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?

Based on data from past approved or rejected loans in our bank we have to determine whether new customers are creditworthy or not. We have to select appropriate variables and the model with the best accuracy.

- What data is needed to inform those decisions?
    - Variables from all past applications (loans approved or rejected)
    - Variables from current applications
      Variables should be related at least to age, assets, employment and financial situation.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We need a binary model because there are two possible outcomes: creditworthy or non-creditworthy. The potential model candidates are:
    - Logistic Regression
    - Decision Tree
    - Forest Model
    - Boosted Tree

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.***

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

As shown in Figure 1, the Field Summary tool was used to find variables with missing data or low variability.

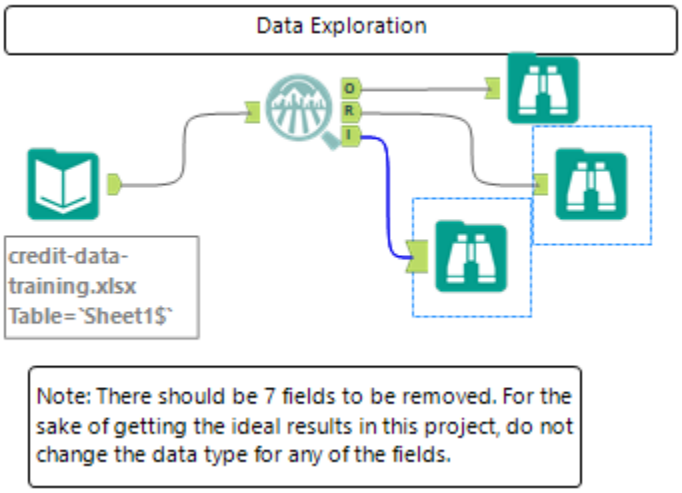

Figure 1. Field Summary tool to find variables with missing values or low variability.

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”.

With the **Association Analysis tool** the Pearson correlation values between numeric values can be obtained. As shown in Figure 2 the highest correlation value between variables (amount of credit and months of credit) is 0.574. Therefore there is no need to discard variables because of high correlation — very similar information given.

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Type.of.apartment	Age.years
Duration.of.Credit.Month	1.000000	0.573980	0.068106	0.299855	0.152516	-0.064197
Credit.Amount	0.573980	1.000000	-0.288852	0.325545	0.170071	0.069316
Instalment.per.cent	0.068106	-0.288852	1.000000	0.081493	0.074533	0.039270
Most.valuable.available.asset	0.299855	0.325545	0.081493	1.000000	0.373101	0.086233
Type.of.apartment	0.152516	0.170071	0.074533	0.373101	1.000000	0.329350
Age.years	-0.064197	0.069316	0.039270	0.086233	0.329350	1.000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Type.of.apartment	Age.years
Duration.of.Credit.Month		0.0000e+00	1.2830e-01	7.5764e-12	6.2192e-04	1.5175e-01
Credit.Amount	0.0000e+00		4.5919e-11	8.3045e-14	1.3277e-04	1.2164e-01
Instalment.per.cent	1.2830e-01	4.5919e-11		6.8653e-02	9.5961e-02	3.8090e-01
Most.valuable.available.asset	7.5764e-12	8.3045e-14	6.8653e-02		0.0000e+00	5.3979e-02
Type.of.apartment	6.2192e-04	1.3277e-04	9.5961e-02	0.0000e+00		4.0856e-14
Age.years	1.5175e-01	1.2164e-01	3.8090e-01	5.3979e-02	4.0856e-14	

Figure 2. Correlation between numeric variables. Top: Pearson’s correlation variables. Bot: Associated p-values to each correlation.

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
  As shown in Figure 3, **Duration-in-Current-address** has 68.8% of missing values, so it is not considered for the analysis.

