Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

What decisions needs to be made?

The decisions that need to be made are:

We need to know whether or not to approve a loan application.
We also need to know which model is the best for deciding this.

What data is needed to inform those decisions?

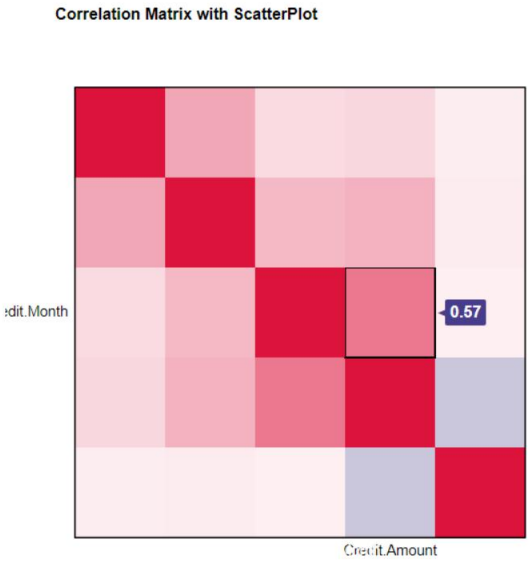
We need historic data on loan applications and the outcome of each loan application. This data will need to include enough information to create variable to be used to train the model.
We also need the same data for the new applications.

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

We need to use a Binary model as the main question to answer is yes or no on whether or not to approve a loan.

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”.

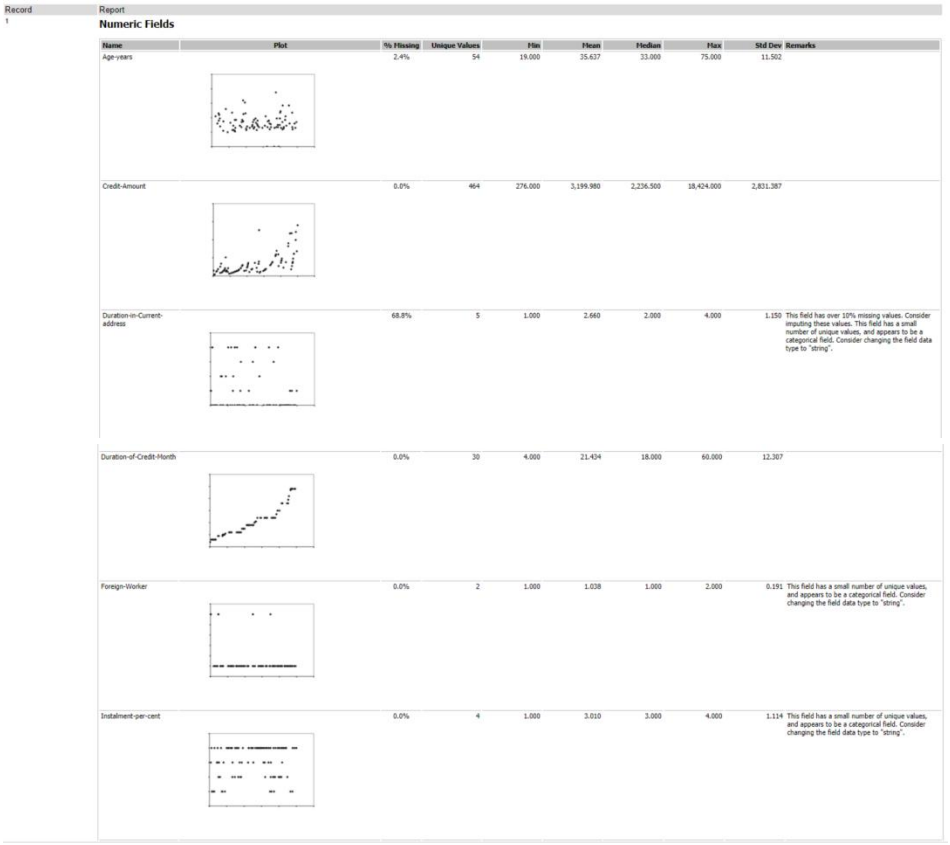


Based on association analysis, no two fields are highly correlated, I.e. >0.7.

