Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- **What decisions needs to be made?**

Determining if customers are creditworthy to give a loan to

- **What data is needed to inform those decisions?**

- Data on all past applications

- The list of customers that need to be processed in the next few days

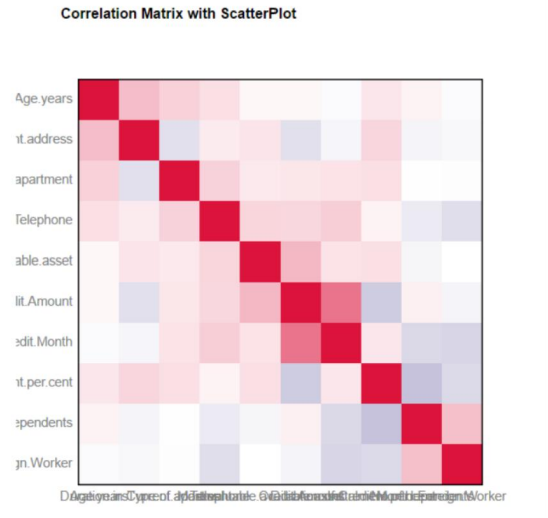
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

Binary classification model such as Logistic Regression, Decision Tree, Forest Model and Boosted Tree needed to use for making the decision.

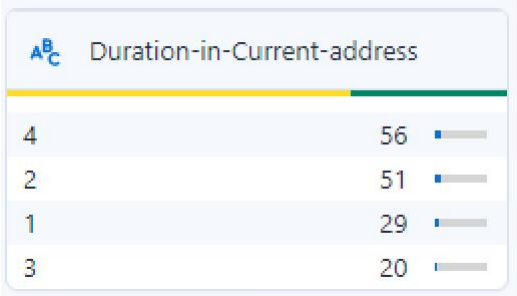
Step 2: Building the Training Set

- **For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".**

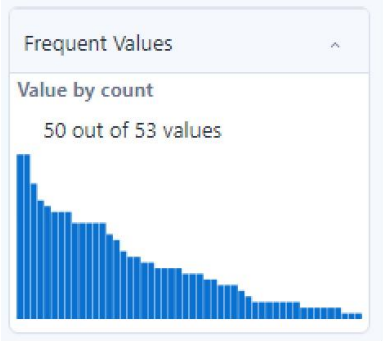
There are no fields which are highly correlated with each other.



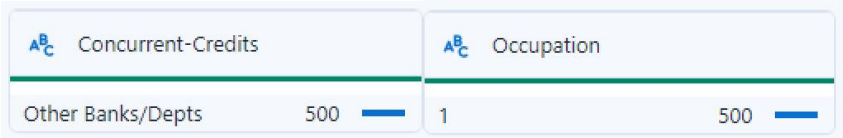
- **Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed.**

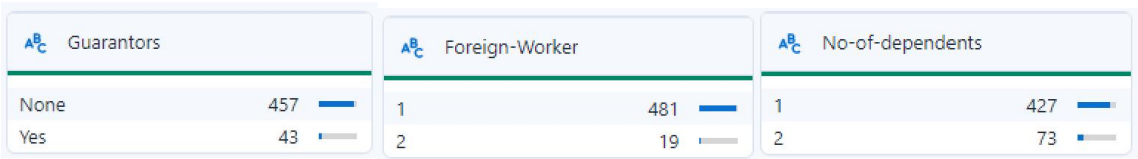


There are two column *"Duation-in-Current-address"* (68.8% missing data) and *"Age-years"* (2.4% missing data). Therefore, the column *"Duation-in-Current-address"* would be removed and the column *"Age-years"* would be imputed using median due to right skewed distribution as the image below.



