Project: Creditworthiness

CHANDAN MISHRA

Step 1: Business and Data Understanding

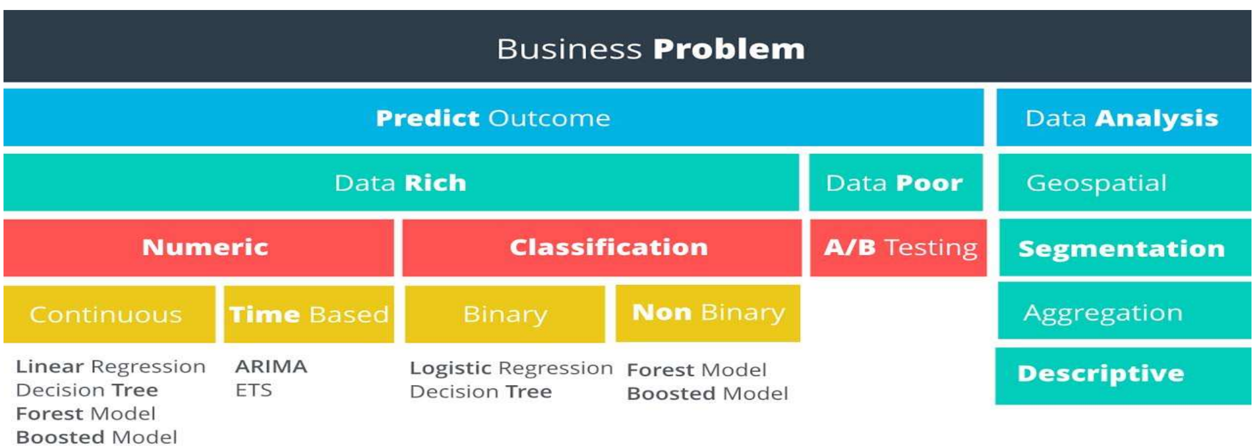
- What decisions needs to be made?

  The goal is to determine if a customer who is applying for the loan is creditworthy or not and provide a list to the manager of all the creditworthy customers.

- What data is needed to inform those decisions?

  - Data on all past applications where we have information whether a customer was creditworthy or not, customer's account balance, customer's credit amount etc.
  - List of customers that are needed to be processed

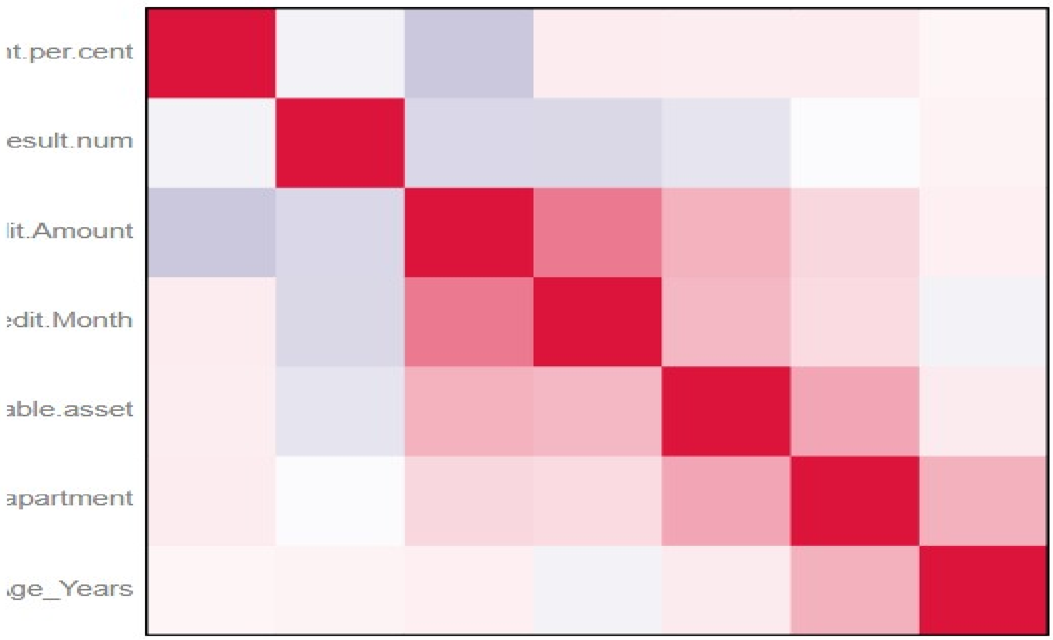
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?