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

Duration.in.Current.address has 69% missing values so this field will be removed.
Age-years has 2% missing values so will not be removed but rather imputed with the median Age-years replacing the 2% missing values.

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.

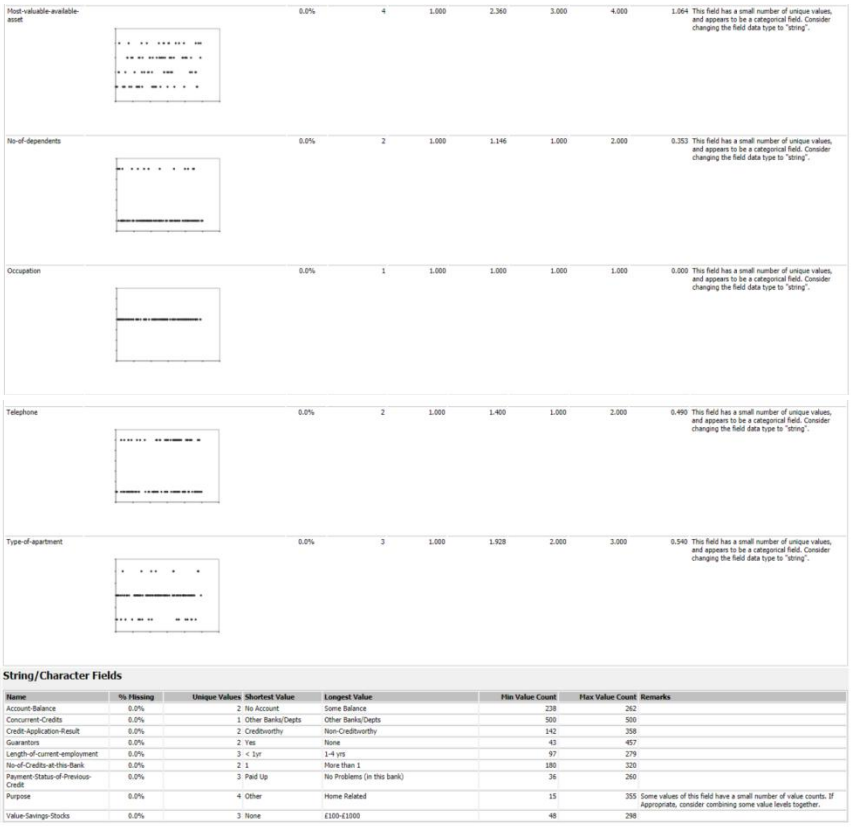


Record | Report

1

Numeric Fields

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Credit-Amount		0.0%	464	276.000	3,199.980	2,236.500	18,424.000	2,831.387	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Duration-of-Credit-Month		0.0%	30	4.000	21.434	18.000	60.000	12.307	
Foreign-Worker		0.0%	2	1.000	1.038	1.000	2.000	0.191	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Instalment-per-cent		0.0%	4	1.000	3.010	3.000	4.000	1.114	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".



Host-valuable-available-asset	0.0%	4	1.000	2.360	3.000	4.000	1.064	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
No-of-dependents	0.0%	2	1.000	1.146	1.000	2.000	0.353	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Occupation	0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Telephone	0.0%	2	1.000	1.400	1.000	2.000	0.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Type-of-apartment	0.0%	3	1.000	1.928	2.000	3.000	0.540	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

2

String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Account-Balance	0.0%	2	No Account	Some Balance	238	262	
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts	500	500	
Credit-Application-Result	0.0%	2	Creditworthy	Non-Creditworthy	142	358	
Guarantors	0.0%	2	Yes	None	43	457	
Length-of-current-employment	0.0%	3	< 1yr	1-4 yrs	97	279	
No-of-Credits-at-this-Bank	0.0%	2	1	More than 1	180	320	
Payment-Status-of-Previous-Credit	0.0%	3	Paid Up	No Problems (in this bank)	36	260	
Purpose	0.0%	4	Other	Home Related	15	355	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Value-Savings-Stocks	0.0%	3	None	£100-£1000	48	298	

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

The following data fields have low variability and will be removed:
Foreign worker, no of dependents, occupation, concurrent credits and guarantors.

