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?

Need to evaluate the creditworthiness of new loan applications based on customer data.

: Awesome: Good job identifying the key decision to be made.

- What data is needed to inform those decisions?

We will decide the final creditworthiness based on Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, and so on.

: Awesome: All these data should indeed prove useful in our analysis. Well done coming up with these example data.

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

We will evaluate creditworthiness using Binary classification models, including Logistic Model, Decision Tree, Forest Model and Boosted Model.

: Awesome: The correct model type has been identified.

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.***

<u>Here are some guidelines to help guide your data cleanup:</u>

- 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”.
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- 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.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

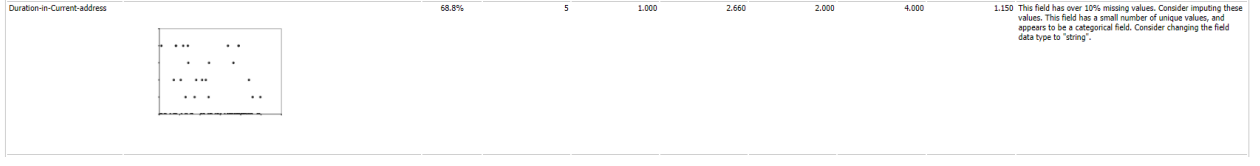
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

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

I imputed Age-years with median, to prevent huge impact of age outliers, using median instead of average is appropriate. Additionally, I excluded No-of-dependents, Occupation, Foreign-Worker, Concurrent-Credits, Guarantor and Type-of-apartment due to low variability. If we forced our model consider these data fields, it may result in over-fitting. Duration-in-current-address is also removed due to 68.8% missing values.



: Awesome: This decision is correct. We impute because not much data is missing here. And we use the median because of the presence of a slight skew in it's distribution.

: Awesome: All the appropriate fields have been removed alongwith the correct justifications.

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.

(a.) Logistic Regression Model

Amount-Balance is the most important predictor, which-value is 0.00014, quite small.

: Suggestion: Whenever using the Logistic Regression model, it's suggested to also mention if the stepwise tool was used. Recall from the lesson that stepwise automates the process of coming up with the best predictor variables, thereby improving our efficiency of coming up with the final solution.

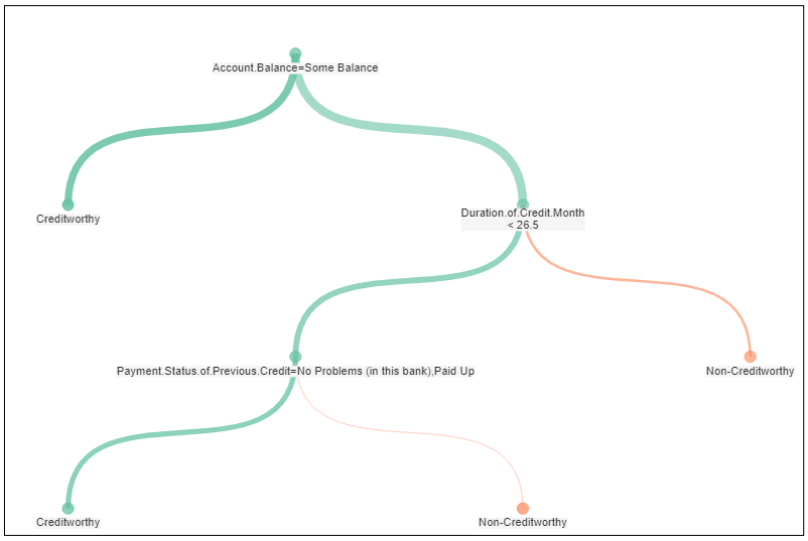
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.7915335	1.154e+00	-4.15364	3e-05	***
Account.BalanceSome Balance	-1.2201931	3.210e-01	-3.80133	0.00014	***
Credit.Amount	0.0001889	6.907e-05	2.73543	0.00623	**
Duration.of.Credit.Month	0.0165561	1.359e-02	1.21848	0.22304	
Instalment.per.cent	0.4728272	1.525e-01	3.09986	0.00194	**
Length.of.current.employment4-7 yrs	0.5801341	5.222e-01	1.11094	0.26659	
Length.of.current.employment< 1yr	0.9028676	4.337e-01	2.08164	0.03738	*
Most.valuable.available.asset	0.2948866	1.516e-01	1.94523	0.05175	.
No.of.Credits.at.this.BankMore than 1	-0.3391001	4.046e-01	-0.83813	0.40196	
Payment.Status.of.Previous.CreditPaid Up	0.3374750	4.128e-01	0.81756	0.41361	
Payment.Status.of.Previous.CreditSome Problems	1.5429005	5.582e-01	2.76420	0.00571	**
PurposeNew car	-1.9774633	6.670e-01	-2.96476	0.00303	**
PurposeOther	-0.3552814	1.070e+00	-0.33201	0.73988	
PurposeUsed car	-0.4611278	4.250e-01	-1.08497	0.27793	
Telephone	-0.2200734	3.204e-01	-0.68689	0.49215	
Value.Savings.StocksNone	1.0534489	5.110e-01	2.06168	0.03924	*
Value.Savings.Stocks £ 100- £ 1000	0.4582878	5.684e-01	0.80624	0.4201	
Age_years	-0.0012172	1.429e-02	-0.08519	0.93211	

