Farida Muhammad

Project Creditworthiness Submission

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?**

  As a loan officer at a small bank, I must determine if customers are creditworthy to give a loan. I must accurately identify people who qualify and do not qualify for loans for my manager. I will need to:

  - Come up with an efficient solution to classify new customers on whether they can be approved for a loan or not.
  - Use a series of classification models to figure out the best model
  - Provide my manager with a list of creditworthy customers

- **What data is needed to inform those decisions?**

  I will use the Alteryx software to build predictive classification models using Logistic Regression, Decision Tree, Random Forest, and Boosted Model to evaluate the creditworthiness of these new loan applicants. To properly build the models and select predictor variables, I will explore and cleanup my data. The following data will be used to build the predictive classification models and make the decisions needed for this project:

  - ***credit-data-training.xlsx*** - This file contains all credit approvals from past loan applicants the bank has ever completed.
  - ***customers-to-score.xlsx*** - This is the new set of customers that I will need to score on the classification model I will create.

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

  The outcome to determine whether customers are creditworthy or not to give a loan, which is binary. Therefore, I would use a binary classification model.

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.

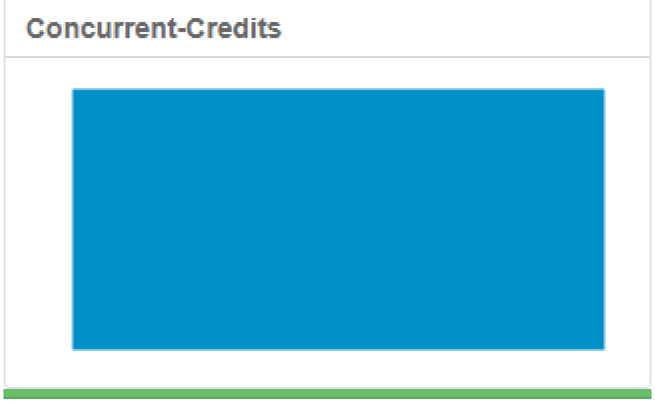
Answer this question:

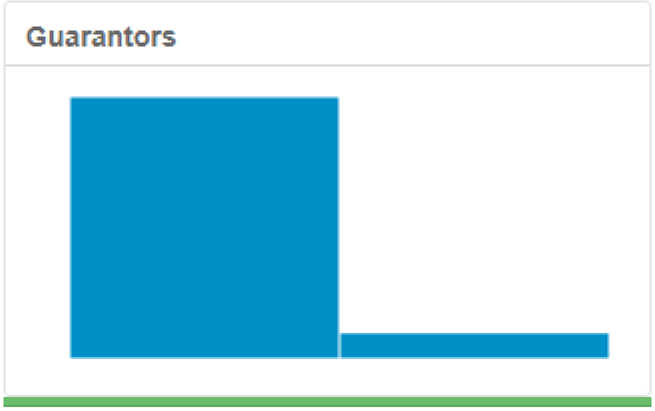
- **In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.**

I did not find any numerical data fields that were highly-correlated (.70) with each other. I removed seven fields from the dataset.

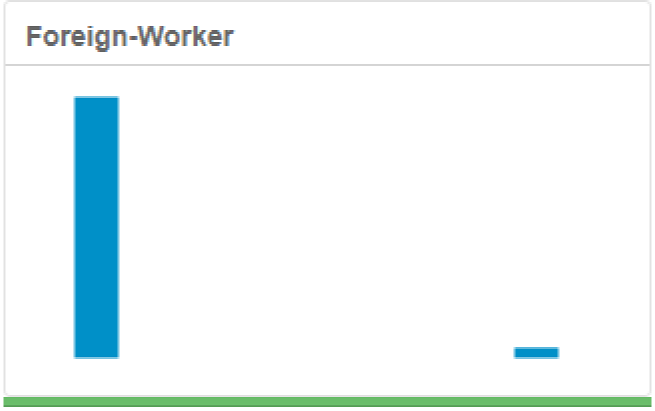
The following fields with few values (**low variability**) in the subset of my data were removed:

- Concurrent-credits, guarantors, Foreign-worker, No-of-dependents, and occupation.

