UDACITY ND.Business Analytics
Francihelena Uzcategui

December 12, 2018.

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.

- What decisions needs to be made?

Due to a Financial scandal linked to a competitor bank, fortunately, this week we received 500 loan applications, representing 300 applications more than our usual demand. Processing them by hand will take a while, therefore we'll compute the data using Alteryx tool to process the application loan list of creditworthy customers to approve the loan or not.

- What data is needed to inform those decisions?

We need to calculate how many new creditworthy customers can get a loan this week. First, to build the Prediction model, mainly we already have the information about our Customers, such as Credit Application Result, Duration of Credit Month, Payment Status of Previous Credit, Purpose, Credit Amount, Value Savings Stocks, Length of current employment, and Instalment per cent. Second, to apply our prediction model we need new data from the 500 customers, covering the same variables as the available data.

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

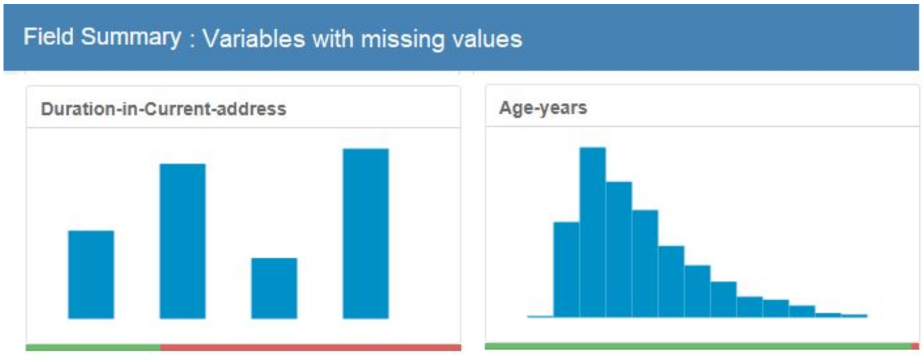
Our goal is to predict outcomes from Data available. We'll predict the categories as a customer falls, a Binary: creditworthy and non-creditworthy customers.

Step 2: Building the Training Set

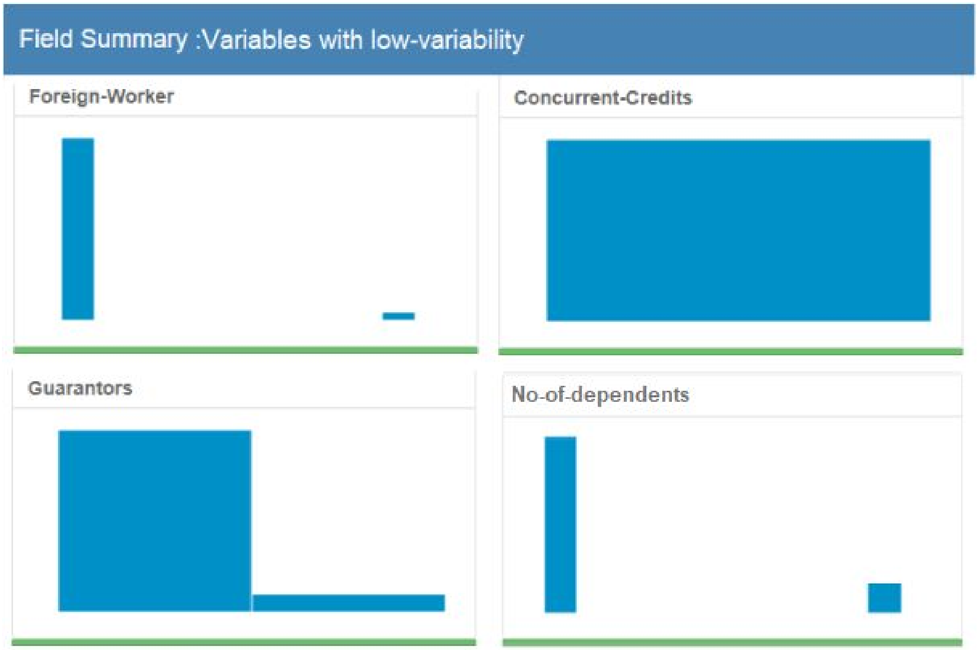
Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

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

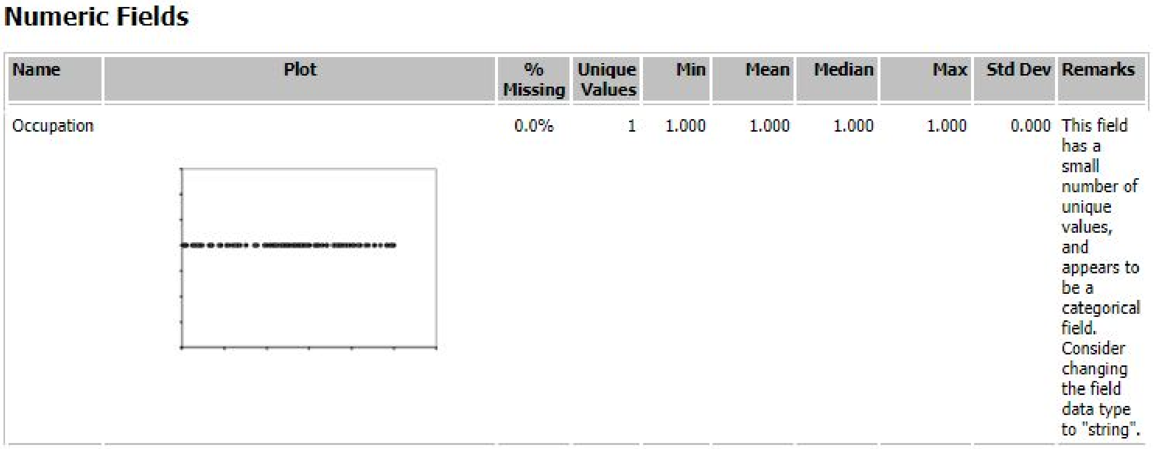
The Field Summary tool provided a landscape of all variables. The Field summary report below shows histograms and statistic summaries, the red sign depicts missing values, conversely, the green color represents sufficient values. First, the field Duration-in-Current-address has a lot of missing values 68%, so we discarded it. Although, we kept the Age-years variable that has just 2% of missing values; initially we tried to manage it by Data Cleansing tool, but finally, we choose the Imputation tool with her median 33.



Due the low-variability, we removed the fields Foreign-Worker, Concurrent-Credits, Guarantors, and No_ of_dependents.



We removed the Occupation because of is a uniform data, as well.

