Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?
  Ans: Which new customers can be approved for a loan or not.

- What data is needed to inform those decisions?
  Ans: Account Balance, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Credit Amount, Value Saving Stocks, Length of Current Employment, Instalment per-cent, Guarantors, Duration in Current Address, Most Valuable Available Asset, Age, Type of Apartment, Number of Credits at this Bank, Occupation, Number of Dependents, Foreign Worker.

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

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.***

Here are some guidelines to help guide your data cleanup:

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

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

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)*

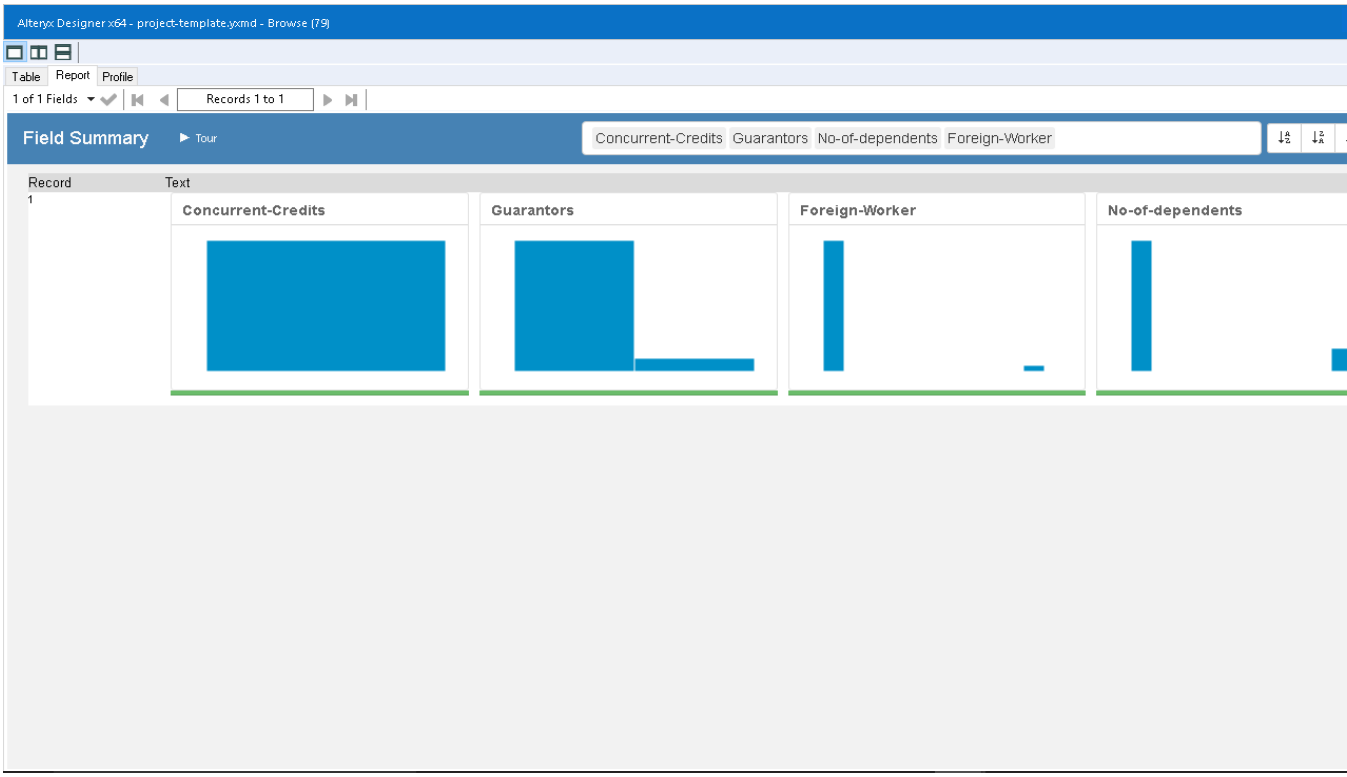
Note: *For students using software other than Alteryx, please format each variable as:*

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

To achieve consistent results reviewers expect.

