Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?

Ans: A small bank, where I am working, has recently faced a sudden influx of loan applications due to a competitor bank facing some scandal. Previously, bank used to **manually** process almost 200 loan applications per week. But due to an abrupt increase in loan applications, the manager wants an efficient way to categorize the applicants as creditworthy or not creditworthy. I have been entrusted to perform this task by using my knowledge of classification models and give a list of people who are creditworthy to my manager in the next two days.

- What data is needed to inform those decisions?

Ans: To make an informed decision regarding applicant's credit worthiness, the following two datasets are required:

1. A training dataset where classification of creditworthy and noncredit worthy applicants is already done based on different parameters.
2. A dataset of new applicants on which the classification needs to be done.

Both of these datasets should have information such as applicants bank balance, savings, duration of balance in his account, etc. to determine their credit worthiness.

Note: We are not determining any significant variables/parameters yet that would help us in predicting the default risk.

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

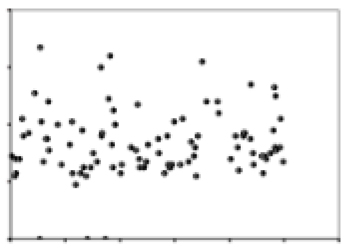
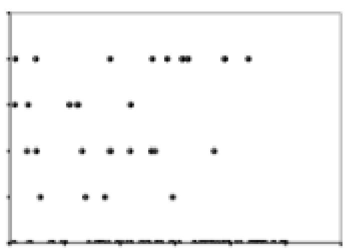
Ans: We would use a **Binary Classification Model** as the result that we are looking for has only two options, either creditworthy or not creditworthy.

Step 2: Building the Training Set

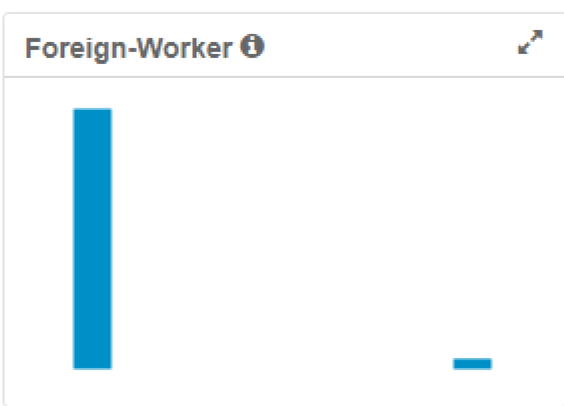
Data Cleaning Process:

Step 1: I used the **Field Summary** tool to investigate my data for any null / missing values and low variability data. The following graphs show you the results

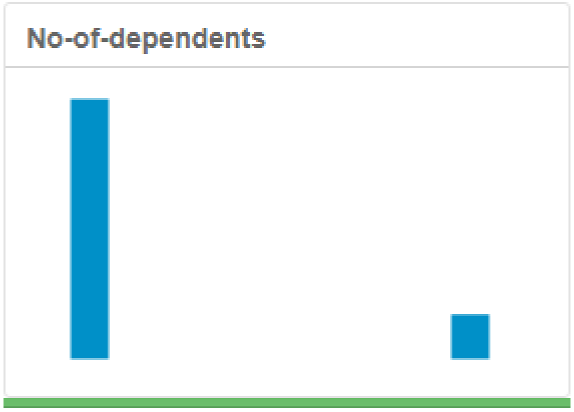
Missing Data

Name	Plot	% Missing	Unique Values	Min	Mean
Age-years		2.4%	54	19.000	35.637
Duration-in-Current-address		68.8%	5	1.000	2.660

