Project: Creditworthiness

Step 1: Business and Data Understanding

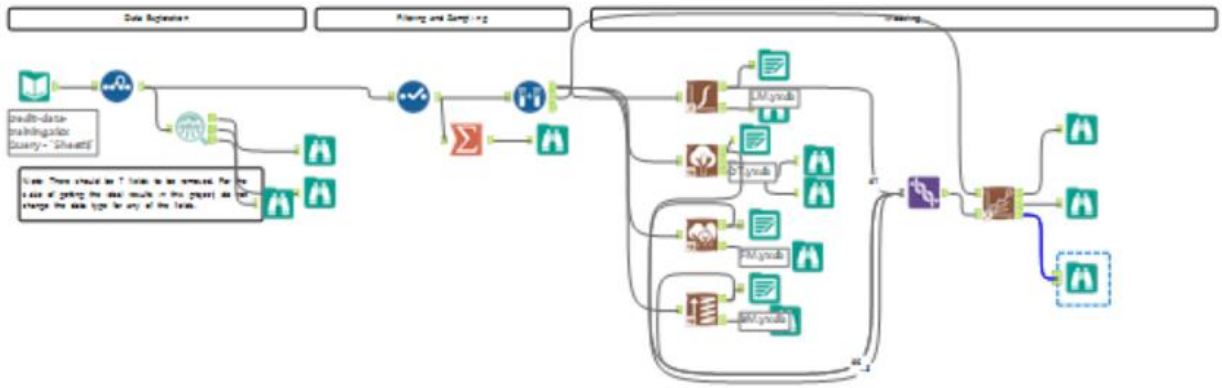
Key Decisions:

Answer these questions

- What decisions needs to be made?
    1. Selecting the best model/process to determine the creditworthiness of a customer for loan approval
    2. To determine the creditworthiness of the new 500 customers using the selected model
- What data is needed to inform those decisions?
    1. Data of past loan approvals for training the model
    2. Data of new customers for which creditworthiness is to be classified.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

    The Target variable (Creditworthy) is a binary categorical variable. Thus, a binary classification model should be used.

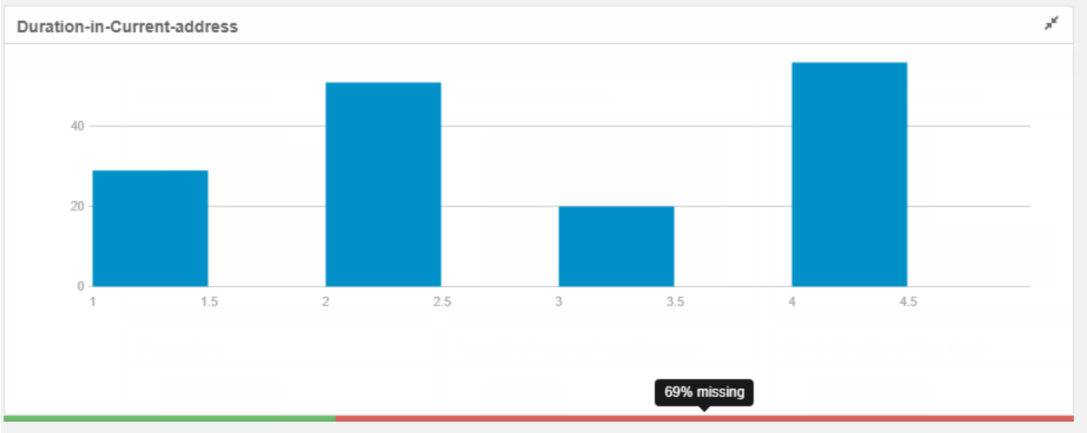
Step 2: Building the Training Set



Answer this question:

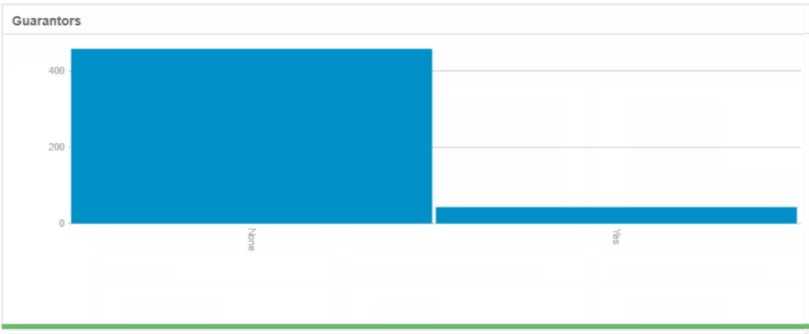
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