Telephone has also been removed based on its irrelevance to the target.

Age-years has been imputed with the median Age-years due to the small number of missing values.

In total I have removed the following seven variables:
Duration in current address, Foreign worker, no of dependents, occupation, concurrent credit, guarantors and telephone.

Step 3: Train your Classification Models

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

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_years	-0.0141206	1.535e-02	-0.9202	0.35747	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

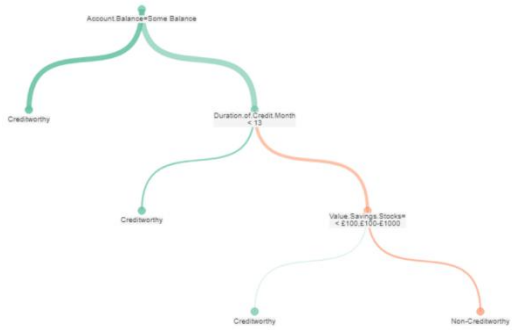
(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3

Number of Fisher Scoring iterations: 5

Using logistic regression the following variables were found to be statistically significant.
Account.Balance
Payment.Status.of.Previous.Credit
Purpose
Credit.Amount
Length.of.current.employment
Installment.per.cent
Most.valuable.available.asset

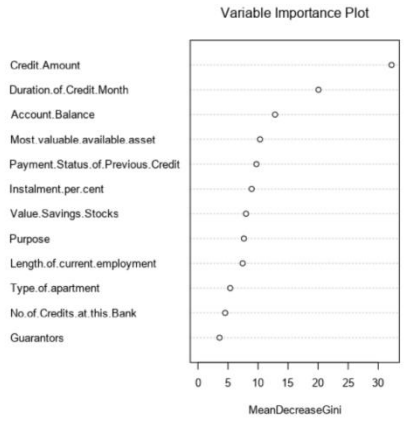
Decision Tree



Using a decision tree the following variables were found to be important:
Account.Balance
Duration.of.Credit.Month
Value.Savings.Stocks

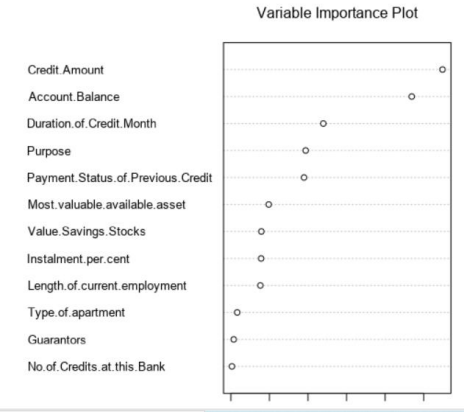
Added Credit.Amount due to importance in Decision Forest and Boosted models.

Decision Forest



Using a decision forest model, all variables were found to be important to the model. However, the top three were:
Credit.Amount
Duration.of.Credit.Month
Account.Balance