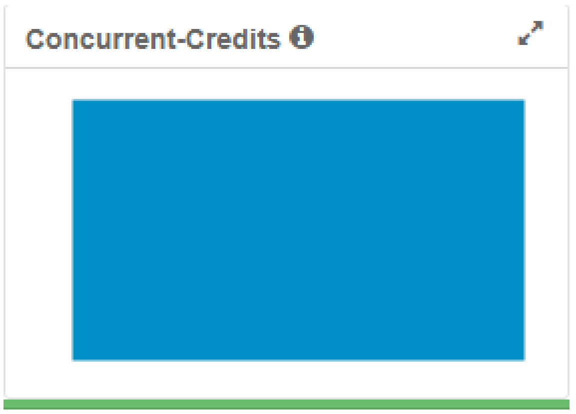
Low Variability Data



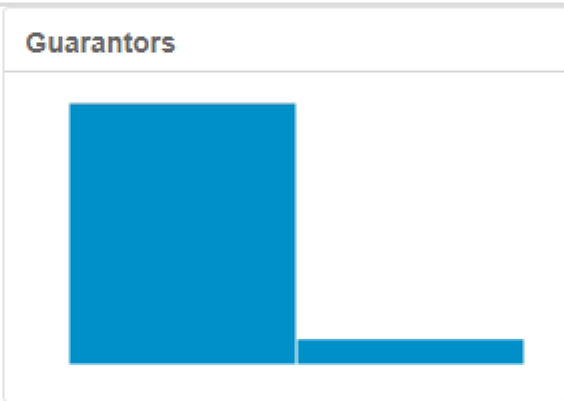
Number of Unique Values: 2



Number of Unique Values: 2

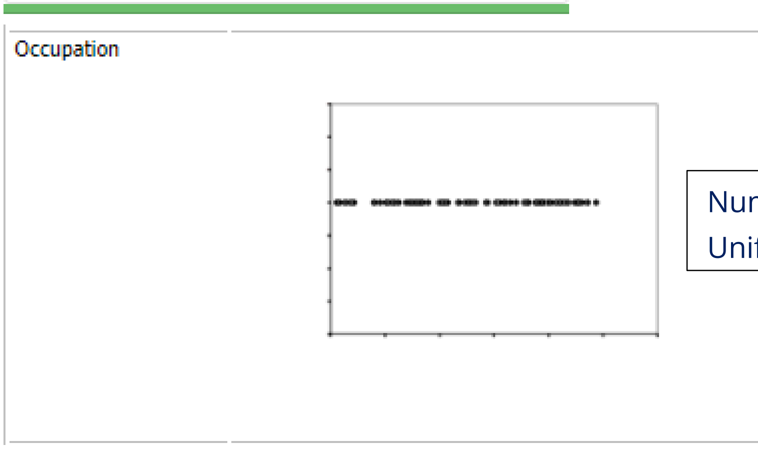


Number of Unique Values: 1
Uniform data



Number of Unique Values: 2

The data here is skewed as one value has greater count



Number of Unique Values: 1
Uniform data

Step 2: Remove low variability data columns from the dataset

	Field	Type
☑	Credit-Application-Result	V_String
☑	Account-Balance	V_String
☑	Duration-of-Credit-Month	Double
☑	Payment-Status-of-Previous-Credit	V_String
☑	Purpose	V_String
☑	Credit-Amount	Double
☑	Value-Savings-Stocks	V_String
☑	Length-of-current-employment	V_String
☑	Instalment-per-cent	Double
☐	Guarantors	V_String
☐	Duration-in-Current-address	Double
☑	Most-valuable-available-asset	Double
☑	Age-years	Double
☐	Concurrent-Credits	V_String
☑	Type-of-apartment	Double
☑	No-of-Credits-at-this-Bank	V_String
☐	Occupation	Double
☐	No-of-dependents	Double
☐	Telephone	Double
☐	Foreign-Worker	Double
☑	*Unknown	Unknown

Removed as per the instruction in the project supporting material
" *for the sake of this project please exclude Telephone from your data with the reasoning that there is no logical reason for including the variable."*

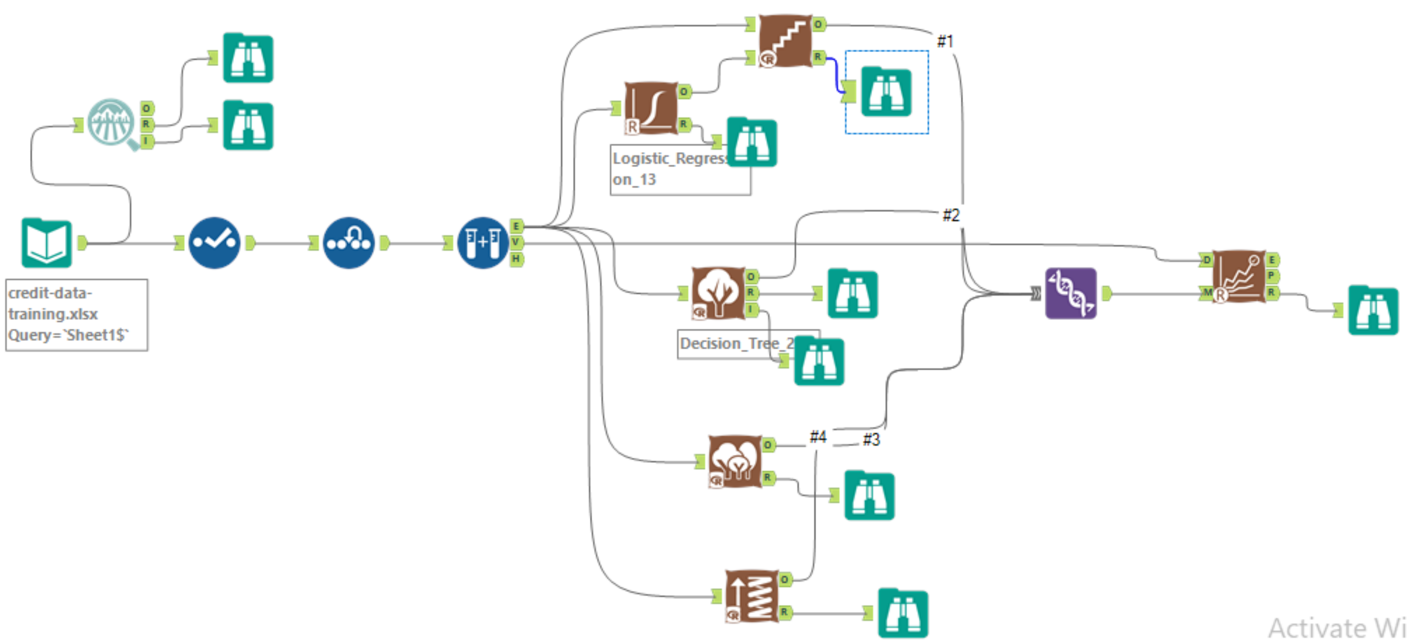
Step 3: Impute value in the column "Age-years".
Since the percentage of missing values in the column "Age-years" is only **2.4%,** it is not a good idea to drop the column. Instead we have imputed median value of the Age-year column where the data was missing. We have done this using the **Imputation tool** in Alteryx.

