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?

  A small bank which typically gets 200 loan applications per week and generally approve by hand suddenly received nearly 500 loan applications due to a financial scandal that hit a competitive bank last week. As a loan officer, my manager wants me to figure out how to process all of these loan applications within a week. I need to systematically evaluate the creditworthiness of these new loan applicants and provide a list of creditworthy customers to my manager in the next two days. The decision at hand is determining which customers are creditworthy to give a loan to.

- What data is needed to inform those decisions?

  1. Data on all past applications

     They include 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.

  2. The list of customers that need to be processed in the next few days

     Their information including 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.

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

  We need to build a binary classification model which will predict two outcomes: whether a customer is creditworthy or not. To achieve this, we will be comparing the following models to select the most accurate model:

  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Boosted Model

  Accuracy is the top priority for the company as this will give information about creditworthy/uncreditworthy customers.

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:*

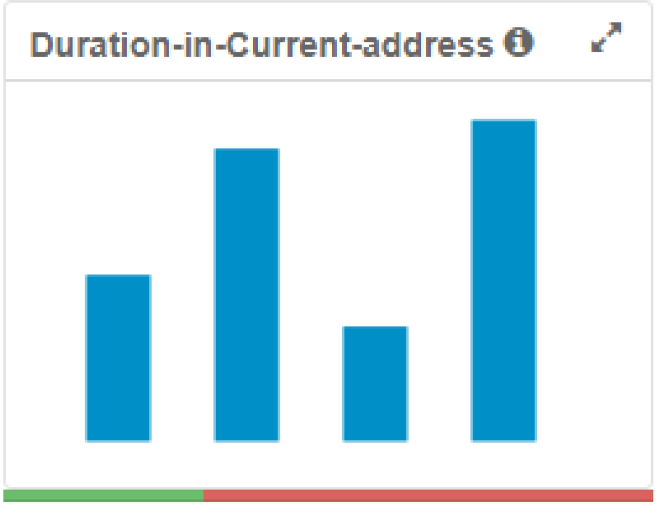
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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

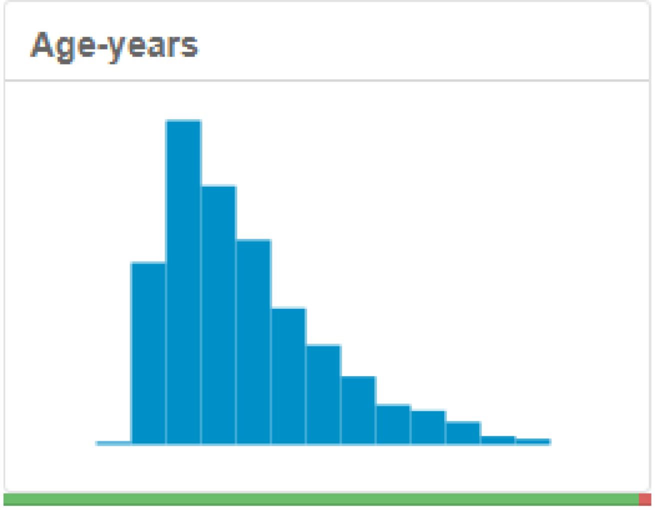
To achieve consistent results reviewers expect.

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.

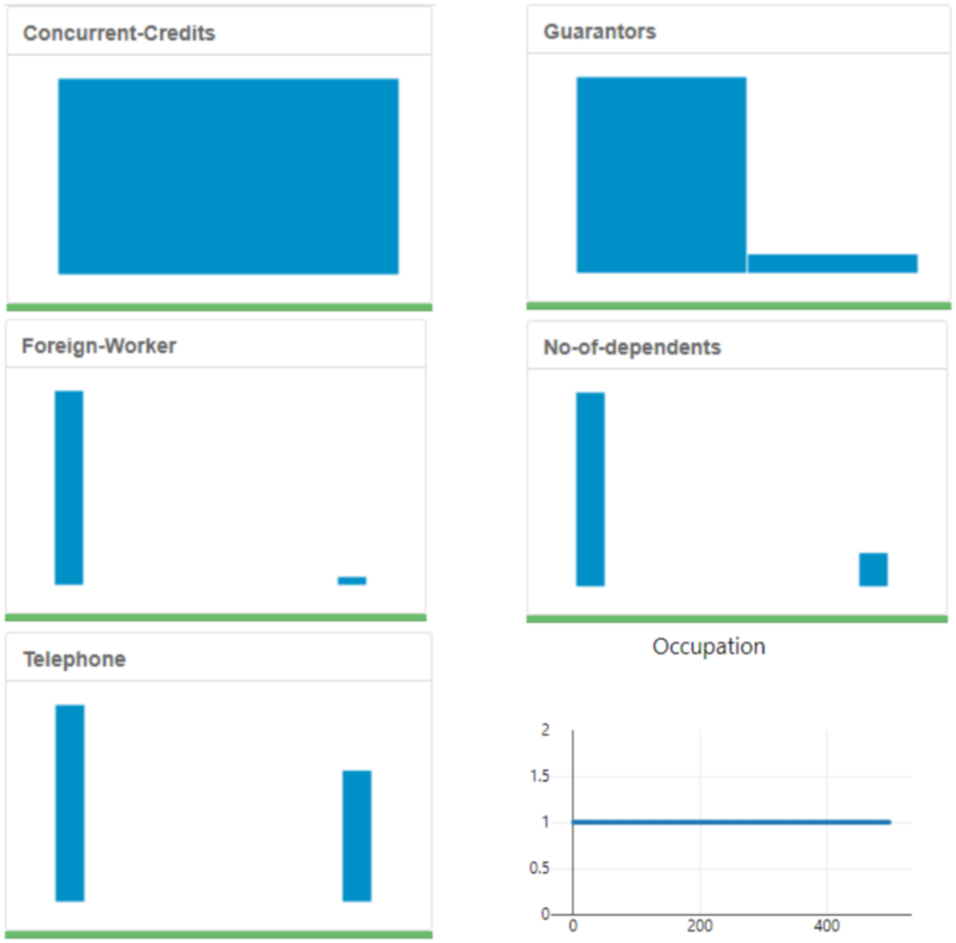
I used the Field Summary tool to peek into the dataset. I found that there were missing values in two fields: Duration in Current Address (68.8% missing data) and Age-Years (2.4% missing data). The following are the outputs from the field summary tools:





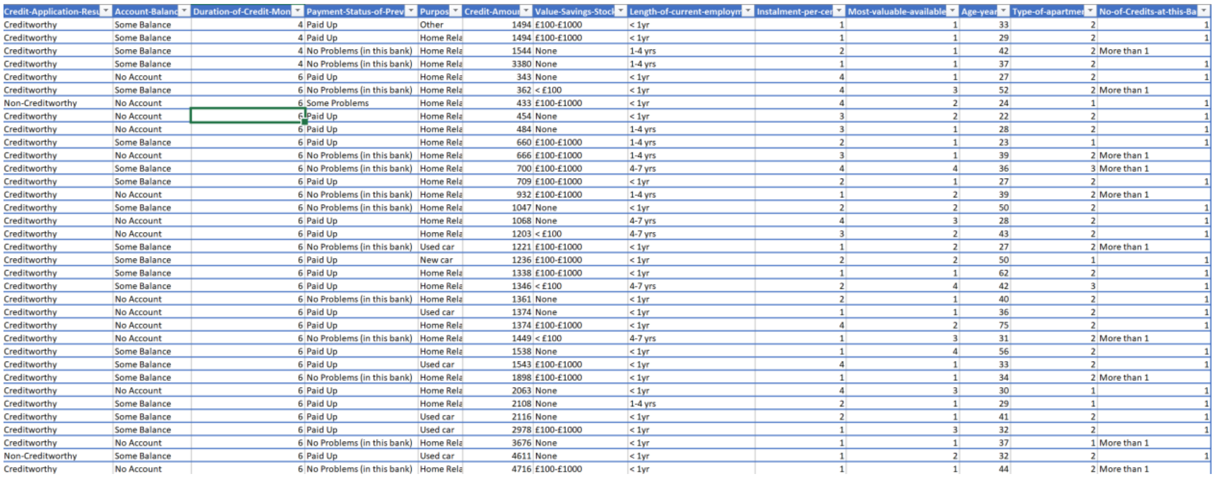
I removed the field "Duration in Current Address" due to large percentage of missing values since it would not mean any significance to the model. For Age-Years, we can see that the distribution of this data is right skewed, hence the imputing the missing values with median would provide a better estimation of the central tendency of the data. If our data distribution would be bell-shaped, imputing with average values would be a better representation of the central tendency of the data. For our case, median would represent age between the youngest and oldest groups.

After analysing the other variables, I found several variables in our dataset which demonstrated low variability. Below are variables which were removed on this criterion.



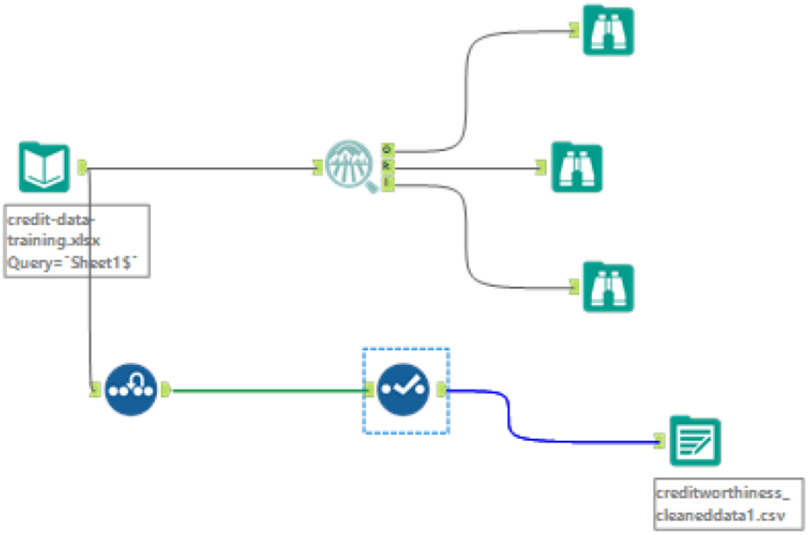
From above histograms, we can see that Concurrent credits have one uniform value, which will skew the data towards this value. Hence, it shows there is no variation in the data. Guarantors has skewed value towards None. The same holds true for No. of dependents, foreign workers and occupation. The telephone field is likely to not show any predictive power to determine the creditworthiness of a customer, hence this variable was also removed.

Below is the final cleaned dataset consisting of 13 columns:



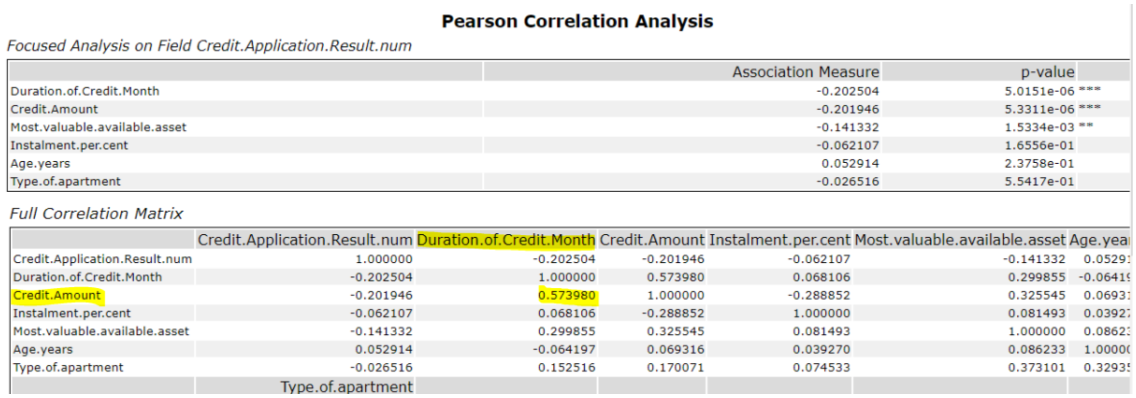
Credit-Application-Resu	Account-Balanc	Duration-of-Credit-Mon	Payment-Status-of-Prev	Purpos	Credit-Amour	Value-Savings-Stock	Length-of-current-employm	Instalment-per-cer	Most-valuable-available	Age-year	Type-of-apartmer	No-of-Credits-at-this-Ba
Creditworthy	Some Balance	4	Paid Up	Other	1494	£100-£1000	< 1yr	1	1	33	2	1
Creditworthy	Some Balance	4	Paid Up	Home Rela	1494	£100-£1000	< 1yr	1	1	29	2	1
Creditworthy	Some Balance	4	No Problems (in this bank)	Home Rela	1544	None	1-4 yrs	2	1	42	2	More than 1
Creditworthy	Some Balance	4	No Problems (in this bank)	Home Rela	3380	None	1-4 yrs	1	1	37	2	1
Creditworthy	No Account	6	Paid Up	Home Rela	343	None	< 1yr	4	1	27	2	1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Rela	362	< £100	< 1yr	4	3	52	2	More than 1
Non-Creditworthy	No Account	6	Some Problems	Home Rela	433	£100-£1000	< 1yr	4	2	24	1	1
Creditworthy	No Account	6	Paid Up	Home Rela	454	None	< 1yr	3	2	22	2	1
Creditworthy	No Account	6	Paid Up	Home Rela	484	None	1-4 yrs	3	1	28	2	1
Creditworthy	Some Balance	6	Paid Up	Home Rela	660	£100-£1000	1-4 yrs	2	1	23	1	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	666	£100-£1000	1-4 yrs	3	1	39	2	More than 1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Rela	700	£100-£1000	4-7 yrs	4	4	36	3	More than 1
Creditworthy	Some Balance	6	Paid Up	Home Rela	709	£100-£1000	< 1yr	2	1	27	2	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	932	£100-£1000	1-4 yrs	1	2	39	2	More than 1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Rela	1047	None	< 1yr	2	2	50	2	1
Creditworthy	No Account	6	Paid Up	Home Rela	1068	None	4-7 yrs	4	3	28	2	1
Creditworthy	No Account	6	Paid Up	Home Rela	1203	< £100	4-7 yrs	3	2	43	2	1
Creditworthy	Some Balance	6	No Problems (in this bank)	Used car	1221	£100-£1000	< 1yr	1	2	27	2	More than 1
Creditworthy	Some Balance	6	Paid Up	New car	1236	£100-£1000	< 1yr	2	2	50	1	1
Creditworthy	Some Balance	6	Paid Up	Home Rela	1338	£100-£1000	< 1yr	1	1	62	2	1
Creditworthy	Some Balance	6	Paid Up	Home Rela	1346	< £100	4-7 yrs	2	4	42	3	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	1361	None	< 1yr	2	1	40	2	1
Creditworthy	No Account	6	Paid Up	Used car	1374	None	< 1yr	1	1	36	2	1
Creditworthy	No Account	6	Paid Up	Home Rela	1374	£100-£1000	< 1yr	4	2	75	2	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	1449	< £100	4-7 yrs	1	3	31	2	More than 1
Creditworthy	Some Balance	6	Paid Up	Home Rela	1538	None	< 1yr	1	4	56	2	1
Creditworthy	Some Balance	6	Paid Up	Used car	1543	£100-£1000	< 1yr	4	1	33	2	1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Rela	1898	£100-£1000	< 1yr	1	1	34	2	More than 1
Creditworthy	No Account	6	Paid Up	Home Rela	2063	None	< 1yr	4	3	30	1	1
Creditworthy	Some Balance	6	Paid Up	Home Rela	2108	None	1-4 yrs	2	1	29	1	1
Creditworthy	Some Balance	6	Paid Up	Used car	2116	None	< 1yr	2	1	41	2	1
Creditworthy	Some Balance	6	Paid Up	Used car	2978	£100-£1000	< 1yr	1	3	32	2	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	3676	None	< 1yr	1	1	37	1	More than 1
Non-Creditworthy	Some Balance	6	Paid Up	Used car	4611	None	< 1yr	1	2	32	2	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Rela	4716	£100-£1000	< 1yr	1	1	44	2	More than 1

The alteryx workflow for the data exploration process:



Before diving into training the classification models, I ran a Pearson Correlation Analysis to determine if any fields have any bivariate association with one another. If the correlation between any two predictor variables is greater than 0.7, we would use one of the variables which is statistically more significant.

The following was the **output** of the analysis:



Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

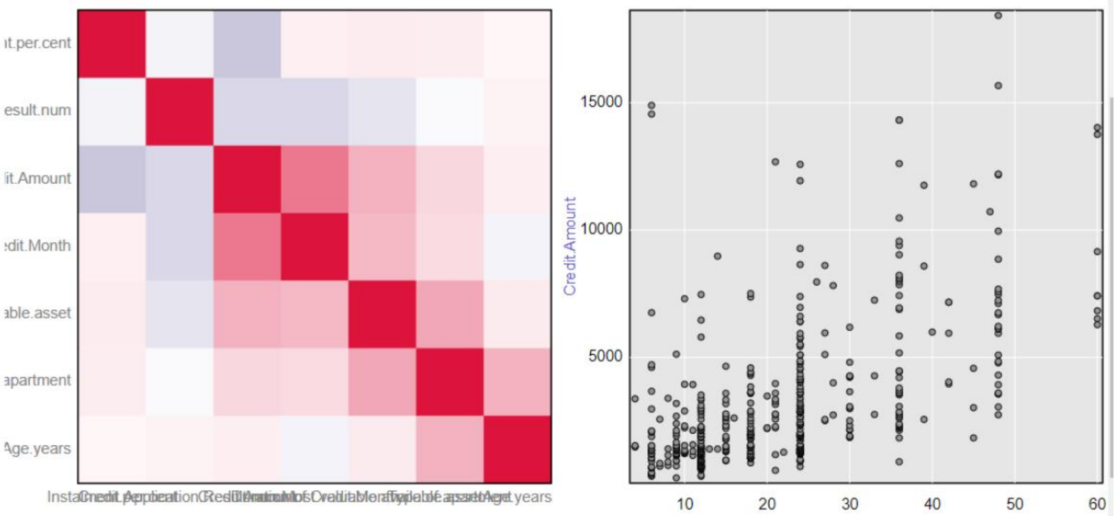
	Association Measure	p-value
Duration.of.Credit.Month	-0.202504	5.0151e-06 ***
Credit.Amount	-0.201946	5.3311e-06 ***
Most.valuable.available.asset	-0.141332	1.5334e-03 **
Instalment.per.cent	-0.062107	1.6556e-01
Age.years	0.052914	2.3758e-01
Type.of.apartment	-0.026516	5.5417e-01

