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?

To classify new customer on whether they can be approved for a loan or not and provide a list of creditworthy customers.

- What data is needed to inform those decisions?

Credit application result, account balance, payment of status of previous credit, purpose, number of credit at this bank, value saving stocks, length of current employment, guarantors, most valuable available asset, age in years, type of appartment, credit amount, duration of credit in month, telephone from past application data and list of customers who have applied to get a loan

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

Binary

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.***

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".
- 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.

Based on Field Summary module results:
- Duration-in-Current-address field is removed due to 68.8% missing value.
- Age-years field is imputed using median as suggested due to 2.4% missing value.
- Occupation, Guarantors, and Concurrent-Credits is removed due to low variability.
Based on Association Analysis module results, low correlation fields are removed, i.e. Foreign-Worker, Telephone and Number-of-Dependents.

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value	
Duration.of.Credit.Month	-0.2025036	5.0151e-06	***
Credit.Amount	-0.2019458	5.3311e-06	***
Most.valuable.available.asset	-0.1413324	1.5334e-03	**
Instalment.per.cent	-0.0621068	1.6556e-01	
Age.years	0.0529139	2.3758e-01	
No.of.dependents	-0.0410479	3.5969e-01	
Telephone	-0.0289707	5.1807e-01	
Type.of.apartment	-0.0265155	5.5417e-01	
Foreign.Worker	0.0091861	8.3765e-01	

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

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

Predictor variables are sorted from the most significant or important.
Logistic Regression variables: Account Balance, Purpose, Payment Status of Previous Credit, Most Valuable Available Asset, and can be added Length of Current Employment, and Credit Amount based on stepwise.
Decision Tree variables: Account Balance, Duration of Credit Month,Value Saving Stocks and Purpose.
Forest Model variables: Credit Amount, Age Years, Duration of Credit Month and Account Balance if Mean Decrease Gini > 10 considered.
Boosted Model variables: Account Balance and Credit Amount if Relative Importance > 10 considered.

- 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?

Logistic Regression: 0.76. If PPV = 93/(93+24) = 0.795 and NPV = 21/(21+12) = 0.636, then there is any bias.
Decision Tree: 0.7467. If PPV = 93/(93+26) = 0.782 and NPV = 19/(19+12) = 0.613, then there is any bias.
Forest Model: 0.7867. If PPV = 99/(99+26) = 0.792 and NPV = 19/(19+6) = 0.78, then there is no bias.
Boosted Model: 0.7867. If PPV = 101/(101+28) = 0.783 and NPV = 17/(17+4) = 0.81, then there is no bias.

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

Report for Logistic Regression Model Logistic_Credit

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Duration.of.Credit.Month + Credit.Amount + Type.of.apartment + Age.years + Most.valuable.available.asset + Length.of.current.employment + Value.Savings.Stocks + No.of.Credits.at.this.Bank + Purpose + Payment.Status.of.Previous.Credit + Account.Balance, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.032	-0.735	-0.449	0.694	2.490

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.1840436	9.247e-01	-2.3619	0.01818	*
Duration.of.Credit.Month	0.0118887	1.329e-02	0.8943	0.37118	
Credit.Amount	0.0001135	6.001e-05	1.8914	0.05857	.
Type.of.apartment	-0.1861682	2.918e-01	-0.6381	0.52342	
Age.years	-0.0147558	1.522e-02	-0.9696	0.33222	
Most.valuable.available.asset	0.3607760	1.553e-01	2.3234	0.02016	*
Length.of.current.employment4-7 yrs	0.5794020	4.880e-01	1.1873	0.23511	
Length.of.current.employment< 1yr	0.7601757	3.931e-01	1.9337	0.05315	.
Value.Savings.StocksNone	0.6066625	5.073e-01	1.1958	0.23176	
Value.Savings.Stocks£100-£1000	0.1739142	5.616e-01	0.3097	0.75681	
No.of.Credits.at.this.BankMore than 1	0.3510908	3.790e-01	0.9264	0.35423	
PurposeNew car	-1.7860902	6.242e-01	-2.8613	0.00422	**
PurposeOther	-0.1956457	8.033e-01	-0.2436	0.80757	
PurposeUsed car	-0.8647570	4.051e-01	-2.1345	0.0328	*
Payment.Status.of.Previous.CreditPaid Up	0.3956520	3.813e-01	1.0377	0.29941	
Payment.Status.of.Previous.CreditSome Problems	1.2200584	5.345e-01	2.2826	0.02246	*
Account.BalanceSome Balance	-1.5219305	3.193e-01	-4.7657	1.88e-06	***