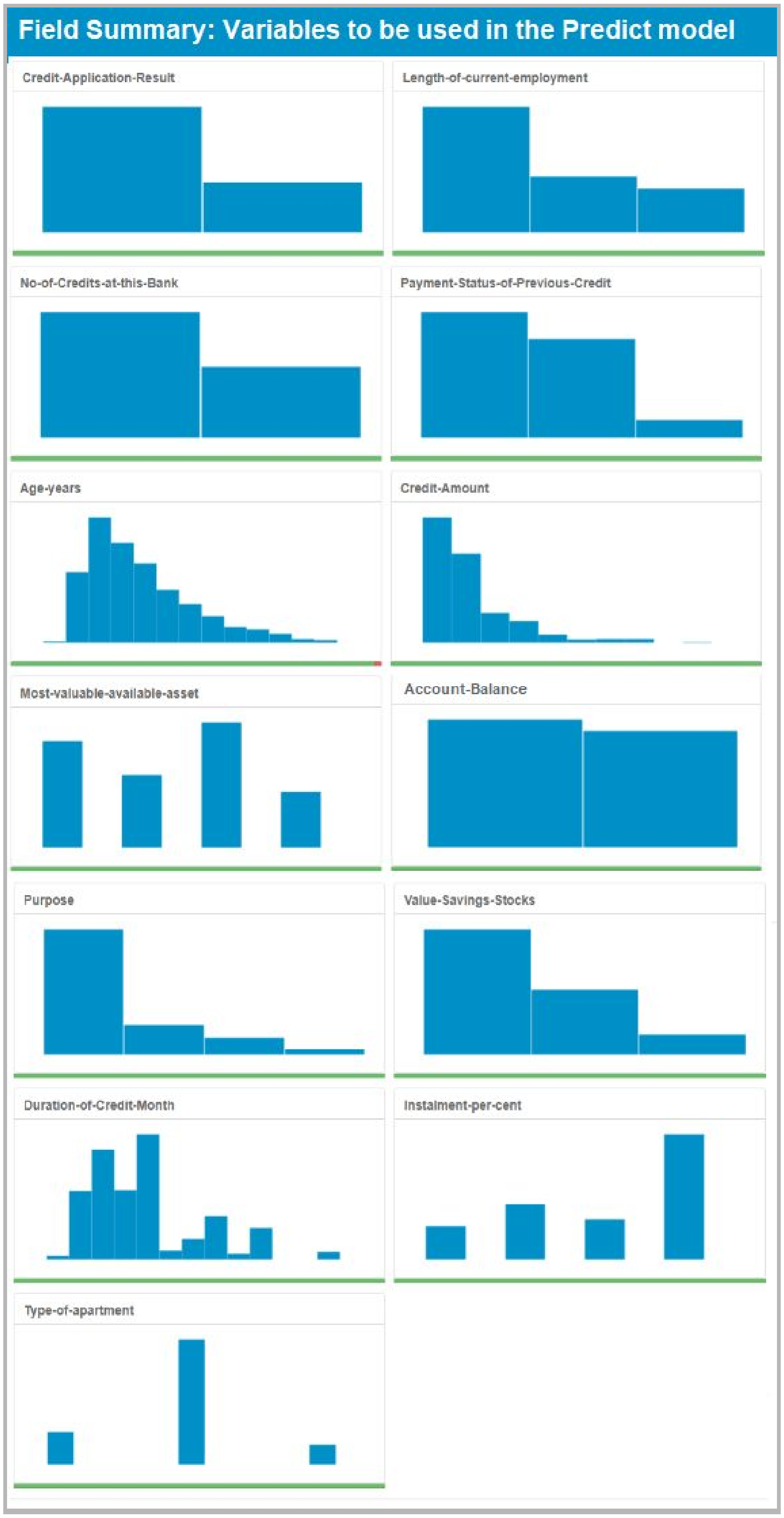


Numeric Fields

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
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".

The Telephone field was removed because included private information and is not relevant to classification.

After the cleansing and preparation, the model keeps only thirteen variables to be evaluated.


Field Summary: Variables to be used in the Predict model
Credit-Application-Result
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit
Age-years
Credit-Amount
Most-valuable-available-asset
Account-Balance
Purpose
Value-Savings-Stocks
Duration-of-Credit-Month
Instalment-per-cent
Type-of-apartment

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.

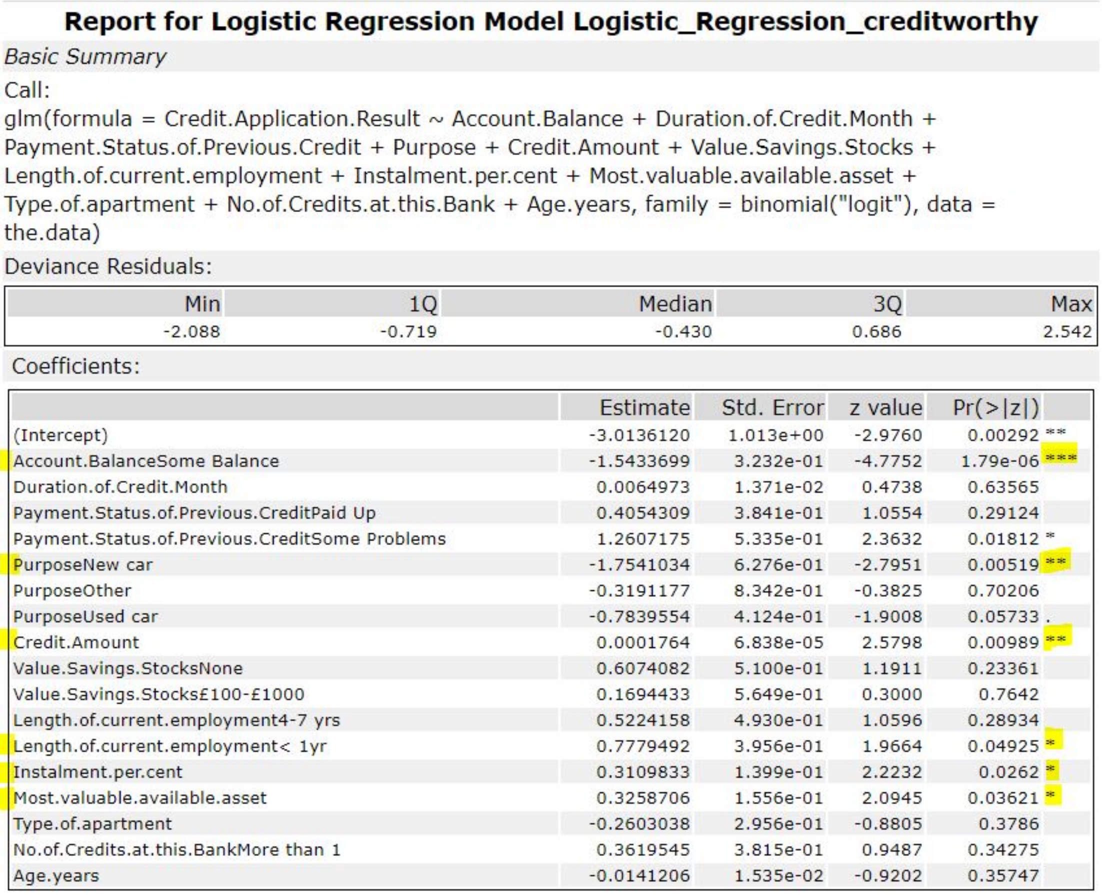
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model Answer these questions for **each model** you created:*

