Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

The key decision to be made is to determine which of the new customers who applied for loan are "Creditworthy" and who are not.

- What data is needed to inform those decisions?

We need data on previous applicants who were processed and given loans. We need information like employment, do they have other debts, age, income, location, purpose of applying for loan, Guarantors, Type of other property/assets they own, Concurrent credits, current account balance etc.;

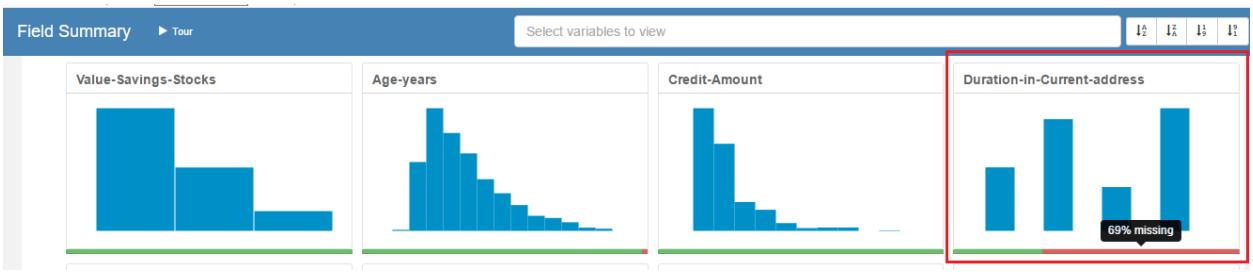
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

The decision we need to make is a binary classification. In order to make our decision we can use Logistic Regression, Decision Trees, Decision Forests or Boosted model to make our decision.

Step 2: Building the Training Set

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

We need to remove the "Duration-in-current address" field as 69% of data is missing here which is a huge amount of missing data.

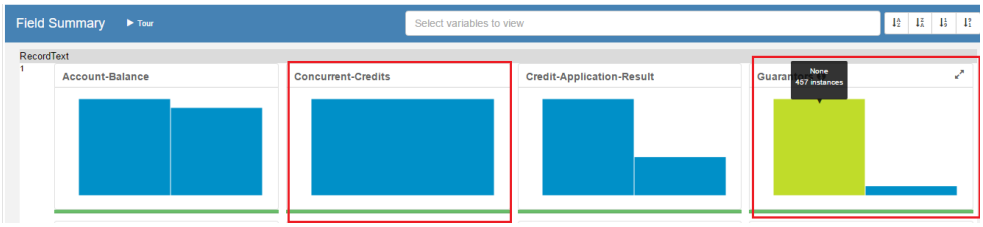


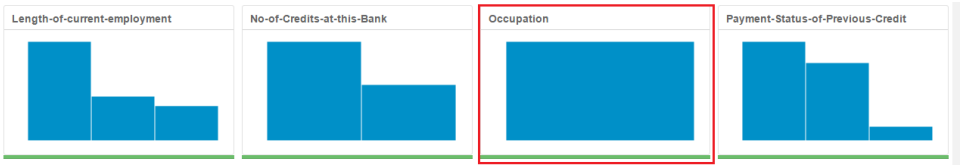
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

The following fields have been removed due to no or low variability:

- ➔ Concurrent-credits
- ➔ Guarantors
- ➔ Foreign-worker
- ➔ No-of-dependents
- ➔ Occupation

The “concurrent credits” field doesn’t have any variability in the data hence needs to removed.
In the “Guarantors” field the data is skewed towards one value “none” and hence there is low variability here.
We also need to remove the “Foreign Worker” and “No-of-dependents” fields as the variability in the data in low and they seemed to be skewed towards one value.
The “Foreign worker” field data is highly skewed towards a “1” value and “No-of-dependents” field is skewed towards “1” and hence both are fields are removed as predictors due to low variability in the data.
The “occupation” field as it has no variability in the data. Hence I removed this as predictor variable.

