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

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?

  I am an official of a bank, and I am responsible for determining if the clients are solvent to grant them loans, generally, the bank receives 200 weekly loan requests and the approval is by hand.
  After a financial scandal that affected the rival bank, suddenly the team receives more than 500 loan applications per week. The manager sees this as a great opportunity and wants to figure out how to process all loan applications in no more than a week.
  Based on the ranking models I recently learned, I need to systematically assess the creditworthiness of these new loan applicants, and provide the manager with the list of creditworthy clients.

- What data is needed to inform those decisions?

  • The Dataset on all past credit applications
  • The list of clients to be processed in the next few days

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

  We need to use the binary model to make our decisions, because what we seek is to identify the people who qualify and do not qualify for bank loans.

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.***

<u>Here are some guidelines to help guide your data cleanup:</u>

- 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

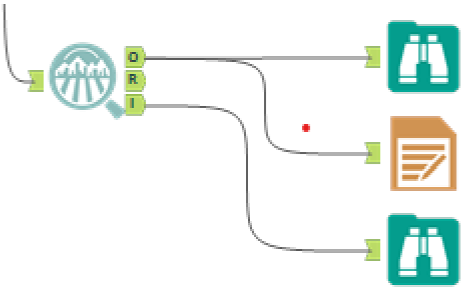
To achieve consistent results reviewers expect.

Answer this question:

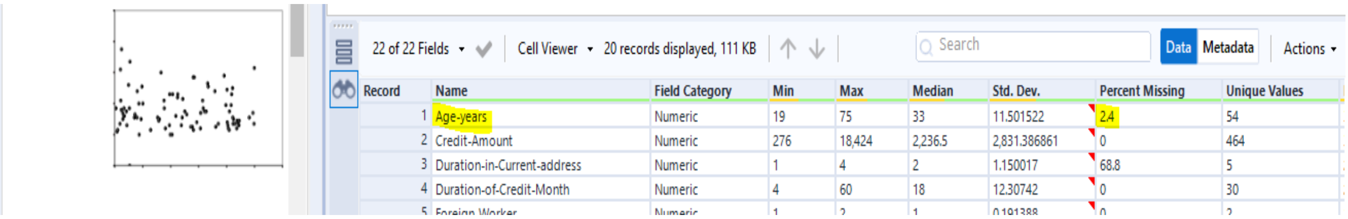
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  - I am going to impute Age-Years
  - I am going to remove Duration-in-Current-address, Occupation, Concurrent-Credits, Guarantors, Foreign-Worker, No-of-dependents, and Telephone.

  I was able to identify missing data and low variability fields using the "Field_Summary" tool and analyzing the interactive output report.



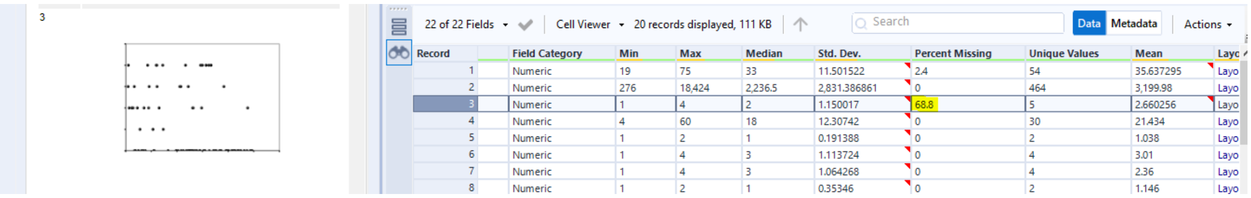
Missing data


22 of 22 Fields · Cell Viewer · 20 records displayed, 111 KB

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values
1	Age-years	Numeric	19	75	33	11.501522	2.4	54
2	Credit-Amount	Numeric	276	18,424	2,236.5	2,831.386861	0	464
3	Duration-in-Current-address	Numeric	1	4	2	1.150017	68.8	5
4	Duration-of-Credit-Month	Numeric	4	60	18	12.30742	0	30
5	Foreign Worker	Numeric	1	2	1	0.191388	0	2

The Age-Years field has 2.4 percent of missing data, using the "Field_Summary" tool I got the result, I am going to impute the field using the median of the entire field.