- 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)

1.- Logistic Model and Stepwise:

Logistic Regression model classified 17 variables. The relevant predictor variables with low P-value are: Account.Balance, Purpose, Credit.Amount, Length.of.current.employment, Instalment.per.cent, Most.valuable.available.asset, respectively.



Report for Logistic Regression Model Logistic_Regression_creditworthy

Basic Summary

Call:
glm(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, family = binomial("logit"), data = the.data)

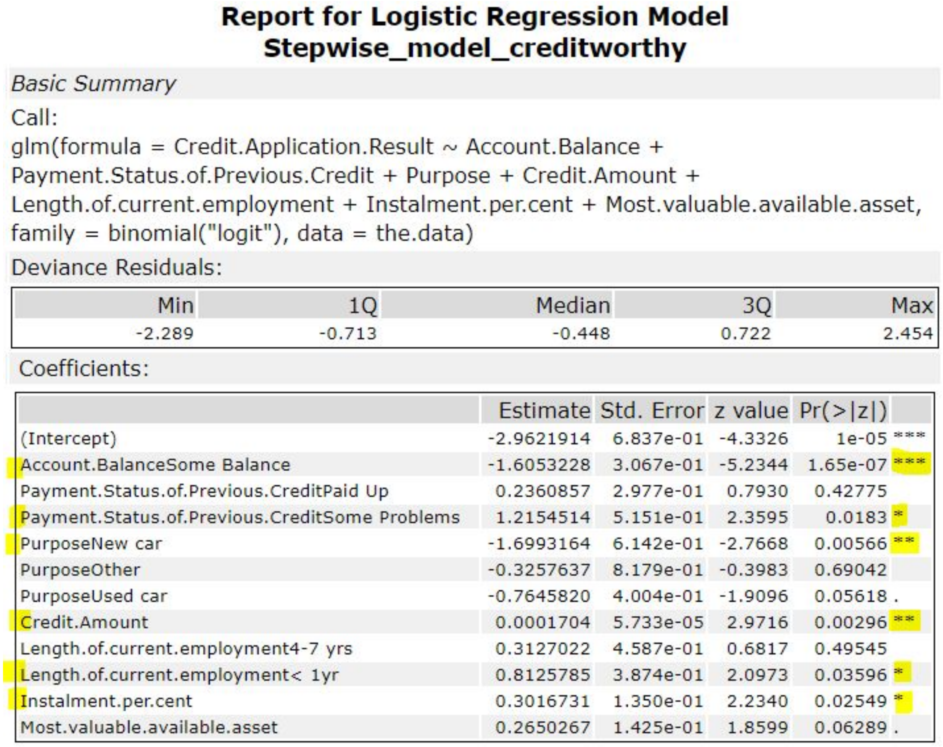
Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	

Applying the Stepwise model the variables were reduced to 11, keeping almost the same variables aforementioned: Account.Balance, Purpose, Credit.Amount, Payment.Status.of.Previous.Credit, Instalment.per.cent, and Length.of.current.employment. Although Account.Balance increased her significance to 0.000000165.



Report for Logistic Regression Model
Stepwise_model_creditworthy

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)

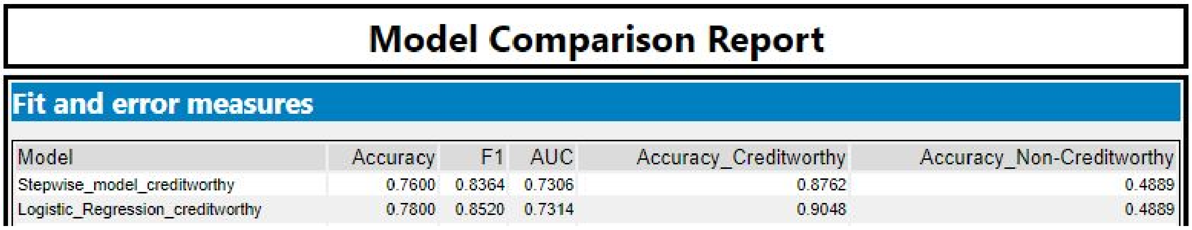
Deviance Residuals:

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	.

Logistic regression and Stepwise model share the third and fourth position as the higher Accuracy values among all models.

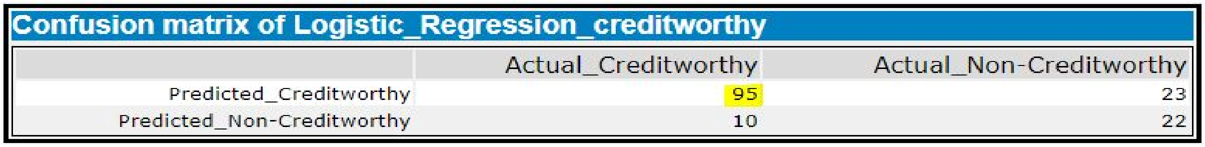


Model Comparison Report

Fit and error measures

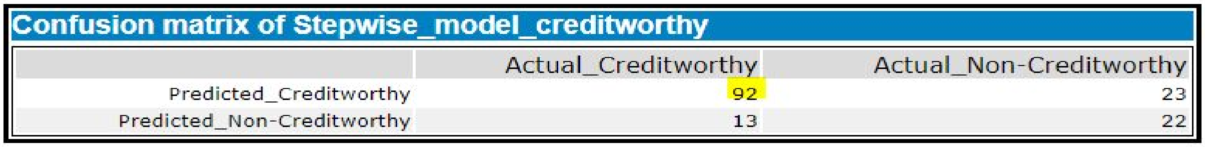
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_model_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889
Logistic_Regression_creditworthy	0.7800	0.8520	0.7314	0.9048	0.4889

Both models are biased to Creditworthy with high True positive values. It is confirmed by ROC curve and Gain chart where all models going to True positive rate.
Logistic Regression true positive rate: TP/ actual yes = 95/105 ⇒ 0.9048; being the second higher Accuracy_creditworthy value.
Stepwise model true positive rate: TP/actual yes = 92/105 ⇒ 0.8762; it is the third higher Accuracy_creditworthy value.
Logistic.R and Stepwise models' false positive value is 23, being the lowest among the rest of the models.



