Project 3: Predicting Default Risk | Ioanna Vasilopoulou

Step 1: Business and Data Understanding

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

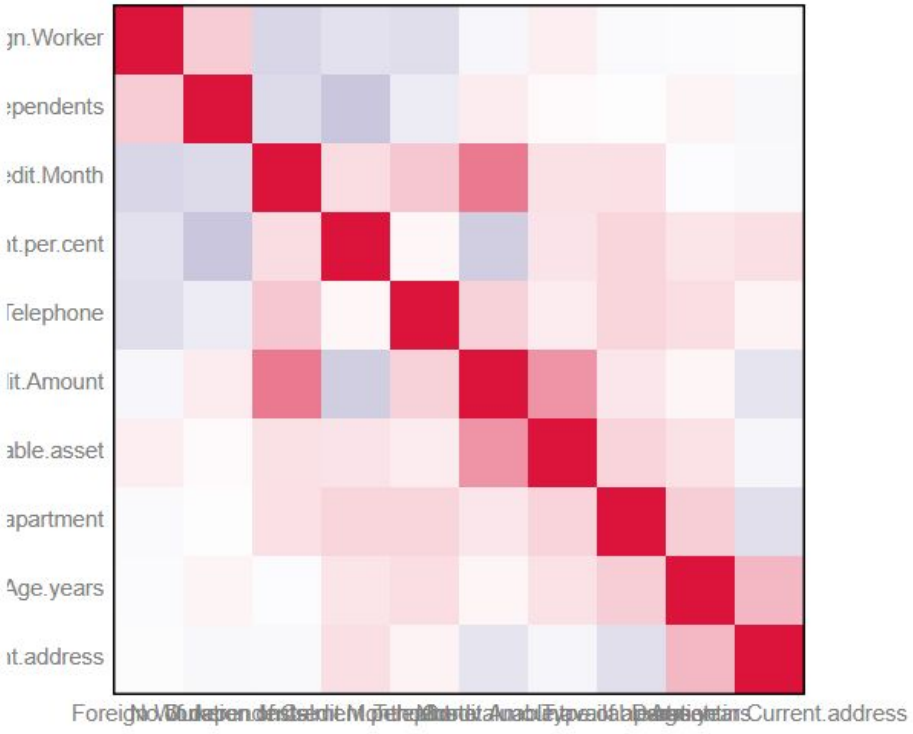
Key Decisions:

- ***What decisions needs to be made?***
  We need to determine if customers who apply for a loan are creditworthy to give a loan to.
- ***What data is needed to inform those decisions?***
  - Data on all past applications
  - List of customers
- ***What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?***
  In order to analyse and determine which clients are creditworthy we will use:
  - Logistic Regression
  - Decision Tree
  - Forest Model
  - Boosted Tree

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".**

  According to the association analysis performed (Correlation Matrix with ScatterPlot, below) for the numerical data fields, there are no fields that highly-correlate (correlation ≧ 0.70).



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

According to our fields summary below, we can see that:

- **Age-years:** 2% missing data. This 2% of missing data should be replaced. As the data is is skewed to the left it would be better to replace the missing data with "middle" value (median), rather than the average age (mean). Median is 33.

Record #	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295

- **Duration-in-Current-address:** 69% missing data. This data should be removed as this is an irrelevant factor to the potential customer's creditworthiness.

