Chris Hom

Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?
    - Identify which loan applications should be approved or denied.

- What data is needed to inform those decisions?
    - Past loan applications needed to help inform the decision. The last applications should have multiple customer features and whether the loan was approved or denied
    - Current Loan applications that need to be scored.
    - Data points like: age in years, purpose of the loan, credit amount, type of apartment is needed to build the model and score the applications.

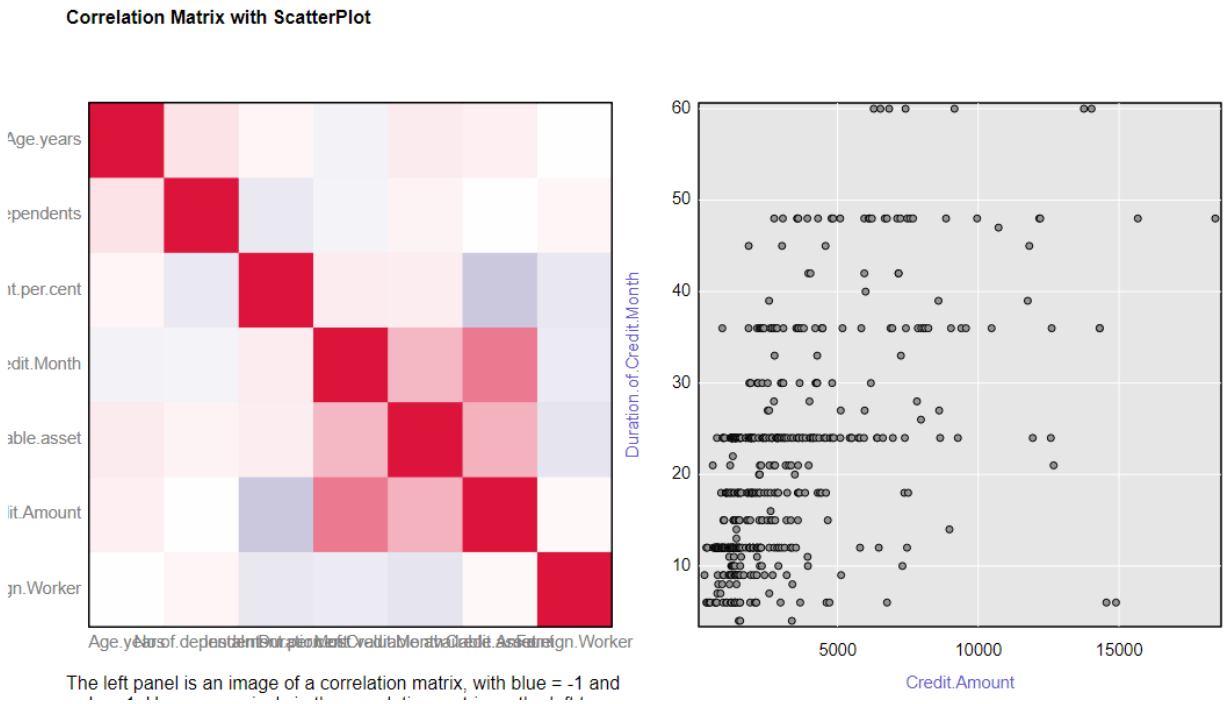
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

    - This is a binary model. There are two outcomes: Approved (creditworthy), not approved (not creditworthy). We need to use a predictive model to determine if an application is creditworthy.