Confusion matrix of Logistic_Regression_creditworthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22



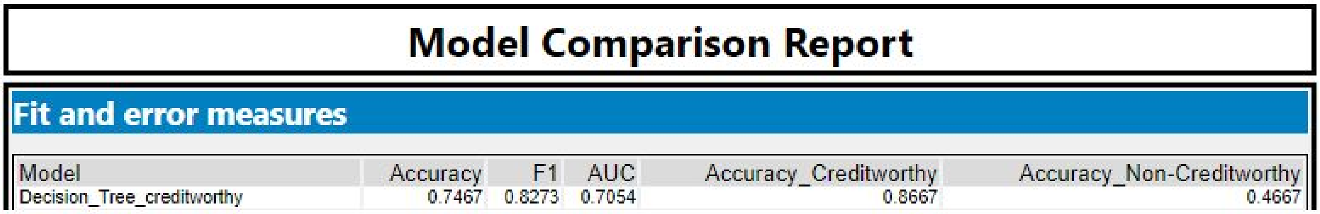
Confusion matrix of Stepwise_model_creditworthy

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

2.- Decision Tree:

Root Node Error is close to 28%, more than a quarter of values went to the incorrect terminal node.

Decision Tree accuracy is the lowest value among all the models, being 0.7467. This Accuracy creditworthy class has the lowest value, as well.

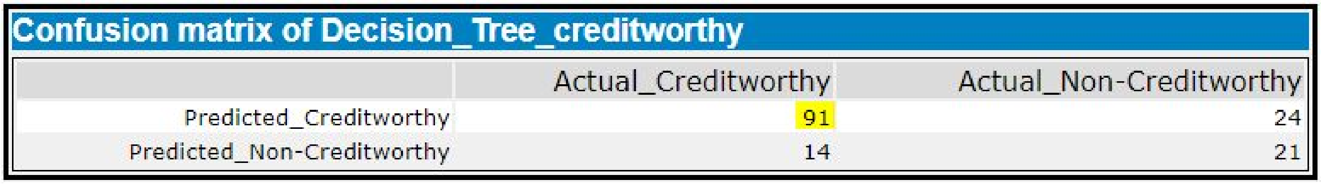


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_creditworthy	0.7467	0.8273	0.7054	0.8667	0.4667

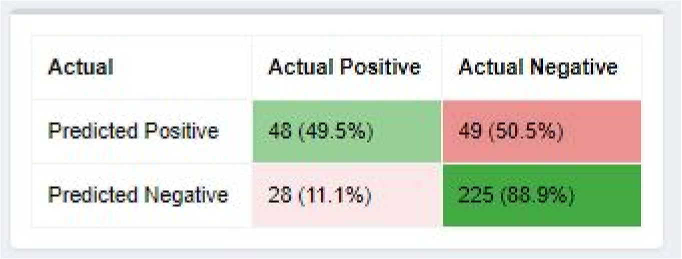
Decision Tree model true positive rate: TP/actual yes = 91/105 ⇒ 0.8667. ROC curve and Gain chart shows the model goes to the left corner, nonetheless, the black lines are close to the baseline, being this a reason for low accuracy. AUC value confirms that the Decision Trees line is far to number 1 and closer to the baseline, it means low accuracy.



Confusion matrix of Decision_Tree_creditworthy

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

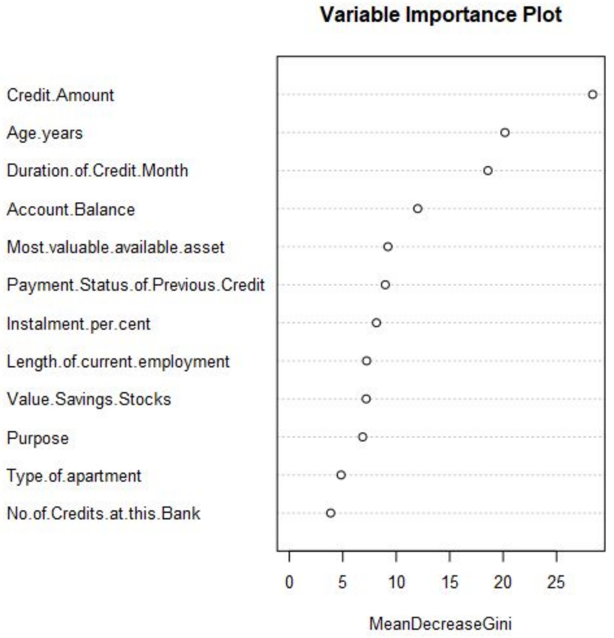
For the Classification Tree the Misclassification section shows the confusion matrix with high False Positives more than a half of set, and a low False Negative, it means a High recall and low precision.



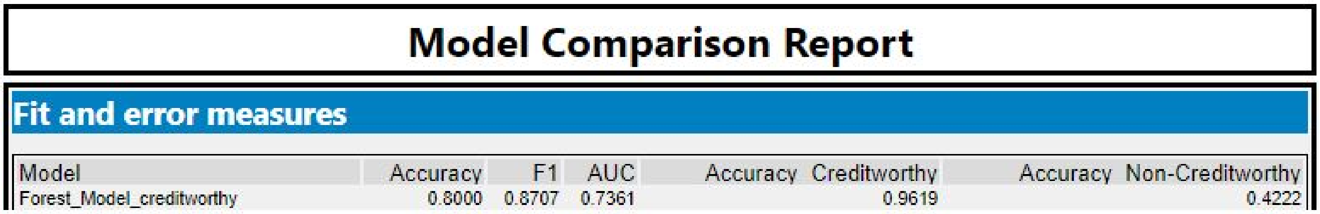
Actual	Actual Positive	Actual Negative
Predicted Positive	48 (49.5%)	49 (50.5%)
Predicted Negative	28 (11.1%)	225 (88.9%)

3.- Forest Model:

According to the Variable Importance plot, the top 4 variables are Credit amount, Age years, Duration of credit month, and Account Balance, respectively. These variables have a large Mean Decrease in Gini Values.