- **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.**

Further observations:

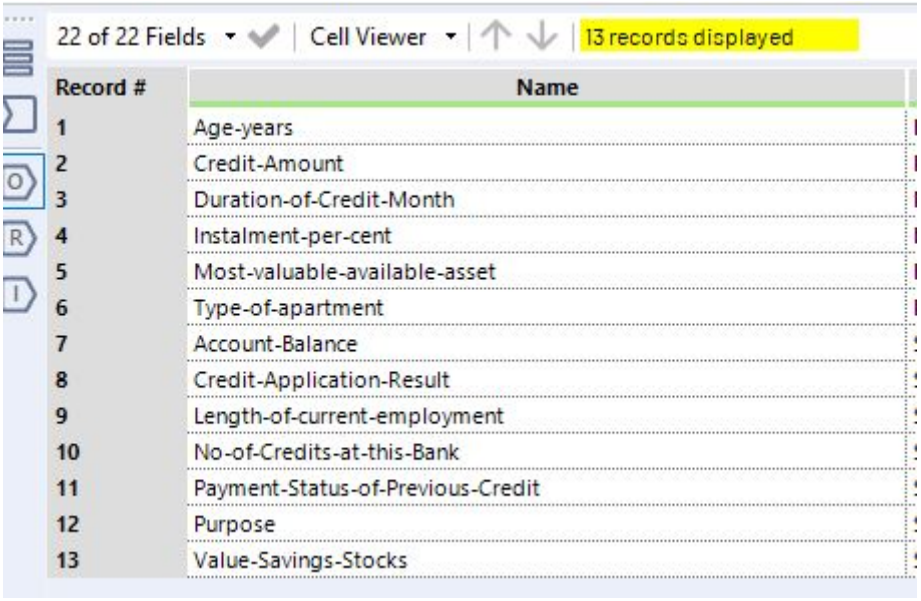
- **Concurrent-credits:** There is only 1 value
- **Occupation:** There is only 1 value
- **Foreign-worker:** There are only 2 values, one of which has 96.2% of the data behind it.
- **Guarantors:** There are only 2 values, one of which has 91.4% of the data behind it.
- **No-of-dependents:** There are only 2 values, one of which has 85.4% of the data behind it.
- **Telephone:** This data is irrelevant with the decision we need to make which is whether a customer is creditworthy or not.

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

Record #	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295

Out of the 20 columns that were initially included in the file, and after our decisions above to not include specific

Columns in our final and clean data set, we are left with a total of 13 columns, which are:



Record #	Name
1	Age-years
2	Credit-Amount
3	Duration-of-Credit-Month
4	Instalment-per-cent
5	Most-valuable-available-asset
6	Type-of-apartment
7	Account-Balance
8	Credit-Application-Result
9	Length-of-current-employment
10	No-of-Credits-at-this-Bank
11	Payment-Status-of-Previous-Credit
12	Purpose
13	Value-Savings-Stocks

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

Fields removed:

- **Duration-in-Current-address:** 69% missing data. This data should be removed as this is an irrelevant factor to the potential customer's creditworthiness.
- **Concurrent-credits:** There is only 1 value
- **Occupation:** There is only 1 value
- **Foreign-worker:** There are only 2 values, one of which has 96.2% of the data behind it.
- **Guarantors:** There are only 2 values, one of which has 91.4% of the data behind it.
- **No-of-dependents:** There are only 2 values, one of which has 85.4% of the data behind it.
- **Telephone:** This data is irrelevant with the decision we need to make which is whether a customer is creditworthy or not.

Fields imputed:

- **Age-years:** We replaced the 2% of missing data with the median which is 33

Record #	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	Age-years	Numeric	19	75	33	11.501522	2.4	54	35.637295

Step 3: Train your Classification Models

Logistic Regression:

Report for Logistic Regression Model X

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)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

- **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**

  Our target variable is Credit Application Result; The most important variables are Account Balance, Purpose and Credit amount, as they all have a p-value which is less than 0.05.

- **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?**

  The overall accuracy is 76%. Accuracy for creditworthiness and non-creditworthiness is 87.62% and 48.89%, respectively. We can see that the model is biased towards predicting customers as non-creditworthy.