Boosted Model

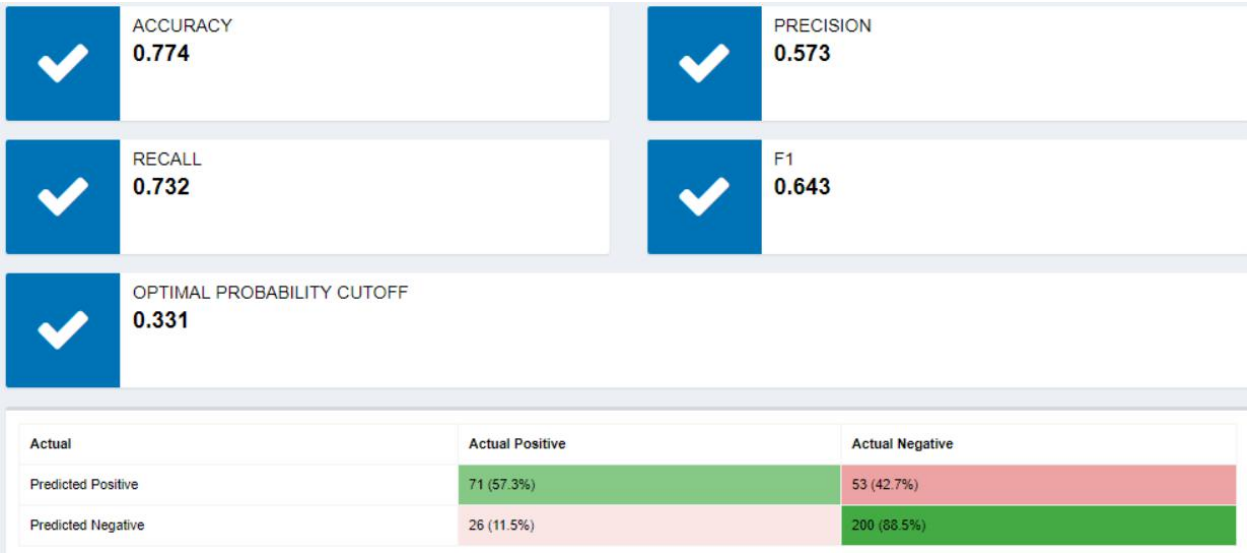


Using a boosted model, all variables contributed however the following three were the most important:
Credit.Amount
Account.Balance
Duration.of.Credit.Month

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?

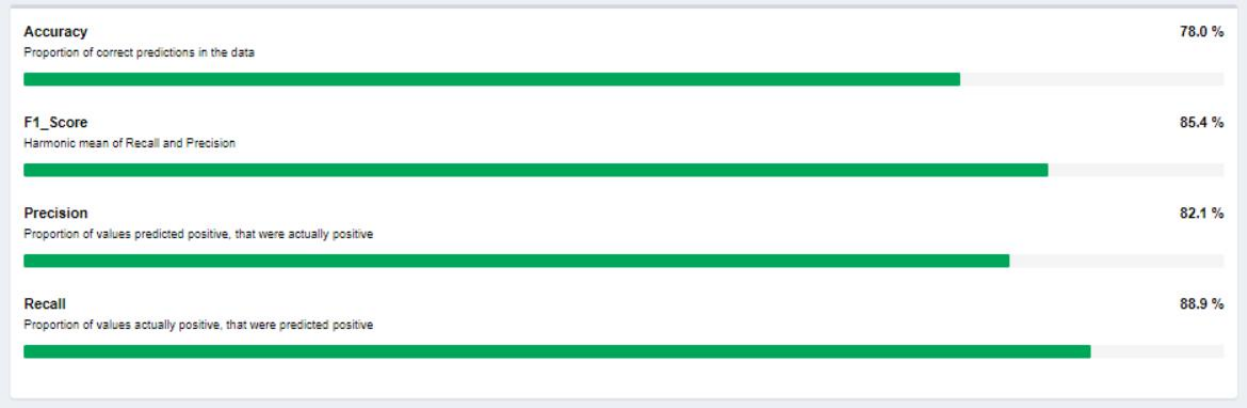
Using the statistically significant variables, a model was created of each of the four types below and tested against the 30% validation set.

Logistic Regression:



Actual	Actual Positive	Actual Negative
Predicted Positive	71 (57.3%)	53 (42.7%)
Predicted Negative	26 (11.5%)	200 (88.5%)

Decision Tree:



Actual	Actual Positive	Actual Negative
Predicted Positive	48 (49.5%)	49 (50.5%)
Predicted Negative	28 (11.1%)	225 (88.9%)

Decision Forest:

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.639	62	35

Boosted Model:

Confusion matrix of Boosted_Model

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

All models are bias toward Actual_Negative showing much higher accuracy when predicting Negative values while the accuracy when predicting Positive values is much lower.