The overall accuracy is 0.80, being the highest Accuracy value among all models.

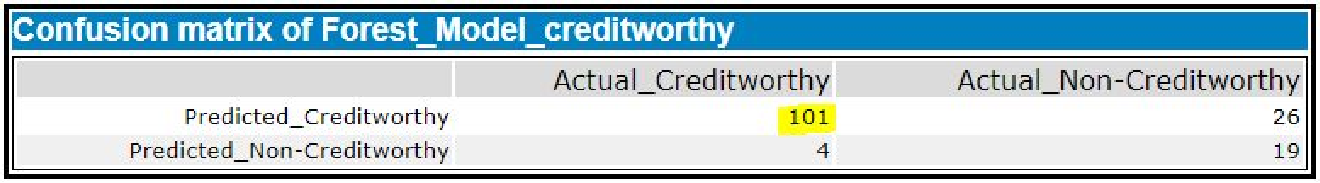


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy Creditworthy	Accuracy Non-Creditworthy
Forest_Model_creditworthy	0.8000	0.8707	0.7361	0.9619	0.4222

Confusion matrix depicts the highest True positive value, contrary to a tiny amount of false negative value (the lowest among all models). The matrix⇒ 101/105= 0.9619 predicts the Creditworthy, being a test with high sensitivity which means rare misdiagnosis.

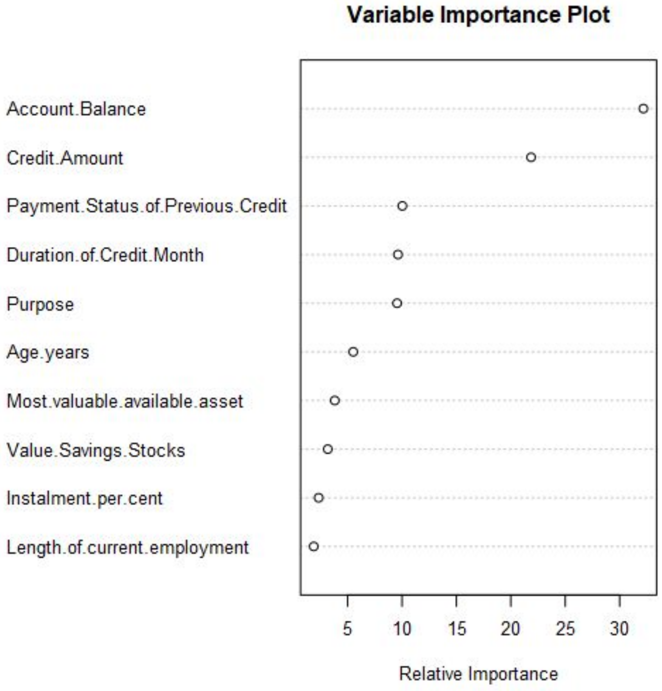


Confusion matrix of Forest_Model_creditworthy

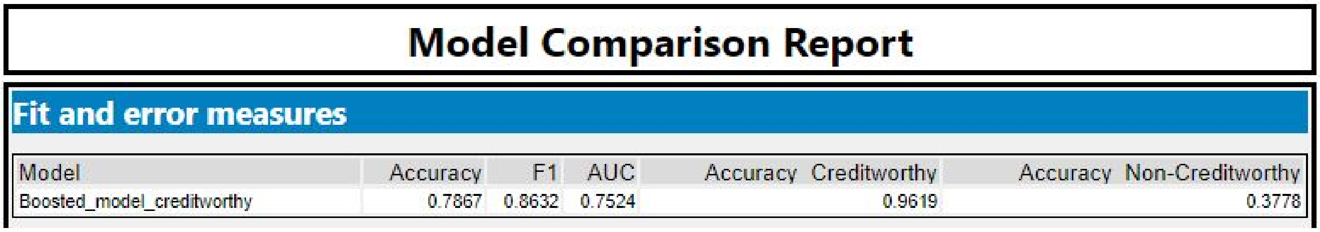
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

4.- Boosted Model:

According to the Variable Importance plot, the top 2 variables are Account Balance and Credit Amount. This classification is similar than the previous model-Forest, which included them on his top 4 of importance.



Boosted model had the second higher accuracy value 0.7867, among all models used.

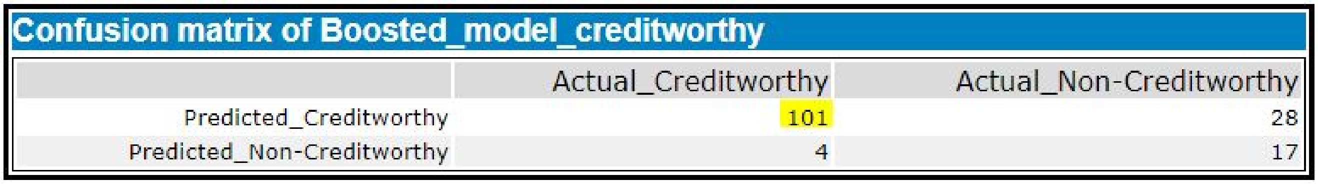


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_model_creditworthy	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix shows the high True positive value 101 in contrast to a tiny amount of false negative value 4. The matrix⇒ 101/105= 0.9619 predicts the Creditworthy, showing high sensitivity test with low failure rate; being the same value than Forest model.



