Project: Creditworthiness

Step 1: Business and Data Understanding

The decisions should be made are completely related and should be support the main decision of deciding which applicants are creditworthy to give loan. There are relevant key decisions should be made from point of the analyst like decide to clean or reformat the data and choose the model according to the needed result.

Key Decisions:

- What decisions needs to be made?
  The key decision that need to be made are 1- which applicants are creditworthy to give them loans 2- is a binary case or non-binary 3- Which predictable models to choose 4- Which target and predictor variables to choose to support point 1.

- What data is needed to inform those decisions?
  The data needed are all about applicant related information like ages, years of employment, job contract type, balances, credits status, amounts, credit ratings and other relevant data.

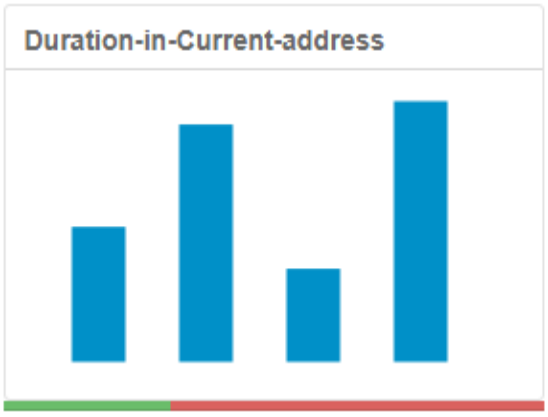
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary classification model since we want to decide either if the applicant is a creditworthy or not.

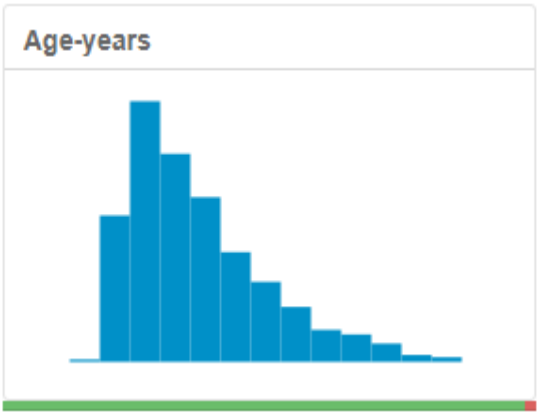
Step 2: Building the Training Set

There is no high correlation between any data fields as shown in the following table:

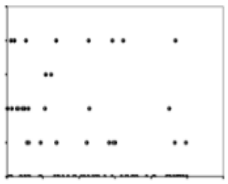
FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years
Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	-0.064197
Credit-Amount	0.57398	1	-0.288852	0.325545	0.069316
Instalment-per-cent	0.068106	-0.288852	1	0.081493	0.03927
Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.086233
Age-years	-0.064197	0.069316	0.03927	0.086233	1

There are data missing values for fields of **Duration-in-Current-address** and **Age-years** as shown in the following photos (missing values are visualized with red color of lines below each chart):

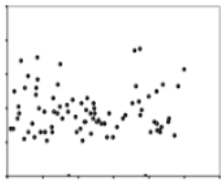




The **Duration-in-Current-address** field is removed because the missing values data percentage is high, it is 68.8%

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".

The **Age-years** missing values percentage is 2.4% and they are imputed by the median of 33 since it is a right skewed distribution

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	

- The following data fields are low variability fields:
  1 - Concurrent-Credits (entirely uniformed only one variable)
  2 - Guarantors (highly skewed for one variable)
  2 - Foreign-Worker (highly skewed for one variable)
  3 - No-of-dependents (highly skewed for one variable)
  4 - Occupation (entirely uniformed)
  Note: Telephone field is removed because it is not a relevant

Step 3: Train your Classification Models

Most Important Variables for each model:

Model	Most Important Variables
Logistic Regression	Account.BalanceSome Balance Payment.Status.of.Previous.CreditSome Problems PurposeNew car Credit.Amount Instalment.per.cent Length.of.current.employment< 1yr
Forest Model	Credit.Amount Age.years Account.Balance Duration.of.Credit.Month
Decision Tree	Account.Balance Duration.of.Credit.Month Credit.Amount
Boosted Model	Account.Balance Credit.Amount

Logistic Regression

p-values for the model

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + No.of.Credits.at.this.Bank + Age.years, family = binomial(logit), data = the.data)

Deviance Residuals:

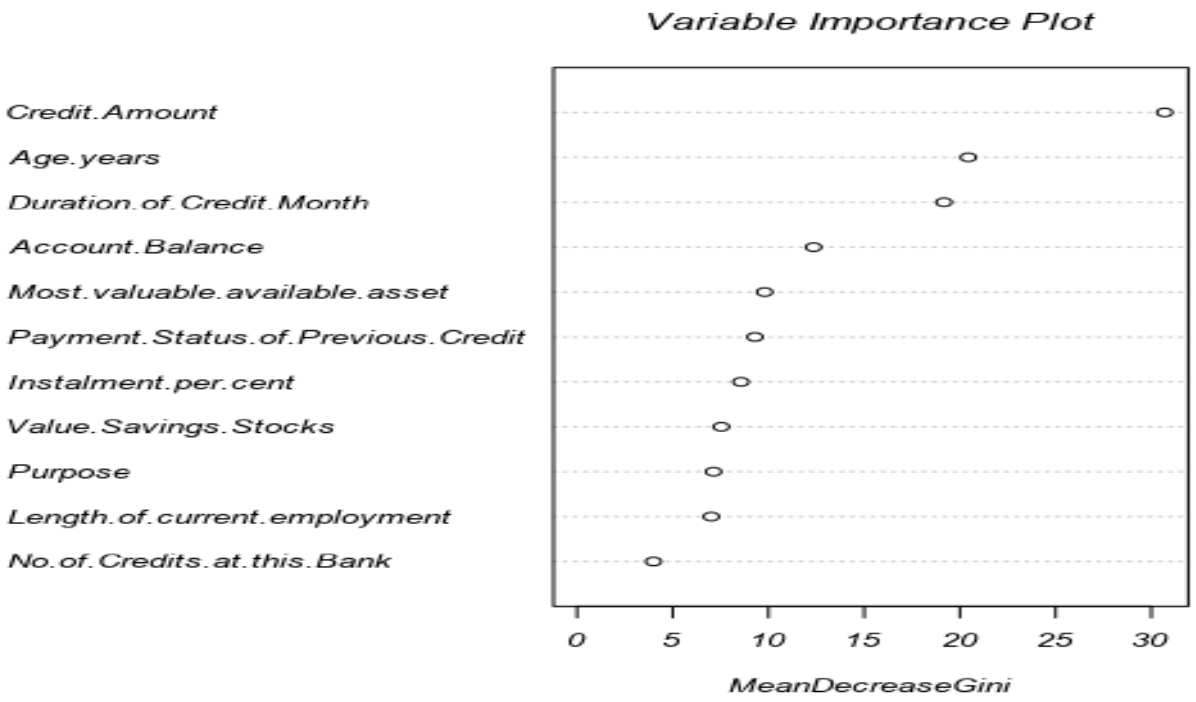
Min	1Q	Median	3Q	Max
-2.064	-0.721	-0.421	0.736	2.473

Coefficients:

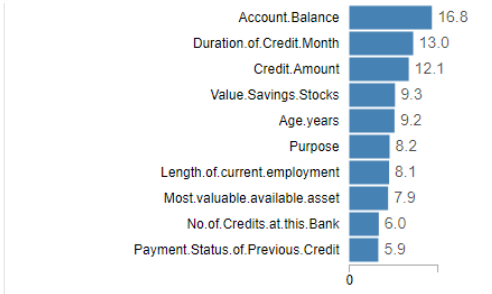
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.2290394	9.845e-01	-3.2800	0.00104 **
Account.BalanceSome Balance	-1.5843791	3.200e-01	-4.9511	7.38e-07 ***
Duration.of.Credit.Month	0.0058321	1.365e-02	0.4272	0.6692
Payment.Status.of.Previous.CreditPaid Up	0.4306851	3.847e-01	1.1195	0.26294
Payment.Status.of.Previous.CreditSome Problems	1.2872278	5.339e-01	2.4109	0.01591 *
PurposeNew car	-1.7472435	6.271e-01	-2.7862	0.00533 **
PurposeOther	-0.2780516	8.305e-01	-0.3348	0.73778
PurposeUsed car	-0.7651003	4.108e-01	-1.8624	0.06255 .
Credit.Amount	0.0001734	6.833e-05	2.5375	0.01116 *
Value.Savings.StocksNone	0.5996934	5.065e-01	1.1840	0.2364
Value.Savings.Stocks£100-£1000	0.1818563	5.621e-01	0.3236	0.74628
Length.of.current.employment4-7 yrs	0.5259720	4.934e-01	1.0660	0.28642
Length.of.current.employment< 1yr	0.7776684	3.951e-01	1.9681	0.04906 *
Instalment.per.cent	0.2969774	1.384e-01	2.1457	0.0319 *
Most.valuable.available.asset	0.2877408	1.488e-01	1.9337	0.05315 .
No.of.Credits.at.this.BankMore than 1	0.3918288	3.812e-01	1.0280	0.30397
Age.years	-0.0180861	1.475e-02	-1.2259	0.22022

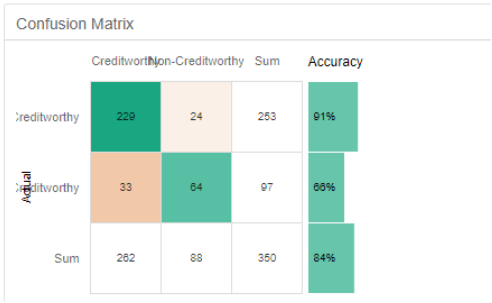
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Forest Model