The Duration-in-Current-address field has 68.8% missing data, therefore it will be removed

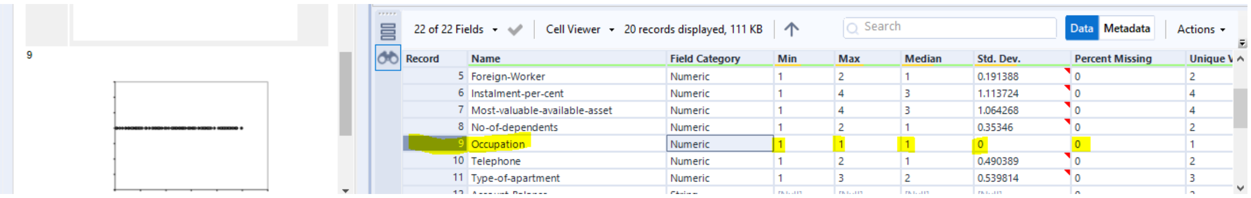


22 of 22 Fields · Cell Viewer · 20 records displayed, 111 KB

Record		Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean	Layo
1		Numeric	19	75	33	11.501522	2.4	54	35.637295	Layo
2		Numeric	276	18,424	2,236.5	2,831.386861	0	464	3,199.98	Layo
3		Numeric	1	4	2	1.150017	68.8	5	2.660256	Layo
4		Numeric	4	60	18	12.30742	0	30	21.434	Layo
5		Numeric	1	2	1	0.191388	0	2	1.038	Layo
6		Numeric	1	4	3	1.113724	0	4	3.01	Layo
7		Numeric	1	4	3	1.064268	0	4	2.36	Layo
8		Numeric	1	2	1	0.35346	0	2	1.146	Layo

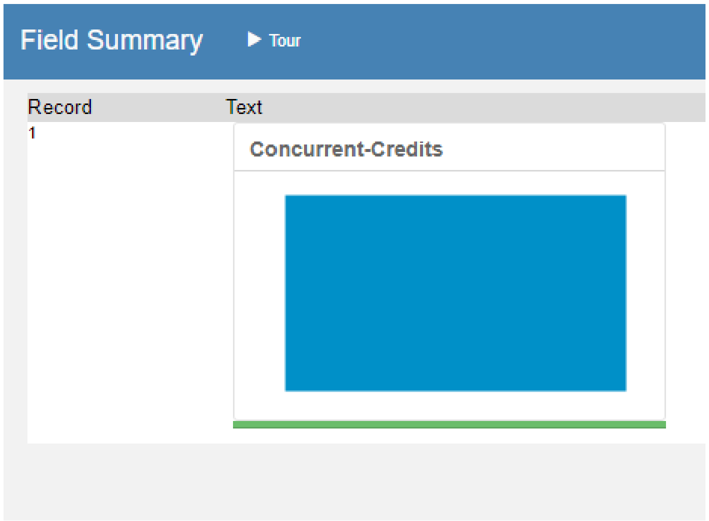
There is low variability: our data are uniform and there is no variation in the data.

The occupation field only has "1" as a value



Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique V
5	Foreign-Worker	Numeric	1	2	1	0.191388	0	2
6	Instalment-per-cent	Numeric	1	4	3	1.113724	0	4
7	Most-valuable-available-asset	Numeric	1	4	3	1.064268	0	4
8	No-of-dependents	Numeric	1	2	1	0.35346	0	2
9	Occupation	Numeric	1	1	1	0	0	1
10	Telephone	Numeric	1	2	1	0.490389	0	2
11	Type-of-apartment	Numeric	1	3	2	0.539814	0	3

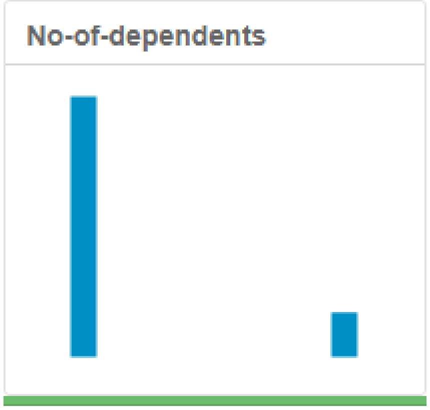
The concurrent credits field only has a value "Other banks / deposits", 500 instances in total.



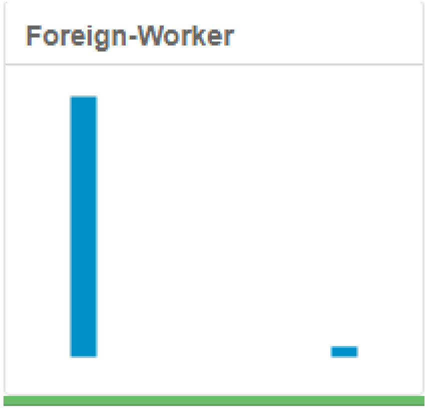
Record	Text
1	Concurrent-Credits

Low variability: the data field is biased towards one data type

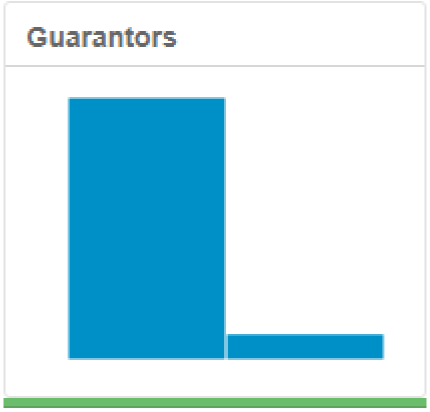
- The telephone field should be removed because it does not contribute anything about solvency.
- Number of dependents should be eliminated because there is a bias towards "1"



- Foreign worker should be removed because most of the data is skewed by “1”.