Confusion matrix of Boosted_model_creditworthy

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

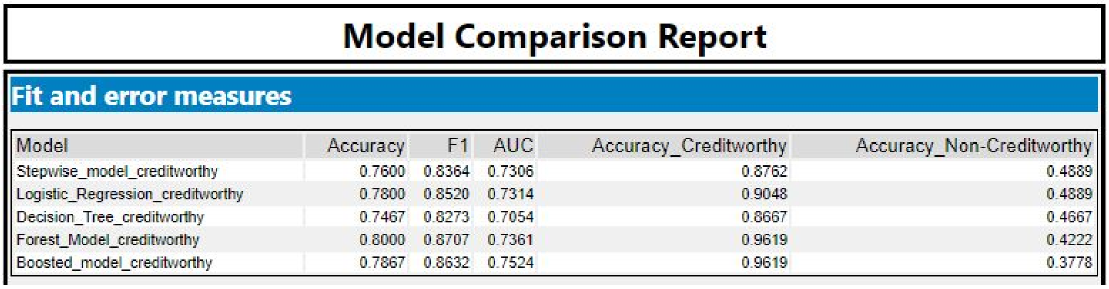
Step 4: Writeup

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



Model Comparison Report

Fit and error measures

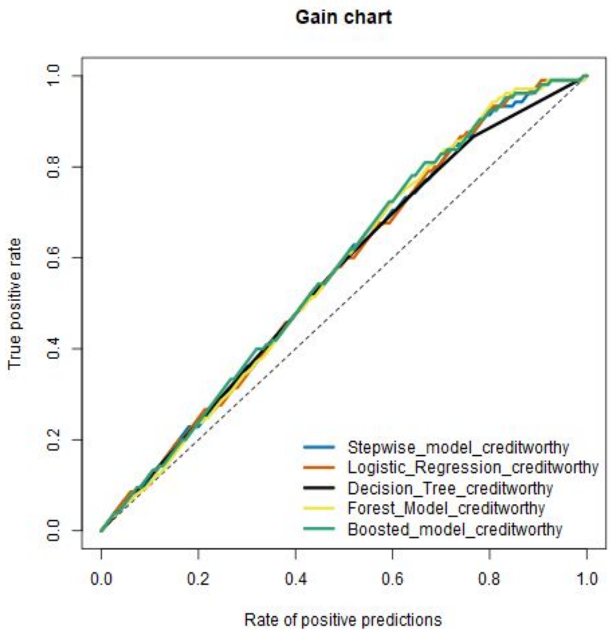
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_model_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889
Logistic_Regression_creditworthy	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree_creditworthy	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model_creditworthy	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_model_creditworthy	0.7867	0.8632	0.7524	0.9619	0.3778

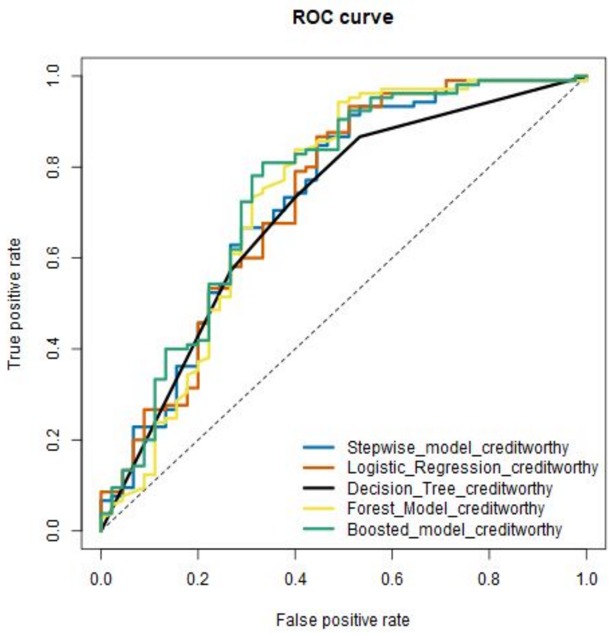
According to the above report, the four models are biased to Creditworthy, therefore we used the overall Accuracy value to select the highest value, it is the **Forest Model**.

● The model 's accuracy is 0.80.

● Accuracy Creditworthy rate= TP/ actual yes ⇒ 101/105= 0.9619. Being a test with high True positives rate and high sensitivity, it means rarely fail diagnosis.

● In ROC curve and Gain chart, the Forest model's yellow lines excel from the rest of variables, because they have a constant grow to True positive rate axes and the left corner. Furthermore, the area under the curve (AUC) is the second most far from baseline and close to 1, meaning a high true positive rate.





- How many individuals are creditworthy?

According to the model score that included the whole population, there are **406** individuals Creditworthy and **94** Non-creditworthy. We obtained this number through the logic of contradiction: IF (Score_model_Creditworthy > Score_model_Non-Creditworthy, “Creditworthy”, “Non-Creditworthy”). Also, we ran and compared the performance of all models. Finally, we selected the Forest model such as a prudent solution to the business problem with the highest Accuracy value, furthermore, it is biased to Creditworthy.

Works cited

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

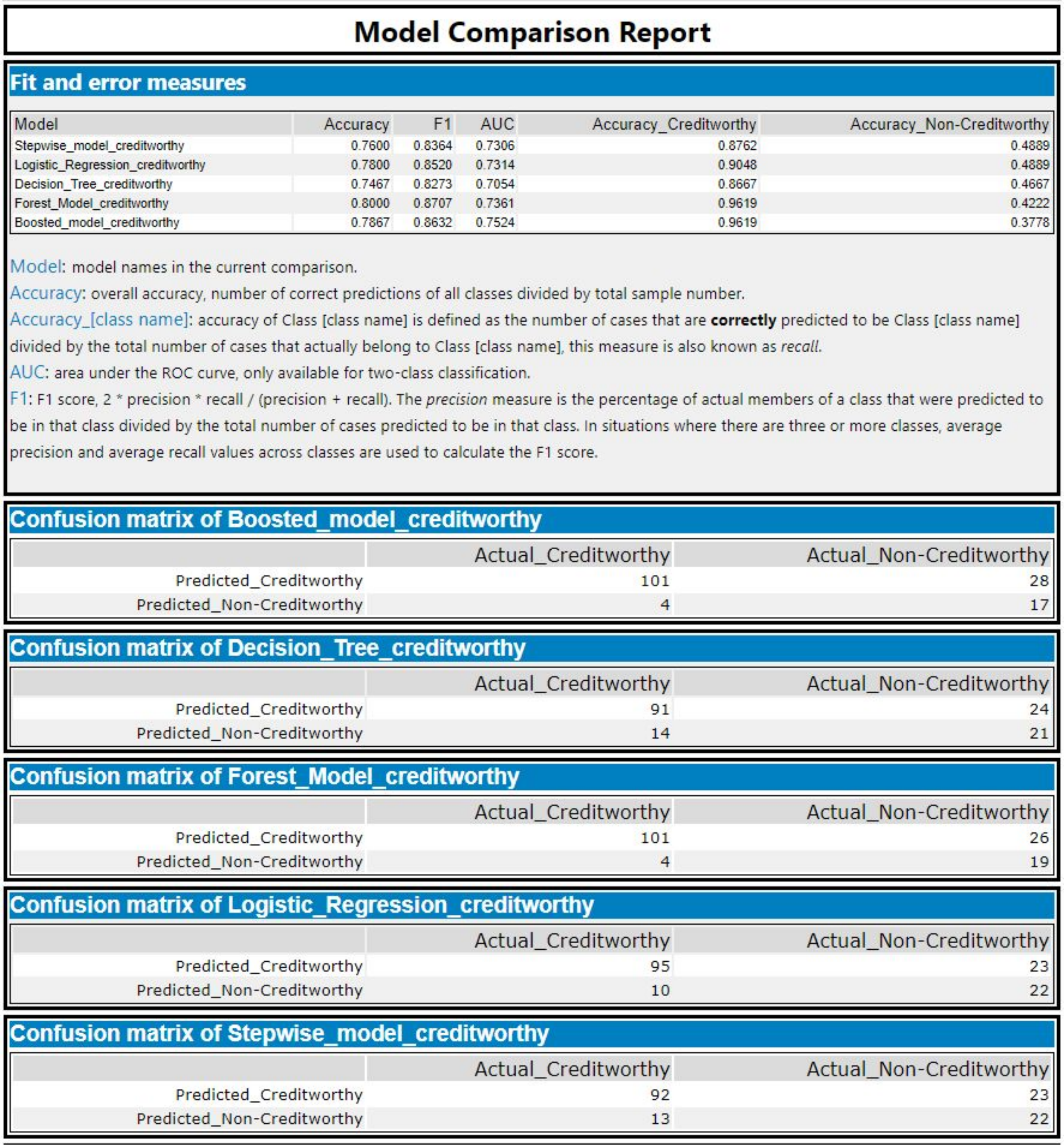
4 Reasons Your Machine Learning Model is Wrong (and How to Fix It)
https://www.kdnuggets.com/2016/12/4-reasons-machine-learning-model-wrong.html