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"

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_14	0.7467	0.8273	0.7054	0.8667	0.4667
Decision_Forest	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778
Logistic_Regression_32	0.7800	0.8520	0.7314	0.9048	0.4889

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Forest

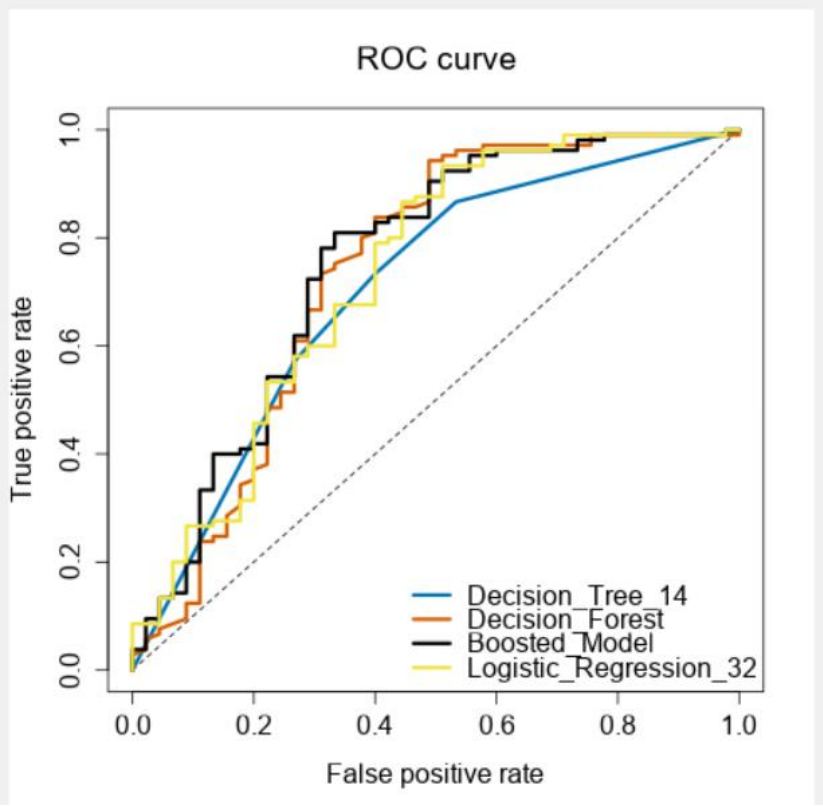
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Decision_Tree_14

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Logistic_Regression_32

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



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:

The model I have chosen to use is the forest model. This model has the highest overall accuracy, the third highest AUC from the ROC curve, the second highest accuracy when predicting non-creditworthy and the highest accuracy when predicting creditworthy applicants.

Overall Accuracy against your Validation set

Overall the decision forest model was the most accurate with an overall accuracy of 80.00%. The decision tree, boosted and logistic regression models were 74.67%, 78.67% and 78% respectively.

Accuracies within "Creditworthy" and "Non-Creditworthy" segments

All models showed a higher accuracy predicting creditworthy with accuracies higher on Actual_Creditworthy applicants. The logistic regression model had the highest accuracy when predicting negative applicants with an accuracy of 48.89%.

ROC graph

The ROC graphs shows that the boosted model is the best model when using ROC as a metric. The AUC of the boosted model is 75.24% which is higher than the other models.

Bias in the Confusion Matrices

All models show a bias toward predicting creditworthy applicants which is a bit of a worry as this means several actual non-creditworthy applicants are being classified as creditworthy. The boosted model is the most biased while the other models are all similarly biased.

How many individuals are creditworthy?

I have decided on a Forest model as the best model I retrained the Forest model on the entire dataset and ran the new data through.
Based on the new model with the new data, **408** individuals are creditworthy.

Alteryx Workflow