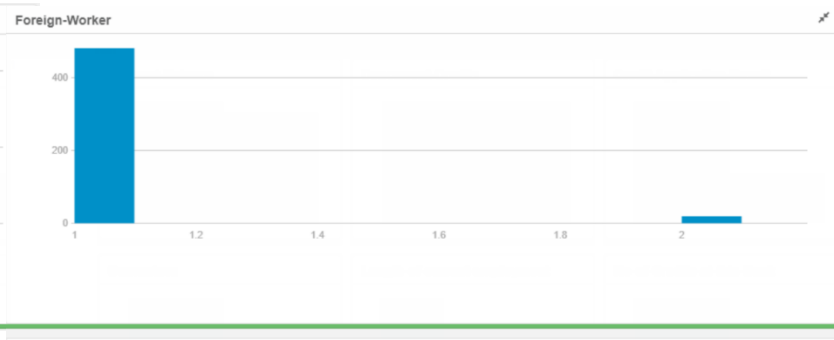
1. Duration in Current Address: This field was **removed** since it has mostly (≈70%) missing values

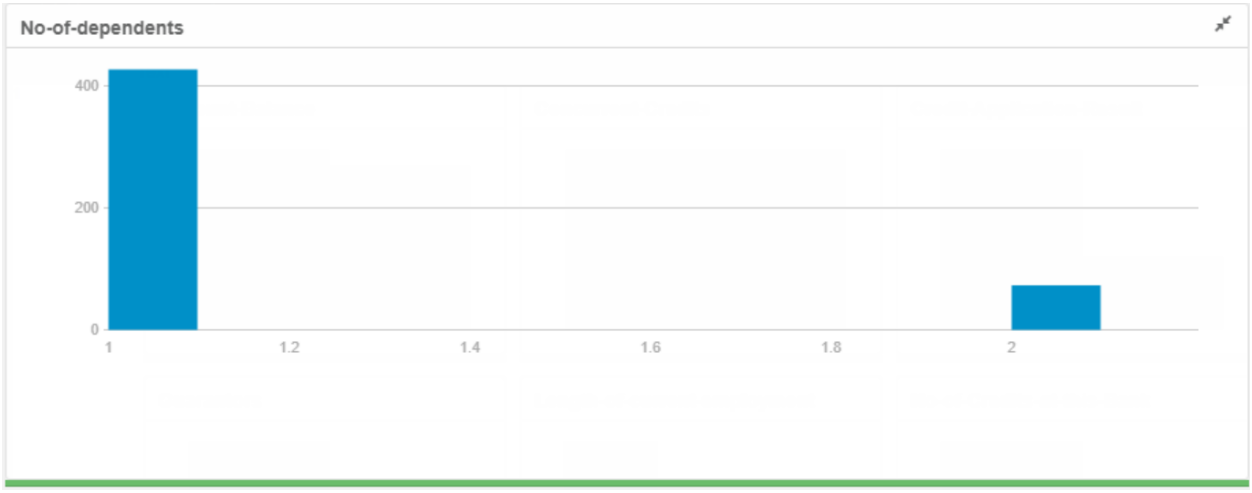


2. Concurrent Credits: The field was **removed** since it is a 100% uniform model with only 'Other Banks/Debts' value in all records.

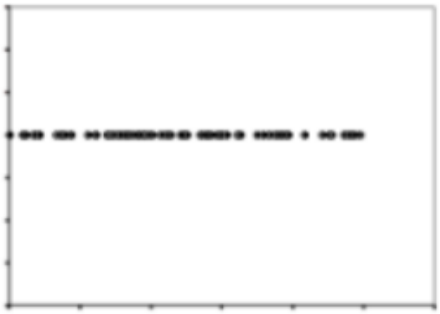
3. Guarantors, Foreign Worker, No of Dependents, Occupation, Telephone: These fields have low-variability in their values due to domination of one value, hence **removed**