BankThink Look Beyond Hard Numbers to Define Creditworthines
https://www.americanbanker.com/opinion/look-beyond-hard-numbers-to-define-creditworthiness

IF with a string and empty field
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/IF-with-a-string-and-empty-field/td-p/70769

Help!... Mean Decrease in Gini for dummies
https://community.alteryx.com/t5/Alteryx-Designer-Discussions/Help-Mean-Decrease-in-Gini-for-dummies/td-p/197223

Annexes



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_model_creditworthy	0.7600	0.8364	0.7306	0.8762	0.4889
Logistic_Regression_creditworthy	0.7800	0.8520	0.7314	0.9048	0.4889
Decision_Tree_creditworthy	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model_creditworthy	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_model_creditworthy	0.7867	0.8632	0.7524	0.9619	0.3778

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_model_creditworthy

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

Confusion matrix of Decision_Tree_creditworthy

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

Confusion matrix of Forest_Model_creditworthy

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

Confusion matrix of Logistic_Regression_creditworthy

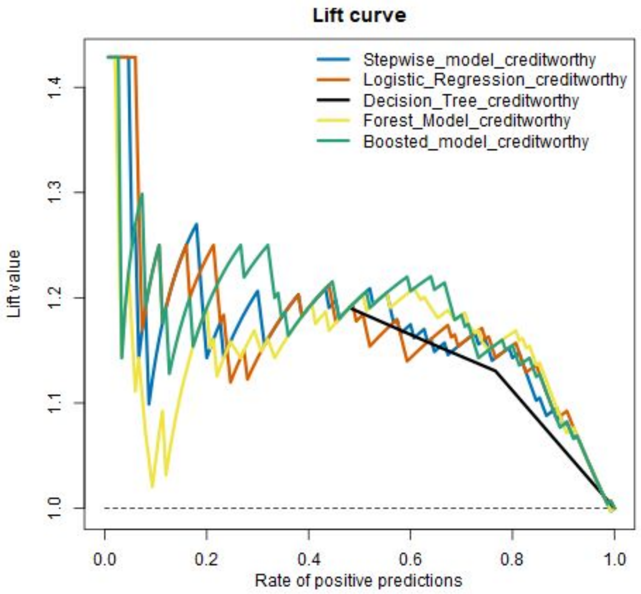
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

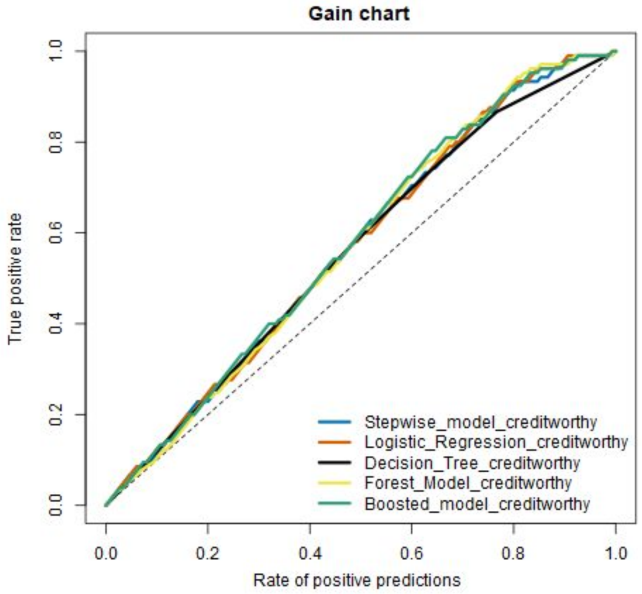
Confusion matrix of Stepwise_model_creditworthy

