Project: Creditworthiness
Obalana Oluwatosin

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

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

The decisions to be made are:

1. How to process all of these loan applications within one week.
2. Evaluating the creditworthiness of the new loan applicants.
3. Providing a list of creditworthy customers to my manager in the next two days.

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

The data to inform these decisions are:

- Data on all past applications: credit-data-training.xlsx, the file contains all credit approvals from the past loan applicants the bank has ever completed.
- [ ] **Credit-Application Result:** If applicant is Creditworthy or Non-Creditworthy
- [ ] **Account-Balance**: Account balance of the applicant:
- [ ] **Duration-of-Credit-Month:** Duration of credit applied for Month
- [ ] **Payment-Status-of- Previous-Credit:** Status related to previous
- [ ] **Purpose:** Purpose for which the credit is taken
- [ ] **Credit Amount:** Credit applied for
- [ ] **Age-years;** Age in years
- [ ] **Most Valuable Asset:** Applicant Valuable asset
- [ ] **Value-Saving-Stocks:** Savings
- [ ] **Installment-per-cent:** Installment percent
- [ ] **Duration-in-Current-address:** Time in current address
- [ ] **Length-of-current-employment:** Length of employment in range
- [ ] **No-of-Credits-at-this-Bank:** Number of credit at the bank

- The list of customers that need to be processed in the next few days[customers-to-score.xlsx, this is the new set of customers that you need to score on the classification model you will create.] The only difference in the variables with the one above would be the exception of the **Credit-Application Result**because that's what we are to predict

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

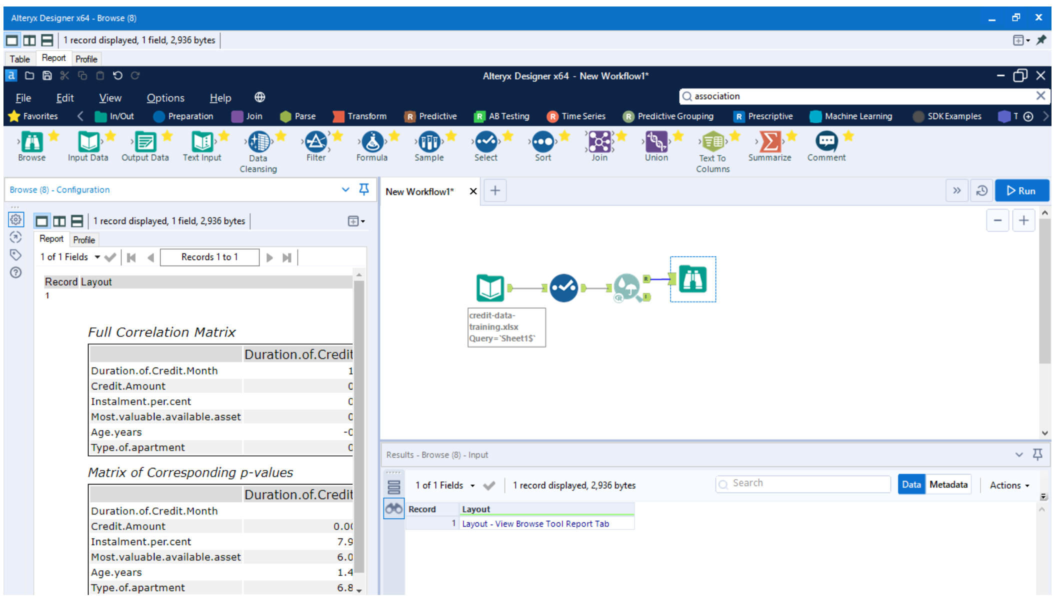
We are going to be using alteryx to create the classification models, that means we would be using the binary classification model. . The variable to be predicted would be the Credit-Application-Result

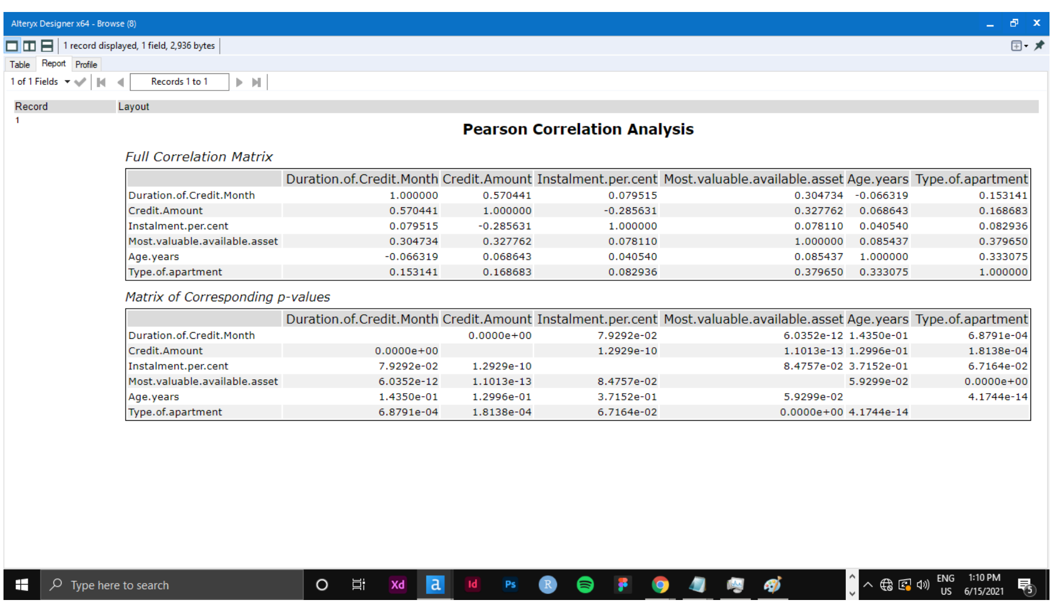
Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't need to convert any data fields to the appropriate data types.

