Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

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

  I need to find efficient solutions for 500 new loans applications and classify new customers on whether they can be approved for a loan or not, and figure out which classification method is suitable for this issue.

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

  The data needed are credit-data-training.xlsx and customers-to-score.xlsx files, from which I could perform training set and find suitable model to follow.

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

  It is a Binary 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.***

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

  I have removed Concurrent-Credits, Guarantors, Foreign Worker, Occupation and No of dependents due to low variability; Duration-in-Current-address due to a lot of null values; Telephone because it is not relevant.

Also, I have imputed missing values for Age-years and set null values to be replaced with the median. After preparation all of mentioned data I've got 13 variables which I have used for further analysis.

Step 3: Train your Classification Models

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

  In the Logistic Regression Model the most significant variables are: Account.Balance, Payment.Status.of.Previous.Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent, Most.valuable.available.asset

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.StocksL100-L1000	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*

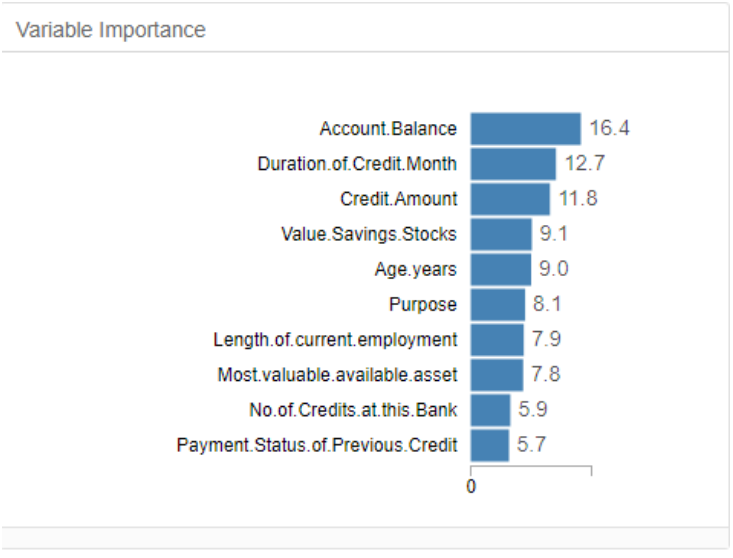
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275

In the Stepwise model the most significant variables are: Account.Balance, Payment.Status.of.Previous.Credit, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent.

Coefficients:

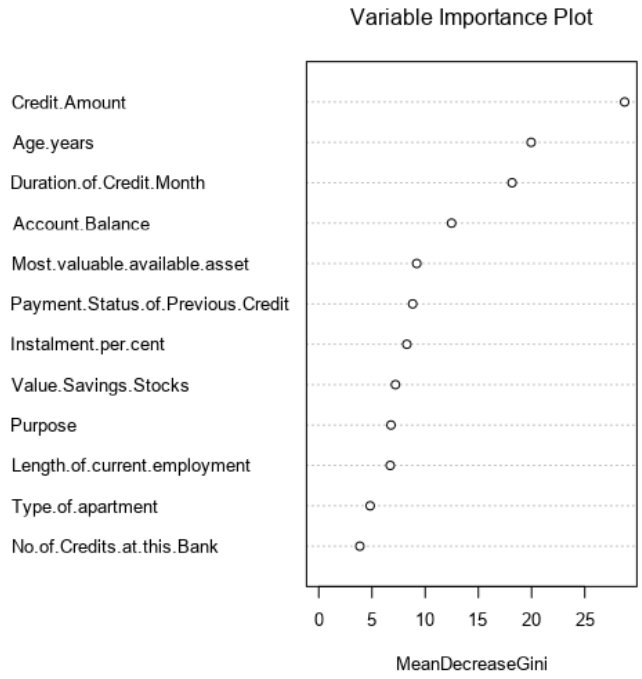
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	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.