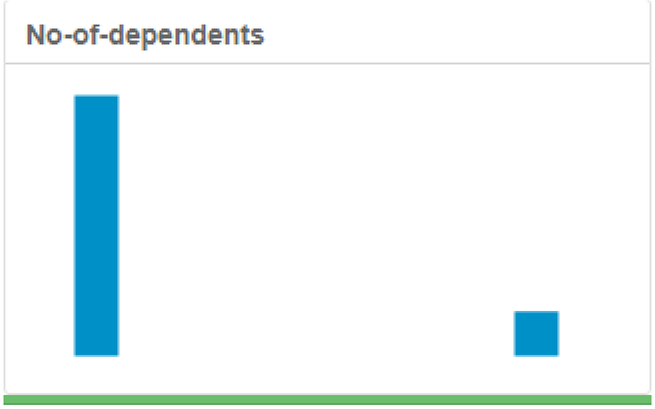




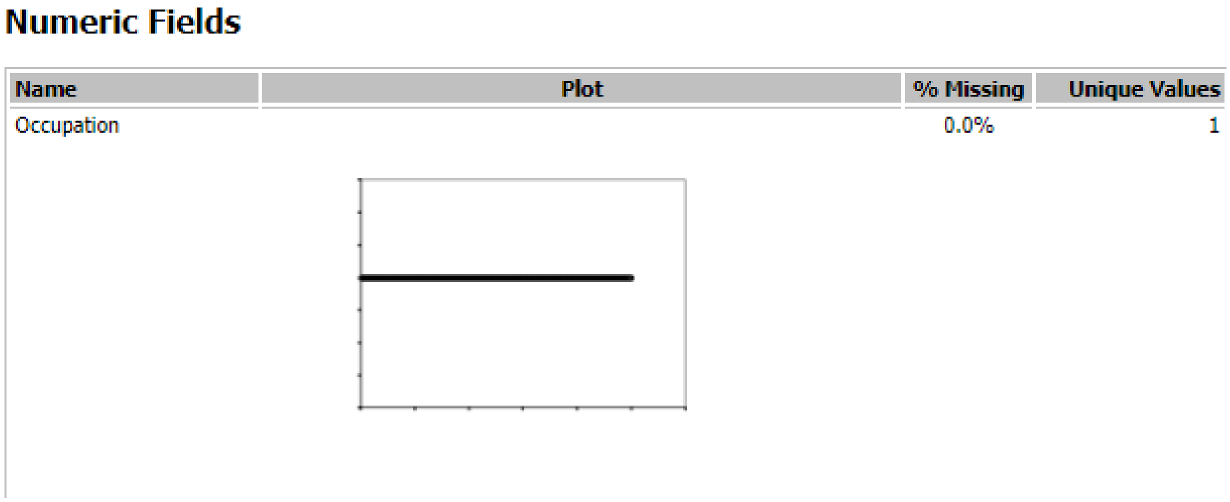
Guarantors



Foreign-Worker



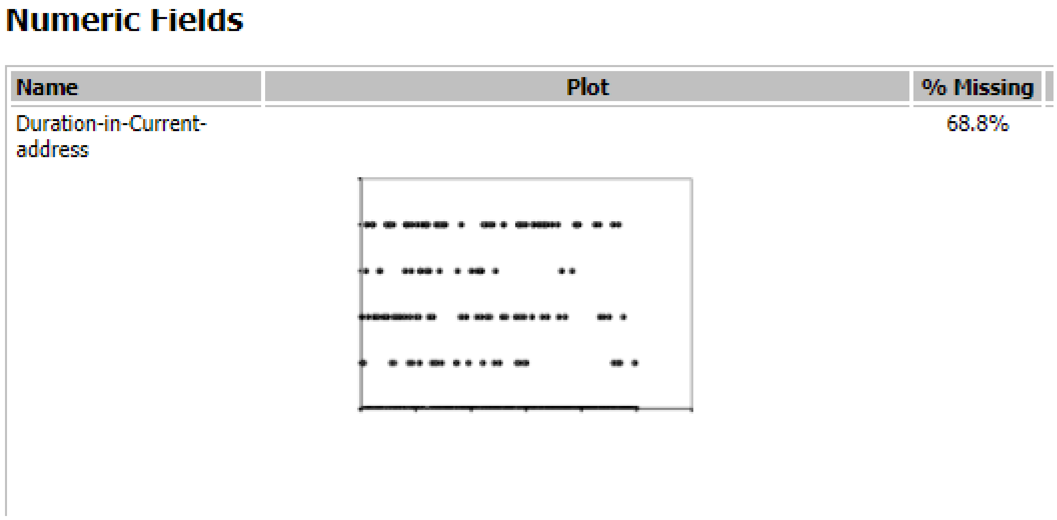
No-of-dependents



Numeric Fields

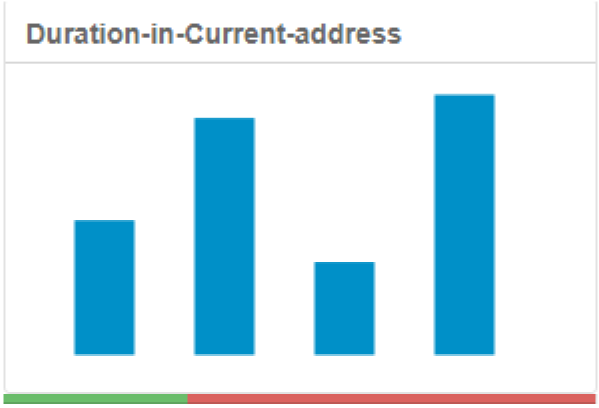
Name	Plot	% Missing	Unique Values
Occupation		0.0%	1

- The following field was removed because it had too many **missing values** 68.8**%**: **Duration-in-Current-address**