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".

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





Pearson Correlation Analysis

Full Correlation Matrix

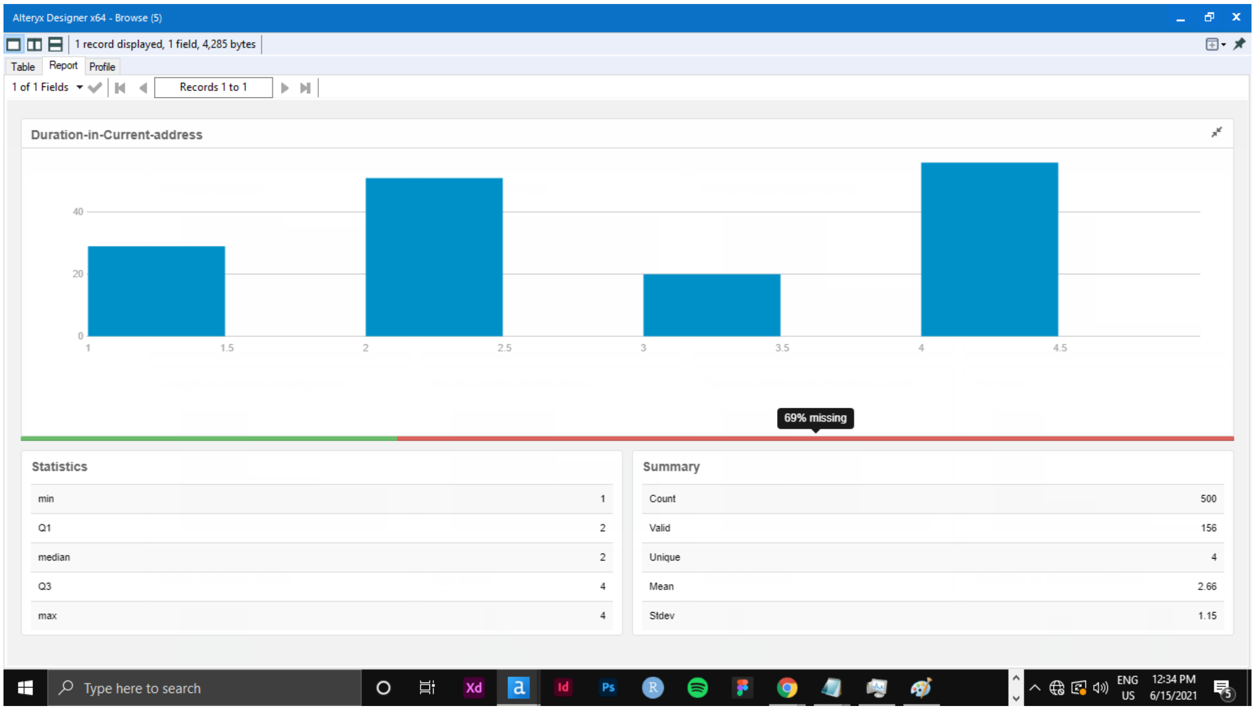
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Age.years	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