Since we have to predict and classify customers as creditworthy or not creditworthy, so it's a binary classification. We will use binary classification model such as Logistic regression, Decision tree, forest model and boosted model.

Step 2: Building the Training Set

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



Association analysis was checked for the numeric variables and none of the variables were highly correlated.

	Association Measure
Duration.of.Credit.Month	-0.204317
Credit.Amount	-0.200990
Most.valuable.available.asset	-0.137917
Instalment.per.cent	-0.065345
Age_Years	0.056737
Type.of.apartment	-0.021860

Field Summary report:

Concurrent credits and **Occupation** have just one level and hence this variable is of no importance.

Guarantors – 457 none and 43 yes, shows that its heavily skewed towards none.

Duration in current address – 67% of the data in this column is missing hence we can ignore this column.

Foreign workers – 481 instances of 1 and 19 instances of 2 shows its heavily biased towards 1 and hence it can be removed.

No of Dependents – 427 instances of 1 and 73 instances of 2 shows that its heavily biased towards 1 and hence can be removed.

Telephone was removed because of its irrelevancy.

Impute values:
Since we have 2% values in age missing, we can impute these using median. Median age is used rather than mean since data is skewed towards the left and not distributed normally.

Step 3: Train your Classification Models

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

Logistic Regression Model (Stepwise)

Report for Logistic Regression Model stepwise_log

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 .

Account balance, Purpose, Credit amount are the most important variables having p-values less than 0.05 for predicting the credit application result.

Fit and error measures

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

Confusion matrix of stepwise_log

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

Overall model accuracy is 76.0% while accuracy for creditworthy is higher than non-creditworthy at 87.6% and 42.9% respectively. The model is biased towards predicting customers as non-creditworthy.

Decision Trees

Account balance, value savings stocks and duration of credit month are the top three most important predictor variables for predicting credit application result.

Using model comparison report, the accuracy is 74.7%. Creditworthy accuracy is 86.7% while not creditworthy accuracy is 46.7%. The model is biased towards predicting customers as non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Predict	0.7467	0.8273	0.7054	0.8667	0.4667