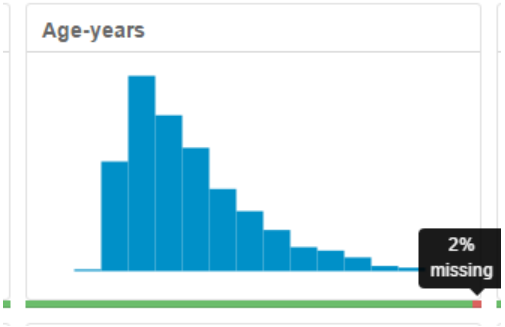




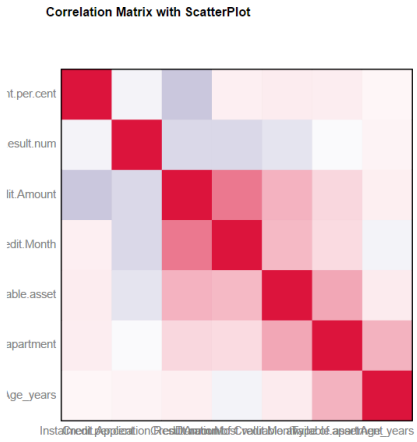
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

I removed the “Telephone” field from the dataset as there is no logical reason to include this as a predictor variable.

I have imputed the “Age-years” field with ‘median of values in this field’ to impute the data that was missing.
I choose to impute the data because, the more data we have we have a good chance of building a better model. Hence since only 2 % of data was missing in this field I decided to impute it with median of age-years. Also since the data is skewed, median would be an appropriate measure of center and hence imputed the missing data with median of values in this field.



- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered “high”.



After removing all the fields with low variability, missing data and imputing the missing values, I run the association analysis on the remaining fields to check if any fields are highly correlated (>0.70). Based on the correlation matrix, I can see that there are no fields that are high correlated with each other. Hence we can include all this fields as predictors to build our model.

<u>Step 3:</u> Train your Classification Models

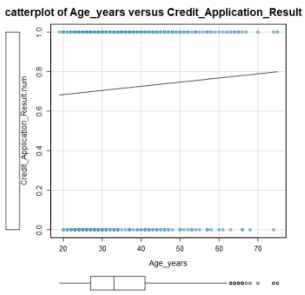
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

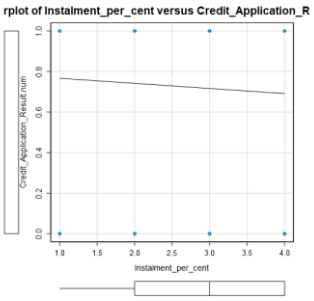
Focused Analysis on Field Credit.Application.Result.num

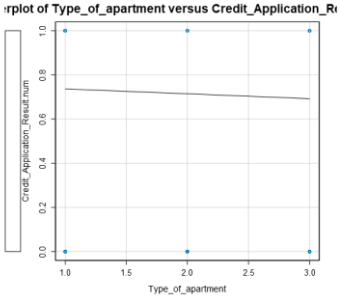
	Association Measure	p-value	
Duration.of.Credit.Month	-0.202504	5.0151e-06	***
Credit.Amount	-0.201946	5.3311e-06	***
Most.valuable.available.asset	-0.141332	1.5334e-03	**
Instalment.per.cent	-0.062107	1.6556e-01	
Age_years	0.052914	2.3758e-01	
Type.of.apartment	-0.026516	5.5417e-01	

Based on the p-values in the above report, I conclude that “Duration_of_credit_month”, “credit_amount” and “most-valuable-available-assets” are statistically significant variables.

To determine whether the remaining numeric field variables are statistically significant, I am using the scatter plots

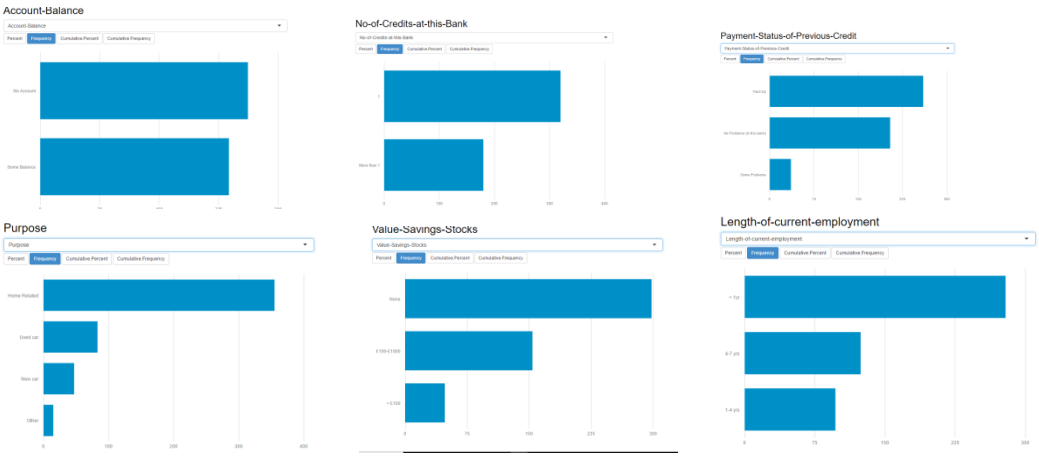






All there of these other variable “Age-years”, “Type-of-apartment” and “Installment-percent” have a linear relationship (either positive or negetive) with the target variable (which is “credit-application-result” considering “Creditworthiness” as the target of interest), I choose to include these variables also as predictors for building the model.

