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 need to be made?
  Our bank needs to produce an efficient solution to classify new customers on whether they can be approved for a loan or not. I systematically evaluated the creditworthiness of new loan applicants through classification modelling to determine if customers were creditworthy to give loans to. I used a series of classification models to figure out the best model and then generated a list of creditworthy customers.

- What data is needed to inform those decisions?
  I had data on all past applications and the list of customers which needed to be processed. The following pieces of information might be useful:
    - Availability of account balance
    - Duration of credit month
    - Payment status of previous credit
    - Purpose
    - Credit amount
    - Value of savings and stocks
    - Length of current employment
    - Instalment percent
    - Availability of guarantors
    - Duration in current address
    - Most valuable available asset category
    - Age in years
    - Type of apartment
    - Number of credits at this bank
    - Number of dependents
    - Foreign worker category

I did a quick check of the data on Excel and thought the concurrent credits and occupation categories were irrelevant as their data was entirely uniform. Additionally, there was no logical reason to include the telephone category.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We need to use a Binary Classification model as we need to determine whether a credit application is creditworthy or non-creditworthy (i.e. the credit application result is a binary variable).

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

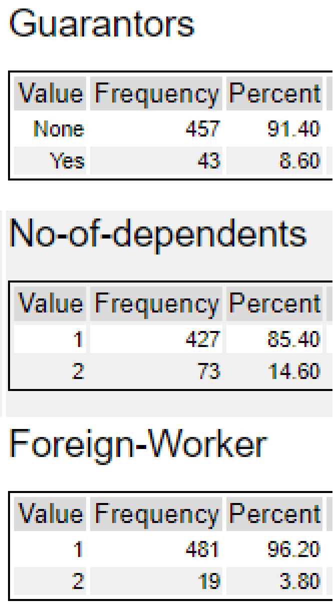
To achieve consistent results reviewers expect.

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.

Data Field Name	**Removed / Imputed**	**Reason**	**Explanation**	**Visualization**
Occupation; Concurrent-Credits	Removed	Contained only 1 unique value		
Duration-in-Current-address	Removed	68.8% data was missing		
Telephone	Removed	No logical reason to include it		
Guarantors; No-of-dependents; Foreign-Worker	Removed	Heavily skewed towards one value		Figure 2.1
Age-years	Imputed	2.4% data was missing	Missing values were imputed with the median (33) since	Figure 2.2

			the frequency distribution for the data was skewed. The median was more representative of the central location as the mean was dragged towards the skewed values.	



