Project: Creditworthiness

Step 1: Business and Data Understanding

1. What decisions needs to be made?
   The decision that needs to be made is whether the new load applicants are creditworthy or not, and accordingly should they be granted a loan.

2. What data is needed to inform those decisions?
   Some data that is needed to inform such decisions include, annual income, credit score, age, account balance, and loan amount.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
   A Binary Classification model should be used since predicting the results of new customers (as creditworthy or not) are based on available data.

Step 2: Building the Training Set

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  One variable was imputed and seven other variables were deleted, as follows:
  - First, the Age-Years field was imputed, which is one of the data we depend on in our decisions, as only about 2% is missing. Therefore, as instructed, the average value was used when *null* value is encountered.
  - The deleted variables are due to missing data, such as:
    - Duration-in-Current-address (69% missing data).
  - Or also variables were deleted due to valuelessness to our desired model as their values are distinct, such as:
    - Concurrent-Credits
    - Occupation
    - Guarantors
    - No-of-dependents
    - Foreign-Worker
  - Finally, one variable was deleted as its irrelevant to our desired model, which is:
    - telephone

Step 3: Train your Classification Models

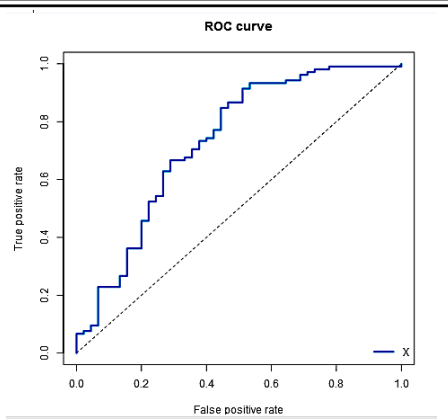
- Logistic Regression: using the Stepwise tool, the selected predictor variables are Credit Amount, Account Balance, Installment per cent, Purpose, Payment status of previous, and Length of current employment. The accuracy of this model reaches 78% even though the R-squared value is under 0.1936. Finally, the model has a slight bias in predicting the statuses for some clients who are creditworthy as non-creditworthy.

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	*

Confusion matrix of Logistic Regression:

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



- Decision Tree: some of the important predictor variables include Account Blanca, Value Saving Stocks, Duration of Credit Month, and Credit Amount. This model reaches an accuracy rate of 74.67% with false prediction for forty-nine creditworthy clients as non-creditworthy.

EA: Awesome:

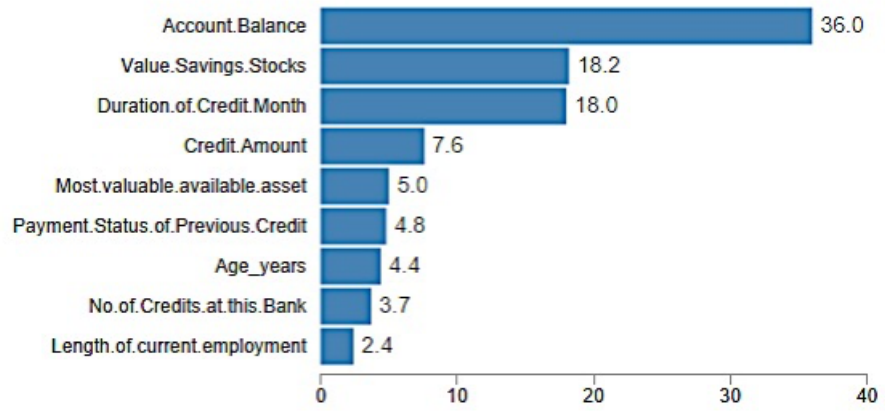
Confusions matrices are included.

Good job identifying bias in the models predictions.

If there is a big difference between the accuracy of the two classes, it's an indication of biased predictions. E.g decision tree cannot predict the non-creditworthy class with the same accuracy as the creditworthy class.

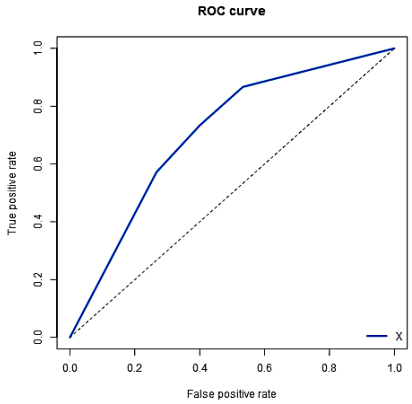
As you point out, logistic regression misclassifies many creditworthy applicants. Decision tree has even more false negatives (14). Forest has only 3. So decision tree and logistic regression would deny loans to many creditworthy applicants.