- Guarantors most of the data are biased towards “1”



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

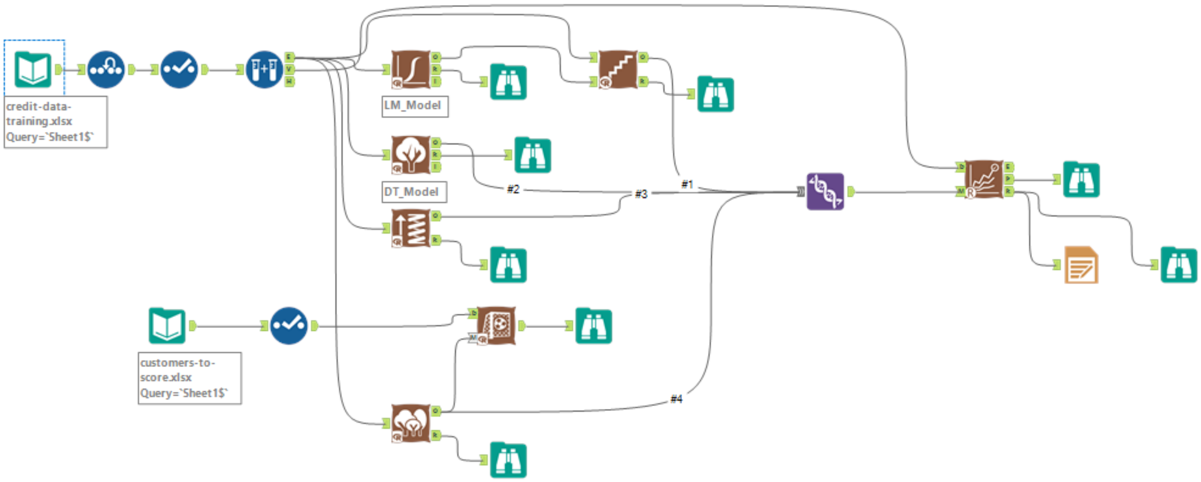
Answer these questions for ***each model*** *you created:*

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

- 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?

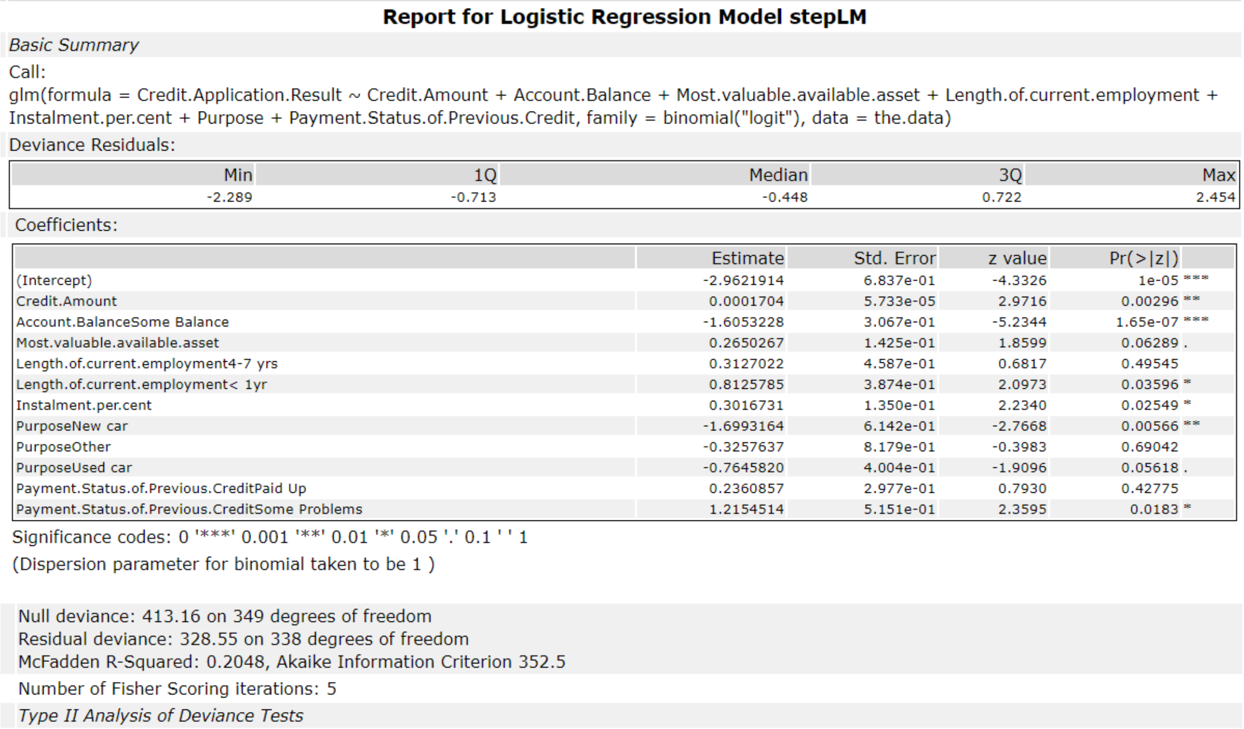
You should have four sets of questions answered. (500 word limit)

I built all four models in Alteryx and used the tool to compare the models to validate Logistic Stepwise, Decision Tree, Forest Model and Boosted Model.