Report for Logistic Regression Model Log_Step_Credit

Basic Summary

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

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.228	-0.732	-0.461	0.723	2.358

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.0056725	5.160e-01	-3.8868	1e-04	***
Credit.Amount	0.0001221	5.176e-05	2.3593	0.01831	*
Most.valuable.available.asset	0.3220738	1.403e-01	2.2961	0.02167	*
Length.of.current.employment4-7 yrs	0.3560956	4.541e-01	0.7843	0.43289	
Length.of.current.employment< 1yr	0.7784357	3.844e-01	2.0251	0.04286	*
PurposeNew car	-1.7443937	6.107e-01	-2.8562	0.00429	**
PurposeOther	-0.2284958	7.914e-01	-0.2887	0.77279	
PurposeUsed car	-0.8513778	3.933e-01	-2.1650	0.03039	*
Payment.Status.of.Previous.CreditPaid Up	0.2085991	2.940e-01	0.7096	0.47796	
Payment.Status.of.Previous.CreditSome Problems	1.1662580	5.119e-01	2.2783	0.02271	*
Account.BalanceSome Balance	-1.5789854	3.040e-01	-5.1949	2.04e-07	***

Summary Report for Decision Tree Model Decision_Tree_Credit

Call:
rpart(formula = Credit.Application.Result ~ Duration.of.Credit.Month + Credit.Amount + Type.of.apartment + Age.years + Most.valuable.available.asset + Length.of.current.employment + Value.Savings.Stocks + No.of.Credits.at.this.Bank + Purpose + Payment.Status.of.Previous.Credit + Account.Balance, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary

Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose 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	0.87629	0.082704

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)
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) *

Basic Summary

Call:
randomForest(formula = Credit.Application.Result ~ Duration.of.Credit.Month + Credit.Amount + Age.years + Most.valuable.available.asset + Length.of.current.employment + Value.Savings.Stocks + No.of.Credits.at.this.Bank + Purpose + Payment.Status.of.Previous.Credit + Account.Balance, data = the.data, ntree = 500, replace = TRUE)

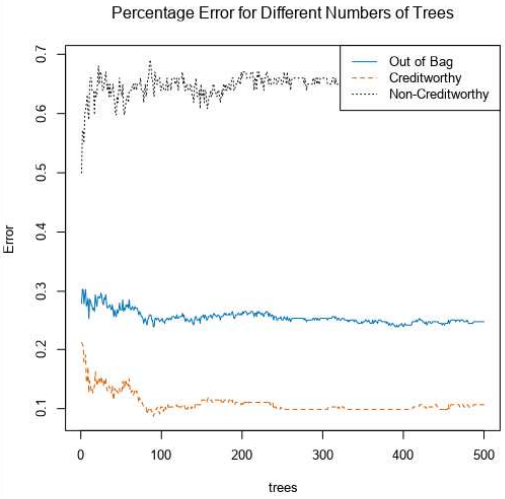
Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

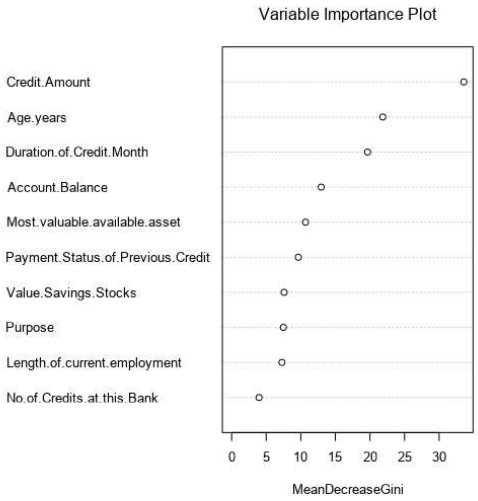
OOB estimate of the error rate: 24.9%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.107	226	27
Non-Creditworthy	0.619	60	37

