Project: Creditworthiness

Step 1: Business and Data Understanding

A small bank responsible for determining the creditworthiness of customers to receive loans suddenly experience an influx of new customers applying for loans. To meet up with the increased demand, the bank chose to use this opportunity to process nearly 500 loans in one week using analysis.

Key Decisions:

- What decisions needs to be made?
  The bank needs to decide which of the 500 loan applications are credit worthy for loans within one week. This new business opportunity for the bank can only be achieved by taking up a process that can achieve loan application processes within one week.
- What data is needed to inform those decisions?
  To make the necessary decisions the data needed for the analysis are:
  Data on all past applications ***credit-data-training.xlsx;*** which contains approval of past loan applicants the bank has ever completed – this file is needed to build a suitable model that can make the needed predictions.
  Data on the new set of customers ***customers-to-score.xlsx;*** this file contain the 500 new customers applying for loans, the model built will be scored on this file to predict their creditworthiness.
- What kind of model? Based on the expected result of this analysis a binary classification model will be suitable as we are classifying the loan applications to individuals that are creditworthy and non-creditworthy.

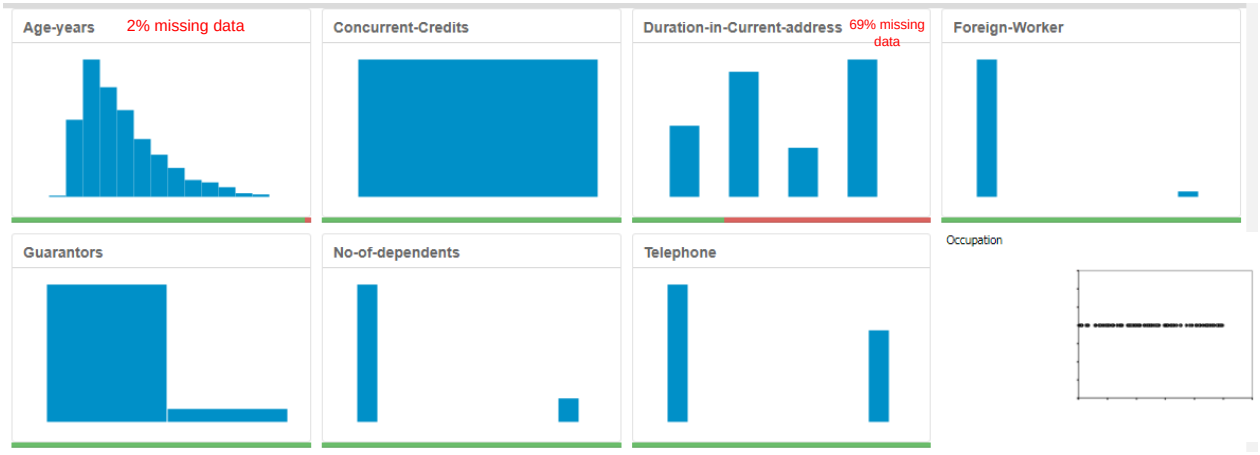
Step 2: Building the Training Set

- To build the credit-data-training.xlsx dataset some fields were removed and imputed; no fields correlated at 0.70. Table 1&Figure 1. below shows these:

Table 1.

COLUMN FIELDS REMOVED/IMPUTED	REASON
Removed: Duration-in-current-address	69% Missing data.
Removed: Guarantors, Foreign worker, No-of-dependents, Concurrent-Credits, and Occupation.	Low Variability
Removed: Telephone	Not useful as predictor variable
Imputed: Age-years	2% Missing Data imputed by median value (33) because the distribution is skewed.

Figure 1.