Guarantors

Value	Frequency	Percent
None	457	91.40
Yes	43	8.60

No-of-dependents

Value	Frequency	Percent
1	427	85.40
2	73	14.60

Foreign-Worker

Value	Frequency	Percent
1	481	96.20
2	19	3.80

Figure 2.1: Frequency Tables

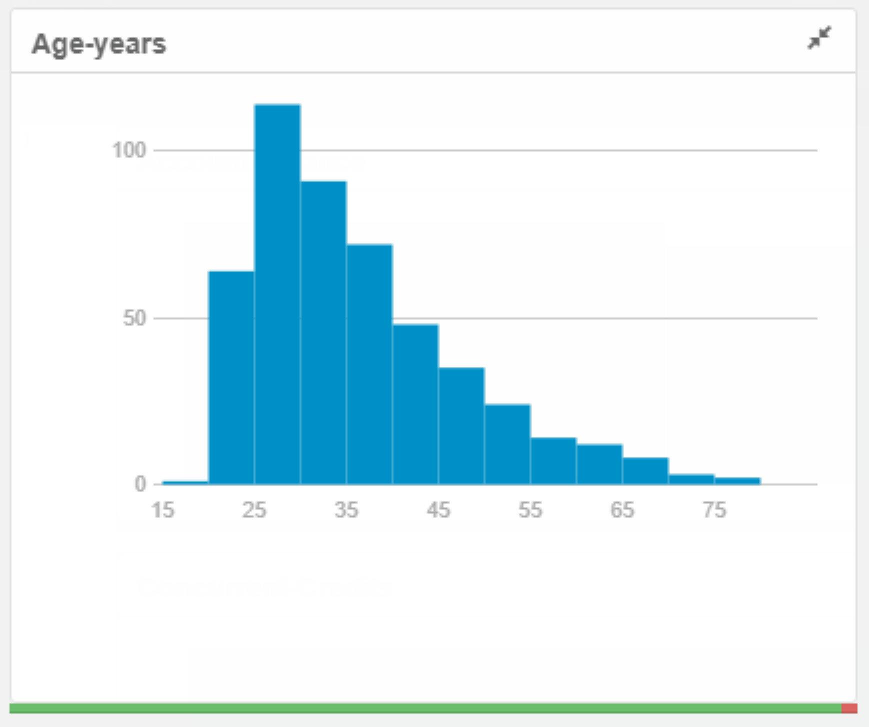


Figure 2.2: Right-Skewed Distribution of Age

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

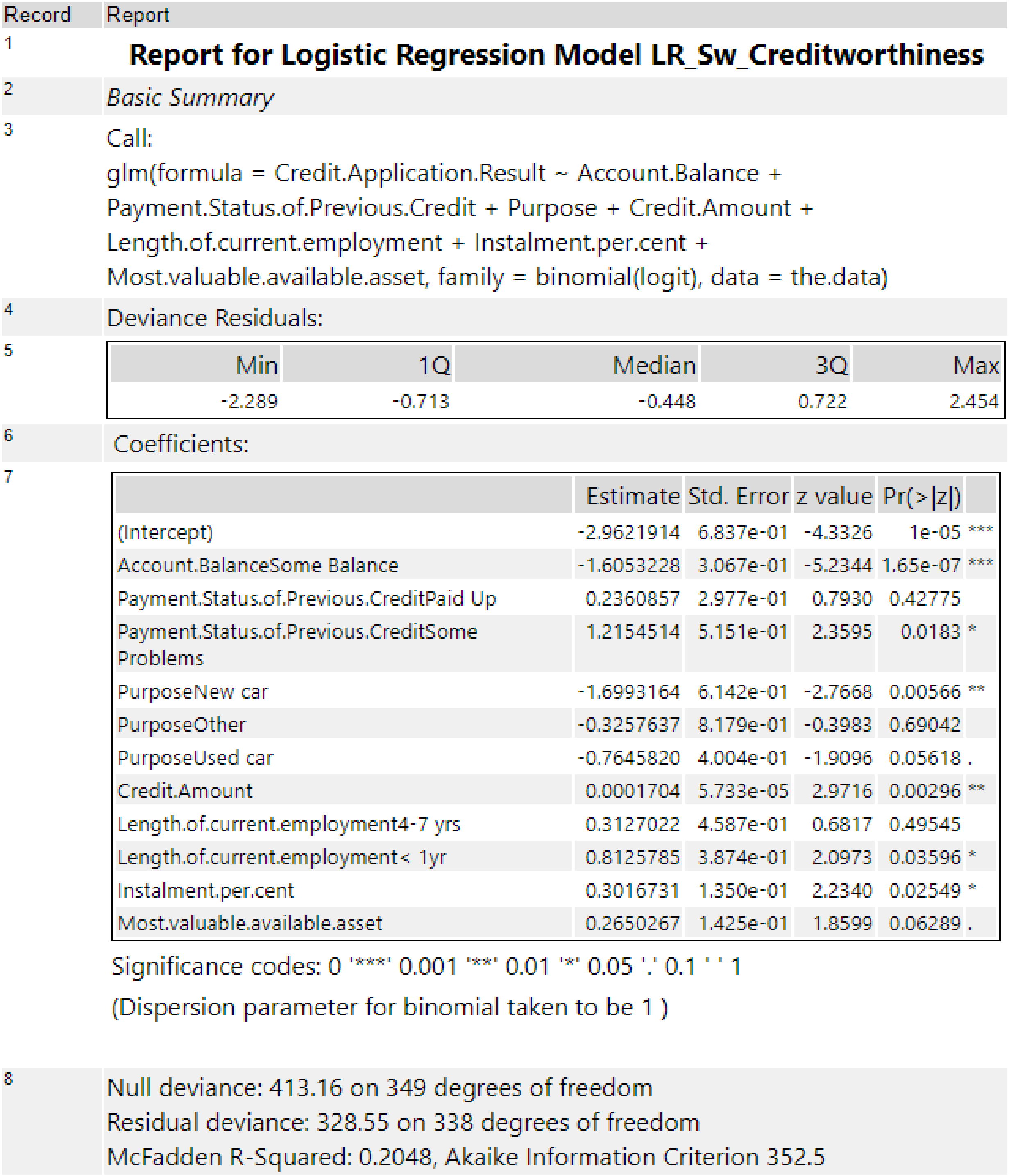
*Answer these questions for **each model** you created:*

- 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. Is there any bias seen in the model's predictions?

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

Logistic Regression - Stepwise Model



Record Report

1 **Report for Logistic Regression Model LR_Sw_Creditworthiness**

2 *Basic Summary*

3 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)

4 Deviance Residuals:

5

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

6 Coefficients:

7

	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)

8 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

Figure 3.1: Report for Logistic Regression - Stepwise Model

In Figure 3.1, Record 8 shows a very low R-Squared value of 0.20, and Record 7 shows significant predictor variables with p-values <= 0.05 (shown above as Pr(>|z|)):

Account-Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Length-of-current-employment, Instalment-per-cent and Most-valuable-available-asset.