Full Correlation Matrix

	Credit.Application.Result.num	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.yea
Credit.Application.Result.num	1.000000	-0.202504	-0.201946	-0.062107	-0.141332	0.0529
Duration.of.Credit.Month	-0.202504	1.000000	0.573980	0.068106	0.299855	-0.0641
Credit.Amount	-0.201946	0.573980	1.000000	-0.288852	0.325545	0.0693
Instalment.per.cent	-0.062107	0.068106	-0.288852	1.000000	0.081493	0.0392
Most.valuable.available.asset	-0.141332	0.299855	0.325545	0.081493	1.000000	0.0862
Age.years	0.052914	-0.064197	0.069316	0.039270	0.086233	1.0000
Type.of.apartment	-0.026516	0.152516	0.170071	0.074533	0.373101	0.3293
	Type.of.apartment					

As we can see from the results, duration of credit month and credit amount are fairly correlated with correlation value 0.57. Apart from these two variables, there is no significant correlation among the predictor variables. Hence, we can conclude that all the variables are good to move forward to be used in predictive modelling because there is no bivariate association with one another with over a value of 0.7.

Correlation Matrix



The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see

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.

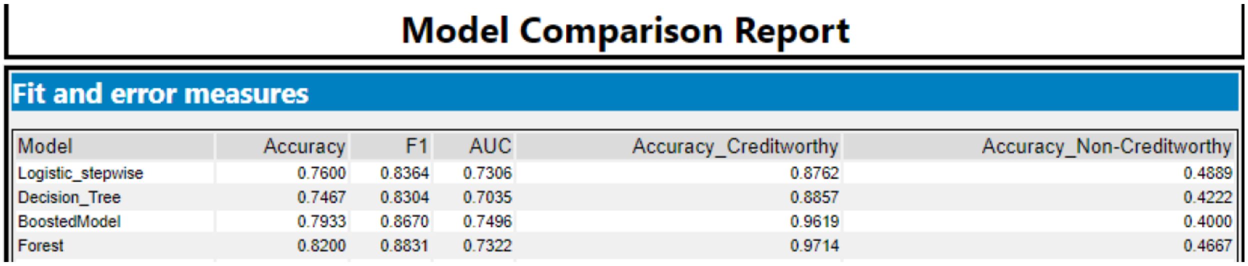
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

You should have four sets of questions answered. (500 word limit)

Now that we are ready with our predictor variables and a clean dataset, I have built the classification models and compared the results of each of these models:

- Logistic Regression
- Decision Tree
- Random Forest
- Boosted Model

Model Comparison Results for the 4 classification models



Model Comparison Report

Fit and error measures

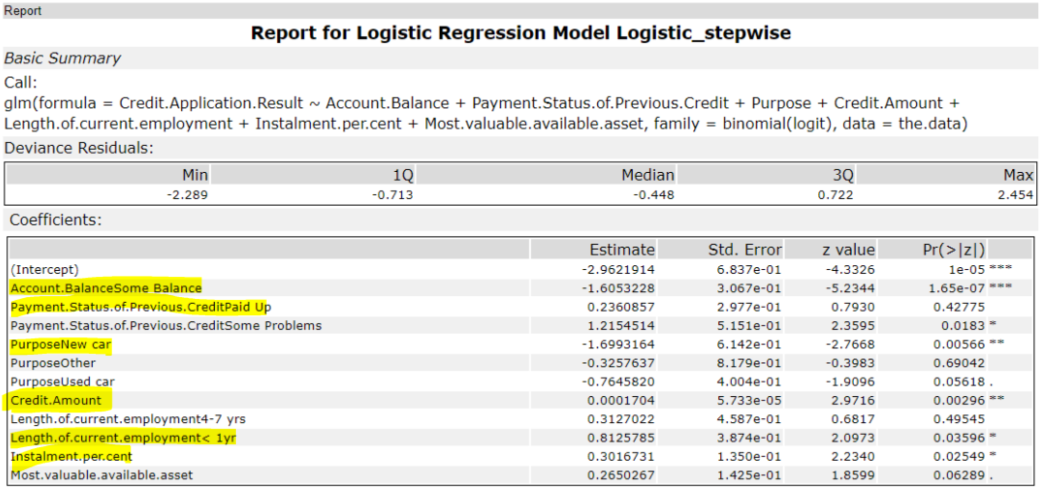
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
BoostedModel	0.7933	0.8670	0.7496	0.9619	0.4000
Forest	0.8200	0.8831	0.7322	0.9714	0.4667

Logistic Regression Model

I built a logistic regression model and used stepwise regression to determine the most statistically significant predictor variables.