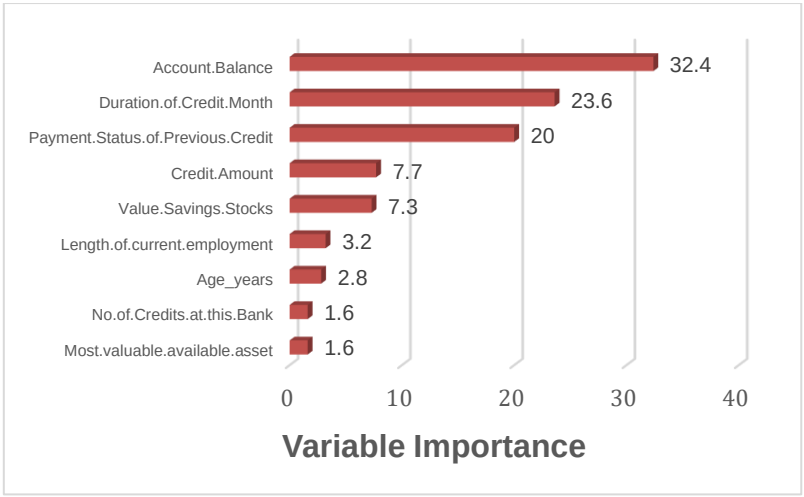
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

(b.) Decision Tree Model

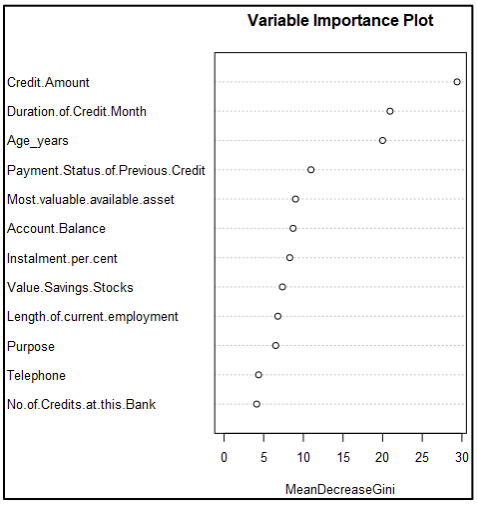
In this decision tree model, the most significant feature is **Amount-Balance**.





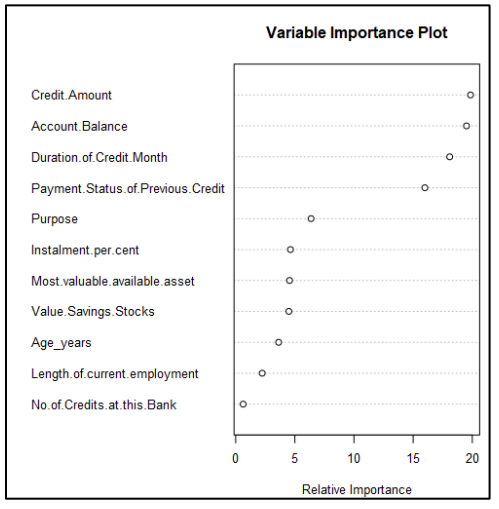
(c.) Forest Model

In Forest Model, the most importance variable is **Credit-Amount**.



