PREDICTING DEFAULT RISK

Classification Models

TABLE OF CONTENTS

PREDICTING DEFAULT RISK

Business and Data Understanding

The analysis is to determine if customers are creditworthy for a new loan. The team typically gets 200 loan applications per week and approves them all by hand.

Due to sudden increase, there are nearly 500 loan applications to process this week due to a financial scandal that hit a competitive bank last week. Your manager sees new influx is a great opportunity to figure out to process all these loan applications in one week. I need to systematically evaluate the creditworthiness of these new loan applications.

Key Decisions:

To be able to make an informed decision if customers are creditworthy, I need to have the following:

- Data on past loan applicants
- List of customers that need to process.

We have been provided two data sets:

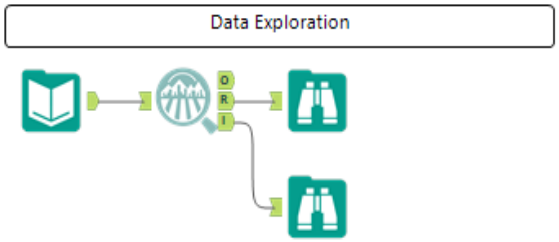
- Credit-data-training contains approvals from the past loan applicants the bank has processed.
- Customers-to-score contains the new customers that I need to score on the classification model.

I will create a Binary classification model, to help make the decision if the customers are creditworthy or not.

Building the Training Set

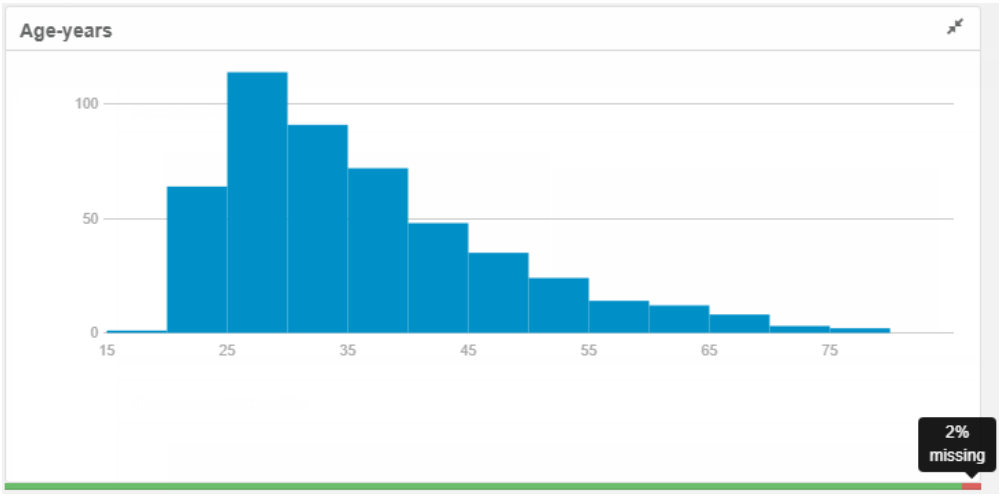
First, I need to explore the data to see if any columns that should be removed or imputed.

Using Alteryx to explore the data using the field comparison tool.

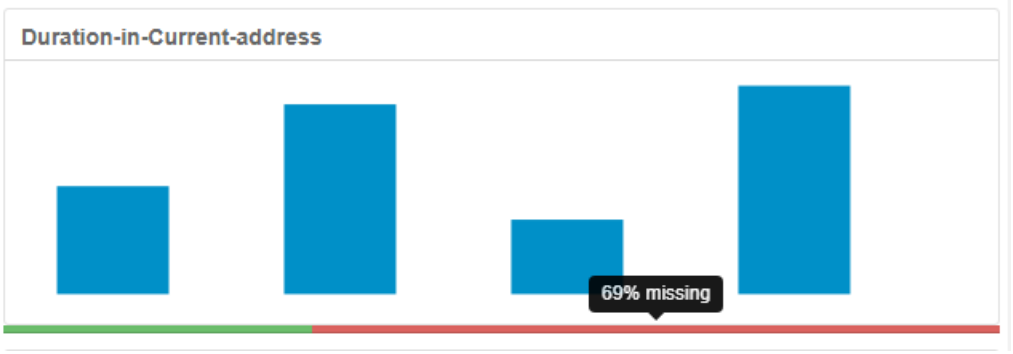


Missing Data

- Age-years 2.4% missing.
    - Impute the miss data using the median age of 33 from the entire field.