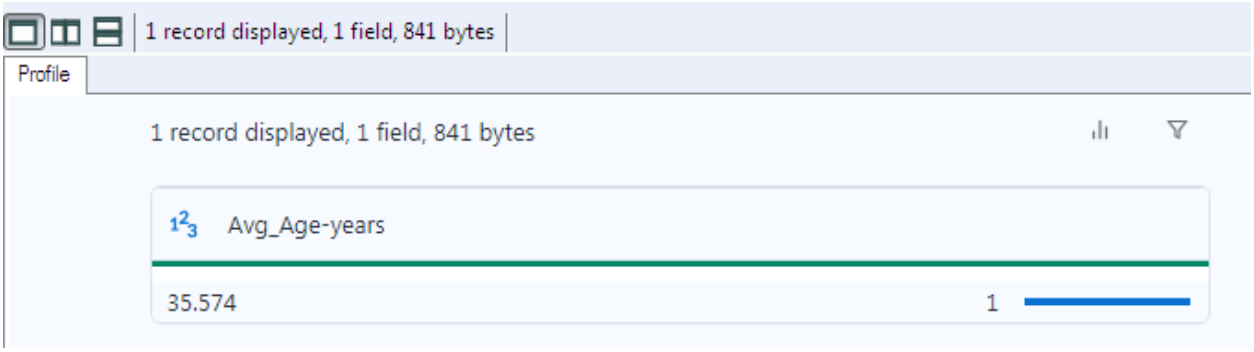
There were 13 columns left and mean of age-years is 35.574 approx. 36. As shown in Table 2 and 3. below:

Table 2.

13 of 13 Fields | Cell Viewer

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

Table 3.



Step 3: Train your Classification Models

Trained the *credit-data-training.xlsx* dataset with sample creation tool, shared the data into 70% for Estimation and 30% reserved for Validation.

Four classification models were applied to the estimation data **Logistic Regression, Decision Tree, Forest Model, Boosted Model.**

Logistic Regression: The determined predictor variables significant to the target variable (**Credit.Application.Result**) are:

Account.Balance (p-value: 1.65e-07***), Payment.Status.of.Previous.CreditSome Problems (p-value: 0.0183*), PurposeNewCar (p-value: 0.00566**), Credit.Amount (p-value: 0.00296**), Length.of.current.employment < 1yr (p-value: 0.03596*), and Instalment.per.cent (p-value: 0.02549*).

Table 4: p_values of the logistic regression stepwise result.

Report for Logistic Regression Model CW_Step

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

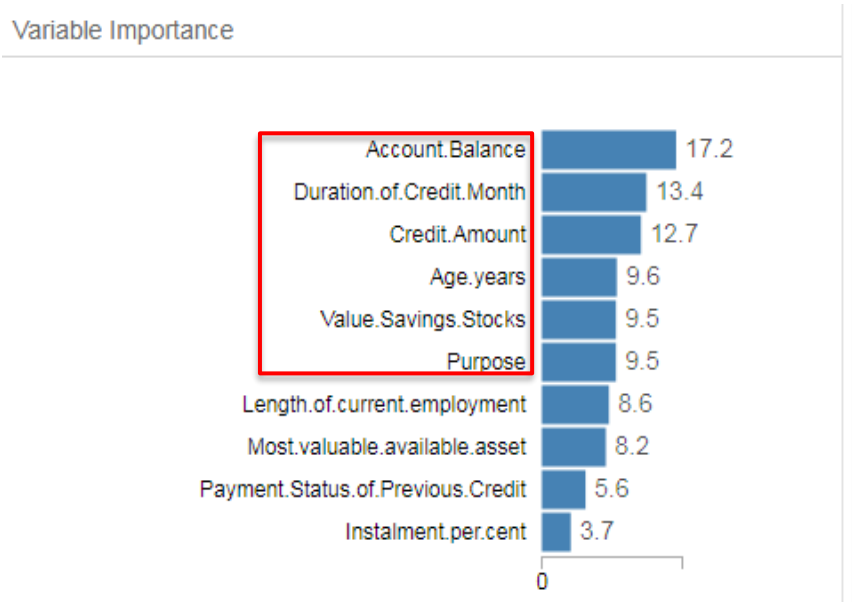
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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	.

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

(Dispersion parameter for binomial taken to be 1)