Decision Tree

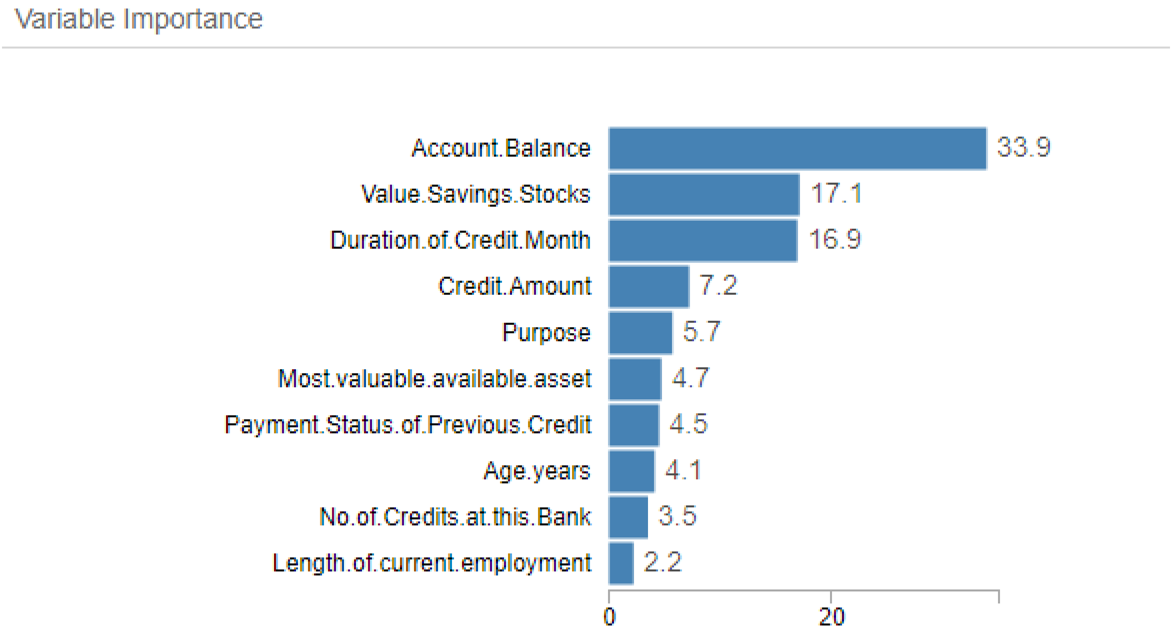


Figure 3.2: Variable Importance Plot for Decision Tree

Figure 3.2 shows the most important predictor variables are Account-Balance, Value-Savings-Stocks and Duration-of-Credit-Month.

