Project: Creditworthiness

BY AYAME GOD'SWILL CLAUDE

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 need to be made?
- What data is needed to inform those decisions?
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Answers:

We need to decide if the 500 new loan applications of the current week should be approved.

To make this decision, we need to determine if the 500 loan applicants are creditworthy to give a loan to.

To help to make this determination, we will need to employ the skills of classification modelling, with a view to systematically evaluate the creditworthiness of these new loan applicants. This way, we will be able to provide a list of creditworthy customers to the manager in the next two days.

Since all we need to determine is whether or not they are creditworthy (a simple YES or NO situation), the binary classification model is most fitting.

The data required to build our model and make our prediction include:

i. Data on all past applications (training dataset to build our model), which include key relevant fields like: Account-Balance of the applicants, the

Purpose for which loan is being taken, Credit-Amount applied for, Age-years of the applicants, Payment-Status-of-Previous-Credit, Duration-of-Credit-Month, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Guarantors, Duration-in-Current-address, Most-valuable-available-asset, Concurrent-Credits, Type-of-apartment, No-of-Credits-at-this-Bank, Occupation, No-of-dependents, Telephone and Foreign-Worker. Out of these fields, the actual significant predictor variables which provide good insight to the target variable (Credit-Application-Result field) will later be determined.

ii. The dataset containing the list of customers that need to be processed in the next few days (dataset on which predictions will be made), which includes similar fields like those in the training dataset.

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

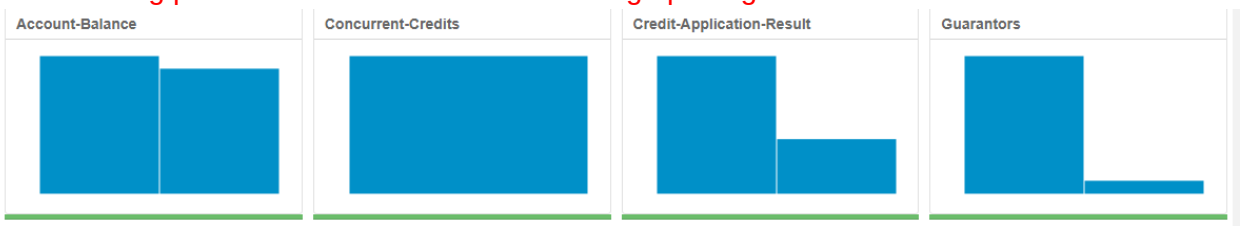
To achieve consistent results reviewers expect.

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.

Answers:

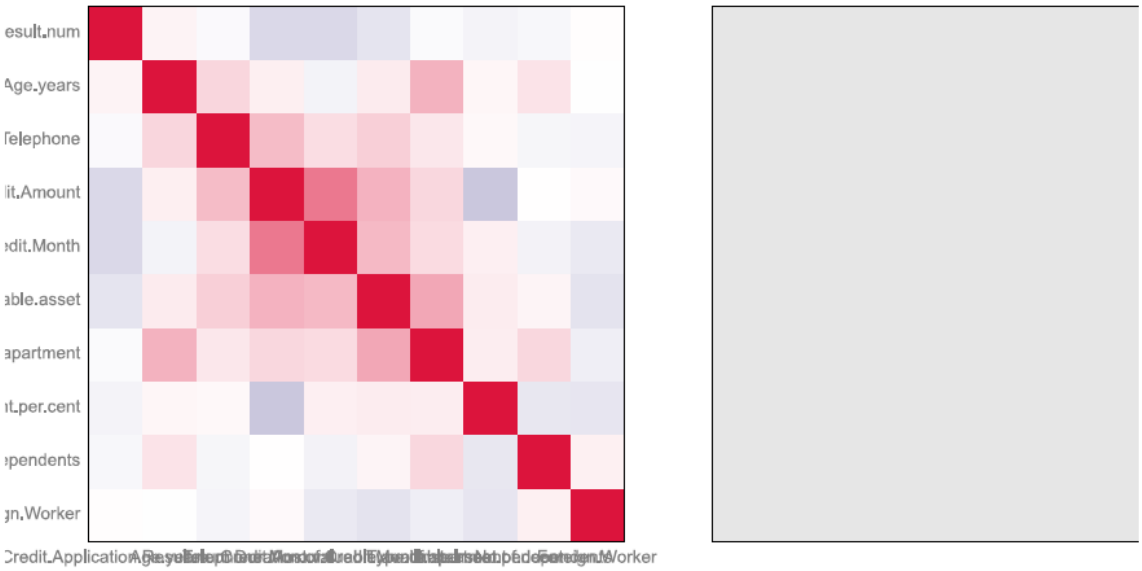
The following processes were involved in cleaning up the given dataset:



- First, the field summary tool was used to check the *categorical variables* to see which ones have too many categories. None fell into this group hence no predictor variable was removed based on this consideration**.**
- Again, the results of the field summary tool shows that the field **Duration-in-Current-address** has too many missing data values. Hence, it was removed.
- **Age-years** has just a couple of null values, hence its field was retained and the missing values were imputed using the median of the entire data field, instead of removing a few data points.
- More so, the fields **Concurrent-Credits** and **Occupation** exhibits uniformity (a specific case of low variability) since there is only one value for the entire field. Consequently, these field was also removed.

- Hovering the cursor over the charts in the field summary results shows that the field **Guarantors** has two unique values, but data is almost uniformly distributed to one of the values (457 NONE and only 43 YES). And as for **No-of-dependents**, its data stream is heavily skewed – there are 427 ones (1's) as against 73 twos (2's). More so, the field **Foreign Worker** has 481 2's as against 19 1's. Hence, all three fields were removed for exhibiting low variability (one value happening much more than the other).
- The field **Telephone** has no logical basis for being included in the training dataset since it is not relevant to predicting creditworthiness of an applicant.
- As for the *numerical data fields*, the ASSOCIATION ANALYSIS tool was used to check if there were any fields that highly correlate with each other (with a correlation ≥ ±0.70). The correlation matrix (heat plot) and corresponding scatterplots from the I (interactive) output of the ASSOCIATION ANALYSIS tool helped me to determine if there were pairs of predictor variables with inner correlation with each other (duplicate variables, since they are basically one and the same thing), one of which should be removed and not used in building our model to avoid redundancy and errors in our model. The one that will be removed should be the one that has a lesser association with the target variable (as determined by the p-value and number of stars) in the R (report) output of the ASSOCIATION ANALYSIS tool.
  The results showed that none of the predictor variables have an inner correlation ≥ ±0.70

Correlation Matrix with ScatterPlot



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 the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

But then again, looking at the R output of the ASSOCIATION ANALYSIS tool, it can be observed that the fields DURATION OF CREDIT MONTH, CREDIT AMOUNT and MOST VALUABLE AVAILABLE ASSET are the most statistically significant variables (with p-values all less than 0.05, and a minimum of two stars **) to the target variable CREDIT APPLICATION RESULT.

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value	
Duration.of.Credit.Month	-0.2025036	5.0151e-06	***
Credit.Amount	-0.2019458	5.3311e-06	***
Most.valuable.available.asset	-0.1413324	1.5334e-03	**
Instalment.per.cent	-0.0621068	1.6556e-01	
Age.years	0.0529139	2.3758e-01	
No.of.dependents	-0.0410479	3.5969e-01	
Telephone	-0.0289707	5.1807e-01	
Type.of.apartment	-0.0265155	5.5417e-01	
Foreign.Worker	0.0091861	8.3765e-01	

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)

ANSWERS:

After creating my Estimation and Validation samples (with 70% of my entire training dataset going to Estimation and 30% reserved for Validation) and setting the Random Seed to 1 in the CREATE SAMPLES tool in Alteryx, the following models were created with the resulting statistics for each included below.

1. **Logistic regression:** After connecting the LOGISTIC REGRESSION TOOL to the estimation dataset, the STEPWISE tool was connected to the output of the LOGISTIC REGRESSION tool to automate the process of determining which predictor variables are statistically significant. The R output of the STEPWISE TOOL showed an R-squared value of 0.2048, which is far less than the ideal value of 1. And then again, looking at the

upper part of the report, we can see that there are 6 predictor variables which show some statistical significance to the building of the model (with p-values of less than 0.05 and a range of 1 to 3 stars), namely Account.Balance, Payment.Status.of.Previous.Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent and Most.valuable.available.asset.

Report for Logistic Regression Model Stepwise_tool_3

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

2. **Decision Tree Model**: The R output showed a Root node error: 97/350 = 0.27714 or 27.7%. This gives us an idea of the actual percentage of the out of our overall dataset which fell in the correct terminal node. Thus, 97 out of 350 (27.7%) fell in the wrong terminal node, whilst 72.3% fell in the correct terminal node.

