Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- **What decisions need to be made?**
  Determining whether the 500 new loan applicants are creditworthy for a loan.
- **What data is needed to inform those decisions?**
  - Account balance and Number of credits at our bank.
  - Personal information (age, employment status and length, number of dependents).
  - Credit Amount and payment status of previous credits.
  - Purpose of loan.
  - Value of savings stocks.

- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**
  A binary model, since we are looking for whether a new loan applicant is creditworthy or not.

Step 2: Building the Training Set

Guidelines for data cleanup:

- **For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least 0.70 to be considered "high".**
  There were **no high-correlation** between any of the numerical data fields with according to the following report (Pearson Correlation tool was used)

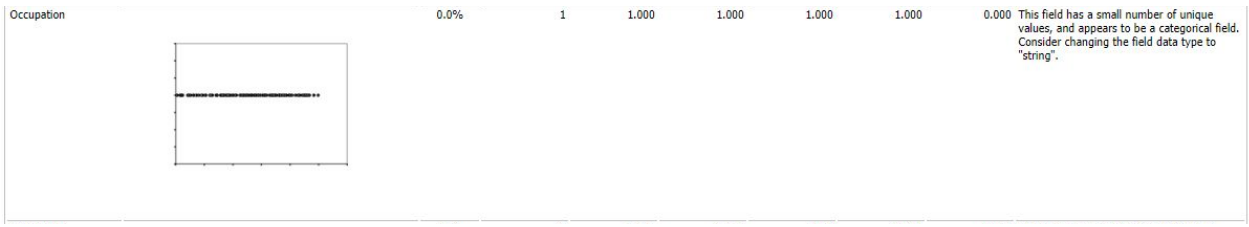
Fields | Cell Viewer

	FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Type-of-apartment	Age_years
1	Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	0.152516	-0.064197
2	Credit-Amount	0.57398	1	-0.288852	0.325545	0.170071	0.069316
3	Instalment-per-cent	0.068106	-0.288852	1	0.081493	0.074533	0.03927
4	Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.373101	0.086233
5	Type-of-apartment	0.152516	0.170071	0.074533	0.373101	1	0.32935
6	Age_years	-0.064197	0.069316	0.03927	0.086233	0.32935	1

- **Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed**
  Duration in current address field has been removed, since most of its data were missing.

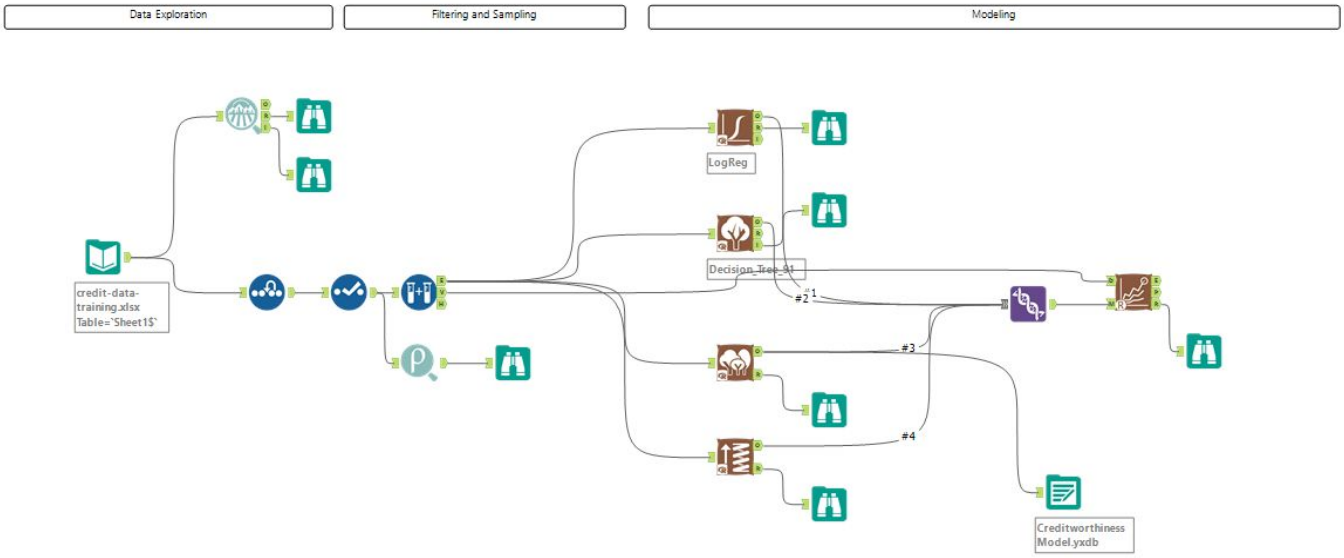
- **Are there only a few values in a subset of your data field? Does the data field look very uniform (low variability)**
  Low-variability field to be removed:
  - ➢ Occupation ~ Has only one value.
  - ➢ Concurrent credits ~ Has only one value.
  - ➢ Guarantors ~ the majority of data was skewed towards one value.
  - ➢ Foreign worker ~ the majority of data was skewed towards one value.
  - ➢ Number of dependents ~ the majority of data was skewed towards one value.
  - ➢ Also, I removed the Telephone field since it is irrelevant.

- **Average of *Age Years* should be 36 (rounded up)**
  Age years field was imputed *using the median* of values.