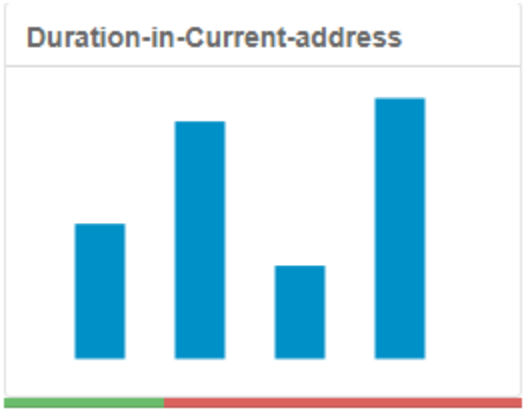

Figure 3. The variable Duration-in-Current-address has around 69% of missing values. It is one of the non-considered variables.

Figure 4 shows that **Age_years** has 2.4% missing values. This is a very small amount of missing values so I decided to impute the median value — average is not appropriate because of the highly right skewed shape— with an Imputation tool.

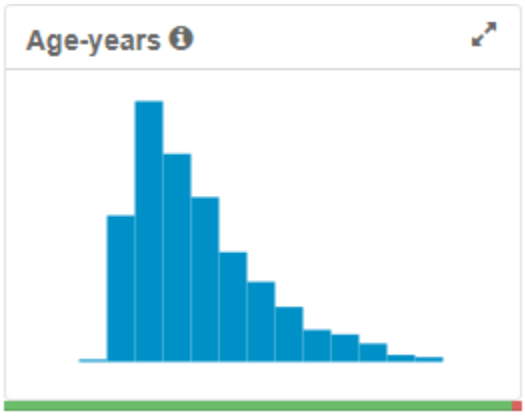

Figure 4. The variable Age-years has around 2% of missing values. Due to the highly skewed shape imputing the median value is the best option.

- 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.

As shown in Figure 5 **Concurrent-Credits** and **Occupation** have one unique value and must be discarded.

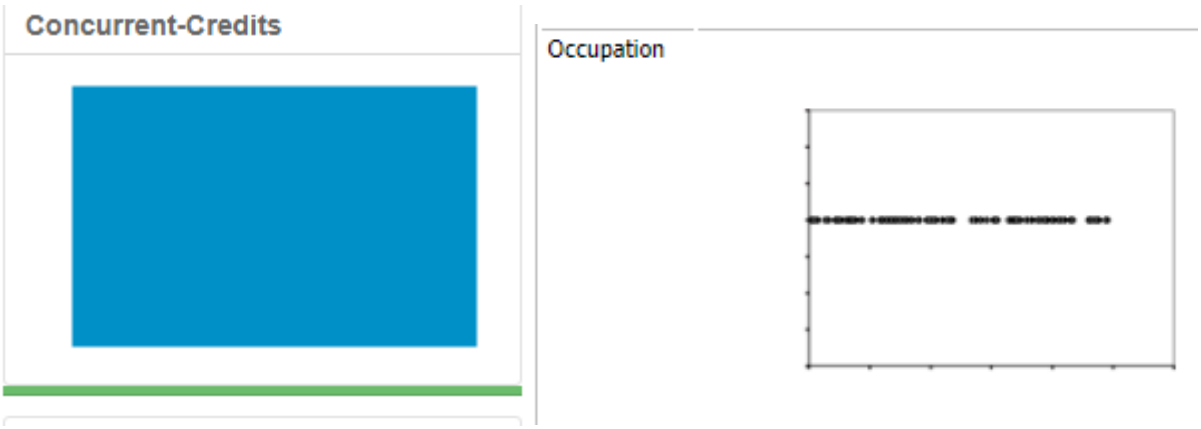

Figure 5. Variables with one unique value: Concurrent-Credits and Occupation.

Figure 6 reveals that **Guarantors, No-of-dependents** and **Foreign-Worker** have very low variability, and therefore should not be considered in the analysis.