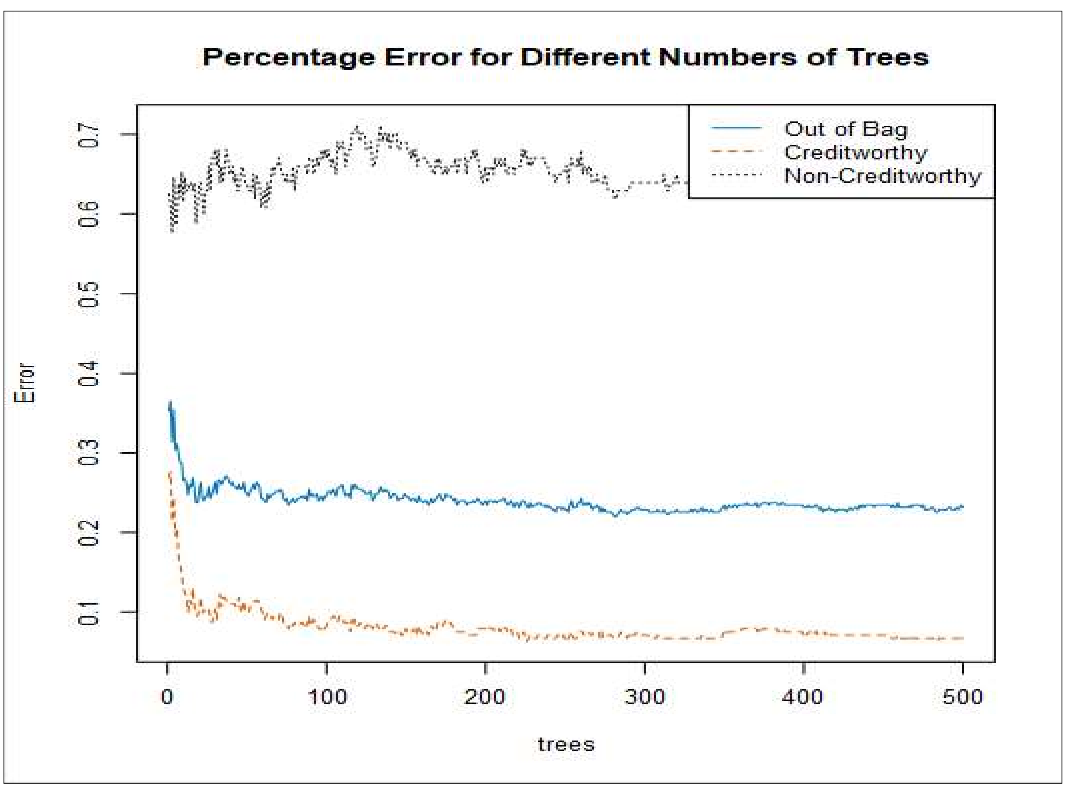
Confusion matrix of DT_Predict

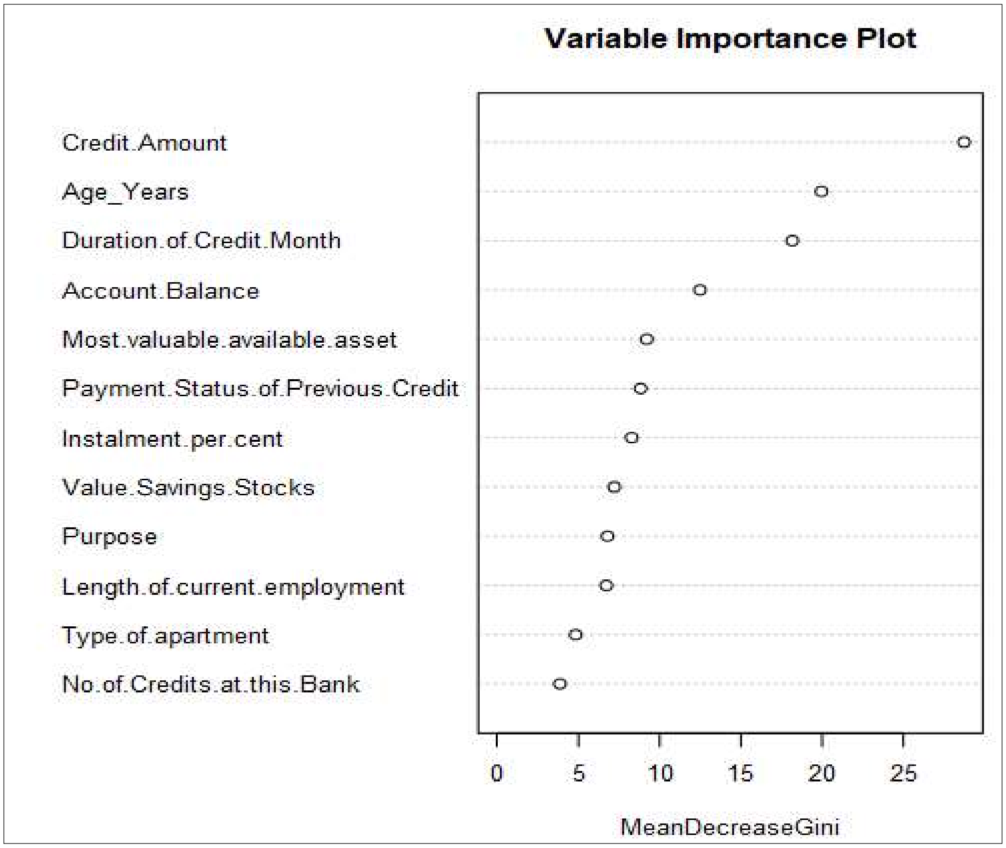
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Forest Model

Credit Amount, Age Years and duration of credit month are the top three most important predictor variables for predicting credit application result.

Overall model accuracy is 79.3% and creditworthy accuracy is 97.1% and not creditworthy accuracy is 37.8%. The model is biased towards predicting customers as non-creditworthy.





Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Mod	0.7933	0.8681	0.7368	0.9714	0.3778

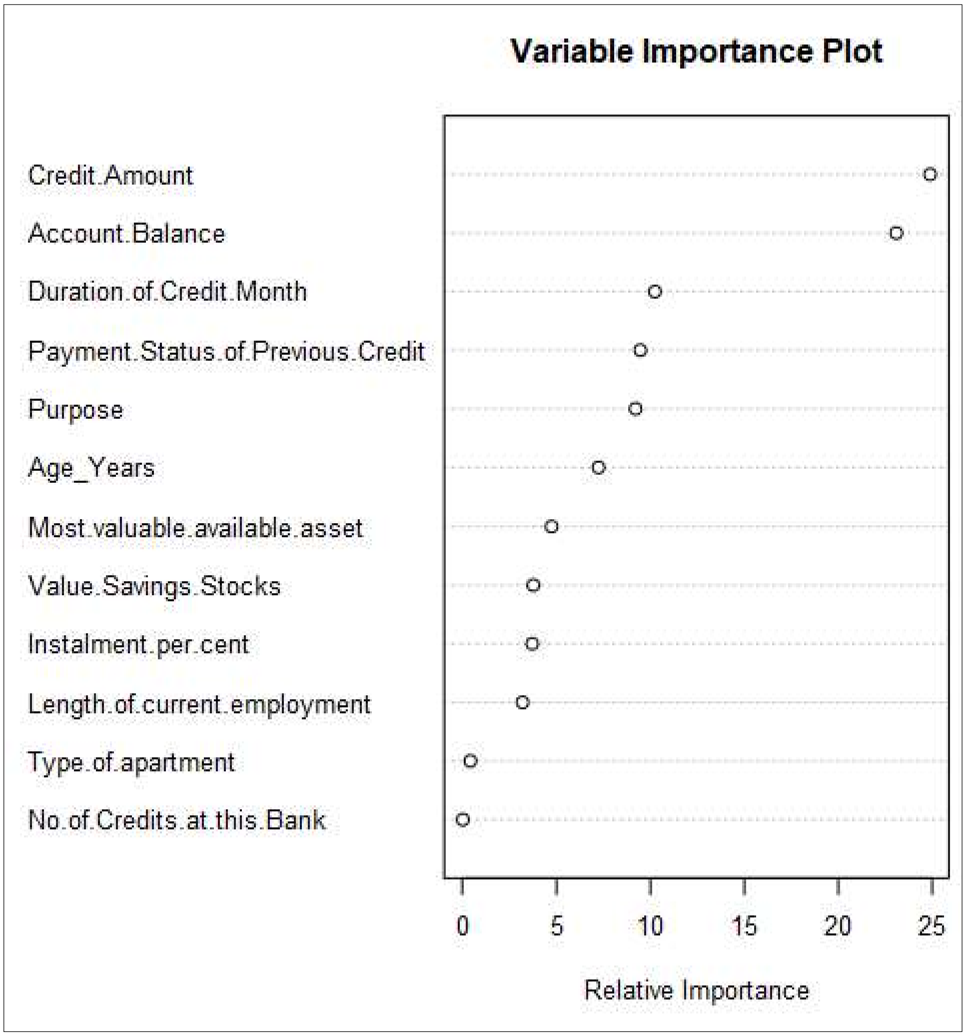
Confusion matrix of Forest_Mod

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

Boosted Model

Credit Amount, Account balance and duration of credit month are the top three most important predictor variables for predicting credit application result.

Overall model accuracy is 79.3% and creditworthy accuracy is 96.1 and not creditworthy accuracy is 40%. The model is biased towards predicting customers as non-creditworthy.



Fit and error measures

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

Confusion matrix of Boosted_Mod

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

Step 4: Writeup

- 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.

Forest model is chosen since its accuracy is 79.3% which is same as boosted model but has higher creditworthy accuracy at 97.1% compared to 96.1% which signifies more business for the company as it can lend more to creditworthy customers. The non- creditworthy accuracy for forest model is 37.7% compared to 40% in boosted model, which is close enough. It allows the bank to avoid lending it to the customers who will default. Since the creditworthy accuracy and non-creditworthy accuracy has a significant difference, the model is slightly biased towards non-creditworthy customers.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Predict	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Mod	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Mod	0.7933	0.8670	0.7509	0.9619	0.4000
stepwise_log	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted_Mod

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

Confusion matrix of DT_Predict

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

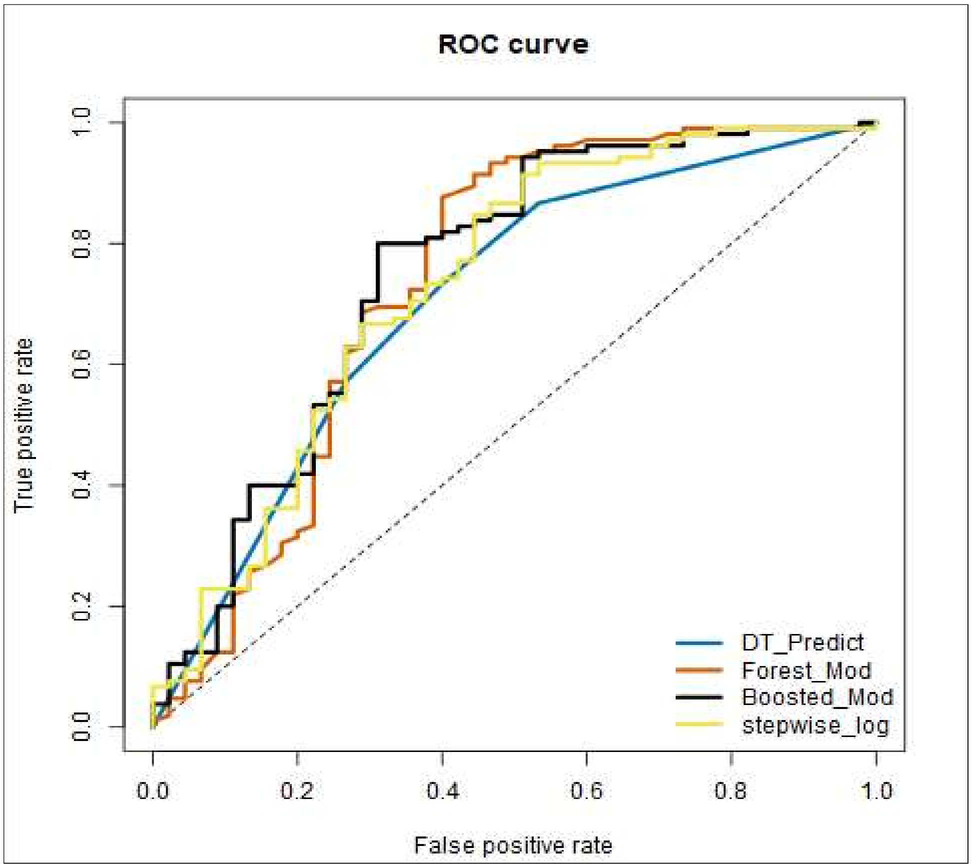
Confusion matrix of Forest_Mod

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

Confusion matrix of stepwise_log

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

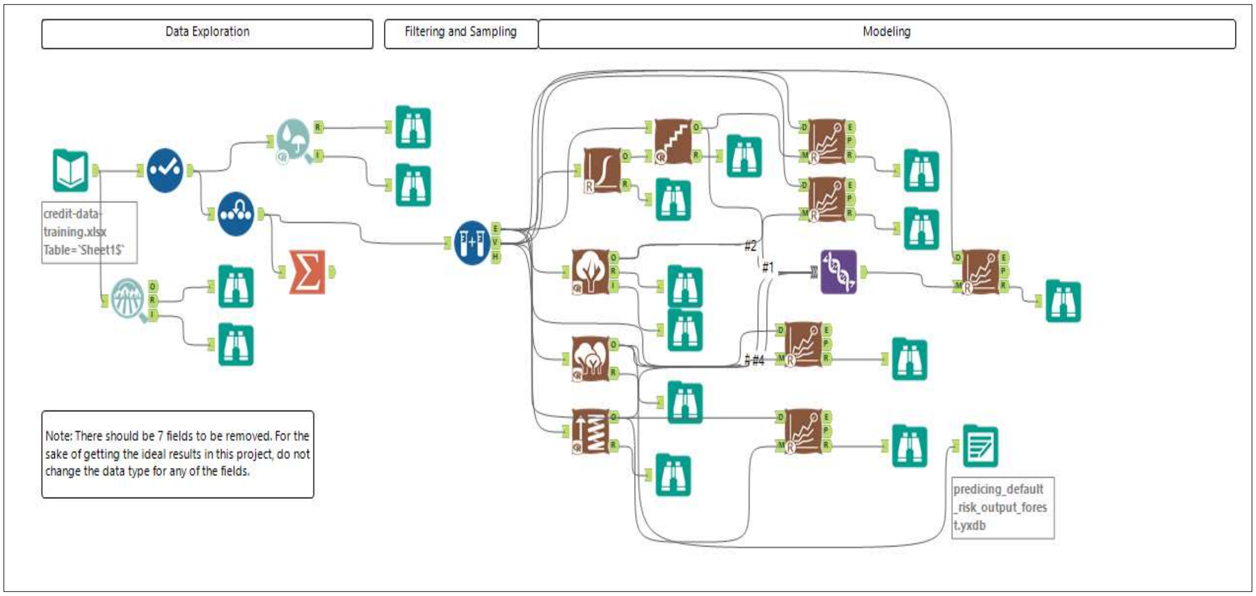
Forest and boosted model are almost same in accuracy but in ROC curve, we can see forest model reaches the True positive rate (TPR) at the fastest rate amongst all the models and hence is selected.



- How many individuals are creditworthy?
  408

Alteryx workflow:

- Building the model



- Scoring the model