Matrix of Corresponding p-values

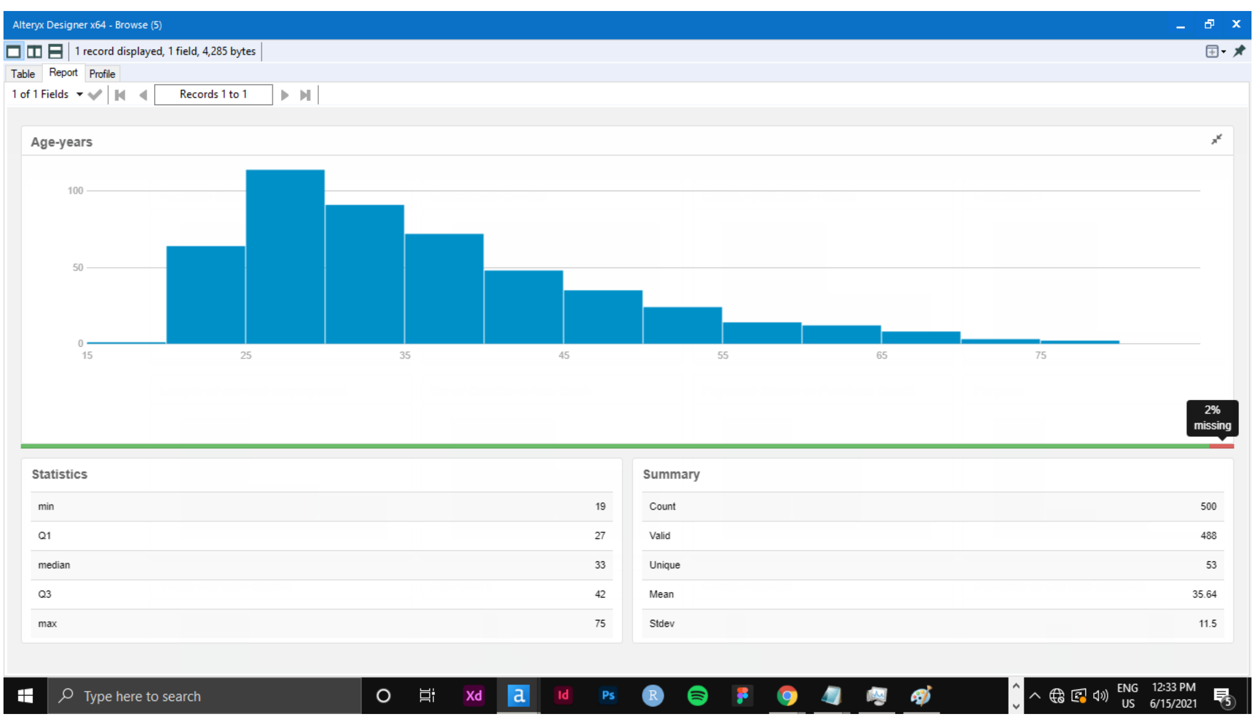
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month		0.0000e+00	7.9292e-02	6.0352e-12	1.4350e-01	6.8791e-04
Credit.Amount	0.0000e+00		1.2929e-10	1.1013e-13	1.2996e-01	1.8138e-04
Instalment.per.cent	7.9292e-02	1.2929e-10		8.4757e-02	3.7152e-01	6.7164e-02
Most.valuable.available.asset	6.0352e-12	1.1013e-13	8.4757e-02		5.9299e-02	0.0000e+00
Age.years	1.4350e-01	1.2996e-01	3.7152e-01	5.9299e-02		4.1744e-14
Type.of.apartment	6.8791e-04	1.8138e-04	6.7164e-02	0.0000e+00	4.1744e-14	

We can see that there's no correlation greater than 0.7 for the numerical fields

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



- The **Duration-in-Current-address column** should be excluded by unchecking it using the **select tool** because we have **69%** of the data missing .



The **Age.years column** has **2%** of missing data, the solution to this is to impute the median of the Age .years which is 33, using the mean can cause bias, so therefore the median is an appropriate representation to replace the missing values. I renamedthe new column to be Age_yrs

Checking whether the **numerical data fields correlate with each other**, after imputing and removing the missing values

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Type.of.apartment
Duration.of.Credit.Month	1.000000	0.573980	0.068106	0.299855	0.152516
Credit.Amount	0.573980	1.000000	-0.288852	0.325545	0.170071
Instalment.per.cent	0.068106	-0.288852	1.000000	0.081493	0.074533
Most.valuable.available.asset	0.299855	0.325545	0.081493	1.000000	0.373101
Type.of.apartment	0.152516	0.170071	0.074533	0.373101	1.000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Type.of.apartment
Duration.of.Credit.Month		0.0000e+00	1.2830e-01	7.5764e-12	6.2192e-04
Credit.Amount	0.0000e+00		4.5919e-11	8.3045e-14	1.3277e-04
Instalment.per.cent	1.2830e-01	4.5919e-11		6.8653e-02	9.5961e-02
Most.valuable.available.asset	7.5764e-12	8.3045e-14	6.8653e-02		0.0000e+00
Type.of.apartment	6.2192e-04	1.3277e-04	9.5961e-02	0.0000e+00	

