Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?
  The key business decision that needs to be made is with regards to the processing of the 500 new loan applications - to decide which loan applications to approve by deciding whether the applicant is creditworthy based on available data. Because of the sudden increase in the number of new applications, there is a need to automate the process, rather than approve them by hand, so that all these applications can be processed in one week.

- What data is needed to inform those decisions?
  Data on past applications as well as the new applicants are needed for the following data points – account balance, duration of the credit month, payment Status of previous credit, purpose for which loan is required, the credit amount required, value of savings and stocks, length of current employment, the instalment percent, if there are any guarantors, duration of stay in the current address, most valuable available asset, age years, concurrent credits (if any), type of apartment living in, number of credits at this bank, occupation, number of dependents, whether they are a foreign worker or not.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Since the target variable is to decide the creditworthiness, which can take only two values – Creditworthy or Non-creditworthy, the model needed is a classification problem that is binary. Hence, we need a Binary Classification 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.***

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".

  There are no variables that highly-correlate with each other. Credit month and Duration of Credit Month have a correlation with a coefficient of 0.57 but not high (not > 0.70). The other pairs of variables show low correlation.

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

  - The field "Age-years" has 2% missing data. This missing data is imputed with the median of the rest of the field, to fill in the null values.

  - The field "Duration in Current address" has 69% data missing. Since such a high percentage of the field is missing, it is being filtered out from consideration since it will not provide useful information for the model.

- 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" and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

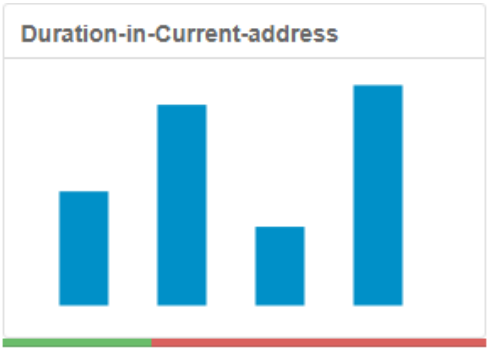
  - "Guarantors" field has low variance - 457 instances of None and only 43 instances of Yes.
  - "Concurrent-Credits" field has no variance at all - 500 instances of Other Banks/Depts
  - "Type of Apartment" field has low variance - 352 instances of 2, and only 92 of 1 and 56 of 2 respectively.
  - "No. of dependents" field has low variance - 427 instances of 1 and 73 instances of 2
  - "Foreign worker" field has low variance - 481 instances of 1 and only 19 instances of 2

  Since these 5 fields have low variance, they can heavily skew towards one type of data. For this reason, they are being filtered out from consideration.

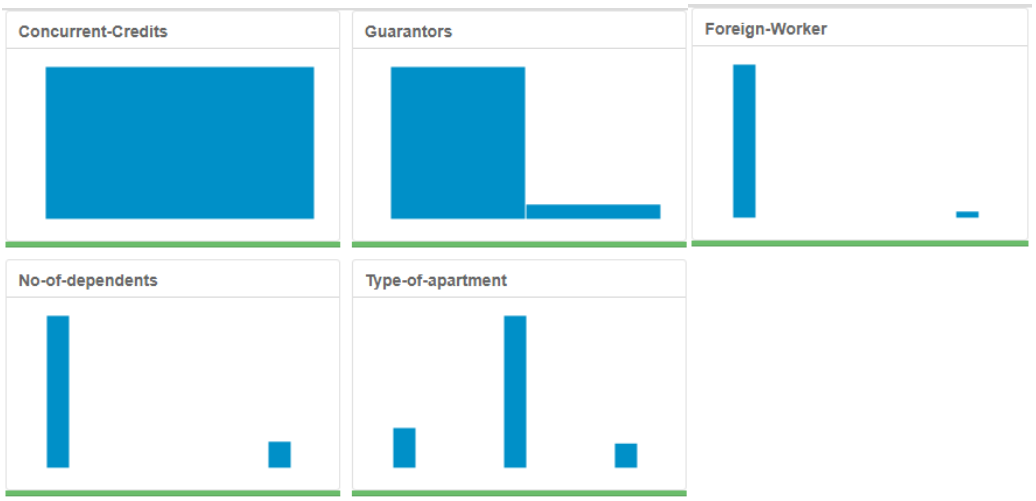
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Finally, the “Telephone” field is being removed since it will not give any useful information about the creditworthiness of the applicant and hence, will not be a strong predictor of the target variable.

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  - The “Duration in Current address” field was removed because it had 69% missing data.



  - The 5 fields - “Guarantors”, “Concurrent-Credits”, “Type of Apartment”, “No. of dependents” and “Foreign worker” – were removed because of low variance of values in the respective fields because of which they can heavily skew towards one type of data.



- The “Telephone” field is being removed since it will not give any useful information about the creditworthiness of the applicant and hence, will not be a strong predictor of the target variable.
- The “Age-year” field has 2% missing data; therefore, the null values are imputed in Alteryx with the median (value = 33) of the entire data field. The decision was taken to choose the median to impute since the range of values of the field is high with several outliers above the upper fence of 64.5. This could skew the data upwards if the imputation is done with the mean. The imputation gave the average of the “Age-years” field as 36 (35.574).