- **Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.**





"Concurrent-Credits" and *"Occupation"* field are uniform since there is only one value for the entire field while *"Guarantors"*, *"Foreign Worker"* and *"No of Dependents"* show low variability

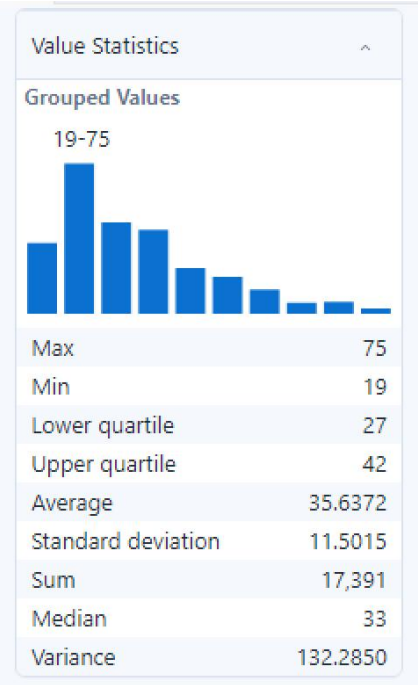
where more than 80% of the data skewed towards one data. Therefore, these 5 columns would be removed due to low variability.

"Telephone" field should also be removed due to its irrelevancy to the customer creditworthy.

- **Your clean data set should have 13 columns where the Average of Age Years should be 36 (rounded up).**

In a conclusion, the seven columns that would be removed are *"Duation-in-Current-address", "Concurrent-Credits", "Occupation", Guarantors"*, *"Foreign Worker", "No of Dependents",* and *"Telephone"*.

The column *"Age-years"* would be imputed using median. The average of Age Years is 36 (rounded up) according to the image below.



Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)****Note:*** *For students using software other than Alteryx, please format each variable as:*

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String

Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

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.**

➢ Logistic Regression (Stepwise)

Using *"Credit Application Result"* as the target variables, top three most important predictor variables are *"Account Balance", "Payment Status"* and *"Purpose"*.

Report

Report for Logistic Regression Model logistic_stepwise

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 + Age.years_Indicator, family = binomial("logit"), data = the.data)

Deviance Residuals:

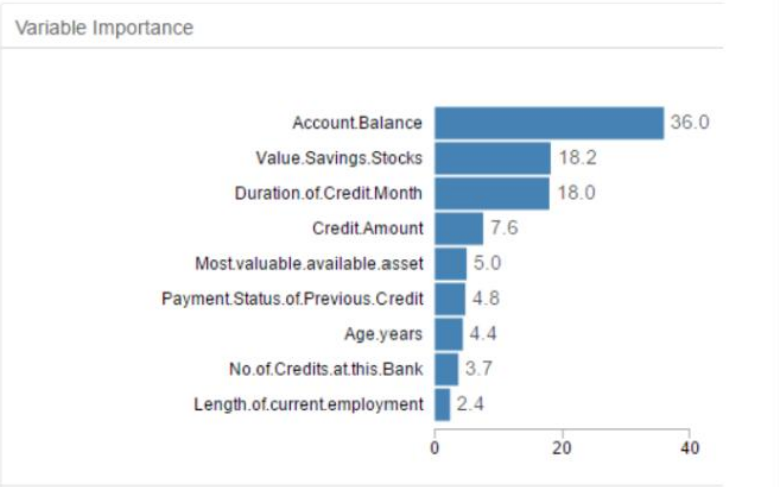
Min	1Q	Median	3Q	Max
-2.546	-0.729	-0.492	0.566	2.594

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.243e+00	8.023e-01	-4.042941	5e-05	***
Account.BalanceSome Balance	-1.089e+00	2.999e-01	-3.631551	0.00028	***
Duration.of.Credit.Month	2.776e-02	1.405e-02	1.975959	0.04816	*
Payment.Status.of.Previous.CreditPaid Up	1.148e-01	3.008e-01	0.381766	0.70264	
Payment.Status.of.Previous.CreditSome Problems	1.947e+00	5.427e-01	3.587239	0.00033	***
PurposeNew car	-1.123e+00	5.568e-01	-2.016915	0.0437	*
PurposeOther	1.710e+01	2.659e+03	0.006431	0.99487	
PurposeUsed car	-1.045e-01	3.744e-01	-0.279088	0.78018	
Credit.Amount	1.203e-04	7.182e-05	1.675453	0.09385	.
Value.Savings.StocksNone	5.931e-01	5.062e-01	1.171694	0.24132	
Value.Savings.Stocks£100-£1000	-3.637e-03	5.671e-01	-0.006414	0.99488	
Length.of.current.employment4-7 yrs	2.251e-01	4.653e-01	0.483845	0.6285	
Length.of.current.employment< 1yr	7.525e-01	3.890e-01	1.934163	0.05309	.
Instalment.per.cent	2.380e-01	1.432e-01	1.662224	0.09647	.
Age.years_Indicator	-3.333e+01	2.868e+03	-0.011624	0.99073	