Variable Importance

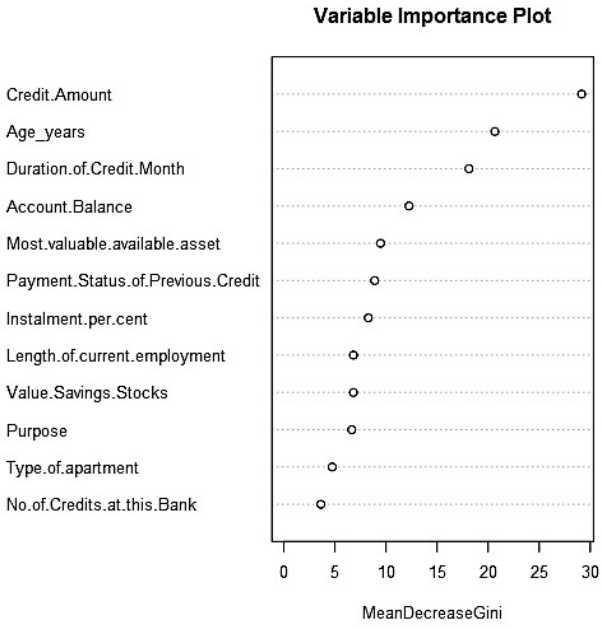


Confusion matrix of Decision Tree:

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

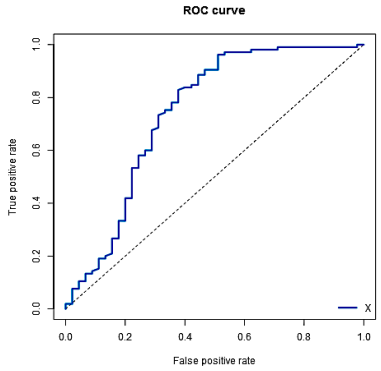


- Forest Model: some of the important variables include Credit Amount, Age Years, Duration of Credit Month, and Account Balance. The model reaches an accuracy rate of 82% with slight improvement in predicting non-creditworthy clients.

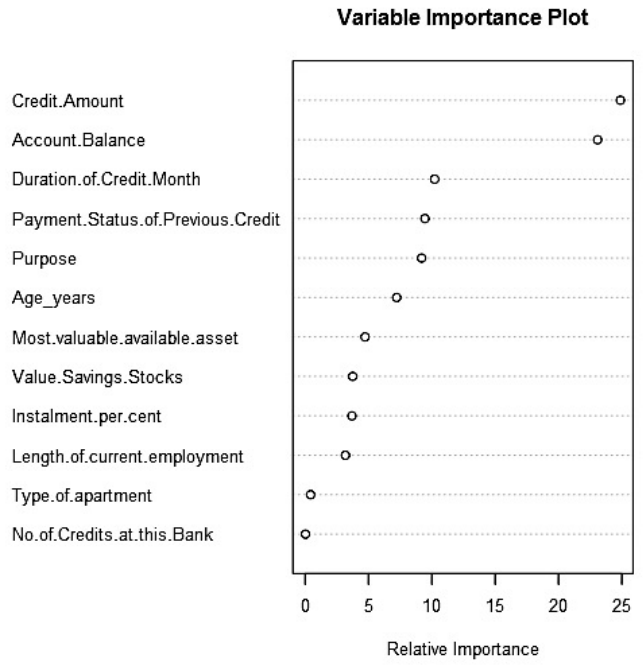


Confusion matrix of Forest Model:

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

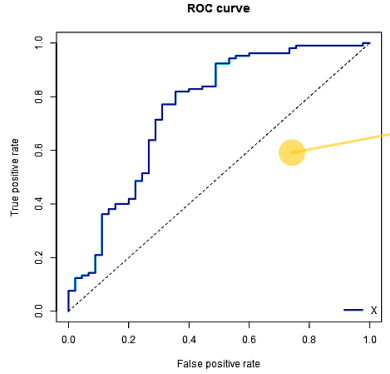


- Boosted Model: some of the important predictor variables include Credit Amount, Account Balance, and Duration of Credit Month (among others). The accuracy rate of this model reaches 79.33% with no bias in prediction.



Confusion matrix of Boosted Model:

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18



EA: Suggestion:

By using the union tool in Alteryx, we can visualize all curves in the same graph. It makes the comparison a little easier.

Step 4: Writeup

By comparing the four models using the Union tool, it can be concluded that the Forest model has the highest rate of accuracy (i.e. 82%), and also the rates of creditworthy and non-creditworthy clients are higher than other models in term of accuracy, which can be examined though the confusion matrix.

The most important reason is due to the shape of ROC curve where the ROC curve of the Forest model has raised true positive rate and unchanged false positive rate. In contrast to the ROC curves of other models, the ROC curve of Forest model remains its trend throughout.

Finally, by using the Score tool, the prediction of creditworthy clients is 408 while the other 92 clients would be considered non-creditworthy, otherwise.

EA: We should look closer at Accuracies within “Creditworthy” and “Non-Creditworthy” segments. In the creditworthy segment, all models have an accuracy of around 80%. When looking at the non-creditworthy accuracy, we can see that forest has a big advantage. Forest has a non-creditworthy accuracy well above 80%. E.g decision tree has only 60%