There's no correlation greater than 0.7 for the numerical fields

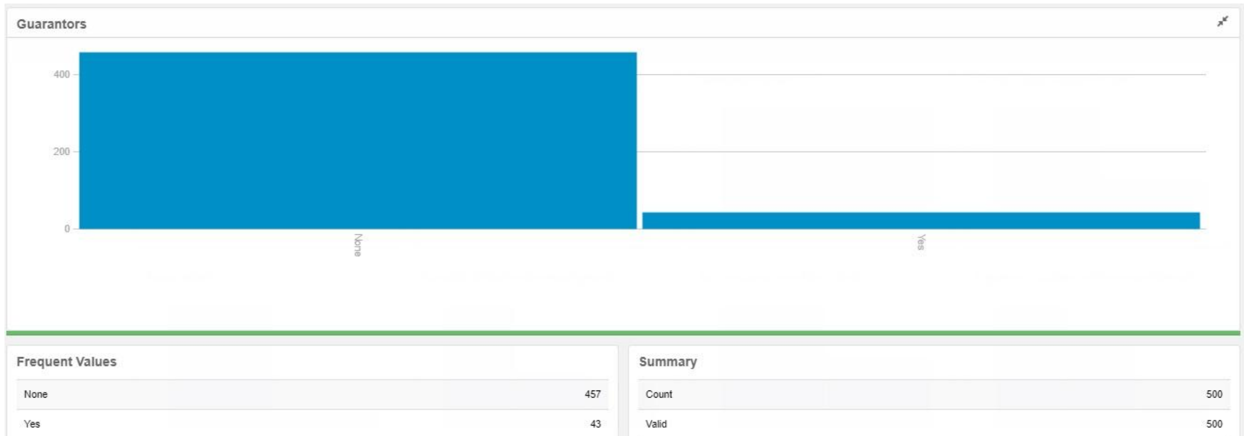
Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and you should remove fields that have low variability. Refer to the "Tips" sectionto find examples of data fields with low-variability.

Concurrent-Credits

- low variability: it has only one value(500)

Guarantors

- Low variability: it is skewed towards **none** (457 to 43)



Foreign-Worker

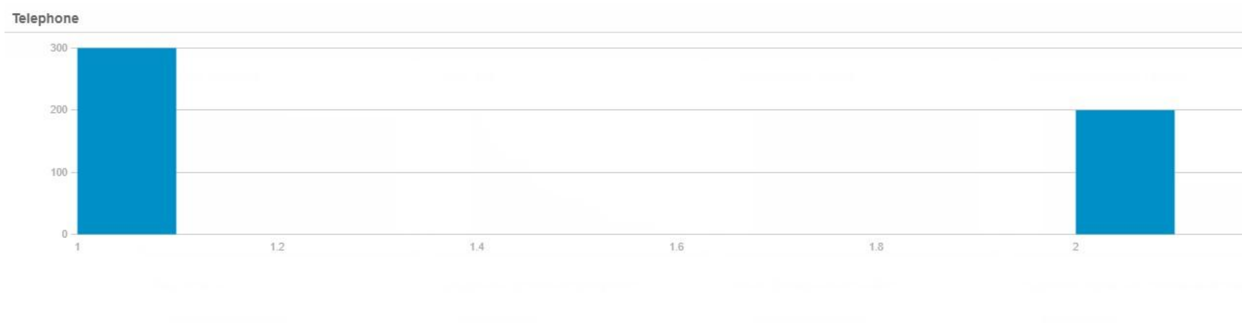
- Low variability: it is skewed towards **one(1)** (481 to 19))

No-of-dependents

- Low variability: it is skewed towards **one(1)** (473 to 73))

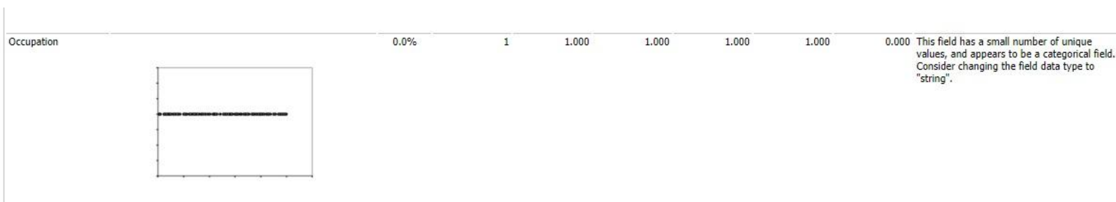
Telephone

- Low variability:



Occupation

- Low variability: Only one value

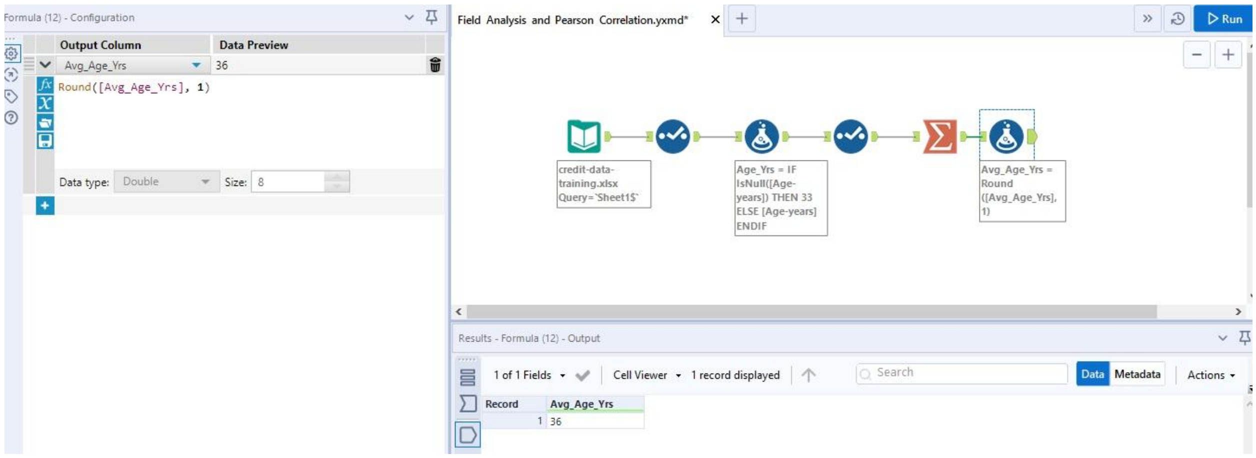


All these are unselected/removed.

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

Fields

	Name	Type	Size	Source
1	Credit-Application-Result	V_String	255	File: C:\Users\Ogrey\Video
2	Account-Balance	V_String	255	File: C:\Users\Ogrey\Video
3	Duration-of-Credit-Month	Double	8	File: C:\Users\Ogrey\Video
4	Payment-Status-of-Previous-Credit	V_String	255	File: C:\Users\Ogrey\Video
5	Purpose	V_String	255	File: C:\Users\Ogrey\Video
6	Credit-Amount	Double	8	File: C:\Users\Ogrey\Video
7	Value-Savings-Stocks	V_String	255	File: C:\Users\Ogrey\Video
8	Length-of-current-employment	V_String	255	File: C:\Users\Ogrey\Video
9	Instalment-per-cent	Double	8	File: C:\Users\Ogrey\Video
10	Most-valuable-available-asset	Double	8	File: C:\Users\Ogrey\Video
11	Type-of-apartment	Double	8	File: C:\Users\Ogrey\Video
12	No-of-Credits-at-this-Bank	V_String	255	File: C:\Users\Ogrey\Video
13	Age_Yrs	Double	8	Formula: IF IsNull([Age-yea



Note: For the sake of consistency in the data cleanupprocess, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)

Step 3: Train your Classification Models

First, create your Estimation and Validation sampleswhere 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:

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.
2. Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any biases seen in the model's predictions?

The Process

1. I built four models [**Logistic Regression Model, DecisionTree, Random Forest, Boosted Model** and then joined them all together using the join tool.
2. I used the Model Comparison Tool to **validate the models**, **compare accuracies**, and check for the **important variables**

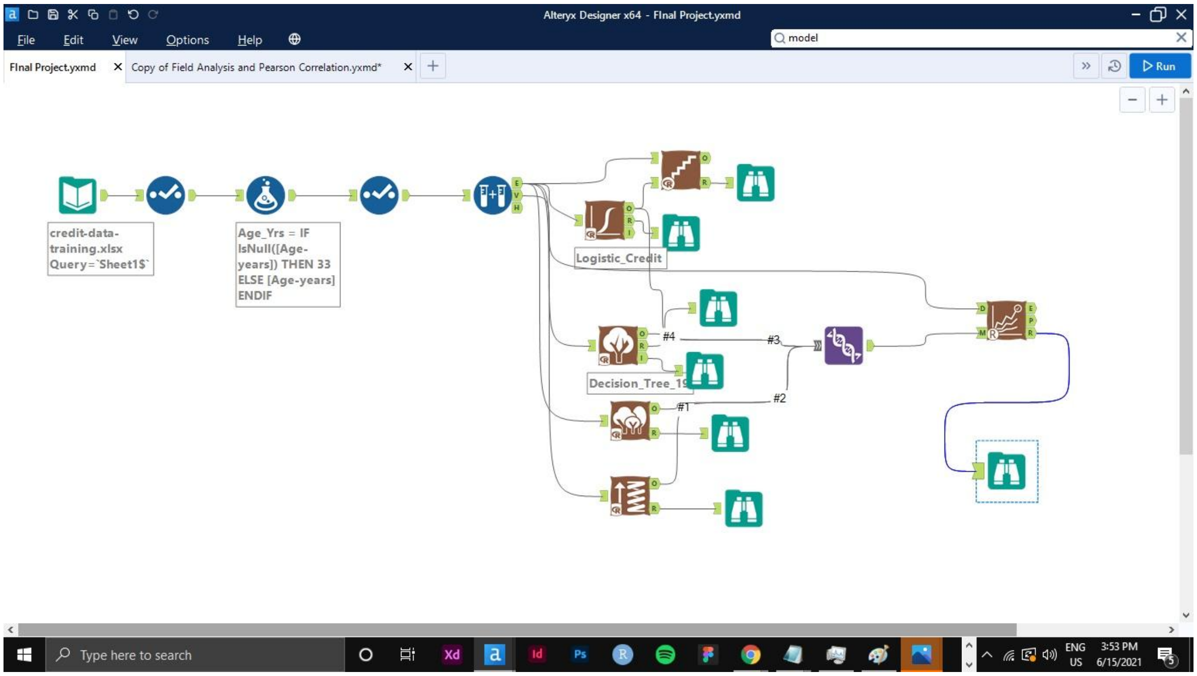
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
rf_credit	0.8000	0.8707	0.7361	0.9619	0.4222
dt_credit	0.6667	0.7685	0.6272	0.7905	0.3778
Logistic_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
boosted_credit	0.7933	0.8670	0.7505	0.9619	0.4000

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.