As for the remaining categorical variables, I used the Frequency table tool to check the frequency distribution of data. Based on the frequency maps, I conclude to use all these categorical variables as predictors.



- 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?

Based on the model comparison report, the Forest model performs better than other models with an overall accuracy at 80%, followed by boosted model with overall accuracy at 78%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
P4_DesTree	0.7467	0.8273	0.7054	0.7913	0.6000
P4_ForestModel	0.8000	0.8707	0.7419	0.7953	0.8261
p4_BoostedModel	0.7867	0.8632	0.7524	0.7829	0.8095
P4_Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Based on the report there is little bias in the confusion matrix.

Confusion matrix of P4_DesTree

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

Confusion matrix of P4_ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of P4_Stepwise

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

Confusion matrix of p4_BoostedModel

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

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
P4_DesTree	0.7467	0.8273	0.7054	0.7913	0.6000
P4_ForestModel	0.8000	0.8707	0.7419	0.7953	0.8261
p4_BoostedModel	0.7867	0.8632	0.7524	0.7829	0.8095
P4_Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

If we look at the report, the logistic regression stepwise and the decision tree are slightly biased towards identifying "Creditworthy" class than the "Not-creditworthy".
The Forest model and the boosted model are biased towards identifying "not creditworthy" class than the "creditworthy class".
If we look at our most accurate model (forest model), this is slightly biased towards identifying "not creditworthy" at 82% than identifying "creditworthy" which is at 79%. So there is a little bias in the forest model for correctly identifying the "not creditworthy".

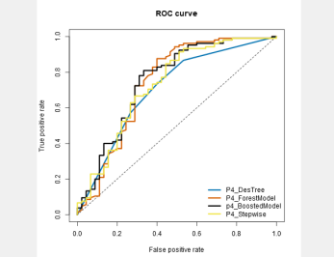
Step 4: Write up

- 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

Based on the above observations of Model Comparison report, if we look at the Fit and Error measures, the overall accuracy of "Forest model" is more than other models which is close to 80% so I would say this is a best model to score the new customers against.

Also, as we observe the Positive predicted rate and Negative predicted rate i.e.; accuracies against "Creditworthy" and "Non-creditworthy" I see that again forest model does a better job with Accuracy of "Creditworthy" at 0.79 and Accuracy of "Non-creditworthy" at 0.82. These values are better than other models, suggesting that the Forest model does a good job at classifying the classes correctly without too much bias.

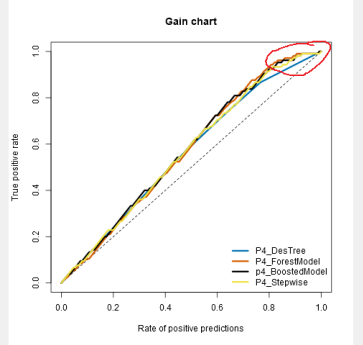
After examining the ROC curve, as shown below



As we examine the ROC curve, it is clear that ROC curve for Forest model has the highest true positive rate and the curve is close to the upper left corner. This means that the area under the curve for Forest model is greater than other models and hence it does a good job at actually classifying the two classes correctly.

For most of the given threshold values, the Forest model has higher True positive rate than other models as shown on the ROC curve.

Now looking at the gains charts below



If we examine closely, Forest model has a high True positive rate when compared to other models. All the models have very close true positive rate however, Forest model seems to have overall highest values the most times as well as this has reached the highest point quicker than other models.

- How many individuals are creditworthy?

Now that I concluded Forest model to be the best model for this business problem, I now score my new customer data set using the forest model object. As mentioned, I use the formula tool to set Score_Creditworthy greater than Score_NonCreditworthy, to label the person as “Creditworthy”. I generate a new column which shows whether the applicant from the “Customer to score” dataset are either “Creditworthy” or “Not creditworthy”. Based on the results I see that 415 customers are “creditworthy” and other 85 customers are “Not creditworthy”.