Step 2: Building the Training Set

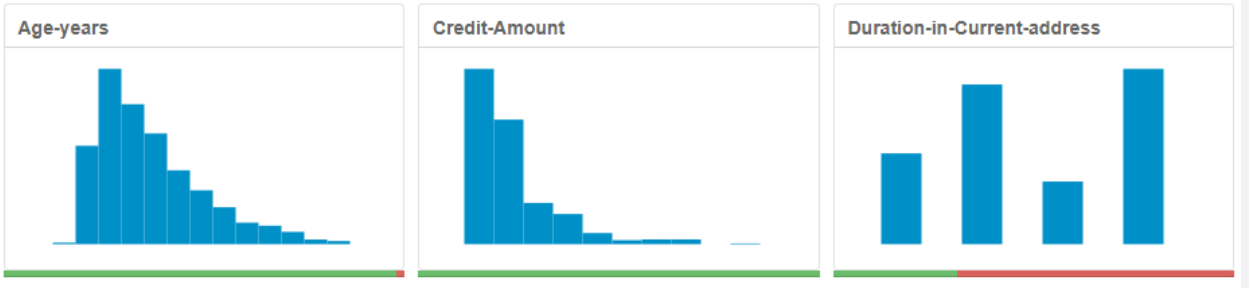
Here are some guidelines to help guide your data cleanup:

- 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 highly correlated fields. However, there is a medium correlation between Duration of Credit and Credit Amount (57%). I will not remove any fields due to correlation.



- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
    - Yes there are two fields with missing data: Age and Duration in Current Address.

Duration in Current Address has 69% missing rows and is to be removed completely from the analysis. Age will be imputed using the median of the dataset. We use Median because the distribution is skewed to the right and the mean will not be a correct representation of “center” of the distribution.

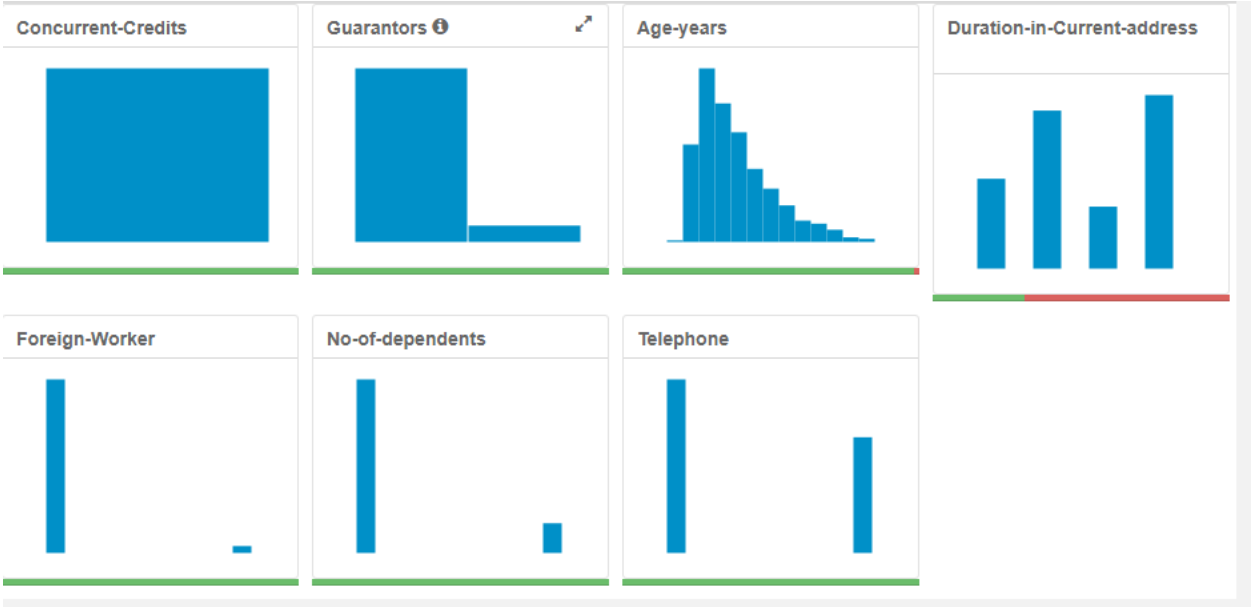


- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

    - Concurrent-Credits
    - Guarantors
    - Occupation
    - Foreign-Worker
    - No-of-dependents
    - Telephone

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.
    - Age: Imputed because only a few records were missing. Used the median because distribution was skewed right (see below)
    - Concurrent-Credits, Guarantors, , Foreign-Worker, No-of-dependents, and Type-of-apartment: Removed because variability was low (see below)
    - Duration in Current Address was removed because a significant amount of data was missing (see below)
    - Telephone is removed because of low relevance to the model



Step 3: Train your Classification Models

Please reference the importance charts for relevant importance for the Decision Tree, Random Forest, and Boosted Models. The Stepwise Logistic model shows that account balance is very significant.