Significant predictive variables

Logistic_Stepwise: According to the report, the significant predictors are account balance, previous credit payment status, purpose, credit amount, current job duration and percentage of installment payments.



Report for Logistic Regression Model stepLM

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Credit.Amount + Account.Balance + Most.valuable.available.asset + Length.of.current.employment + Instalment.per.cent + Purpose + Payment.Status.of.Previous.Credit, 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 ***
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .
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 *
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 .
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 *

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

(Dispersion parameter for binomial taken to be 1)

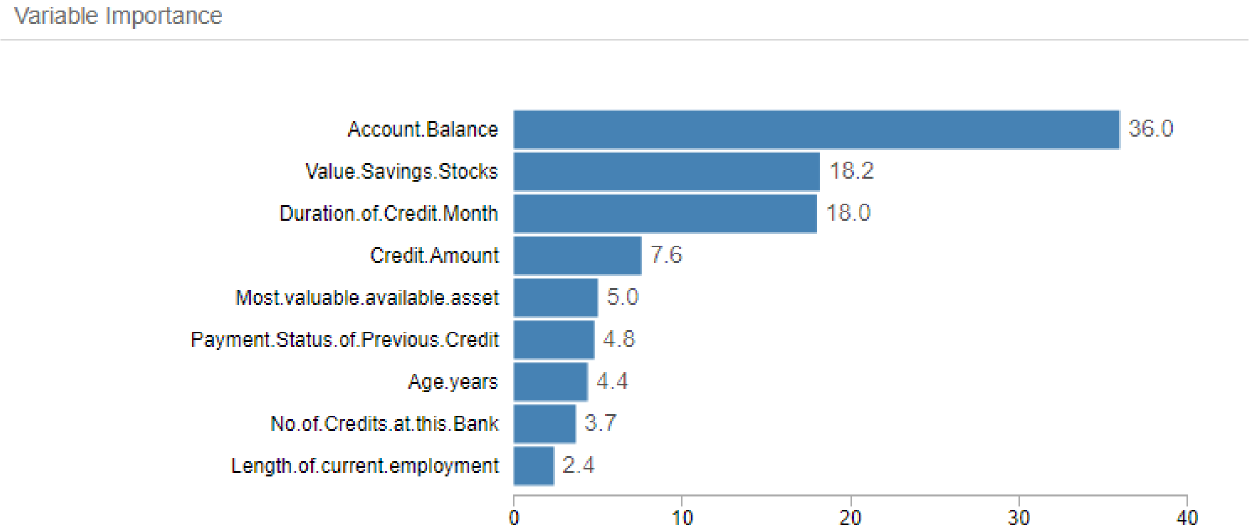
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