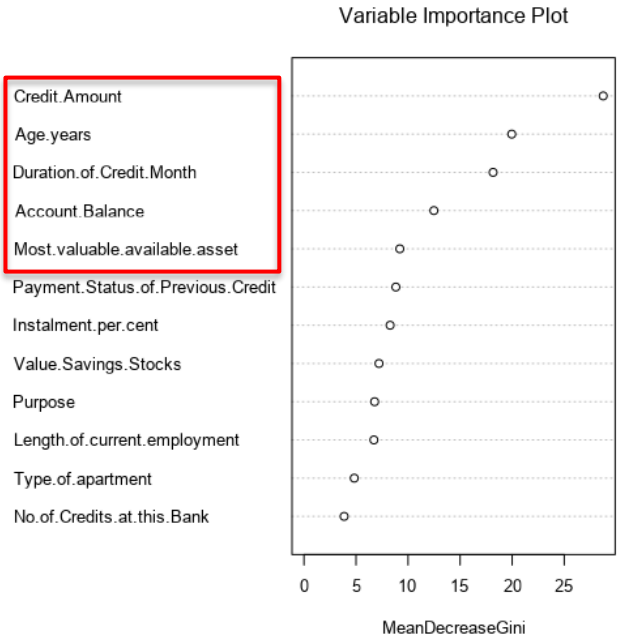
Decision Tree: The determined predictor variables that are most significant to the target variable (**Credit.Application.Result**), as shown by the variable comparison plot are: Account.Balance (17.2), Duration.of.Credit.Month (13.4), Credit.Amount (12.7), Age.years (9.6) Value.Savings.Stocks (9.5), Purpose (9.5).

Figure 2: Decision tree model variable importance plot.



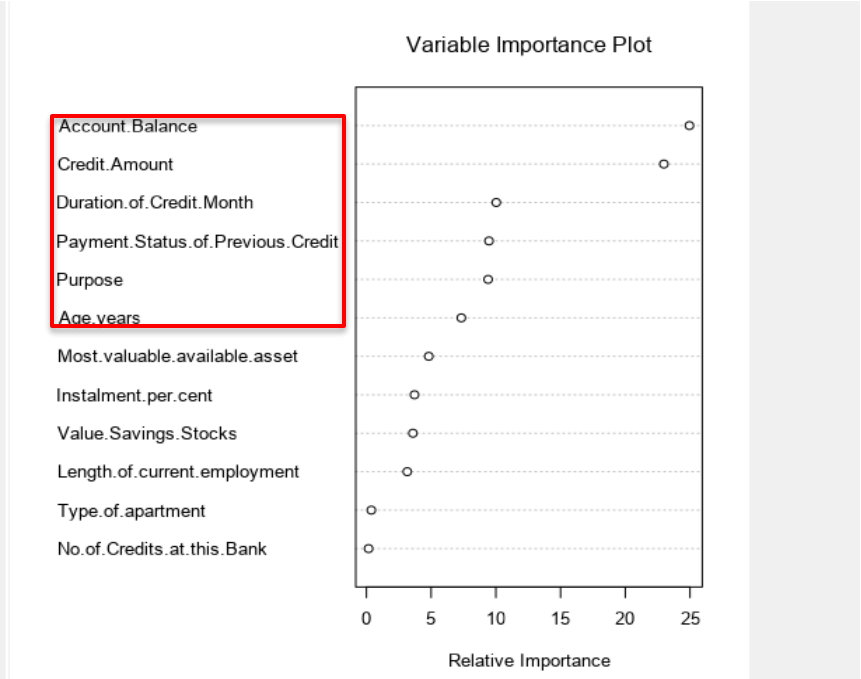
Forest Model: The most important predictor variables significant to the target variable (**Credit.Application.Result**) are:
Credit.Amount, Age.years, Duration.of.Credit.Month, Account.Balance, Most.valuable,available. asset.

Figure 3: The forest model variable importance plot.



Boosted Model: The determined predictor variables that are most significant to the target variable (**Credit.Application.Result**), are:
Account.Balance, Credit.Amount, Duration.of.Credit.Month, Payment.Staus.of.Previous.Credit, Purpose, Age.years.

Figure 4: The boosted model variable importance plot.



- **Model Validation:**

The models (Logistic Regression, Decision tree, and Forest, Boosted) were validated using the Model Comparison tool against the reserved 30% validation sample.

Table 5: The model with highest overall accuracy is Forest Model followed by Boosted Model and least is the Decision Tree model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_CW	0.6867	0.7854	0.6270	0.8190	0.3778
Forest_CW	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_CW	0.7867	0.8632	0.7515	0.9619	0.3778
CW_Step	0.7600	0.8364	0.7306	0.8762	0.4889

Table 6: Shows the confusion matrices of the four models using the model comparison tool.