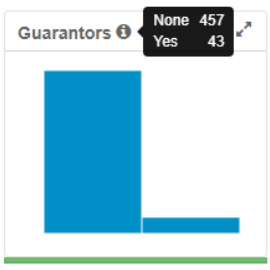
- Duration-in-Current-address 68.8% missing.
    - This field will be removed, as there is a lot of missing values.

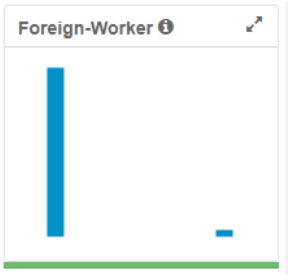


Clean-up

There are 3 fields that are heavily skewed towards toward one type of data, these should be removed.

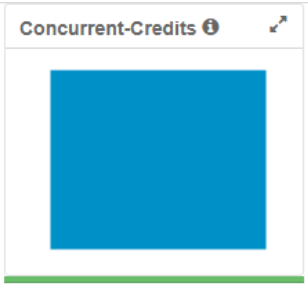
- Guarantors: is heavily towards ‘None’
- Foreig-Worker: heavily towards ‘1’
- No-of-dependents: heavily towards ’1’

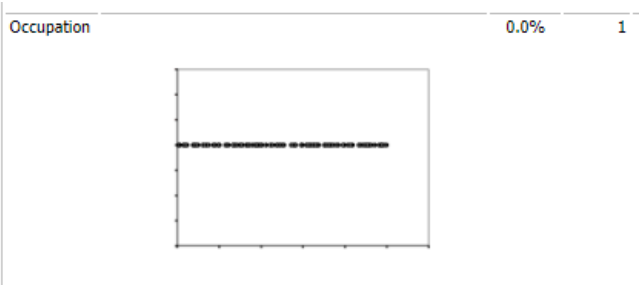




2 fields have low variability, no variations and is entirely uniform, there is no variation and should be removed.

- Concurrent-Credits, has only 1 unique value.
- Occupation also contains only 1 unique value.





The last field that would have no significance for the model would be Telephone, this can also be removed.

Field Summary

Name	Field Category	Percent Missing	Unique Values
Age-years	Numeric	2.4	54
Credit-Amount	Numeric	0	464
Duration-in-Current-address	Numeric	68.8	5
Duration-of-Credit-Month	Numeric	0	30
Foreign-Worker	Numeric	0	2
Instalment-per-cent	Numeric	0	4
Most-valuable-available-asset	Numeric	0	4
No-of-dependents	Numeric	0	2
Occupation	Numeric	0	1
Telephone	Numeric	0	2
Type-of-apartment	Numeric	0	3
Account-Balance	String	0	2
Concurrent-Credits	String	0	1
Credit-Application-Result	String	0	2
Guarantors	String	0	2
Length-of-current-employment	String	0	3
No-of-Credits-at-this-Bank	String	0	2
Payment-Status-of-Previous-Credit	String	0	3
Purpose	String	0	4
Value-Savings-Stocks	String	0	3

Result after imputing Age-years.

Name	Percent Missing	Mean
Age-years	0	35.574

Age-years validation 35.574 rounded up = 36

Train Classification Models

Creating the estimation and validation samples where 70% of the data set will be used for estimation and the remaining 30% is reserved for validation.

4 Models will be created: Logistic Regression, Decision Tree, Forest Model, Boosted Model.

Stepwise will be used, to improve the efficiency as this tool will automate finding the best predictor variables.

Target Variable in all models is Credit Application Result

Logistic Regression: Stepwise

The top 3 predictor variables that statistically significant with p-values of less that = 0.05.