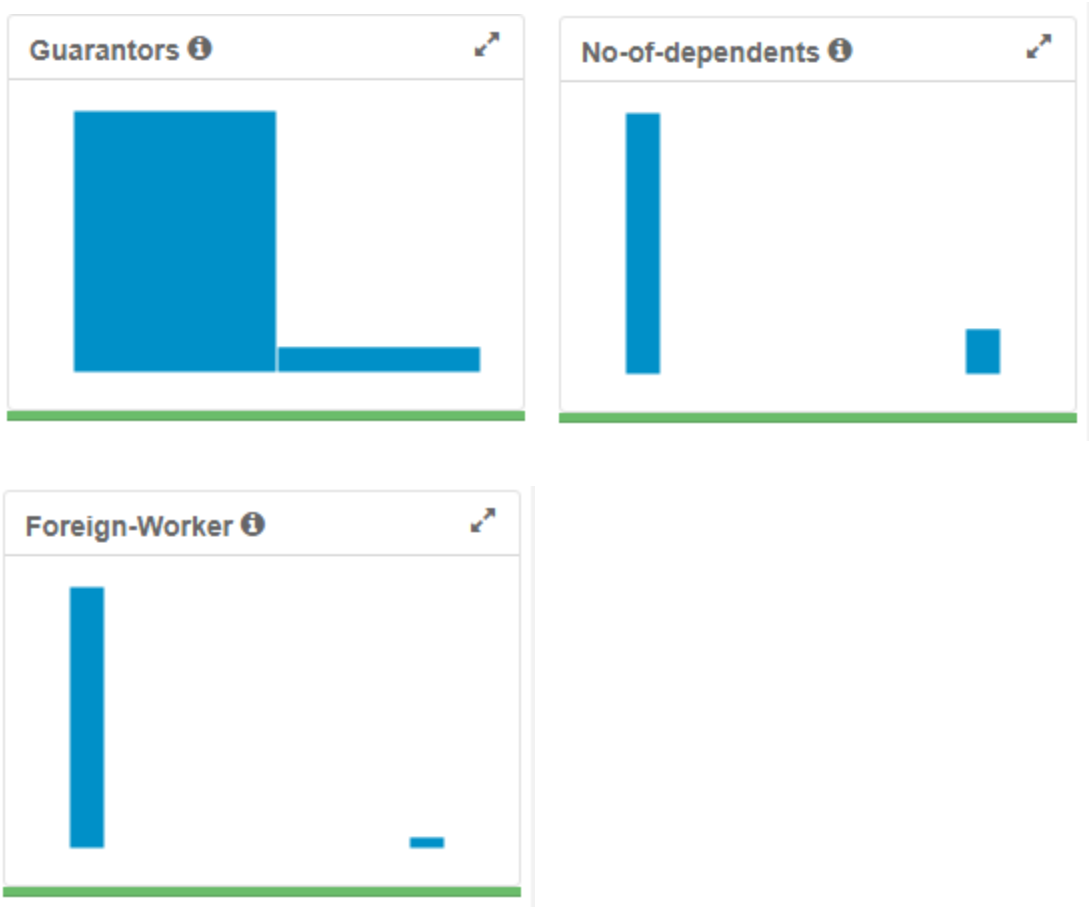

Figure 6. Variables with low variability: Guarantors, No-of-dependents and Foreign-Worker.

Finally, we should remove **Telephone** since it is not a logical variable to predict if a person is creditworthy.

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:*

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

3.A.1. Logistic Regression Model

In the logistic regression model the variables employed in the model are: Account-Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Length-of-current-employment, Instalment-per-cent and Most-valuable-available-asset. In the right column of Figure 7 the p-values are shown: variables with at least one asterisk or dot are significant. Some categorical variables are not significant in all the categories.

Report for Logistic Regression Model Stepwise_Log

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

Figure 7. Information about the predictor variables in the logistic regression model.

3.A.2. Decision Tree Model

As shown in Figure 8, for the decision tree model only the three variables with more importance (Figure 9) were used in the tree construction: Account-Balance, Duration-of-Credit-Month and Value-Savings-Stocks.

```
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 +
Type.of.apartment + No.of.Credits.at.this.Bank + Age.years, data = the.data, minsplit = 20,
minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)
```

Model Summary

```
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month 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	0.92784	0.084295

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) *
```

Figure 8. Information about the predictor variables in the decision tree model.