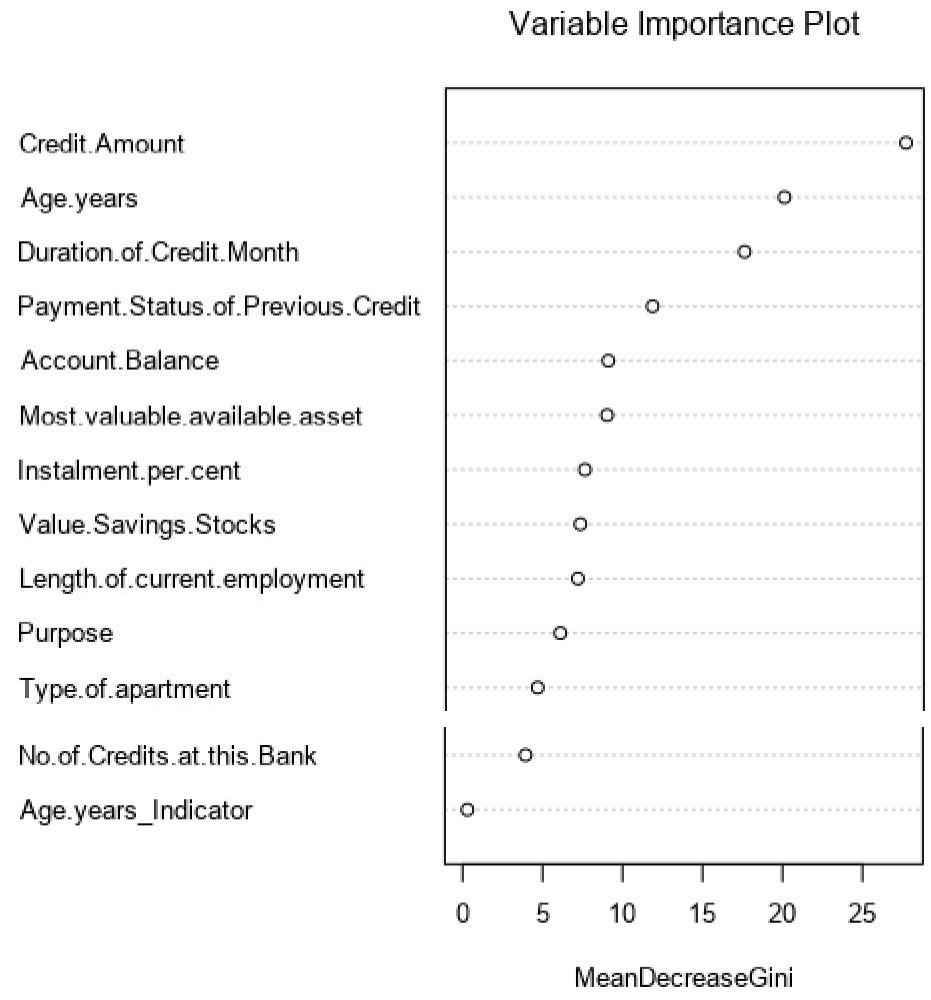
- Decision Tree

Using *"Credit Application Result"* as the target variables, top three most important predictor variables are "*Account Balance", "Value Savings Stocks"* and *"Duration of Credit Month".*



➢ Forest Model

Using *"Credit Application Result"* as the target variables, top three most important predictor variables are *"Credit Amount", "Age years"*, and *"Duration of Credit Month".*