Variable	p-value
Account balance	-2.41e-08
Purpose	0.00665
Credit amount	0.00167

For full report Reference figure 1

LR Model accuracy & Confusion Matrix

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
dt_model	0.7467	0.8304	0.7035	0.8857	0.4222
fm_model	0.7933	0.8681	0.7368	0.9714	0.3778
boost_model	0.7867	0.8632	0.7515	0.9619	0.3778
lr_model	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of lr_model

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

- Credit accuracy = actual_creditworthy / (predicted_creditworthy)

  = 92 / (92+23) = 0.8 or = 80%

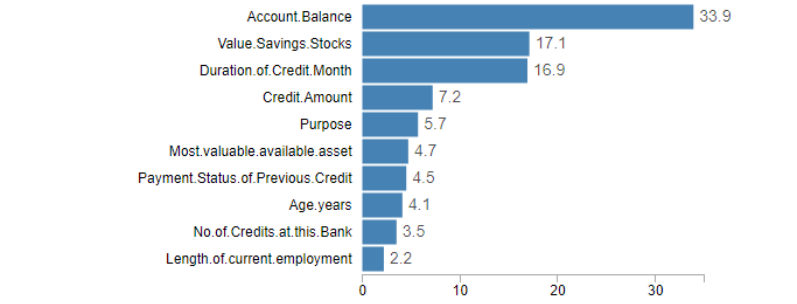
- Non-Credit accuracy = actual_Non-creditworthy / (predicted_Non-creditworthy)

  = 22 / (13+22) = 0.6286 or = 62.86%

The linear regression model shows bias in predicting customers as non-creditworthy

Decision Tree

The top 3 predictor variables that statistically significant are:



- Account Balance
- Vale Savings Stocks
- Duration of Credit Month

Full report details [Reference figure 2]

DT Model Accuracy & Confusion Matrix

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
dt_model	0.7467	0.8304	0.7035	0.8857	0.4222
fm_model	0.7933	0.8681	0.7368	0.9714	0.3778
boost_model	0.7867	0.8632	0.7515	0.9619	0.3778
lr_model	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of dt_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

- Credit accuracy = actual_creditworthy / (predicted_creditworthy)

  = 93 / (93+26) = 0.7815 or = 78.15%

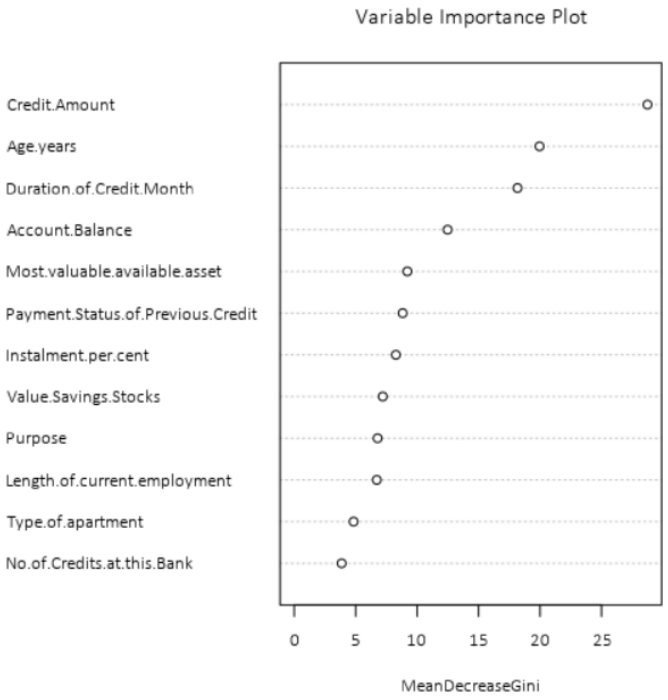
- Non-Credit accuracy = actual_Non-creditworthy / (predicted_Non-creditworthy)

  = 19 / (12+19) = 0.6129 or = 61.29%

Similar to the linear regression model, the Decision Tree shows bias in predicting customers as non-creditworthy.

Random Forest Model

The top 3 predictor variables that statistically significant are:



- Credit Amount
- Age-Years
- Duration of Credit Month

Full report details [Reference figure 3]

RF Model Accuracy & Confusion Matrix

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
dt_model	0.7467	0.8304	0.7035	0.8857	0.4222
fm_model	0.7933	0.8681	0.7368	0.9714	0.3778
boost_model	0.7867	0.8632	0.7515	0.9619	0.3778
lr_model	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of fm_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

- Credit accuracy = actual_creditworthy / (predicted_creditworthy)