Step 3: Train your Classification Models

Logistic Regression Model

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

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.2290394	9.845e-01	-3.2800	0.00104 **
Account.BalanceSome Balance	-1.5843791	3.200e-01	-4.9511	7.38e-07 ***
Duration.of.Credit.Month	0.0058321	1.365e-02	0.4272	0.6692
Payment.Status.of.Previous.CreditPaid Up	0.4306851	3.847e-01	1.1195	0.26294
Payment.Status.of.Previous.CreditSome Problems	1.2872278	5.339e-01	2.4109	0.01591 *
PurposeNew car	-1.7472435	6.271e-01	-2.7862	0.00533 **
PurposeOther	-0.2780516	8.305e-01	-0.3348	0.73778
PurposeUsed car	-0.7651003	4.108e-01	-1.8624	0.06255 .
Credit.Amount	0.0001734	6.833e-05	2.5375	0.01116 *
Value.Savings.StocksNone	0.5996934	5.065e-01	1.1840	0.2364
Value.Savings.Stocks£100-£1000	0.1818563	5.621e-01	0.3236	0.74628
Length.of.current.employment4-7 yrs	0.5259720	4.934e-01	1.0660	0.28642
Length.of.current.employment< 1yr	0.7776684	3.951e-01	1.9681	0.04906 *
Instalment.per.cent	0.2969774	1.384e-01	2.1457	0.0319 *
Most.valuable.available.asset	0.2877408	1.488e-01	1.9337	0.05315 .
Age.years	-0.0180861	1.475e-02	-1.2259	0.22022
No.of.Credits.at.this.BankMore than 1	0.3918288	3.812e-01	1.0280	0.30397
Occupation	NA	NA	NA	NA

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 350 degrees of freedom
Residual deviance: 323.08 on 333 degrees of freedom
McFadden R-Squared: 0.218, Akaike Information Criterion 357.1

The most significant predictor variables are Account Balance (Some Balance), Purpose (new car), Credit Amount, Payment status of previous credits (Some problems), Instalment percent and Length of current employment (< 1 year) in decreasing order.

- 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?

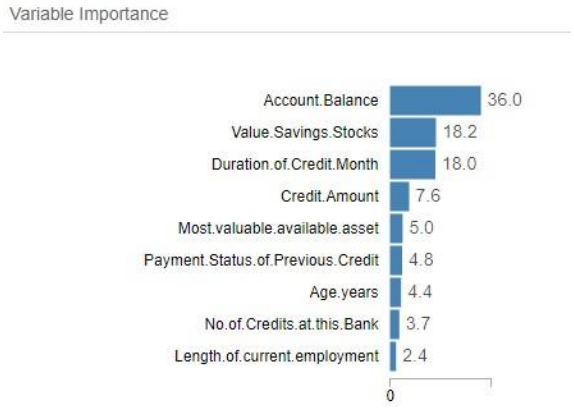
  The overall accuracy is 76%.

Confusion matrix of Logistic_Step_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

The confusion matrix is shown above. It shows a bias towards Creditworthy with a higher accuracy of prediction (87.62%) compared to Non-creditworthy (48.89%).

Decision Tree Model

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



The most important predictor variables according to the Decision Tree model are Account Balancce, Value Savings Stocks, Duration of Credit Month and Credit Amount. The other ones shown as significant are Most valuable available asset, Payment status of previous credit, Age (years), No. of credits at this bank and length of current employment.

- 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?

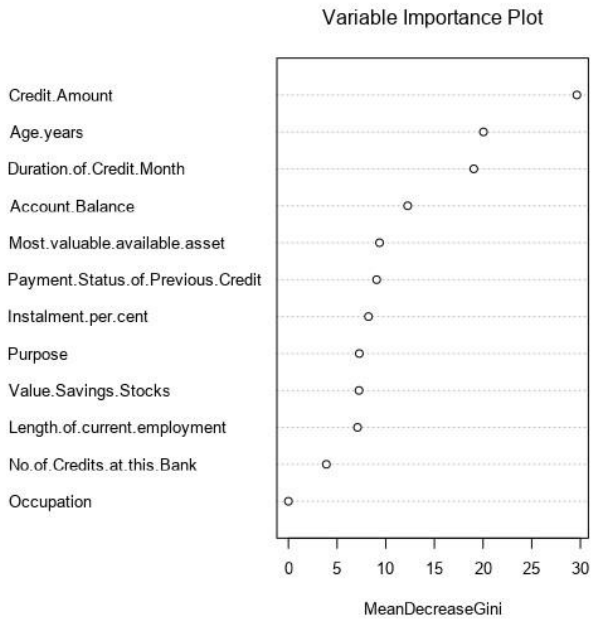
  The overall accuracy is 74.67%.

Confusion matrix of DT_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

The confusion matrix is shown above. It shows a bias towards Creditworthy with a higher accuracy of prediction (86.67%) compared to Non-creditworthy (46.67%).