(d.) Boosted Model

In this Boosted Model, the most important features are **Credit-Amount** and **Account-Balance**.



: Awesome: Excellent work appropriately setting up the four models to come up with the correct set of the most significant variables for each of them.

- 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 **forest model** presented the best accuracy for validation data. From the confusion matrix, I think these four models are comparable.

Model	Accuracy
Logistic	76.00%
Decision_Tree	76.00%
Forest	77.33%
Boosted	74.67%

Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	32
Predicted_Non-Creditworthy	6	12

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	29
Predicted_Non-Creditworthy	7	15

Confusion matrix of Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	29
Predicted_Non-Creditworthy	5	15

Confusion matrix of Logistic

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	29
Predicted_Non-Creditworthy	7	15

: Awesome: Great job appropriately validating the models to come up with the correct confusion matrices.

You should have four sets of questions answered. (500 word limit)

Step 4: Writeup

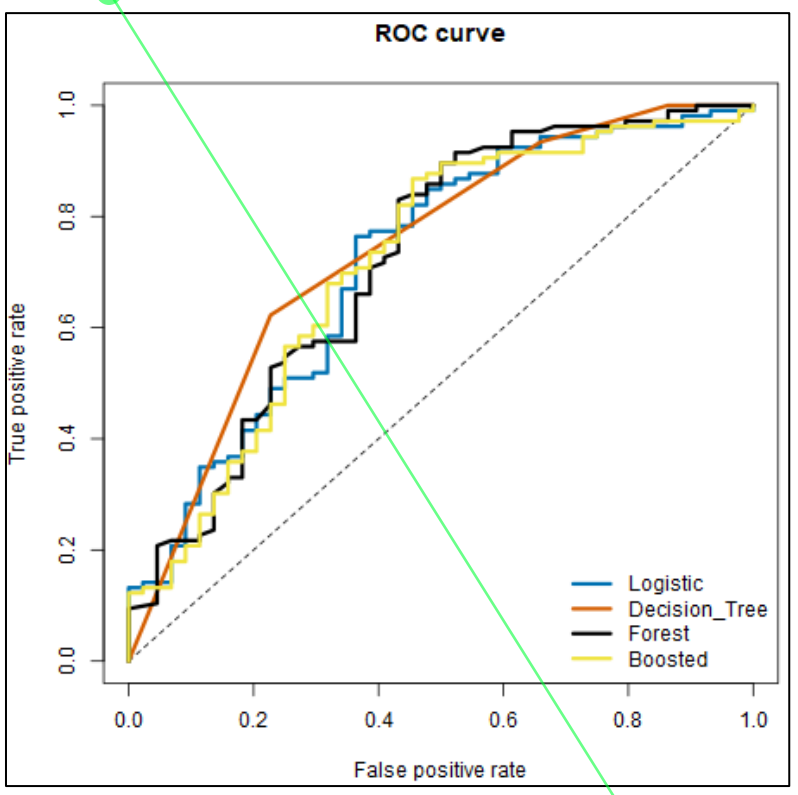
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

For our business, we will focus on the accuracy of Creditworthy and Non-Creditworthy segments, although Decision Tree has higher value of AUC. From the two tables shown below, the Forest model presented the best performance for overall, creditworthy and non-creditworthy accuracy. As for bias in the confusion metrices, there is no huge difference between four models. Thus, the forest model is recommended.



: Comment: Note that we are also interested in the speed at which a model rises in the ROC curve. Note that the forest model rises the fastest, meaning that we are getting a higher rate of true positive rates vs. false positives. This is important because we do not want to extend loans to people who are not creditworthy.

: Comment: This statement isn't entirely correct. Do note that the "accuracy_creditworthy" in the Alteryx report actually represents the True Positive Rate, i.e. the cases when the model predicted creditworthy when it was actually creditworthy. Similarly the "Accuracy_noncreditworthy" represents the True Negative Rate - i.e. cases where the model predicted non-creditworthy when it was actually non-creditworthy. However, we do not use only these parameters to detect bias in a model. The place where you should be looking for bias is in the confusion matrix. We calculate two values - True Negative Value and Precision. If these values are close to each other, we say that the model is almost unbiased, and if these values differ by a substantial amount, then the model is biased (the calculation of these values weren't explicitly covered in the lesson. If you want, you may skim through the following Wikipedia article - https://urlzs.com/QEqES). Let's look at two examples - check out the confusion matrix for your forest model. We calculate True Negative Value which is defined as (Num of true negatives)/(Total num of negatives). From your table, this will be 15/20 = 75%. Similarly, we calculate the Precision = (Num of true positives) /