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.
  Ans: **Duration in Current Address** has a lot of missing values (68.8%) and so is dropped. **Guarantors**, **Concurrent-Credits**, **Occupation**, **No-of-dependents**, and **Foreign-Worker** all have **low-variability** and so are dropped. **Telephone** is not relevant for the analysis and is also dropped.
  Age-years has few missing values (**2.4%**) and is imputed with the median as the median is not affected by outliers.



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.

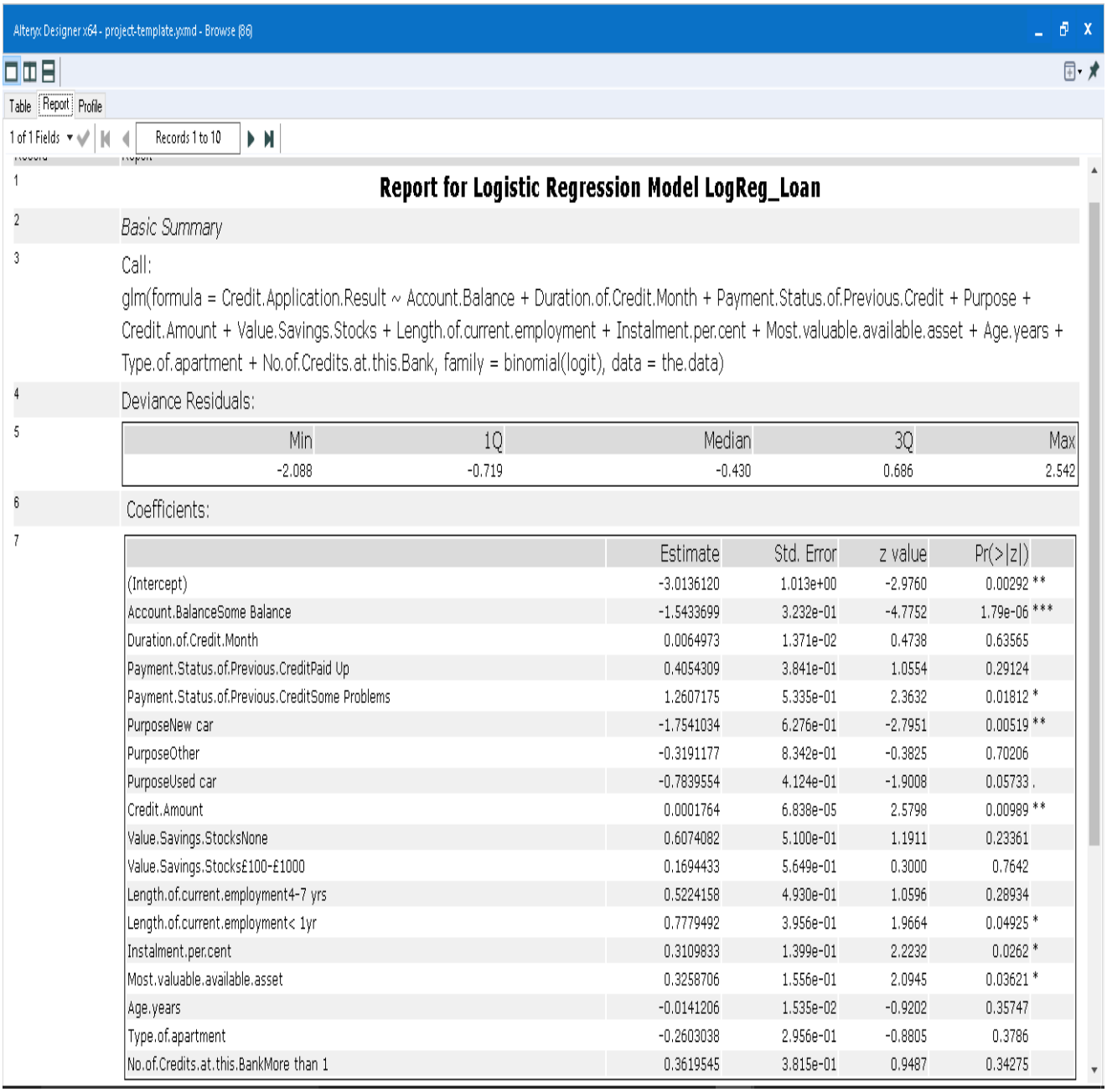
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*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.