Plots



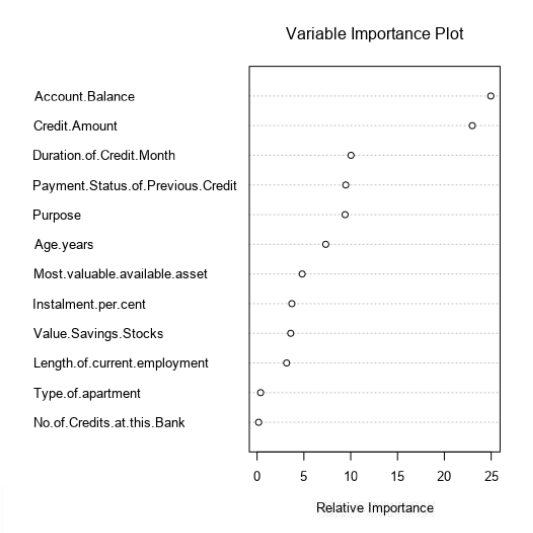


Report for Boosted Model Boosted_Credit

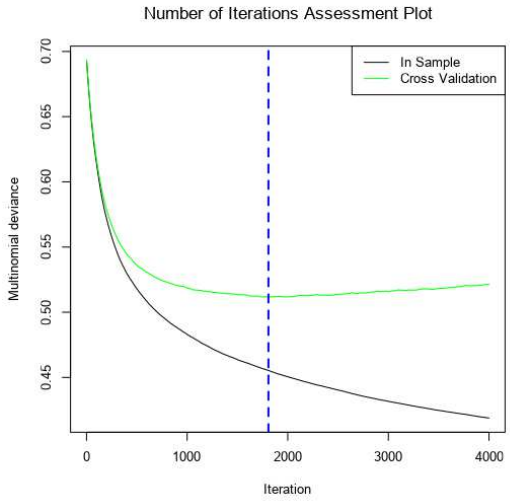
Basic Summary:

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

Plots:



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.



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).

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Credit	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Credit	0.7867	0.8609	0.7400	0.9429	0.4222
Boosted_Credit	0.7867	0.8632	0.7515	0.9619	0.3778
Log_Step_Credit	0.7600	0.8378	0.7181	0.8857	0.4667

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_Credit

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

Confusion matrix of Decision_Tree_Credit

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

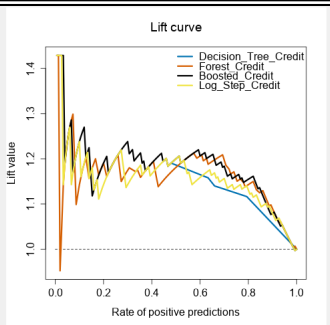
Confusion matrix of Forest_Credit

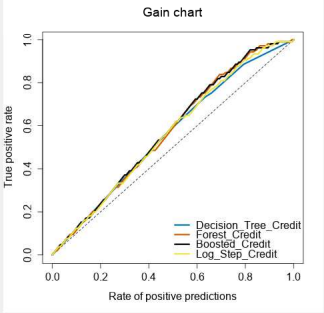
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	99	26
Predicted_Non-Creditworthy	6	19

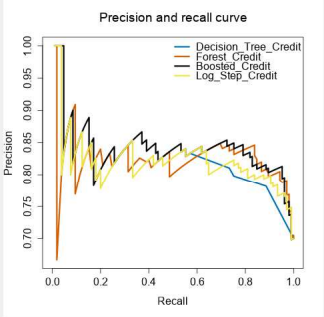
Confusion matrix of Log_Step_Credit

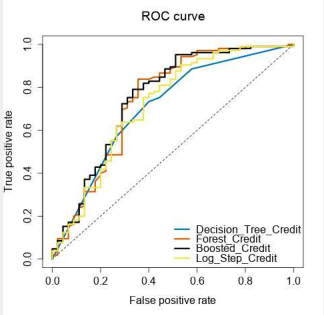
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	24
Predicted_Non-Creditworthy	12	21

Performance Diagnostic Plots









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

I choose to use Forest model based on
○ Overall Accuracy against your Validation set: 0.7867 which is the highest.
○ Accuracies within "Creditworthy" and "Non-Creditworthy" segments: 0.9429 for Creditworthy and 0.4222 for Non-Creditworthy.
○ ROC graph: it has AUC 0.7400.
○ Bias in the Confusion Matrices: it has no bias so creditworthy is almost as accurate as non-creditworthy.

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

- How many individuals are creditworthy?

There are 410 creditworthy individuals.