Abdullah Ficici

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.

Key Decisions:

- What decisions need to be made?

I am a data scientist working for a small bank. This week the number of loan applications our bank received increased by 150% from 200/week to 500/week. I'll use Alteryx to process the loan applications. I will classify the loan applications as "Creditworthy" or "Not-Creditworthy".

- What data is needed to inform those decisions?

In order to predict the "Creditworthy" customers, I need the data on all past applications to build and train the model. The past application data has the following information: "**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", "Guarantors", "Duration-in-Current-address", "Most-valuable-available-asset", "Age-years", "Concurrent-Credits", "Type-of-apartment", "No-of-Credits-at-this-Bank", "Occupation", "No-of-dependents", "Telephone",** and **"Foreign-Worker"**.
After building the model, I will apply the model to new data on 500 new applications.

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

We are going to classify new customers as **"Creditworthy"** or **"Not-Creditworthy"**. This is a classification problem. Since we have two classes we can use Binary and/or Non-Binary Models.

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.***

- 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.
  "Duration-in-current-address" field has 69% missing data. This field will 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.

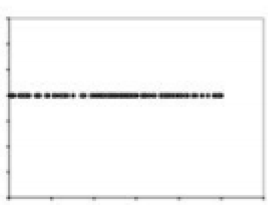
The Field summary report shows histograms and summaries, the red sign represents missing values, and the green color represents available values.
The field **“Duration-in-Current-address”** has 69% missing values, so it will be removed. The **“Age-years”** variable has 2% of missing values, they will be imputed with median 33.

“Foreign-Worker”, “Concurrent-Credits”, “Guarantors”, and “No_ of_dependents” fields will be removed due to low-variability.

The clean data set has 13 columns and the Average of "**Age-years**" is 36.

The **"Occupation"** field will be removed due to uniform data.

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 does not contribute to the target variable, it will be removed.


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

- 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?

1- Logistic Regression Model

Logistic Regression Model classified 17 variables. The following predictor variables are significant with low P-values: "**Account.BalanceSome Balance", 'Purpose", "Credit.Amount", "Length.of.current.employment", "Instalment.per.cent",** and **"Most.valuable.available.asset"**

Report for Logistic Regression Model loan_logit_Reg

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 + Age.years + No.of.Credits.at.this.Bank + Telephone, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.094	-0.734	-0.424	0.762	2.547

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.6041138	1.036e+00	-3.4786	5e-04 ***
Account.BalanceSome Balance	-1.6152718	3.229e-01	-5.0016	5.68e-07 ***
Duration.of.Credit.Month	0.0072250	1.369e-02	0.5276	0.59777
Payment.Status.of.Previous.CreditPaid Up	0.4475591	3.863e-01	1.1587	0.24658
Payment.Status.of.Previous.CreditSome Problems	1.3374204	5.356e-01	2.4972	0.01252 *
PurposeNew car	-1.7349564	6.274e-01	-2.7654	0.00569 **
PurposeOther	-0.1926841	8.355e-01	-0.2306	0.8176
PurposeUsed car	-0.7804912	4.126e-01	-1.8915	0.05856.
Credit.Amount	0.0001507	7.096e-05	2.1240	0.03367 *
Value.Savings.StocksNone	0.6188301	5.067e-01	1.2213	0.22199
Value.Savings.StocksÂ£100-Â£1000	0.1726049	5.623e-01	0.3070	0.75887
Length.of.current.employment4-7 yrs	0.5313580	4.916e-01	1.0809	0.27973
Length.of.current.employment< 1yr	0.8040089	3.939e-01	2.0411	0.04124 *
Instalment.per.cent	0.2882110	1.393e-01	2.0683	0.03861 *
Most.valuable.available.asset	0.2671762	1.498e-01	1.7840	0.07442.
Age.years	-0.0199363	1.491e-02	-1.3375	0.18107
No.of.Credits.at.this.BankMore than 1	0.3897906	3.826e-01	1.0188	0.30828
Telephone	0.3786710	3.138e-01	1.2068	0.22752

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
loan_logit_Reg	0.7867	0.8559	0.7244	0.9048	0.5111

The Logistic Regression Model is biased to Creditworthy with high True positive values. It is confirmed by the ROC curve and Gain chart are going to True positive rate.
Logistic Regression true positive rate: TP/ actual yes = 95/105 ⇒ 0.9048 is the second highest Accuracy_creditworthy value.
Logistic Regression Models' false positive value is 23, is the lowest among the rest of the models.

Confusion matrix of loan_logit_Reg

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

2- Decision Tree Model

Root node error: 97/350 = 0.27714. Approximately 28% of the values went to the incorrect terminal node.