Forest Model

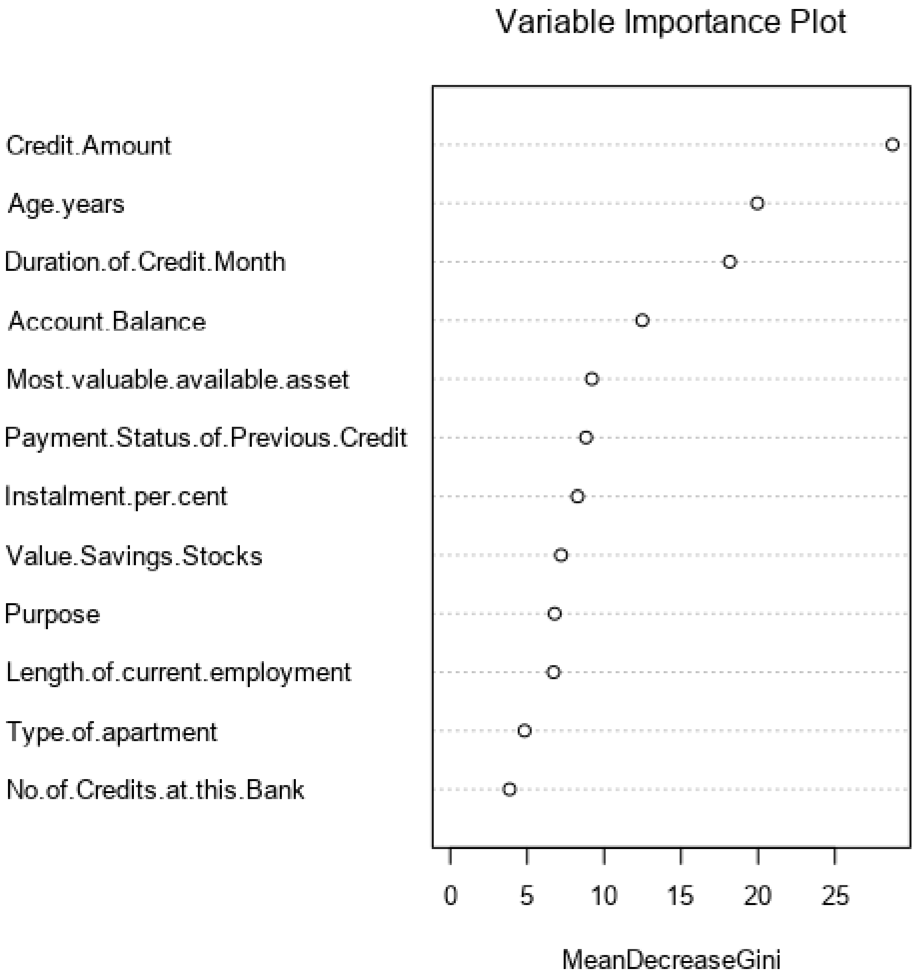


Figure 3.3: Variable Importance Plot for Forest Model

Figure 3.3 shows the most important predictor variables are Credit-Amount, Age-years, Account-Balance, Duration-of-Credit-Month and Account-Balance.

Boosted Model

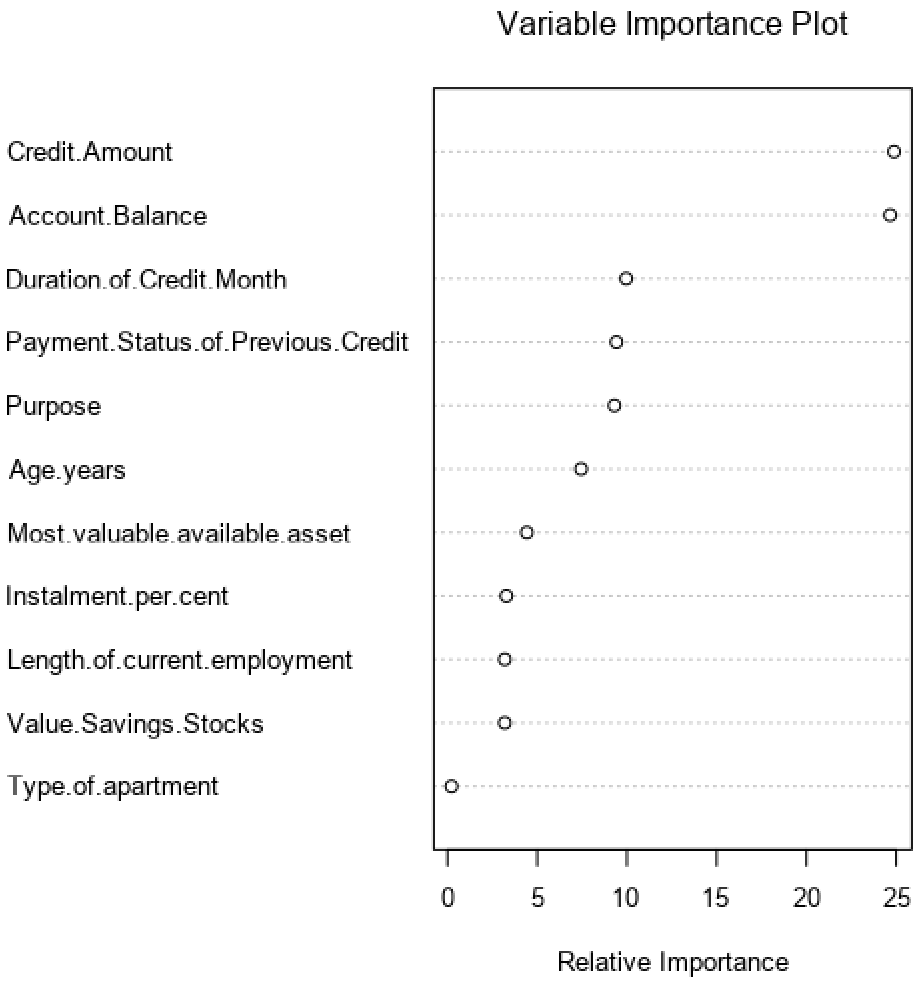
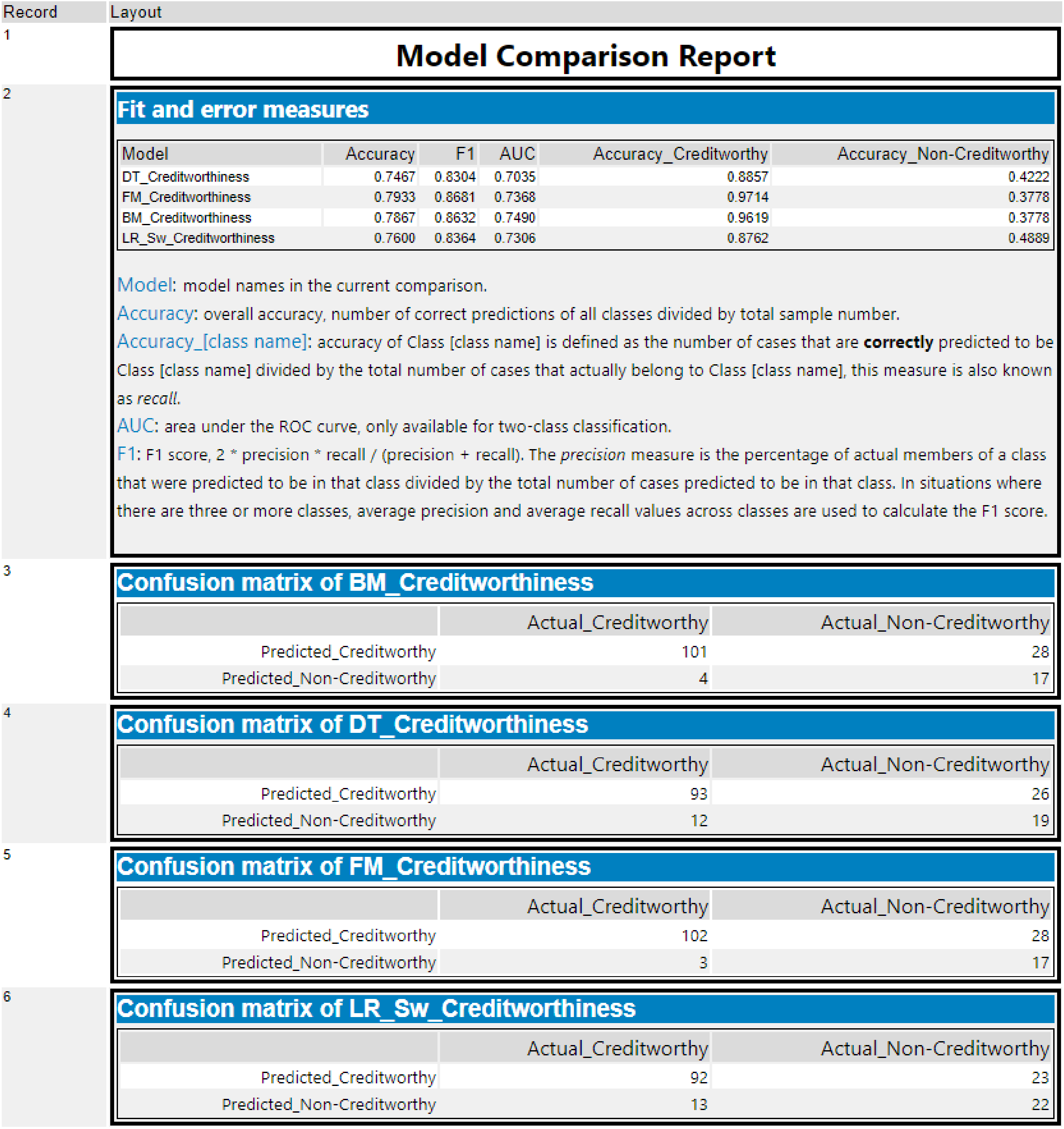