  = 102 / (102+28) = 0.7846 or = 78.46%

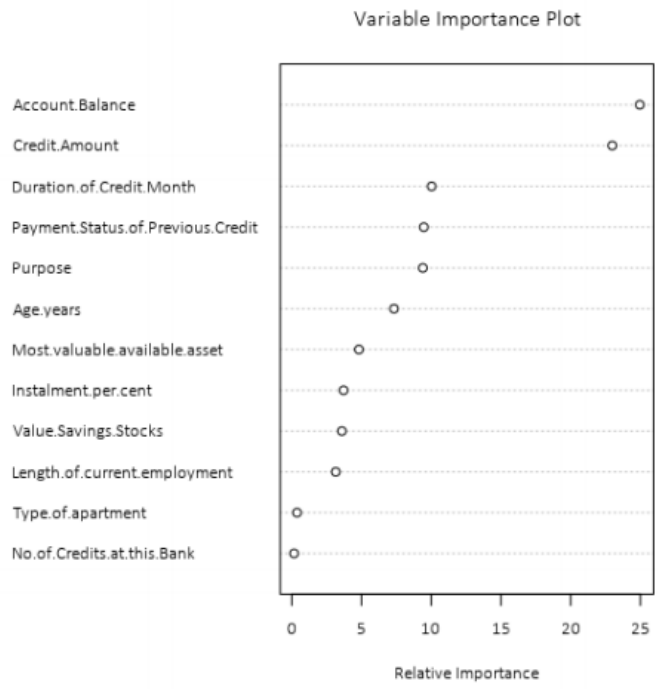
- Non-Credit accuracy = actual_Non-creditworthy / (predicted_Non-creditworthy)

  = 17 / (3+17) = 0.85 or = 85.00%

The Random forest does not seem to build biased as the accuracies are quite close to each other.

Boosted Model

The top 3 predictor variables that statistically significant are:



- Account Balance
- Credit Ammount
- Duration of Credit Month

Full report details [Reference figure 4]

Boosted Model Accuracy & Confusion Matrix

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
dt_model	0.7467	0.8304	0.7035	0.8857	0.4222
fm_model	0.7933	0.8681	0.7368	0.9714	0.3778
boost_model	0.7867	0.8632	0.7515	0.9619	0.3778
lr_model	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of boost_model

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

- Credit accuracy = actual_creditworthy / (predicted_creditworthy)

  = 102 / (102+28) = 0.7829 or = 78.29%

- Non-Credit accuracy = actual_Non-creditworthy / (predicted_Non-creditworthy)

  = 17 / (3+17) = 0.8095 or = 80.95%

The Boosted Model does not seem to be biased as the accuracies are quite close to each other.

Writeup

The Forest model has the highest accuracy of 79.33% from all the four models and has the highest F1 score of 86.81%. However, the AUC is lower than the next highest, the Boosted Model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
dt_model	0.7467	0.8304	0.7035	0.8857	0.4222
fm_model	0.7933	0.8681	0.7368	0.9714	0.3778
boost_model	0.7867	0.8632	0.7515	0.9619	0.3778
lr_model	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.

The Linear Regression (Stepwise) and Decision Trees seem to create bias towards classifying customers as Non-Creditworthy whereas the Forest and Boosted Models seems to be comparable in classify credit and Non-Creditworthy customers.

Confusion matrix of boost_model

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

Confusion matrix of dt_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of fm_model

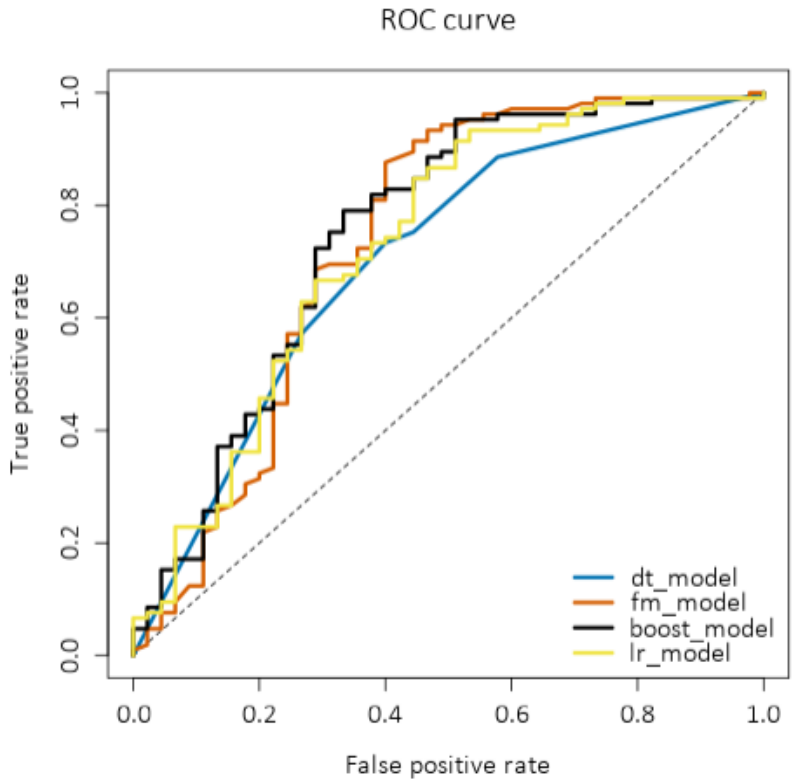
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of lr_model