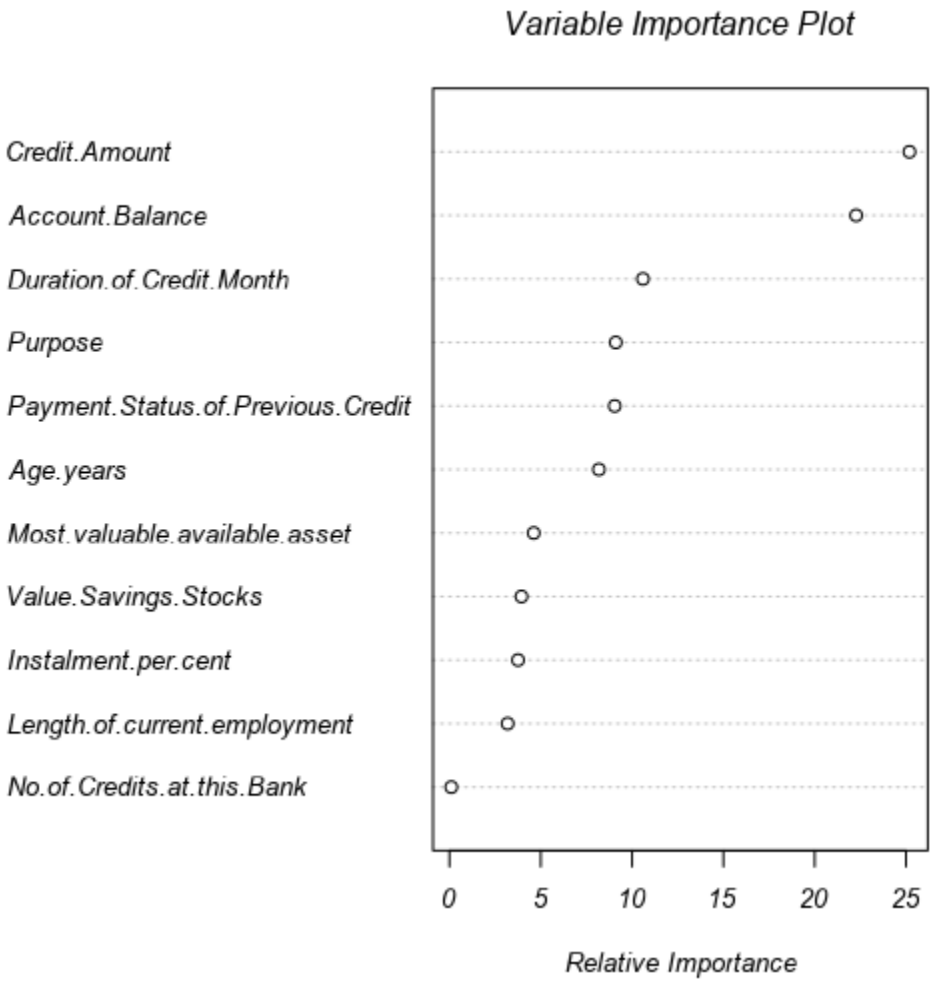


Decision Tree





Boosted Model



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogisticModel	0.7600	0.8364	0.7306	0.8762	0.4889
ForestModel	0.8200	0.8841	0.7414	0.9810	0.4444
DecisionTree	0.6667	0.7685	0.6272	0.7905	0.3778
BoostedModel	0.7800	0.8584	0.7524	0.9524	0.3778

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BoostedModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of DecisionTree

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

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	25
Predicted_Non-Creditworthy	2	20

Confusion matrix of LogisticModel

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

According to the report of the comparison tool, the highest accuracy goes for the Forest Model with 82%. Its accuracy for creditworthy is 98% and for non-creditworthy is 44%.

The second model comes after the Forest Model is the Boosted Model with 78% accuracy. It is creditworthy accuracy is 95% and for non-creditworthy is 37%.

The third model in its accuracy is the logistic Regression by 76%. It is creditworthy accuracy is 87% and non-creditworthy is 48%.

The lowest model in its accuracy is the Decision Tree Model with 66%. It is creditworthy accuracy is 79% and non-creditworthy accuracy is 37%

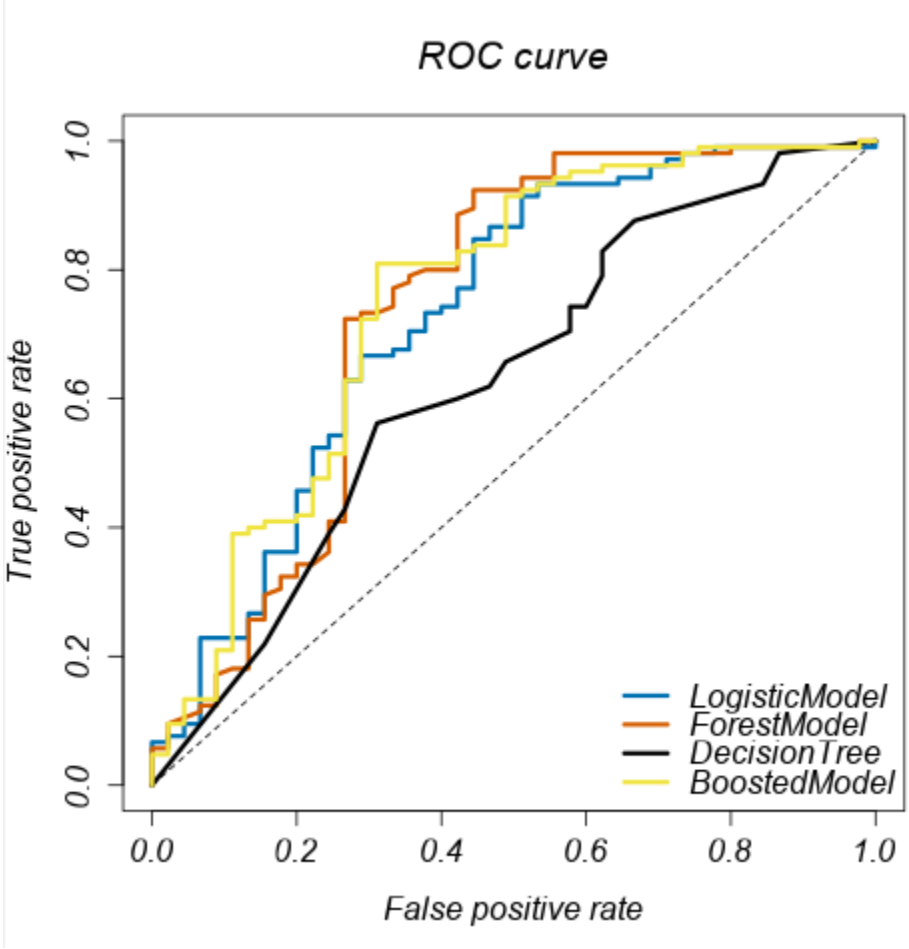
I will use the Forest Model since it has the highest accuracy between all used models as explained above