Figure 3.4: Variable Importance Plot for Boosted Model

Figure 3.4 shows the most important predictor variables are Credit-Amount, Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit and Purpose.

Model Comparison



Record	Layout
1	Model Comparison Report
2	Fit and error measures

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Creditworthiness	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Creditworthiness	0.7867	0.8632	0.7490	0.9619	0.3778
LR_Sw_Creditworthiness	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.

3

Confusion matrix of BM_Creditworthiness

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

4

Confusion matrix of DT_Creditworthiness

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

5

Confusion matrix of FM_Creditworthiness

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

6

Confusion matrix of LR_Sw_Creditworthiness

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

Figure 3.5: Model Comparison Report for Logistic Regression - Stepwise Model (LR_Sw_Creditworthiness), Decision Tree Model (DT_Creditworthiness), Forest Model (FM_Creditworthiness) and Boosted Model (BM_Creditworthiness)

Figure 3.5 Record 2 shows the overall accuracy of Decision Tree Model (74.67%), Forest Model (79.33%), Boosted Model (78.67%) and Logistic Regression - Stepwise Model (76.00%). Record 3 shows the confusion matrix of the Boosted Model, with True Positive Rate of 96.19% (101/105) and True Negative Rate of 37.78% (17/45). Record 4 shows the confusion matrix of the Decision Tree, with True Positive Rate of 88.57% (93/105) and True Negative Rate of 42.22% (19/45). Record 5 shows the confusion matrix of the Forest Model, with True Positive Rate of 97.14% (102/105) and True Negative Rate of 37.78% (17/45). Record 6 shows the confusion matrix of the Logistic Regression - Stepwise Model, with True Positive Rate of 87.62% (92/105) and True Negative Rate of 48.89% (22/45). There is some bias towards correctly identifying creditworthy individuals in all of the models' predictions as the accuracy of creditworthy individuals is much higher than the accuracy of non-creditworthy individuals.