These were:

- Account-Balance
- Payment-Status-of-Previous-Credit
- Purpose
- Credit-Amount
- Length-of-current-employment
- Instalment-per-cent



Report

Report for Logistic Regression Model Logistic_stepwise

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)
```

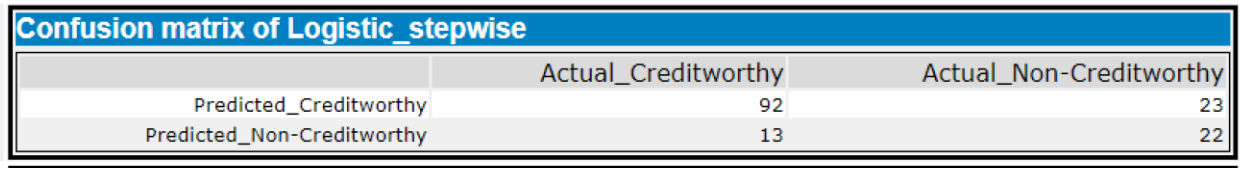
Deviance Residuals:

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	.

Confusion matrix

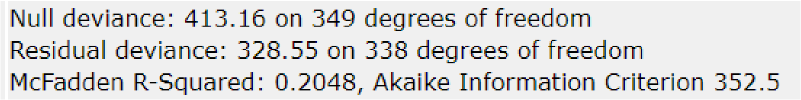


Confusion matrix of Logistic_stepwise

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

The overall percent accuracy of logistic regression model is 76%. We can see that the logistic regression shows bias where the accuracy to predict creditworthy customers is more than the accuracy of the non-creditworthy customers based on the validation set.

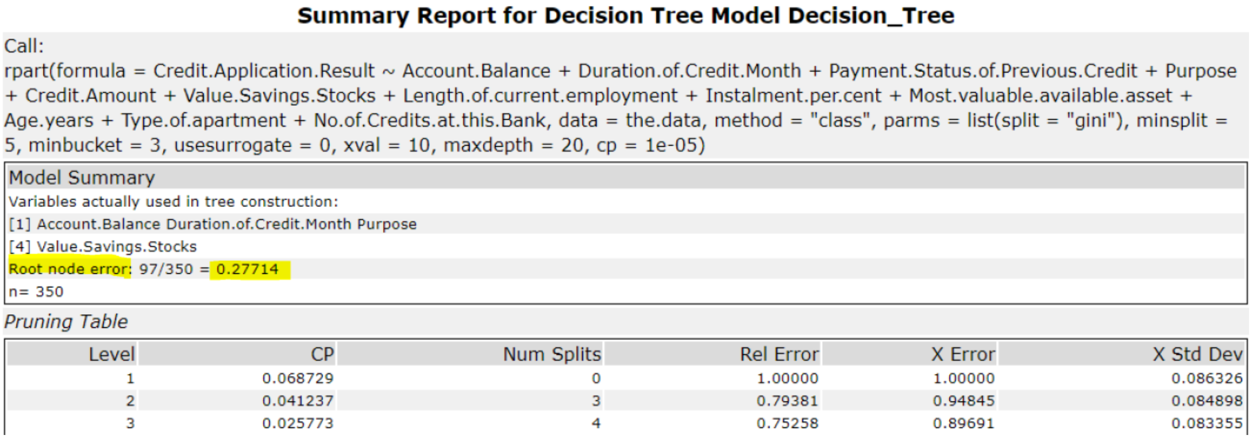
The R-Squared value sounds not good at all, with a value of 0.2048.



```
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
```

Decision Tree Model

The Decision Tree model produced a root node error of 27.7%, which considered as an acceptable error.



Summary Report for Decision Tree Model Decision_Tree

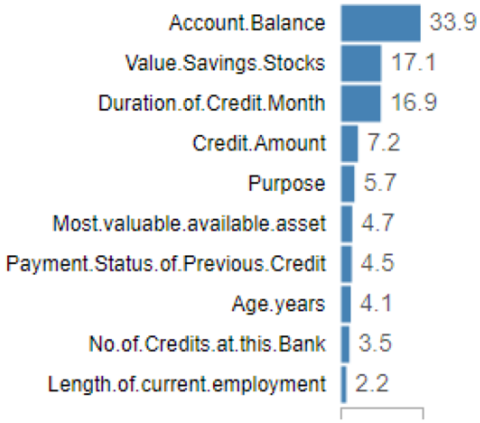
```
Call:
rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose
+ Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset +
Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, method = "class", parms = list(split = "gini"), minsplit =
5, minbucket = 3, usesurrogate = 0, xval = 10, maxdepth = 20, cp = 1e-05)
```

```
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose
[4] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.94845	0.084898
3	0.025773	4	0.75258	0.89691	0.083355

