Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- **What decisions needs to be made?**

  The decision that needs to be made is to classify the 500 loan applicants as creditworthy or non-creditworthy based on the data.

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

  We can use data such as length of employment, income, account balance, age, credit amount, duration of credit month, purpose of loan, assets amounts, number of credits, etc. to help determine if a loan application should be classified as creditworthy or not creditworthy,

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

  We will use a binary classification model to determine if the customer is creditworthy or not creditworthy.

Step 2: Building the Training Set

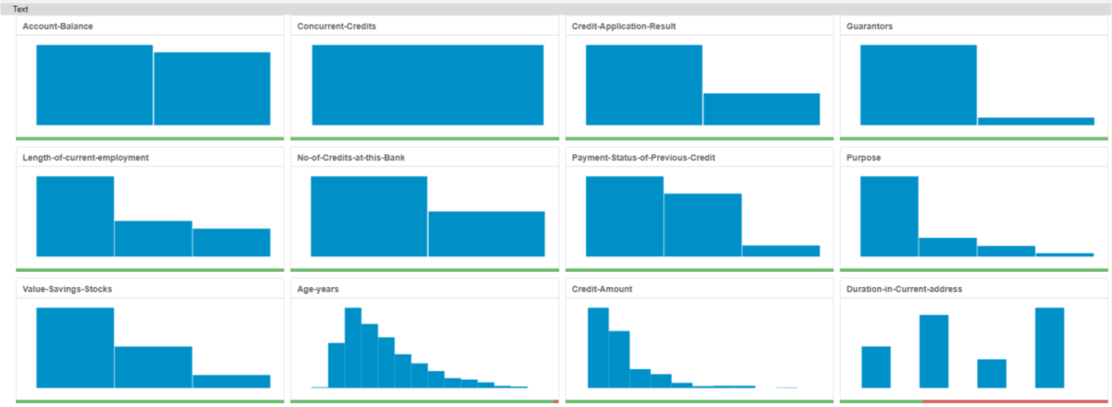
Answer this question:

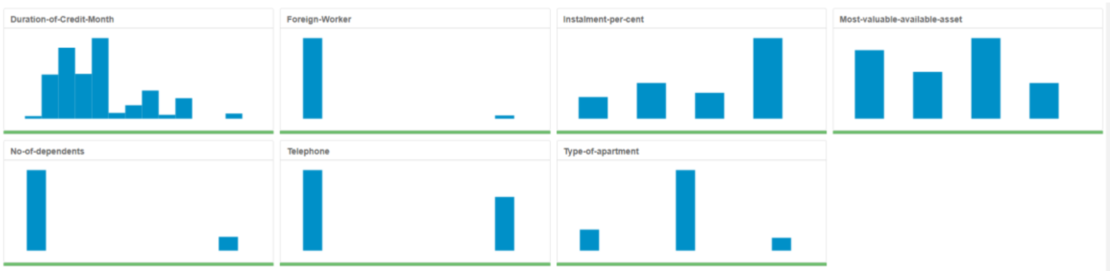
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  During the data cleanup process, I removed the following fields:

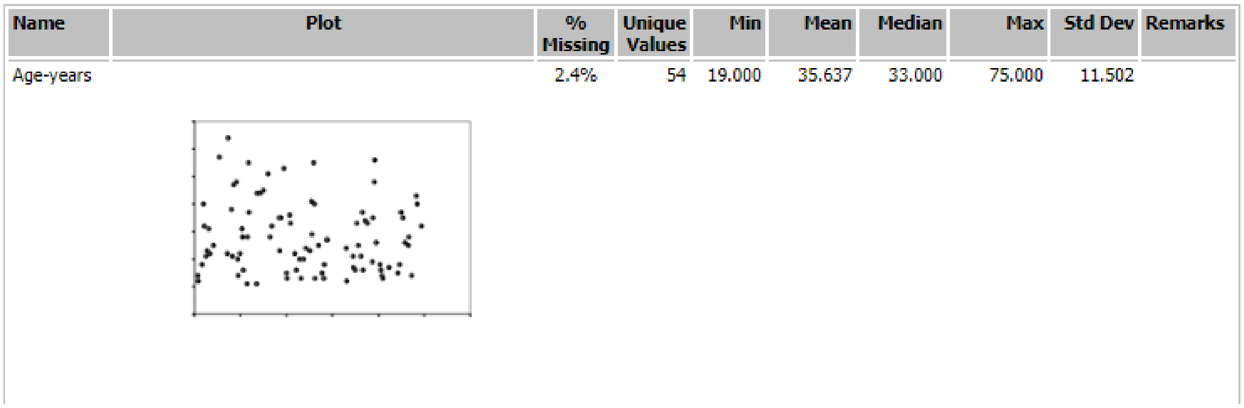
  1) Gaurantors- due to low variability
  2) Duration in current address- due to missing a majority of the data (69%)
  3) Concurrent credit- due to low variability (only 1 unique value)
  4) Occupation- due to low variability (only 1 unique value)
  5) Number of dependents- due to low variability

6) Telephone data- because it is not relevant to the model
7) Foreign worker- due to low variability





Additionally I found null values in the age year field. Because it was only missing 2% of the data, I decided to impute the missing data with the media age of 33. I chose to use the median because the data is shifted towards the left (see below).



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	

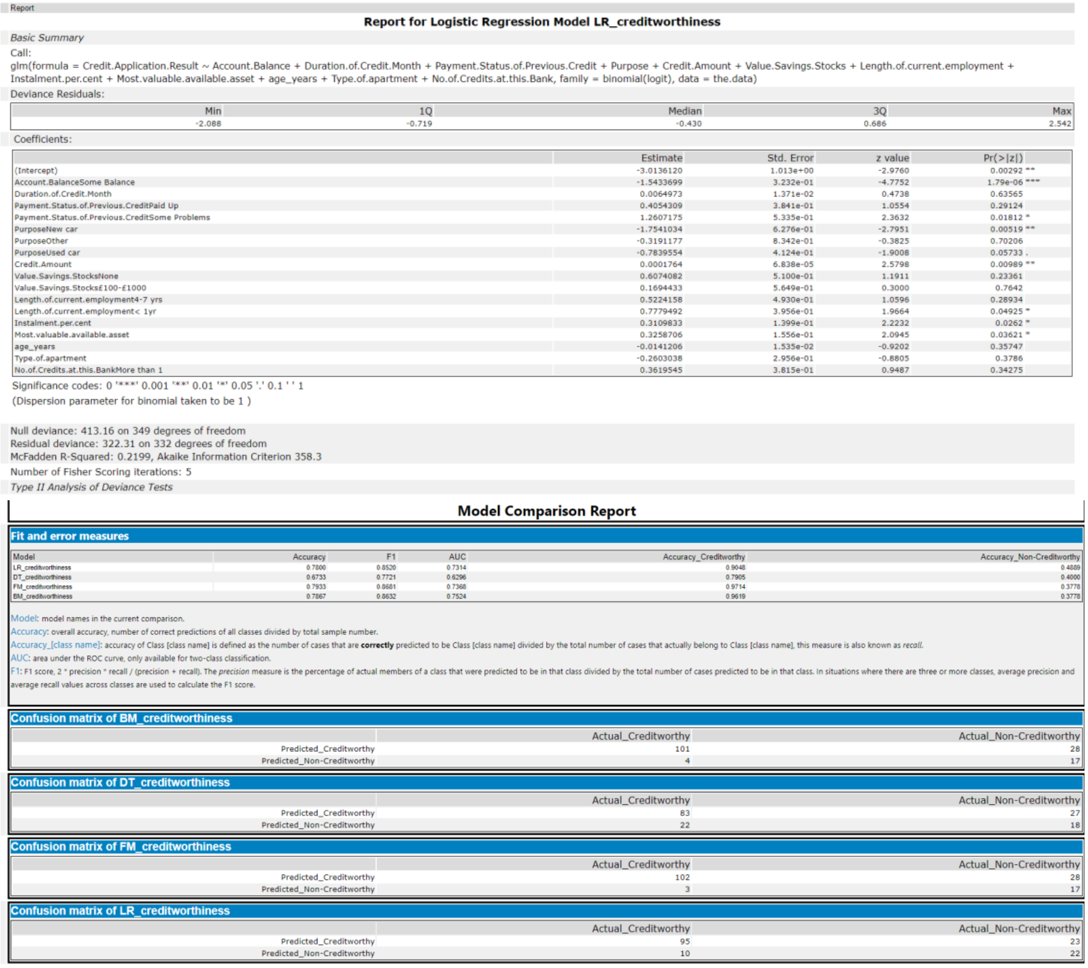
Step 3: Train your Classification Models

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?

Logistic Regression:

The most significant predictor variables are account.balancesome balance, purposenewcar, and credit.amount. Using the model comparison report we can see that the overall accuracy for this model is 78%. Based on the confusion matrix, this model seems to be slightly biased towards classifying as creditworthy.



Report

Report for Logistic Regression Model LR_creditworthiness

Basic Summary

Call:
glm(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, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
age_years	-0.0141206	1.535e-02	-0.9202	0.35747
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275

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: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3
Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthiness	0.6733	0.7721	0.6296	0.7905	0.4000
FM_creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778

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 BM_creditworthiness

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

Confusion matrix of DT_creditworthiness

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

Confusion matrix of FM_creditworthiness

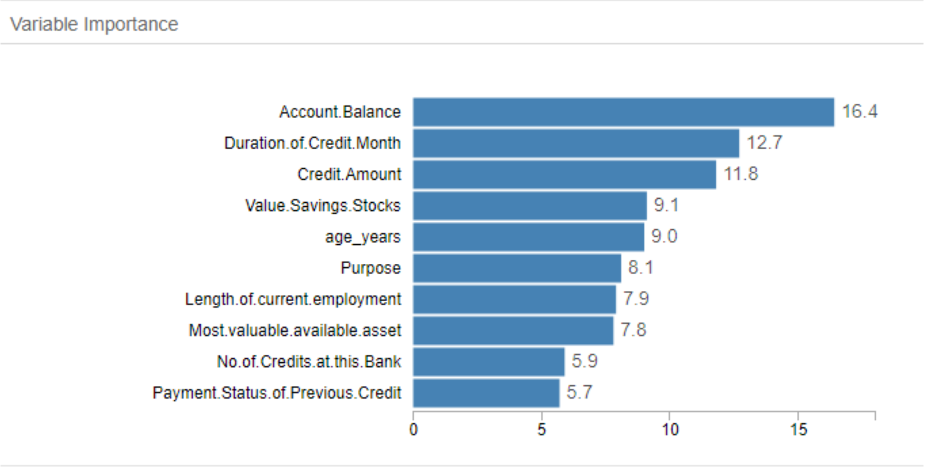
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

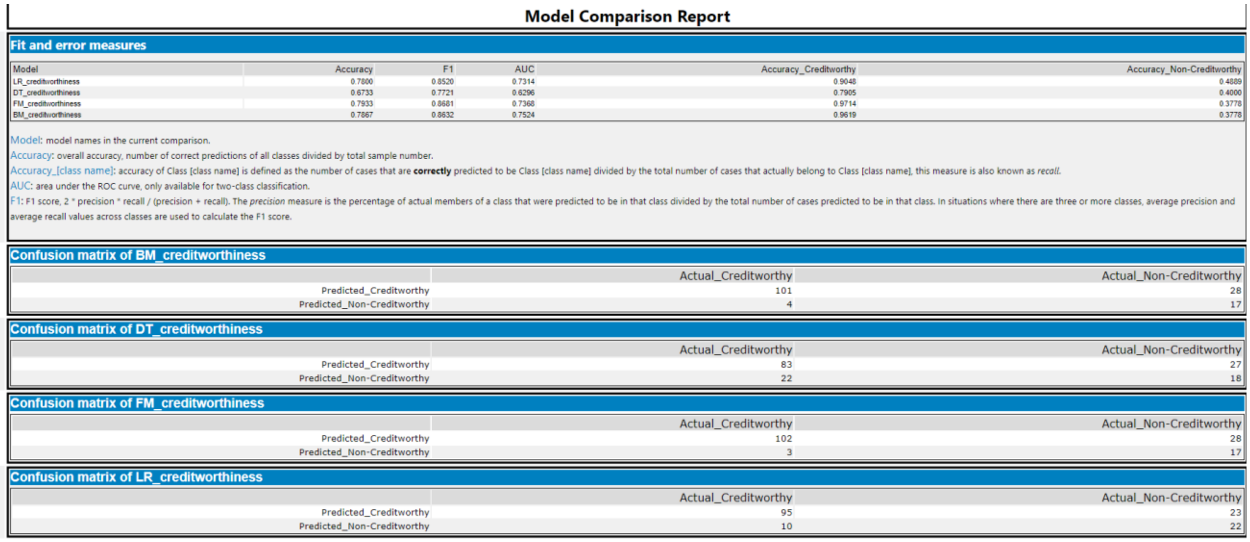
Confusion matrix of LR_creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Decision Tree:

The most significant predictor variables are credit.amount, duration.of.credit.month and account.balance. Using the model comparison report we can see that the overall accuracy for this model is 67.3% which is quite low. The creditworthy accuracy is the lowest compared to the other models at 79.1%.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthiness	0.6733	0.7721	0.6296	0.7905	0.4000
FM_creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778

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 BM_creditworthiness

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

Confusion matrix of DT_creditworthiness

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

Confusion matrix of FM_creditworthiness

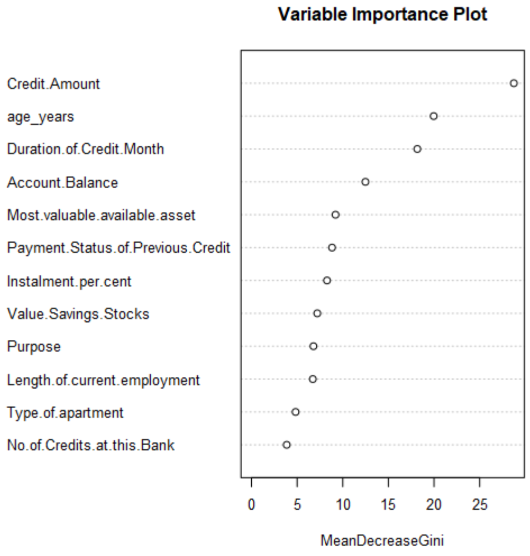
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

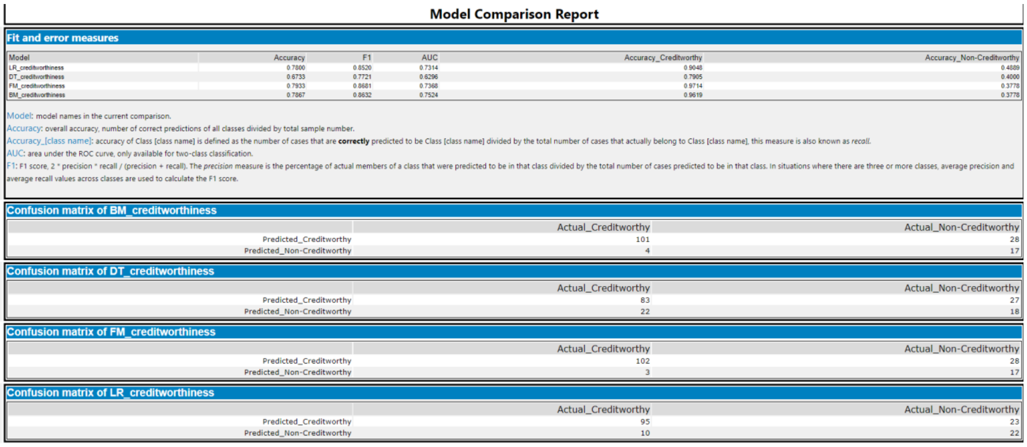
Confusion matrix of LR_creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Forest Model:

The most significant predictor variables are credit.amount, age_years, and duration.of.credit.month . Using the model comparison report we can see that the overall accuracy for this model is 79.3% which is the highest overall accuracy when compared to the other models. This model also scored the highest creditworthy accuracy at 97.1%. Based on the confusion matrix, this model seems to be slightly biased towards classifying as creditworthy.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthiness	0.6733	0.7721	0.6296	0.7905	0.4000
FM_creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778

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 BM_creditworthiness

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

Confusion matrix of DT_creditworthiness

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

Confusion matrix of FM_creditworthiness

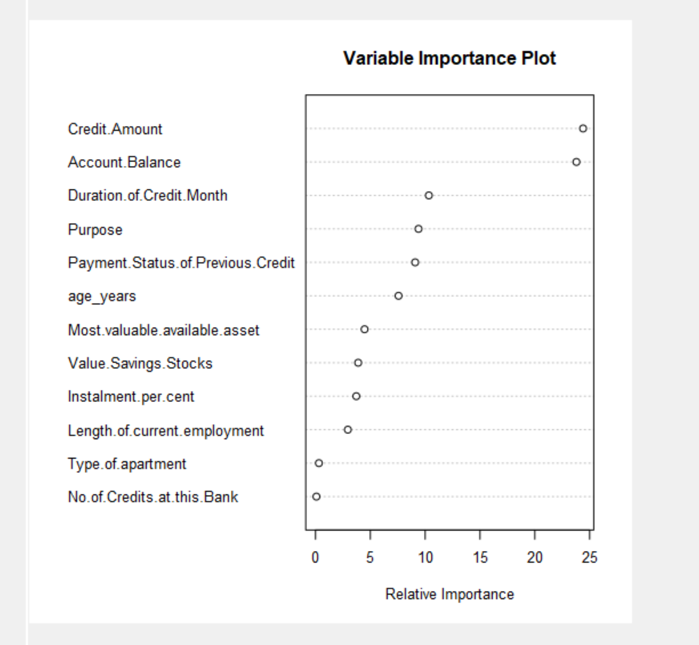
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

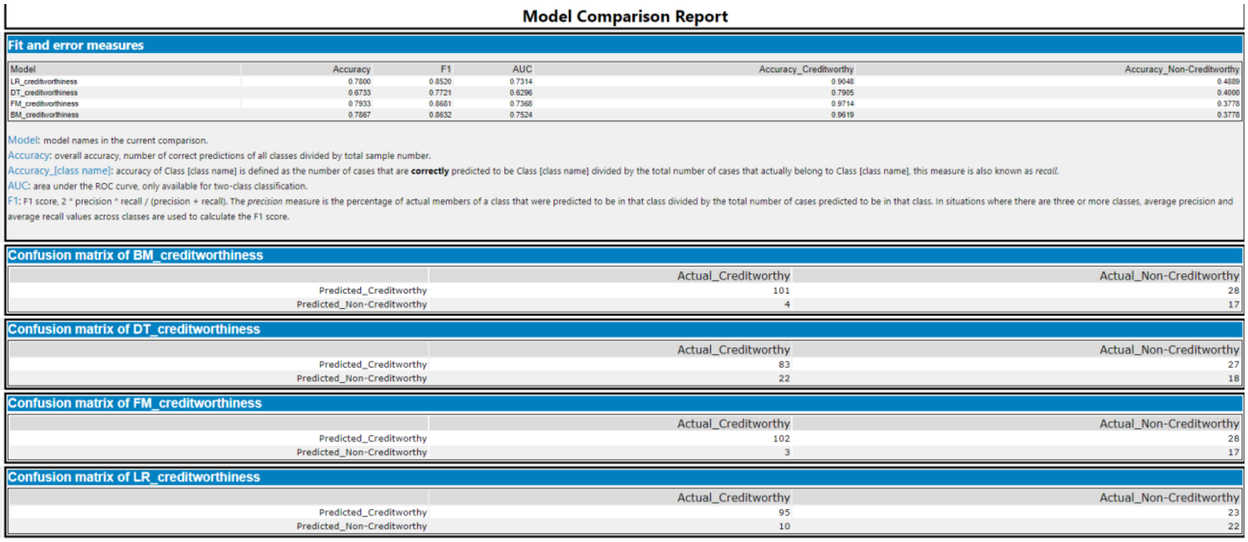
Confusion matrix of LR_creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Boosted Model:

The most significant predictor variables are credit.amount and account.balance . Using the model comparison report we can see that the overall accuracy for this model is 78.7%% which is the second highest overall accuracy when compared to the other models. This model also scored the second highest creditworthy accuracy at 96.2%. Based on the confusion matrix, this model also seems to be biased towards classifying as creditworthy.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthiness	0.6733	0.7721	0.6296	0.7905	0.4000
FM_creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778

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 BM_creditworthiness

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

Confusion matrix of DT_creditworthiness

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

Confusion matrix of FM_creditworthiness

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

Confusion matrix of LR_creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Step 4: Writeup

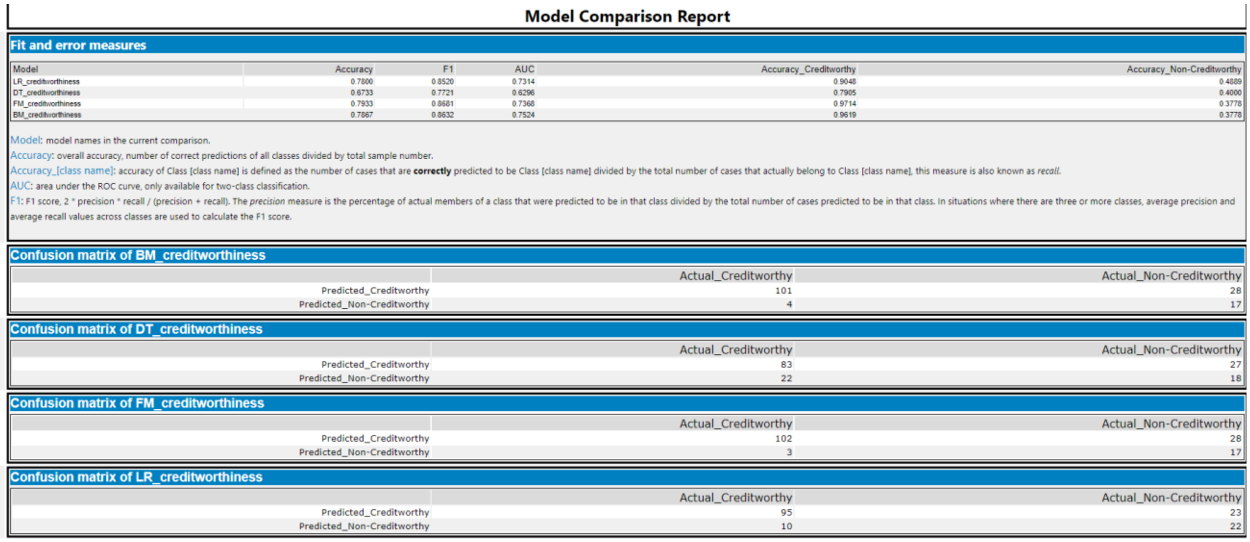
Answer these questions:

- 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:

  The model I chose to use is the forest model.

  1) **Overall Accuracy against your Validation set**: The forest model scored the highest overall accuracy at 79.3%

2) **Accuracies within “Creditworthy” and “Non-Creditworthy” segments**: The forest model had the highest creditworthy accuracy at 97.1%.
3) **ROC graph**: Based on the graph the forest model produces the best results because it is the highest and also reached the true positive rate the quickest.
4) **Bias in the Confusion Matrices**: There seems to be a slight bias towards classifying as creditworthy. However, this is likely due to the fact that our training dataset had a significantly smaller sample size of non-creditworthy customers than creditworthy customers.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_creditworthiness	0.7800	0.8520	0.7314	0.9048	0.4889
DT_creditworthiness	0.6733	0.7721	0.6296	0.7905	0.4000
FM_creditworthiness	0.7933	0.8681	0.7368	0.9714	0.3778
BM_creditworthiness	0.7867	0.8632	0.7524	0.9619	0.3778

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 BM_creditworthiness

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

Confusion matrix of DT_creditworthiness

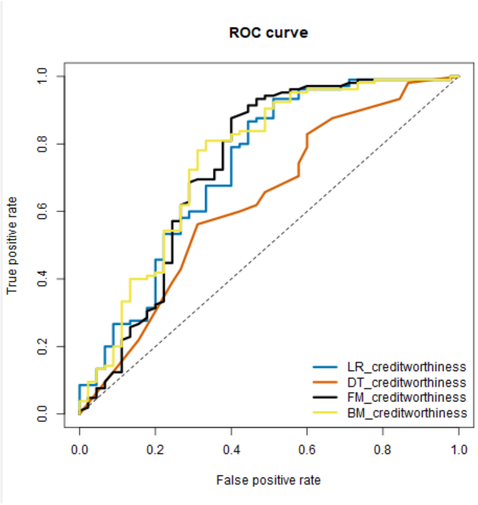
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of FM_creditworthiness

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

Confusion matrix of LR_creditworthiness

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22



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

- How many individuals are creditworthy?

Using the forest model, 408 applications classified as creditworthy.