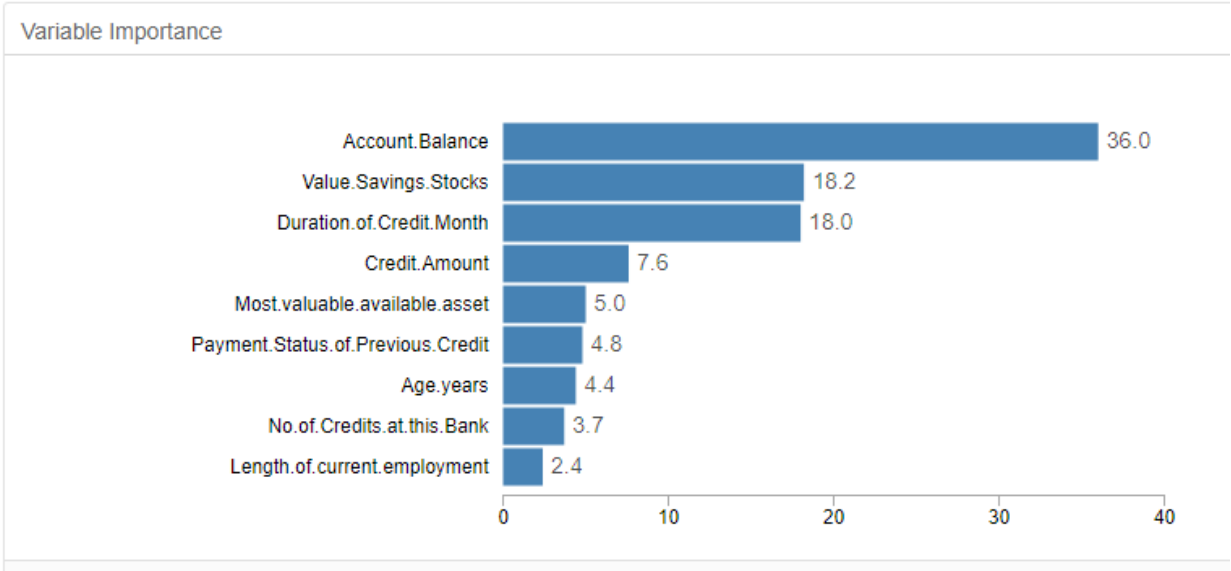

Figure 9. Variable importance for decision tree model.

3.A.3. Random Forest Model

Figure 10 shows the formula with the variables employed in the random forest model with 500 trees and Figure 11 show the variable importance (ordered from top to bottom).

```
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 +
Type.of.apartment + No.of.Credits.at.this.Bank + Age.years, 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: 24%
```

Figure 10. Information about the predictor variables in the random forest model.

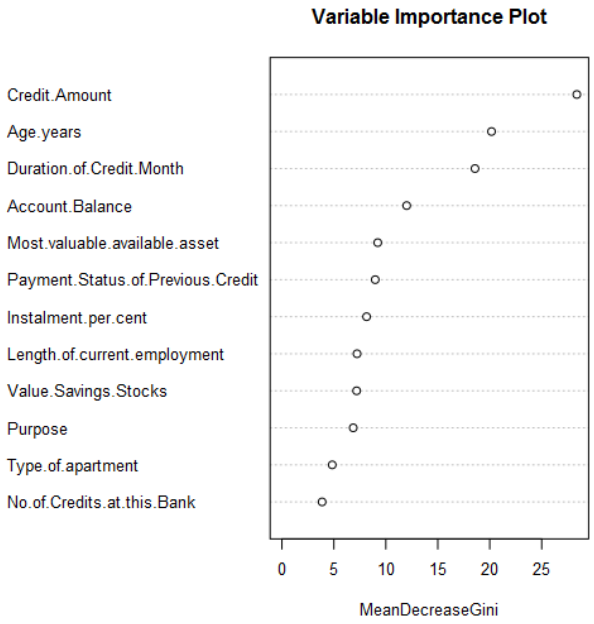

Figure 11. Variable importance for random forest model.

3.A.4. Boosted Model

The loss function distribution was Bernoulli and 4,000 trees were used although 2,036 would be the best number of trees according to 5-fold cross validation. Figure 12 shows the importance for each variable, normalized to sum 100.