- 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?

You should have four sets of questions answered. (500 word limit)
Ans: **Logistic Regression**- The most important predictor variables are **Account BalanceSomeBalance**, **Payment Status of Previous CreditSome Problems**, **PurposeNew Car**, **Credit Amount**, **Length of current employment< 1yr**, **Instalment,per.cen**t, **Most.valuable.available.asset.**



Report for Logistic Regression Model LogReg_Loan

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

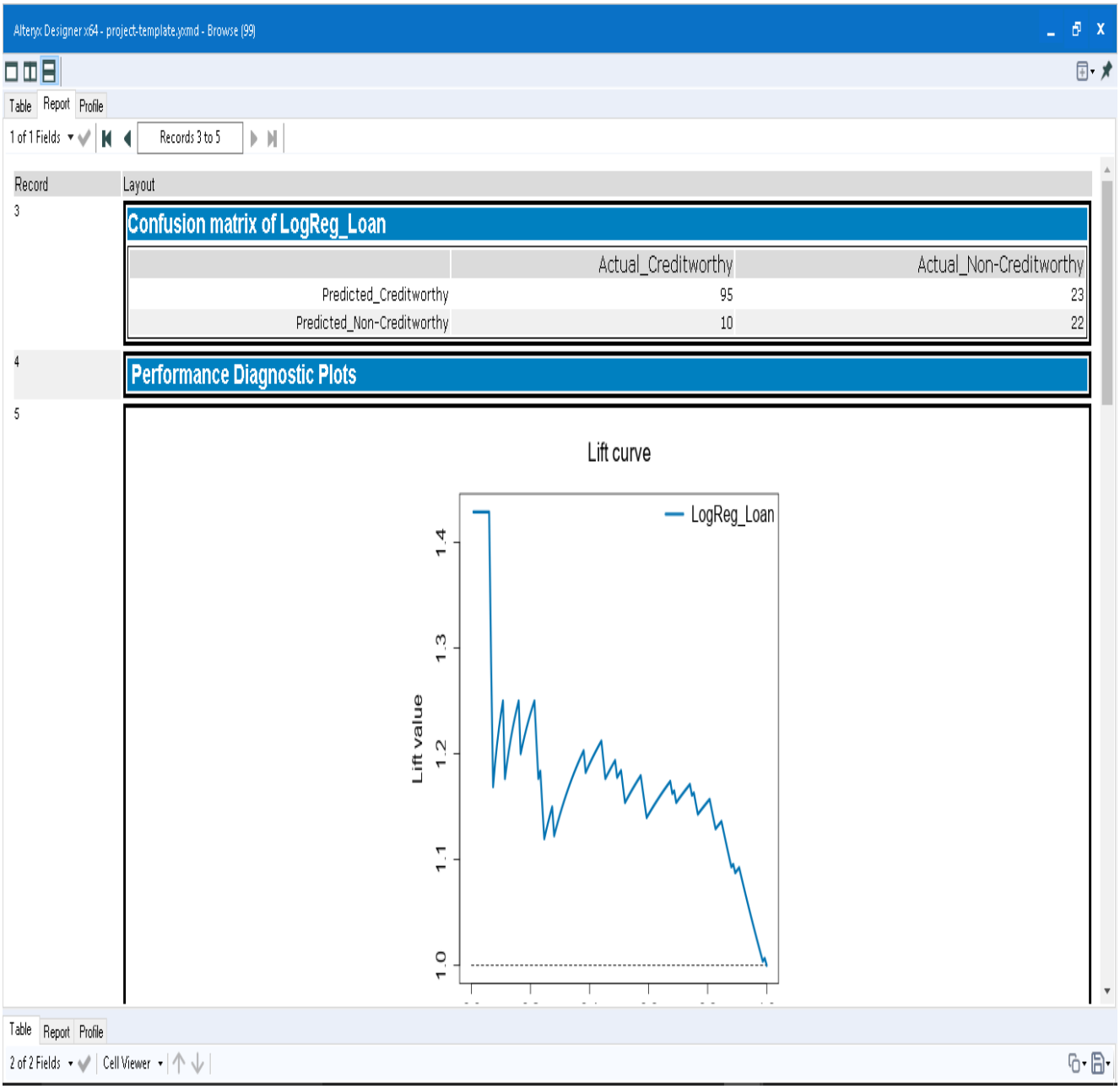
Deviance Residuals:

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

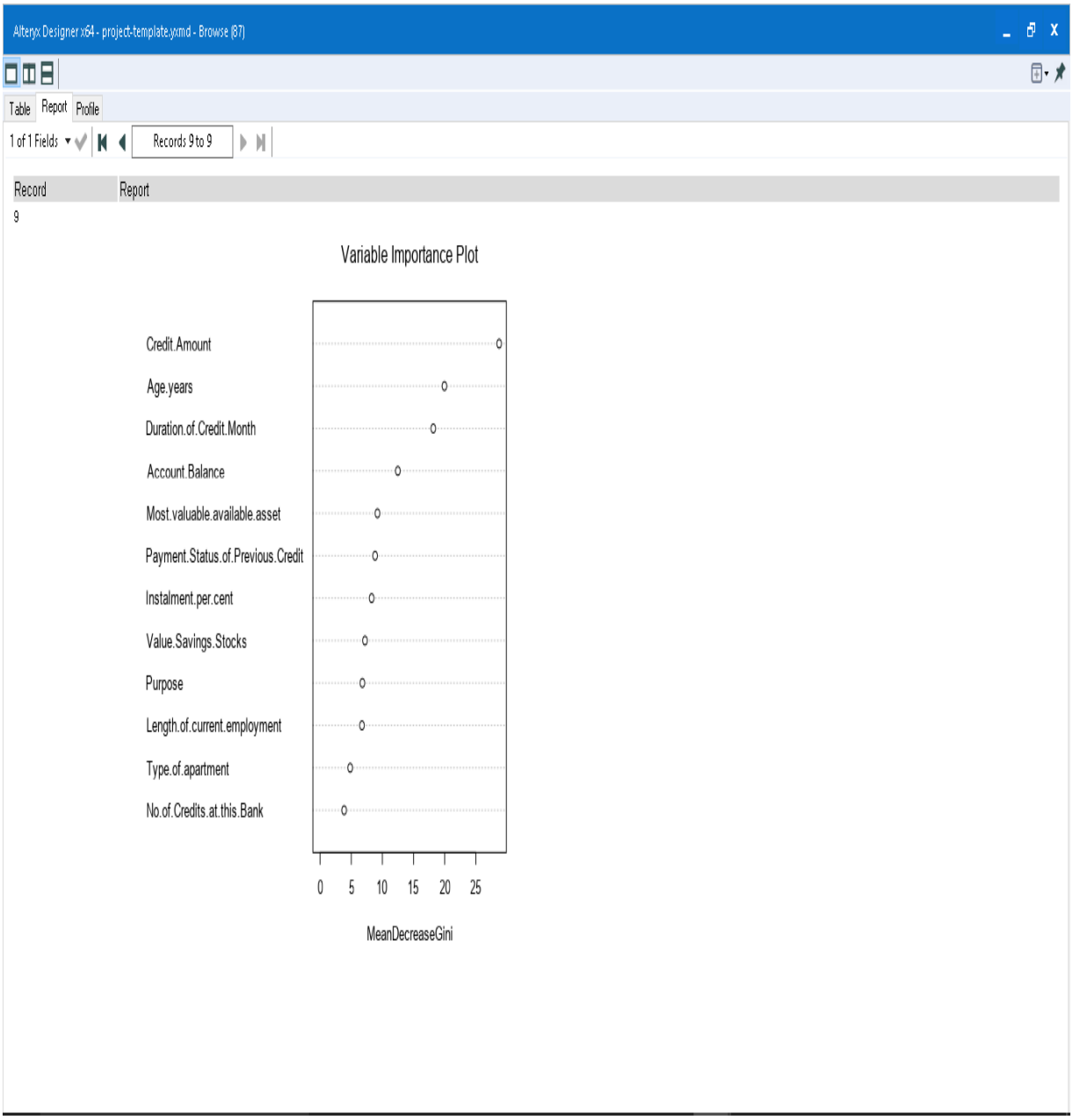
Coefficients:

	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	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
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	

Overall % Accuracy for Logistic Regression Model is 78%. The model is biased as it predicts the Creditworthy class better than it does the Non-Creditworthy class.



