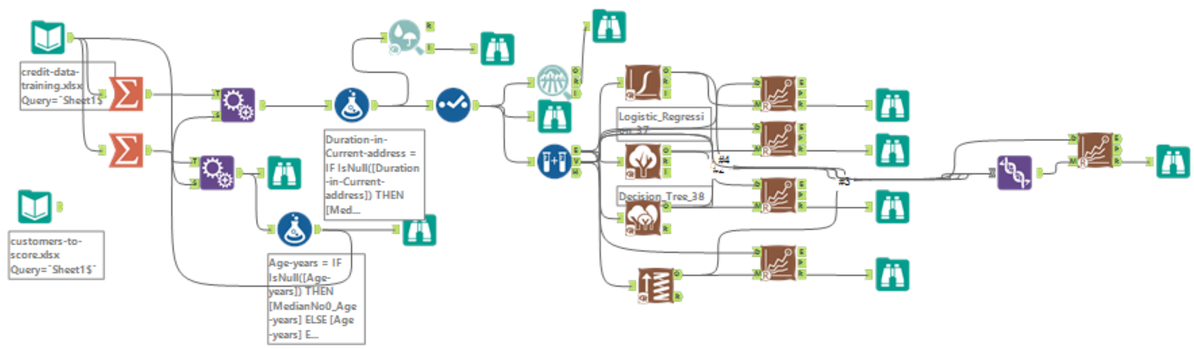
Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

Alteryx Flow



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 objective is to identify whether customers who applied for loan are creditworthy or not.

- What data is needed to inform those decisions?
  The data needed will come from "credit-data-training.xlsx". Data on past applications such as Account Balance and Credit Amount and list of customers to be processed are required in order to inform those decisions

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary classification models such as logistics regression, decision tree, forest model and boosted tree will be used to analyze and determine creditworthy customers

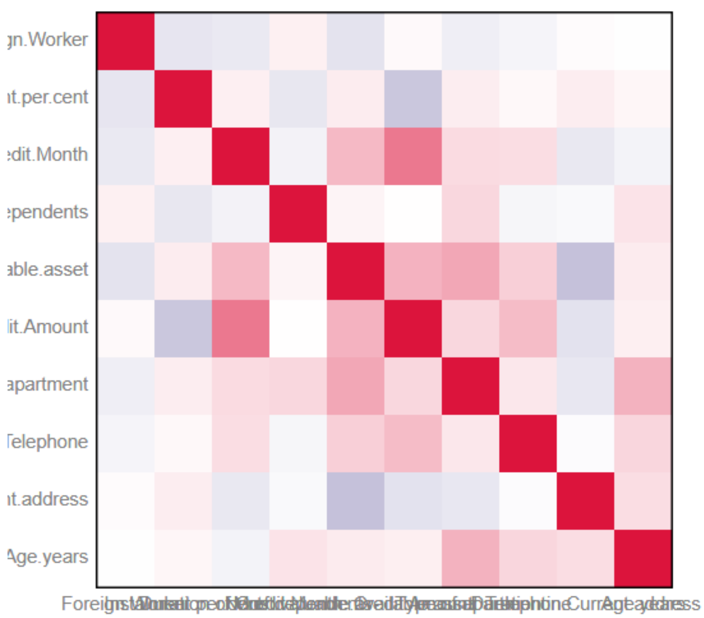
Step 2: Building the Training Set

When summarizing all data fields, Duration in Current Address has 69% missing data and should be removed. While Age Years has 2% missing data, it is appropriate to impute the missing data with the median age. Median age is used instead of mean as the data is skewed to the left as shown below. In addition, Concurrent Credits and Occupation has one value while Guarantors, Foreign Worker and No of Dependents show low variability where more than 80% of the data skewed towards one data. These data should be removed in order not to skew our analysis results.
Below are columns that potentially show low variability due to the majority of its data being one sided

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

Telephone field should also be removed due to its irrelevancy to the customer creditworthy.

An association analysis is performed on the numerical variables and there are no variables which are highly correlated with each other.



Step 3: Train your Classification Models

Using Credit Application Result as the target variables, Account Balance, Purpose and Credit Amount are the top 3 most significant variables with p-value of less than 0.05. In order to create the models, a 70/30 split was done to create an estimation and validation dataset. The models were run and each of the model summaries are below.

Logistic Regression Report :

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

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	.