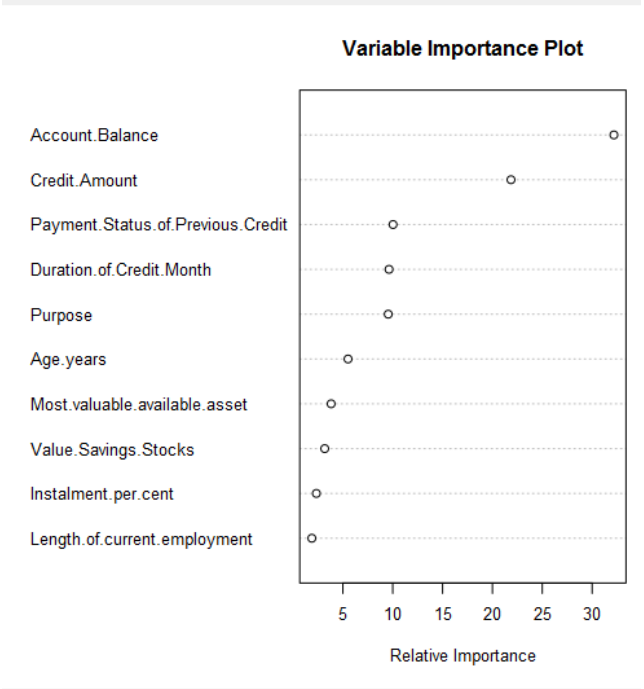

Figure 12. Variable importance for the boosted model.

B) 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?

- Regarding accuracy, as shown in Figure 13, random forest model shows the higher value (0.8), also the highest F1 score and a good area under the ROC curve (AUC) of 0.7361. The four models have higher accuracy for creditworthy than for non-creditworthy. In the past applications, 358 out of 500 application resulted in given loans, so the creditworthy people are more represented in the sample and the models show a bias towards determining "creditworthy" and underpredicting "non-creditworthy". The null error rate — how often we will we wrong if we always predict the majority class, creditworthy — is 0.3.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Random_Forest	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_Log	0.7600	0.8364	0.7306	0.8762	0.4889

Figure 13. Performance metrics for logistic regression, decision tree, random forest and boosted models. Accuracy shows the number of correct predictions of all classes divided by total sample number; 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; AUC: area under the ROC (receiver operator curve) curve. It is feasible to lend a loan to creditworthy people.

3.B.1. Logistic Regression Model

Figure 14 shows the confusion matrix for the logistic regression model, with a true positive (TP) rate of 0.88, false positive (FP) rate of 0.51 and true negative (TN) rate of 0.49. Creditworthy people are better predicted. Due to the low TN rate, we can affirm that this model underpredicts non-creditworthy people. The overall accuracy, as shown in Figure 13 is 0.76.

Confusion matrix of Stepwise_Log		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Figure 14. Confusion matrix for the logistic regression model.

3.B.2. Decision Tree Model

Figure 15 shows the confusion matrix for the decision tree model, with a true positive (TP) rate of 0.87, false positive (FP) rate of 0.53 and true negative (TN) rate of 0.47. Because of the low TN rate, we can affirm that this model underpredicts non-creditworthy people. The overall accuracy, as shown in Figure 13 is 0.7467.

Confusion matrix of Decision_Tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Figure 15. Confusion matrix for the decision tree model.

3.B.3. Random Forest Model

Figure 16 shows the confusion matrix for the random forest model, with a true positive (TP) rate of 0.96, false positive (FP) rate of 0.58 and true negative (TN) rate of 0.42. The low TN rate shows that this model underpredicts non-creditworthy people. The overall accuracy, as shown in Figure 13 is 0.80.

Confusion matrix of Random_Forest		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Figure 16. Confusion matrix for the random forest model.

3.B.4. Boosted Model

Figure 17 shows the confusion matrix for the boosted model, with a true positive (TP) rate of 0.96, false positive (FP) rate of 0.62 and true negative (TN) rate of 0.38. This low TN rate reflects that this model underpredicts non-creditworthy people. The overall accuracy, as shown in Figure 13 is 0.7867.

