Predictive Analytics for Business Nanodegree

Project: Predicting Default Risk

Laila Hussain Alqawain

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

   The decision to need to be made is that we should decide if a person who wants a loan is creditworthy or not creditworthy.

2. What data is needed to inform those decisions?

   We need to know the information about the people who applied for a loan before like their age, the reason which makes them take a loan, and how long have they employed.

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

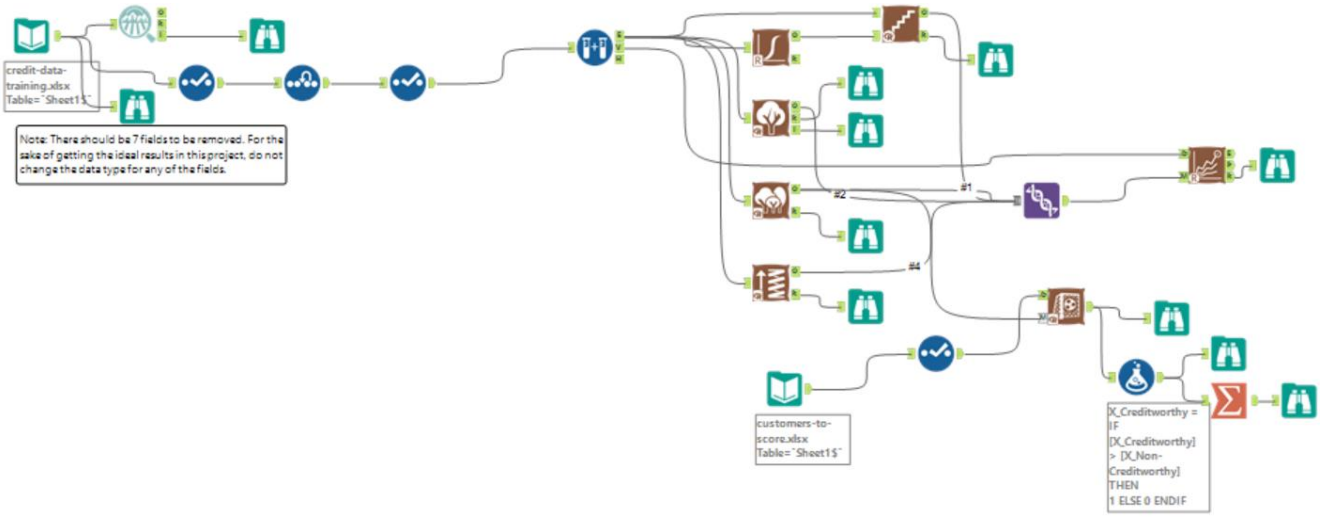
   As we want to decide if the person who applied for the loan is creditworthy or not creditworthy, so the model is binary due to there are two options.

Step 2: Building the Training Set

Let's see how the data looks like after I visualized the data by the field summary tool.

As we see above, Concurrent-Credits, Guarantors, Occupation, No-of-dependents, and Foreign-Worker are low variability. Duration-in-Current-address has many missing data and the Telephone is not relevant data. So, I removed these fields. Age-years has a few missing data, so I imputed by the median to ensure that is not affected by outliers.

Step 3: Train your Classification Models



As we see the workflow above, I built the required models: Logistic Regression, Decision Tree, Forest Model, and Boosted Tree.

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

Logistic Regression Model

Report for Logistic Regression Model LRM

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	.

Decision Tree Model

Summary Report for Decision Tree Model DTM

Call:
rpart(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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

