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.

You have the following information to work with:

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

Step 1: Business and Data Understanding

Key Decisions:

- **What decisions needs to be made?**

  We need to determine whether the new customers based on the data provided are creditworthy for a loan.

- **What data is needed to inform those decisions?**

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

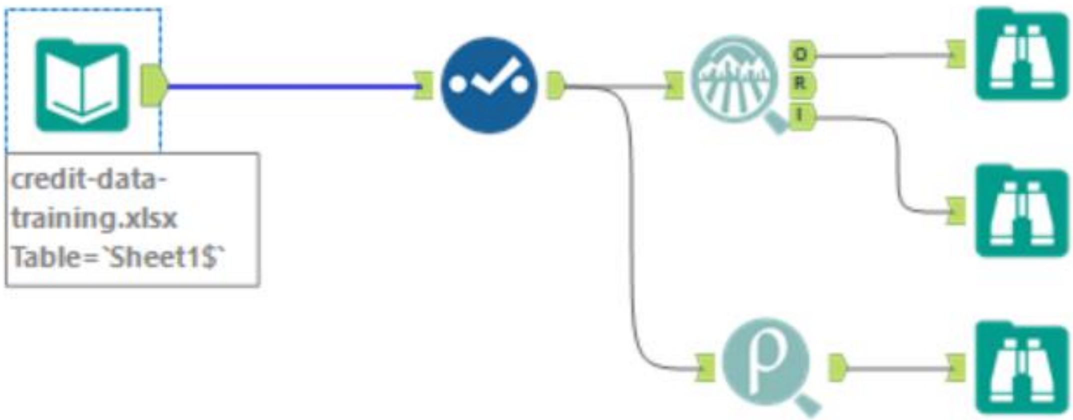
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

  We are trying to determine whether the new customers based on the data provided are creditworthy for a loan. **Involve Binary 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.***

To build the training set, we started to visualize the data thought the field summary and pearson correlation functions.

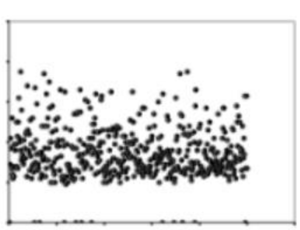


Then we decoded to remove the following field for the following reason with illustration below:

Removed Field	Reason
Duration-in-Current-address	Many Missing data
Concurrent-Credits	Low variability, only "Other Banks/Depts"
Guarantors	Low variability
Occupation	Low variability, only "1"
No-of-dependents	Low variability
Telephone	No relevant data
Foreign-Worker	Low variability

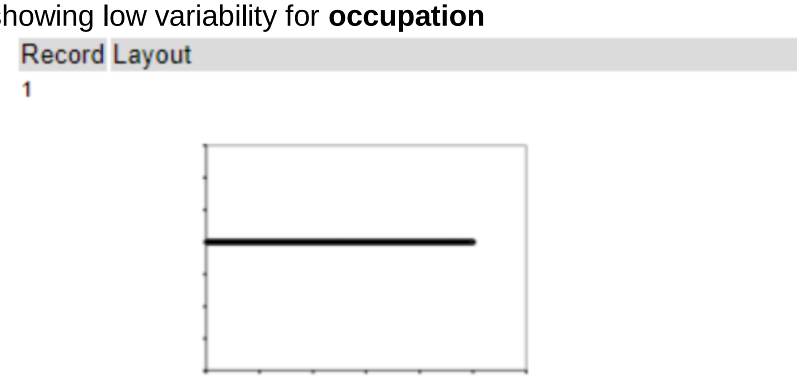
We also decided to impute **age-years** for the few missing data to the median

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean	La
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295	La

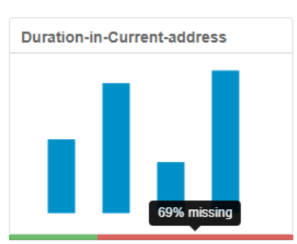


Field summery showing low variability for concurrent credit, foreign worker, non of dependent, guarantor.

