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)

As a loan officer I need to identify which new customers can be approved for loan or not.

Key Decisions:

Answer these questions

- What decisions needs to be made?
  Identification of new customers who are eligible for loan or not.
  Need an efficient system to classify customers as credit worthiness and non-credit worthiness.

- What data is needed to inform those decisions?
  a. All past data of customers are required who were eligible for loan such as account balance, duration of credit month, purpose, credit amount, length of current employment etc.
  b. List of all new customers.

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

  Binary Model.

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 average 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

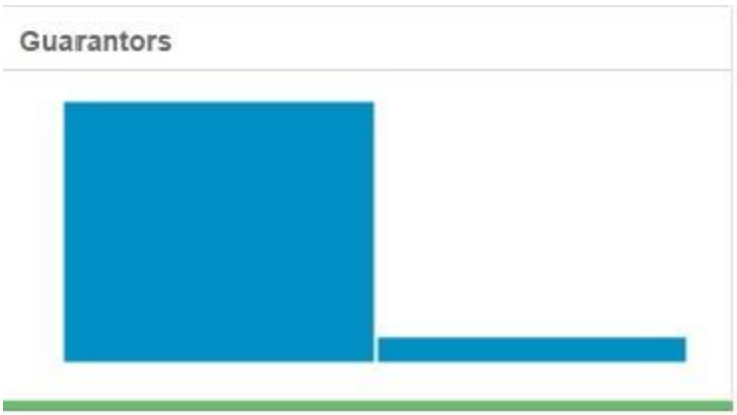
To achieve consistent results reviewers expect.

Answer this question:

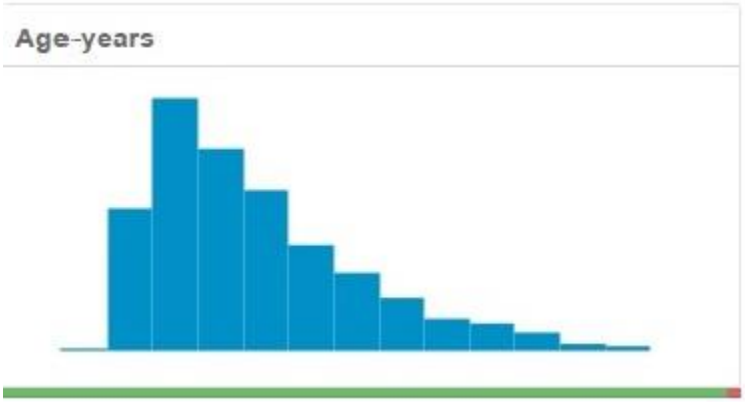
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

  For the purpose of this work, we need to remove 7 variables and keep 13 variables. The fields with lot of missing data was removed, for example duration in current address is one such data field.

  The data field with low variability was removed.



Gurantors field for example is a low variability field.



The Age-Years was imputed with its median value since the data is skewed and new column was generated as Age_Years since the missing data was very less which was around 2.4%.

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.

Model Summary

Logistic Regression

A stepwise logistic model was run, and from the table below it can be concluded that these are the most important variable based on the significance value.

1. Account balance
2. Purpose
3. Credit Amount

	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

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

Reports are attached as separate sheet (Report LR)

Decision Tree

A decision tree model was run and through the table of model summary it can be found that

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age_years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

1. Account Balance
2. Age_years
3. Credit Amount

Report is attached as separate sheet (Report DT)

