Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here:

Step 1: Business and Data Understanding

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

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand. Due to a financial scandal that hit a competitive bank last week, All of a sudden you have nearly 500 loan applications to process this week!

What data is needed to inform those decisions?

- I need to get all the data of past applicants

- The list of customers that need to be processed in the next few days

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

I need to use a Binary model because out decision is going to be either "Qualify" or "Not Qualified"

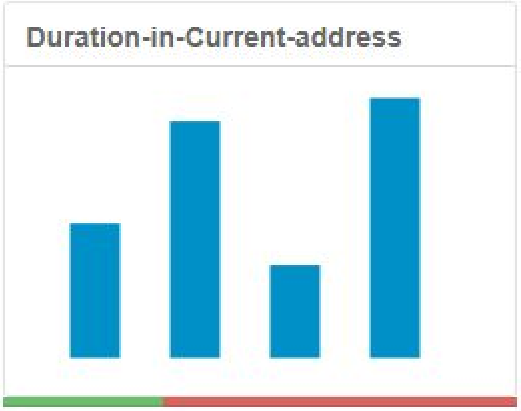
Step 2: Building the Training Set

Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

I have found some data that need to be imputed and removed from the training data set. I believe some of the data will not make an impact on the final model.

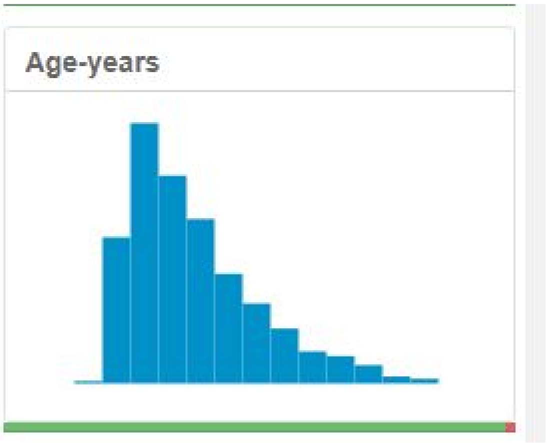
- **Duration-in-current-address**

This data has a 69% missing data which will not have a positive impact on the training data, therefore I will remove it by deselecting it.



- **Age-years**

Age-years have a 2% missing data which still makes it a good variable but it needs to be 100% full. Therefore I decided to impute the missing data the median of Age-years 33, this way it is closer to its possible missing data.



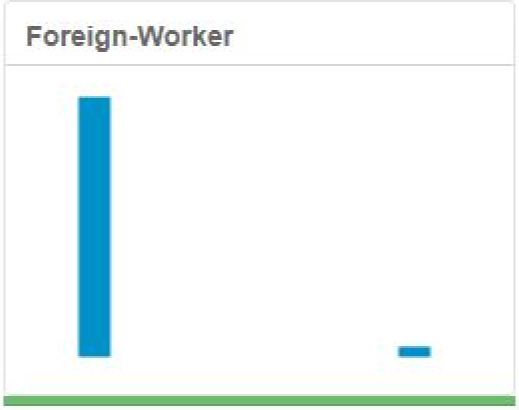
Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability”

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.

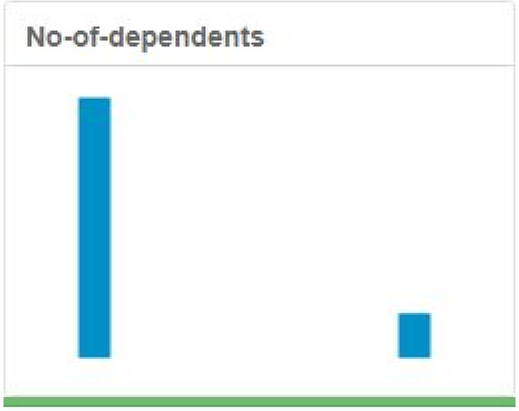
There are low variability in some of the column which are:

- **Foreign-Worker: has a low variability with right skewed of 481 instances (1.0-1.1) and 19 instances (2.0-2.1) Therefore, I will remove it**

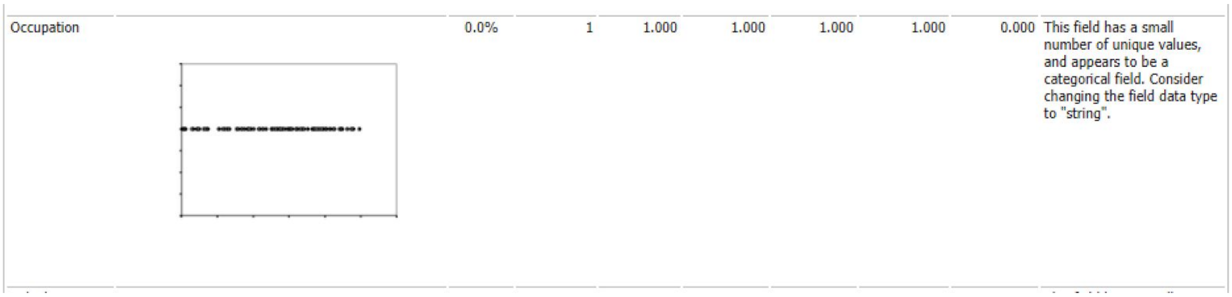


- **No-of-dependents**

This one has right skewed of 427 instances (1.0-1.1) compared to 73 instances (2.0-2.1). Therefore it will not be a good fit to the training data and I will remove it



- **Occupation:**
  I will remove occupation since it has a one value of all records and it does not mean anything to the data nor has positive correlation.



Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

- **Concurrent-Credits**

Type of category “Otherbanks/debt” with 500 instances which will not be a “high variability”. **Therefore, I will remove it**

- **Guarantors**

Majority of the answers are 457 instances “No” compared to the “Yes” with 43 instances. Therefore, I will remove it.

- **Telephone**

I will remove Telephone because it is not a valid predictor variable.

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.

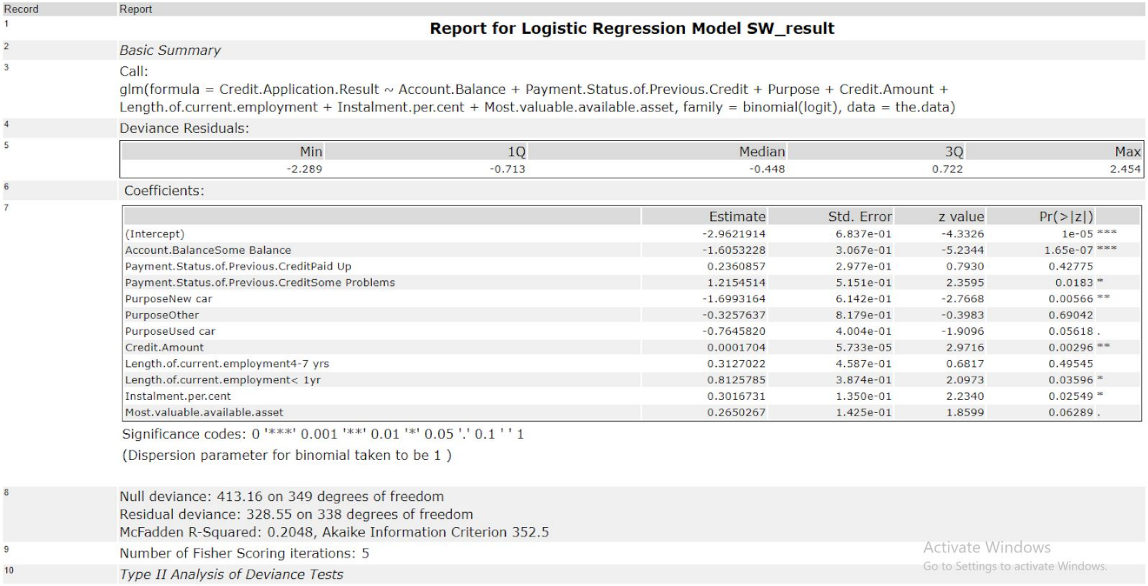
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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.

- ### Logistic Regression

As it shows in the report of the LOG_result,
Predictive variable: Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset



Record Report

1 **Report for Logistic Regression Model SW_result**

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)

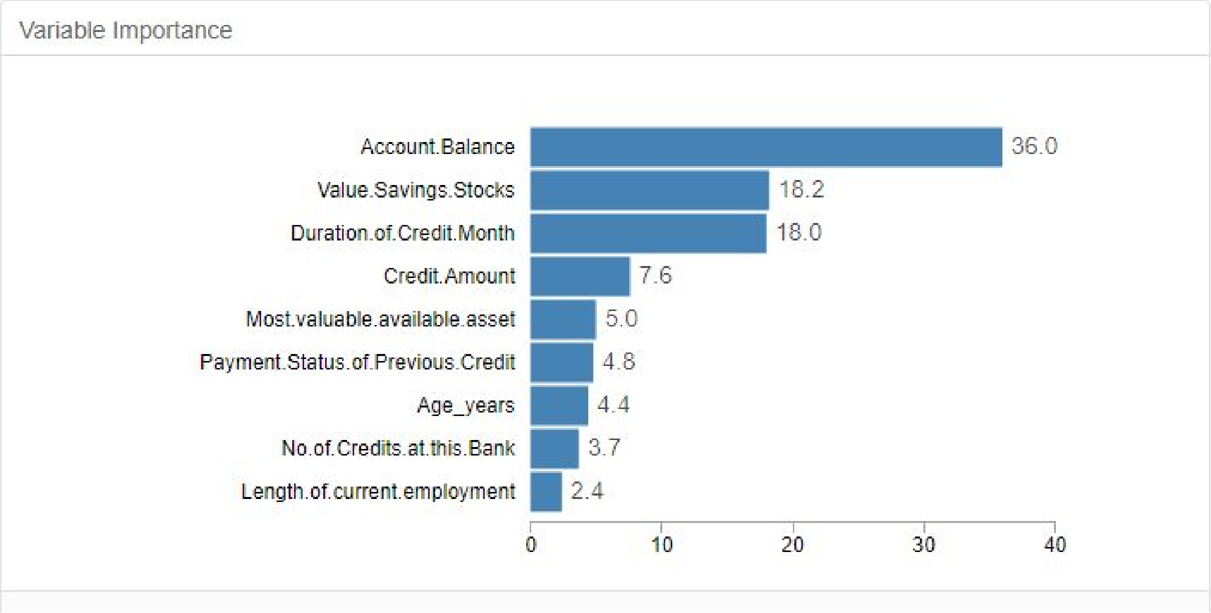
8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

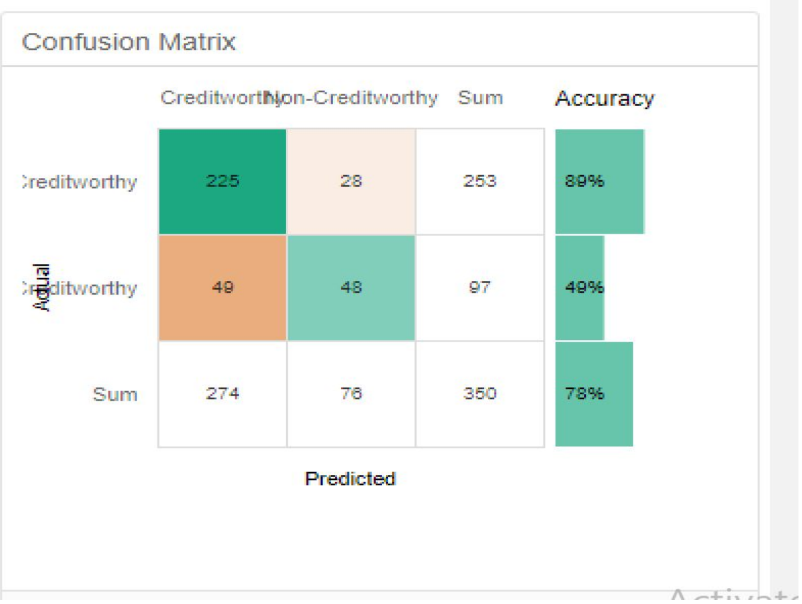
- ### Decision Tree

The variable importance shows the account.balance as the top variables. Also it seems that value.savings.stocks and duration.of.credit.month has a close connection graph.



Also,

In DT report, the confusion matrix shows Creditworthy with an accuracy of 89%. And not creditworthy 49%

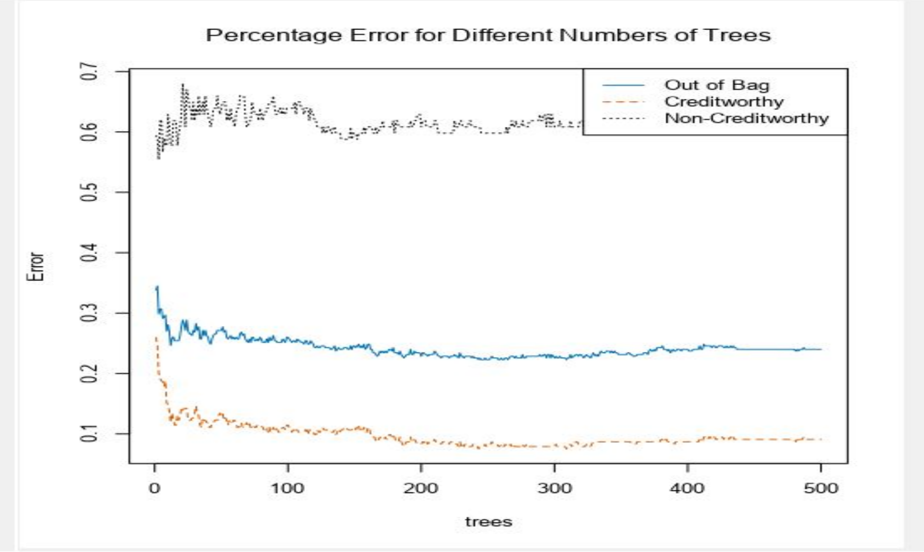


Confusion Matrix

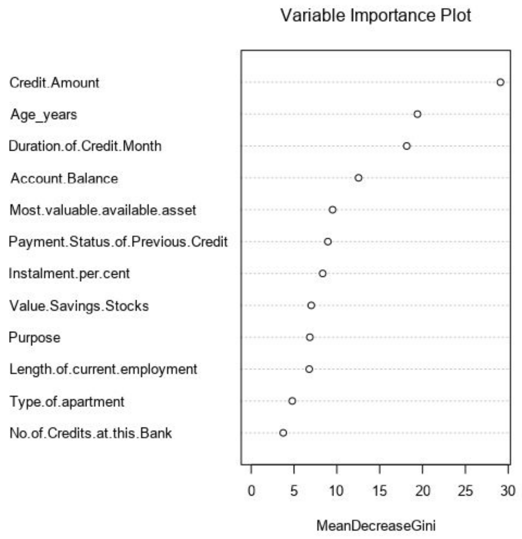
Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
Creditworthy	225	28	253	89%
Creditworthy	49	48	97	49%
Sum	274	76	350	78%

Predicted

- **Forest Model:**
The FM_result report shows many insightful visualizations as that would be a very positive aspect of the model for the predictor variables.
In the Repost, OOB estimate of the error rate: 24% which has some errors that impacted the model for the different Trees.
Also, the Out of Bag seems flat as this shows how the different error of the different number of trees

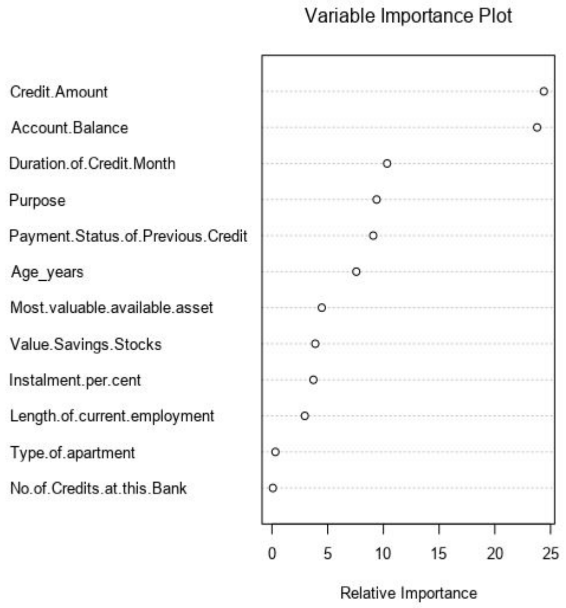


Moreover,
The variable importance shows the mean decrease of some variables
And the top 3 predictor variables are credit.amount, age.years, duration.of.credit.month



Boosted Model:

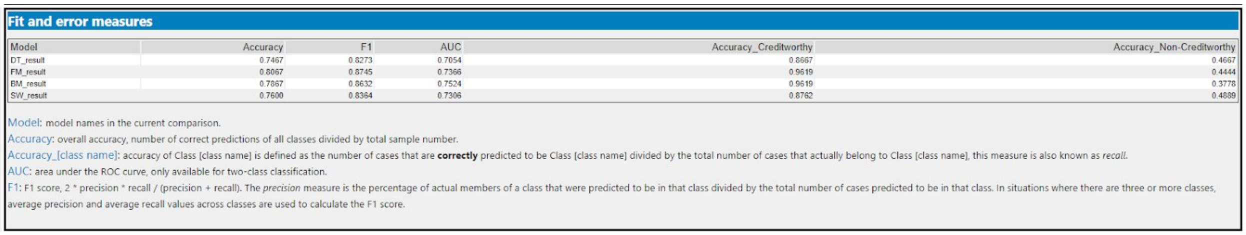
The variable importance plot shows how different variables plotted relative to each predictor field. And the top 2 important predictive variables are credit.amount and account balance.



Validate your model against the Validation set. What was the overall percent accuracy?

After joining the 4 models, I have added a model comparison and linked all the output to the model comparison. Also I linked the sample validation to model comparison to test the accuracy

The report shows that the highest accuracy percent is Forest Model with 80% accuracy and with 96% accuracy of Creditworth. Also the second is Boost Model with 79% accuracy with the same the same accuracy of Creditworth 96%



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_result	0.7467	0.8273	0.7054	0.8667	0.4667
FM_result	0.8067	0.8745	0.7366	0.9619	0.4444
BM_result	0.7867	0.8632	0.7524	0.9619	0.3778
SW_result	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.

Show the confusion matrix. Are there any bias seen in the model's predictions?

Yes and more explanation below

Step 4: Writeup

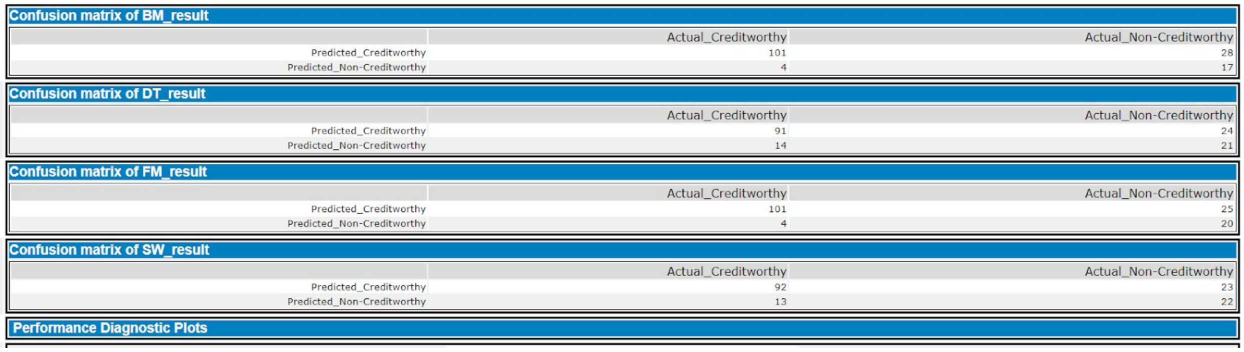
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

This is the confusion matrix of the four models stating the predicted_creditworth and predicted_non_creditworth. And I think there is bias in the creditworth side as it has more prediction than non_creditworthy



Confusion matrix of BM_result

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

Confusion matrix of DT_result

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of FM_result

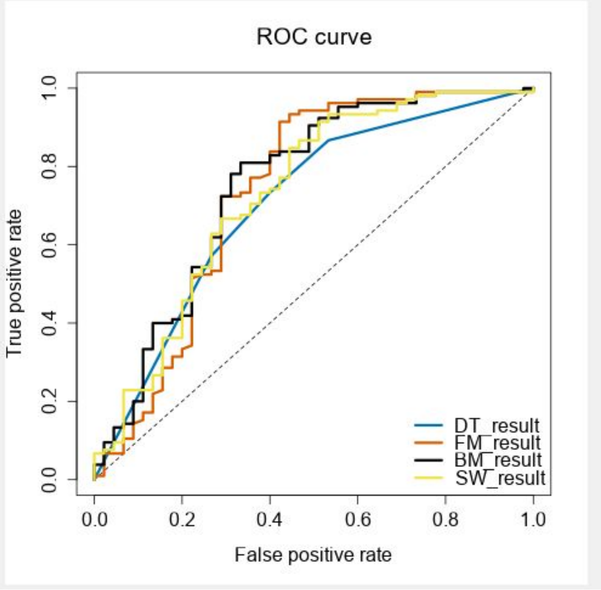
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

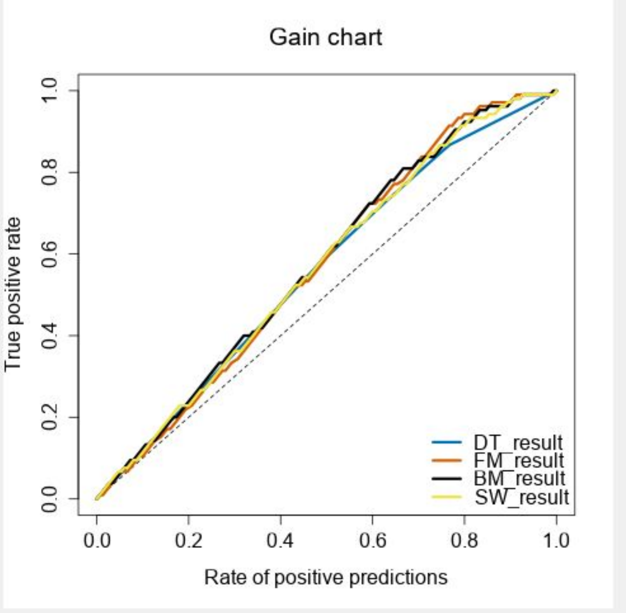
Confusion matrix of SW_result

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

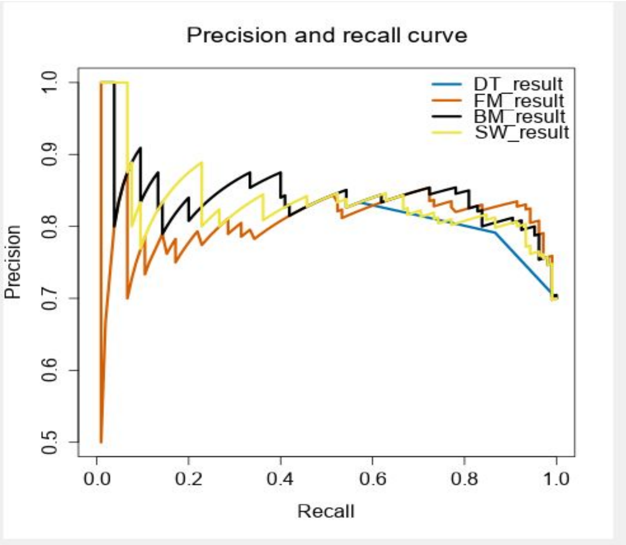
Performance Diagnostic Plots

The ROC shows a significant demonstration of how FM_result is better than other models by reaching to the top. Although most models are relatively close to each other, FM_result is representing a strong model at the top where its becoming the best performance of binary classifier.





In conclusion, I decided to go with Forest Model since it has the highest accuracy rate and more predicted creditworth. and I will build a model to come up with the number of "creditworthy" applicants.

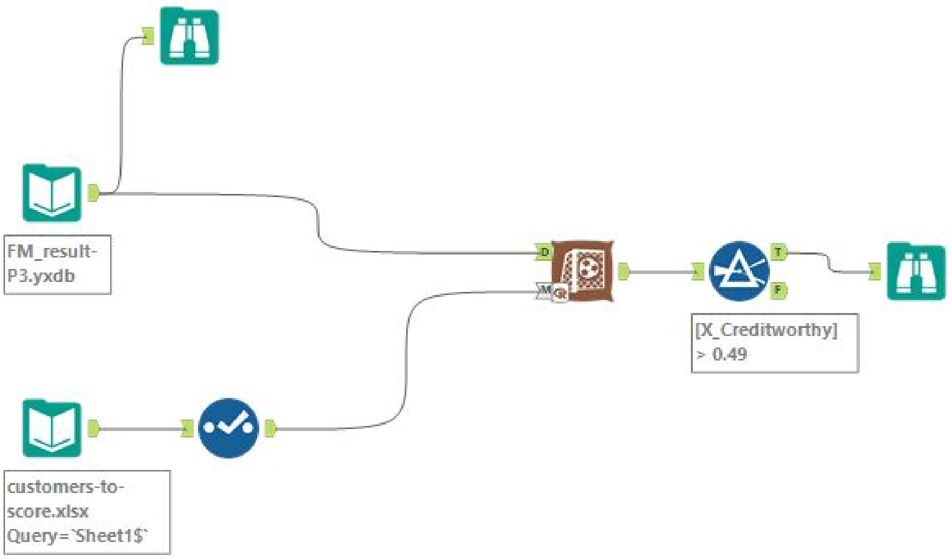


How many individuals are creditworthy?

After choosing the Forest Model to be the right fit model to predict applicants, I have used the score model with customer-to-score with FM to get how much individuals are creditworthy.

This is how I did it:

The number of individual applicants are 407. I filtered the score model to only show any applicant with approved score > 0.49 and above.


FM_result-
P3.yxdb
D
M
[X_Creditworthy]
> 0.49
T
F
customers-to-
score.xlsx
Query=`Sheet1$`