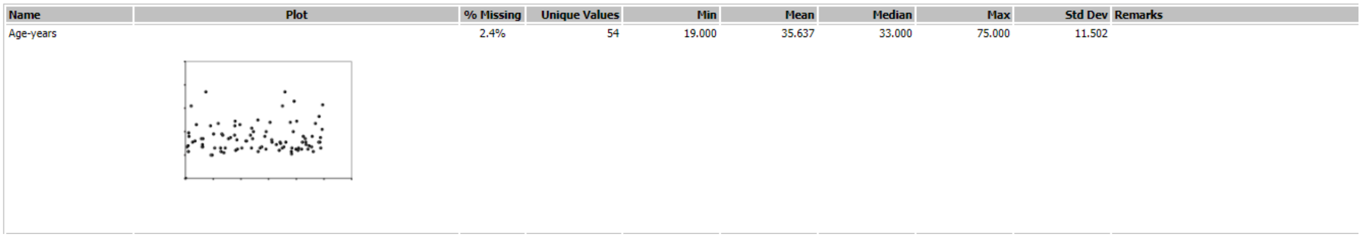
Numeric Fields

Name	Plot	% Missing
Duration-in-Current-address		68.8%

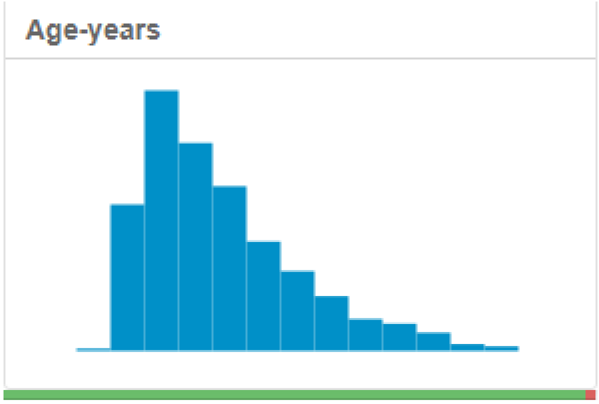


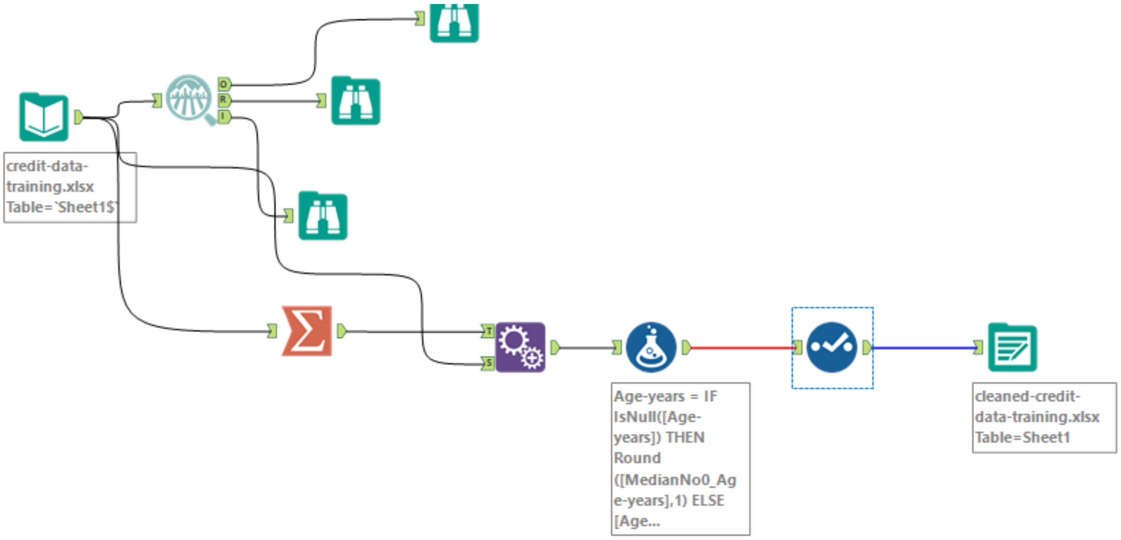
- The removed **Telephone** because there is **no logical reason** for including the variable - **Telephone**

- I imputed **Age-years** with the median value of 33 because it only had 2.4% missing values.



Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	




credit-data-training.xlsx Table=`Sheet1$`
O
R
I
T
S
Age-years = IF IsNull([Age-years]) THEN Round ([MedianNo0_Age-years],1) ELSE [Age...
cleaned-credit-data-training.xlsx Table=Sheet1

Overall, my clean data set has 13 columns.

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:*

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

2. **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**?

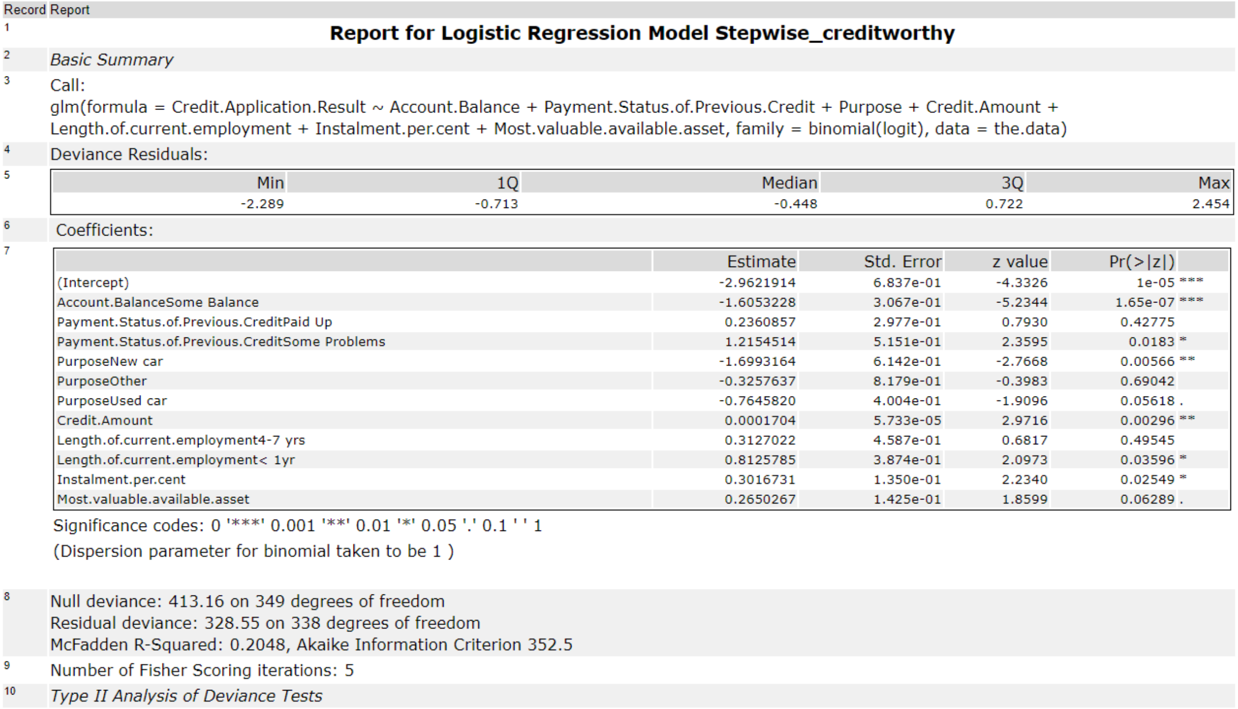
Logistic Regression Model

Significant Predictor Variables

After running the Logistic Regression Model through the stepwise tool, the following **predictor variables** were **removed**: Duration-of-Credit-Month, Value-Savings-Stocks, Age-years, and Type-of-apartment, and No-of-Credits-at-this-Bank

The following **significant predictor variables** remained in the prediction model: Account-Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset

The most important predictor variables based on the **significance** codes are: Account-Balance (Some Balance), Payment-Status-of-Previous-Credit (Some Problems), Purpose (new car), Credit-Amount, Length-of-current-employment (< 1 yr), Instalment-per-cent



Record Report

1 **Report for Logistic Regression Model Stepwise_creditworthy**

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)

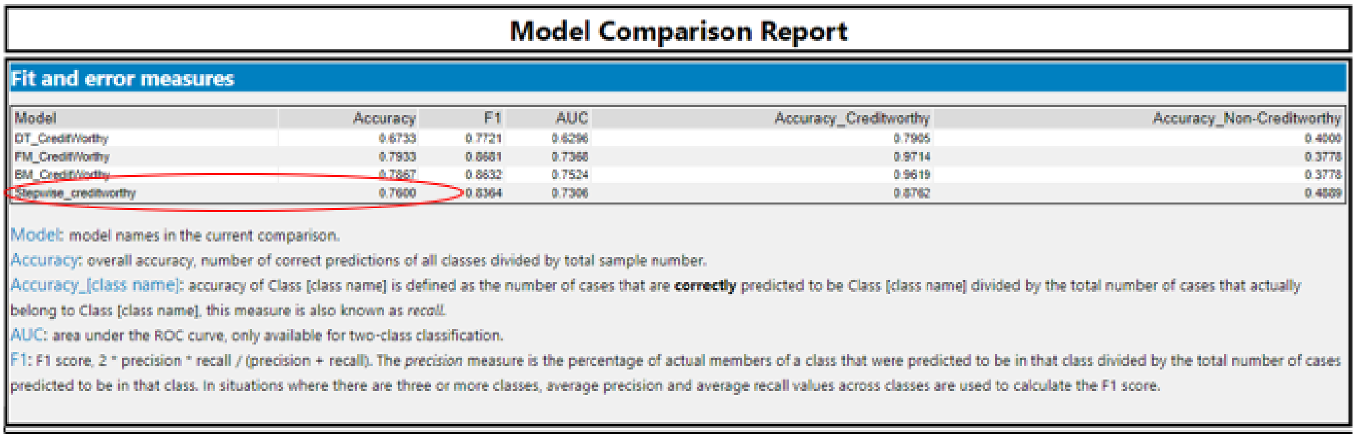
8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