Forest Model

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



The most important variables as per the Forest Model are Credit Amount, Age (years), Duration of Credit Month and Account Balance. The other ones are Most valuable available asset, Payment Status of Previous Credit, Instalment percent, Purpose, Value Savings stocks, length of current employment. No. of credits at this bank show very low importance whereas Occupation is not important at all.

- 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?

The overall accuracy is 81.33%.

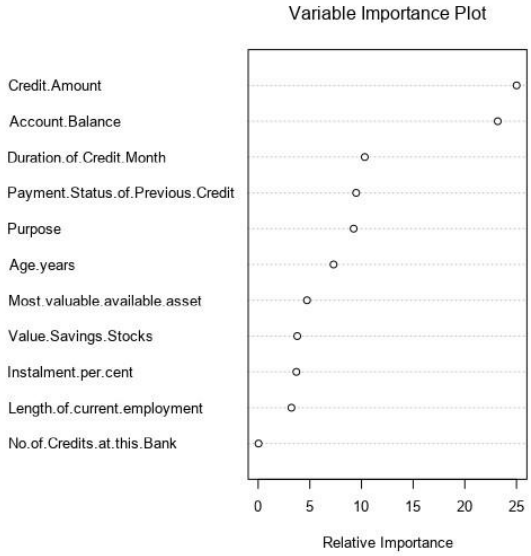
Confusion matrix of RF_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	26
Predicted_Non-Creditworthy	2	19

The confusion matrix is shown above. It shows a greater bias than the previous two models towards Creditworthy with a higher accuracy of prediction (98.1%) compared to Non-creditworthy (42.22%).

Boosted Model

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



The most important variables as per the Boosted Model are Credit Amount and Account Balance. This is followed by Duration of Credit Month, Payment Status of Previous Credit, Purpose, Age (years), Most valuable available asset, Value Savings stocks, Instalment percent, length of current employment. No. of credits at this bank is not important at all.

- 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?

  The overall accuracy is 79.33%.

Confusion matrix of Boosted_Credit

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

The confusion matrix is shown above. It shows a greater bias than the first two models towards Creditworthy with a higher accuracy of prediction (96.19%) compared to Non-creditworthy (40.0%).

Step 4: Writeup

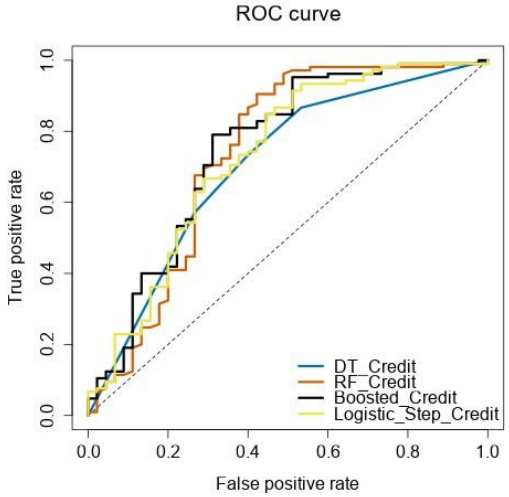
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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)

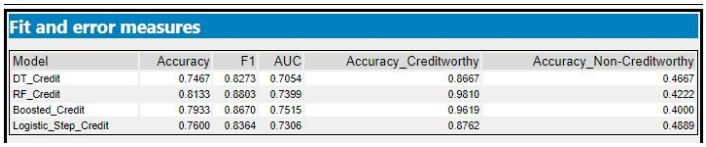
Answer these questions:

- 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

  The Forest Model has been chosen.
    - It has the highest overall accuracy of 81.33%, which is higher by 2% compared to the next most accurate model (Boosted model with 79.33% accuracy).
    - Its accuracy for the Creditworthy segment at 98.1% is the highest for all models. Though the accuracy for Non-creditworthy segment at 42.22% is lower than that of Logistic (48.89%) and Decision Tree (46.67%), this reduction is lower compared to the gain in the accuracy for the Creditworthy segment.
    - The ROC graph shows that the Forest Model is generally higher than other models on average and reaches the top true positive rate the fastest. This is an indicator of a good predictive model. The ROC graph is shown below:



- If we observe the accuracy scores for the different models (indicating the bias in the confusion matrices), we see that the Logistic and Decision Tree models show the small bias (with a difference of about 40% in accuracy of predictions between Creditworthy and Non-creditworthy segments). But these models have much lower overall accuracy. Between Forest and Boosted Models, the difference in accuracy of predictions is about 56% but Forest Model has higher overall accuracy as well as segment-wise accuracy levels.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
RF_Credit	0.8133	0.8803	0.7399	0.9810	0.4222
Boosted_Credit	0.7933	0.8670	0.7515	0.9619	0.4000
Logistic_Step_Credit	0.7600	0.8364	0.7306	0.8762	0.4889

Based on the above considerations, the Forest Model was chosen.

- How many individuals are creditworthy?

  Based on the Forest Model, 411 individuals are creditworthy whereas 89 are not creditworthy.