Model Comparison Report showing the different accuracies



My Alteryx Workflow containing the four models [**Logistic Regression Model, Decision Tree, Random Forest, Boosted Model**]

Confusion Matrix for the **Logistic Regression Model, Decision Tree, Random Forest, Boosted Model**

N.B : dt_credit = Decision Tree Model and rf_credit = Random Forest Model

Confusion matrix of Logistic_Credit

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

Confusion matrix of boosted_credit

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

Confusion matrix of dt_credit

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

Confusion matrix of rf_credit

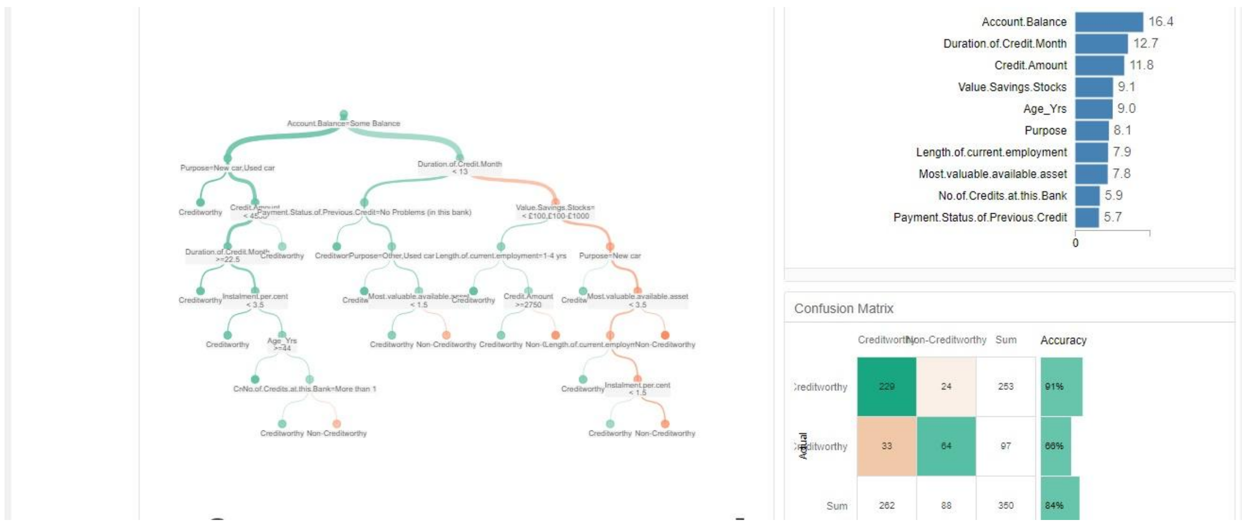
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Model	**Accuracy**
Random Forest	0.8000

Boosted	0.7933
Logistic Regression	0.7800
Decision Tree	0.6667

Decision Tree

Decision Tree model ranks last in our analysis with an accuracy of **66.67%**



The important variables are:

- Account Balance
- Credit Amount
- Duration.of.Credit.Month

Confusion matrix of dt_credit

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

Calculating the **ppv[Positive Predicted Value]** we have

Prec = $TP / TP \ + FP$

(83/83+28) * 100 = **75%**

Calculating the **npv[Negative Predicted Value]** we have

FPR = FP/TN + FP

(17/22+17) * 100 = **44%**

75%-44% = **31%** Model is biased to the Creditworthy

Logistics Regression Model

Logistic Regression has an accuracy of **78%** it comes third in our ranking.

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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age_Yrs, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	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_Yrs	-0.0141206	1.535e-02	-0.9202	0.35747

Significant Variables	**P-values**
Account.BalanceSome Balance	***
Payment.Status.of.Previous.CreditSome Problems	*
PurposeNew car	**
Credit.Amount	**
Length.of.current.employment< 1yr	*
Instalment.per.cent	*
Most.valuable.available.asset	*