Field summary showing low variability for **occupation**



Field summary showing 69% of missing data for duration in **current address**

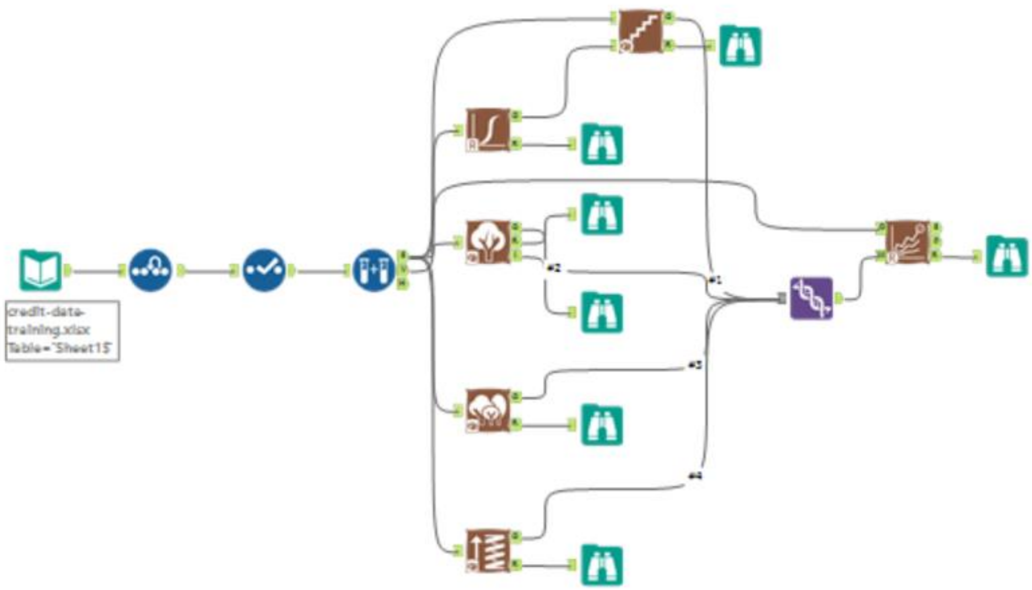


However, the Pearson correlation table does not show high correlation between data (*>%70*)

Record	FieldName	Duration-of...	Credit-A...	Instalment...	Duration-in-Curr...	Most-valua...	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worl
1	Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269	0.143176	-0.115916
2	Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986	0.286338	0.025493
3	Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894	0.029354	-0.133411
4	Duration-in-Current-address	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]
5	Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454	0.203509	-0.146005
6	Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
7	Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738	0.101443	-0.089848
8	Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
9	No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1	-0.048559	0.065943
10	Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559	1	-0.055516
11	Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943	-0.055516	1

Step 3: Train your Classification Models

We create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. And we Create the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model.



- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Below is a summary table of the significant predictor variable under the difference model.

Model Type	significant predictor variable
Logistic Regression	Account balance Payment status of previous credit Purpose Credit amount Length of current employment Instalment per cents
Decision Tree	Account Balance Duration of the credit in months Value savings stocks
Forest Model	Credit amount Age in years Duration of the credit in months Account Balance
Boosted Model	Credit amount Account Balance Duration of the credit in months Payment status of previous credit

Report

Report for Logistic Regression Model Logisic_Regression

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Length.of.current.employment + Credit_Amount + Instalment_per_cent + Most_valuable_available_asset, family = binomial(logit), data = the.data)

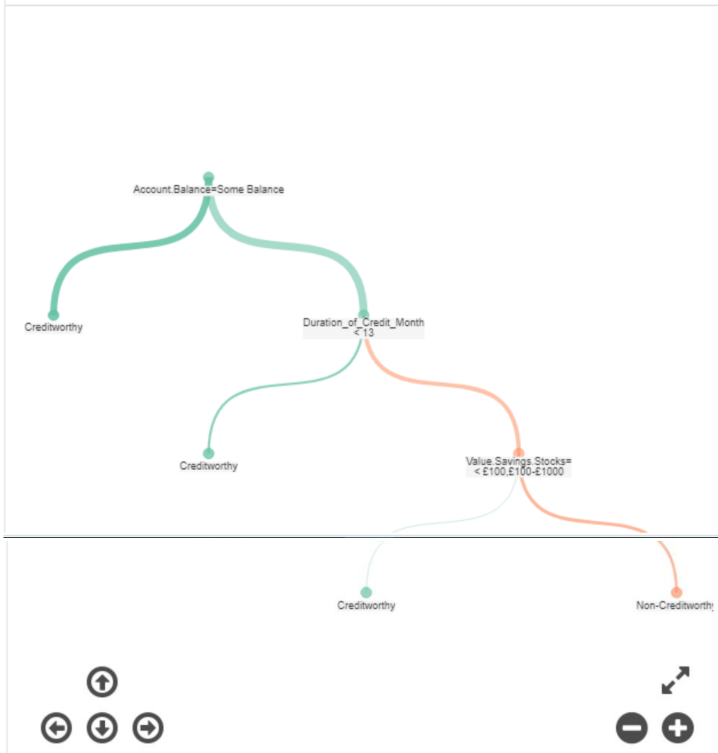
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

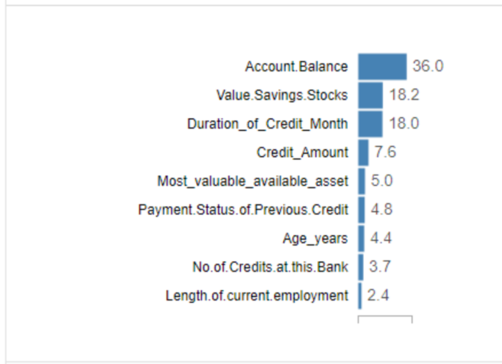
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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 .
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 *
Credit_Amount	0.0001704	5.733e-05	2.9716	0.00296 **
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 .