Confusion matrix of Boosted		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Figure 17. Confusion matrix for the boosted model.

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"
The bias we are looking for is a short discussion about whether the model over or under predicts either result.

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:
  I chose the random forest model.
  - Overall Accuracy against your Validation set
    The overall accuracy of random forest is the highest, 0.8 followed by boosted model (0.7867), logistic regression (0.76) and decision tree (0.7467).

  - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    As shown in Figure 13, random forest presents the best tradeoff in accuracies creditworthy - non-creditworthy.

  - ROC graph
    Random forest has the second higher AUC, after boosted model, as shown in Figure 18.

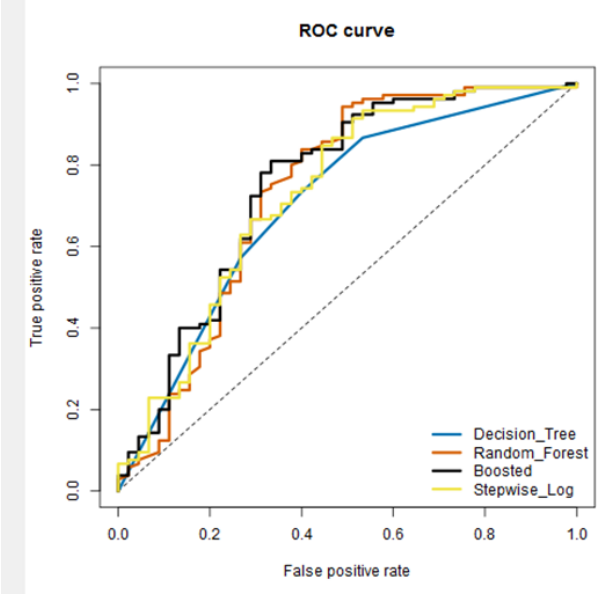

Figure 18. ROC curve for decision tree, random forest, boosted and stepwise logistic regression models.

- Bias in the Confusion Matrices
  The four models are biased, they underpredict the non-creditworthy people because most of the people got a loan in the past, but based on the accuracies, random forest is the best model.

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

- How many individuals are creditworthy? There are 406 creditworthy individuals out of 500, with the criterion than if the score of worthiness is greater than the non-worthiness the loan is given.

The workflows are shown in Figures 19 and 20.

