

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?

Based on data from current application, we'd like to use a classification modeling to evaluate 500 applications per week by system, instead of manual way.

- What data is needed to inform those decisions?

First, we need data from past loan customers which already have a decision. This data may include many independent variables, such as account-balance, revenue, length of current address, employment status and others. We may need p-value, correlation matrix and plot to find predictors.

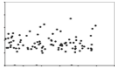
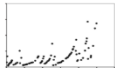





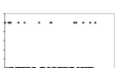



- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

I may use binary model to predict, because this is a 'yes' or 'no' question for scoring.

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.***

Numeric Fields

Name	Plot	% Missin	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Credit-Amount		0.0%	464	276.000	3,199.980	2,236.500	18,424.000	2,831.387	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Duration-of-Credit-Month		0.0%	30	4.000	21.434	18.000	60.000	12.307	
Foreign-Worker		0.0%	2	1.000	1.038	1.000	2.000	0.191	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Instalment-per-cent		0.0%	4	1.000	3.010	3.000	4.000	1.114	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Most-valuable-available-asset		0.0%	4	1.000	2.360	3.000	4.000	1.064	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
No-of-dependents		0.0%	2	1.000	1.146	1.000	2.000	0.353	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Telephone		0.0%	2	1.000	1.400	1.000	2.000	0.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Type-of-apartment		0.0%	3	1.000	1.928	2.000	3.000	0.540	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Account-Balance	0.0%	2	No Account	Some Balance	238	262	
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts	500	500	
Credit-Application-Result	0.0%	2	Creditworthy	Non-Creditworthy	142	358	
Guarantors	0.0%	2	Yes	None	43	457	
Length-of-current-employment	0.0%	3	< 1yr	1-4 yrs	97	279	
No-of-Credits-at-this-Bank	0.0%	2	1	More than 1	180	320	
Payment-Status-of-Previous-Credit	0.0%	3	Paid Up	No Problems (in this bank)	36	260	
Purpose	0.0%	4	Other	Home Related	15	355	Some values of this field have a small number of value counts. If appropriate, consider combining some value levels together.
Value-Savings-Stocks	0.0%	3	None	£100-£1000	48	298	

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.
After cleanup process, I decide to remove the columns, Concurrent-credits, Guarantors, No-of-dependents, foreign-worker because those are low variability. Duration-in-current-address is coming with 69% of missing value, so I don't use this. The column named Telephone number only have 1 or 2, it's not meaning anything, so I don't use this. Also, for just a few missing values for age-years, I used median value to replace.

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

1. Regression model

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

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

The most significant variables according to regression tables are account balance, payment status of pervious credit, purpose, credit amount, length of current employment, installment per cent. The reason why I chosen these variables, because p value over than 0.05.

- 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?

Confusion matrix of Reg_Model_		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

The overall percent accuracy is 76%. This model has bias on the results of creditworthy.

2. Decision tree



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

Based on the decision tree created, account balance, duration of credit month and value saving stocks are used.

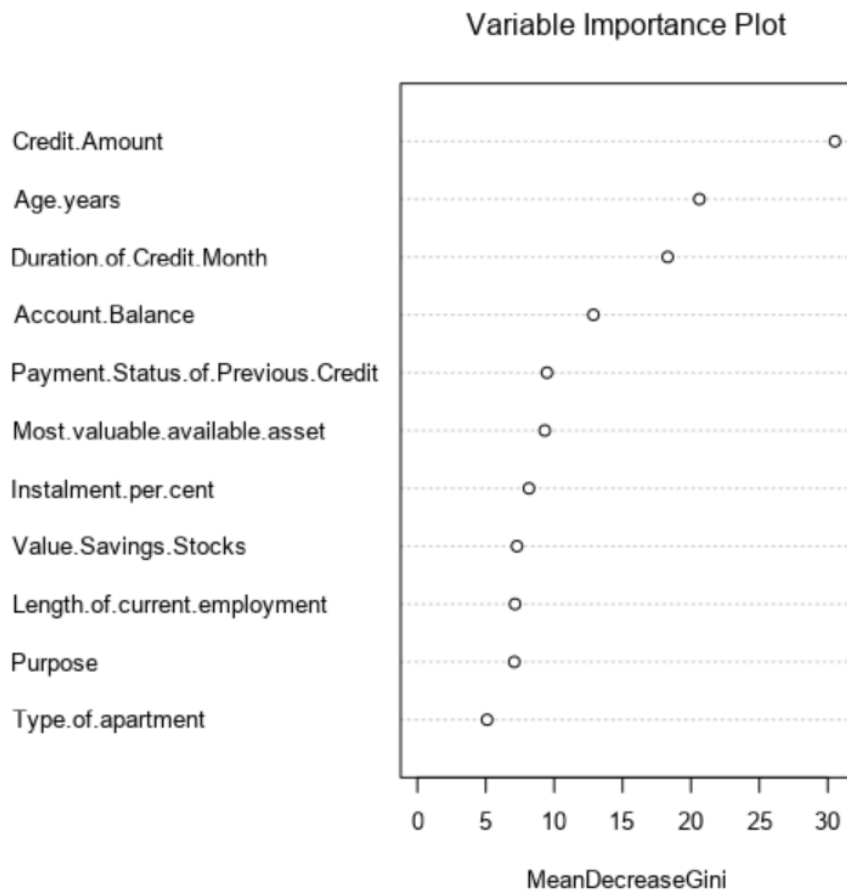
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