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.

Model	**Overall Accuracy**	**Accuracy Within Creditworthy Segment**	**Accuracy Within Non-Creditworthy Segment**	**AUC**	**Bias in the Confusion Matrices**	**F1**
Forest Model	79.33% [1]	97.14% [1]	37.38% [3]	73.68% [2]	6.54% [2]	86.81% [1]
Boosted Model	78.67% [2]	96.19% [2]	37.38% [3]	74.90% [1]	2.66% [1]	86.32% [2]
Logistic Regression - Stepwise	76.00% [3]	87.62% [4]	48.89% [1]	73.06% [3]	17.14% [4]	83.64% [3]

Decision Tree	74.67% [4]	88.57% [3]	42.22% [2]	70.35% [4]	16.86% [3]	83.04% [4]

- Overall accuracy: the fraction of Creditworthy and Non-Creditworthy predictions the model got right.
- Accuracy within creditworthy/non-creditworthy segment: the fraction of Creditworthy/Non-Creditworthy predictions the model got right.
- AUC: area under the ROC curve, ranging from 0 to 1. The higher the AUC, the better the model.
- Bias in the confusion matrices: the difference between positive predictive value (PPV: the proportion of positive identifications which was actually correct) and negative predictive value (NPV: the proportion of negative identifications which was actually correct).
- F1: score calculated as 2 * *precision* * *recall / (precision* + *recall)*. Precision is the same as PPV, whereas recall is the proportion of actual positives which was identified correctly.

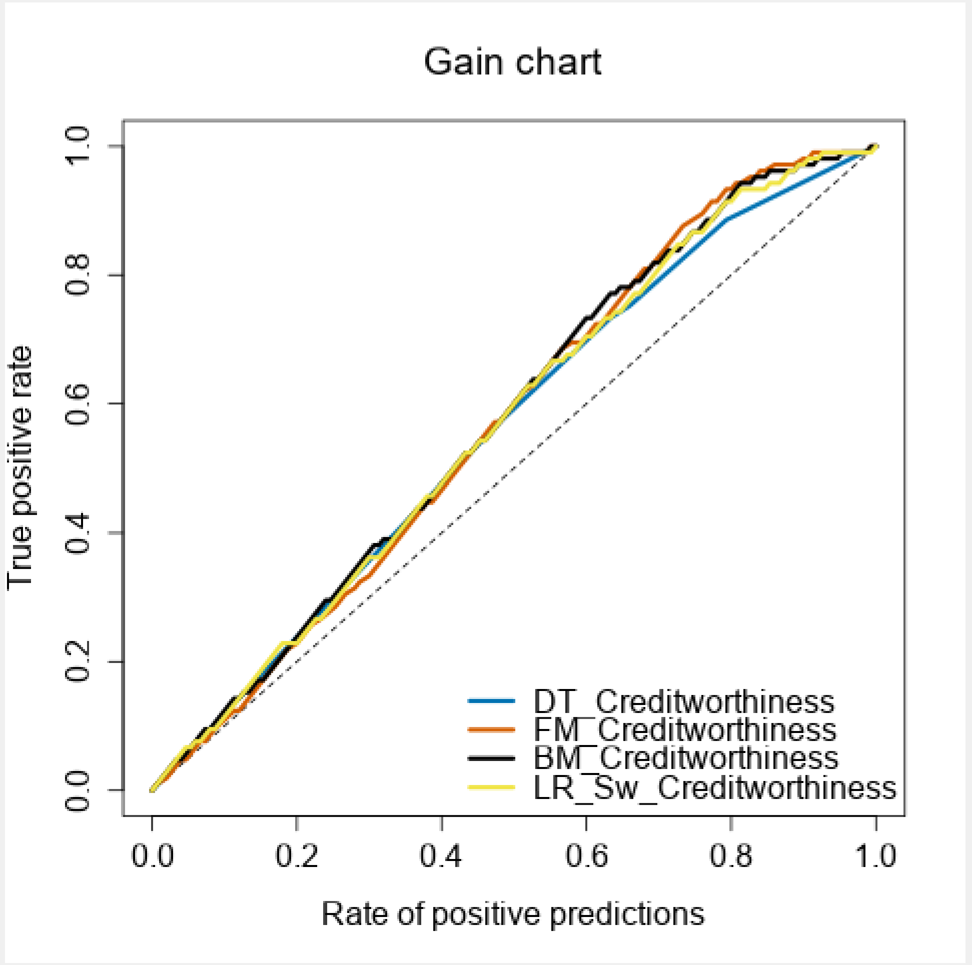


Figure 4.1: Gain Chart

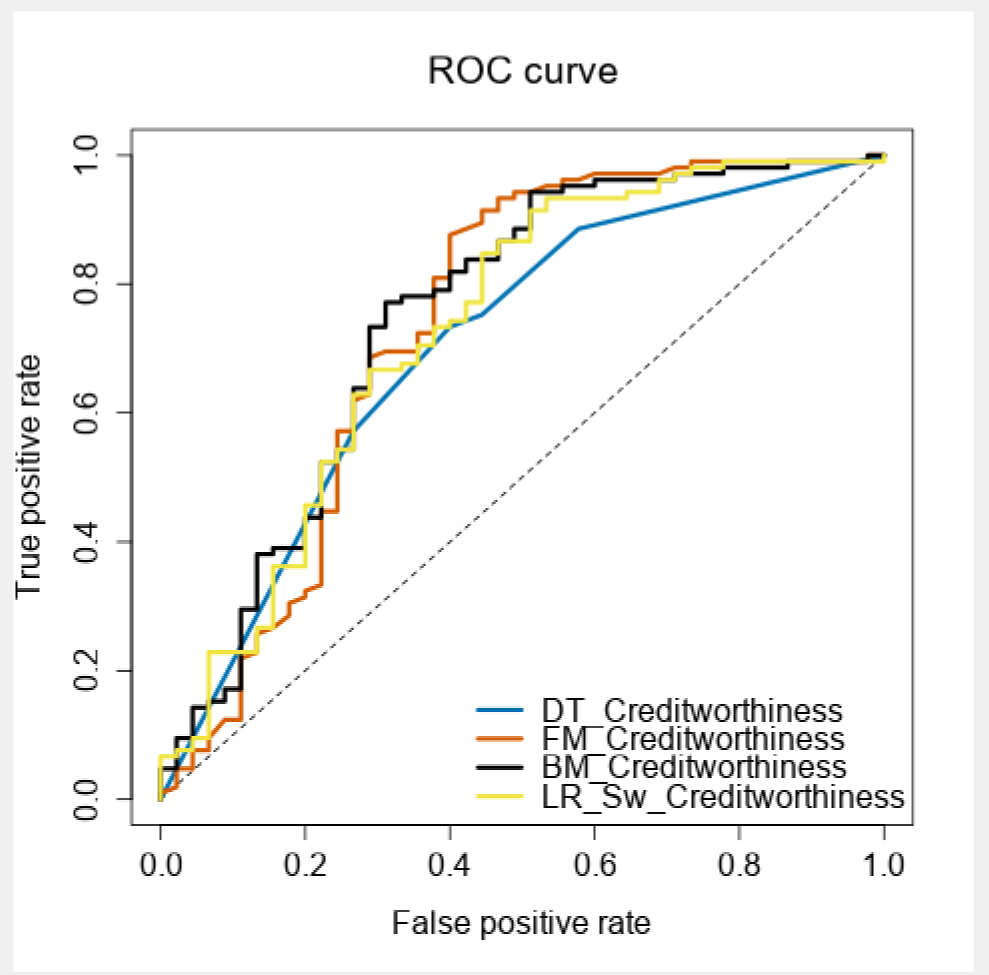
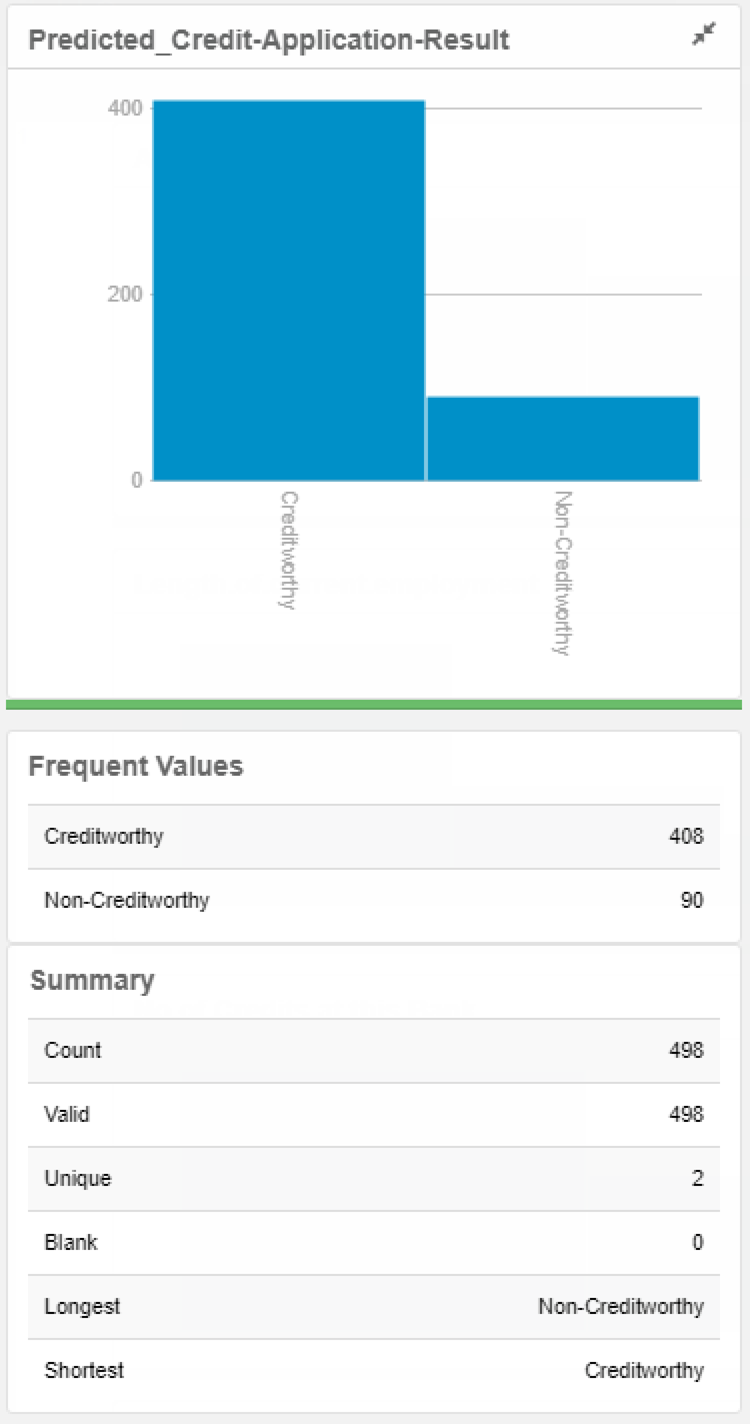


Figure 4.2: ROC Curve

I decided to choose the Forest Model. This model has the highest overall accuracy, accuracy within creditworthy segment and F1, as well as the second highest AUC and the second lowest bias in the confusion matrices. Additionally, this model appears to reach the top the quickest in Figure 4.1.

- How many individuals are creditworthy?



Predicted_Credit-Application-Result

Frequent Values

Creditworthy	408
Non-Creditworthy	90

Summary

Count	498
Valid	498
Unique	2
Blank	0
Longest	Non-Creditworthy
Shortest	Creditworthy

Figure 4.3: New Customers’ Credit Application Result Summary

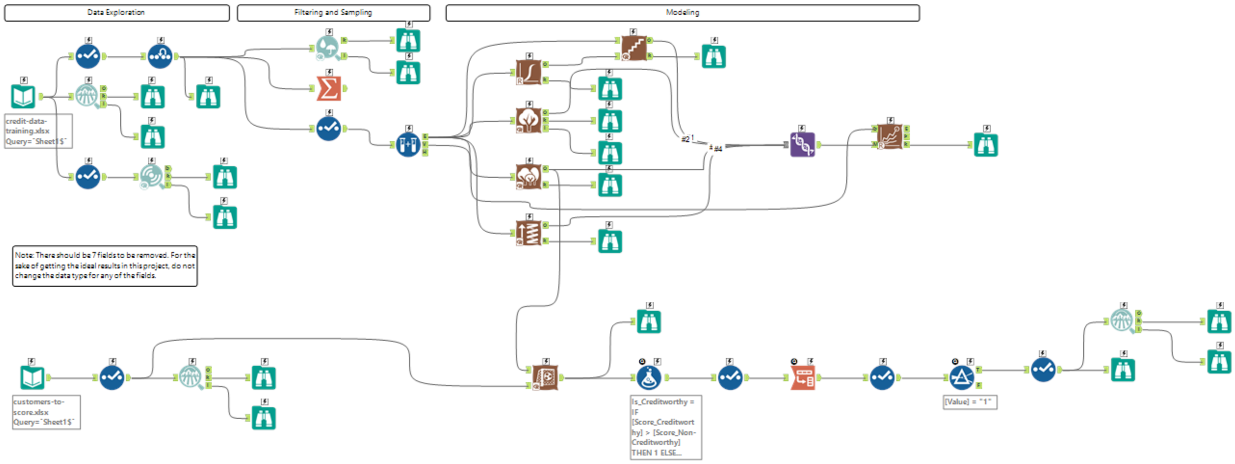
There are 408 creditworthy and 90 non-creditworthy individuals, as shown in Figure 4.3.

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.

Appendix

My Alteryx workflow:



Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.