To summarize:

- We identified missing data and low variability data
- Removed columns with high percentage of missing values
- Removed columns with low variability
- Imputed median value in the missing fields of Age-Years column

Step 3: Train your Classification Models

<u>My Alteryx Workflow:</u>



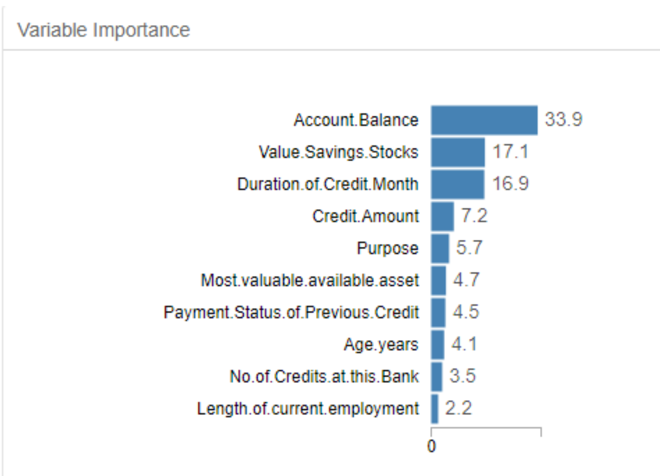
<u>Model # 1: Logistic Stepwise</u>

Significant Variables: Account Balance, Payment Status of Previous Credit, Purpose, Length of Current Employment, Installment

	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	.

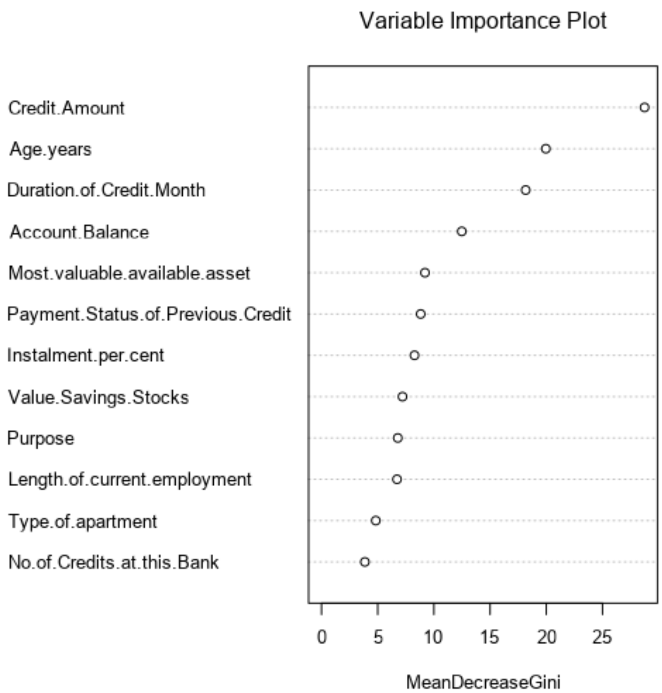
Model # 2: Decision Trees

Top 3 Significant Variables: Account Balance, Value Savings Stocks, Duration of Credit Month