Decision Tree Model has the second lowest Accuracy value (0.7467) among all 4 models.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8304	0.7035	0.8857	0.4222

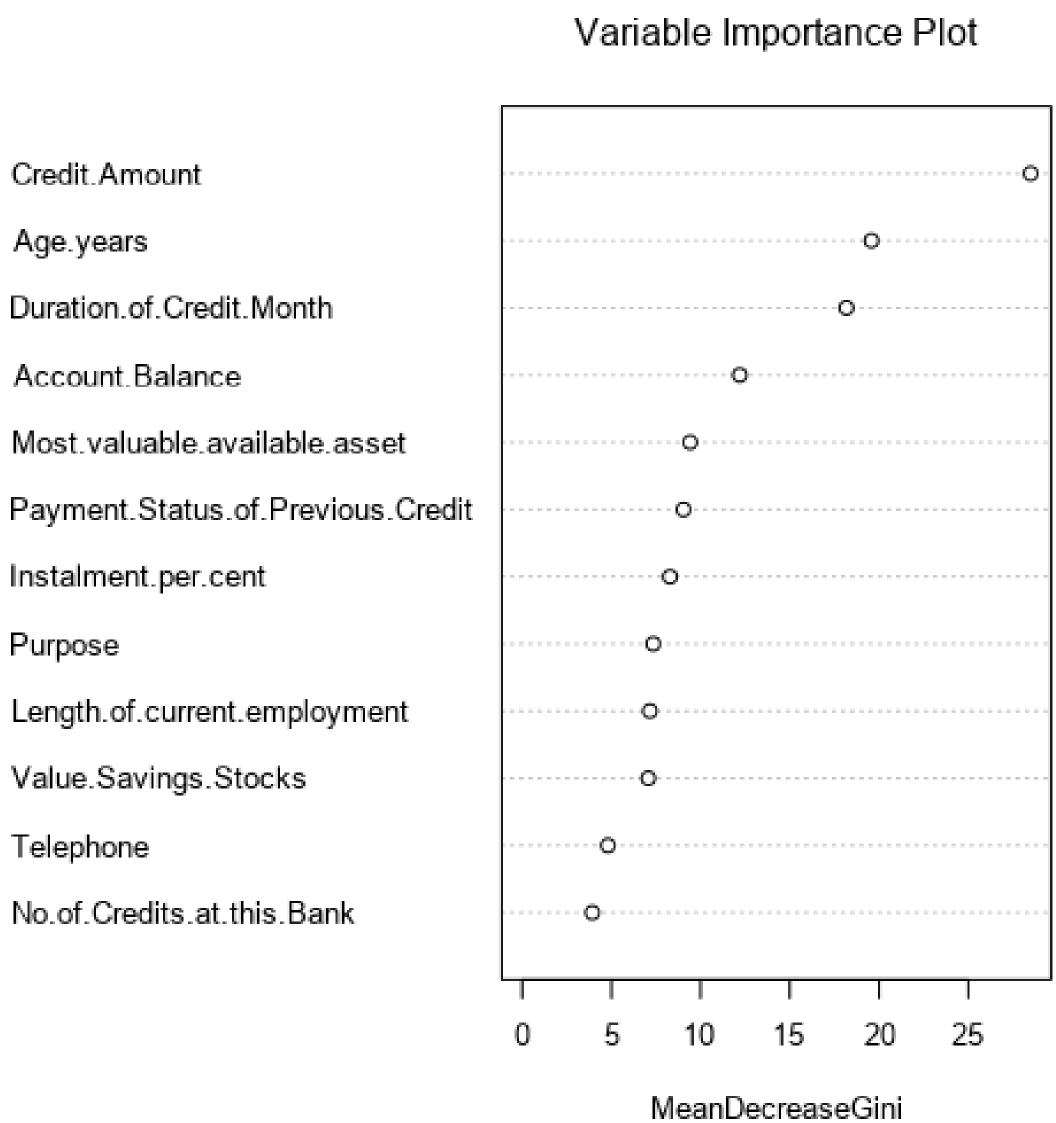
Decision Tree model true positive rate: TP/Actual Positive = 93/105 ⇒ 0.8857. ROC curve and Gain chart shows the model goes to the left corner, nonetheless, the black lines are close to the baseline, being this a reason for low accuracy. AUC value confirms that the Decision Trees line is far to number 1 and closer to the baseline, it means low accuracy

Confusion matrix of DT_Credit

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

3- Forest Model

Variable Importance Plot Shows top 4 variables with large MeanDecreaseGini Values: **Credit amount, Age years, Duration of credit month**, and **Account Balance**.