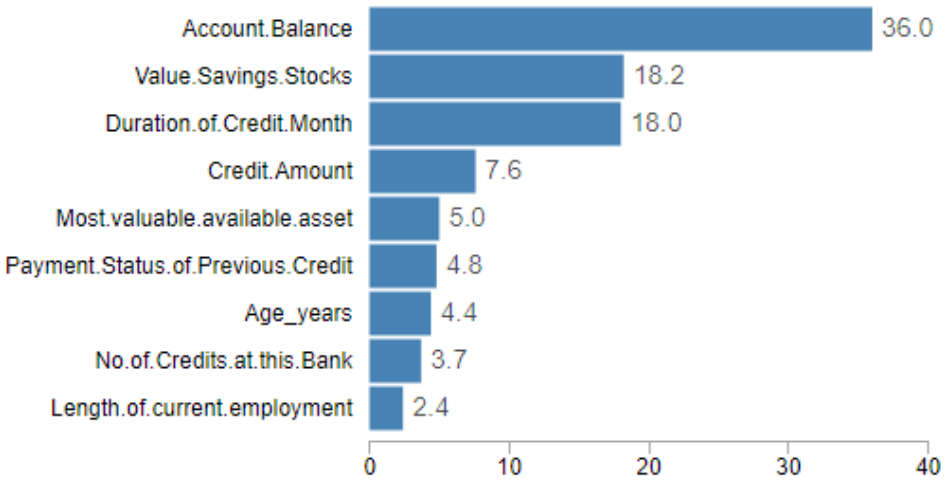
Stepwise Logistic Model

Coefficients:

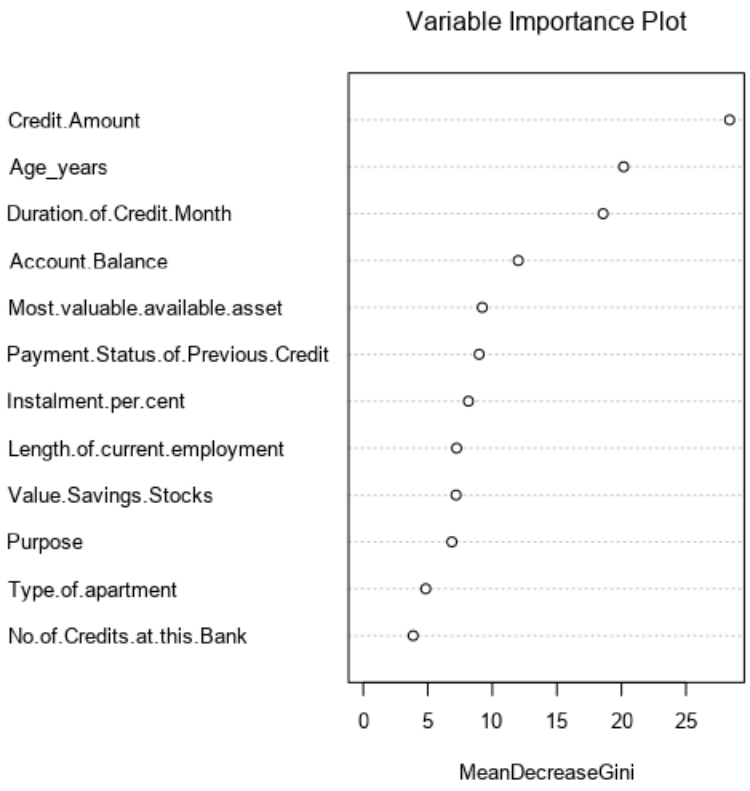
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01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