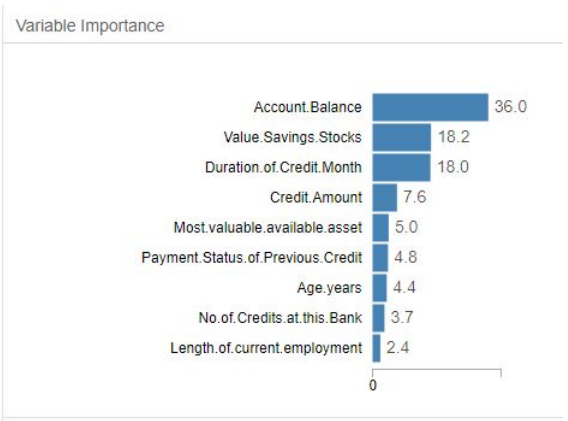
Fit and error measures

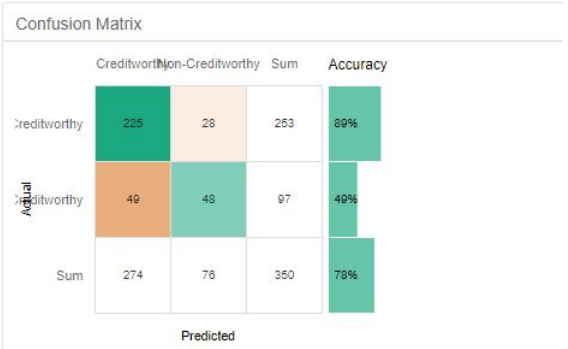
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Logistic

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

Decision Tree:





- **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**
  Having Credit Application Results as our target variable we can see that the 3 most important values are: Account Balance, Value Savings Stocks and Duration of Credit Month with an overall accuracy of 72.20%.

- **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?**
  Accuracy for creditworthy is 86.67% and accuracy for non-creditworthy is 46.67%. In the model we can see a bias towards predicting customers as non-creditworthy.

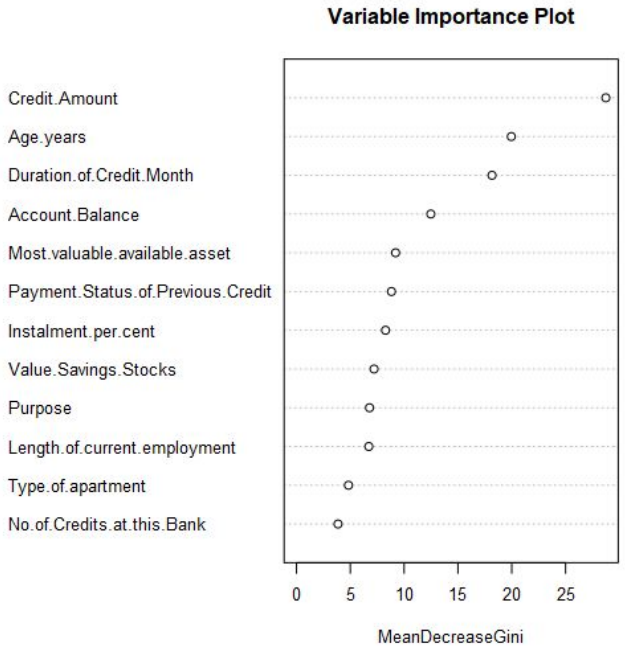
Fit and error measures

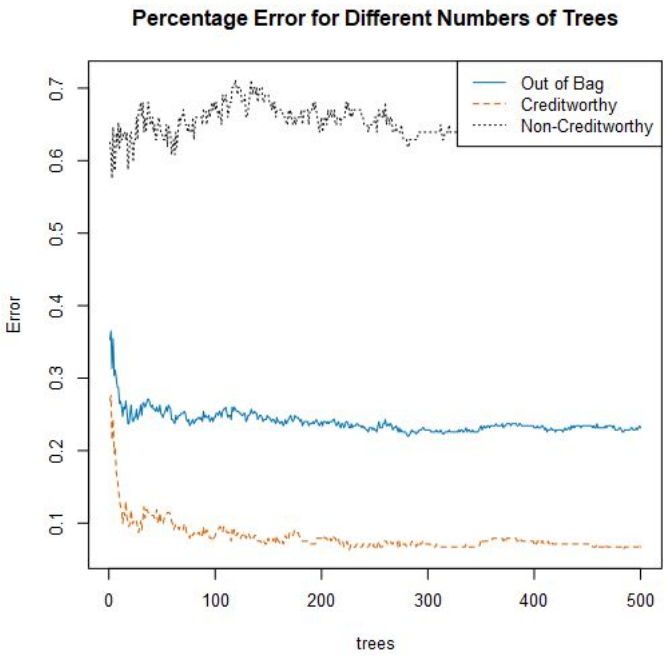
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision	0.7467	0.8273	0.7054	0.8667	0.4667