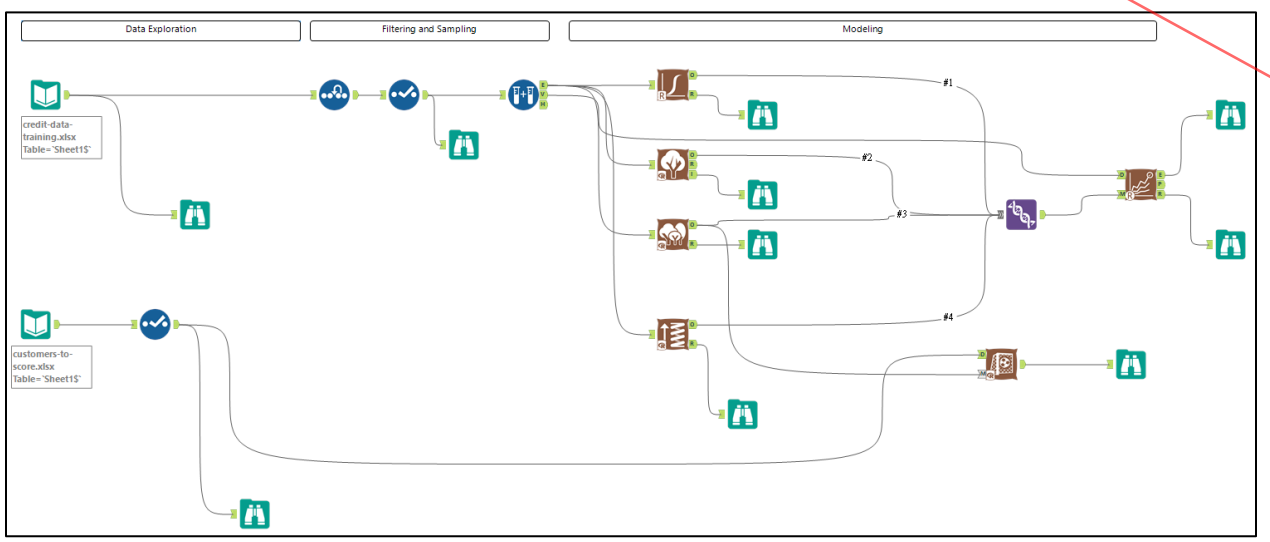
Model	AUC
Logistic	0.71848
Decision_Tree	0.74099
Forest	0.72202
Boosted	0.71677

Model	Accuracy	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic	0.760000	0.773438	0.681818
Decision_Tree	0.760000	0.773438	0.681818
Forest	0.773333	0.776923	0.750000
Boosted	0.746667	0.757576	0.666667

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?
  Based on forest model, **421** individuals are creditworthy.

Appendix



(Total num of positives). From your table, this will be 101/130 = 77.6%. Notice that there isn't too big of a difference in these two values for the forest model. Hence, we say that this model is low in bias. As another example, let's calculate these values for your decision tree model. The True Negative Value = (15/22) = 68.1%. And the Precision = (99/128) = 77.3%. Now notice the big difference in these values. Hence, we say that the decision tree is biased. I would suggest you similarly calculate these values for the other two models as well. You should be able to note that the decision tree and logistic regression models are biased while the forest and boosted models are not. This is how we should be interpreting any bias in the models.

: Awesome: The best model has been correctly chosen and the decision appropriately justified.

: Comment: Note that the low bias in the forest model is another of it's advantage. We shouldn't be choosing models that are biased towards predicting individuals who are creditworthy, as they do not predict individuals who are not creditworthy nearly at the same level as those who are. This is bad for 2 reasons: 1. Loans will be extended to people who are not creditworthy leading towards bad loans 2. Opportunity will be missed by not extending loans to people who are creditworthy.

: Required: The final number of creditworthy individuals is not correct. Make sure that all the appropriate predictor variables were used in setting up the forest model. Only those variables that you identified above should be dropped. Also, be sure to double check if the appropriate imputed field containing age was used in the model. Once you use all the correct predictor variables, you should obtain a final value between 405 and 420 for the number of creditworthy individuals.