Project 4: Predicting Default Risk

The Business Problem

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.

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

Key Decisions:

1. **What decisions needs to be made?**

We need to determine whether the new customers based on the data provided are creditworthy for a loan.

2. **What data is needed to inform those decisions?**

We need to know :

- ✓ Their current loan request (amount, tenor, purpose)
- ✓ Their past credit track record
- ✓ Their revenue profile, including employment status, length, salary
- ✓ Their expense profile, including average monthly, number of children in charge or dependent person
- ✓ Their asset profile, including deposit balance in the bank, numbers of properties,
- ✓ Their liability profile, including current debt outstanding
- ✓ Whether they can provide a guarantor for the loan

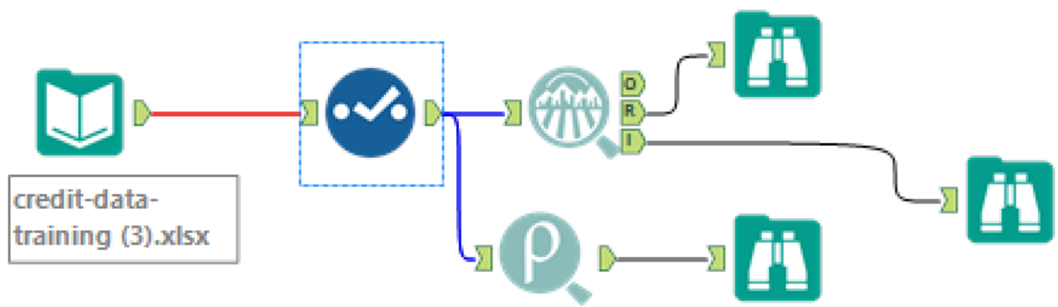
3. **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

Since we are trying to determine the creditworthiness of a loan applicant, the problem would involve a binary model.

Step 2: Building the Training Set

Build your training set given the data provided to you. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

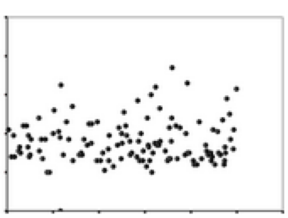
To build the training set, we started to visualize the data through the field summary and pearson correlation functions.



Then we decided to remove the following field for the following reason with illustration below

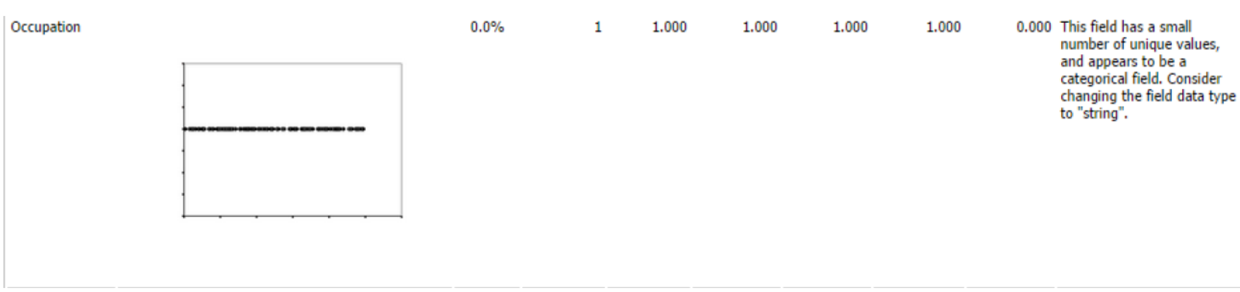
Removed field	Reason
Duration in current address	Many missing data
Concurrent credit	Low variability, only "other banks/dept"
Guarantors	Low variability
Occupation	Low variability, only "1"
No of dependents	Low variability
Telephone	Not relevant data
Foreign worker	Low variability

We also decided to impute age-years for the few missing data to the median

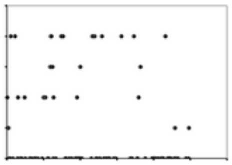
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	

Field summary showing low variability for concurrent credit, foreign worker, nb of dependent, guarantor