Model Validation

After running this model through the stepwise tool and validating it against the validation set, the overall accuracy of the model is 76%.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CreditWorthy	0.6733	0.7721	0.6296	0.7905	0.4000
FM_CreditWorthy	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889

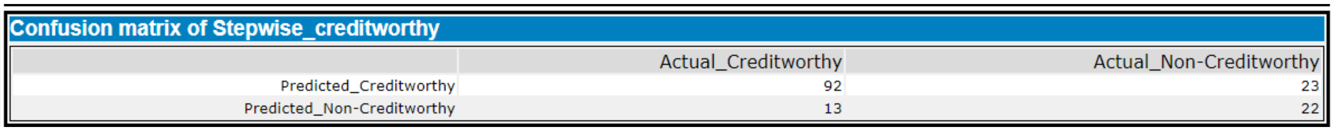
Model: model names in the current comparison.

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.



Confusion matrix of Stepwise_creditworthy

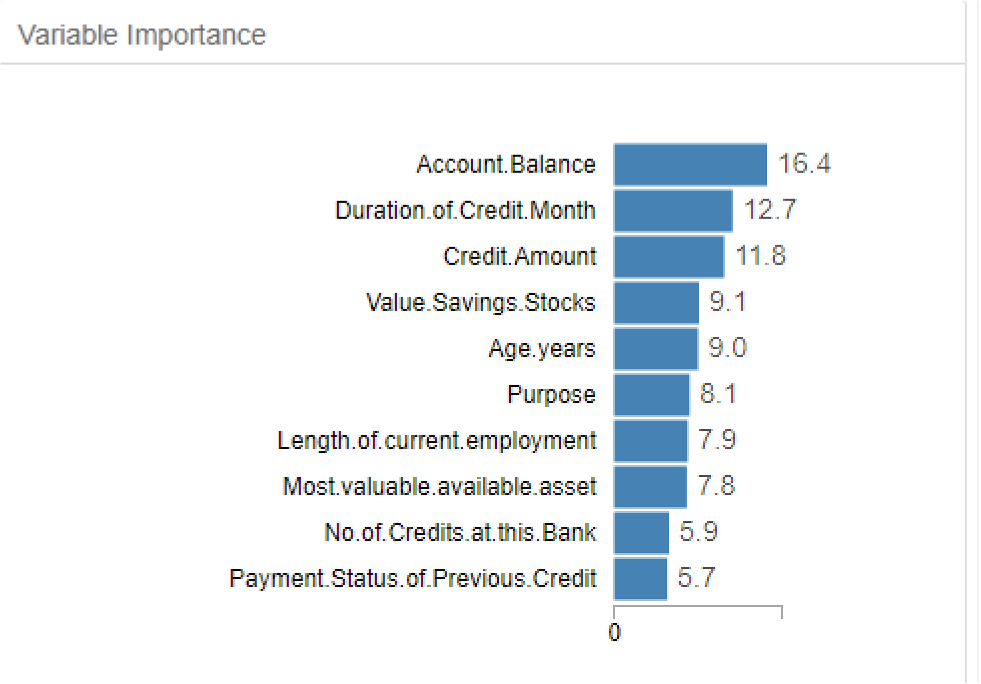
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

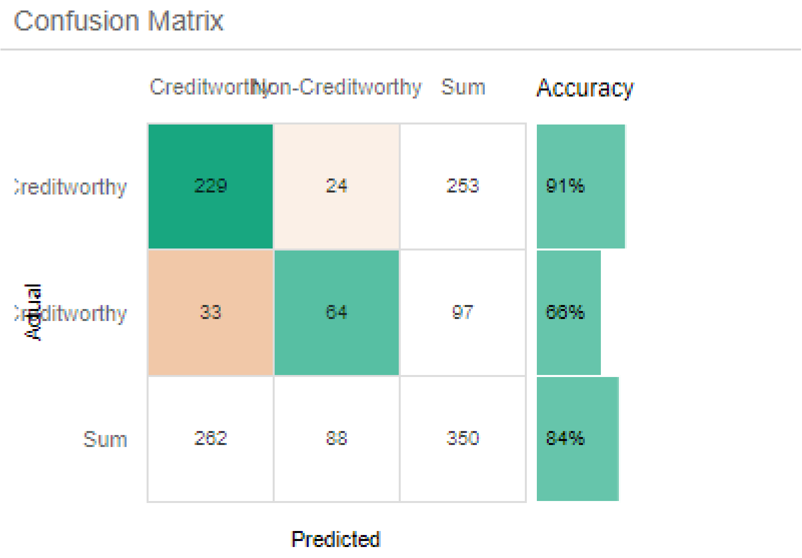
The creditworthy prediction was relatively strong at 87.62%. It was more difficult to predict non-creditworthy applicants accurately at 48.89%.

Decision Tree Model

Significant Predictor Variables

The **most important** predictor variables in the Decision Tree Model are: Account-Balance, Duration-of-Credit-Month, and Credit-Amount.





