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?

Find if a customer is credit worthy for loan after the rise of applicants to the bank.

- What data is needed to inform those decisions?
    1. Data of all past applicants (salary, age, account balance, credit amount …etc).
    2. List of customers who have applied to get a loan. This dataset has been scored with the model to get the list and number of final

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  This is a binary model since we want to predict if its credit worthy or not (two classification).

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.
  Answer:

  Fields to be removed:

  - Duration-in-current-address has 69% of its data are missing and that may make our predictive analysis insufficient.

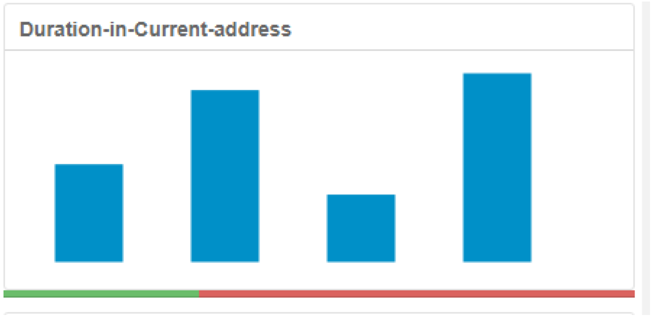


Figure 1 Field summary of Duration-in-current-address

  - No-of-dependents has law variability where data tend to one type.

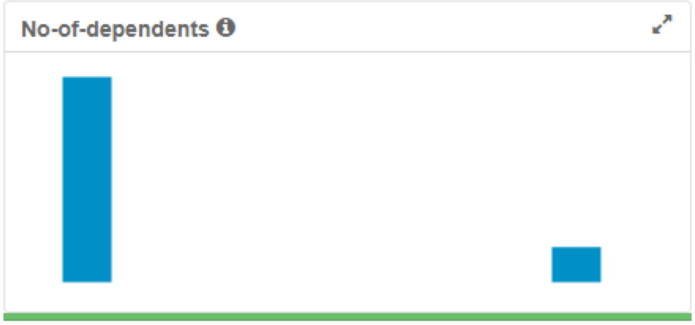


Figure 2 Field summary for No-of-dependents.

  - Guarantors has law variability where data tend to one type.

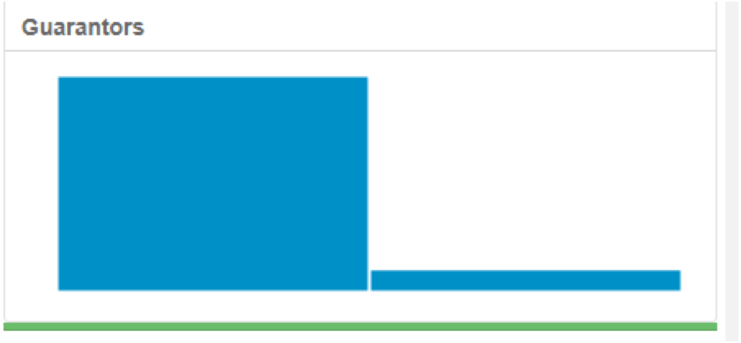


Figure 3 Field summary for Guarantors

- Foreign-worker has law variability where data tend to one type.

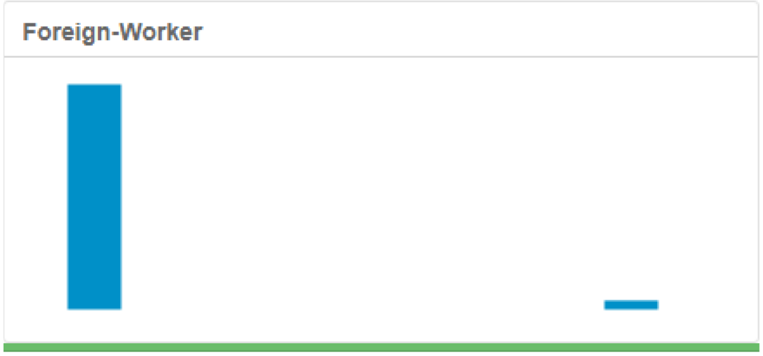


Figure 4 Field summary for Foreign-worker

- Concurrent-credits has law variability where data entirely uniform and there no other variation.

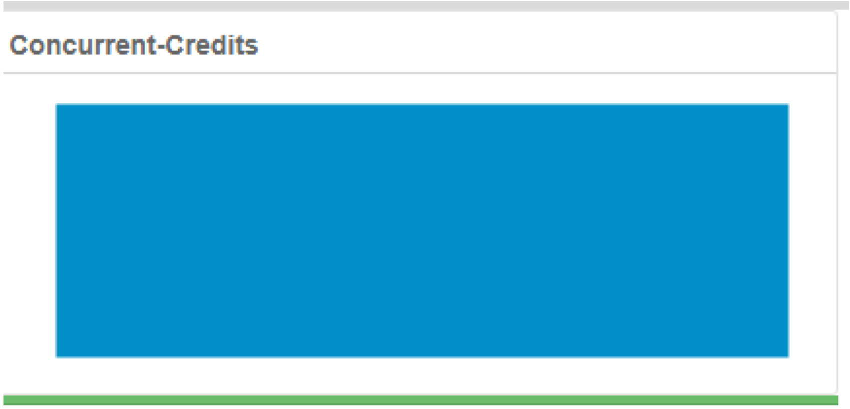


Figure 5 Field summary for Cuncurrent-credits

- Telephone there is no logical reason for including the variable.

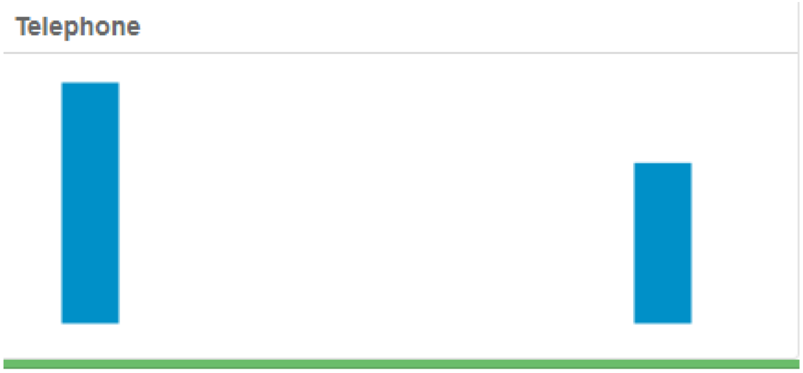


Figure 6 Field summary for Telephone

- Occupation has law variability where data entirely uniform and there no other variation.

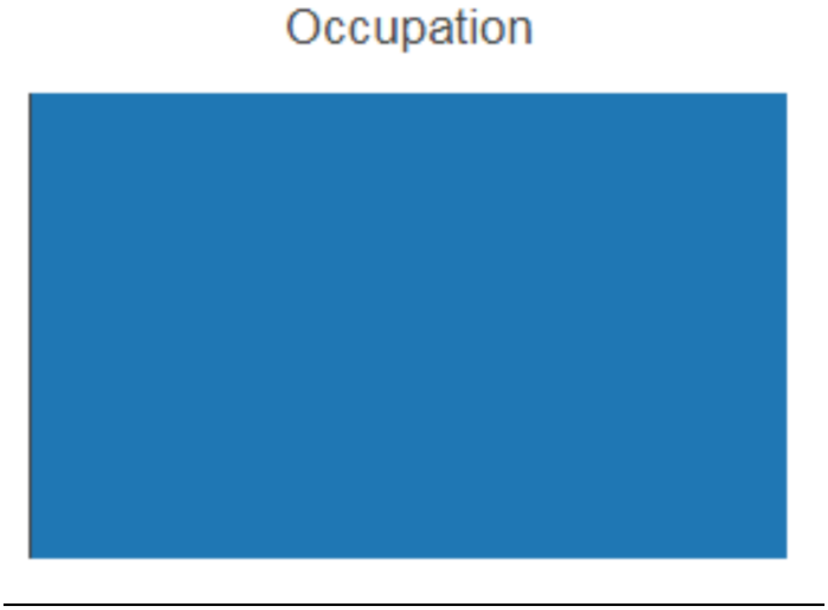


Figure 7 Feild Summary for Occupation

Fields to be impute:

- Age-years there are 2% of the data are missing and since the data are skewed to the right the mean is biased to the right values. We impute the missing values with the median of Age-years and its 33.

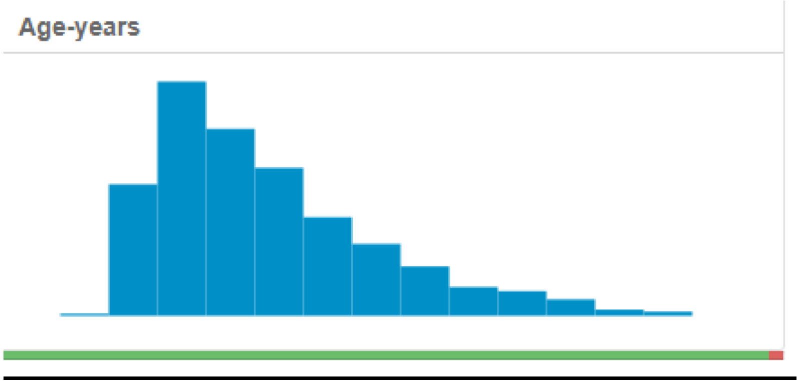


Figure 8 Field summary for Age-years

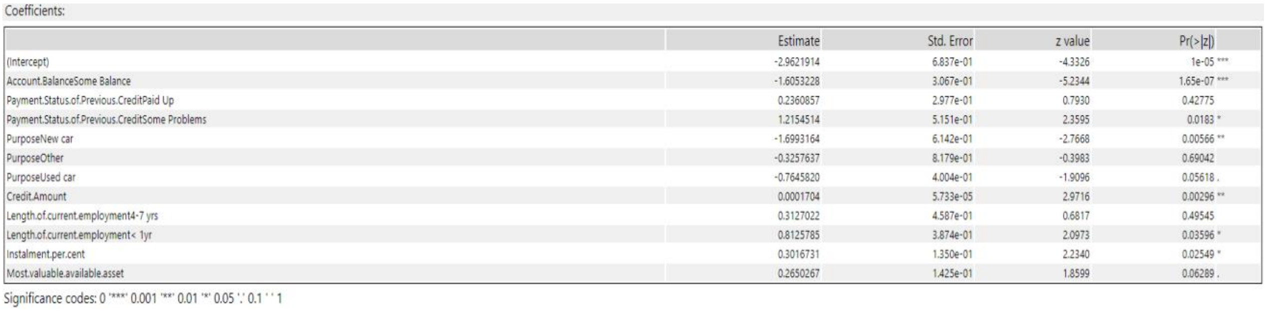
Step 3: Train your Classification Models

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.
    1. **Logistic regression:**
        - Account balance.
        - Purpose.
        - Credit amount



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

Figure 9 Predictor variables foe Logistic regression model

    2. **Decision tree:**
        - Account balance.
        - Value saving stocks.
        - Duration of credit month.

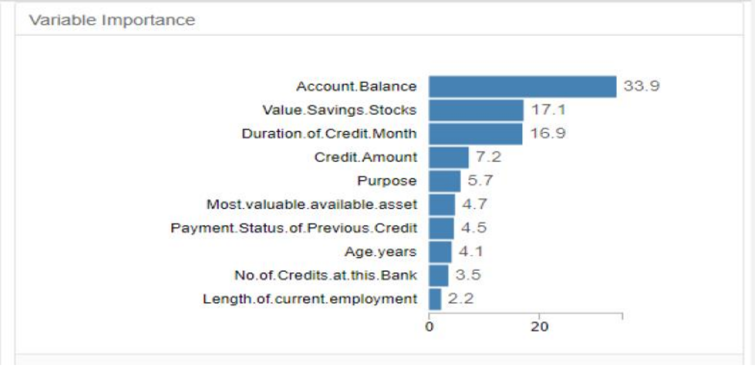


Figure 10 Predictor variables for Decision tree

    3. **Forest model:**
        - Credit amount.
        - Age years.
        - Duration of credit month.

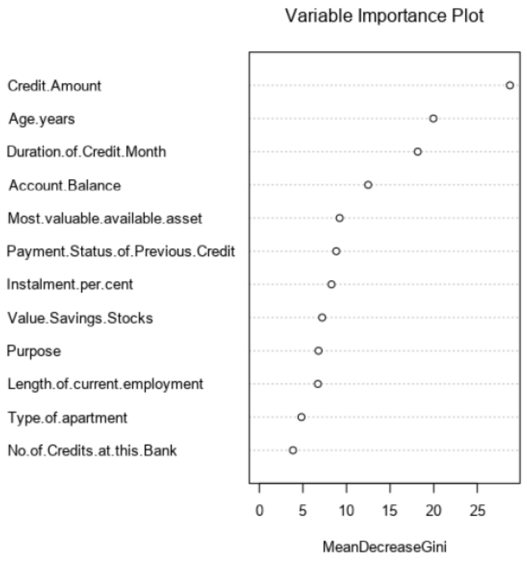


Figure 11 Predictor variables for Forest model

4. **Boosted model:**
   - Credit amount.
   - Account balance.
   - Duration of credit month.

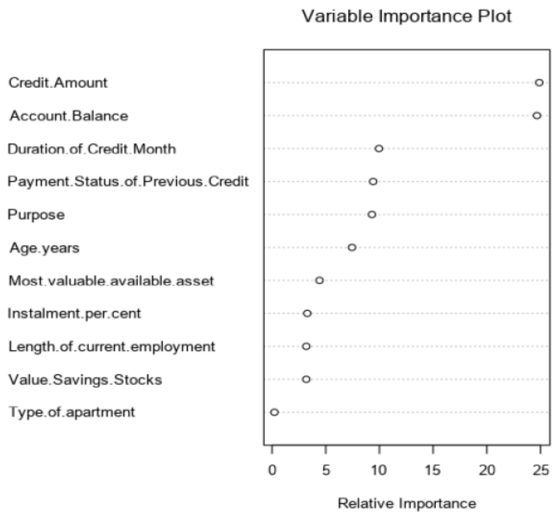
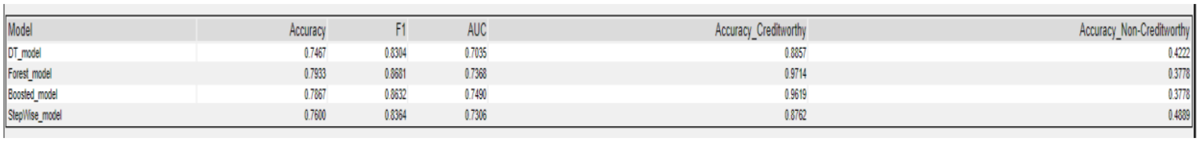


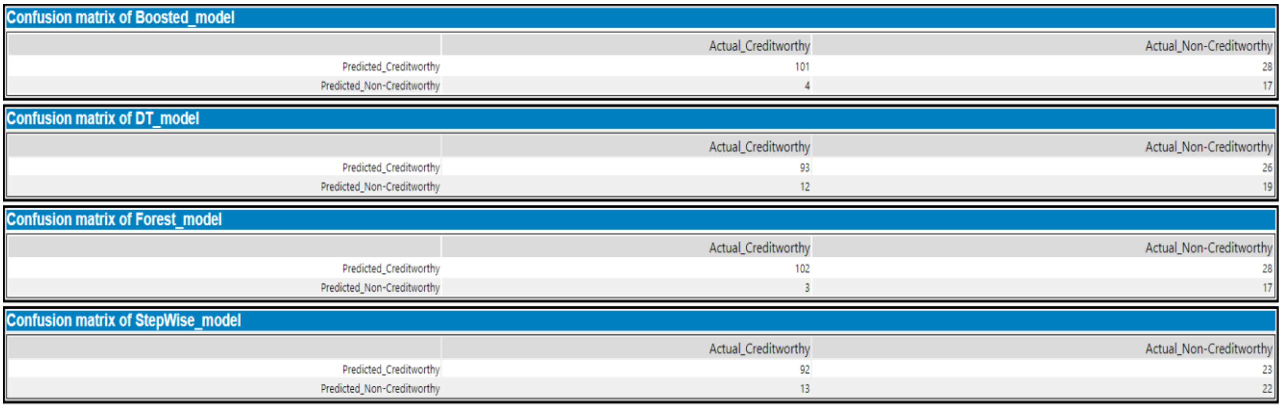
Figure 12 Predictor variables for Boosted model

- 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?



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_model	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_model	0.7867	0.8632	0.7490	0.9619	0.3778
StepWise_model	0.7600	0.8364	0.7306	0.8762	0.4889

Figure 13 Models with their corresponding accuracy



Confusion matrix of Boosted_model

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

Confusion matrix of DT_model

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

Confusion matrix of Forest_model

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

Confusion matrix of StepWise_model

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

Figure 14 Confusion matrix for each model.

1. Logistic regression (StepWise).

The overall accuracy is around 76.0% while accuracy for creditworthy is higher than non-creditworthy at 88% and 49% respectively. The model is biased towards predicting customers as creditworthy.

2. Decision tree.

The overall accuracy is around 75%. Accuracy for creditworthy is 89% while accuracy for non-creditworthy is 42%. The model seems to be biased towards predicting customers as creditworthy.

3. Forest model.

The overall accuracy is around 79%. Accuracy for creditworthy is 97% while accuracy for non-creditworthy is 38%. The model seems to be biased towards predicting customers as creditworthy.

4. Boosted model.

The overall accuracy is around 78%. Accuracy for creditworthy is 96% while accuracy for non-creditworthy is 38%. The model seems to be biased towards predicting customers as creditworthy.

All models are biased because the number of records which are Creditworthy is much higher than Non-Creditworthy in the training dataset.

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"

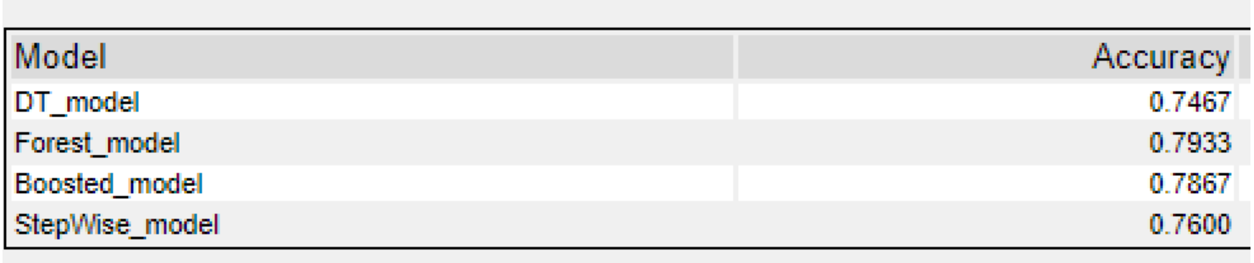
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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

Answer:

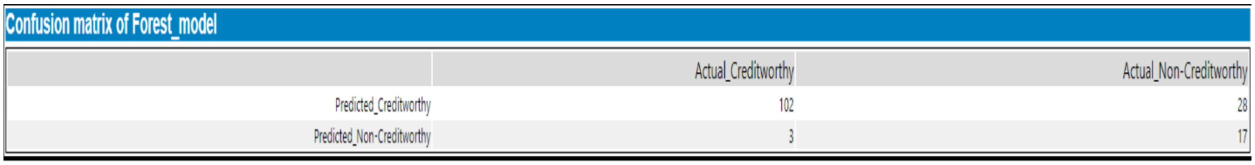
Forest model is chosen as it offers the highest accuracy at 79% against validation set.
Its accuracies for creditworthy and non-creditworthy are among the highest of all.



Model	Accuracy
DT_model	0.7467
Forest_model	0.7933
Boosted_model	0.7867
StepWise_model	0.7600

Figure 15 Each model with its corresponding accuracy.

Forest model has less bias in its prediction compared to the other model, we can see that they are 3 records were predicted as Non-creditworthy and they were actually Creditworthy. And 28 records predicted Creditworthy and they actually Non-Creditworthy



Confusion matrix of Forest_model

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

Figure 16 Confusion matrix for Forest model

Forest model reaches the true positive rate at the fastest rate. This evident is supported by the ROC graph in Figure 17.

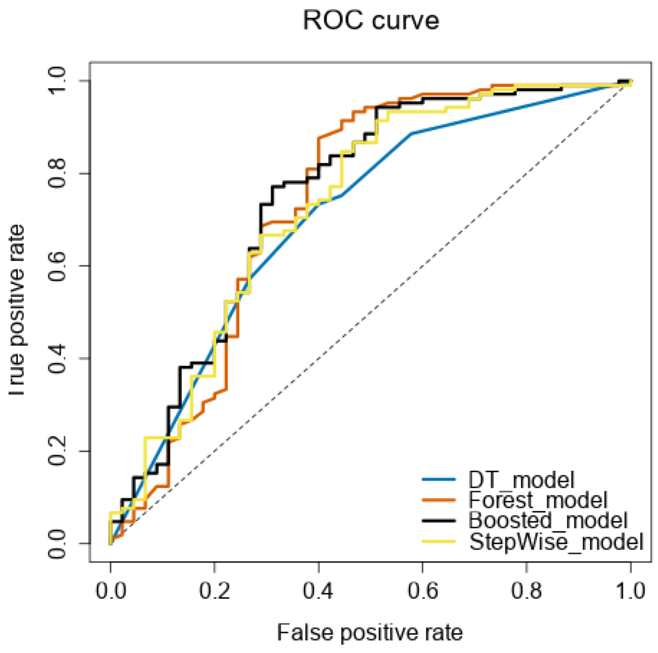


Figure 17 ROC graph

- How many individuals are creditworthy?
  408 Creditworthy

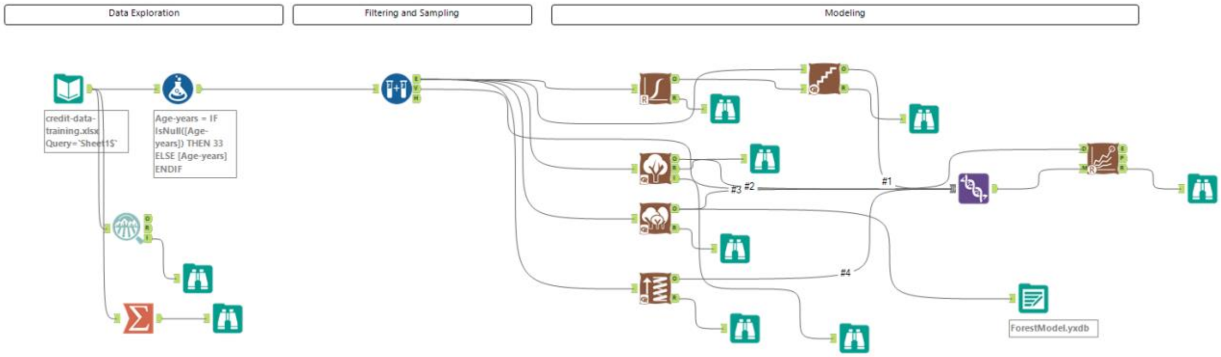

Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx
Query=`Sheet1$`
Age-years = IF IsNull([Age-years]) THEN 33 ELSE [Age-years] ENDIF
#3
#2
#1
#4
ForestModel.yxdb

Figure 18 Project workflow

<u>Before you Submit</u>

Please check your answers against the requirements of the project dictated by the [rubric]here. Reviewers will use this rubric to grade your project.