Confusion matrix of Decision

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

Forest Model:





- **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**
  Having Credit Application Results as our target variable we can see that the 3 most important values are: Credit Amount, Age Years and 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?**
  The overall accuracy is 79.33%. Accuracy for creditworthiness and non-creditworthiness is 97.14% and 37.78%, respectively. We can see that the model is biased towards predicting customers as non-creditworthy.

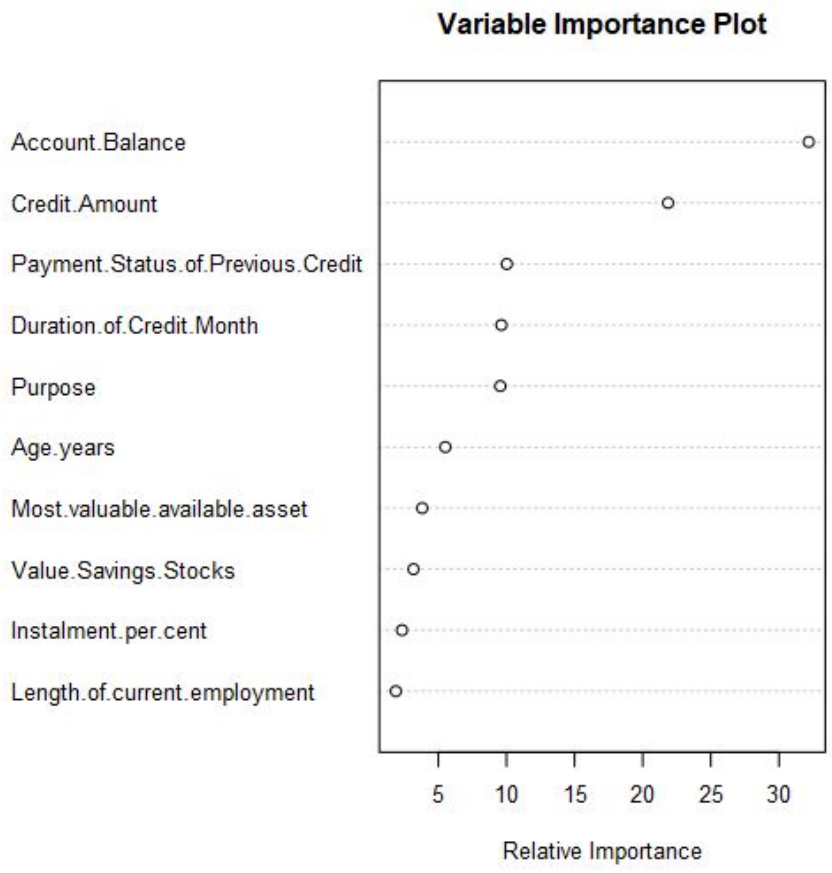
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix of Forest

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

Boosted Model:



- **Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.**
  Having Credit Application Results as our target variable we can see that the 3 most important values are: Account Balance, Credit Amount and Payment Status of Previous Credit

- **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?**
  The overall accuracy is 78.67%. Accuracy for creditworthiness and non-creditworthiness is 96.19% and 37.78%, respectively. We can see that the model is biased towards predicting customers as non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix of Boosted

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

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:**

  We have chosen Forest Model, as the its accuracy is the highest at 79.33%. The accuracy the model has in creditworthiness is the highest amongst all 4 models. Moreover, the Forest Model reached the True Positive Rate the quickest.

