Project 3: Credit Scoring

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.

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days.

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

- What decisions needs to be made? There are 500 new loan applications which needs to be assessed whether approved or rejected for credit.
- What data is needed to inform those decisions? Historical data of loan applications and the list of new loan applications.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions? Binary model, yes or no for credit worthiness.

Step 2: Building the Training Set

The following 7 fields have been removed from the dataset, so only 13 remained from the original 20:

Either uniform or missing data (red boxes below)

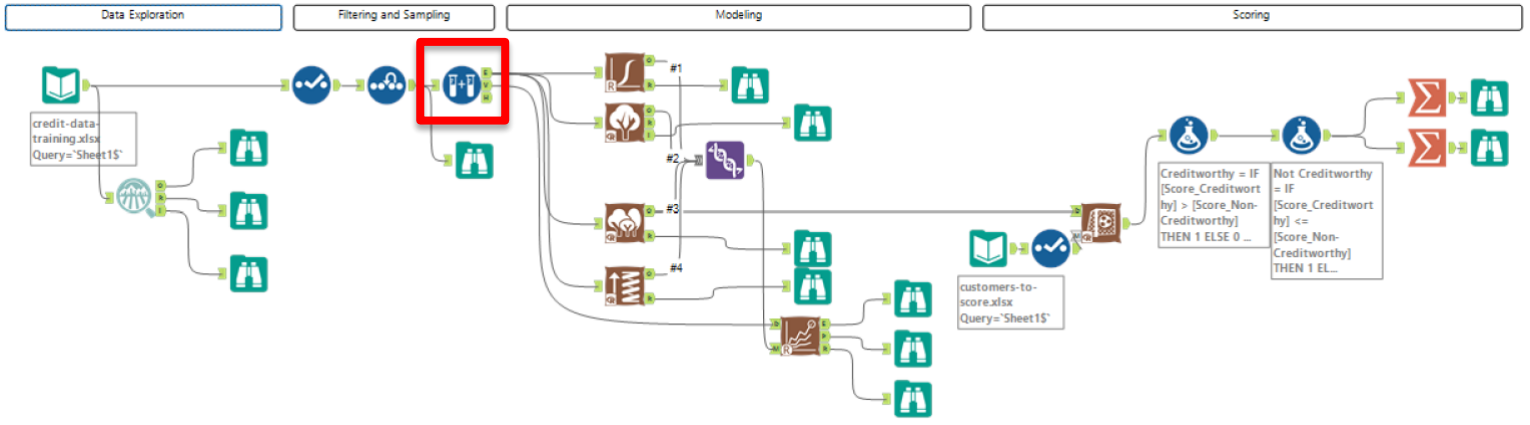
Concurrent-Credits – totally uniform dataset.
Occupation – totally uniform dataset.
Duration-in-Current-Address - 69% of dataset is missing (see green-red bar – red means missing).

Almost uniform dataset (yellow boxes below)

Guarantors – majority of the data is the same
Foreign worker – majority of the data is the same
No-of-dependents – majority of the data is the same

No relation to credit rating: Telephone

Step 3: Training Classification Models



In filtering and sampling stage the model samples (see red box above) 70% of the dataset is for Estimation and 30% of the dataset is reserved for Validation, random Seed is 1.

- Which predictor variables are significant or the most important?

*Logistic Regression – the most significant (*** level of significance) is Account.Balance*

Coefficients:

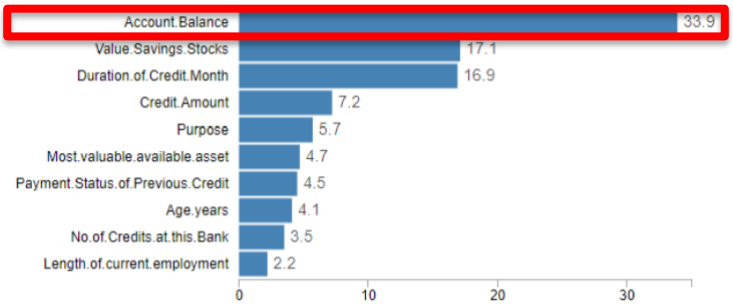
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome 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	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
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	

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

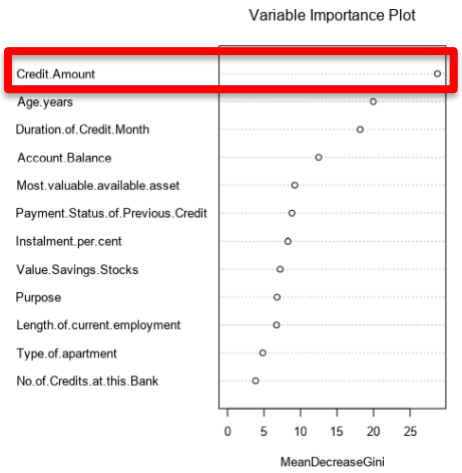
(Dispersion parameter for binomial taken to be 1)

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

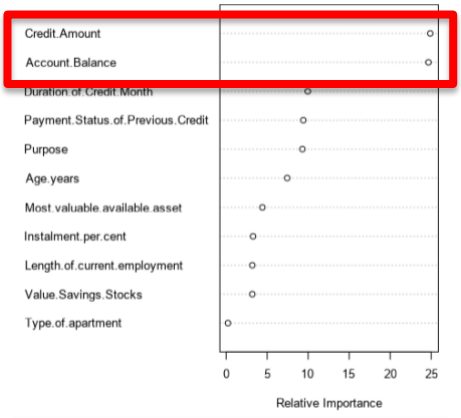
Decision Tree – Variable importance – the most significant predictor is Account.Balance