Confusion matrix of Decision_Tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

The percent accuracy of this model is 74.67%. This model shows the bias of predict on creditworthy.

3. Forest model

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



For the forest model, credit amount, age, duration of credit month and account balance should be used for prediction.

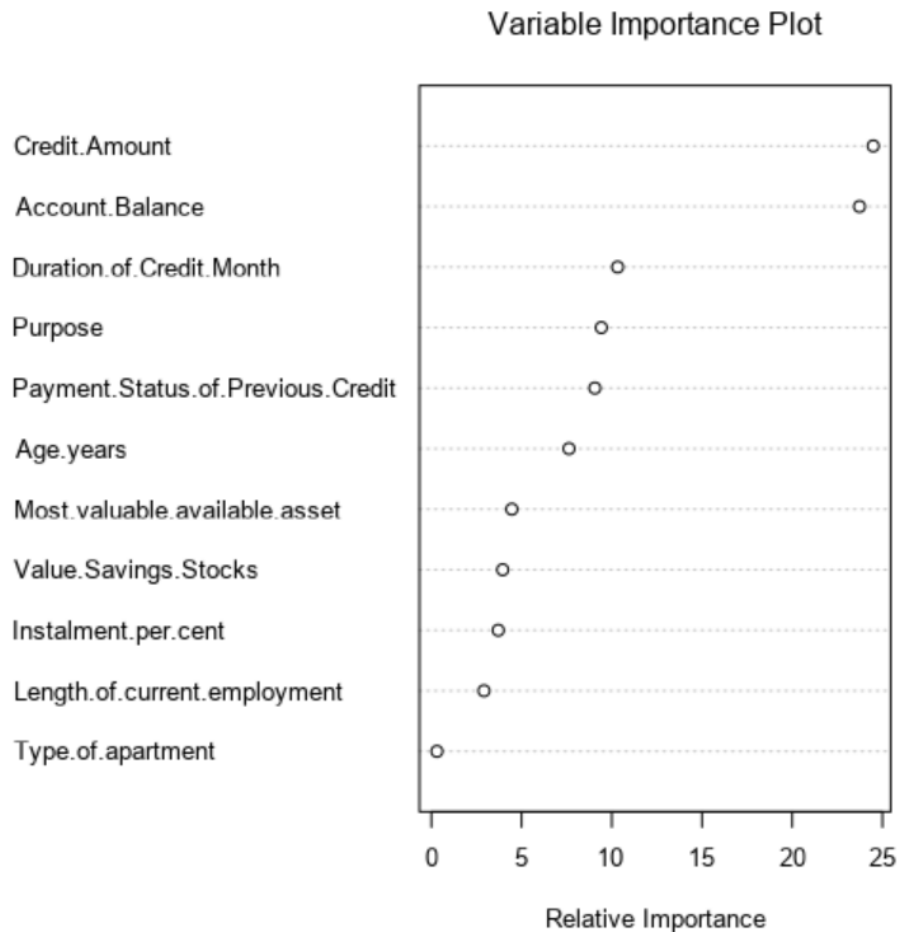
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

Confusion matrix of FM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

For this model, the overall accuracy percent is 81.33% and this is the highest accuracy score. We will use this model after. But there is still bias for predict creditworthy.

4. Boost model

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



For boost model, based on highest importance from plot, I will choose credit amount , account balance ,duration of credit month, purpose and payment status of previous credit to make prediction.

- 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?

Confusion matrix of BM		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

The overall percent is 78.67%. Bias on predict creditworthy and the performance is bad on Non-creditworthy.