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

Model Accuracies,

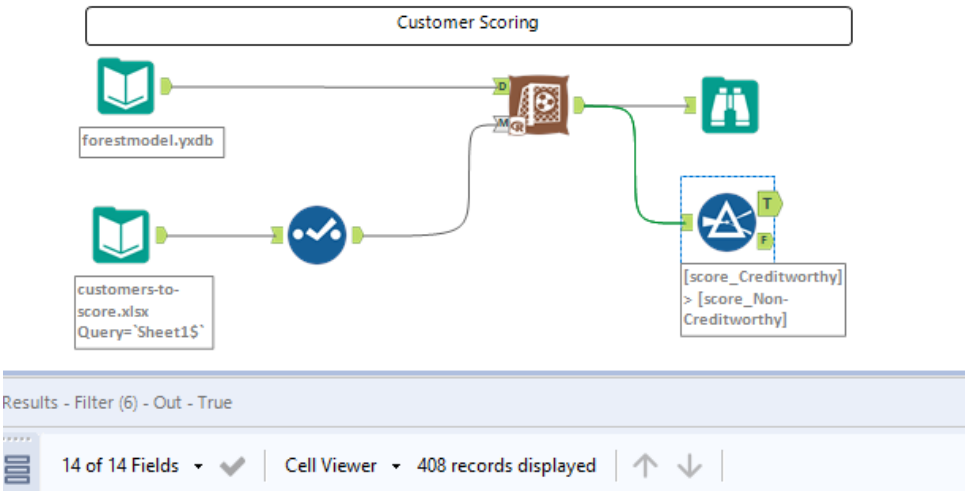
Model	Credit Worthy	Non-Credit Worthy
Logistic Regression	80.00%	62.86%
Decision Tree	78.15%	61.29%
Random Forest	78.46%	85.00%
Boosted	78.29%	80.95%



Looking at the ROC curve, the Decision Tree performed the worst, whereas the Boosted and Forest model performed the best, these two models also reaches the true positive rate the fasted with the Forest model slightly leading.

The Forest model will be chosen to predict the classification for credit and Non-Credit worthy of the new 500 loan applicants.

From the new loan applications 408 (98%) are credit worthy and should be approved for a new loan, 92 have been classified as Non-Credit worthy.



Model Reports

Linear Regression, Stepwise report

Figure 1

Report for Logistic Regression Model lr_model

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)

Null deviance: 413.16 on 349 degrees of freedom

Residual deviance: 328.55 on 338 degrees of freedom

McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)
Account.Balance	31.129	1	2.41e-08***
Payment.Status.of.Previous.Credit	5.687	2	0.05823.
Purpose	12.225	3	0.00665**
Credit.Amount	9.882	1	0.00167**
Length.of.current.employment	5.522	2	0.06324.
Instalment.per.cent	5.198	1	0.02261*
Most.valuable.available.asset	3.509	1	0.06104.

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

Decision Tree

Figure 2

Summary Report for Decision Tree Model dt_model

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 + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, 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 Purpose

[4] 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	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032

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=< Â£100,Â£100-Â£1000 34 11 Creditworthy (0.6764706 0.3235294) *

15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789)

30) Purpose=New car 8 2 Creditworthy (0.7500000 0.2500000) *

31) Purpose=Home Related,Other,Used car 68 22 Non-Creditworthy (0.3235294 0.6764706) *