References

https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

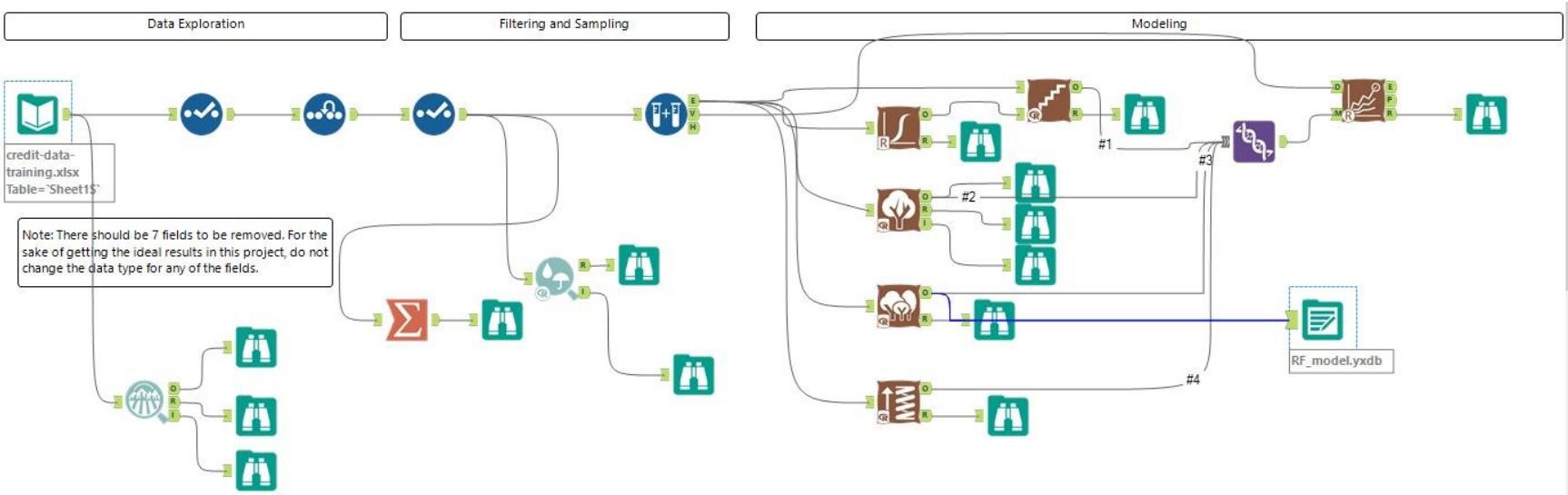

Figure 19. Alteryx workflow to explore, filter, sample and model the data with four binary models: stepwise logistic regression, decision tree, random forest and boosted model. At the end the performance of the models is compared.

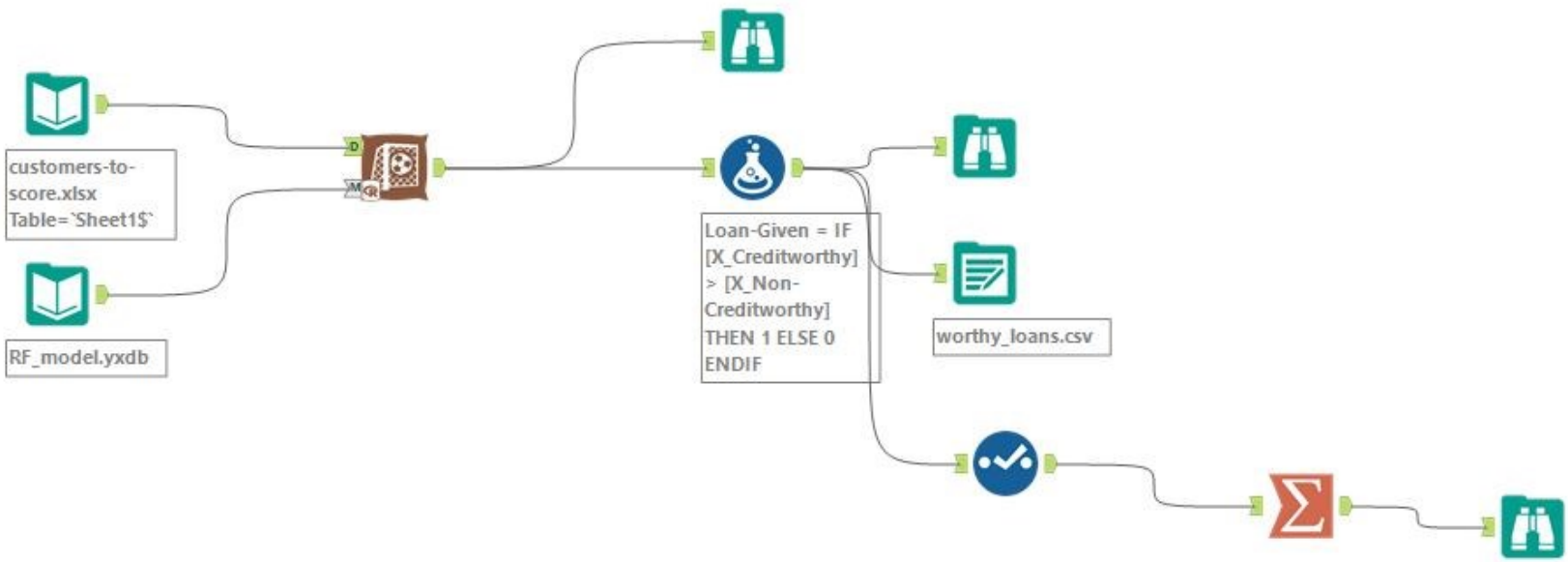

Figure 20. Alteryx workflow to score the new customers with the random forest model and count the number of people who are creditworthy.