Field summary showing low variability for occupation



Field summary showing 68.8% of missing data for duration in current address

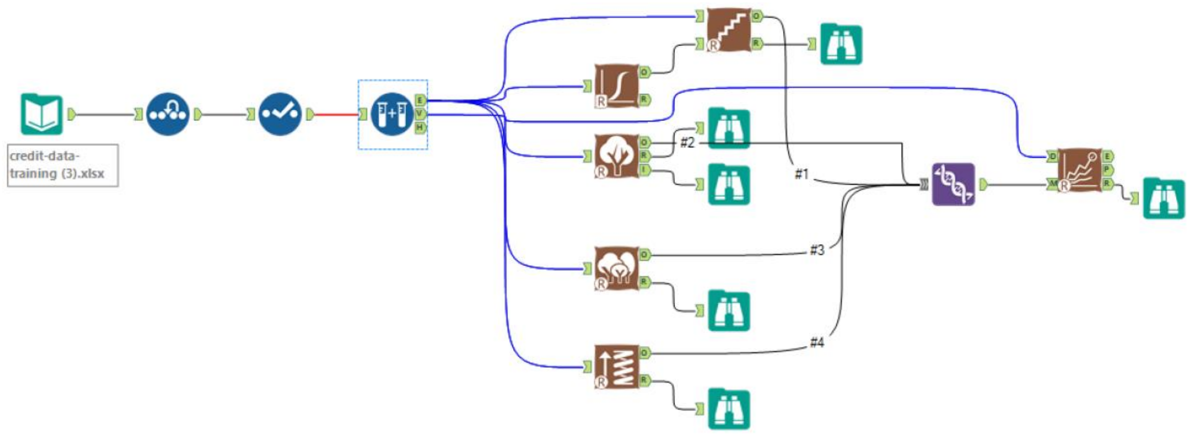
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

However the pearson correlation table does not show high correlation between data (>70%)

FieldName	Duration-...	Credit-Amount	Instalm...	Durati...	Most-valua...	Age-years	Type-of-apa...	Occupation	No-of-de...	Telephone	Foreign-Worker
Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269	0.143176	-0.115916
Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986	0.286338	0.025493
Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894	0.029354	-0.133411
Duration-in-Current-address	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]
Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454	0.203509	-0.146005
Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738	0.101443	-0.089848
Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1	-0.048559	0.065943
Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559	1	-0.055516
Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943	-0.055516	1

Step 3: Train your Classification Models

We create an Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. And we create the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model



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

Below is a summary table of the significant predictor variable under the difference model.

Model type	Significant predictor variables
Logistic Regression	Account balance Payment status of previous credit Purpose Credit amount Length of current employment Instalment per cents
Decision Tree	Account Balance Duration of the credit in months Value savings stocks
Forest Model	Credit amount Age in years Duration of the credit in months Account Balance
Boosted Model	Credit amount Account balance Duration of the credit in months Payment status of previous credit

P-values table for the Logistic Regression Model

Report

Report for Logistic Regression Model Logmodel

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)

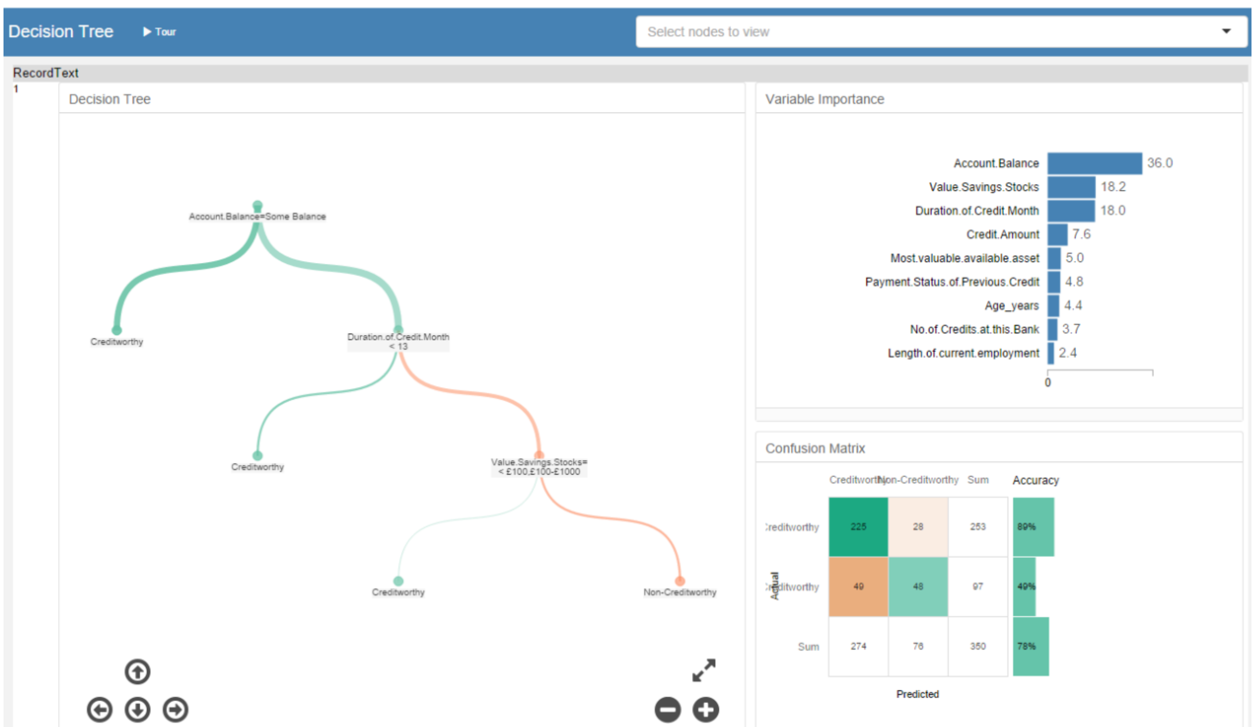
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