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

(Dispersion parameter for binomial taken to be 1)

Decision Tree Report :

Model Summary

Variables actually used in tree construction:

[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks

Root node error: 97/350 = 0.27714

n= 350

Pruning Table

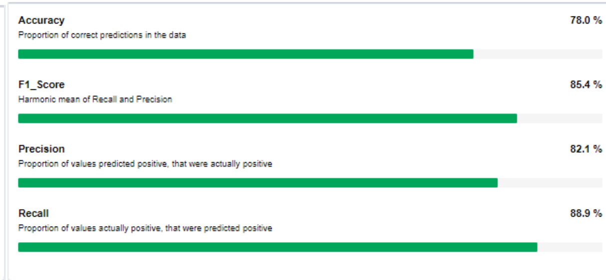
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

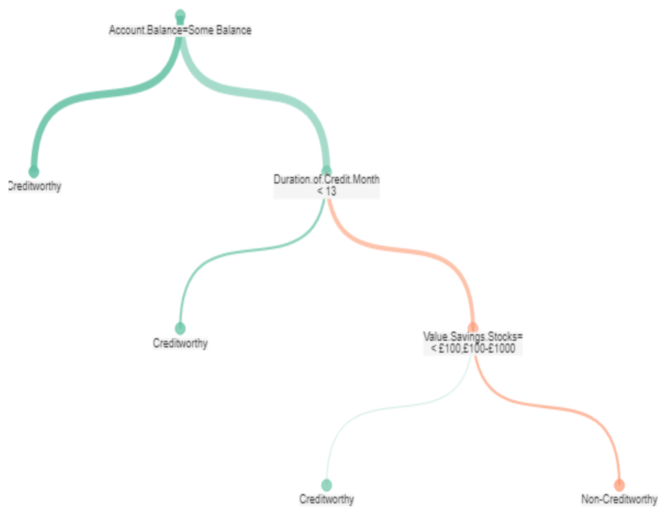
Leaf Summary

```
node), split, n, loss, yval, (yprob)
    * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
  2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
  3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
    6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
    7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
     14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
     15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *
```

Tree and Confusion Matrix of Decision Tree

Actual	Actual Positive	Actual Negative
Predicted Positive	48 (49.5%)	49 (50.5%)
Predicted Negative	28 (11.1%)	225 (88.9%)





Important Features
Account Balance
Duration of Credit Month
Value Saving Stock

Above clearly shows that tree is build on just 3 features.

Forest Tree Report :

Basic Summary

Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 25, replace = TRUE)

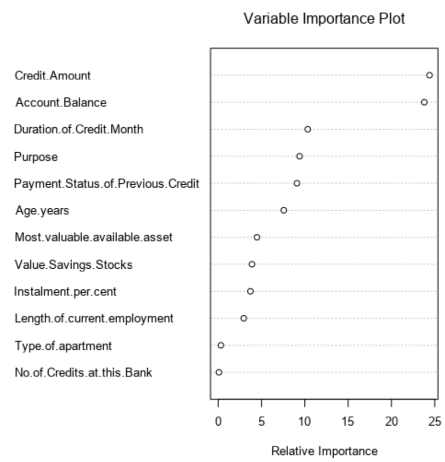
Type of forest: classification
Number of trees: 25
Number of variables tried at each split: 3

OOB estimate of the error rate: 24.6%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.099	228	25
Non-Creditworthy	0.629	61	36

Below is the variable importance chart for the forest tree model.

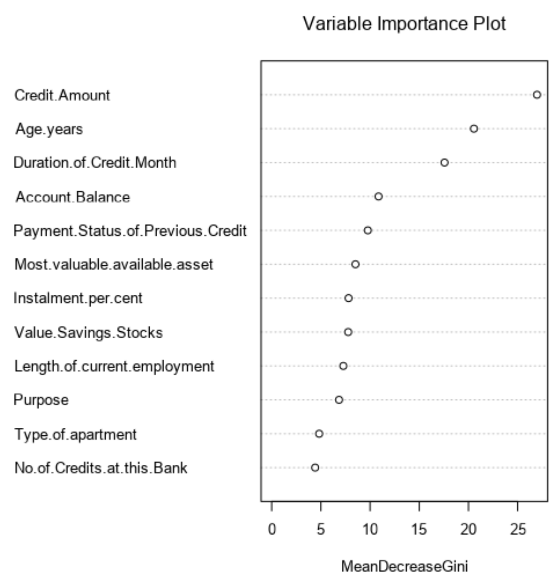


Boosted Tree Report:

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 7-fold cross validation: 2628

Below is the variable importance chart for the boosted tree model.