More so, the variables actually used for the tree construction are: Account.Balance, Duration.of.Credit.Month, Purpose and Value.Savings.Stocks, as revealed by the Model Summary section of the report and the splitted tree diagram below.

Summary Report for Decision Tree Model Decision_Tree_3

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, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary

Variables actually used in tree construction:

[1] Account.Balance Duration.of.Credit.Month Purpose 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	0.88660	0.083032

Leaf Summary

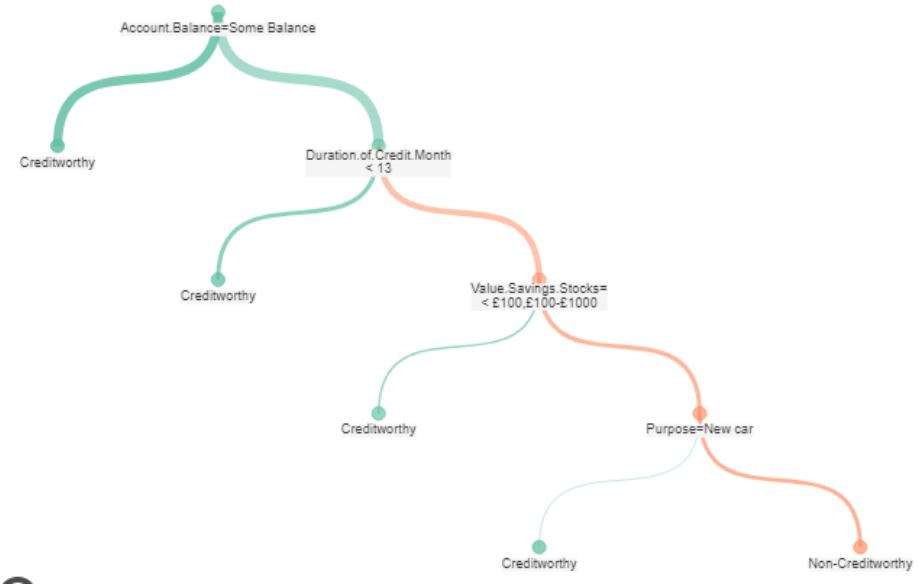
node), split, n, loss, yval, (yprob)

* denotes terminal node

1) root 350 97 Creditworthy (0.7228571 0.2771429)
2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789)
30) Purpose=New car 8 2 Creditworthy (0.7500000 0.2500000) *
31) Purpose=Home Related,Other,Used car 68 22 Non-Creditworthy (0.3235294 0.6764706) *

Plots

The I output of the decision tree revealed that the decision tree was splitted as shown below.



The results of the confusion matrix, which helps in determining where there may be biases in our dataset or if it is skewed to one side, plus the level of Miscalculations, is as shown below

	Actual Positive	Actual Negative
Predicted Positive	46 (67.6%)	22 (32.4%)
Predicted Negative	51 (18.1%)	231 (81.9%)

And the SUMMARY of the I output is as shown below

Accuracy 79.1 %
Proportion of correct predictions in the data

F1 Score 55.8 %
Harmonic mean of Recall and Precision

Precision 67.6 %
Proportion of values predicted positive, that were actually positive

Recall 47.4 %
Proportion of values actually positive, that were predicted positive

But since DECISION TREE model has the downside of overfitting a dataset (wherein the model fits the sample dataset a little too well and as a result does not predict the future

accurately enough), we will need to validate it later and compare with other models to choose the best for our prediction.

3. **Forest Model**: The R output shows an OOB (Out of The Bag) estimate of the error rate of 23.1%, which is quite high and a bit worrisome to trust to predict accurately.
The CONFUSION MATRIX shows that there was only a 6.7% error in predicting the CREDITWORTHY category as against 66% for NON-CREDITWORTHY prediction.