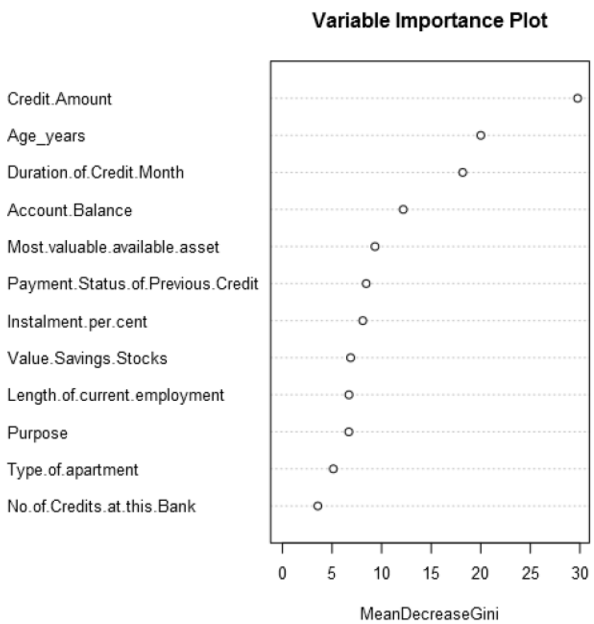
Coefficients:

	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 .

Decision tree summary



Variable importance plot for the Forest model



Variable importance plot for the Boosted model

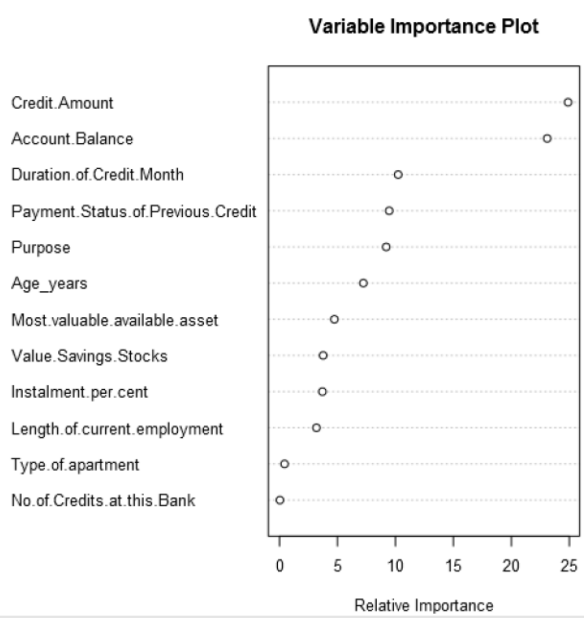
Report

Report for Boosted Model Boosted

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3940

Plots:



2. **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?**

We used the model comparison function to compare the different models showing their respective overall accuracy and confusion matrix.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logmodel	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted	0.7933	0.8670	0.7509	0.7891	0.8182

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted

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

Confusion matrix of Decision_tree

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

Confusion matrix of Forest

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

Confusion matrix of Logmodel

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

For the logistic regression, the overall accuracy is 76%. Though the accuracy to predict creditworthiness is quite good at 80%, however the accuracy to predict non-creditworthiness is quite low at 62.86%.