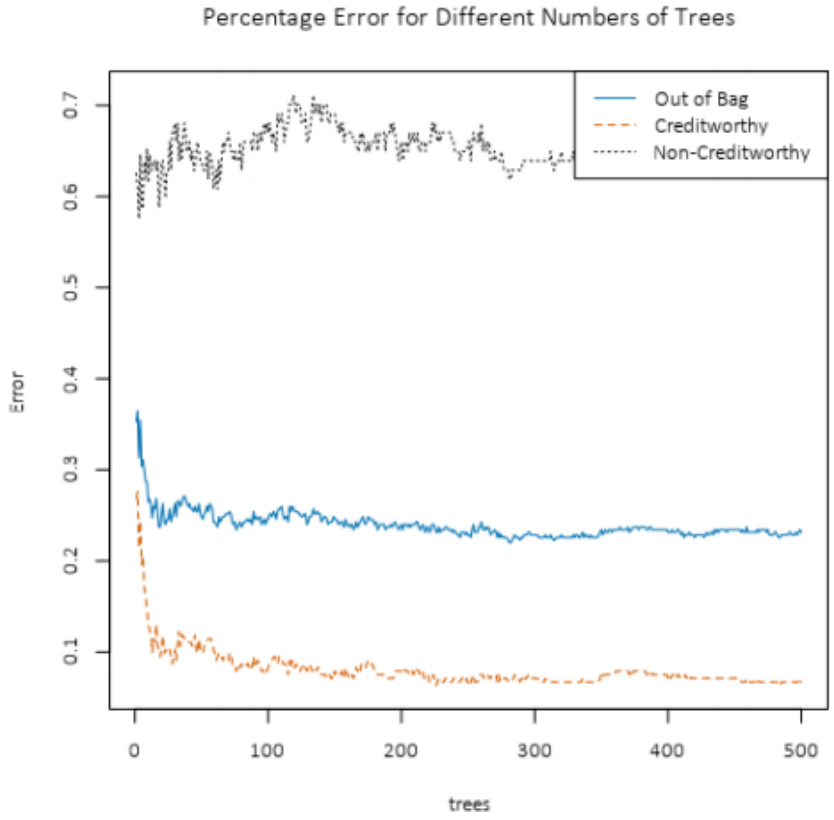
Random Forest

Figure 3

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 = 500, replace = TRUE)
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 23.1%
Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.067	236	17
Non-Creditworthy	0.66	64	33