Confusion matrix of Boosted_CW

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

Confusion matrix of CW_Step

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

Confusion matrix of Decision_Tree_CW

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	86	28
Predicted_Non-Creditworthy	19	17

Confusion matrix of Forest_CW

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

To calculate the bias of each model using the confusion matrix where:

Negative predictive value (NPV) = (No. of true negatives) / (true negatives + false negatives) i.e Predicted_Non-Creditworthy = Actual_Non-Creditworthy / (Actual_Creditworthy + Actual_Non-Creditworthy)

Positive predictive value (PPV) = (No. of true positives) / (true positives + false positives) i.e Predicted_Creditworthy = Actual_Creditworthy / (Actual_Creditworthy + Actual_Non-Creditworthy)

Logistic Regression (CW_Step):
NPV = 22/35 = 0.63 * 100 = 63%
PPV = 92/115 = 0.80 * 100 = 80%
From the calculation, we can see that the PPV (80%) is not close to the NPV (63%) which indicates that the logistic regression model is biased towards Creditworthy.

Decision Tree (Decision_Tree_CW):
NPV = 17/36 = 0.47 * 100 = 47%
PPV = 86/114 = 0.75 * 100 = 75%
From the calculation, we can see that the PPV (75%) is not close to the NPV (47%) indicating that the decision tree model is biased towards Creditworthy.

Forest Model (Forest_CW):
NPV = 17/20 = 0.85 * 100 = 85%
PPV = 102/130 = 0.78 * 100 = 78%
From the calculation, we can see that the PPV (78%) is close to the NPV (85%) indicating that the forest model is non biased towards Creditworthy.

Boosted Model (Boosted_CW):
NPV = 17/21 = 0.81 * 100 = 81%
PPV = 101/129 = 0.78 * 100 = 78%
From the calculation, we can see that the PPV (78%) is close to the NPV (81%) indicating that the boosted model is non biased towards Creditworthy.

Step 4: Writeup

From the overall accuracy report and confusion matrix of the model comparison report of all the models. I chose the **Forest model** for the following reasons:

Overall Accuracy: The forest model overall accuracy against the validation dataset is the highest (0.79) of all indicating it is a strong predictive model.

Accuracy with Creditworthiness: The creditworthiness accuracy (0.97) is significantly better for the Forest Model and Accuracy for the non-creditworthy (0.38) is similar to others but F1 score (0.87) is highest amidst the other models closely followed by boosted model.

ROC graph: For the Forest model, there is a higher area under the ROC curve as compared to the other models and higher F1 score, which is the weighted average of Recall and Precision. Below is Figure 5. (ROC graph) and Figure 6. (Precision and Recall Curve):

Figure 5.