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 = 500, replace = TRUE)

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

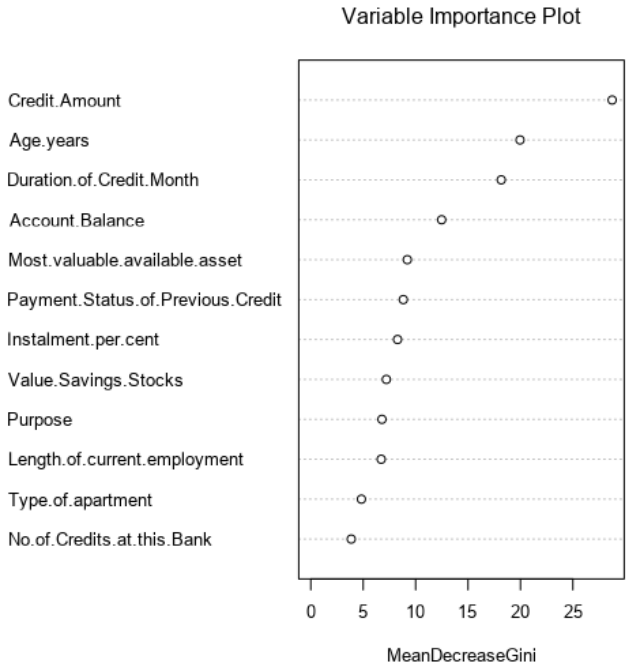
OOB estimate of the error rate: 23.1%

Confusion Matrix:

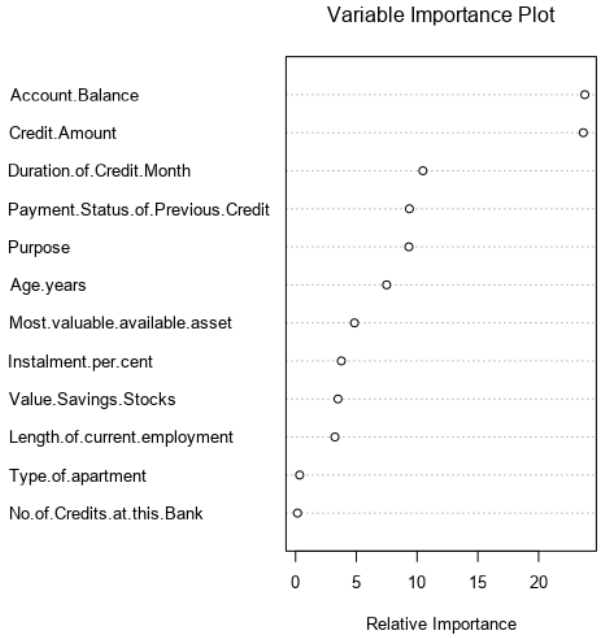
	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.067	236	17
Non-Creditworthy	0.66	64	33

Plots

The VARIABLE IMPORTANCE PLOT shows that all 12 predictor variables are accommodated by this model, with CREDIT AMOUNT, AGE YEARS, DURATION OF CREDIT MONTH, ACCOUNT BALANCE and MOST VALUABLE AVAILABLE ASSET being the most important predictor variables for this model.



4. **Boosted Model**: The VARIABLE IMPORTANCE PLOT of the R output shows 12 statistically significant predictor variables, with ACCOUNT BALANCE and CREDIT AMOUNT being the two most insightful predictors for the target variable in this model.



Next, I used the UNION tool to union all 4 models together, and then finally the MODEL COMPARISON TOOL was used to compare all 4 models side-by-side to ascertain which is the best model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_model	0.7867	0.8632	0.7507	0.9619	0.3778
Forest_Model_3	0.7933	0.8681	0.7368	0.9714	0.3778
Decision_Tree_3	0.7467	0.8304	0.7035	0.8857	0.4222
Stepwise_tool_3	0.7600	0.8364	0.7306	0.8762	0.4889

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.

BIASES IN EACH MODEL

Bias may be considered as the tendency of a model to predict one of its outcomes much more accurately than the others (in this case, Creditworthy vs Non-Creditworthy).

To determine the biases in each model, I considered the accuracies on both the CREDITWORTHY and NON-CREDITWORTHY segments of the Model Comparison Report, plus I checked/worked out the PPV and NPV values using the Confusion Matrix. If one of the

prediction's percentage accuracy is way higher than the other, then the model has a bias towards that higher end, otherwise it has little or no bias.

The following biases were noticed in each model's prediction ability.

Logistic/Stepwise Regression: From Model Comparison Report, the overall percent accuracy of the Logistic model is 76%, which is strong.

Accuracy in predicting Creditworthiness is 0.8762 (87.6%), while that for Non-creditworthiness is 0.4889 (48.9%). The difference is high.

More so, from the Confusion Matrix:
PPV= true positives / (true positives + false positives) = 92 / (92+23) = 0.80 or 80%
NPV= true negatives / (true negatives + false negatives) = 22 / (22+13) = 0.63 or 63%
The difference between the PPV and NPV values is also much.