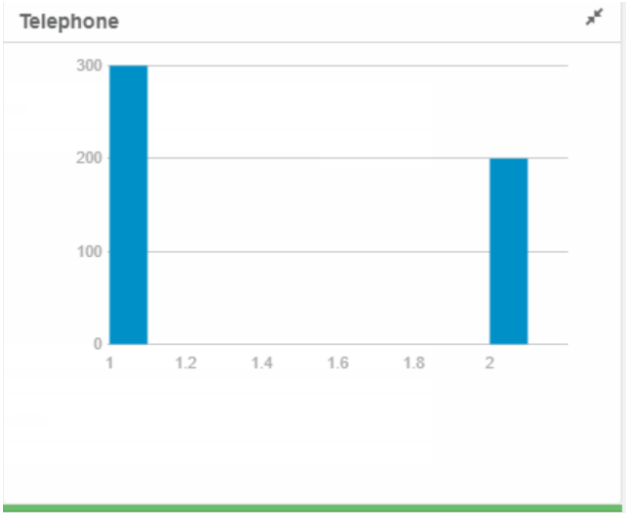




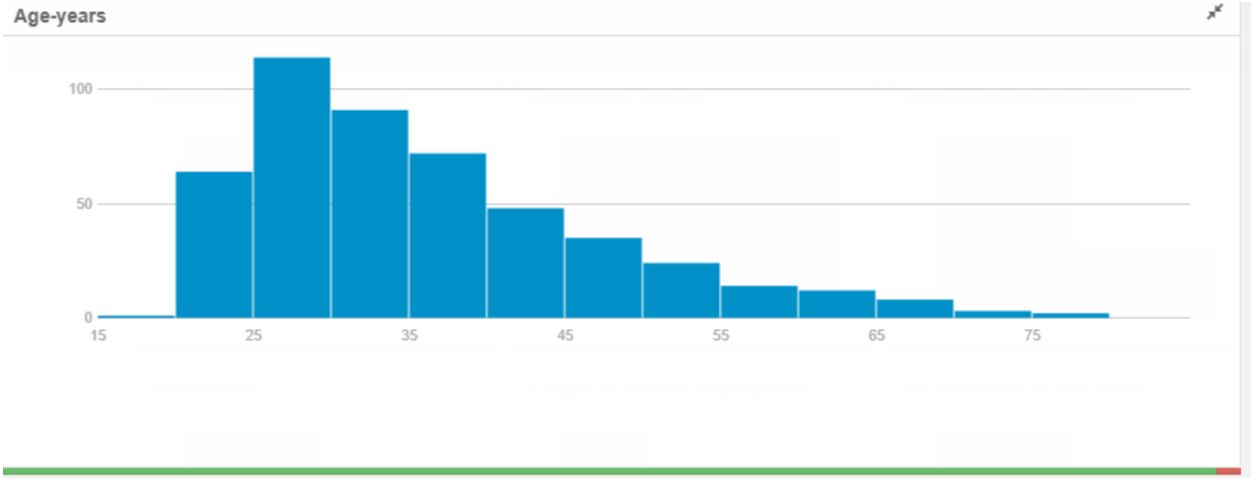


Occupation





4. Age-years: This field has some missing values and these null values are **imputed** with Median value because there might be outliers in the age field, also averaging would give a continuous value like 35.5 but age is distinct number and cannot be integers



Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

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

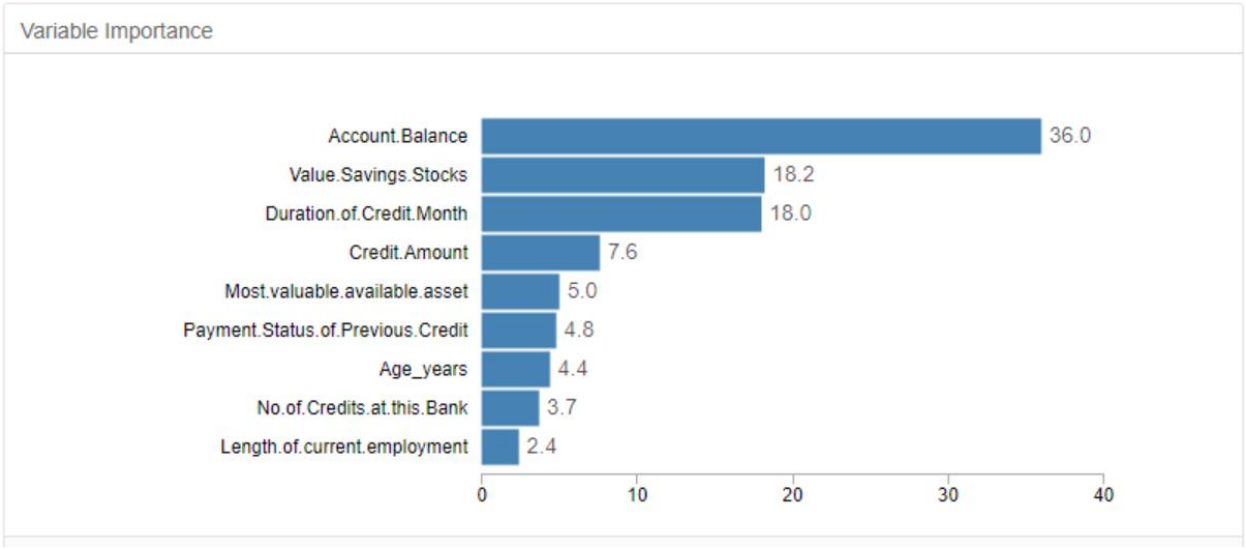
1. Logistic regression:

   Most Significant Predictor Variables are (in descending order): Account Balance, Purpose, Credit Amount, Length of Current Employment, Most Valuable Available Asset, Payment status of previous Credit

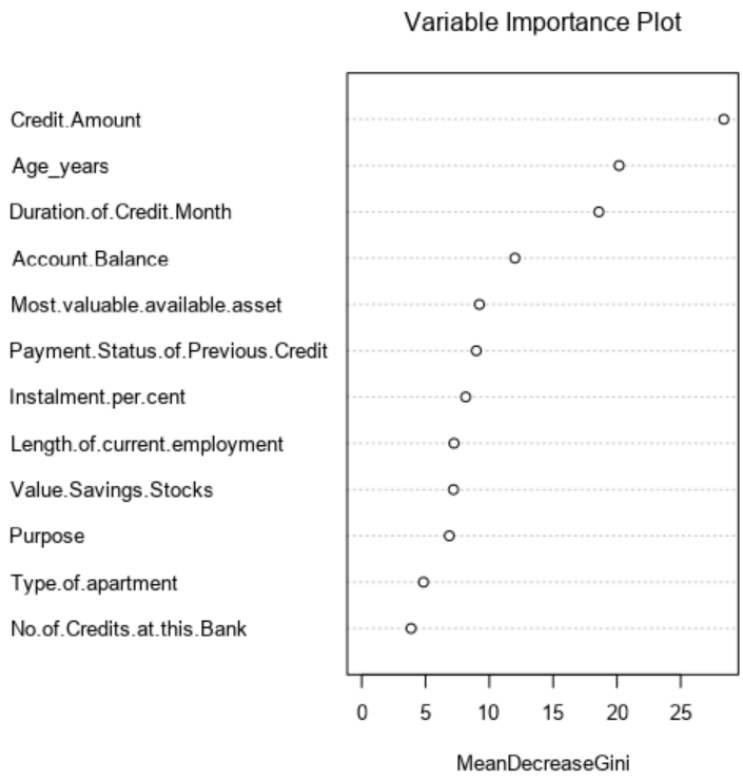
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome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