Step 3: Train your Classification Models

- Estimation and Validation samples where 70% Estimation and 30% Validation of entire dataset.
- **Models created:** 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.**

1- Logistic Regression:

- Account Balance
- Purpose
- Credit Amount

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Number of Fisher Scoring iterations: 5

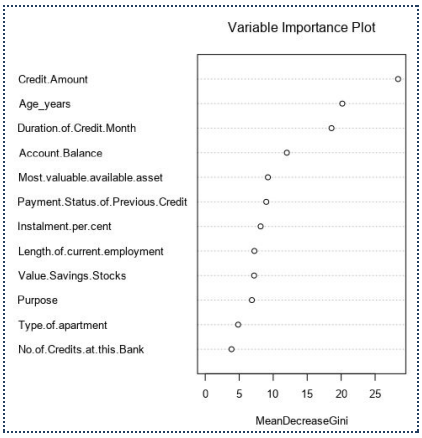
Type II Analysis of Deviance Tests

2- Decision Tree:

- Account Balance
- Duration of Credit per Month
- Credit Amount

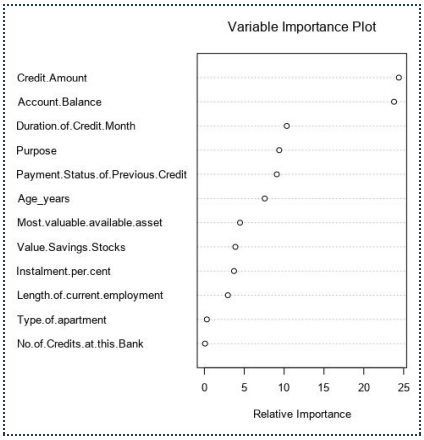
3- Forest Model

- Credit Amount
- Age Years
- Duration of Credit per month



4- Boosted Model

- Credit Amount
- Account Balance
- Duration of Credit per



month

- **Validate your model against the Validation set. What was the overall percent accuracy?**

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogReg	0.7800	0.8520	0.7314	0.9048	0.4889
DT	0.6667	0.7685	0.6272	0.7905	0.3778
Forest	0.8000	0.8707	0.7361	0.9619	0.4222
BM	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix of BM

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

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of Forest

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

Confusion matrix of LogReg

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Performance Diagnostic Plots

- **Forest model** accuracy was the highest at an overall 80%. **Boosted model** also performed well and it was the second best model with an overall accuracy of 78%.
- **Logistic regression** model accuracy is 78%, and has correctly predicted 90% of creditworthy and 48% non creditworthy.
- **Decision tree** model has the lowest accuracy among 4 models at 66%. it has predicted 79% of creditworthy records correctly.

- **Are there any bias seen in the model's predictions?**

Yes there is; since the non-creditworthy records within the data sample were lesser than the creditworthy ones, in other words, creditworthy represented the majority of the data sample we have, which caused the bias within the predicting models.

Step 4: Writeup

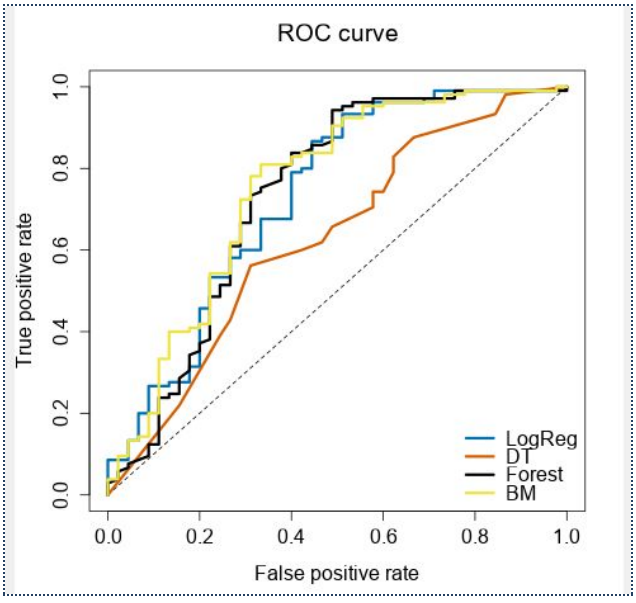
- **Which model did you choose to use? Please justify your decision using all of the following techniques. Please only use these techniques to justify your decision:**
  - **Overall Accuracy against your Validation set &**
  - **Accuracies within "Creditworthy" and "Non-Creditworthy" segments**

  As shown in the screenshots above:

  - **Forest model** accuracy was the highest at an overall 80%, 96% accuracy of predicting creditworthy records and 42% of non-creditworthy ones.
  - **Boosted model** also performed well and it was the second best model with an overall accuracy of 78%, 96% accuracy of predicting creditworthy records and 37% accuracy of predicting non-creditworthy ones.
  - **Logistic regression** model accuracy is 78%, and has correctly predicted 90% of creditworthy and 48% non creditworthy.
  - **Decision tree** model has the lowest accuracy among 4 models at 66%. it has predicted 79% of creditworthy records correctly. The accuracy of predicting non-creditworthy records was 37%.

- **ROC graph**

  According to the graph the forest model and the boosted model(BM) were close in prediction accuracy and the best in performance , while the decision tree (DT) was the worst of all 4 models.



- **Bias in the Confusion Matrices(Predicted_Creditworthy vs. Predicted_Non-Creditworthy)**

  Logistic regression & decision tree predict more non-creditworthy values than the **forest and boosted models,** and the majority of predicted values of these two were for creditworthy.

- **How many individuals are creditworthy?**
  After using the generated forest model with new customers to score data, the sum of creditworthy new applicants is 406.