Project 3: Creditworthiness

Step 1: Business and Data Understanding

1. What decisions needs to be made?

To select the best model in order for our client to specify which requesters deserve a loan and which not.

2. What data is needed to inform those decisions?

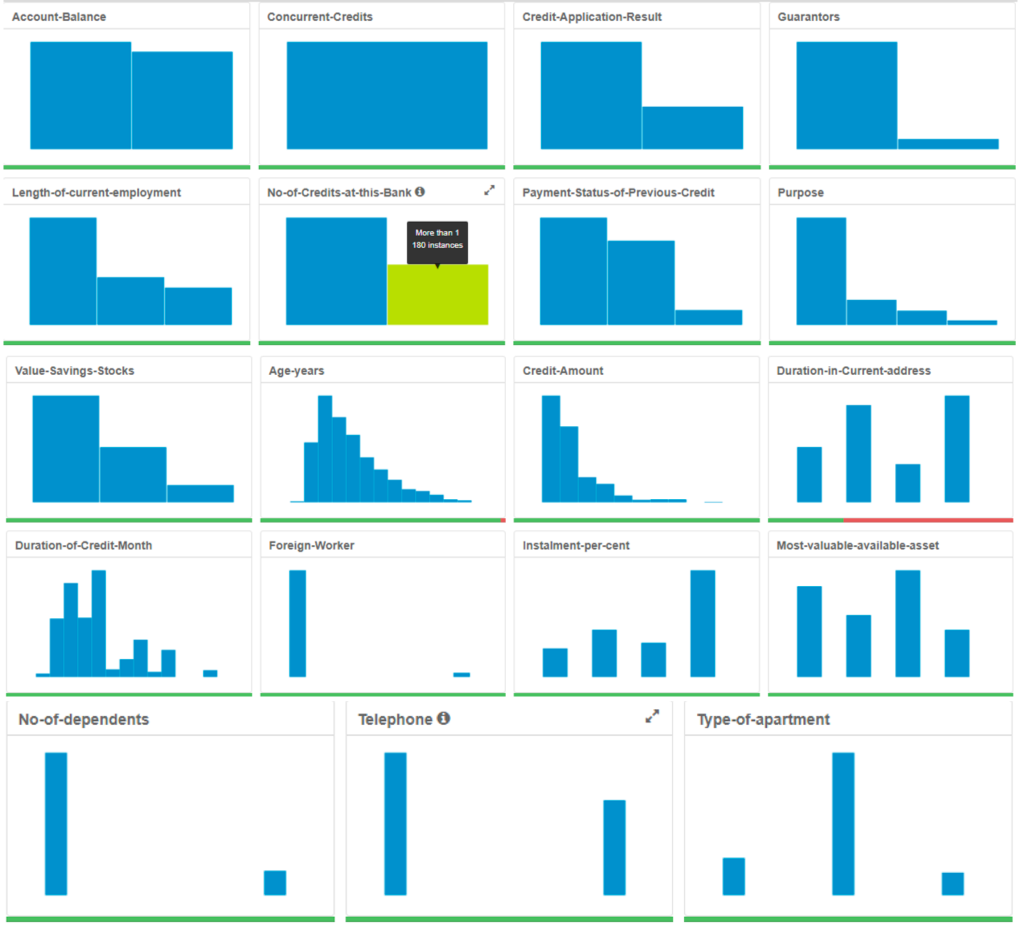
A. Past applications data.
B. Clients proposed for a loan and need to have the decision in the next few days.
C. Predictors need to be used: account balance, duration of credit month, payment status of previous credit, purpose, credit amount, value of stock savings, length of current employment, installment per cent, most valuable available asset, age, type of apartment and number of credits at the bank.

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

Since the decision need to be made is binary (Creditworthy, noncreditworthy), the model need to be used is classification binary model.

Step 2: Building the Training Set

1. **In cleanup process, which fields needed to be removed or imputed?**

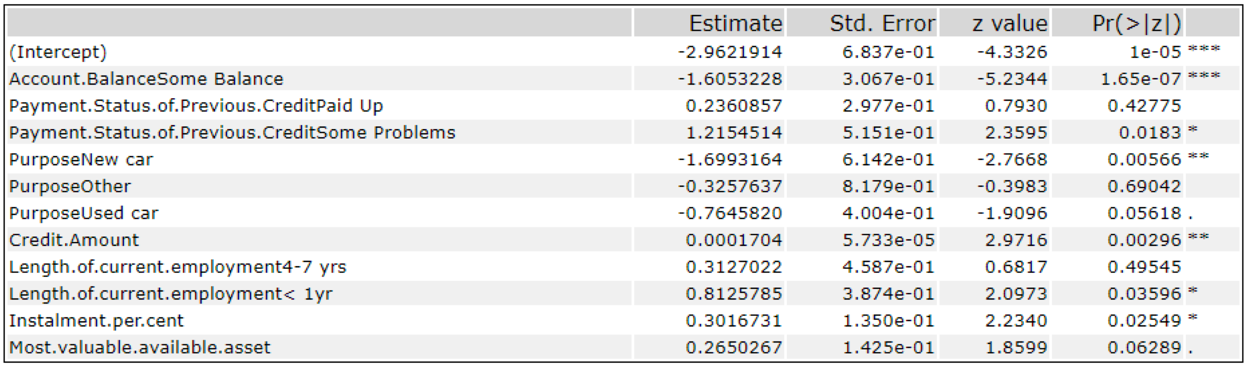


Due to low variability in these fields: foreign worker, guarantors, concurrent credits, no. of dependents, telephone and occupation, they could be removed from our data. Duration in current address should be removed as well due to low variability and the huge number of missing data. 2% of Age data is missing but can be compensated easily with the median (33). After cleaning the data with the steps mentioned above, we have a clean and ready dataset.

Step 3: Classification Models Training

1. Which predictor variables are significant or the most important?

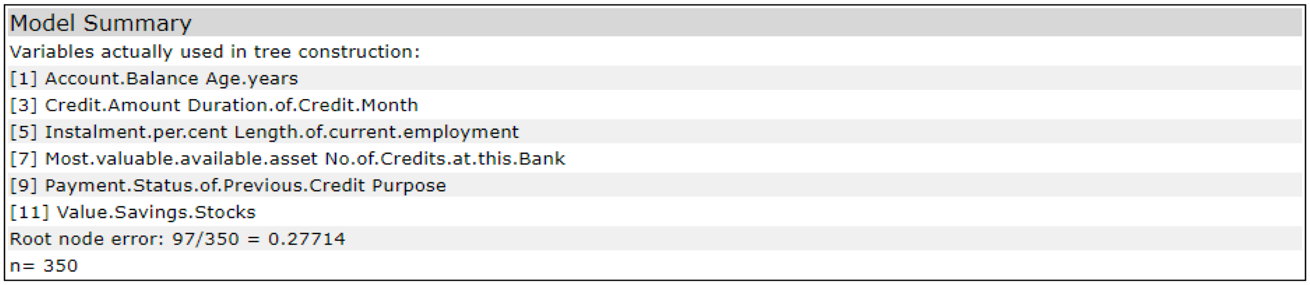
- Logistic Regression:



	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	.

As shown above, six predictors were labeled significant without the intercept. Account balance as the most significant predictor.

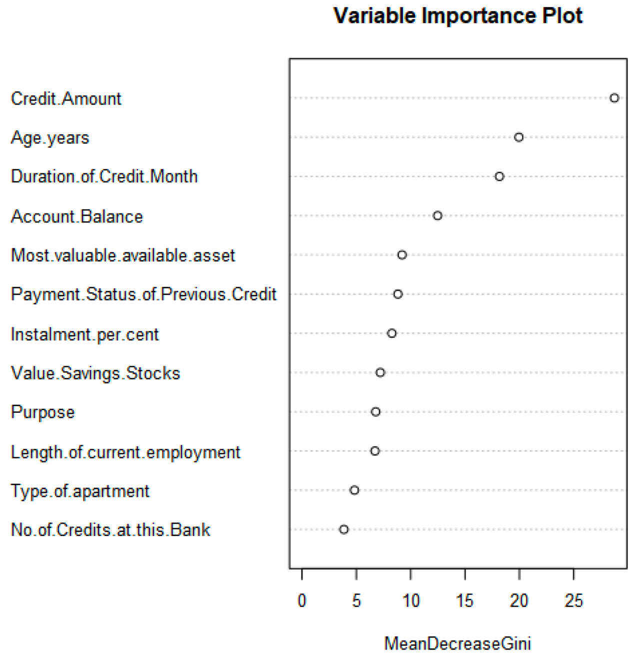
- Decision Tree:



```
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age.years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

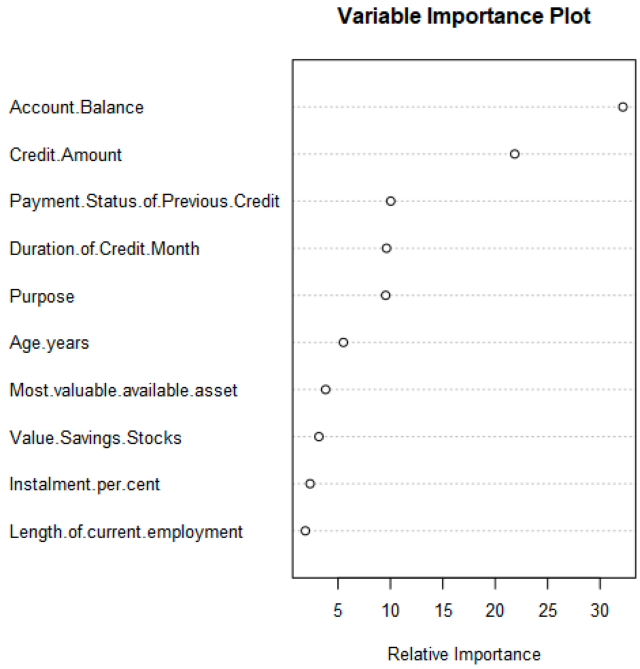
As shown above, eleven predictors were labeled significant. Account balance and age as the most significant predictors.

- Forest Model:



The plot above shows the order of significance in forest model.

- Boosted Model:

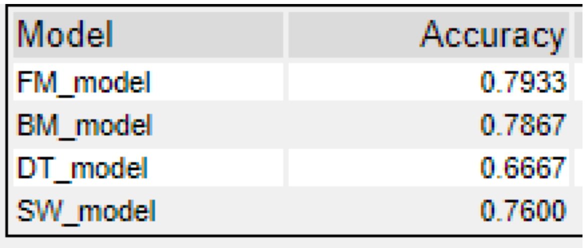


The plot above shows the order of significance in boosted model.

2. What was the overall percent accuracy?

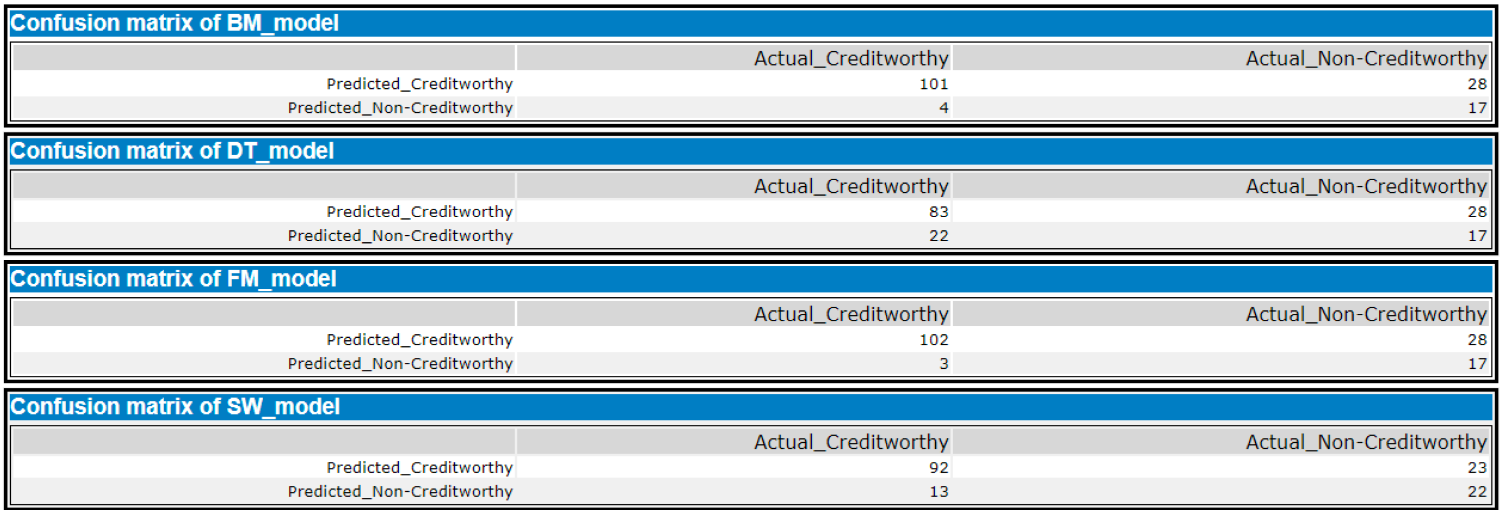
After validating all the models against the validation set, the comparison mdel shows the following:

- Percent Accuracy:



Model	Accuracy
FM_model	0.7933
BM_model	0.7867
DT_model	0.6667
SW_model	0.7600

- Confusion Matrices:



Confusion matrix of BM_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	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of FM_model

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

Confusion matrix of SW_model

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

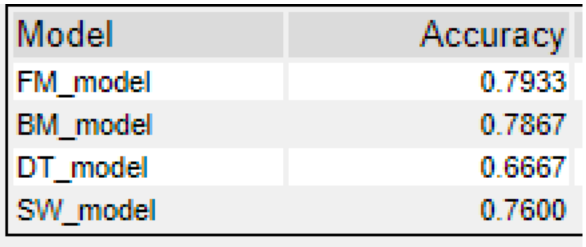
All the models show a clear bias toward creditworthy as all of them tend to predict more creditworthiness, though the most accurate and balanced model is **Forest Model.**

Step 4: Writeup

1. Which model has been chosen to be used?

There are four main techniques in order to select the best model for our analysis:

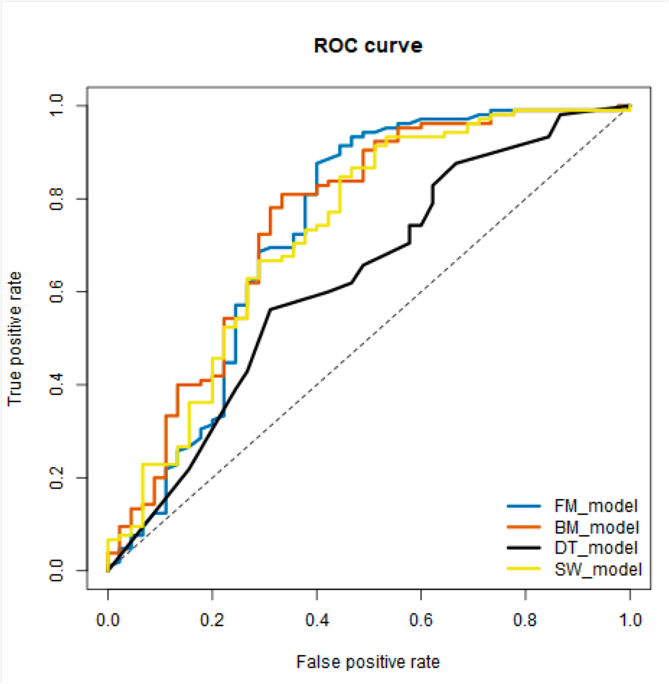
A. Overall Accuracy:



Model	Accuracy
FM_model	0.7933
BM_model	0.7867
DT_model	0.6667
SW_model	0.7600

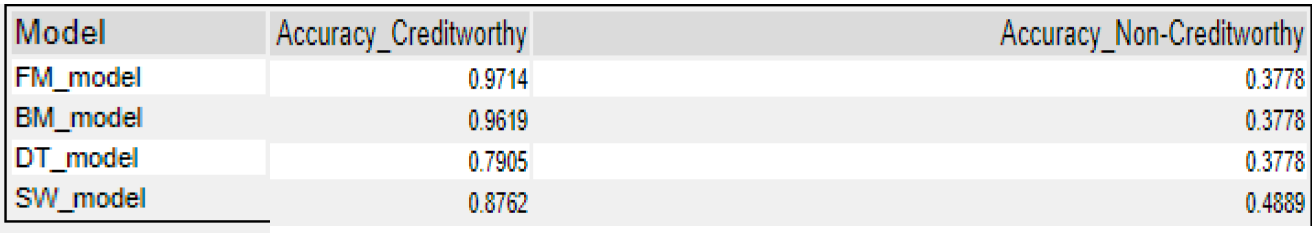
Forest Model has the highest accuracy among all the models.

B. ROC Graph:



Forest Model has the highest true positive rate among all the models.

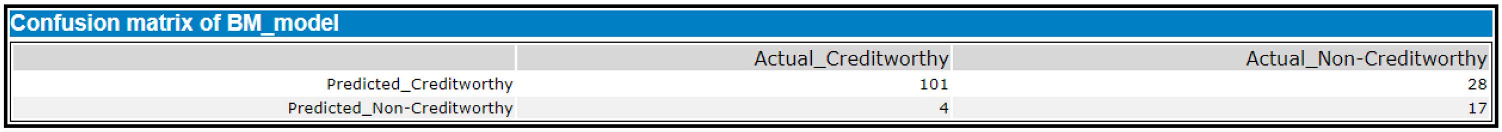
C. Accuracies within “Creditworthy” and “Non-Creditworthy” Segments:



Model	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_model	0.9714	0.3778
BM_model	0.9619	0.3778
DT_model	0.7905	0.3778
SW_model	0.8762	0.4889

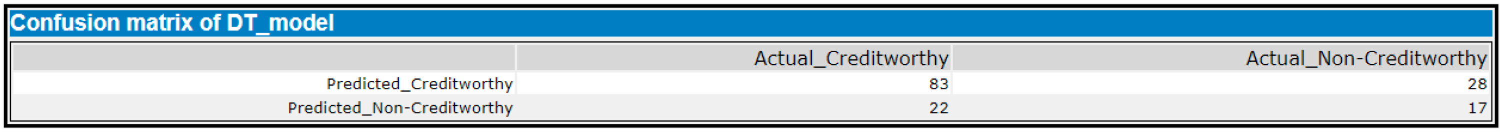
Forest Model has the highest “Creditworthy” and “Non-Creditworthy” accuracy among all the models.

D. Bias in the Confusion Matrices:



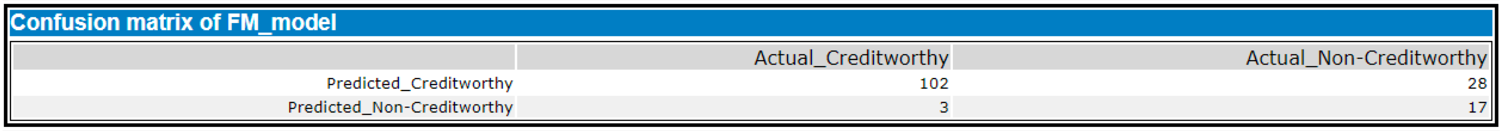
Confusion matrix of BM_model

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



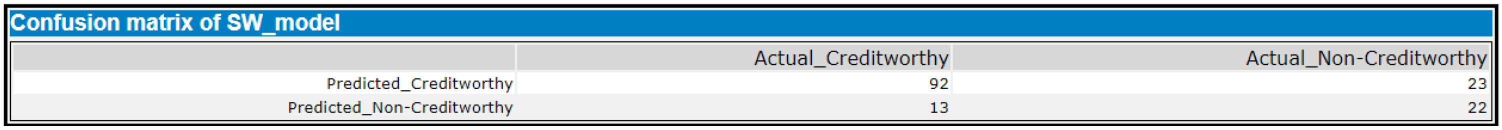
Confusion matrix of DT_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17



Confusion matrix of FM_model

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



Confusion matrix of SW_model

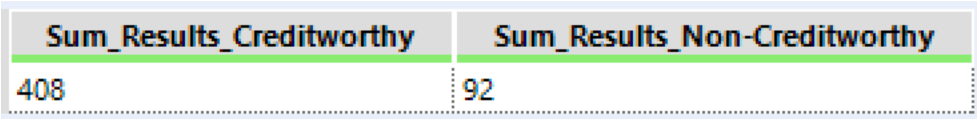
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

All the models show bias to creditworthiness which resulted in misclassifying non-creditworthy customers. Though, **Forest Model** has the most accurate results.

Conclusion:

All the four techniques proved that the most suitable model to be used is **Forest Model** due to the high accuracy among all segments.

2. How many individuals are creditworthy?



Sum_Results_Creditworthy	Sum_Results_Non-Creditworthy
408	92

408 individuals out of 500 are creditworthy depending on our model.