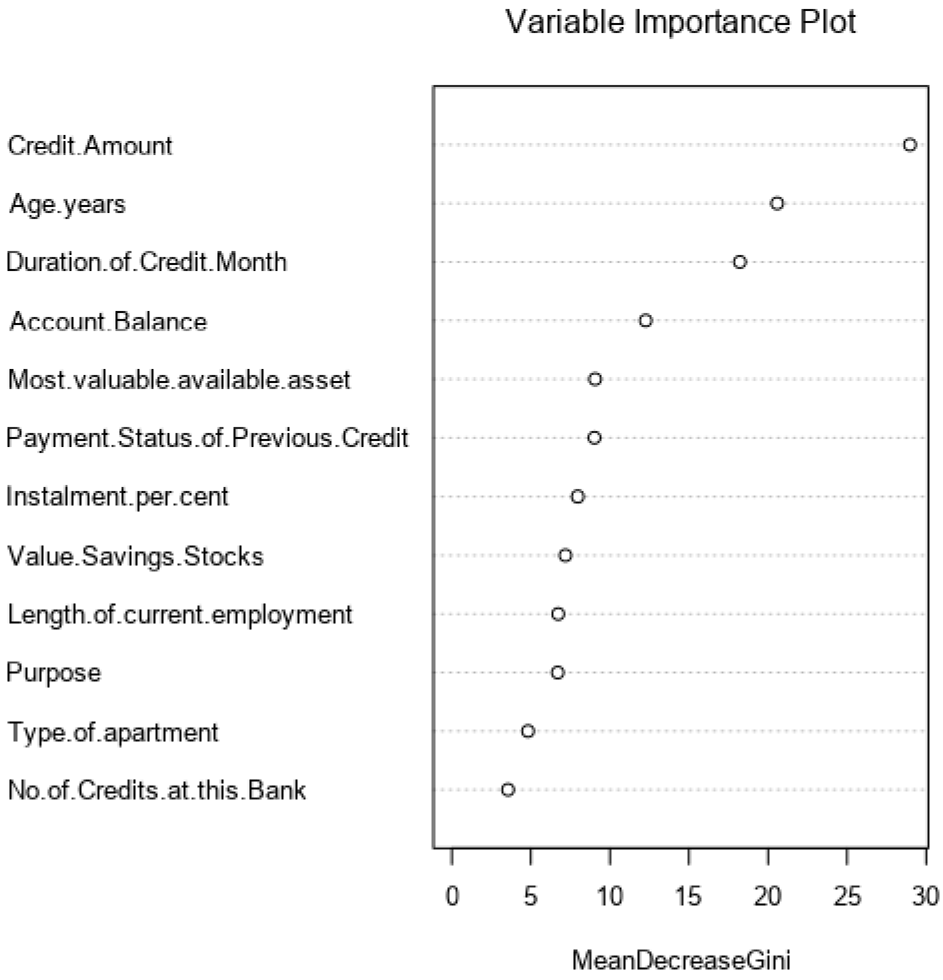
Type II Analysis of Deviance Tests

Decision tree: according to the variable importance report shown in the graph, the 3 main predictive variables are the account balance, value savings actions and the duration of

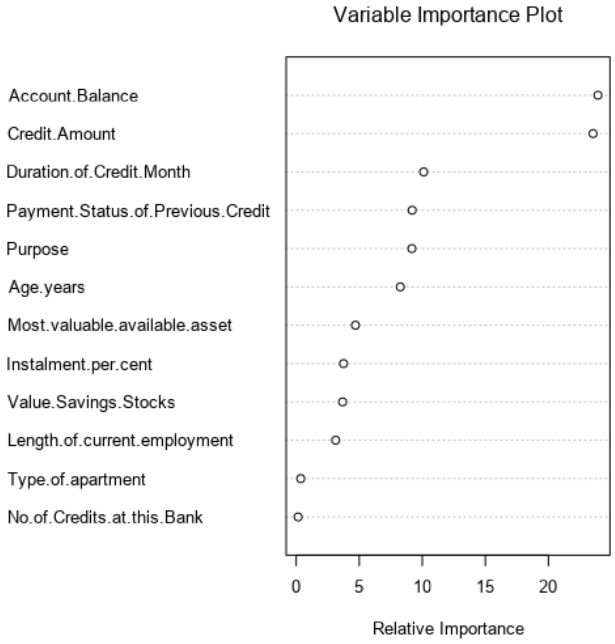
the credit month.



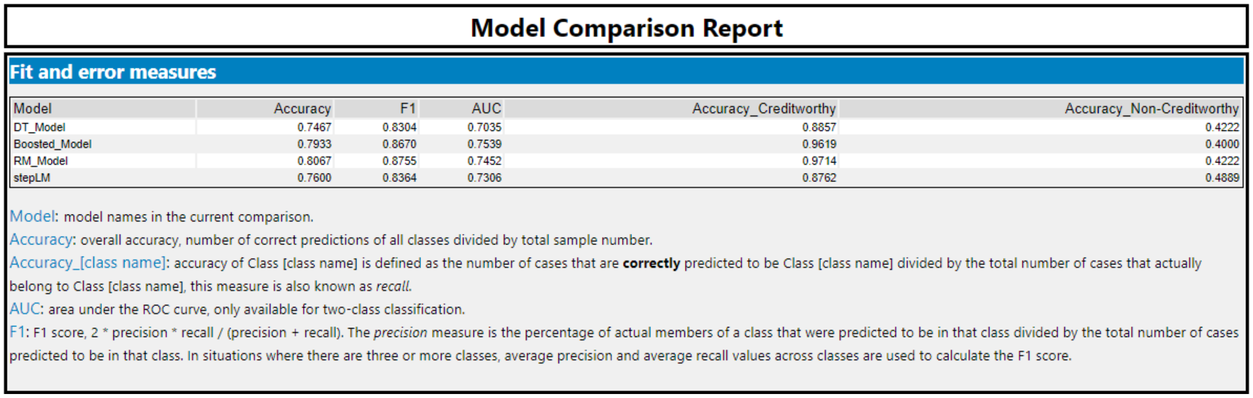
Forest Model: it is observed in the importance graph, the three main predictive variables are Amount of credit, years of age and duration of the month of credit.



Boosted_Model: According to the graph, the three main predictor variables are credit amount, balance amount and length of credit month.



Validate and compare model: I used the model comparison tool to validate and compare the accuracy of the models used.