Forest Model- The most important predictor variables are **Credit Amount, Age-years, Duration.of.Credit.Month, Account Balance**, Most.valuable.available.asset, **Payment Status of Previous Credit, Instalment,per.cent**, Value.Savings.Stocks, **Purpose**, **Length of current employment**.



Overall % Accuracy for Forest Model is 79.33%. The model predicts the Creditworthy class correctly 102 times with 28 False Positives and predicts the Non-Creditworthy class correctly 17 times with 3 False Negatives hence it does fairly well.

Alteryx Designer x64 - project-template.yxmd - Browse (99)

Table | Report | Profile

1 of 1 Fields | Records 1 to 5

Record	Layout
1	

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Loan	0.7933	0.8681	0.7368	0.9714	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*.

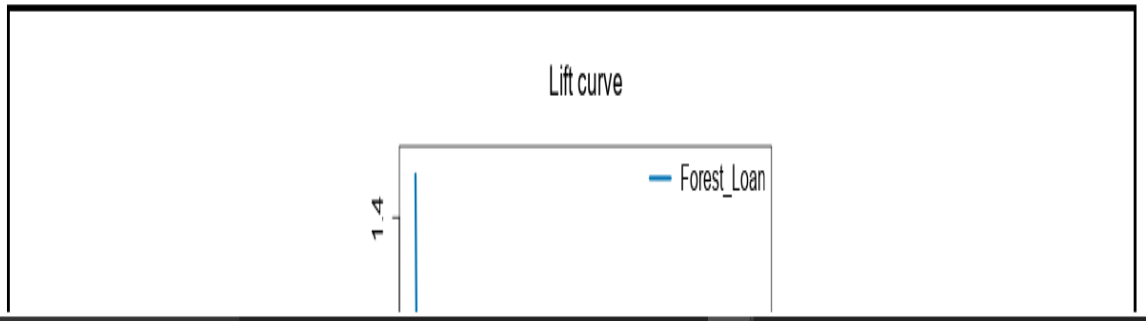
AUC: area under the ROC curve, only available for two-class classification.

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.