In the Decision Tree model the most significant top five variables are: Account.Balance, Duration.of.Credit.Month, Credit.Amount, Value.Savings.Stocks and Age.years

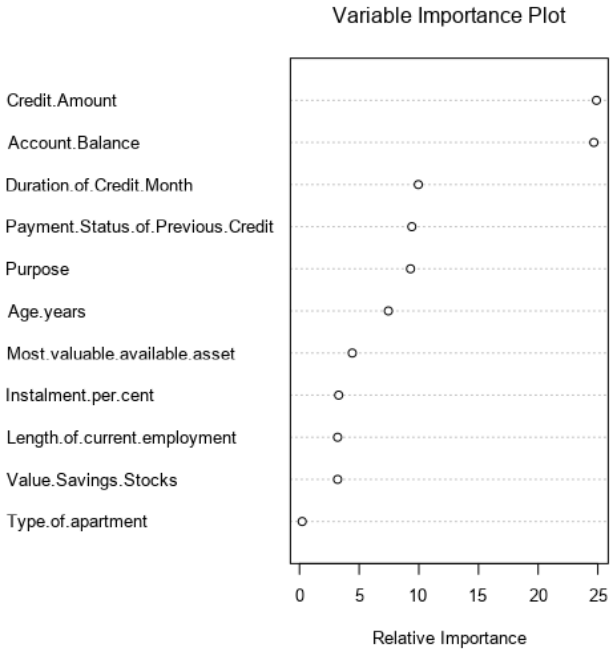


In the Forest model the top five significant variables are:

Credit.Amount, Age.years, Duration.of.Credit.Month, Account.Balance and Most.valuable.available.asset



In the Boosted model the significant two variables are: Credit.Amount and Account.Balance



- **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?**

This is the overall accuracy

Model	Accuracy
Decision Tree	0.6667
Forest Model	0.7933
Boosted model	0.7867
Logical Regression-Strepwise	0.7600

and the Confusion Matrix

Confusion matrix of Boost_M

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

Confusion matrix of DecTree

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

Confusion matrix of For_Mod

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

Confusion matrix of StepW

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

Yes there are bias in all models prediction, but Forest Model supposed to be the best predicted model because the overall accuracy is highest and it has less bias than Boosted Model.

Step 4: Writeup

- **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**

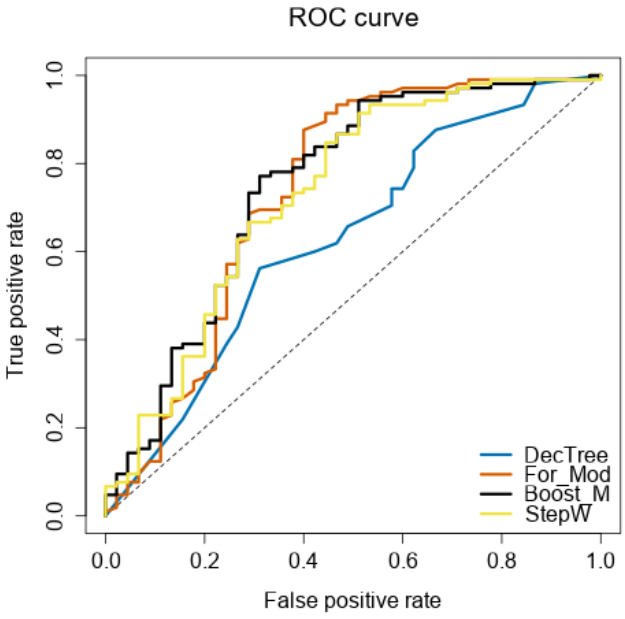
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecTree	0.6667	0.7685	0.6272	0.7905	0.3778
For_Mod	0.7933	0.8681	0.7368	0.9714	0.3778
Boost_M	0.7867	0.8632	0.7490	0.9619	0.3778
StepW	0.7600	0.8364	0.7306	0.8762	0.4889

To predict a customer creditworthiness and to score a model, I have used a Forest model because it has best accuracy level of 80%, and it has less bias compared to the other models. The results show that using the Forest model has greater accuracy of 0.9714 within Creditworthy and less level in Non-Creditworthy (0.3778) segment.

O



ROC graph shows us that Forest model curve is closer to 1 (top left corner), which means that it has greater accuracy level compared to the others.

- **How many individuals are creditworthy?**

After choosing the right model and using the Alteryx Score tool I have come to result of 408 Creditworthy customers.