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"

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Reg_Model_	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
FM	0.8133	0.8793	0.7348	0.9714	0.4444
BM	0.7867	0.8632	0.7526	0.9619	0.3778

Confusion matrix of BM

	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 FM

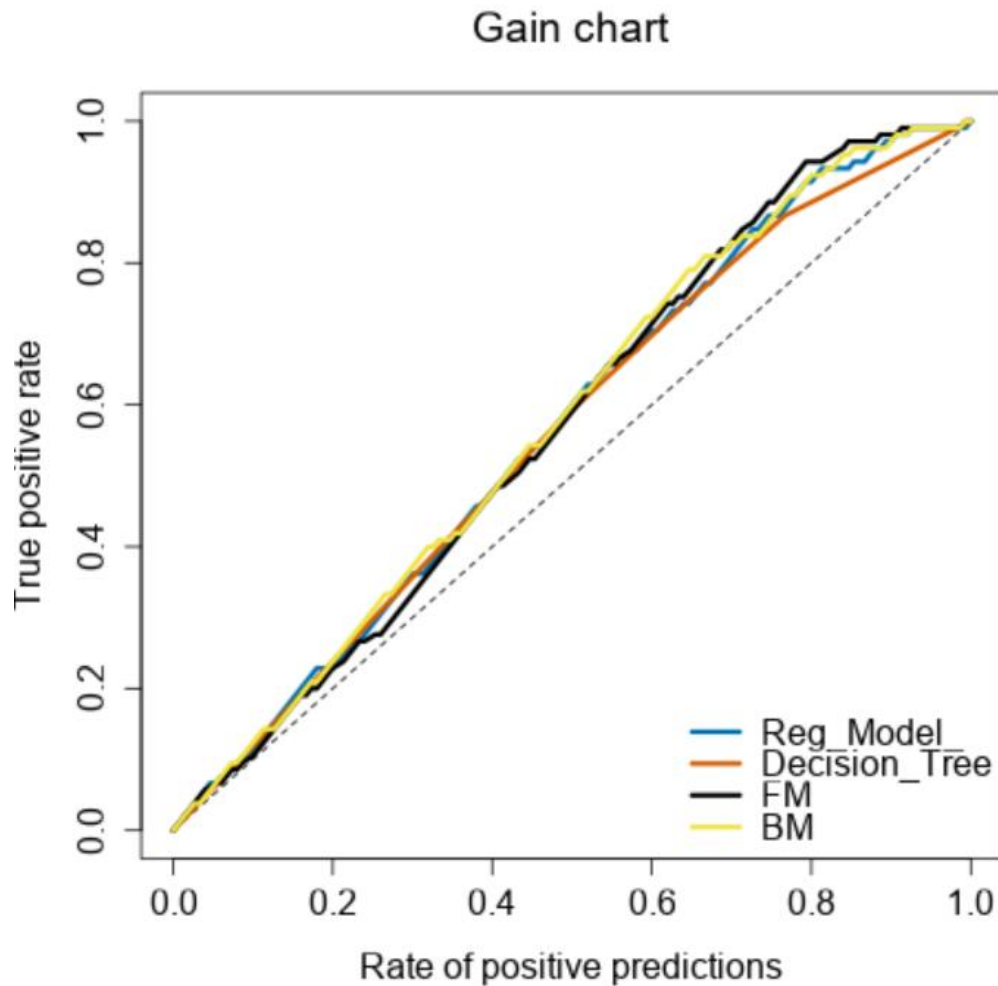
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	25
Predicted_Non-Creditworthy	3	20

Confusion matrix of Reg_Model_

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

I chose the forest model because this model has a highest accuracy rate. Also, when I looked at each accuracy for creditworthy vs non-creditworthy of forest model is still a little bit better than other models.

Overall, four models are good at to predict creditworthy, but poor performance on predict non-creditworthy. So, if we used any of this model to predict non creditworthy. Type ii error will happen a lot.



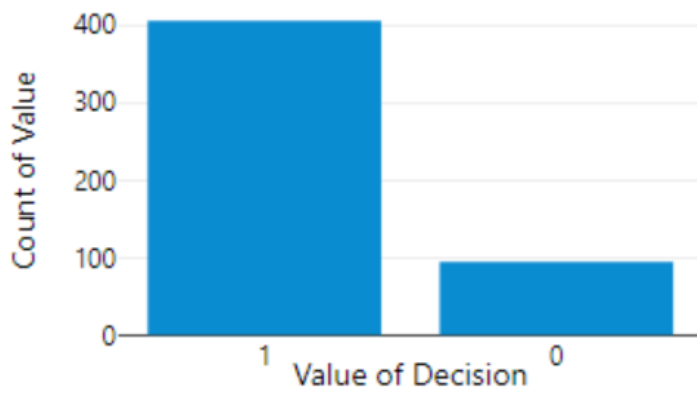
From Gain chart, forest model has a best result by having a largest true positive value than others.

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

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

- How many individuals are creditworthy?



405 customers are creditworthy. 95 are non-creditworthy.