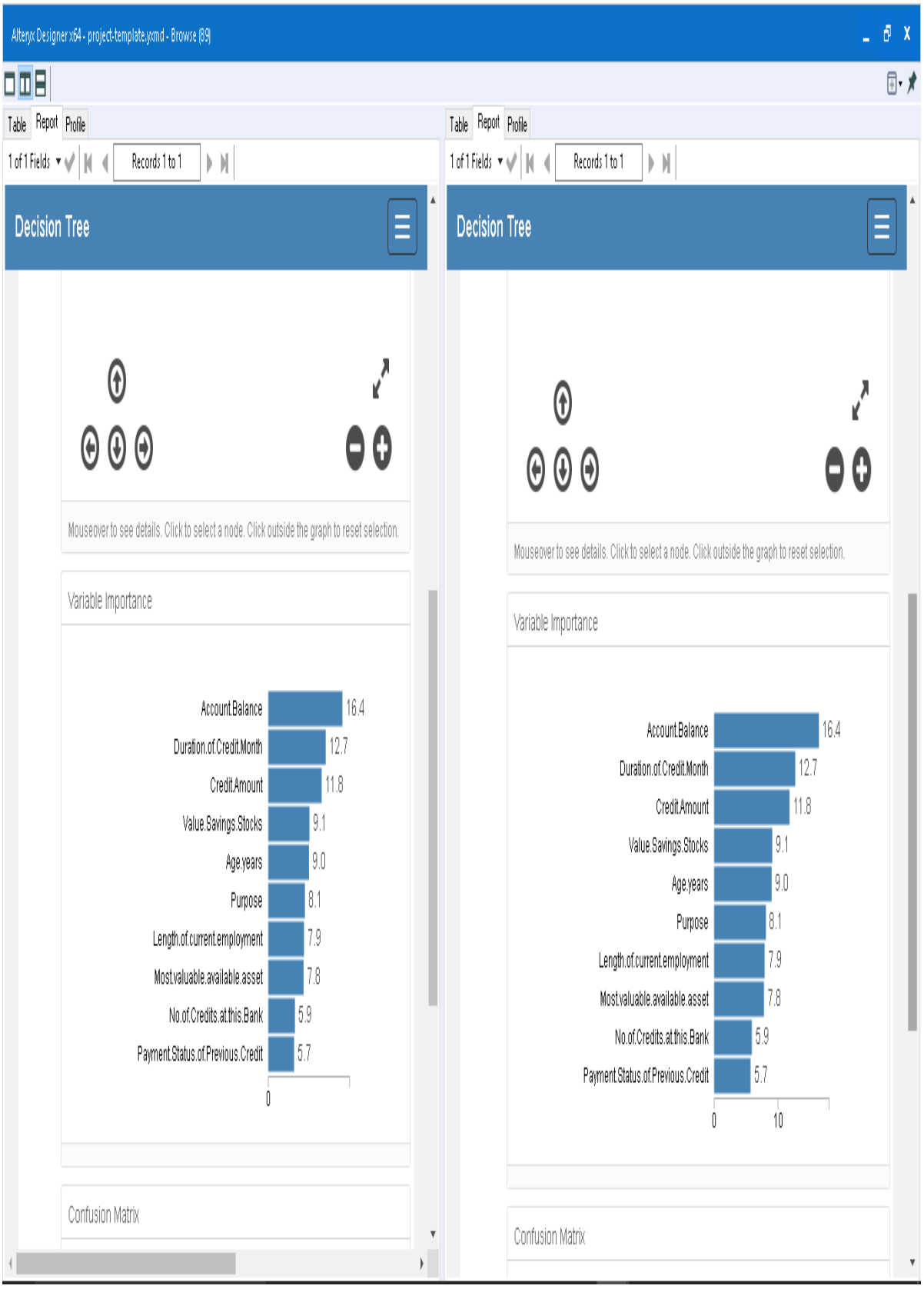
Confusion matrix of Forest_Loan

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

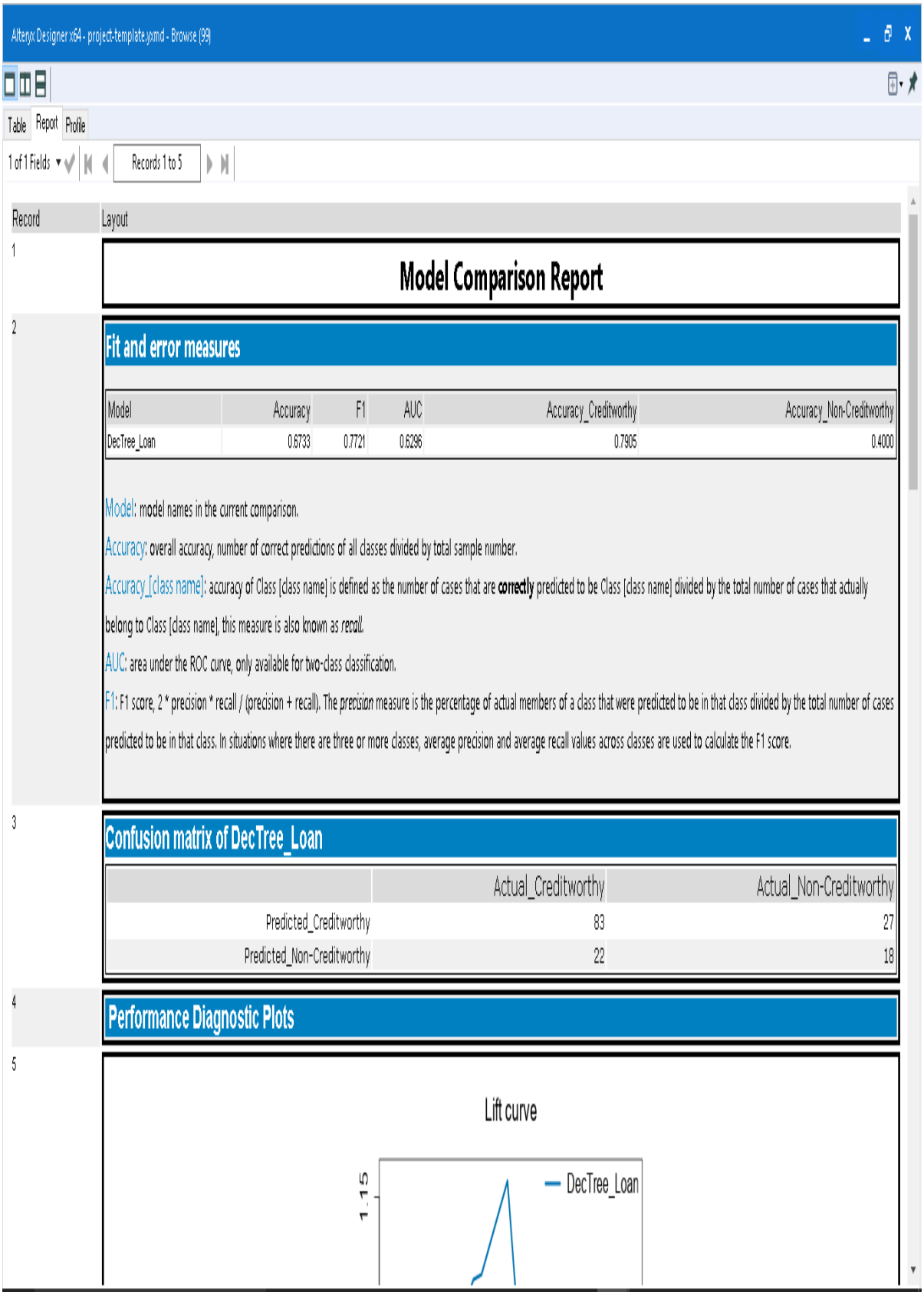
Performance Diagnostic Plots



Decision Tree Model- The most important predictor variables are **Account Balance, Duration.of.Credit.Month, Credit Amount,** Value.Savings.Stocks, **Age-years**, **Purpose**, **Length of current employment**, Most.valuable.available.asset, No.of.Credits.at.this.Bank, **Payment Status of Previous Credit.**


Alteryx Designer x64 - project-template.yxmd - Browse (89)
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Decision Tree
Mouseover to see details. Click to select a node. Click outside the graph to reset selection.
Variable Importance
Account.Balance 16.4
Duration.of.Credit.Month 12.7
Credit.Amount 11.8
Value.Savings.Stocks 9.1
Age.years 9.0
Purpose 8.1
Length.of.current.employment 7.9
Most.valuable.available.asset 7.8
No.of.Credits.at.this.Bank 5.9
Payment.Status.of.Previous.Credit 5.7
0
10
Confusion Matrix

Overall % Accuracy for Decision Tree Model is 67.33%. The model is biased as it predicts the Creditworthy class better (83 True Positives, 27 False Positives) than it does the Non-Creditworthy class (18 True Negatives and 22 False Negatives).