➢ Boosted Model

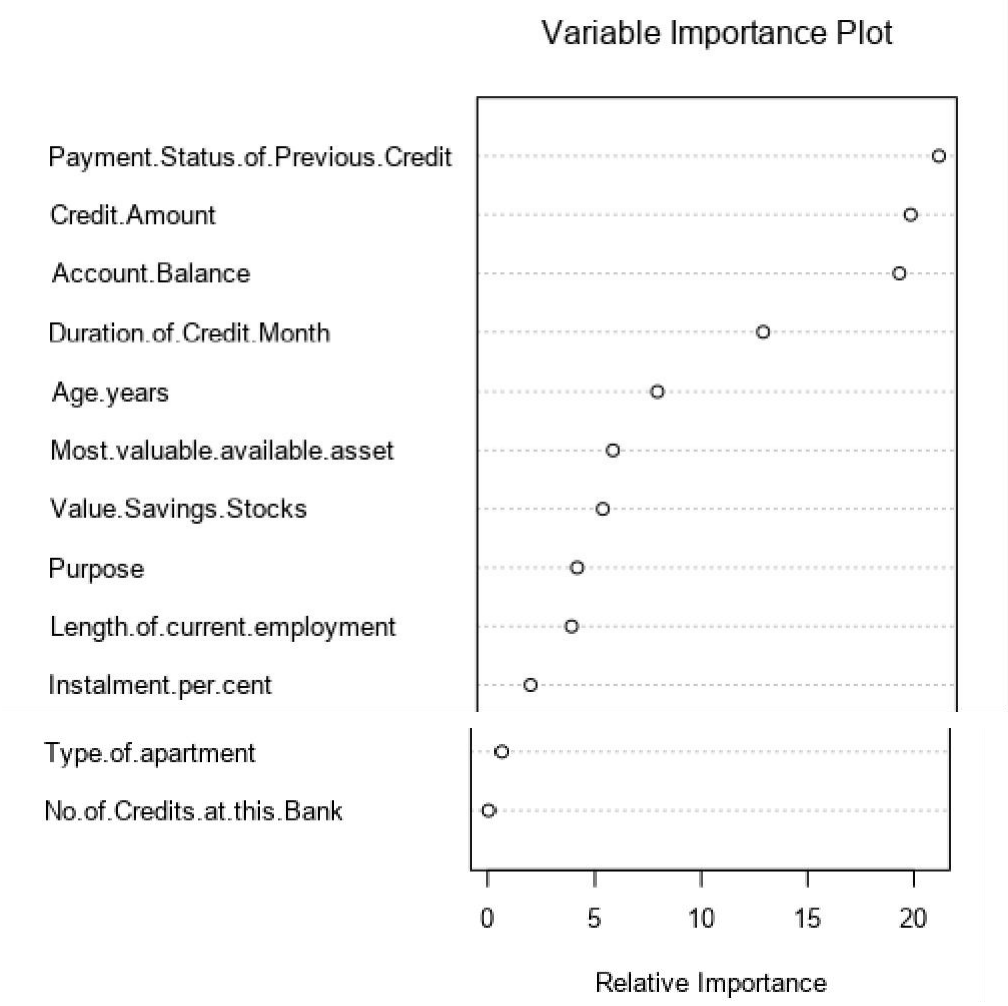
Using *"Credit Application Result"* as the target variables, top three most important predictor variables are *"Payment Status of Previous Credit"*, *"Credit Amount"*, and *"Account Balance"*.

Report

Report for Boosted Model Test_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1329



- **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?**

➢ Logistic Regression (Stepwise)

Overall accuracy is 78.67% while accuracy for creditworthy is twice higher than non-creditworthy at 92.31% and 47.83% respectively. The model is biased towards predicting customers as creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_stepwise	0.7867	0.8571	0.8135	0.9231	0.4783

Confusion matrix of logistic_stepwise

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	24
Predicted_Non-Creditworthy	8	22

➢ Decision Tree

Overall accuracy is 74.67% while accuracy for creditworthy is higher than non-creditworthy at 86.54% and 47.83% respectively. The model is biased towards predicting customers as creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_110	0.7467	0.8257	0.7340	0.8654	0.4783

Confusion matrix of Decision_Tree_110

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	90	24
Predicted_Non-Creditworthy	14	22

➢ Forest Model

Overall accuracy is 80% while accuracy for creditworthy is twice higher than non-creditworthy at 95.19% and 45.65% respectively. The model is biased towards predicting customers as creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.8000	0.8684	0.7656	0.9519	0.4565