So, after checking the confusion matrix there is bias seen in the model's prediction towards Creditworthy.

Forest Model: From Model Comparison Report, the overall percent accuracy of the Forest model is 79.3%, which is strong.

Accuracy in predicting Creditworthiness is 97.1%, while that for Non-creditworthiness is 37.8% - a difference quite high, suggesting bias.

However, from the Confusion Matrix:
PPV= true positives / (true positives + false positives) = 102 / (102+28) = 0.78
NPV= true negatives / (true negatives + false negatives) = 17/ (17+3) = 0.85
I noticed that the PPV and NPV values are quite close. So, after checking the confusion matrix there is little or no bias seen in the model's prediction.

Decision Tree Model: From the Model Comparison Report, the overall percent accuracy of the Decision Tree model is 74.7%, which is strong.

Accuracy in predicting Creditworthiness is 88.6%, while that for Non-creditworthiness is 42.2%. The difference is high.

More so, from the Confusion Matrix:
PPV= true positives / (true positives + false positives) = 93 / (93+26) = 0.78
NPV= true negatives / (true negatives + false negatives) = 19 / (19+12) = 0.61
The difference between the PPV and NPV values is also much.

So, after checking the confusion matrix there is bias seen in the model's prediction towards Creditworthy.

Boosted Model: From the Model Comparison Report, the overall percent accuracy of the Boosted model is 78.7%, which is strong.

Accuracy in predicting Creditworthiness is 96.2%, while that for Non-creditworthiness is 37.8%. The difference is high.

More so, from the Confusion Matrix:
PPV= true positives / (true positives + false positives) = 101 / (101+28) = 0.78
NPV= true negatives / (true negatives + false negatives) = 17 / (17+4) = 0.81
the difference between the PPV and NPV values is also much.

So, just like the forest model, this model is almost not biased at all after checking the confusion matrix.

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:

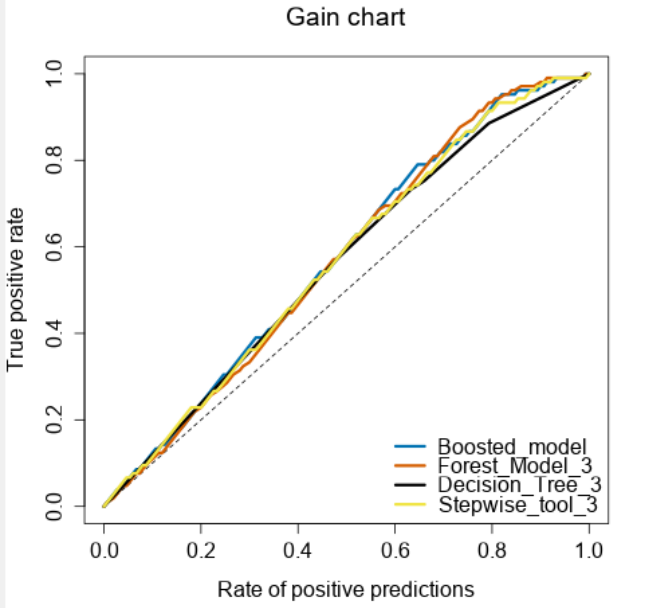
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    - ROC graph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

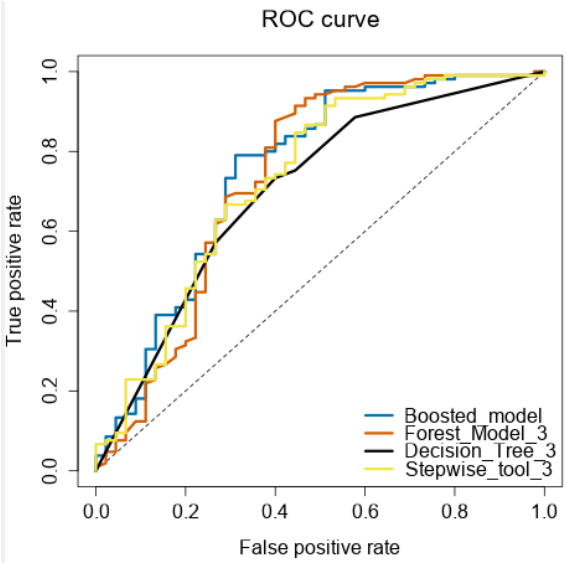
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

ANSWERS
In summary, the FOREST MODEL seems to perform slightly above all other models in all considerations.