Alteryx Designer x64 - project-template.yxmd - Browse (99)
Table
Report
Profile
1 of 1 Fields
Records 1 to 5
Record
Layout
1
Model Comparison Report
2
Fit and error measures
Model
Accuracy
F1
AUC
Accuracy_Creditworthy
Accuracy_Non-Creditworthy
DecTree_Loan
0.6733
0.7721
0.6296
0.7905
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.
3
Confusion matrix of DecTree_Loan
Actual_Creditworthy
Actual_Non-Creditworthy
Predicted_Creditworthy
83
27
Predicted_Non-Creditworthy
22
18
4
Performance Diagnostic Plots
5
Lift curve
DecTree_Loan

Boosted Model- The most important predictor variables are **Credit Amount, Account Balance, Duration.of.Credit.Month**, **Purpose**, **Payment Status of Previous Credit, Age-years,** Most.valuable.available.asset, Value.Savings.Stocks, **Instalment.per.cent, Length of current employment.**

Alteryx Designer x64 - project-template.yxmd - Browse (30)

Table | Report | Profile

1 of 1 Fields | Records 2 to 3

Record	Report
2	Plots:

The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the

Overall % Accuracy for Boosted Model is 78.67%. The model predicts the Creditworthy class correctly 101 times with 28 False Positives and predicts the Non-Creditworthy class correctly 17 times with 4 False Negatives hence it does fairly well.