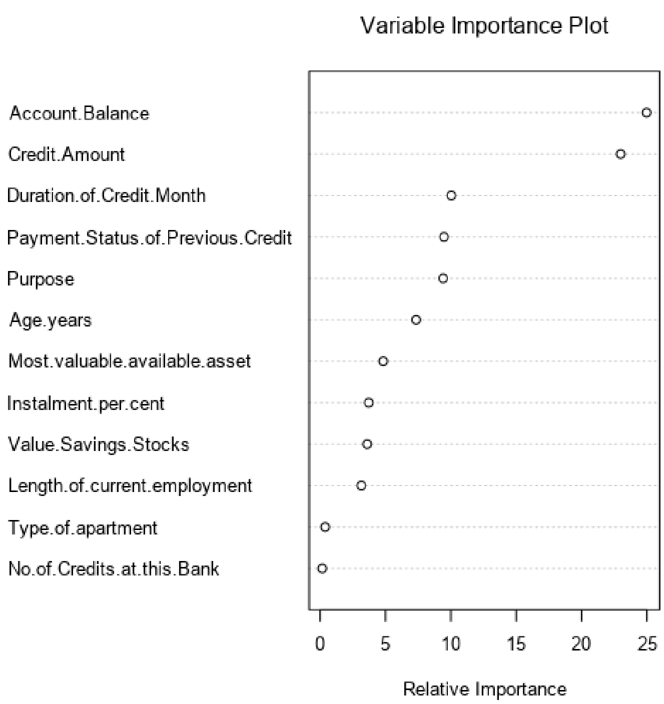
Model # 3: Random Forest

Top 3 Significant Variables: Account Balance, Age.years, Duration of Credit Month



Model # 4: Boosted Model

Top Significant Variables: Account Balance, Credit Amount



Model Comparison

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SW_CreditWorthiness	0.7600	0.8364	0.7306	0.8762	0.4889
DT_CreditWorthiness	0.7467	0.8304	0.7035	0.8857	0.4222
RF_CreditWorthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthiness	0.7867	0.8632	0.7515	0.9619	0.3778

SW: Logistic Stepwise Model
DT: Decision Tree Model
RF: Random Forest Model
BM: Boosted Model

The overall accuracy of the models is as follow:
Logistic Step Wise: 76%
Decision Tree: 74.67%
Random Forest: 79.33%
Boosted Model: 78.67%

Biasness: Decision Tree and Logistic Stepwise Regression is biased towards their true positives but Random Forest and Boosted Model are not.

Confusion Matrix

Confusion matrix of BM_CreditWorthiness

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

Confusion matrix of DT_CreditWorthiness

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

Confusion matrix of RF_CreditWorthiness

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

Confusion matrix of SW_CreditWorthiness

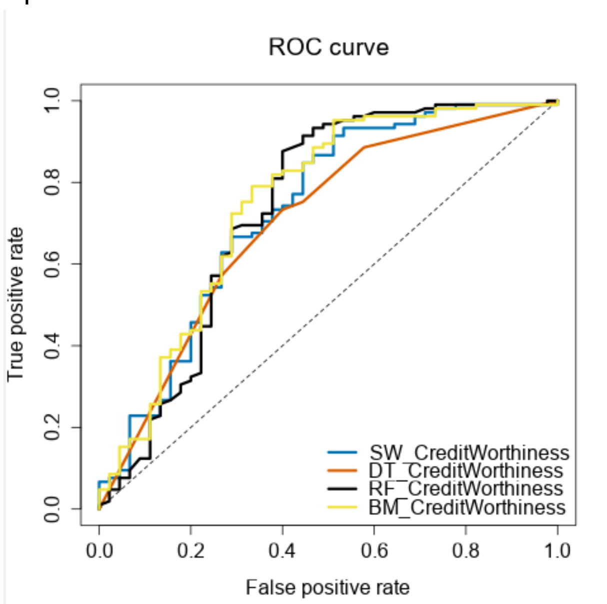
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Step 4: Writeup

- 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:

Ans: I chose **Random Forest Model** because:

- Overall Accuracy against your Validation set
  - The overall accuracy of Random Forest is **79.33%**, which is the highest among all four models.
- Accuracies within "Creditworthy" and "Non-Creditworthy" segments
  - Random Forest Model's accuracy to predict creditworthy people is **97.14%**
  - Its accuracy to predict non-creditworthy people is **37.37%** which is quite low, but given the top two models based on accuracy_creditworthy score, Random Forest being on first and Boosted Model on second, they both have same accuracy of predicting non-creditworthy people. Hence we'll still choose Random Forest.
- ROC graph



Random Forest has the best line curve on the ROC

- Bias in the Confusion Matrices
  - Decision Tree and Logistic Stepwise Regression is biased towards their true positives but Random Forest and Boosted Model are not.

- How many individuals are creditworthy?

Ans: There are **408** people who are credit-worthy and **92** people who are not.