Model Comparison Report

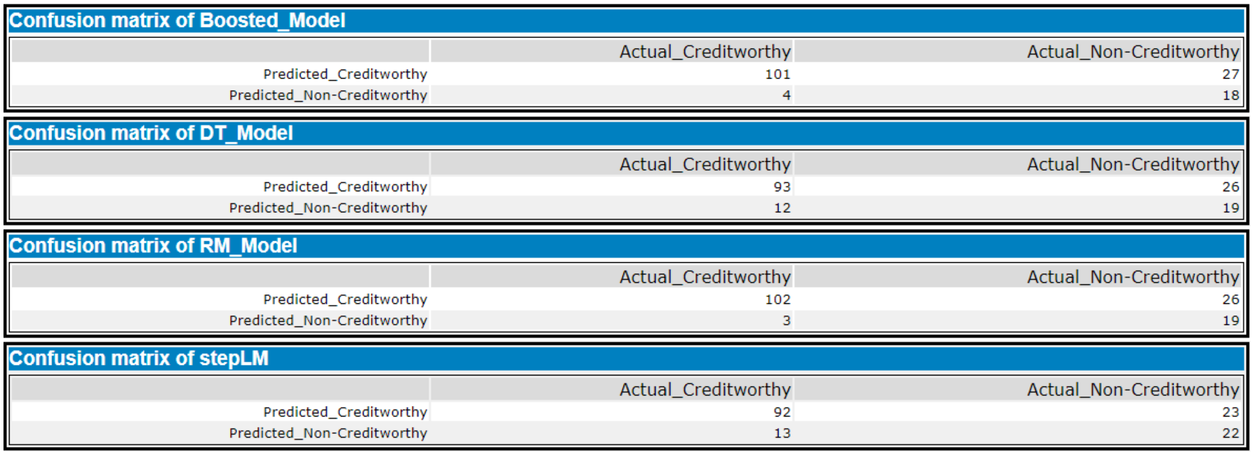
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Model	0.7467	0.8304	0.7035	0.8857	0.4222
Boosted_Model	0.7933	0.8670	0.7539	0.9619	0.4000
RM_Model	0.8067	0.8755	0.7452	0.9714	0.4222
stepLM	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.

According to the comparison report of the previous model, the precision of the Logistic_Stepwise model is 0.76, the decision tree is 0.74, the forest model 0.80 and the boosted model 0.79