Forest Model – Variable importance – the most significant predictor is Credit.Amount



Boosted Model – Variable importance – the most significant predictor is Credit.Amount and Account.Balance



- 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?
  For each model's accuracy, see the red box below:

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7490	0.9619	0.3778
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889

Confusion matrix: by looking at the confusion matrix Decision tree has some bias predicting Non-creditworthy applications, as False Negative (12) is almost the same as True Negative (19), therefore the predictions has low confidence. This is also confirmed by the low 61.3% accuracy for Non-Creditworthy cases with the yellow highlight below,. All other models have lower False Negative and False Positive numbers (things we want to minimize)

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Forest_Model

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

Confusion matrix of Logistic_Regression

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

1. Logistic Regression

Total accuracy: (95+22)((95+10+22+23) = 78%
Creditworthy accuracy: 95/(95+23) = 80.5%
Non-Creditworthy accuracy: 22/(10+22) = 68.75%

2. Decision Tree

Total accuracy: (93+19)((93+12+19+26) = 74.66%
Creditworthy accuracy: 93/(93+26) = 78.15%
Non-Creditworthy accuracy: 19/(12+19) = 61.3%

3. Forest Model

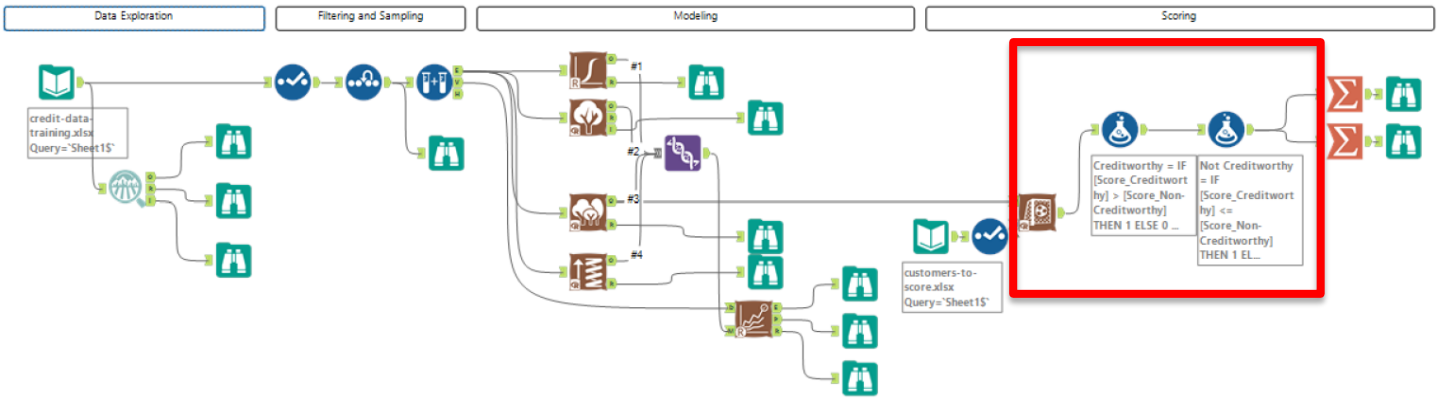
Total accuracy: (102+17)((102+3+17+28) = 79.33%
Creditworthy accuracy: 102/(102+28) = 78.46%
Non-Creditworthy accuracy: 17/(3+17) = 85%

4. Boosted model

Total accuracy: (101+17)((101+4+17+28) = 78.66%
Creditworthy accuracy: 101/(101+28) = 78.3%
Non-Creditworthy accuracy: 17/(4+17) = 90.95%

Step 4: Writeup

If Score_Creditworthy is greater than Score_NonCreditworthy, the person is labeled as “Creditworthy”. The implented scoring logic can be seen in the red box below.

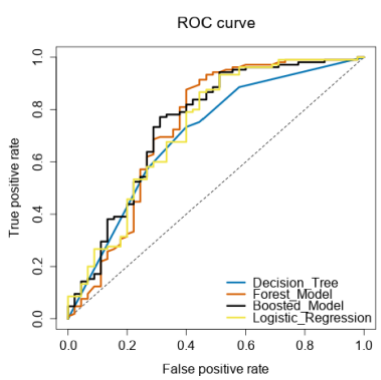


- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set (see red box below) and Accuracies within “Creditworthy” and “Non-Creditworthy” segments (see yellow box below): based on these results, Forest Model has the highest overall accuracy 79.33%, and it is the best among the Creditworthy accuracy and on par with the others in Non-Creditworthy accuracy. For actual hand calculations see previous page, as yellow box does not reflect the correct numbers, while the red box does.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7490	0.9619	0.3778
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889

    - ROC graph: the Decision Tree has the lowest performance, while Boosted model is the best in lower False positive rate regions, the Forest model is the best in higher false positive rates.



    - Bias in the Confusion Matrices for the chosen Forest Model

Confusion matrix of Forest_Model

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

Total accuracy: (102+17)((102+3+17+28) = 79.33%
Creditworthy accuracy: 102/(102+28) = 78.46%
Non-Creditworthy accuracy: 17/(3+17) = 85%

Both the overall and the Creditworthy/Non-Creditworthy accuracy is in the 80% region, which is quite high, so there are no bias in the model.

- How many individuals are creditworthy? There are **<u>408</u>** creditworthy cases based on the Forest Model.