- In the overall accuracy against the validation set, it was the highest at 0.7933, as seen above.
- More so, looking at the GAIN CHART and the ROC curve below, the FOREST MODEL reaches the top (in the y-axis) the quickest slightly above the other models.
  More so, the curve of the FOREST MODEL is the overall highest curve of all. This means that for a given amount of false positive predictions (wrongly predicted creditworthy people), this model will give the best number of true positive predictions (correctly predicted creditworthy people).
  Again, comparing the Area Under the Curve (AUC) values for all the models as seen in the Model Comparison Report, the FOREST MODEL has an AUC of 0.7368 (which is 2nd only to that of the BOOSTED MODEL), and the higher the AUC, the higher the curve, vice versa. Thus, the higher the AUC, the better the model's predicting ability.





- The FOREST'S MODEL accuracy in predicting Creditworthiness is 97.1% (which is the highest of all the 4 models considered), while that for Non-creditworthiness is 37.8% (which is equal to that of the next-best model – the BOOSTED MODEL).

- In terms of Confusion Matrix, the Forest Model has a PPV of 0.78 and an NPV of 0.85, which are quite close. Thus, there is little or no bias seen in the model's prediction ability. And with the exception of the BOOSTED MODEL, the FOREST MODEL ranks the best in the bias consideration. And in actuality, it's rating, bias-wise, is approximately as good as that of the BOOSTED MODEL.

Confusion matrix of Boosted_model

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

Confusion matrix of Decision_Tree_3

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

Confusion matrix of Forest_Model_3

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

Confusion matrix of Stepwise_tool_3

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

Thus, putting all of the four strong factors above together, the RANDOM FOREST MODEL is the best choice.

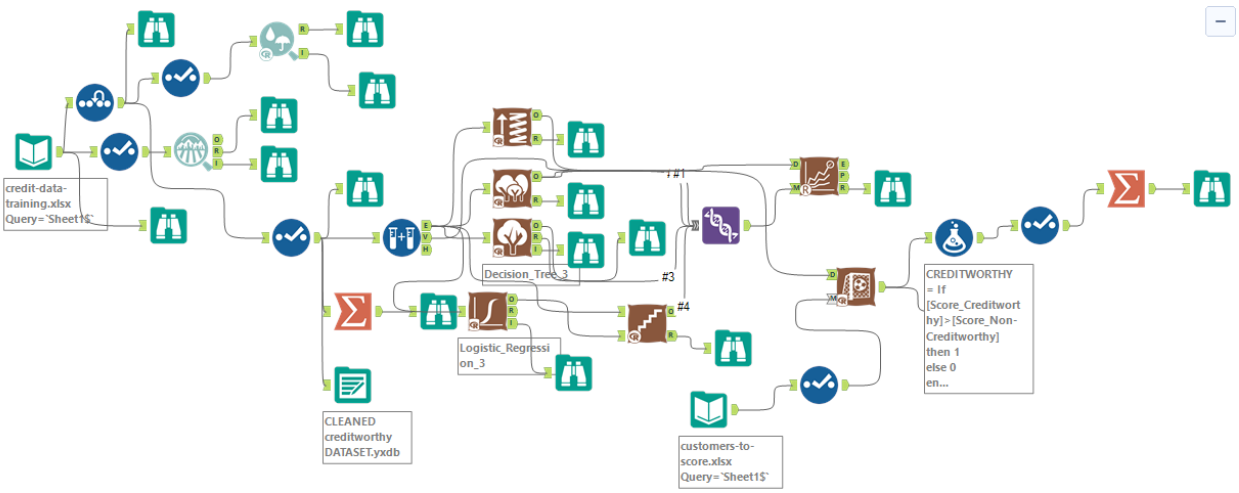
Next, the INPUT DATA TOOL was used to introduce the dataset to be scored to the workflow, which together with the FOREST MODEL was connected to the SCORE TOOL. The results displayed shows the probabilities of Creditworthy and Non-Creditworthy for each record.

The interpretation was done as follows: If Score_Creditworthy is greater than Score_NonCreditworthy, the person was labeled as “Creditworthy”.

A formula tool was then used to find the actual number of “Creditworthy” applicants, using the criteria above. And finally, a SUMMARIZE tool was used to aggregate the number of CREDITWORTHY applicants, which was found to be 408.

Thus, 408 new customers would qualify for a loan.

My solution workflow is shown below.



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

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.