Confusion matrix of X

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	25
Predicted_Non-Creditworthy	5	21

- Boosted Model

Overall accuracy is 74% while accuracy for creditworthy is significantly higher than non-creditworthy at 92.31% and 32.61% respectively. The model is biased towards predicting customers as creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Test_Model	0.7400	0.8312	0.7722	0.9231	0.3261

Confusion matrix of Test_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	96	31
Predicted_Non-Creditworthy	8	15

Step 4: Writeup

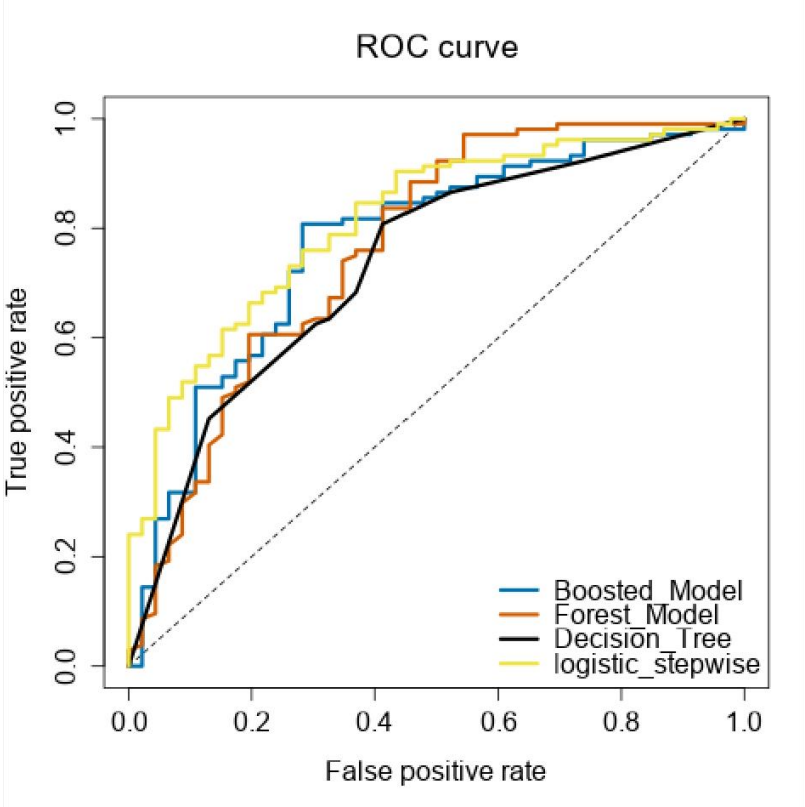
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy".

- **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**

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

Forest model is chosen as its accuracy is highest (80%) against the validation set. In addition, its accuracy for *creditworthy* is among the highest of all with 95.19% and the accuracy for *non-creditworthy* is the second highest with 45.65%.

Regarding ROC curve, the graph below shows that ***Forest model*** reaches the true positive rate at the fastest rate.



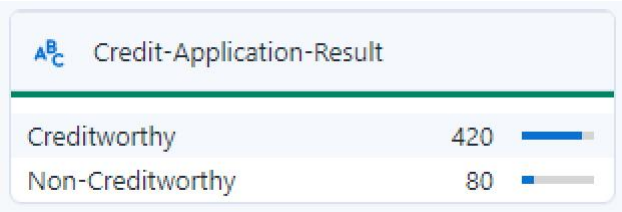
Regarding bias in the confusion matrices, the difference between accuracy of creditworthy and non-creditworthy among the four models are generally the same. Although Forest model is at second ranking in term of bias in the confusion matrices, ***Forest model*** is still chosen as the boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments. However, the boss should be noticed about bias towards any decisions when using this classification model. This is crucial because there is a high chance of lending money to customers with high probability of defaulting.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7400	0.8312	0.7722	0.9231	0.3261
Forest_Model	0.8000	0.8684	0.7656	0.9519	0.4565
Decision_Tree	0.7467	0.8257	0.7340	0.8654	0.4783
logistic_stepwise	0.7867	0.8571	0.8135	0.9231	0.4783

- **How many individuals are creditworthy?**



There are 420 creditworthy customers using forest models to score new customers.

Alteryx Flow