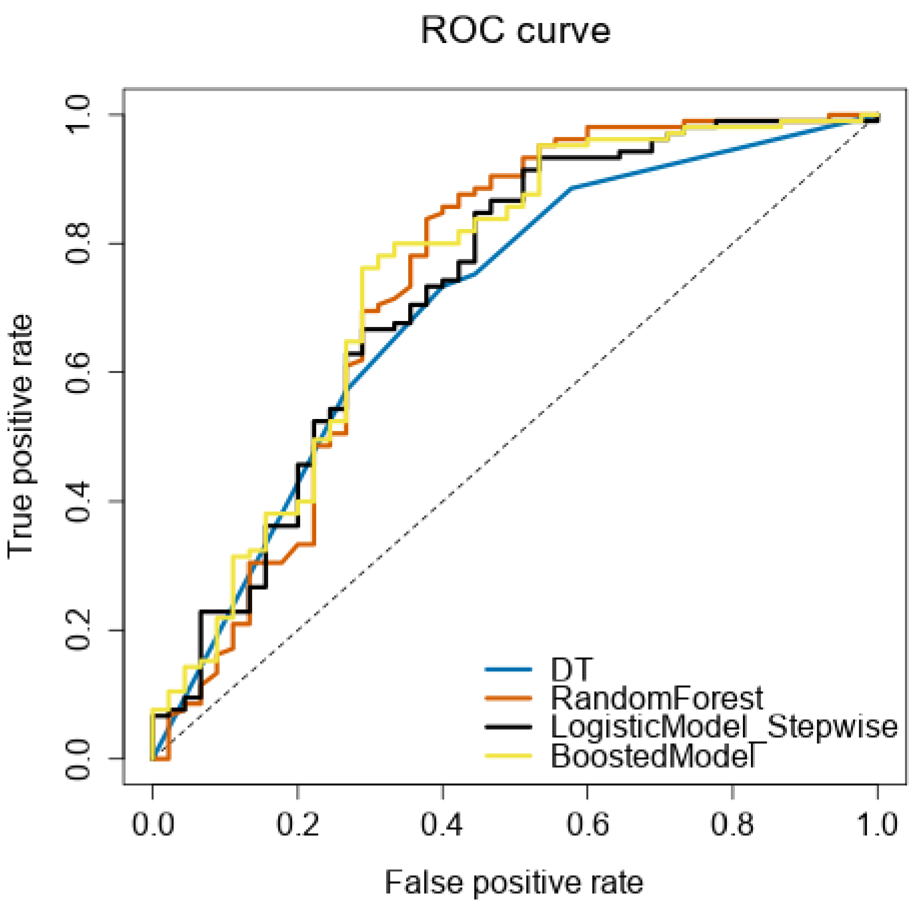
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Overall Accuracy

- We can find that both `Random forest model` and `boosted model` have the top 79.33% accuracy rate.