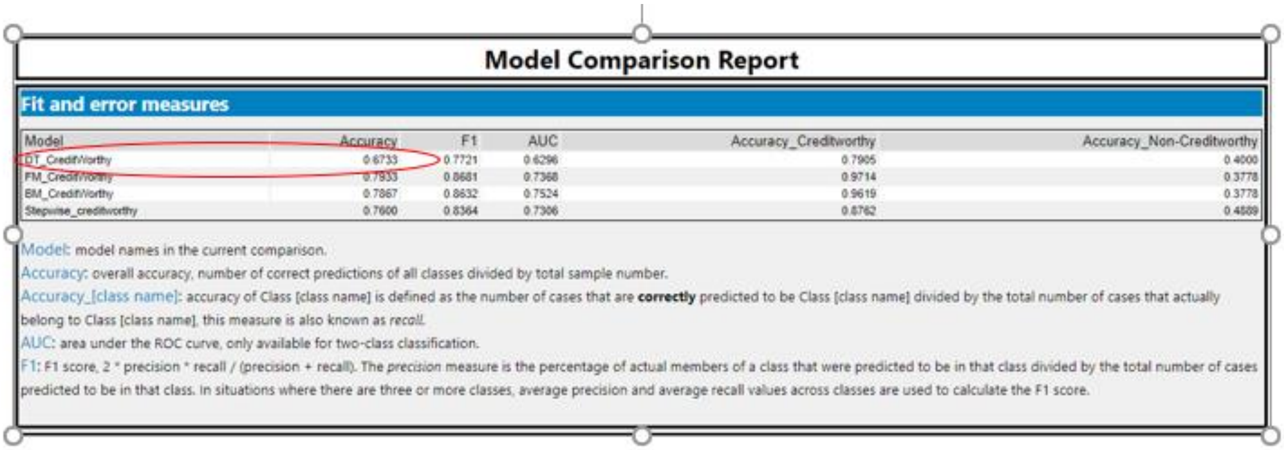
Confusion Matrix

Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	229	24	253	91%
ditworthy	33	64	97	66%
Sum	262	88	350	84%

Predicted

Model Validation

After running this model against the validation set, the overall percent accuracy is 67.33%. This overall accuracy rate (67.33) is lower than the accuracy rate predicted in the estimation set (84%). Decision tree models tends to overfit the sample data.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CreditWorthy	0.6733	0.7721	0.6296	0.7905	0.4000
FM_CreditWorthy	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889

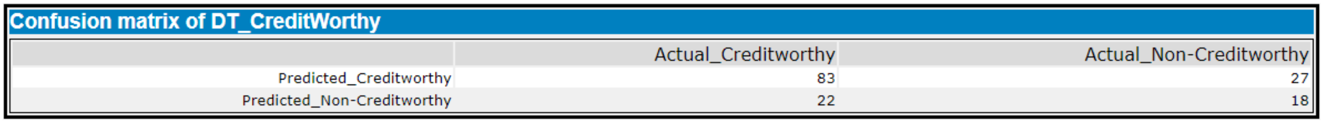
Model: model names in the current comparison.

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.



Confusion matrix of DT_CreditWorthy

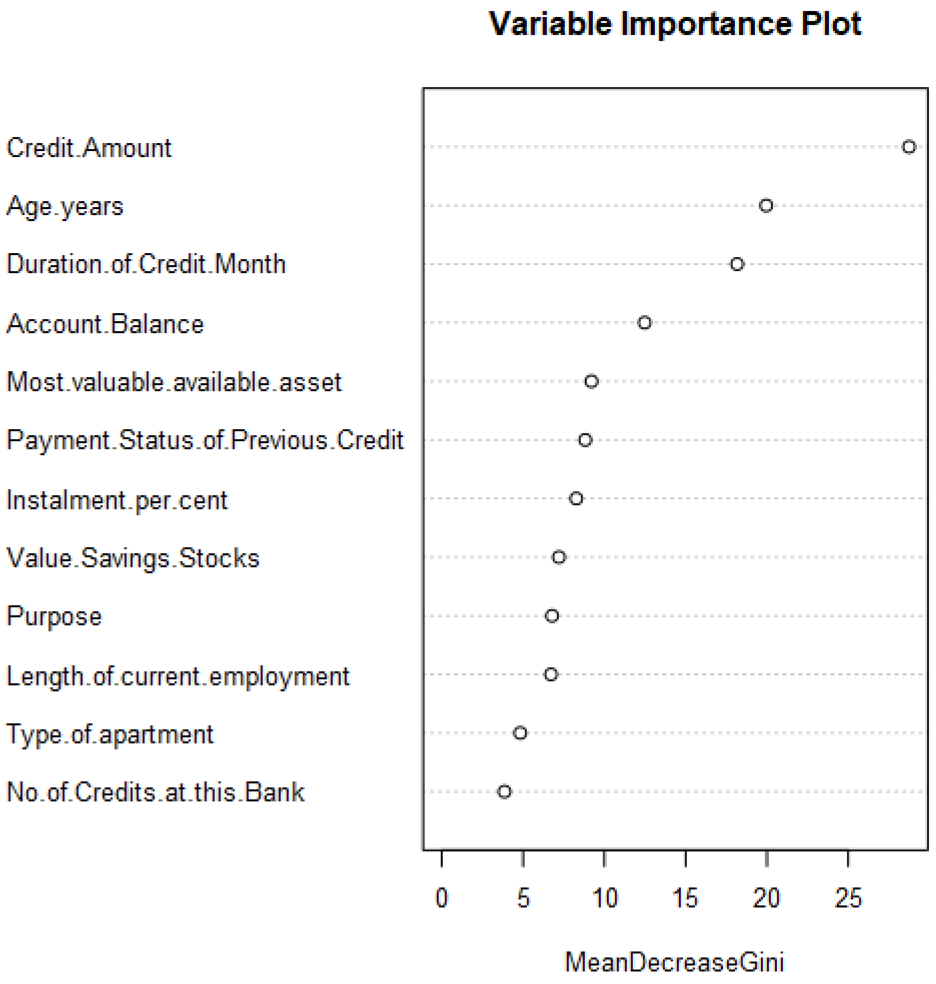
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

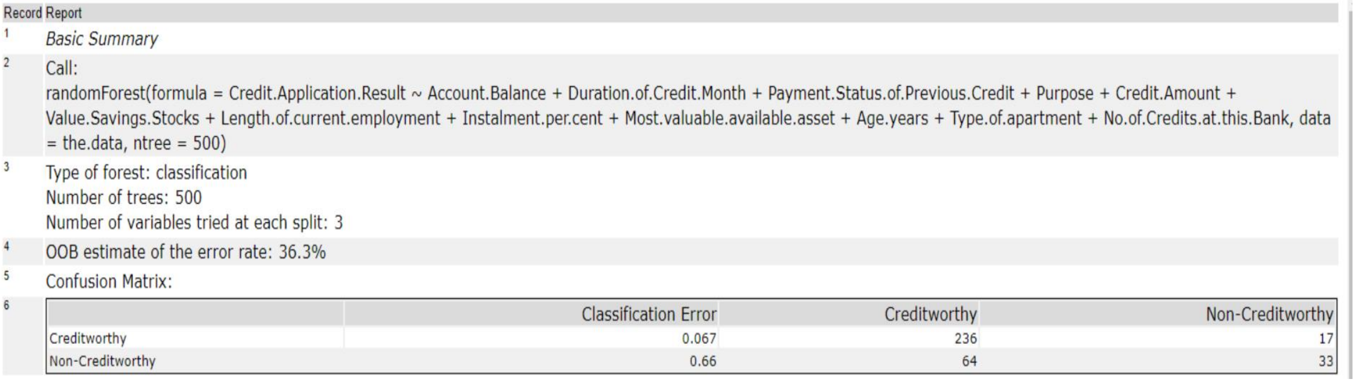
The creditworthy prediction was 79.05%. The model does a decent job in predicting creditworthy applicants. The model under predicts non-creditworthy applicants accurately at 40%.

Forest Model

Significant Predictor Variables

The **most important** predictor variables in the Forest Tree Model are: Credit-Amount, Age-years, Duration-of-Credit-Month, and Account-balance.





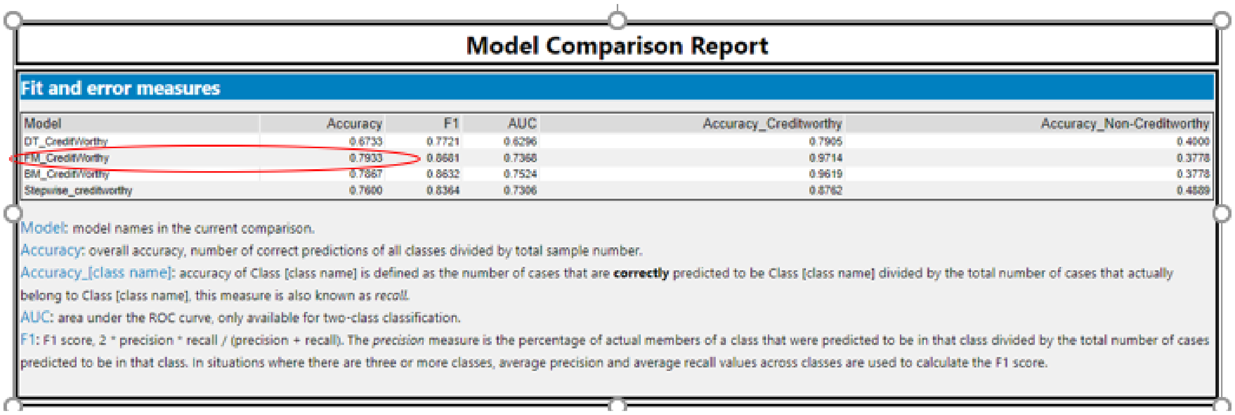
Record	Report
1	*Basic Summary*
2	Call: randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 500)
3	Type of forest: classification Number of trees: 500 Number of variables tried at each split: 3
4	OOB estimate of the error rate: 36.3%
5	Confusion Matrix:
6	

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

The Out of the Bag Error Rate, which explains how well the model performed with the cross-validation set in the estimation data, for the Forest model is 36.3%. Overall, this error rate is high. The model does a better job in predicting creditworthy applicants (classification error 6.7%) than non-creditworthy applicants (classification error 66%)

Model Validation

- After running this model against the validation set, the overall percent accuracy is 79.33%. This model did a better job in predicting the overall percent accuracy than the decision tree model. Forest Models tends to look at the results as a whole to make a prediction. Each individual tree created still has overfitting issues, but when you look at the results as a whole, the overfitting gets averaged out by all of the other trees.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CreditWorthy	0.6733	0.7721	0.6296	0.7905	0.4000
FM_CreditWorthy	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889

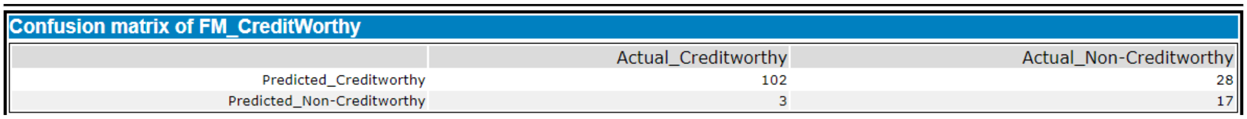
Model: model names in the current comparison.

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.



Confusion matrix of FM_CreditWorthy

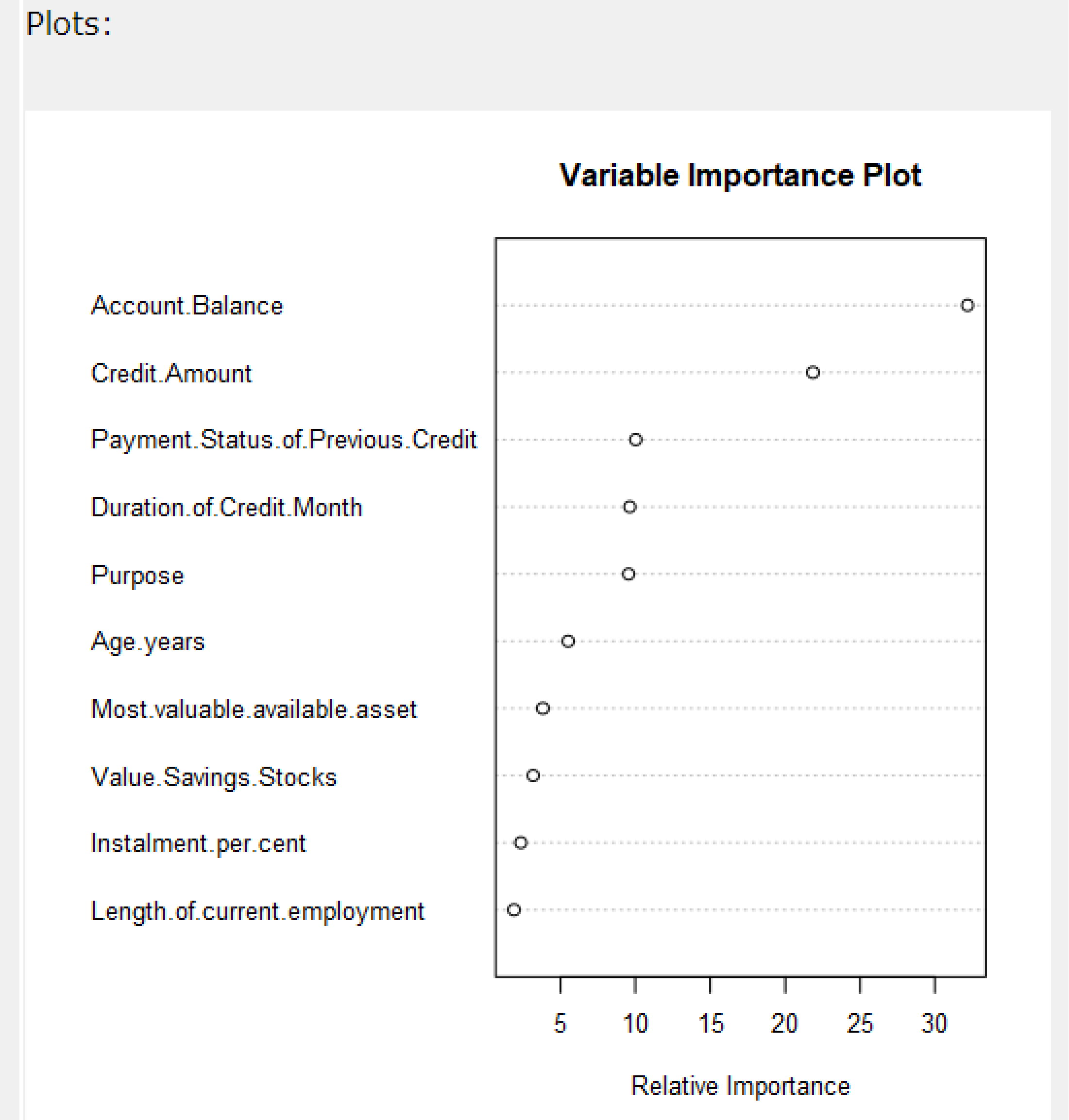
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

The creditworthy prediction was 97.14%, which is very strong. The model under predicts non-creditworthy applicants accurately at 37.78%.

Boosted Model

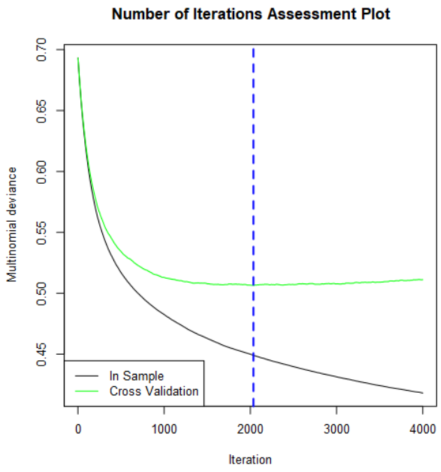
Significant Predictor Variables

The **most important** predictor variables in the Boosted Model are: Account-Balance, Credit-Amount, Payment-Status-of-Previous-Credit, and Duration-of-Credit-Month.



Plots:

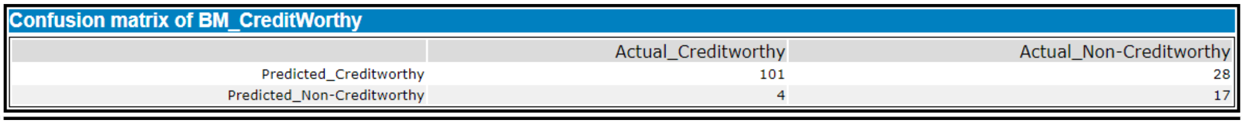
Variable Importance Plot
Account.Balance
Credit.Amount
Payment.Status.of.Previous.Credit
Duration.of.Credit.Month
Purpose
Age.years
Most.valuable.available.asset
Value.Savings.Stocks
Instalment.per.cent
Length.of.current.employment
5
10
15
20
25
30
Relative Importance



The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).

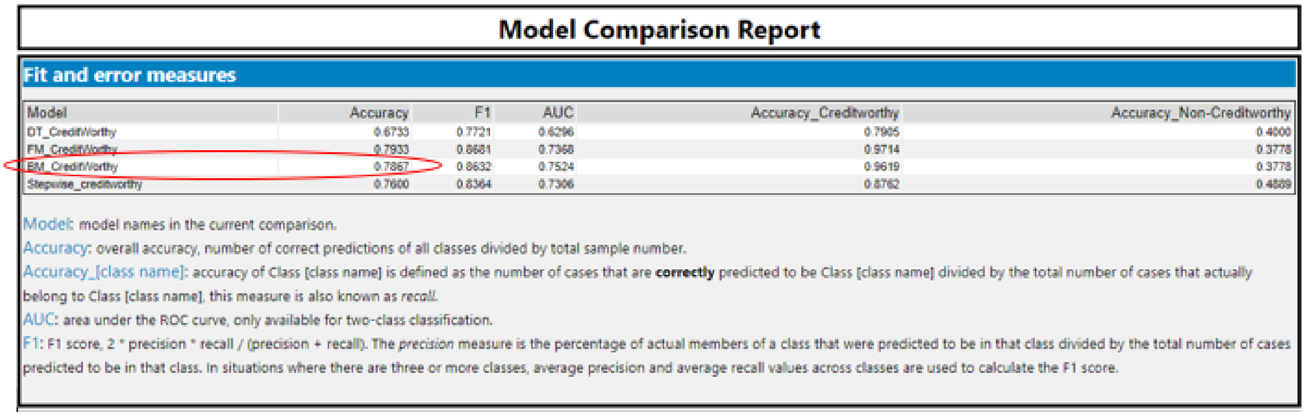
Model Validation

After running this model against the validation set, the overall percent accuracy is 78.67%, slightly less than the Forest Model overall percent accuracy prediction.



Confusion matrix of BM_CreditWorthy

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



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CreditWorthy	0.6733	0.7721	0.6296	0.7905	0.4000
FM_CreditWorthy	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_creditworthy	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.

The creditworthy prediction was 96.19%, which is very strong but slightly less than the Forest Model prediction. This model also under predicts the non-creditworthy applicants accurately at 37.78%.

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

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"

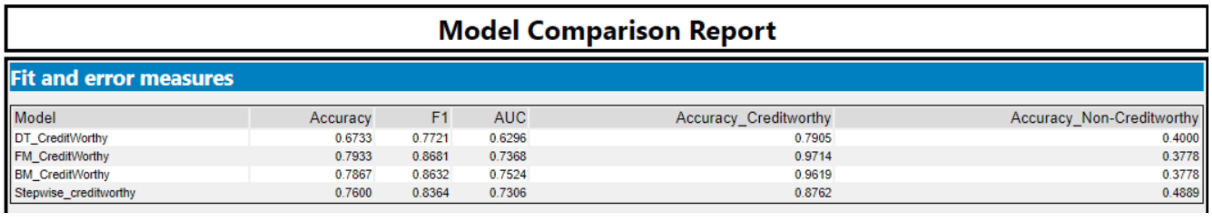
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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

    I chose to use the Forest Model. After looking at model accuracy, overall the forest model performed the best. I chose this model for the following reasons:

    - ○ It had the highest overall accuracy (79.33%) than the other models against the validation set
    - ○ It had the highest creditworthy accuracy rate (97.14%). The accuracy rate for non-creditworthy (37.78%) applicants was not strong but similar to accuracy rate in other models (37.78%, 40% and 48.89%, respectively).

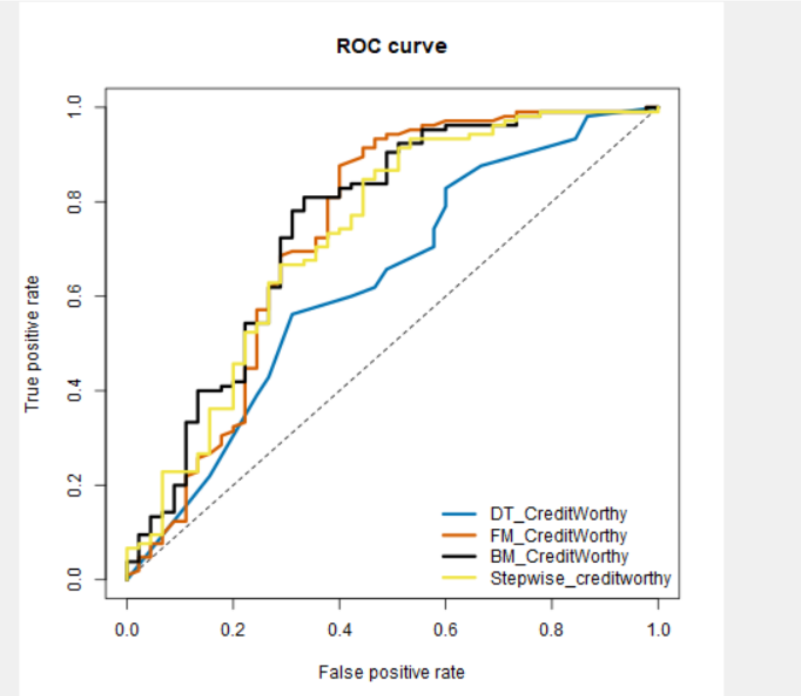


Model Comparison Report

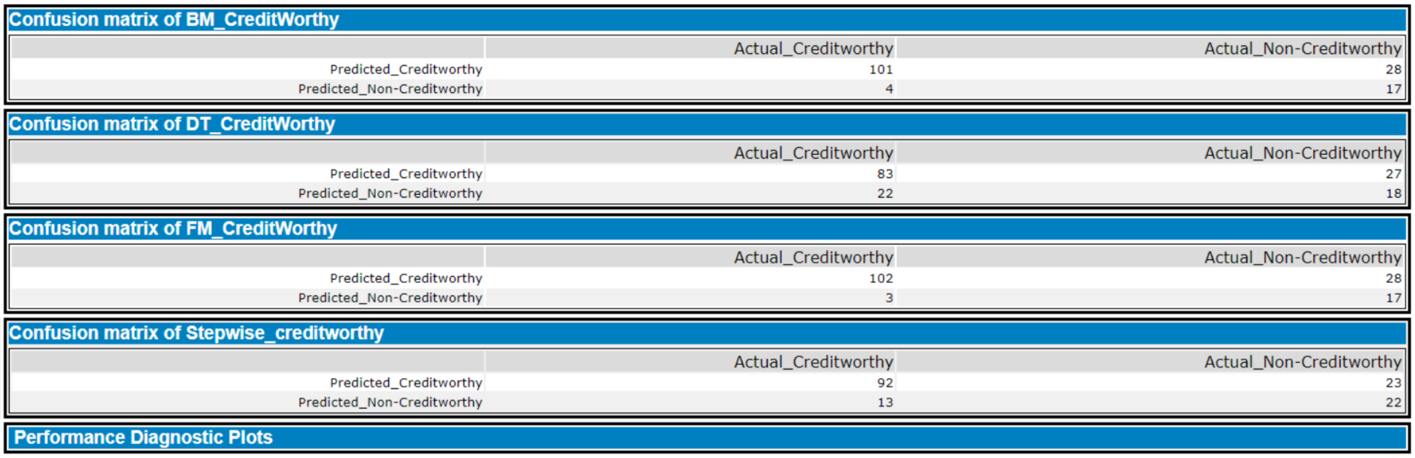
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CreditWorthy	0.6733	0.7721	0.6296	0.7905	0.4000
FM_CreditWorthy	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CreditWorthy	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889

    - ○ **ROC graph** - The forest model (FM_Creditworthy), has a ROC Curve that hugs the upper left corner of the plot the most. The classifier does a good job separating the classes and better at random guessing than the other models.



- ○ There is some bias shown in the Confusion Matrices for all models because they were some applicants that were predicted creditworthy and were actually non-creditworthy and some applicants who were predicted non-creditworthy and were actually creditworthy. The forest model had the best overall accuracy rate at 79.33%.



Confusion matrix of BM_CreditWorthy

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

Confusion matrix of DT_CreditWorthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of FM_CreditWorthy

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

Confusion matrix of Stepwise_creditworthy

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

Performance Diagnostic Plots

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

- How many individuals are creditworthy? 408 applicants are creditworthy.

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.