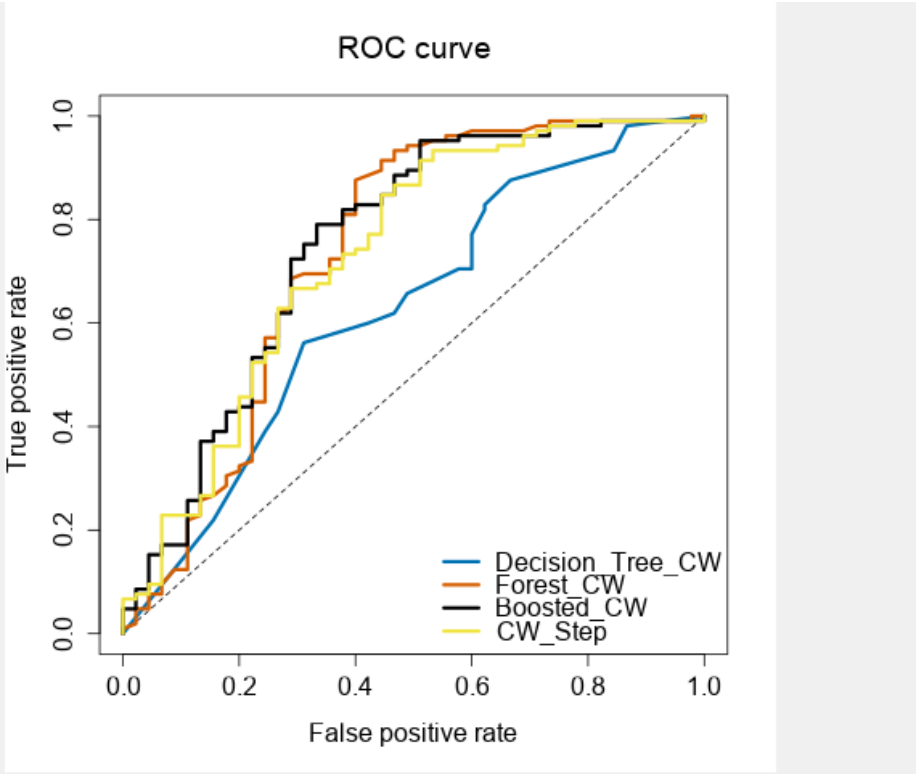
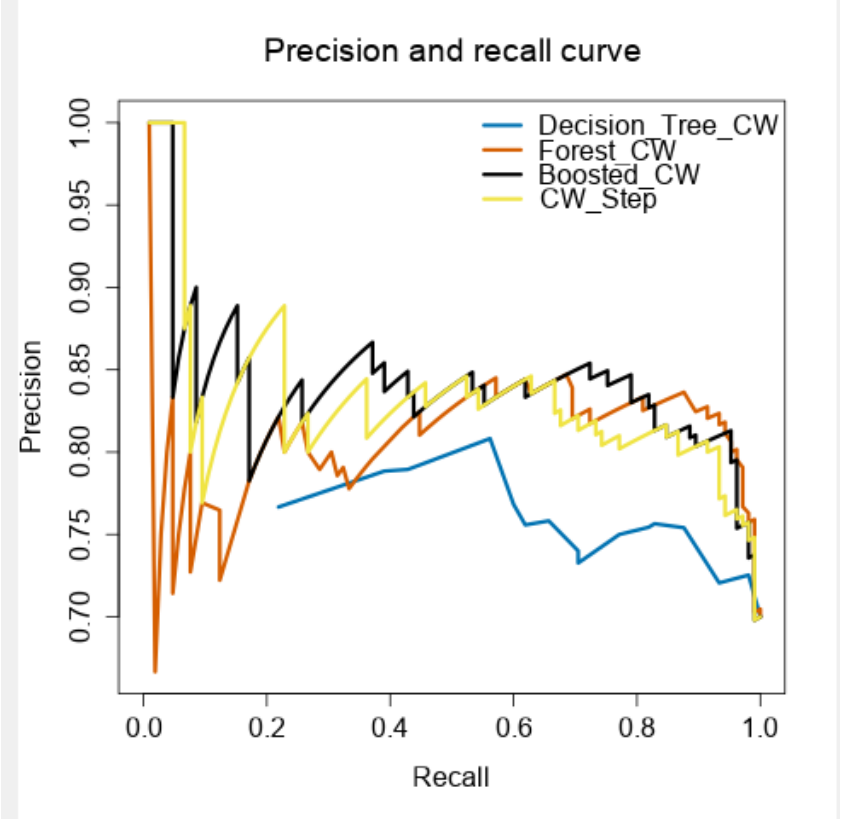


Figure 6.

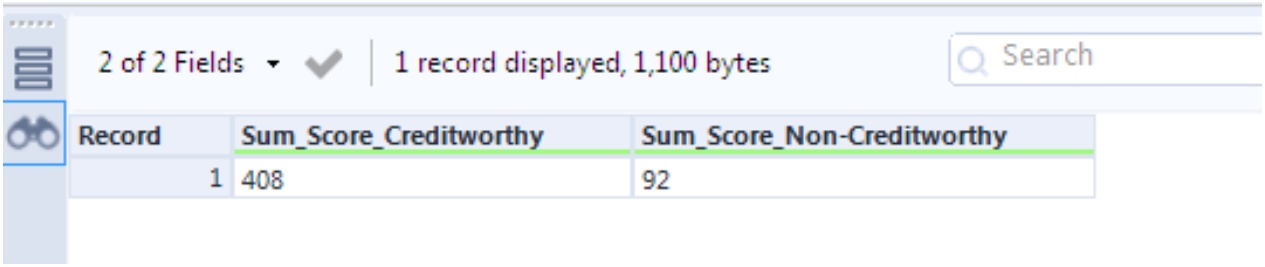


Bias in the Confusion matrices: The difference between PPV (0.78) and NPV (0.85) is small, thus, the forest model is not biased and so it is an accurate predictive model. The prediction accuracy for the credit worthy and non-creditworthy are not biased therefore accurate.

- **Scoring the model**

After scoring the forest model on the new dataset (customers-to-score.xlsx) the number of individuals found to be creditworthy were **408.** As shown in Table 7 below:

Table 7.



Record	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	408	92