Below are the model comparison between results, accuracy and roc. Also confusion matrix of each of the model is shown in comparison report.

Step 4: Writeup

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_38	0.7467	0.8273	0.7054	0.8667	0.4667
Boosting1	0.7867	0.8632	0.7524	0.9619	0.3778
Logistic_Regression_37	0.7800	0.8520	0.7314	0.9048	0.4889
Forest_Tree	0.7733	0.8522	0.7200	0.9333	0.4000

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 Boosting1

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

Confusion matrix of Decision_Tree_38

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

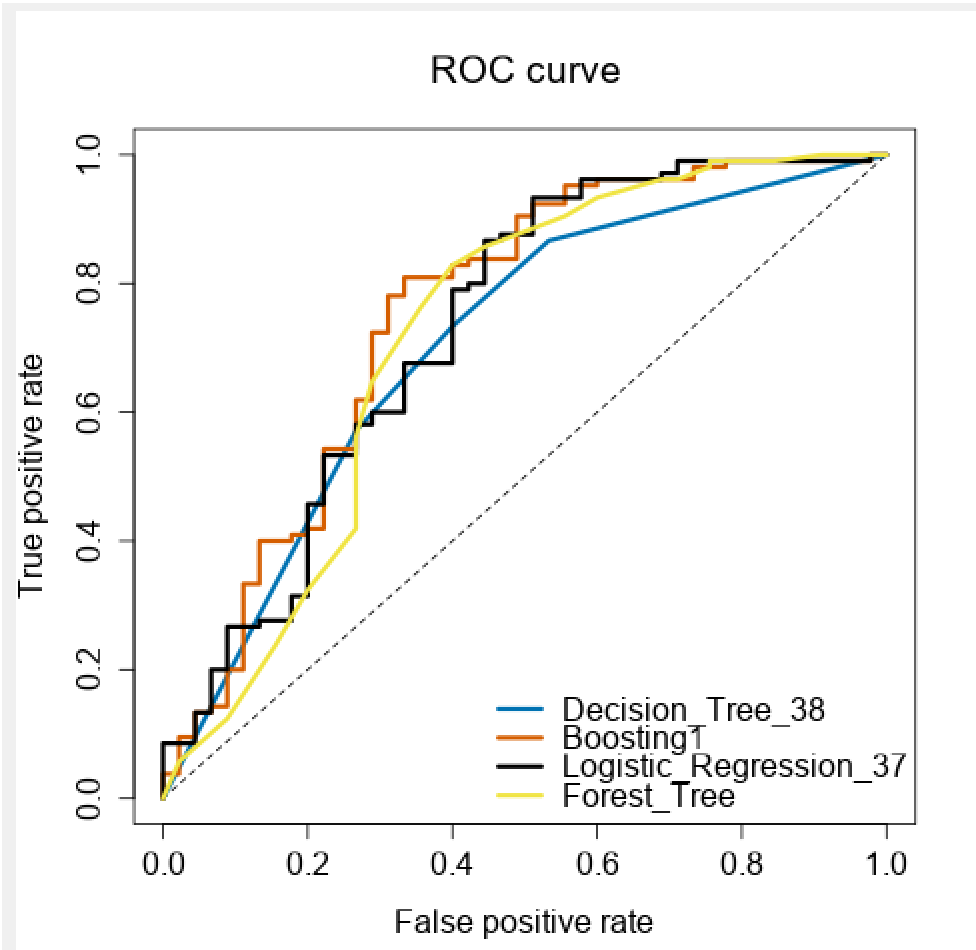
Confusion matrix of Forest_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	98	27
Predicted_Non-Creditworthy	7	18

Confusion matrix of Logistic_Regression_37

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

The final model used for prediction will be the **Logistic Regression** It does not have a high accuracy, 0.7314, but have high score for predicting Creditworthy applicants and also high score for predicting non-creditworthy applicants Below is the ROC chart for the models.



The ROC plots shows the LR model to be the second best with an AUC of 0.7314.

Applying the model to the new dataset, customers-to-score.xls and taking any applicant that has a greater Creditworthy accuracy score than non-creditworthy to mean the applicant should be granted a loan, the final count of individuals whom are creditworthy are 401.