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

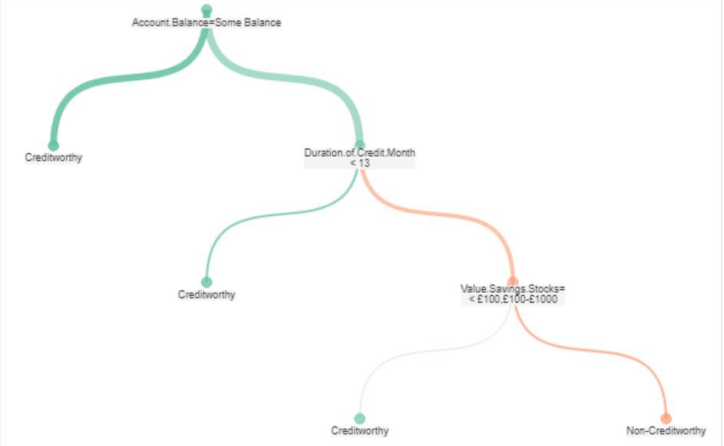
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

Leaf Summary

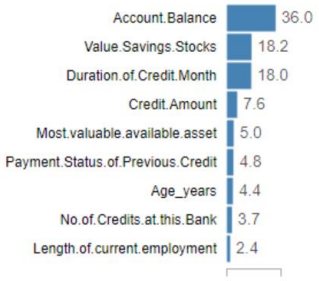
node), split, n, loss, yval, (yprob)
* denotes terminal node

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) *

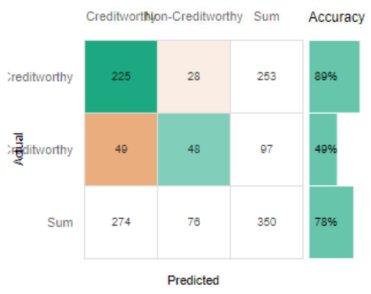
Decision Tree



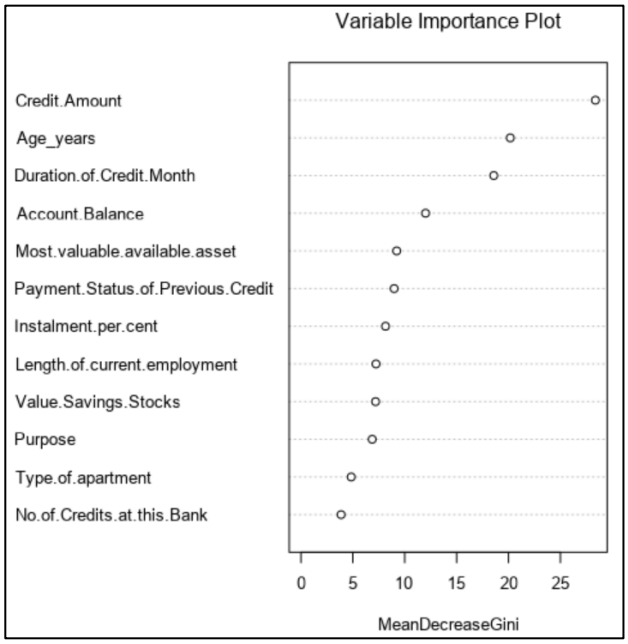
Variable Importance



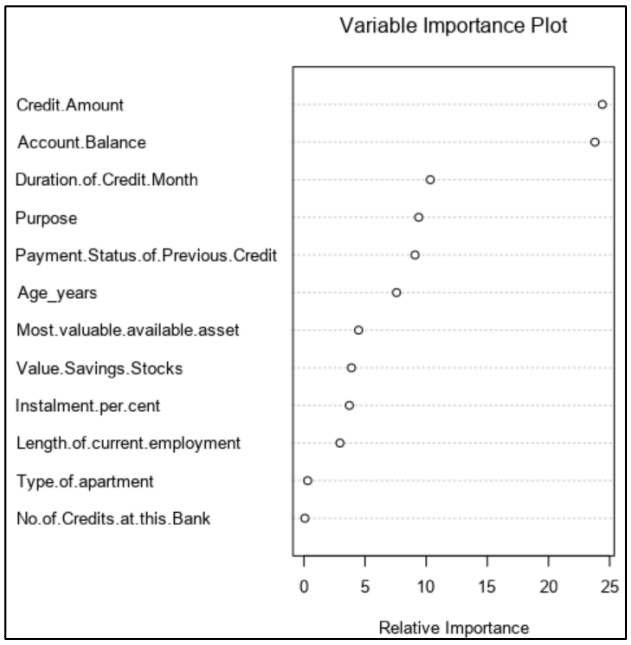
Confusion Matrix



Forest Model



Boosted Tree Model



From the report of models above, the significant of predictor variables for Logistic Regression Model is Account.BalanceSome Balance, Decision Tree Model is Account.Balance, Forest Model is Credit-Amount, and Boosted Tree Model is Credit-Amount.