Decision Tree

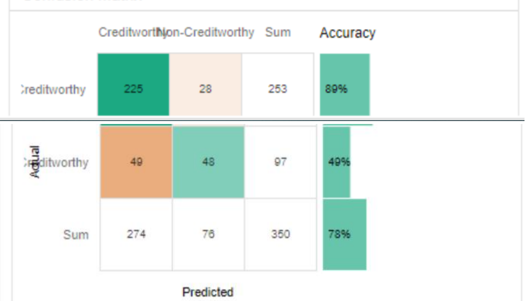


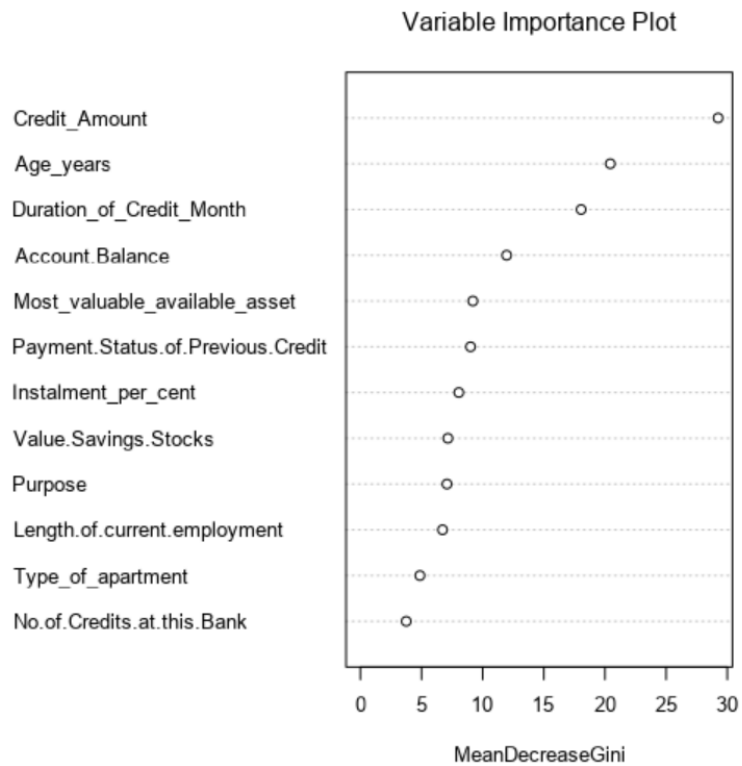
Mouseover to see details. Click to select a node. Click outside the graph to reset selection.