Variable Importance Plot

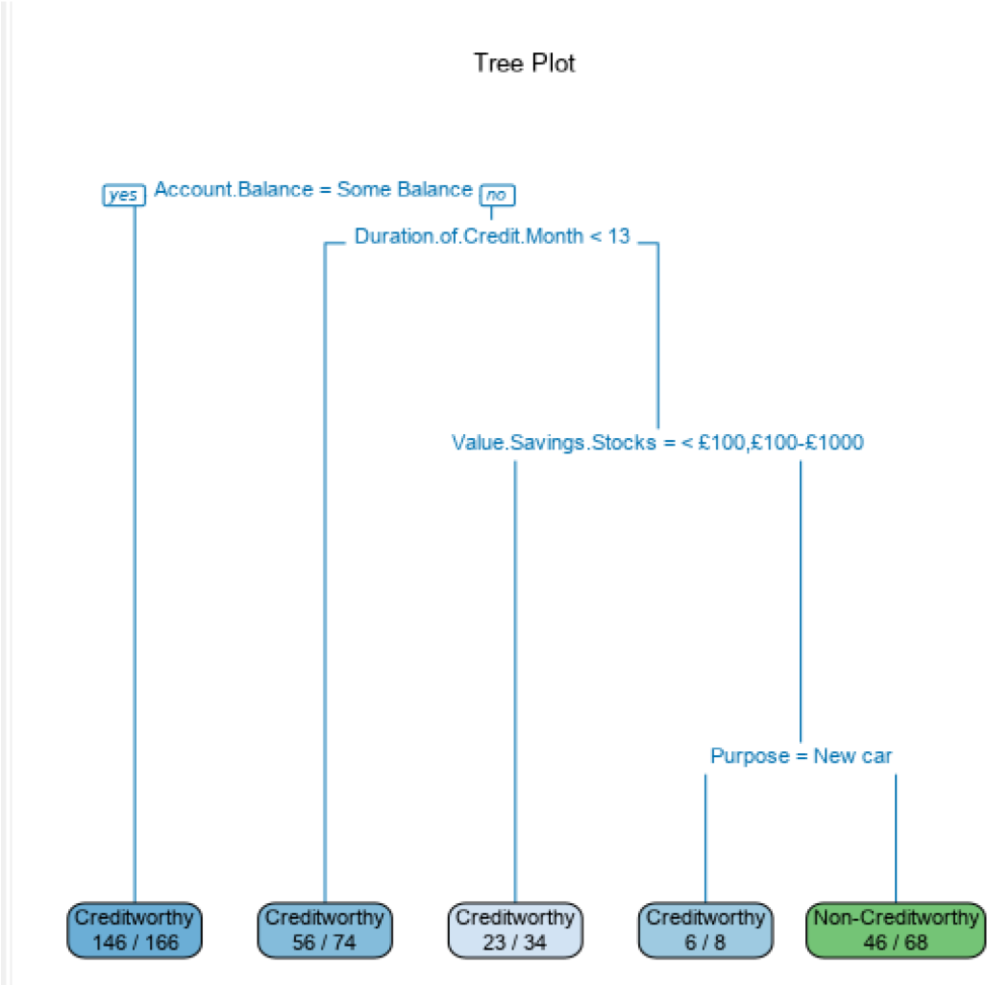


The top important variables were:

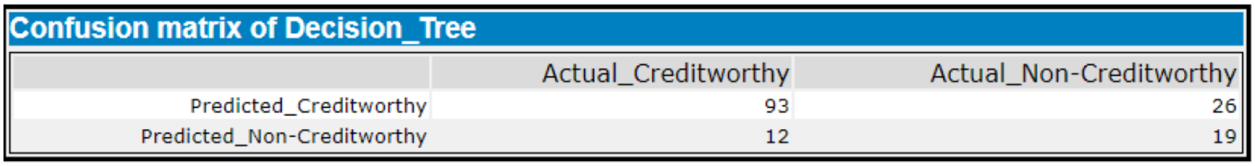
- Account Balance

- Value Savings Stock
- Duration of Credit Month

Tree Plot



Confusion Matrix



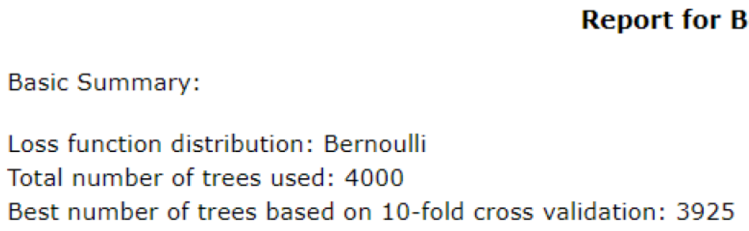
Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

The overall percent accuracy of decision tree model is 74.67%. It is slightly poor than logistic regression model. This model also shows bias where the accuracy to predict creditworthy customers is more than the accuracy of the non-creditworthy customers based on the validation set.

Boosted Model

I obtained the following model summary for boosted model with the loss function Bernouli and number of trees= 4000 to get the best accuracy for Boosted model.

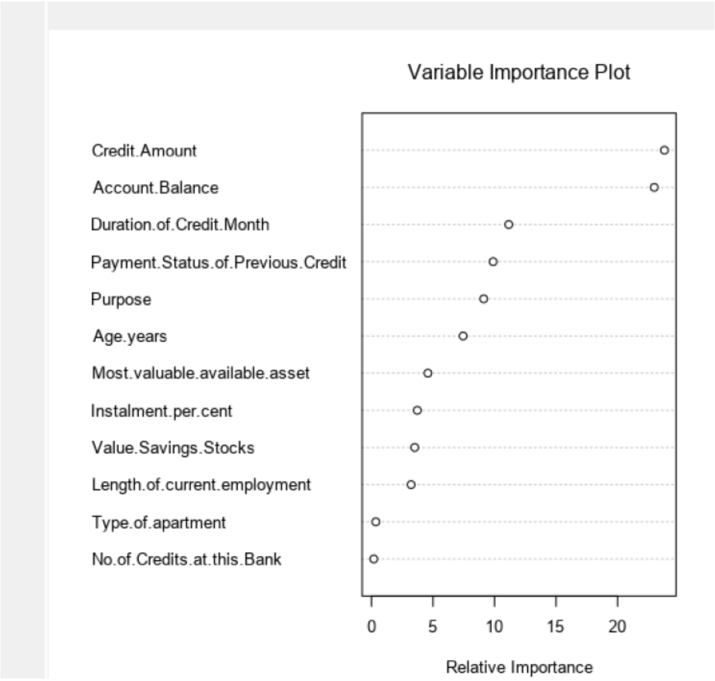


Report for Boosted Model BoostedModel

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 10-fold cross validation: 3925

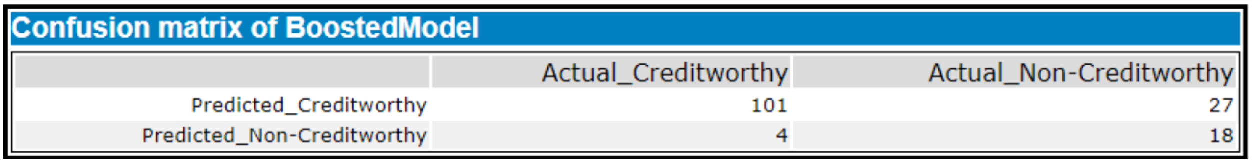
Variable importance plot



The top variables were:

- Credit Amount
- Account Balance
- Duration of Credit Month

Confusion matrix



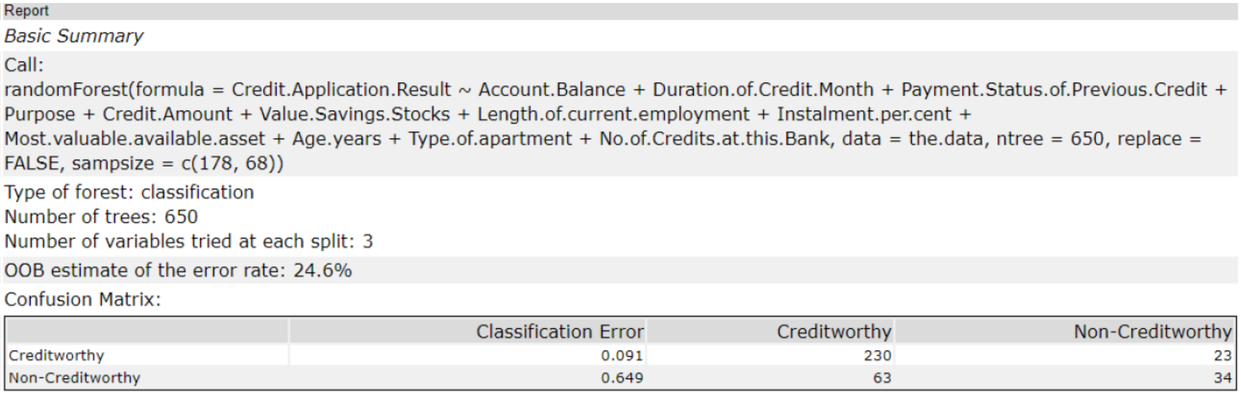
Confusion matrix of BoostedModel

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

The overall percent accuracy of boosted model is 79.33%. It is far better than the previous two models. This model also shows bias where the accuracy to predict creditworthy customers is more than the accuracy of the non-creditworthy customers based on the validation set.

Random Forest Model

The model summary of this model is as follows:



Report

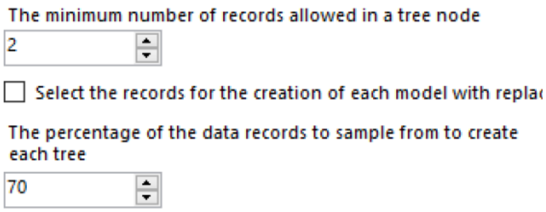
Basic Summary

```
Call:
randomForest(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, data = the.data, ntree = 650, replace =
FALSE, sampsize = c(178, 68))
Type of forest: classification
Number of trees: 650
Number of variables tried at each split: 3
OOB estimate of the error rate: 24.6%
Confusion Matrix:
```

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.091	230	23
Non-Creditworthy	0.649	63	34

To obtain the best accuracy, following configuration was used:

- Number of trees= 650
- Minimum number of records allowed in tree node= 2
- Percentage of data records to sample from to create each tree= 70



The minimum number of records allowed in a tree node

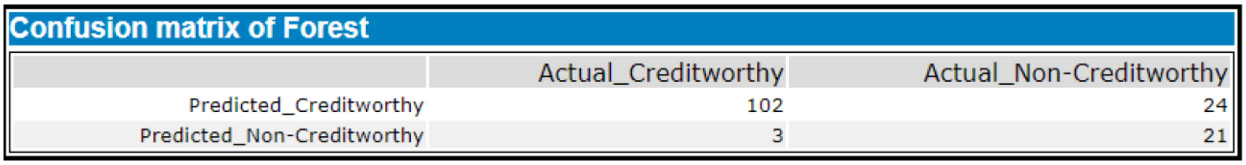
2

□ Select the records for the creation of each model with replac

The percentage of the data records to sample from to create each tree

70

Confusion Matrix



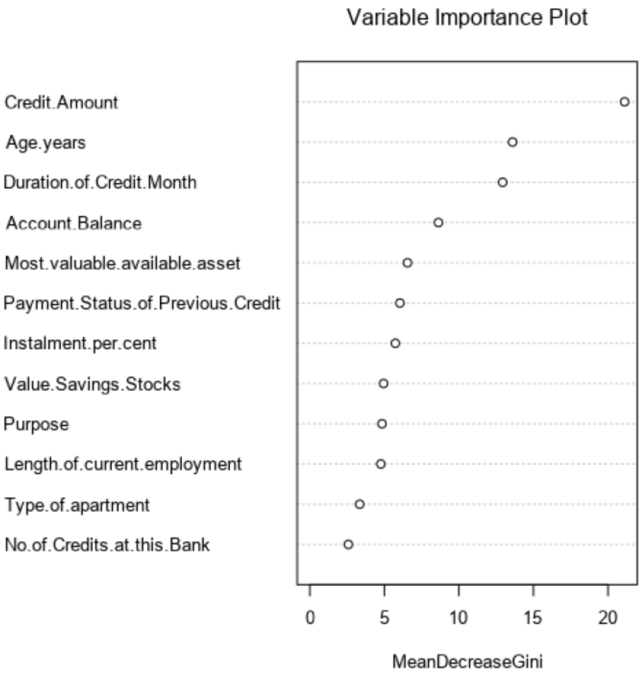
Confusion matrix of Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Variable Importance Plot

The top important variables were:

- Credit Amount
- Age years
- Duration of Credit Month



This model is not significantly different than the previous 3 models in terms of bias. However, the overall accuracy of random forest model is the highest with 82% and the individual accuracies within segments "Creditworthy" and "Non-Creditworthy" is better than the other 3 models.