Decision Tree

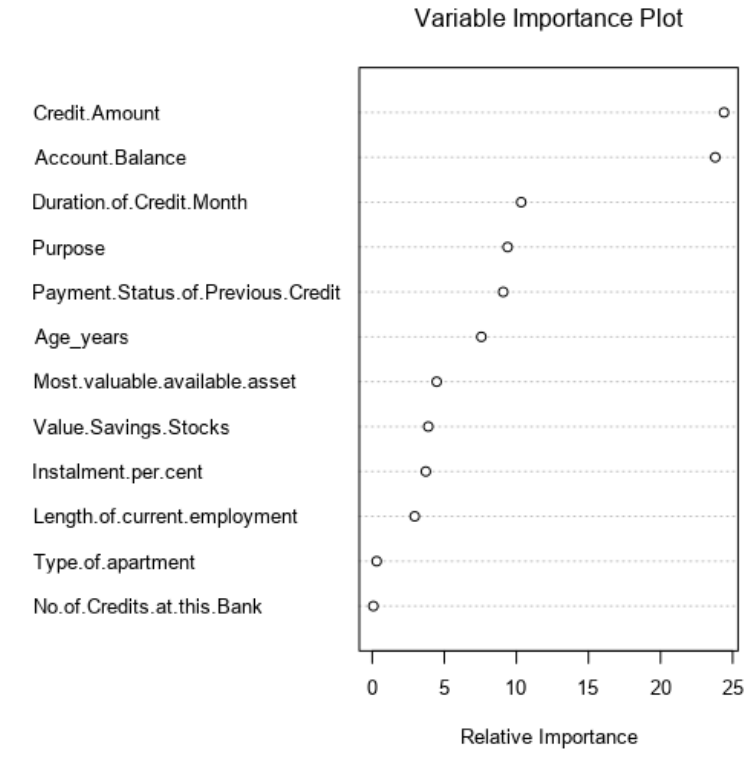
Variable Importance



Random Forest


Variable Importance Plot
Credit.Amount
Age_years
Duration.of.Credit.Month
Account.Balance
Most.valuable.available.asset
Payment.Status.of.Previous.Credit
Instalment.per.cent
Length.of.current.employment
Value.Savings.Stocks
Purpose
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
MeanDecreaseGini

Boosted


Variable Importance Plot
Credit.Amount
Account.Balance
Duration.of.Credit.Month
Purpose
Payment.Status.of.Previous.Credit
Age_years
Most.valuable.available.asset
Value.Savings.Stocks
Instalment.per.cent
Length.of.current.employment
Type.of.apartment
No.of.Credits.at.this.Bank
0
5
10
15
20
25
Relative Importance

The model comparison report shows that the overall accuracy of the random forest model was the highest at 82%. The worst overall performing model was the Decision Tree model. This is likely due to oversampling nature of decision trees. Now when looking at the confusion matrices we can observe the following, we will use "Yes" for creditworthy and "No" for not-creditworthy:

Model Analysis:

- Boosted (Overall accuracy of 79% which is quite strong)
  - This model appears looks to have no bias:
    - PPV: 101 / (28+101) 78%
    - NPV: 17 / (17+4) 81%
- Decision Tree (Overall accuracy of 75% which is strong but the weakest of the bunch)
  - This model appears looks to have bias towards "Creditworthy":
    - PPV: 91 / (91+24) 79%
    - NPV: 21 / (14+21) 60%
- Random Forest (Overall accuracy of 80% which is strongest of all models)
  - This model has no bias
    - PPV: 101 / (101+26) 80%
    - NPV: 19 / (4+19) 83%
- Stepwise-Logistic (Overall accuracy of 76% which is strong)
  - This model has bias towards "Creditworthy"
  - predictions:
    - PPV: 92 / (92+23) 80%
    - NPV: 22 / (13+22) 63%

Confusion matrix of Boosted

	Actual Creditworthy	Actual Non-Creditworthy	Accuracy
Predicted_Creditworthy	101	28	78.29%
Predicted_Non-Creditworthy	4	17	80.95%

Confusion matrix of Decision_Tree

	Actual Creditworthy	Actual Non-Creditworthy	Accuracy
Predicted_Creditworthy	91	24	79.13%
Predicted_Non-Creditworthy	14	21	60.00%