Let's take a look at the confusion matrix of all models



Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of DT_Model

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

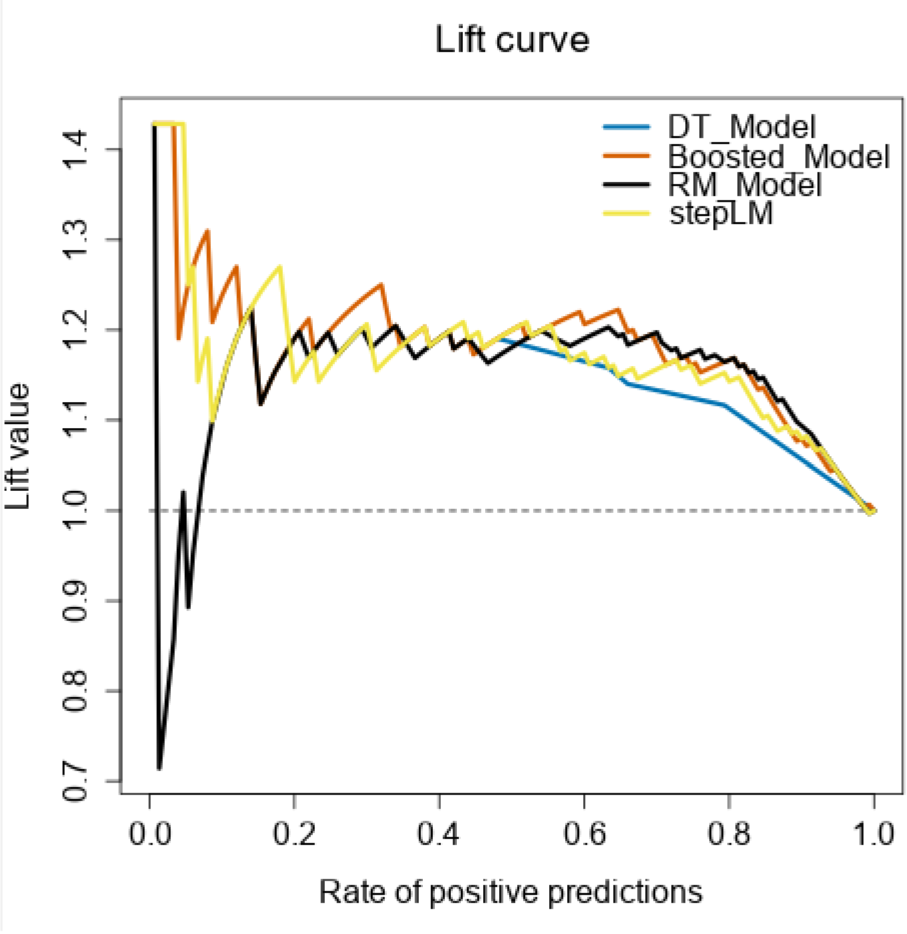
Confusion matrix of RM_Model

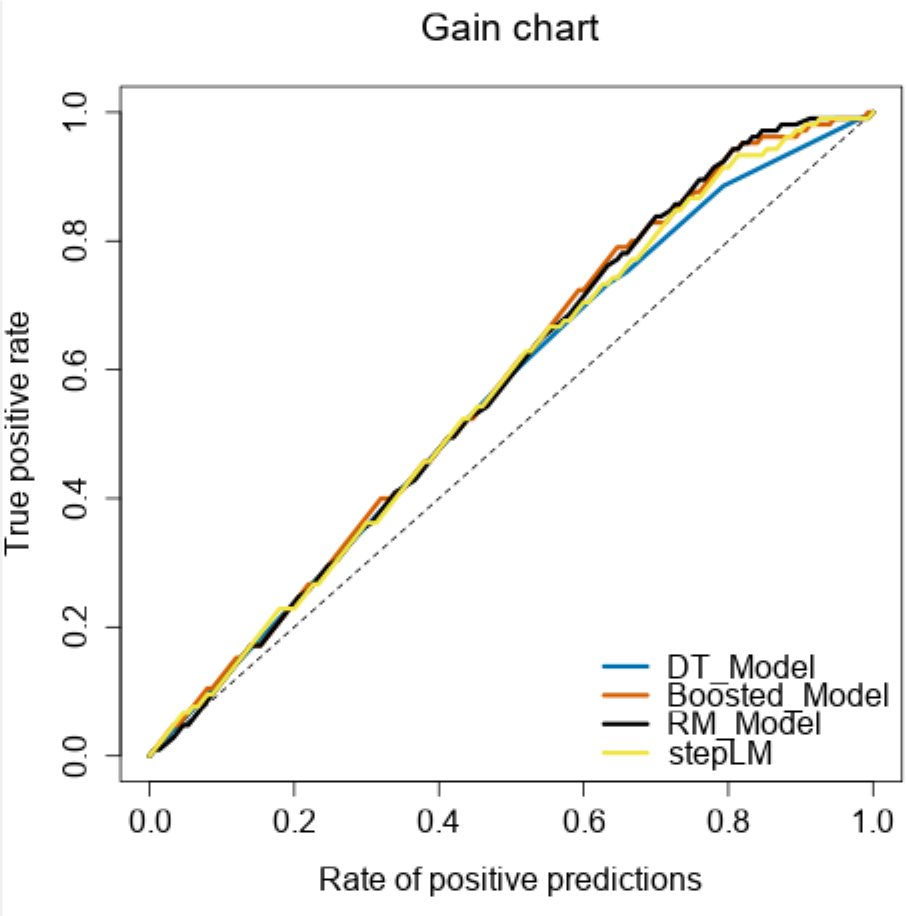
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of stepLM

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

Our dataset is a bit out of balance, there are many solvent applicants than non solvent applicants, so that a predictive model can be selected, we are interested in the overall precision and the positive and negative predictive value, and the f1 score to know which is the best model.





We can deduce that the models that are biased towards the prediction of individuals who are creditworthy, since they do not predict clients who are not creditworthy almost at the same level as those who are.

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)