Step 4: Writeup

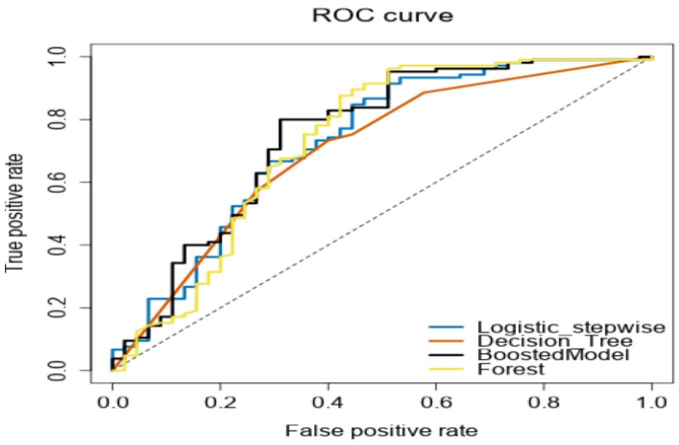
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Answer these questions:

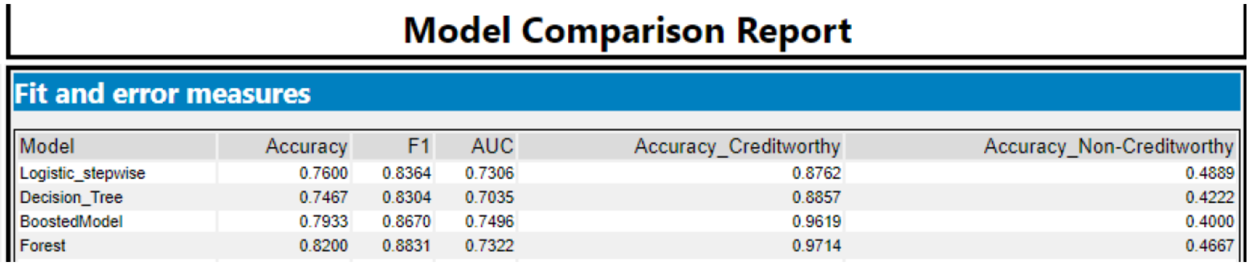
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph
    - Bias in the Confusion Matrices

I would select Random forest model because it has the maximum overall accuracy of 82%. Moreover, looking at the below ROC graph and accuracies within both segments "Creditworthy" and "Non-Creditworthy" we can see that random forest is hugging the top of the ROC curve with maximum true positive rate. Though, the true negative rate of Logistic Regression model (48.8 %) is slightly better than Forest model (46.6 %), it is not enough to select Logistic regression model as our final choice since the overall accuracy is very low. In our case, we have an imbalanced dataset. There are a lot more creditworthy applicants than non-creditworthy. Hence, to select a model for prediction we are interested in the overall accuracy and the Positive Predictive value (Precision), Negative Predictive value and F1 score and random forest model is holistically the best model out of the 4 models.

ROC Graph



Accuracy Results

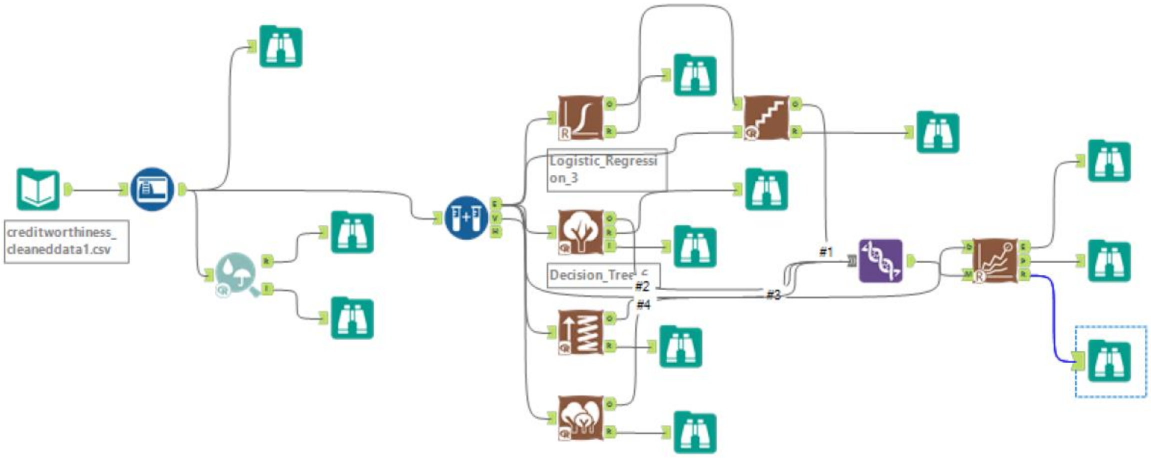


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
BoostedModel	0.7933	0.8670	0.7496	0.9619	0.4000
Forest	0.8200	0.8831	0.7322	0.9714	0.4667

The final alteryx workflow for model comparison:

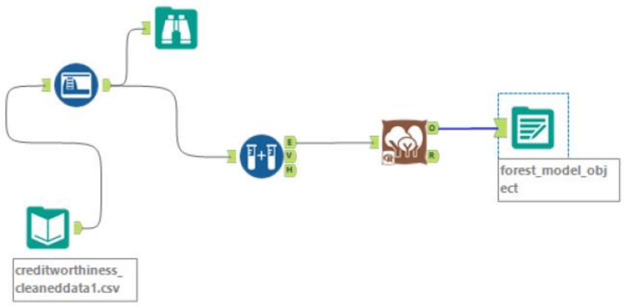


- How many individuals are creditworthy?

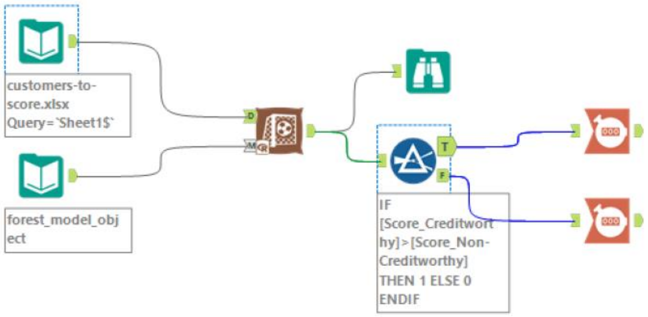
  There are **405** creditworthy individuals.

  The below alteryx workflow saves the best random forest model and predicts the number of creditworthy and non-creditworthy individuals.

Saving the model



Scoring the model



```
IF
[Score_Creditwort
hy]>[Score_Non-
Creditworthy]
THEN 1 ELSE 0
ENDIF
```

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.