Plots

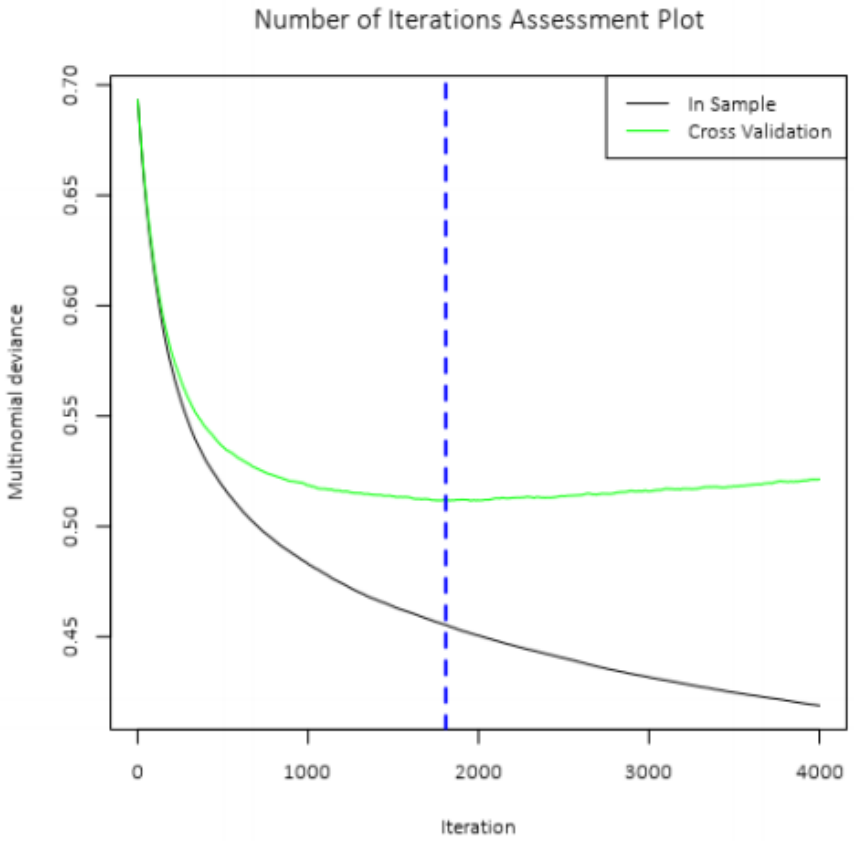


Boosted Model

Figure 4

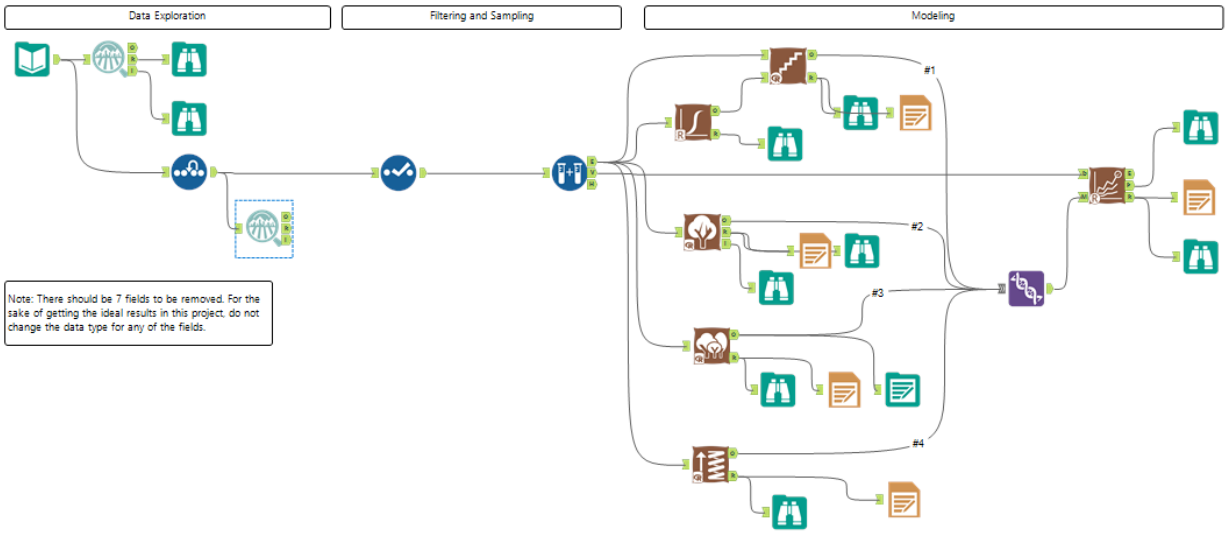
Basic Summary:

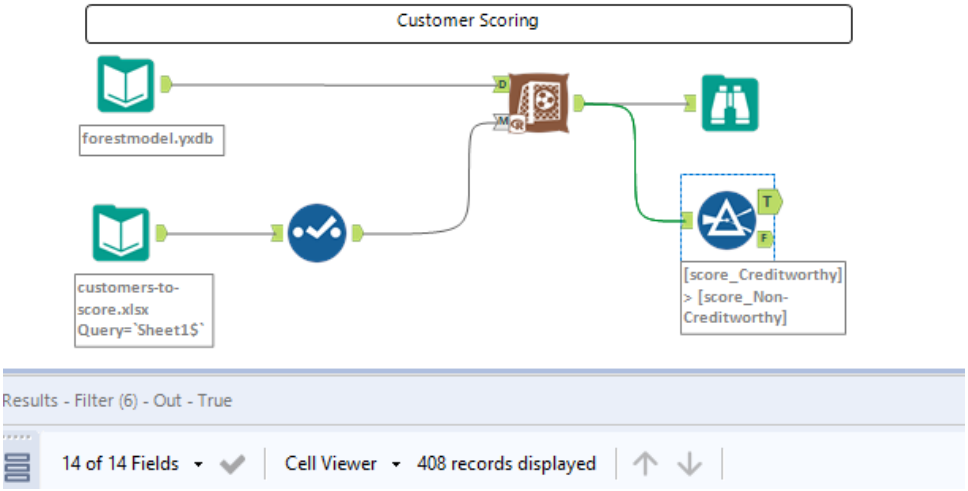
Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1808



The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).

Alteryx Workflows





Resources

Here are some guidelines to help you clean up the data:

- Are any of your numerical data fields highly correlated 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: If you decide to impute any data field, for the sake of consistency in the data clean-up process, impute the data using the median of the entire data field.

Websites

https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Udacity Knowledge

https://knowledge.udacity.com/questions/483519
https://knowledge.udacity.com/questions/462675
https://knowledge.udacity.com/questions/473072