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

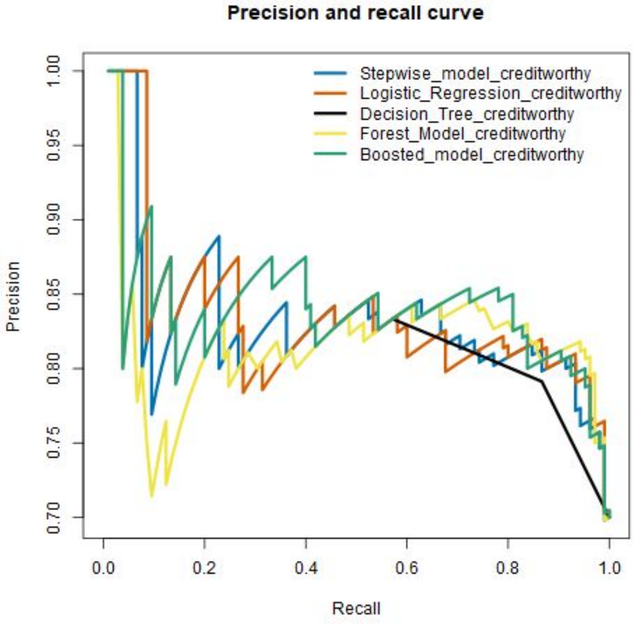
Continued on the next page

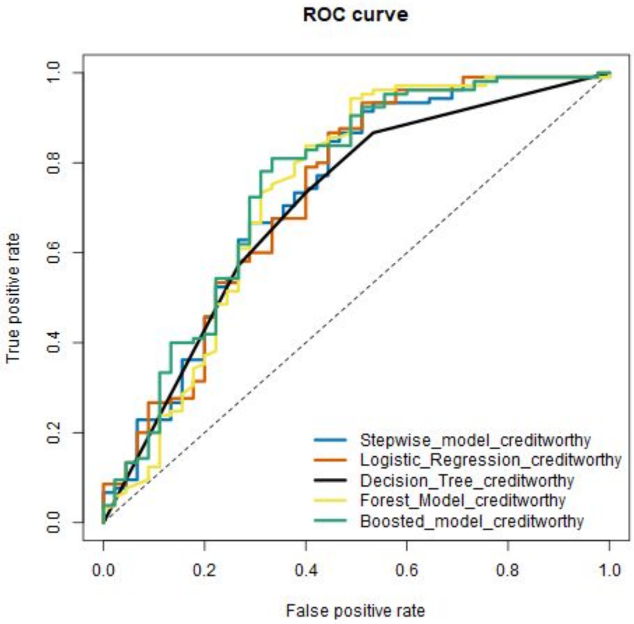
Performance Diagnostic Plots



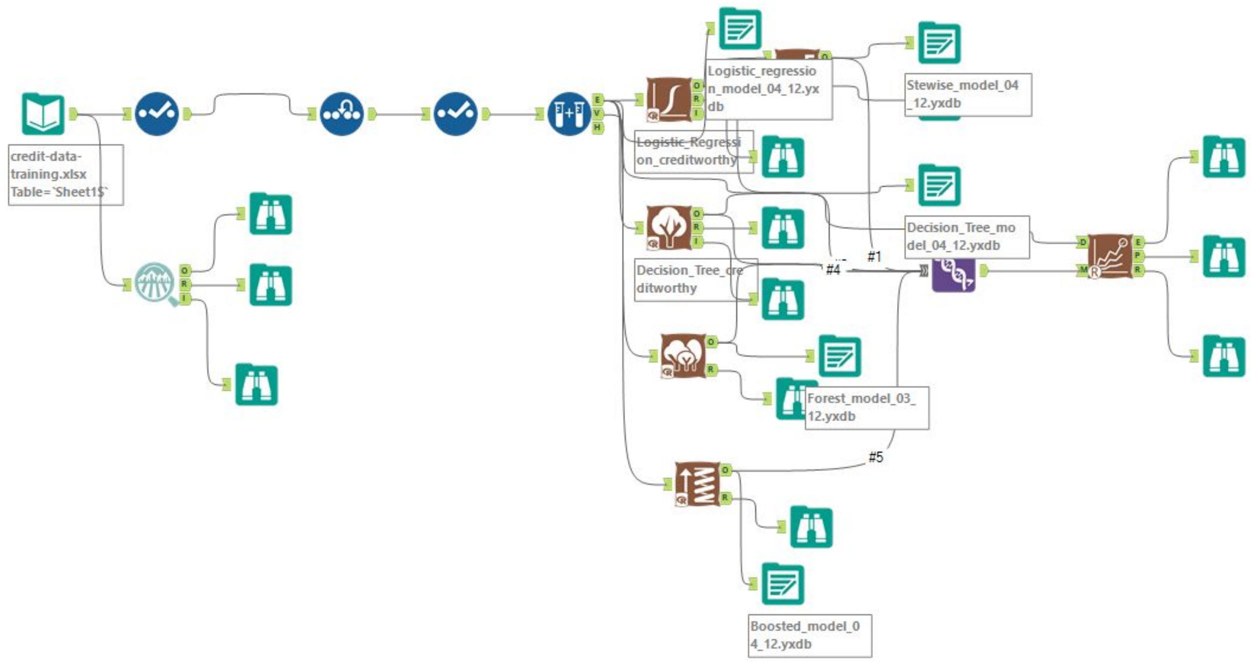


Continued on the next page

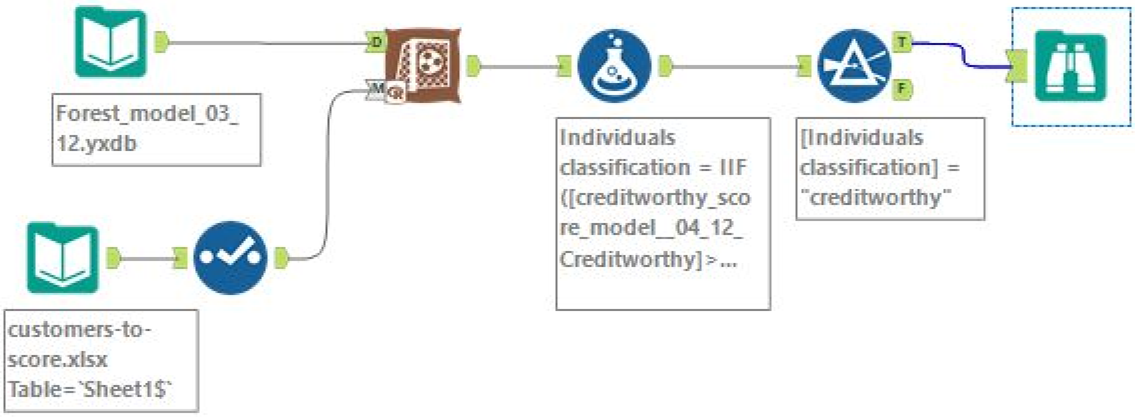

Precision and recall curve
Stepwise_model_creditworthy
Logistic_Regression_creditworthy
Decision_Tree_creditworthy
Forest_Model_creditworthy
Boosted_model_creditworthy
Precision
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.2
0.4
0.6
0.8
1.0
Recall


ROC curve
True positive rate
1.0
0.8
0.6
0.4
0.2
0.0
Stepwise_model_creditworthy
Logistic_Regression_creditworthy
Decision_Tree_creditworthy
Forest_Model_creditworthy
Boosted_model_creditworthy
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

creditworthy_traing_data 03_12.yxmd*


credit-data-training.xlsx Table=`Sheet1$`
O
R
I
E
V
H
Logistic_regression_model_04_12.yxdb
Stewise_model_04_12.yxdb
Logistic_Regression_creditworthy
Decision_Tree_model_04_12.yxdb
Decision_Tree_creditworthy
#1
#4
#5
D
M
E
P
R
Forest_model_03_12.yxdb
Boosted_model_04_12.yxdb

creditworthy_score_model_04_12.yxmd*


D
M
Forest_model_03_12.yxdb
customers-to-score.xlsx Table=`Sheet1$`
Individuals classification = IIF ([creditworthy_score_model__04_12_Creditworthy]>...
T
F
[Individuals classification] = "creditworthy"