The situation is quite similar for the decision tree with an overall accuracy of 74.67% and good accuracy to predict creditworthiness at 79.13% but the accuracy to predict non-creditworthiness is quite low at 60%. The confusion matrix of the decision tree summary above shows also a very low accuracy for prediction of non-creditworthy customers.

There is indeed a bias induced by less sample of non-creditworthy clients. In fact, in our training data only 28.4% of the total customers are tagged non-creditworthy.

However, the forest and boosted model seems to perform better than the logistic regression and decision tree model. Their overall accuracies are 80% and 79.33% respectively and their accuracy rates to predict non-creditworthy customers are even higher than accuracy rates to predict creditworthy customers at 82.61% and 81.62% respectively.

Step 4: Writeup

1. Which model did you choose to use? Please justify your decision using only the following techniques:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logmodel	0.7600	0.8364	0.7306	0.8000	0.6286
Decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted	0.7933	0.8670	0.7509	0.7891	0.8182

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

a. Overall Accuracy against the Validation set

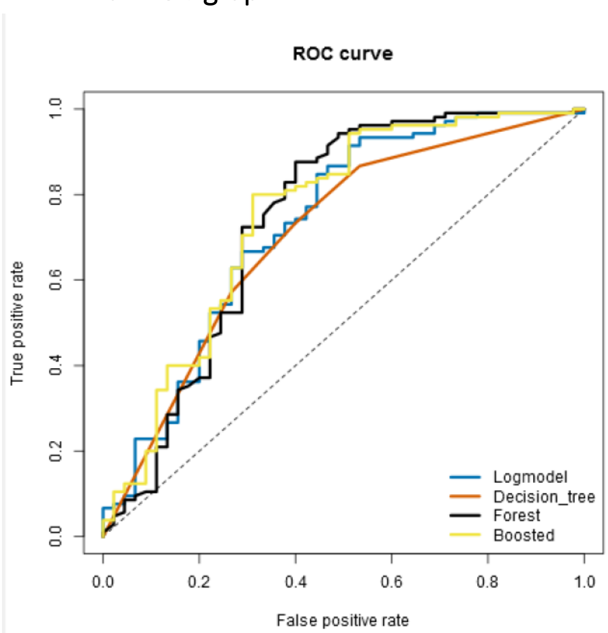
Based on the model comparison report, it appears that the Forest Tree has the highest accuracy rate with 80% compared to other models.

b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments

Within "Creditworthy" and "Non-Creditworthy" segments, the logistic model appears to have the highest accuracy to predict "Creditworthy" however the accuracy for "Non-creditworthy" is quite low at 62.86%.

The Forest Tree model has again the highest accuracy for both "creditworthy" and "non-creditworthy" segments.

c. ROC graph



When comparing the ROC curve of the four models, we see that the decision tree seems to perform the worst, while the forest tree model and boosted model seem to perform the best. The model comparison report shows in fact a higher AUC for the boosted model at 0.7509.

d. Bias in the Confusion Matrices

Confusion matrix of Boosted

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

Confusion matrix of Decision_tree

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

Confusion matrix of Forest

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

Confusion matrix of Logmodel

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

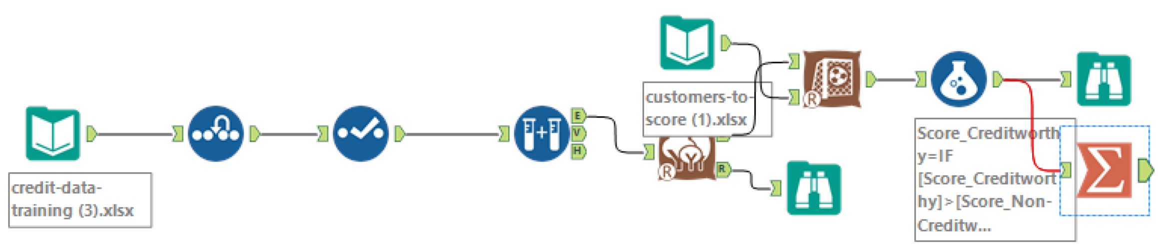
From the confusion matrices, we see that the boosted model and forest model tend to classify more non-creditworthy customers as creditworthy while the decision tree model and logistic model tend to classify creditworthy customer as non-creditworthy.

However since the boss only care for prediction accuracy, we chose the forest model which has the highest accuracy overall.

2. How many individuals are creditworthy?

Finally, we use the score tool to predict the creditworthiness of the new customers.

The model predicts that 415 customers out of 500 are creditworthy.



Record #	Sum_Score_Creditworthy
1	415