2. 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?

Model	Accuracy
DTM	0.7467
FM	0.8000
BM	0.7867
LRM	0.7600

Confusion matrix of BM

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

Confusion matrix of DTM

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

Confusion matrix of FM

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

Confusion matrix of LRM

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

- There is bias in Boosted Tree Model prediction. It predicted the highest number of persons who are 28 persons as creditworthy and actually are non-creditworthy.
- Decision Tree Model predicted the highest number of persons who are 14 persons as non-creditworthy and are creditworthy.
- Forest Model predicted 26 persons as creditworthy and actually are non-creditworthy.
- Logistic Regression Model predicted 13 persons as non-creditworthy and actually are creditworthy. Also, it predicted 23 persons as creditworthy and actually are non-creditworthy.

Step 4: Writeup

1. Which model did you choose to use?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DTM	0.7467	0.8273	0.7054	0.8667	0.4667
FM	0.8000	0.8707	0.7361	0.9619	0.4222
BM	0.7867	0.8632	0.7524	0.9619	0.3778
LRM	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

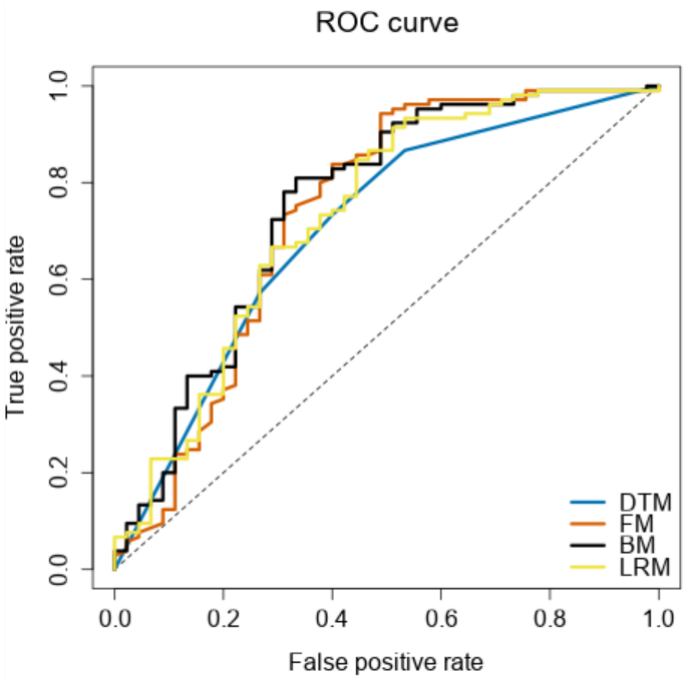
- Overall Accuracy against your Validation set

  From Model Comparison Report, the best model is the Forest Model as we see the highest number of accuracy which is 0.80 for the Forest Model.

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments

  Forest Model and Boosted Tree Model have the highest number of Accuracy_Creditworthy which is 0.96, but for Accuracy_Non-Creditworthy, Boosted Model has the lowest value which is 0.37.
  The highest number of Accuracy_Non-Creditworthy is 0.48 for Logistic Regression Model.

- ROC graph



From ROC curve plot, the Decision Tree Model is the worst model due to it is the nearest to the diagonal line. The Forest Model is the farthest curve from the diagonal line, so it is the best model.

- Bias in the Confusion Matrices

From the Confusion matrix of Models and which I explained before, I can say the Forest Model is the best model for predicting creditworthy comparing with remaining models.

2. How many individuals are creditworthy?

Record	Sum_X_Creditworthy
1	406

The Forest Model predicted 406 persons out of 500 persons are creditworthy.