  - **Overall Accuracy against your Validation set**

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778
Decision	0.7467	0.8273	0.7054	0.8667	0.4667
Logistic_Regression	0.7800	0.8520	0.7314	0.9048	0.4889
Forest	0.7933	0.8681	0.7368	0.9714	0.3778

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

Confusion matrix of Boosted

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

Confusion matrix of Decision

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

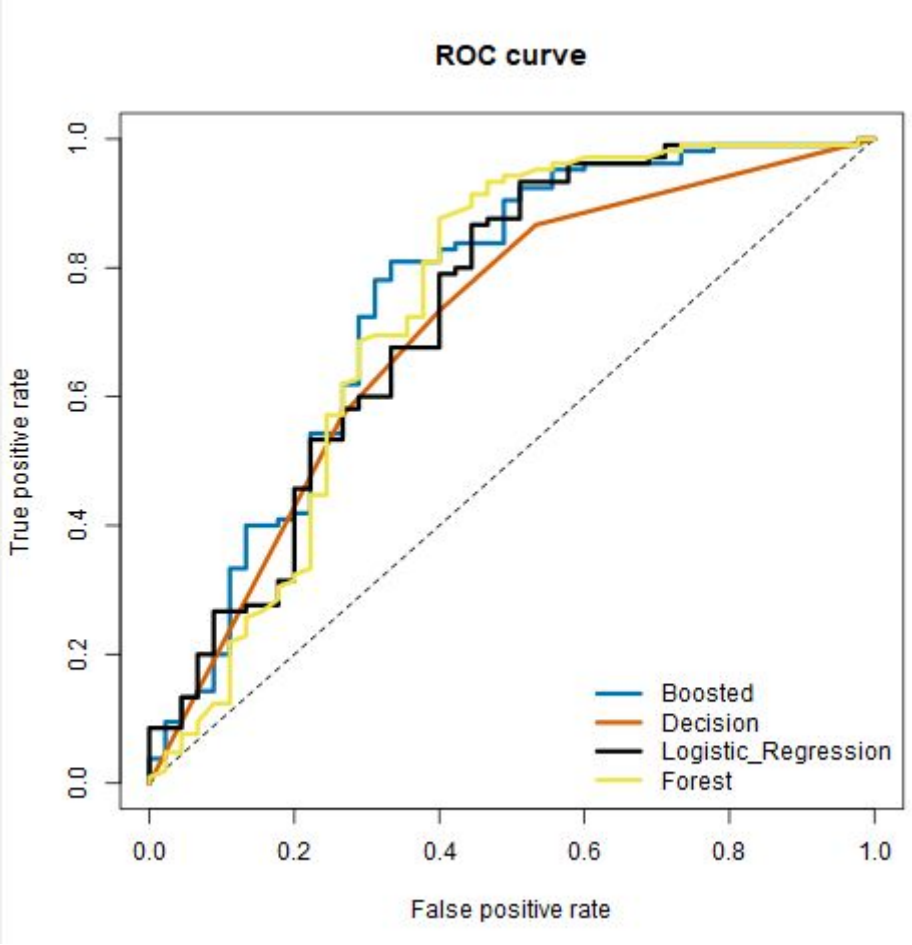
Confusion matrix of Forest

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

Confusion matrix of Logistic_Regression

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

- **ROC graph**



- Bias in the Confusion Matrices
  In the Confusion Matrices we can see that Forest Model Predicts better for both questions; Is a customer creditworthy or not?

- **How many individuals are creditworthy?**

Using the Forest model to scole the creditworthy customers, we found 410 creditworthy individuals.

Record #	Count
1	410