Confusion matrix of Random_Forest

	Actual Creditworthy	Actual Non-Creditworthy	Accuracy
Predicted_Creditworthy	101	26	79.53%
Predicted_Non-Creditworthy	4	19	82.61%

Confusion matrix of Stepwise_Logisitc

	Actual Creditworthy	Actual Non-Creditworthy	Accuracy
Predicted_Creditworthy	92	23	80.00%
Predicted_Non-Creditworthy	13	22	62.86%

See Model Comparison Report (below) for the confusion matrices and overall accuracies.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Random_Forest	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778
Stepwise_Logisitc	0.7600	0.8364	0.7306	0.8762	0.4889

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Random_Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Stepwise_Logisitc

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

Step 4: Writeup

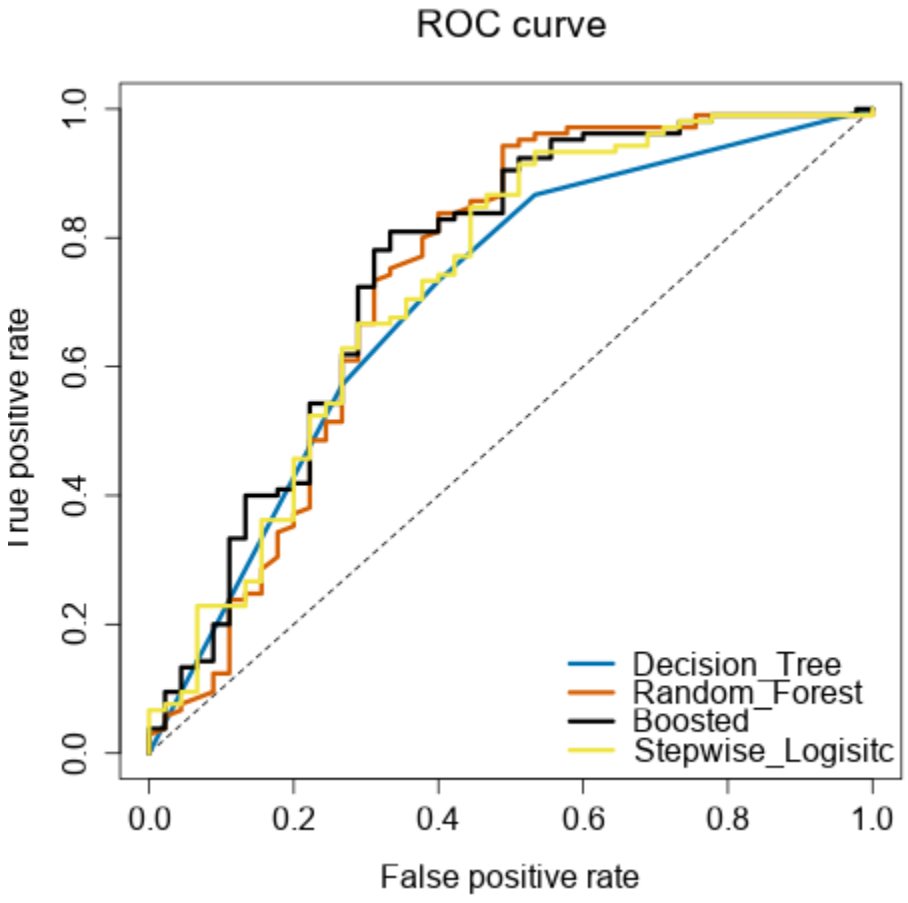
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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:

Model Chosen: **Random Forest**

- Overall Accuracy against your Validation set
  The Random Forest Model had the highest overall accuracy of all 4 models. The accuracy was 80%.
- Accuracies within "Creditworthy" and "Non-Creditworthy" segments
- ROC graph: The Random forest model reaches the top before all other models. The ROC curve at around 0.6 false positive rate shows the random forest model outperforming all over models.



- Bias in the Confusion Matrices

  The accuracies between Creditworth and Non-Creditworth segments were both at or above 80% (80/83). This shows no significant bias that is alarming to the model

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

- How many individuals are creditworthy?
  - 406 individuals are creditworthy