Step 4: Writeup

The model has been chosen after many steps starting from reformatting and imputing some values of the data. Also, deleting those variables with low variability and not relevant variables. Choose the fourth models and apply them on the data. Validate them and choose best one according to highest accuracy.

Answer these questions:

- ROC graph



- Bias in the Confusion Matrices

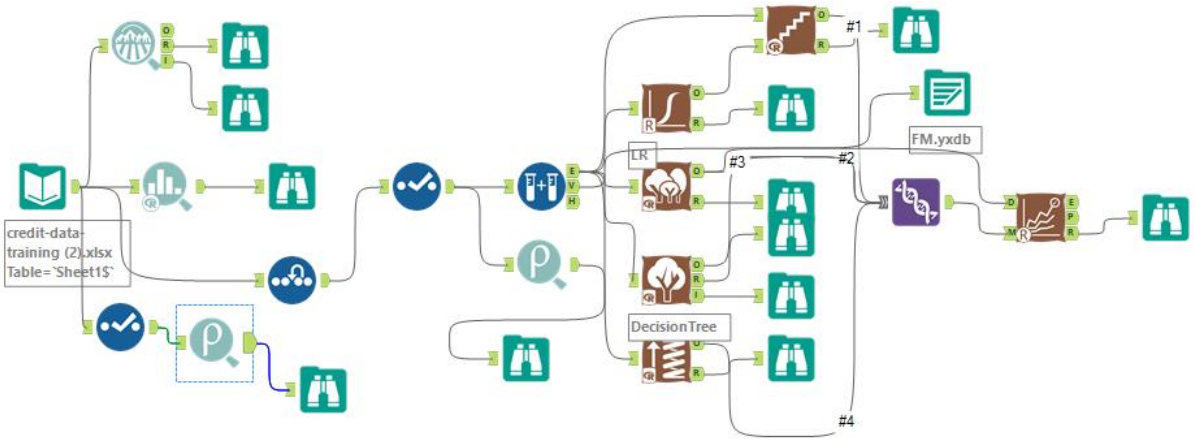
  There is bias in the confusion Metrex since the highest results are on the predicted creditworthy for the actual credit worthy.

<u>Percentages of creditworthy and non-creditworthy segments are shown above below the Model Comparison Report</u>

Credit worthy individuals are 409 as shown

Record	Sum_X_Creditworthy
1	409

The workflow of the all predictable model:



The workflow of the Forest Model that predicted 409 as number of creditworthy: