Project: Creditworthiness

Step 1: Business and Data Understanding

You are a loan officer at a young and small bank (been in operations for two years) that needs to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. You'll use a series of classification models to figure out the best model and provide a list of creditworthy customers to your manager.

Key Decisions:

Answer these questions

- What decisions needs to be made?

We need to predict which customers would be deemed as creditworthy.

- What data is needed to inform those decisions?

We need past data of customers and their categories to build models for prediction, specifically, account balance, duration of credit month, payment status of previous credit, credit amount, value savings stocks, length of current employment, installment percent, purpose, most valuable available asset, age, type of apartment, no of credits at this bank.

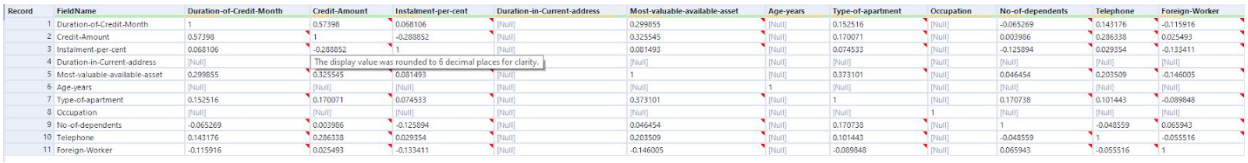
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Binary. Specifically, Logistic Regression, Decision Tree, Forest Model and Boosted Model.

Step 2: Building the Training Set

Data Wrangling:
Duration-in-current-address column was removed as it is missing a lot of data. **Occupation**, **Concurrent-Credits** are also removed as all entries have the same values in these two columns. **Foreign-workers, no-of-dependents** and **guarantor** were also removed due to low variability. **Telephone** was also removed as it is not relevant. For the missing entries in **Age** column, they were replaced with the median age of the column.

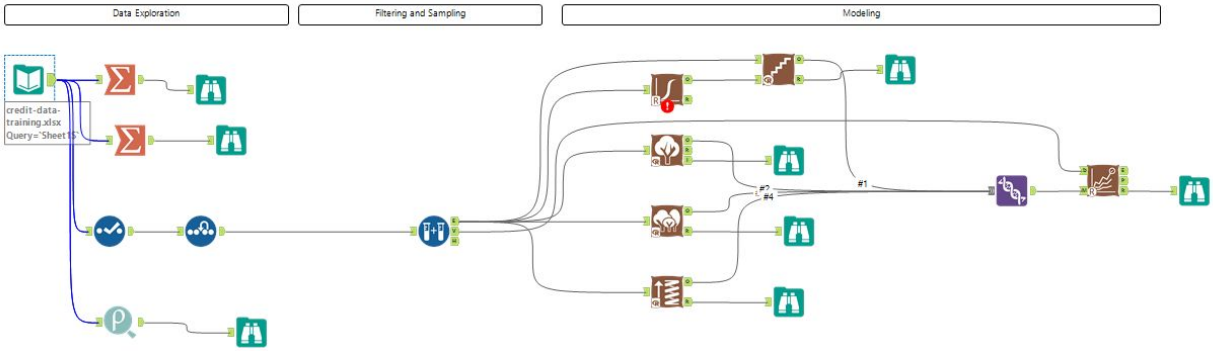
However the Pearson correlation test didn't show any correlations between variables.



Record	FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worker
1	Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269	0.143176	-0.115916
2	Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986	0.286338	0.025493
3	Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894	0.029354	-0.133411
4	Duration-in-Current-address	[Null]	The display value was rounded to 6 decimal places for clarity.		[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]
5	Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454	0.203509	-0.146005
6	Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
7	Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738	0.101443	-0.089848
8	Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
9	No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1	-0.048559	0.065943
10	Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559	1	-0.055516
11	Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943	-0.055516	1

Step 3: Train your Classification Models

I created Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of the entire dataset should be reserved for Validation. And then set the Random Seed to 1 and created the 4 models.



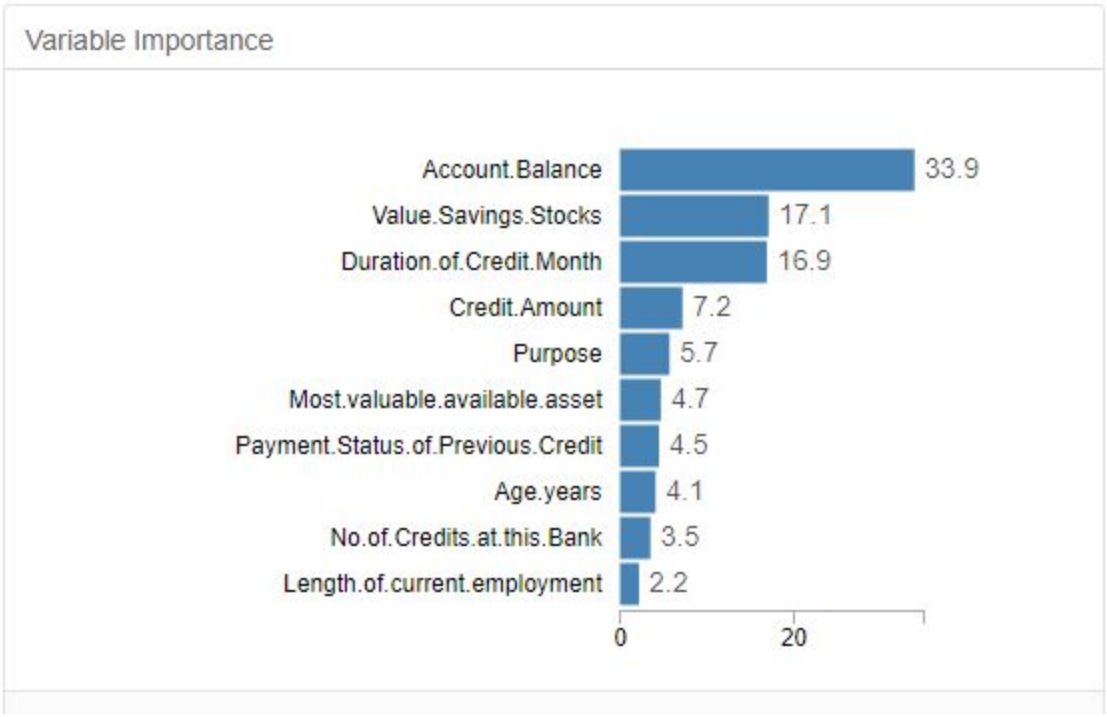
The target variable is the credit application result and predictors are everything else.

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	.

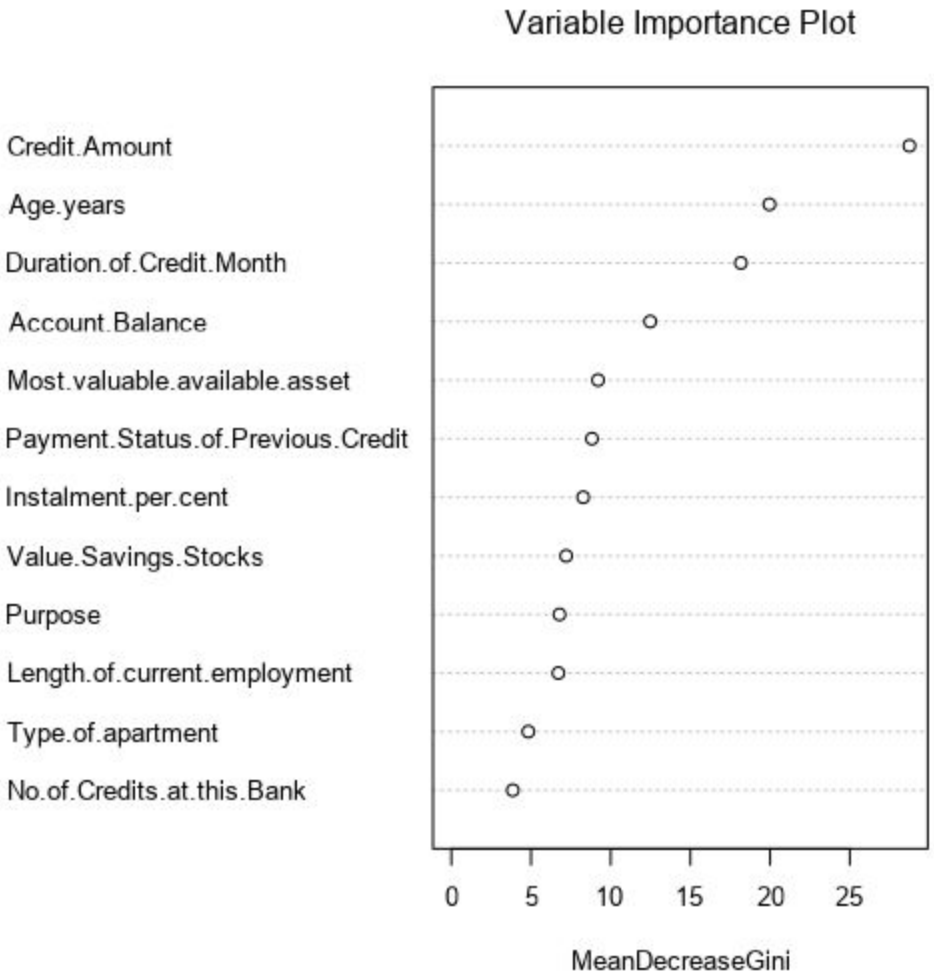
We can see that Account Balance has the smallest P value, which means it is the most important variable in predicting.

Decision Tree:



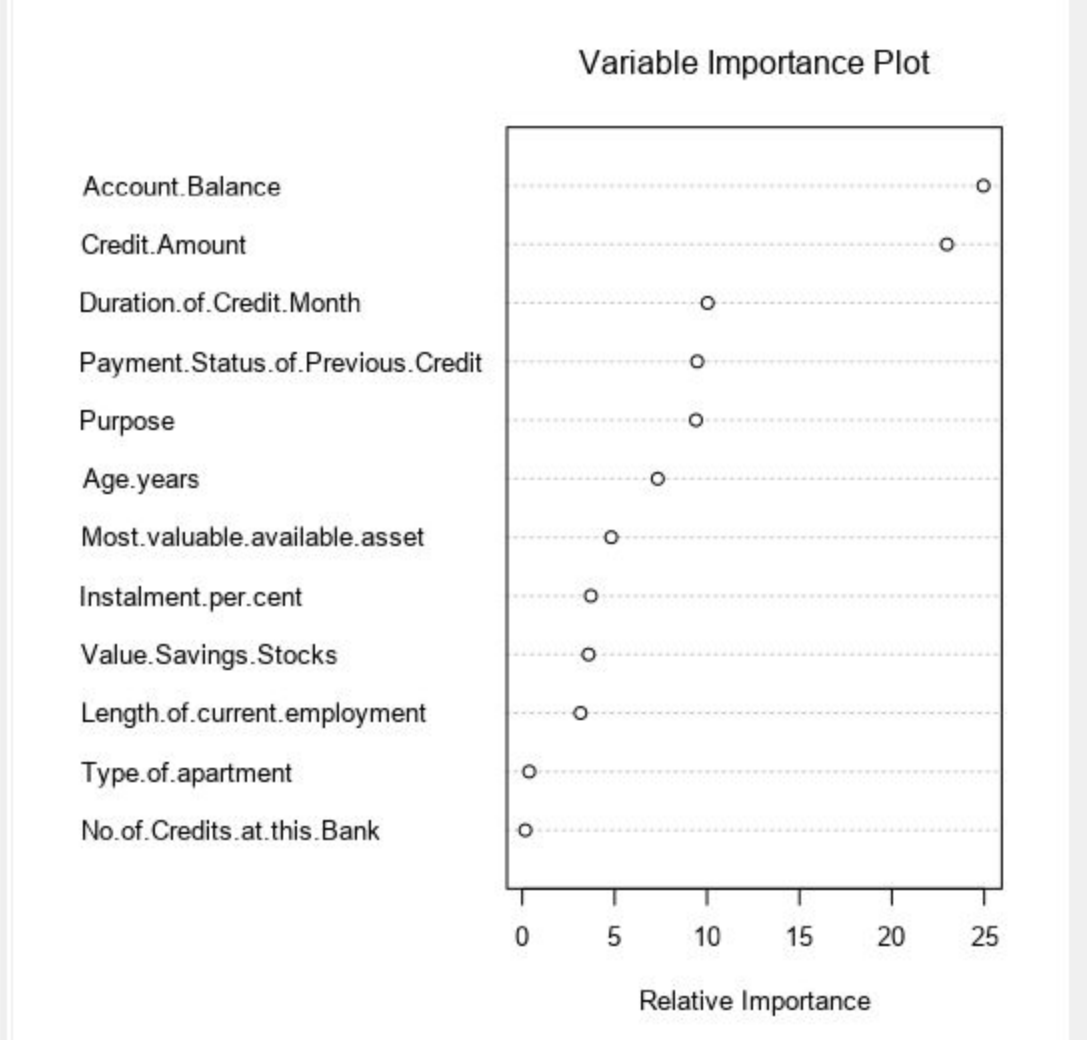
Similar to Logistic Regression, DT model also shows that account balance is the most important predictor, but followed by value savings stocks and duration of credit month.

Forrest Model:



We can see that Credit Amount, Age Years and duration of credit month are the most important.

Boosted Model:



Here account balance, credit amount are the most important predictors.

Finally, we used model comparison to gauge the overall accuracy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	0.4222
FM	0.7933	0.8681	0.7368	0.9714	0.3778
BM	0.7867	0.8632	0.7515	0.9619	0.3778
LR___Step	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of BM

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

Confusion matrix of DT

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

Confusion matrix of FM

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

Confusion matrix of LR___Step

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

Boosted Model:
PPV = 101/(101+28) = 0.78
NPV = 0.80

Decision Tree:
PPV = 0.78
NPV = 0.62

Forrest Model:
PPV = 0.78
NPV= 0.85

Logistic Regressions:
PPV = 0.8
NPV = 0.63

We can see the accuracy of each model from the screenshots above, we can see that logistic regression with step wise is the most accurate one. All of our models are doing a good job at predicting creditworthy, but not so much on non-creditworthy. This is due to the bias induced by the lack of non-creditworthy data in our dataset.

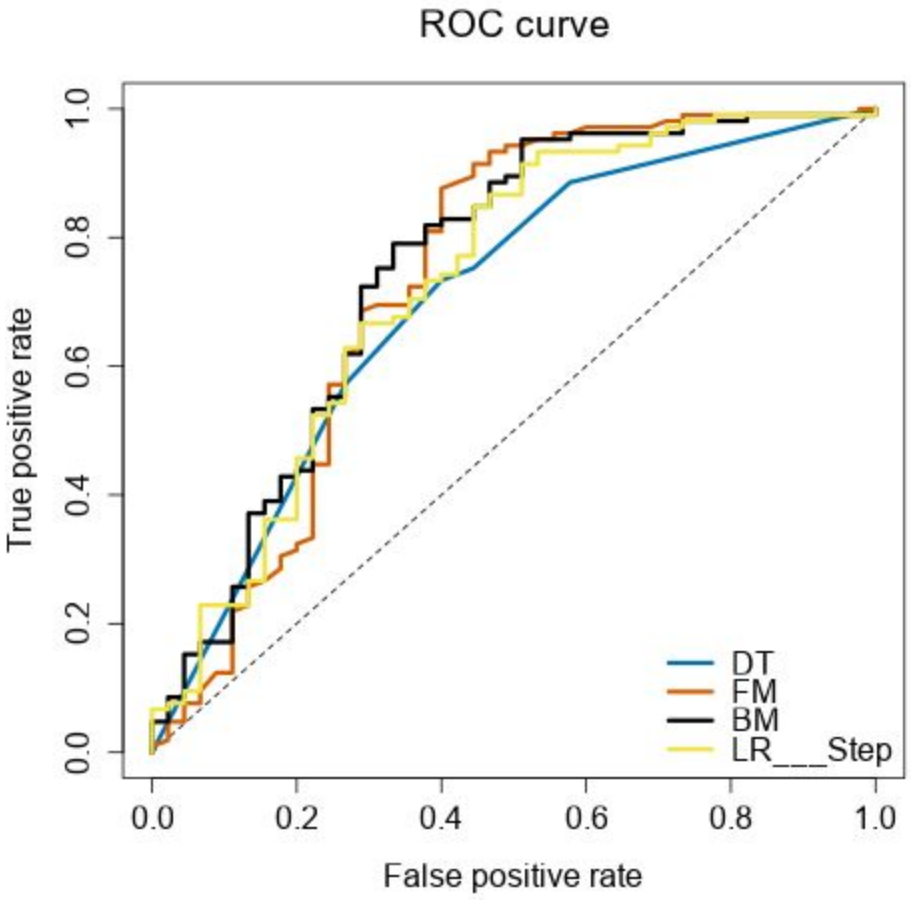
From the confusion matrices and the NPV/PPV values, we can see that Forrest Model overall has the highest values in both segments and they are very close, which means that the model is almost non-biased.

Step 4: Writeup

Answer these questions:

Which model did you choose to use?
I decided to use Forrest model for the following reasons:

- It has the highest overall accuracy against validation set, and the highest F1 score
- Its creditworthy accuracy is the highest and its highest noncreditworthy accuracy is not too far from the rest
- We can also see from the ROC graph that it is doing the best



- Lastly, referring back to our previous discussion on NPV/PPV values, Forrest Model has the highest NPV/PPV values and they are very close to each other - meaning that it is almost non-biased.

After apply the model to the other dataset, we found out that 406 out of 500 to be creditworthy.