Alteryx Designer x64 - project-template.yxmd - Browse (99)

Table | Report | Profile

1 of 1 Fields | Records 1 to 5

Record	Layout
1	**Model Comparison Report**
2	**Fit and error measures**

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Loan	0.7867	0.8632	0.7524	0.9819	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*.

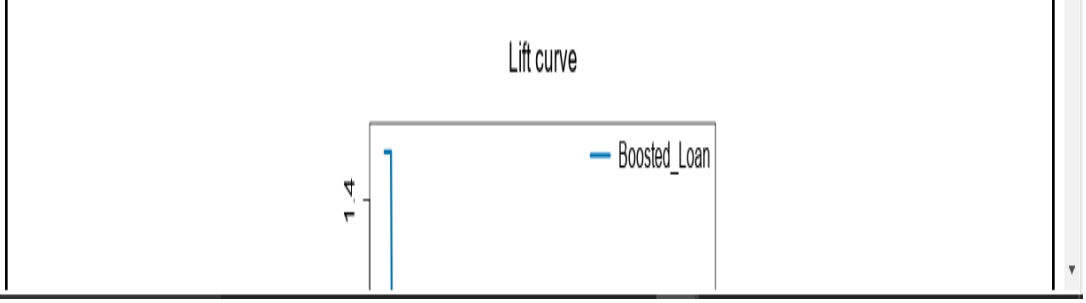
AUC: area under the ROC curve, only available for two-class classification.

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 Boosted_Loan

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

Performance Diagnostic Plots



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 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

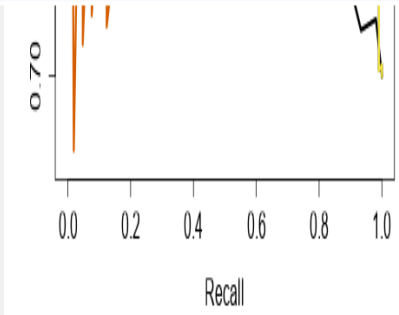
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

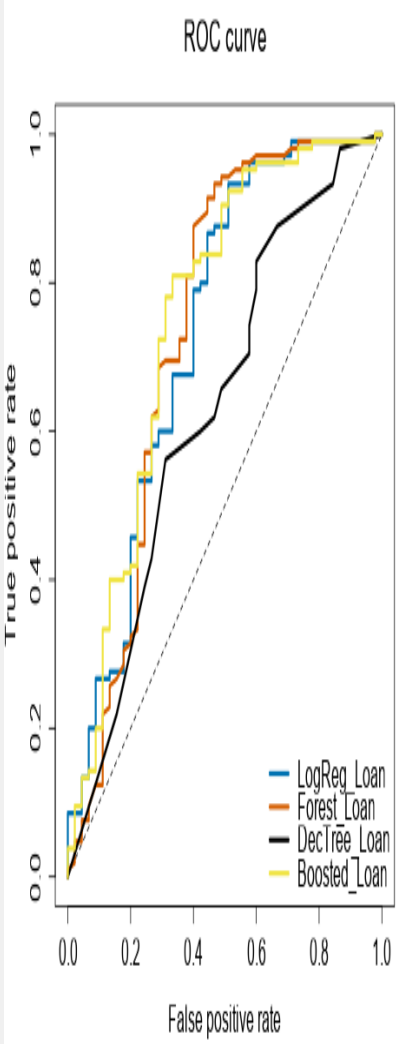
Ans: I chose the Logistic Regression Model as it has a pretty high Accuracy of 78%. The Accuracy for the Creditworthy class is also high with 90.48% and the accuracy for the Non-Creditworthy class is the highest with accuracy of 48.89%. The ROC Curve for the model gives fair results compared to other models.

Alteryx Designer x64 - project-template.yxmd - Browse (93)

Table | Report | Profile

1 of 1 Fields | Records: 1 to 8





- How many individuals are creditworthy?
  401 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.