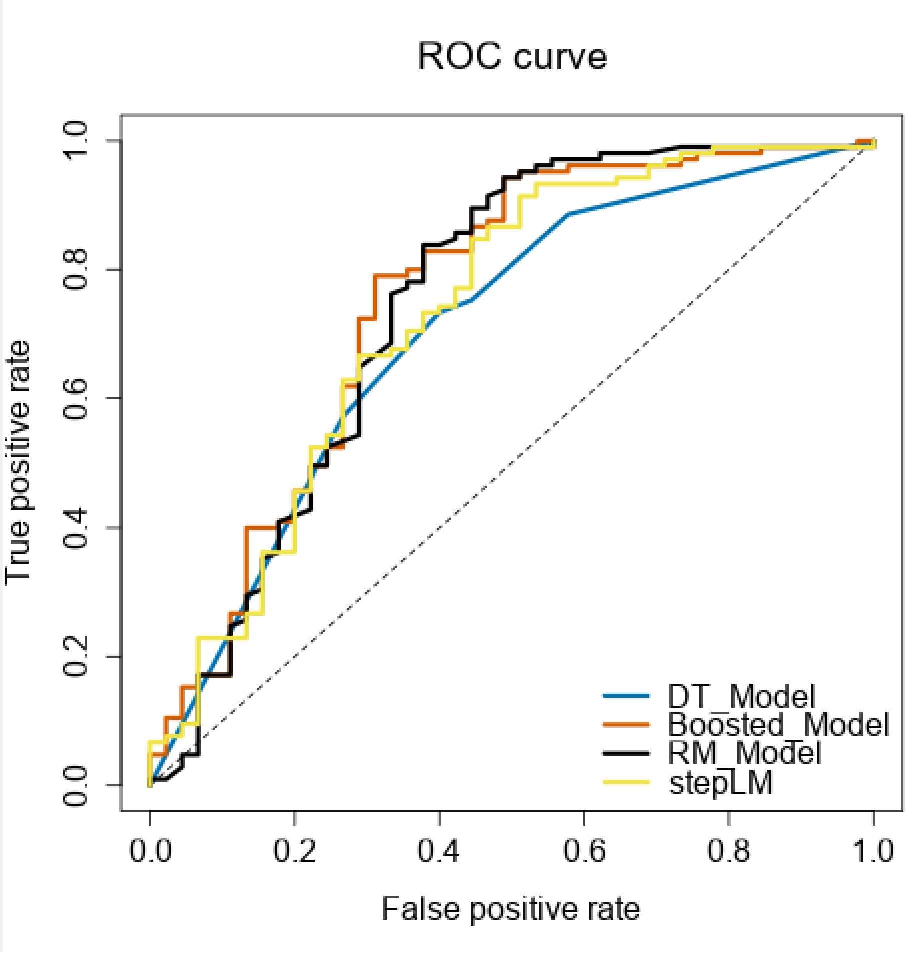
Answer these questions:

- 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

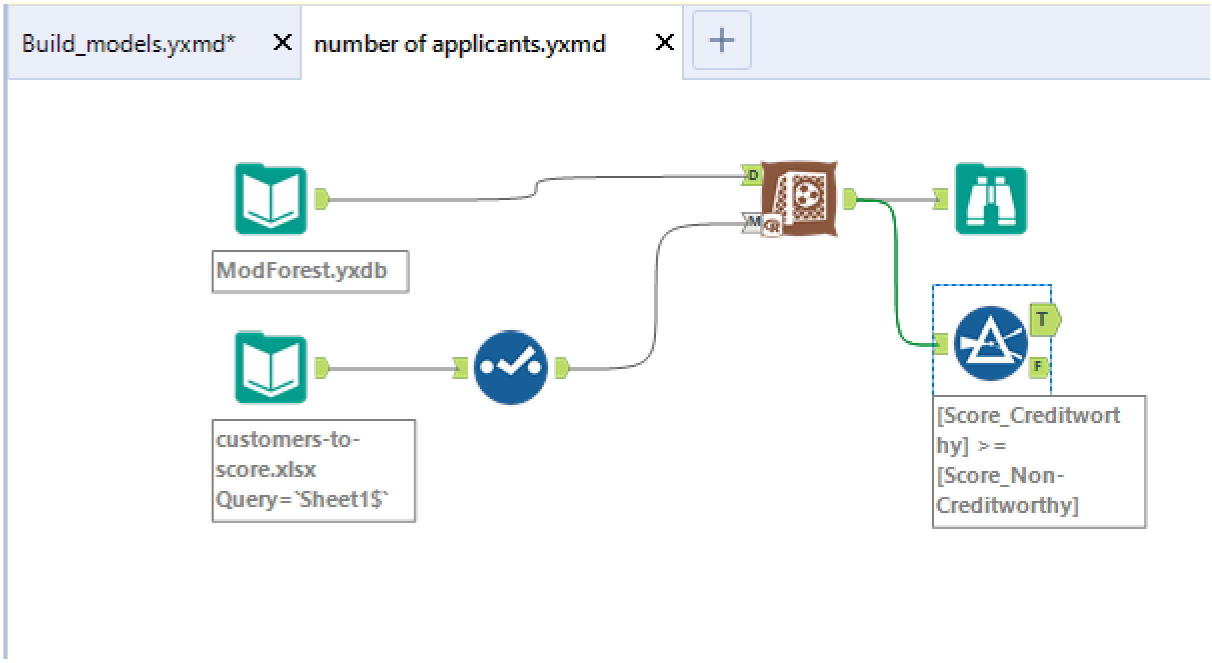
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

I choose to use the Forest Tree model considering the general precision compared to the other models, established at 0.80, having the highest Accuracy_Creditworthy score with 0.95 and 0.42 Non-Credit, in the following graph of the roc curve comparing all The models.



Applicants score
We generate the file "ModForest.yxdb" and we plug in a score together with the "customer-to-score", generating the output file, any request with a score of> = 0.5 is approved, in this case we have 411 of the 500 applicants !!



ModForest.yxdb

customers-to-score.xlsx
Query=`Sheet1$`

[Score_Creditworthy] >= [Score_Non-Creditworthy]

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.