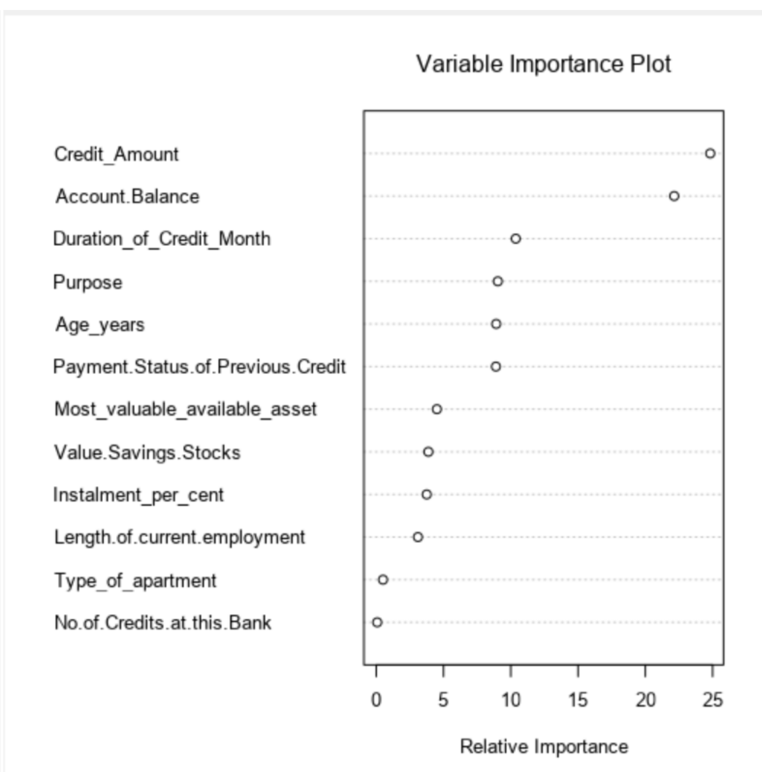
Variable Importance



Confusion Matrix




Variable Importance Plot
Credit_Amount
Age_years
Duration_of_Credit_Month
Account.Balance
Most_valuable_available_asset
Payment.Status.of.Previous.Credit
Instalment_per_cent
Value.Savings.Stocks
Purpose
Length.of.current.employment
Type_of_apartment
No.of.Credits.at.this.Bank
0 5 10 15 20 25 30
MeanDecreaseGini


Variable Importance Plot
Credit_Amount
Account.Balance
Duration_of_Credit_Month
Purpose
Age_years
Payment.Status.of.Previous.Credit
Most_valuable_available_asset
Value.Savings.Stocks
Instalment_per_cent
Length.of.current.employment
Type_of_apartment
No.of.Credits.at.this.Bank
0 5 10 15 20 25
Relative Importance

- 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?

We used the model comparison function to complete the different model showing their respective overall accuracy and confusion matrix.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8067	0.8755	0.7381	0.7969	0.8636
Boosted_Model	0.7867	0.8621	0.7526	0.7874	0.7826
stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

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	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of Decision_Tree

	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 stepwise

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

For the logistic regression, the overall accuracy is 76% though the accuracy to predict creditworthiness is quite good at 80%, however the accuracy to predict non- creditworthiness is quite low at 82.86%.

The situation is quite similar for the decision tree with an overall accuracy of 74.67% and good accuracy to predict creditworthiness at 79.13% but the accuracy to predict non- creditworthiness is quite low at 60%. The confusion matrix of the decision tree summary above shows also a very low accuracy for prediction of non- creditworthiness customers.

There is indeed a bias induced by less sample of non- creditworthiness clients.
In fact, in our training data only 28.4% of the total customers are tagged non- creditworthiness.

However, the forest and boosted model seems to perform better than the logistic regression and decision tree model. Their overall accuracies are 80% and 79.33% respectively and their accuracy rates to predict non- creditworthiness customers are even higher than accuracy rates to predict creditworthiness customers at 82.62% and 81.62% respectively.

Step 4: Writeup

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

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8067	0.8755	0.7381	0.7969	0.8636
Boosted_Model	0.7867	0.8621	0.7526	0.7874	0.7826
stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

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)

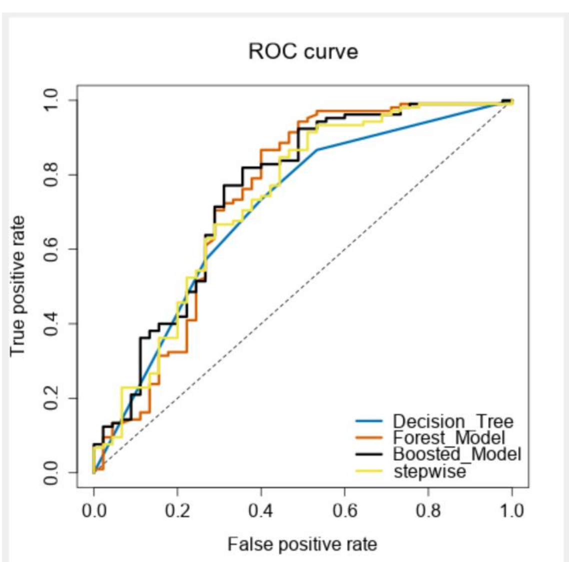
a. **Overall Accuracy against your Validation set:**

Based on the model comparison report, it appears that the Forest Tree has the highest accuracy rate with 80% compared to other models.

b. **Accuracies within "Creditworthy" and "Non-Creditworthy" segments:**

Within "Creditworthy" and "non- Creditworthy" segments, the logistic model appears to have the highest accuracy to predict "Creditworthy" however the accuracy for "Non-Creditworthy" is quite low at 62.86%.
The Forest Tree model has again the highest accuracy for both "Creditworthy" and "Non-Creditworthy"

c. **ROC graph**



d. Bias in the Confusion Matrices

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

	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 stepwise

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

From the confusion matrices, we see that the boosted model and forest model tend to classify more non-creditworthy customers as creditworthy while the decision tree model and logistic model tend to classify creditworthy customer as non- creditworthy.

However, since the boss only care for prediction accuracy, we chose the forest model which has the highest accuracy overall.

- How many individuals are creditworthy?

Finally, we use the score tool to predict the creditworthiness of the new customers.
The model predicts that 415 customers out of 500 are creditworthy.