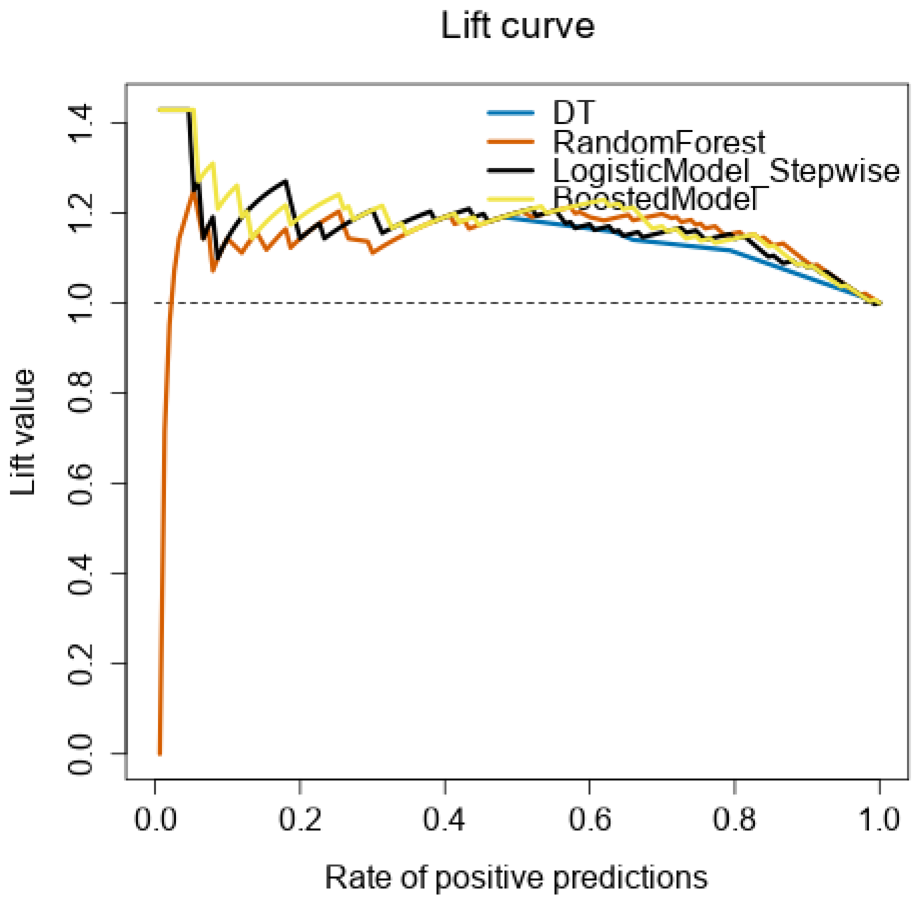
Accuracies in each segments

- Creditworthy: `Random forest model` and `boosted model` 96%
- Non-Creidtworthy: `Logistic Regression Model` 48%

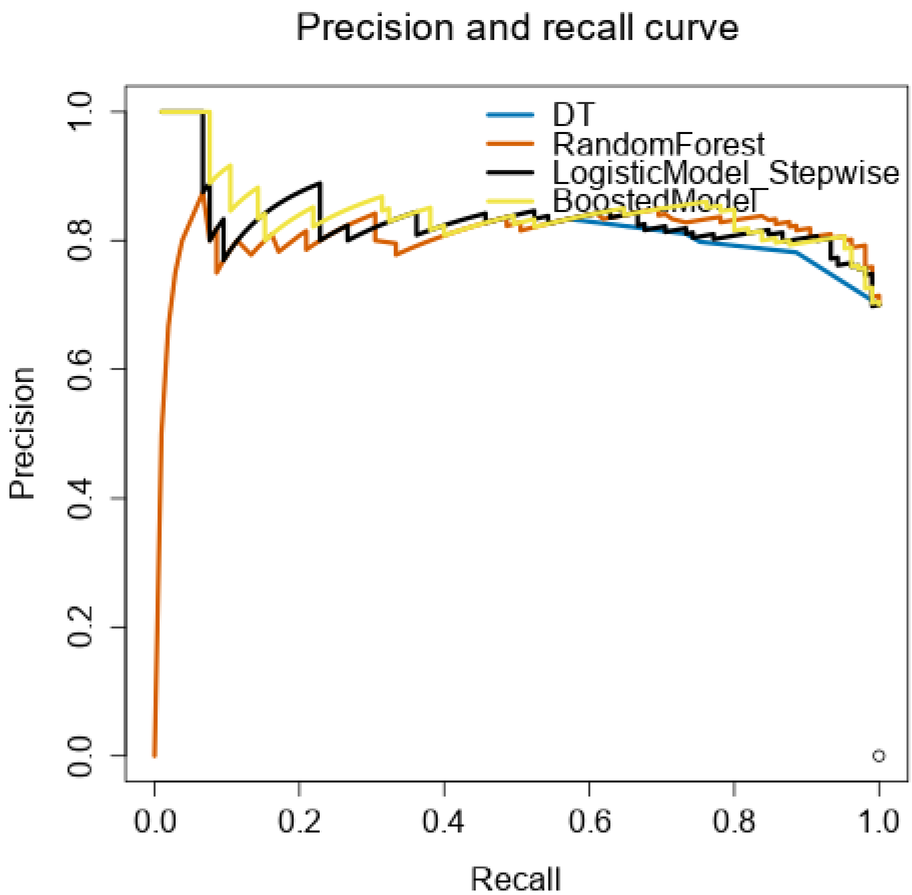


ROC curve

- From the ROC curve we can see that the Random forest performs slightly better than others.



Other supporting Plots



By comparing the above aspects, i choose the random forest model as it has the higher accuracy and less bais among two segments.

Q: How many individuals are creditworthy?

- In this step I used the random forest model to predict the customers from (customer-to-score) file. If the score of creditworthy is greater than score noncredictworthy then the person should be labeled as creditworthy.
- The final result is there are 410 creidtworthy customers and 90 non-creditworthy customers.

Appendix

Alteryx Workflow