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

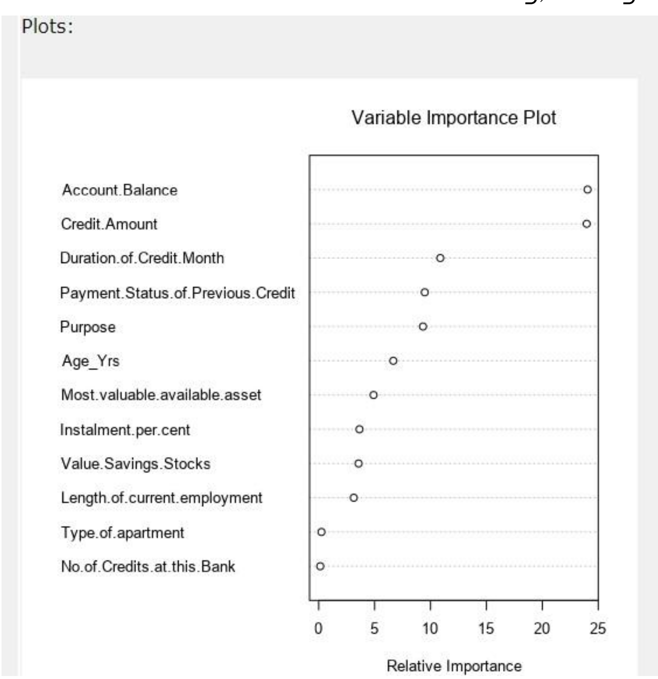
Calculating the **ppv[Positive Predicted Value]** we have
Prec = $TP/TP + FP$
(95/95+23) * 100 = **81%**

Calculating the **npv[Negative Predicted Value]** we have
FPR = FP/TN + FP
(22/10+22) * 100 = **69%**

81% - 69% **= 12% -** Model is biased to the Creditworthy

Boosted Model

The boosted model is second in our ranking, having an accuracy of **79.33%**



Based on the Variable Importance plot above we cansee that

- Account Balance
- Credit Amount
- Duration.of.Credit.Month

  Are the important predictive variables.

Confusion matrix of boosted_credit

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

Calculating the **ppv[Positive Predicted Value]** we have

Prec = $TP/ TP \; + \; FP$

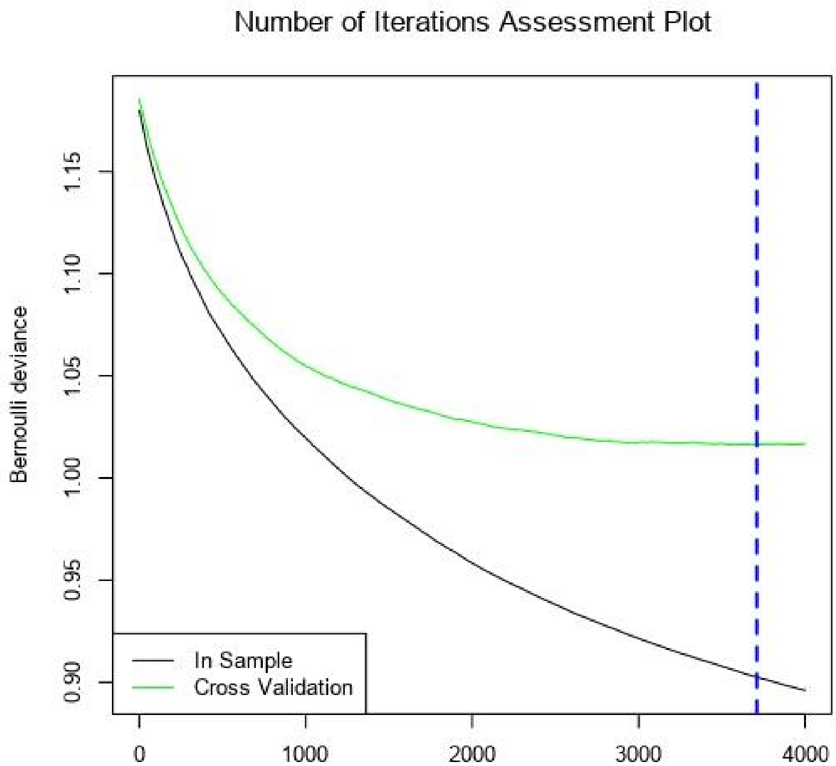
(101/101+27) * 100= **79%**

Calculating the **npv[Negative Predicted Value]** we have

FPR = FP/TN + FP

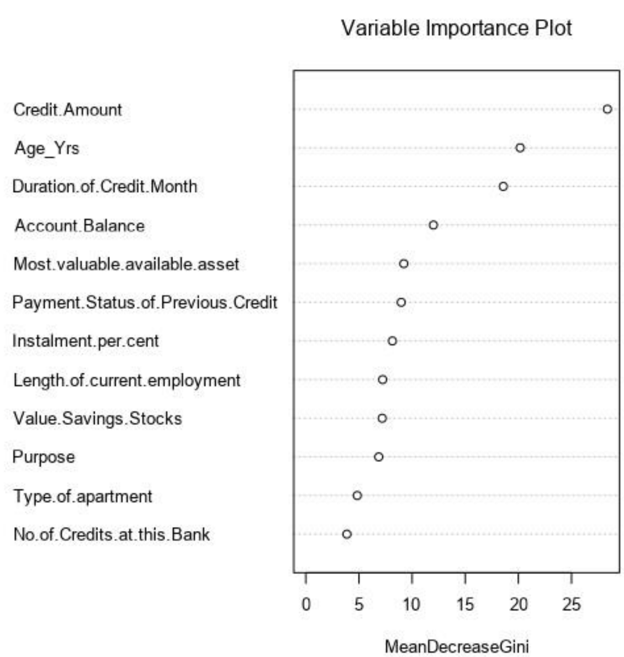
(18/4+18) * 100 = **82%**

3% - Model is unbiased



Random Forest Model:

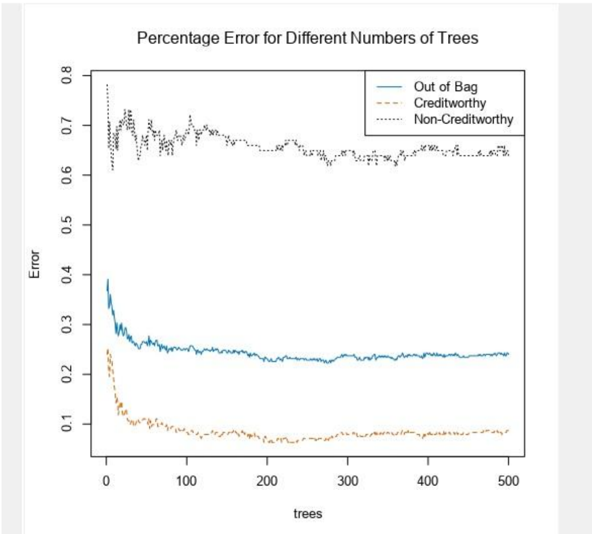
This showed the best accuracy (**80%**), it ranks first in our analysis out of the four models.



Based on the Variable Importance plot above we cansee that

- Age_ Yrs
- Credit Amount
- Duration.of.Credit.Month

  Are the important predictive variables



Calculating the **ppv[Positive Predicted Value]** we have

Prec = $TP/TP + FP$

(101/101+26) * 100= **80%**

Calculating the **npv[Negative Predicted Value]** we have

FPR = FP/TN + FP

(19/4+19) * 100 = **83%**

3% - Model is unbiased

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"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

● Which model did you choose to use? Please justifyyour decision using all of the following techniques. Please only use these techniques to justifyyour decision:

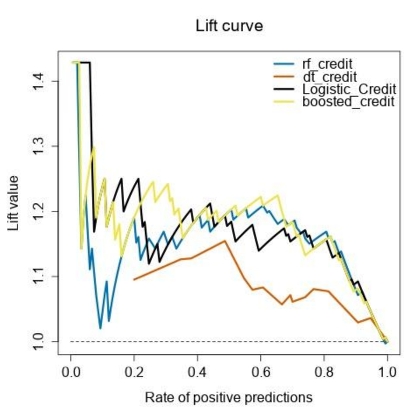
○ Overall Accuracy against your Validation set
○ Accuracies within “Creditworthy” and “Non-Creditworthy” segments
○ ROC graph
○ Bias in the Confusion Matrices

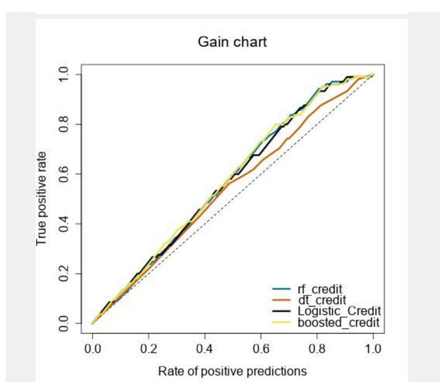
Note: Remember that your boss only cares about predictionaccuracy for Creditworthy and Non-Creditworthy segments.

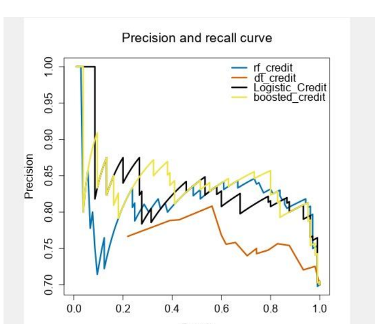
● How many individuals are creditworthy?

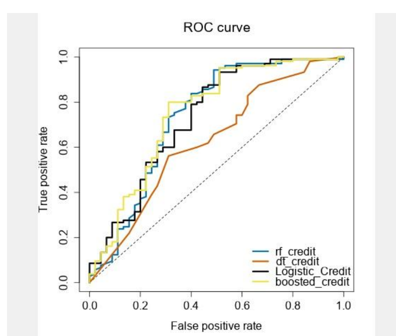
Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.









The forest model has the highest accuracy (0.8000), since my manager only cares about the accuracy, I'll use the random forest model. I alsoanalysed the **Negative Predicted Value** and the **Positive Predicted Value.**Next we can see that the random forest model has the highest F1 score, also we looked at the gain chart. At the ROC we can seethat the blue line [Random Forest] has a high performance compared to others.

How many individuals are creditworthy?
According to the calculations we have 408 out of 500 who are credit worthy