2. Decision Tree:

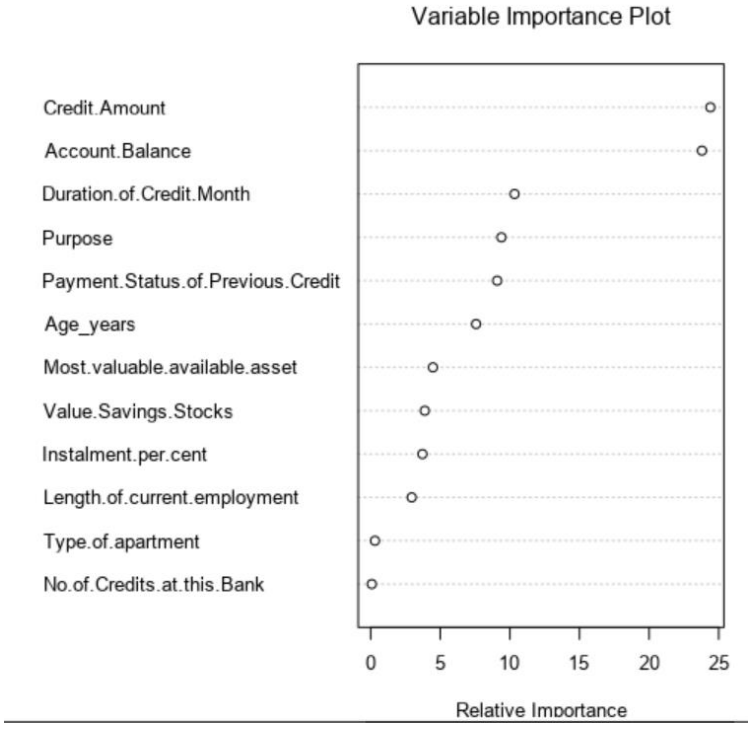
   Account Balance, Value Savings stocks, Duration of Credit Month, Credit Amount, Most Valuable Available Asset, Payment status of previous Credit



3. Forest Model: Credit Amount, Age, Duration of Credit Month, Account Balance, Most Valuable Available asset



4. Boosted Model: Credit Amount, Account Balance, Duration of Credit, Purpose, Previous status of Previous Credit



- 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?

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Log_reg	0.7800	0.8520	0.7314	0.9048	0.4889
DT	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_mod	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	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Forest_mod

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

Confusion matrix of Log_reg

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

BM- Boosted Model; DT- Decision Tree; Forest_mod-Forest Model; Log-reg-Logarithmic Regression

Yes, there is a bias in the model. The accuracy of predicting the Creditworthy is greater (in all models) than predicting Non-creditworthy. The reason maybe due to the domination of the creditworthy values in the field. 358/500 are creditworthy customers.

Step 4: Writeup

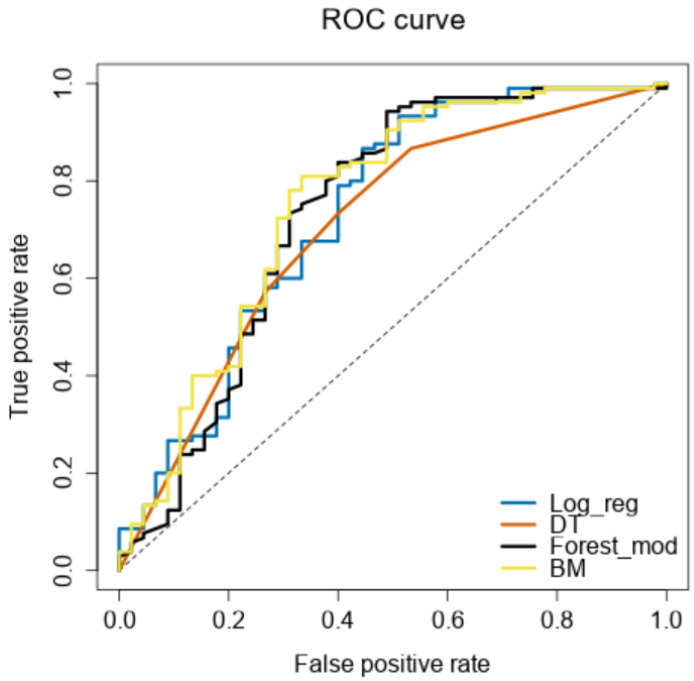
Answer these questions:

- 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

I have chosen the **Forest Model** to carry out further analysis because:

1. It has the highest overall accuracy against the validation set.

2. It has maximum ‘Creditworthy’ accuracy, hence better than Log Regression and Decision Tree Models. Also, it has ‘Non-creditworthy’ accuracy higher than that of Boosted Model.
3. From the ROC plot we see that the black line representing Forest model is reaching the high point soon. It is more inclined to the top side than the other curves, the next close one is Boosted Model



4. From the confusion matrices, we see that the forest model has wrongly predicted only 4 ‘Creditworthy’ customers as ‘Non-creditworthy’ as compared to 4,14,10 with boosted, Decision tree and Logarithmic Regression models respectively.

- How many individuals are creditworthy?
  408 out of the 500 individuals are creditworthy