Project: Creditworthiness

Project Overview

You are a loan officer at a young and small bank (been in operations for two years) that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. You'll use a series of classification models to figure out the best model and provide a list of creditworthy customers to your manager. The manager only cares about how accurate you can identify people who qualify and do not qualify for loans for this problem.

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

  The goal is to build a model which decides whether a customer is creditworthy or not.

- What data is needed to inform those decisions?

  We have a lot of variables we can use to create a model necessary to see if a customer is creditworthy. The kind of data we need is whether the customer has no account or some balance, if they have previous loands, the duration the customer has had a loan, whether the customer has paid up previous loans, for what purpose the customer wants to get a loan, amount of loan, length of the customer's current employment, rate of instalment on loan, age of the customer and so on.

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

  This is a problem where we have historical data and we use it to build a model to predict whether customers are creditworthy or not. So we are data rich and we need a model which classifies our data into two outcomes – meaning we need to use a binary classification model.

Step 2: Building the Training Set

- In your cleanup process, which fields were removed or imputed?

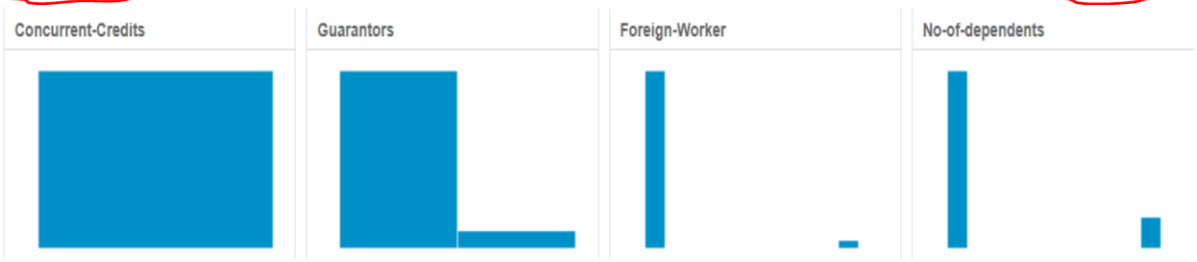
  Fields with missing data:
  Duration-in-current-address – dropped the variable because 68.8% data was missing
  Age years – only 2.4% data was missing so I imputed to add median 33 for missing values – rounded average Age years -> 36

Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values
Age-years	Numeric	19	75	33	11.501522	2.4	54
Credit-Amount	Numeric	276	18424	2236.5	2831.386861	0	464
Duration-in-Current-address	Numeric	1	4	2	1.150017	68.8	5

Fields with low variability: here data tends to
Type 1 -> skew heavily towards one value or
Type 2 -> there is only one value in the field,
such fields do not help in making the model so we can remove them

Guarantors – Type 1
Concurrent-credits – Type 2
Occupation – Type 2
No-of-dependents – Type 1
Foreign worker – Type 2

Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values
Foreign-Worker	Numeric	1	2	1	0.191388	0	2
No-of-dependents	Numeric	1	2	1	0.35346	0	2
Occupation	Numeric	1	1	1	0	0	1
Concurrent-Credits	String	[Null]	[Null]	[Null]	[Null]	0	1
Guarantors	String	[Null]	[Null]	[Null]	[Null]	0	2



Suggested in supporting materials: Telephone-removed
So, from 20 variables, 7 variables have been removed and we go ahead with 13 variables. 70% of data is chosen as sample training data and 30% is chosen as validation data

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.

  The models differ in their opinion of the most important variable

  The most important variables according to stepwise log model -
  account balance, purpose, credit amount, payment status of previous credit, length of current employment, instalment per cent, and most valuable asset.

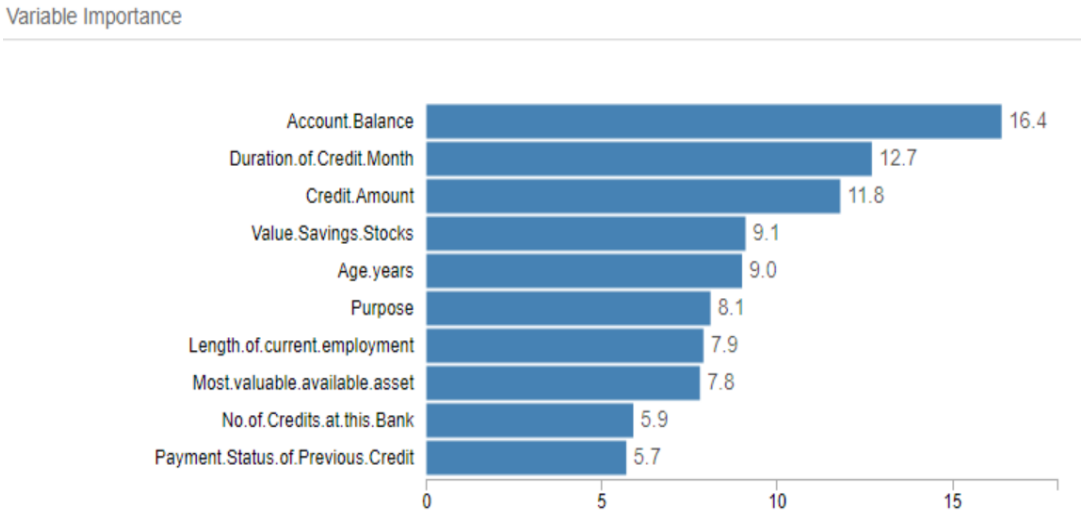
Variables and their p values

Coefficients:

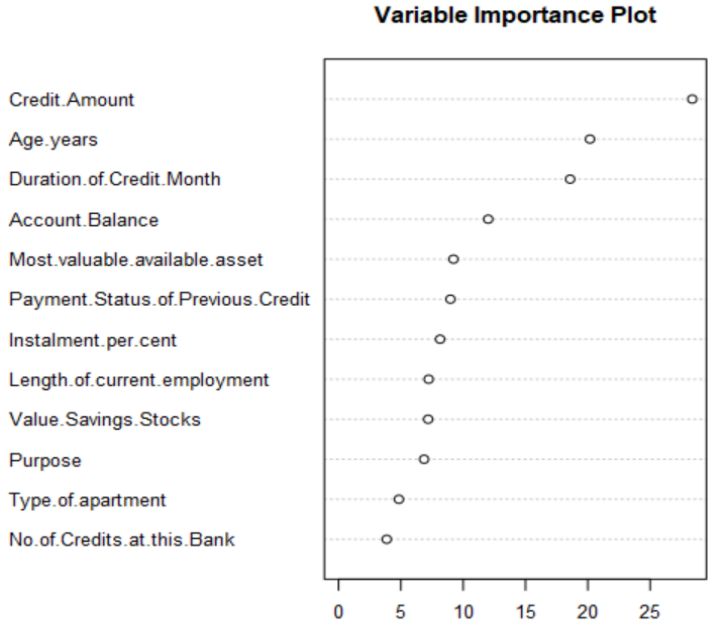
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01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

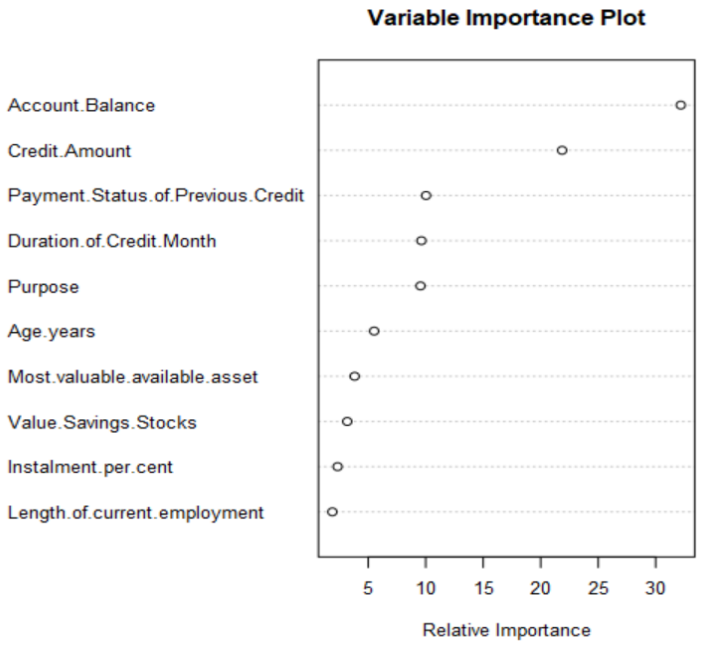
The most important variables according to Decision Tree model



The most important variables according to Random Forest model



The most important variables according to Boosted model



Since all four models seem to consider different variables as more/ less important, I have created a rank score where I take average of ranks of varibles used in the four models and arranged with the most important variable at the top.

	Models (variables ranked by importance)				
Variables	**StepwiseLog**	**Decision Tree**	**Random Forest**	**Boosted Model**	**Average_variable_rank_score**
Account Balance	1	1	4	1	1.75
Credit Amount	2	3	1	2	2.00
Duration of credit month	-	2	3	4	3.00
Age years	-	5	2	6	4.33
Payment Status	4	10	6	3	5.75
Purpose	3	6	10	5	6.00
Most valueable asset	7	8	5	7	6.75
Instalment per cent	5	-	7	9	7.00
Value Savings	-	4	9	8	7.00
Length of current employment	6	7	8	10	7.75
No of credits at this bank	-	9	12	-	10.50
Type of apartment	-	-	11	-	11.00

- 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 accuracy when input with the validation data:

Decision tree model – 67.33 %
Forest model – 80%, most accurate model
Boosted model – 78.67%
Stepwise logistic regression model – 76%

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree_model	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_model	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_logistic_model	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boosted_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of DecisionTree_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	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Stepwise_logistic_model

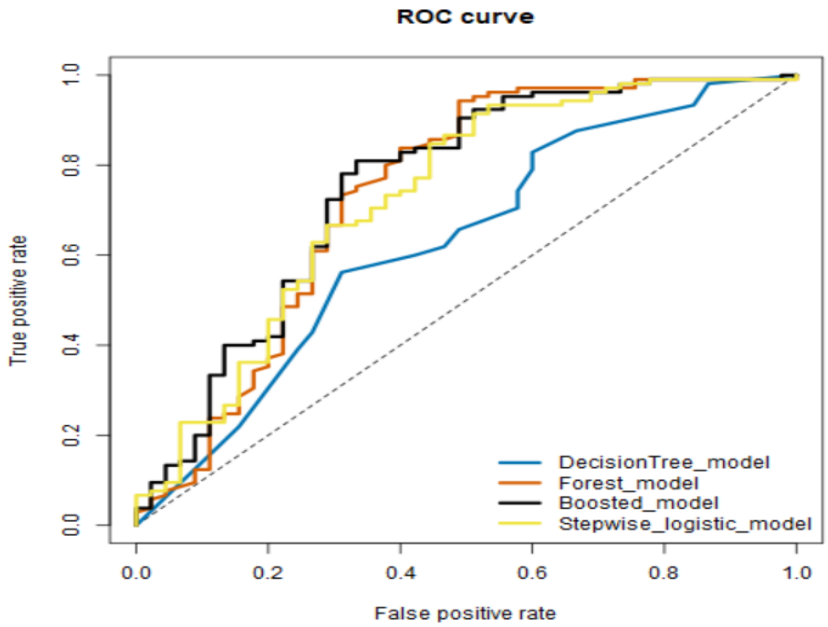
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

All the models have more difficulty is predicting non-creditworthiness i.e. a lot of customers who are not deserving are being labelled as creditworthy. Even the model which best predicts non-creditworthiness only has an accuracy of 48.89%. On the other hand, the models have a much easier times predicting creditworthiness, meaning its easier for the model to predict the deserving customers as creditworthy. The best models have an accuracy of 96.19% and the worst one does it with 79.05% accuracy.

Step 4: Scoring new customers

- Choice of model

  I have chosen Forest Model for the following reasons:
  - Overall Accuracy against your Validation set
    Forest model has the best accuracy of 80%.
  - Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    Forest model ranks #1 for Creditworthy segment and #2 for Non-Creditworthy.
    #1 for Non-Creditworthy, the Stepwise_logistic_model is ranked #3 for Creditworthy, so even considering these two segments Forest Model is a better choice.
  - ROC graph



From the graph, it does look like the Forest model reaches top quite fast, so the Forest Model is a reasonable choice

- Bias in the Confusion Matrices
  All the models seem similarly biased, in the sense that they have higher difficulty is predicting non-creditworthy customers as non-creditworthy and have a very easy time predicting creditworthy customers as creditworthy. Forest Model still better than all of them considering both Creditworthy and Non-Creditworthy segments.

Side note - *Personally, I would have gone for the stepwise logistic model considering the data is regarding loan applications for a bank. My logic for that would have been that for me I would have wanted higher accuracy for rejecting for Non-creditworthy applications. But the problem clearly states that the manager only cares for the higher accuracy of Creditworthy and Non-Creditworthy both, so my logic does not hold and Forest Model is the best choice.*

- How many individuals are creditworthy?

  Using the Forest Model, amongst 500 customers **406** are Creditworthy