The Forest Model has the **highest Accuracy value 0.81** of all models.

Model Comparison Report

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit	0.8133	0.8803	0.7376	0.9810	0.4222

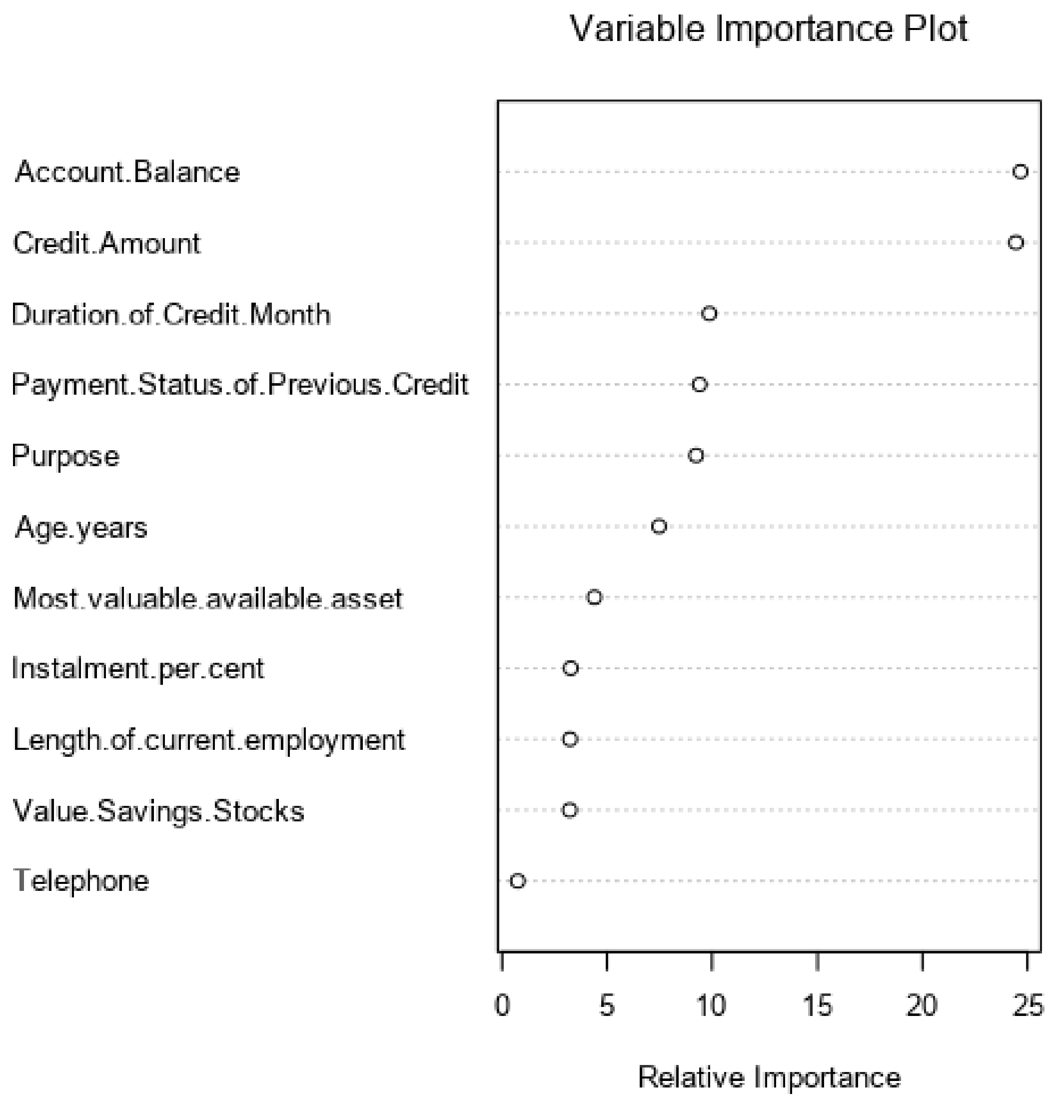
Confusion matrix shows the highest True positive value, and the lowest false negative value among all models. The matrix: 103/105= 0.9810 predicts the Creditworthy.

Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	26
Predicted_Non-Creditworthy	2	19

4- Boosted Model

Variable Importance Plot Shows top 2 variables with large MeanDecreaseGini Values: **Account Balance**, and **Credit amount.**


Variable Importance Plot
Account.Balance
Credit.Amount
Duration.of.Credit.Month
Payment.Status.of.Previous.Credit
Purpose
Age.years
Most.valuable.available.asset
Instalment.per.cent
Length.of.current.employment
Value.Savings.Stocks
Telephone
0
5
10
15
20
25
Relative Importance

Boosted Model has the **second highest Accuracy value 0.7933** of all models.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Credit	0.7933	0.8670	0.7473	0.9619	0.4000

Confusion matrix shows the high True positive value 101. The matrix⇒ 101/105= 0.9619 predicts the Creditworthy

Confusion matrix of Boosted_Credit

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

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"

- 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

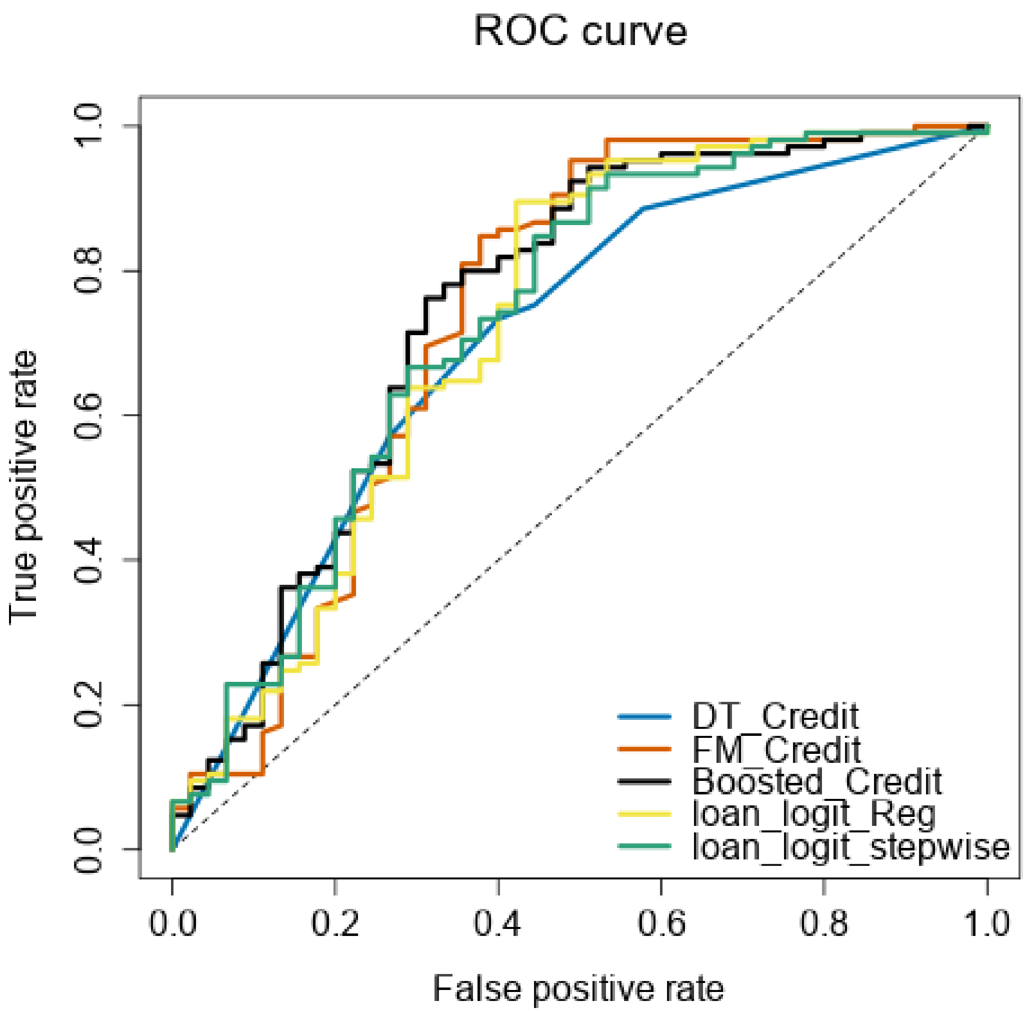
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8304	0.7035	0.8857	0.4222
FM_Credit	0.8133	0.8803	0.7376	0.9810	0.4222
Boosted_Credit	0.7933	0.8670	0.7473	0.9619	0.4000
loan_logit_Reg	0.7867	0.8559	0.7244	0.9048	0.5111
loan_logit_stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Model Comparison Report shows all models are biased to Creditworthy. The **Forest Model** is the best model with the highest overall Accuracy value of 0.8133.

- The Forest Model's Accuracy value is 0.8133.
- Accuracy_Creditworthy rate= TP/ actual yes = 103/105= 0.9810.
- In ROC curve, the Forest model performs better than 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 all 500 new applicants, there are **407** Creditworthy and **93** Non-Creditworthy applicants.