Forest Model

1. Credit Amount
2. Age_years
3. Duration of credit month

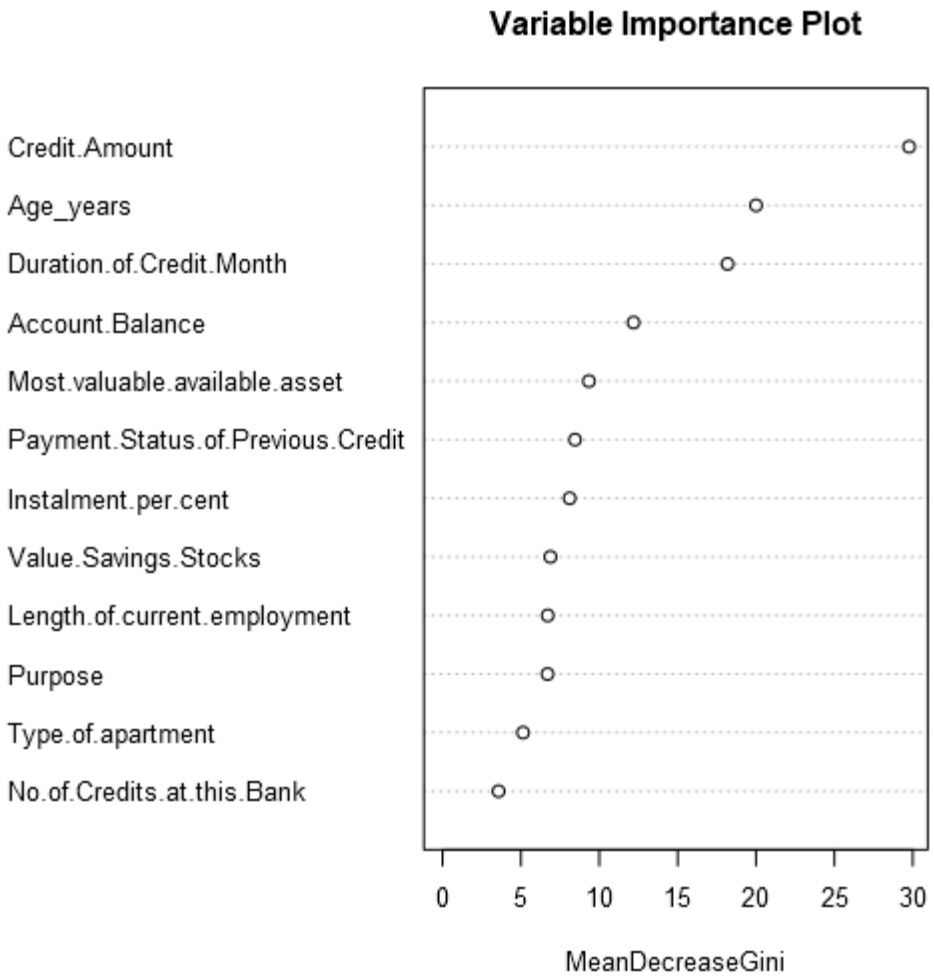
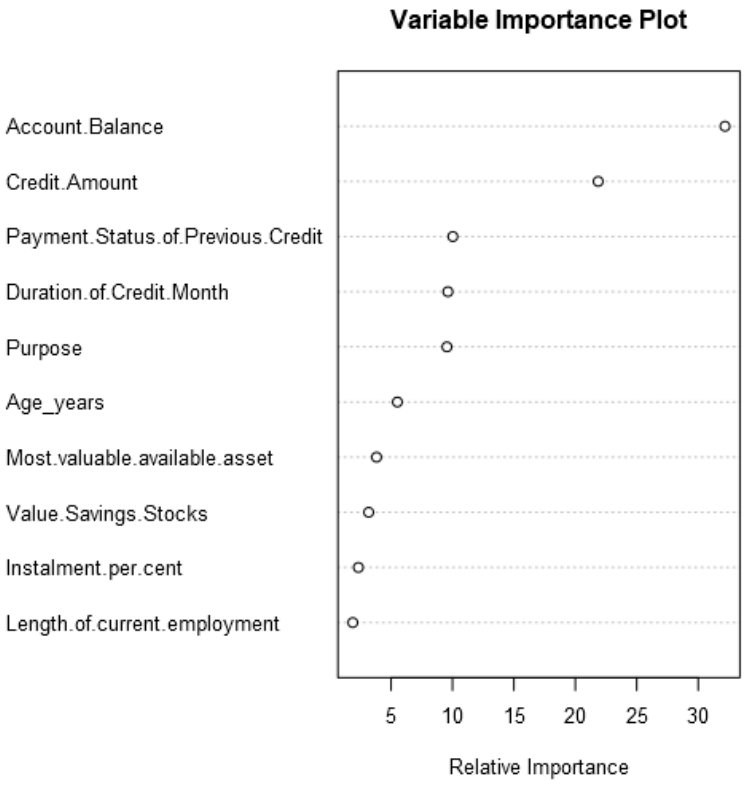


Fig. 2 Variable importance graph for Forest model

Report attached as separate sheet (Report FM)

Boosted Model

1. Account Balance
2. Credit Amount
3. Payment status of previous credit



Report attached as separate sheet (Report BM)

- 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?

Logit Model	0.78
Decision Tree Model	0.67
Forest Model	0.80
Boosted Model	0.78

Confusion matrix

Confusion matrix of BM_Credit

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

Confusion matrix of DT_Credit_worthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of FM_Credit_worthy

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

Confusion matrix of Logit_Credit_worthy

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

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit_worthy	0.6733	0.7721	0.6296	0.7545	0.4500
Logit_Credit_worthy	0.7800	0.8520	0.7314	0.8051	0.6875
FM_Credit_worthy	0.8000	0.8707	0.7419	0.7953	0.8261
BM_Credit	0.7867	0.8632	0.7524	0.7829	0.8095

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

When we look at Table Model comparison report as well as confusion matrix we can understand that Model DT (Decision Tree) has bias of correctly predicting creditworthy individuals of around 75% and 45% accuracy towards Non-creditworthy. Similarly, Logic regression model's accuracy in correctly predicting creditworthy individual is 81% and 69% towards non-creditworthy, which signifies that the model is bias towards correctly predicting creditworthy individual. Now, Forest model's accuracy in correctly predicting creditworthy individual is 80% and 82% towards non-creditworthy. Similarly, Boosted model's accuracy in correctly predicting the creditworthy

individual is 78% and 81% towards non-creditworthy individual. Both the models (forest model and boosted model) are almost not biased at all, because the difference between the accuracies is very small.

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

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

Forest model is the best model out of all the model. The Accuracy is 0.8 which is greater than all other. From the table we can see that accuracies within "Creditworthy" are 0.80 and "Non-Credit worthy" is 0.83 which is far better than other models. Forest model's accuracy in correctly predicting creditworthy individual is 80% and 82% towards non-creditworthy. The model is almost not biased at all, because the difference between the accuracies is very small.

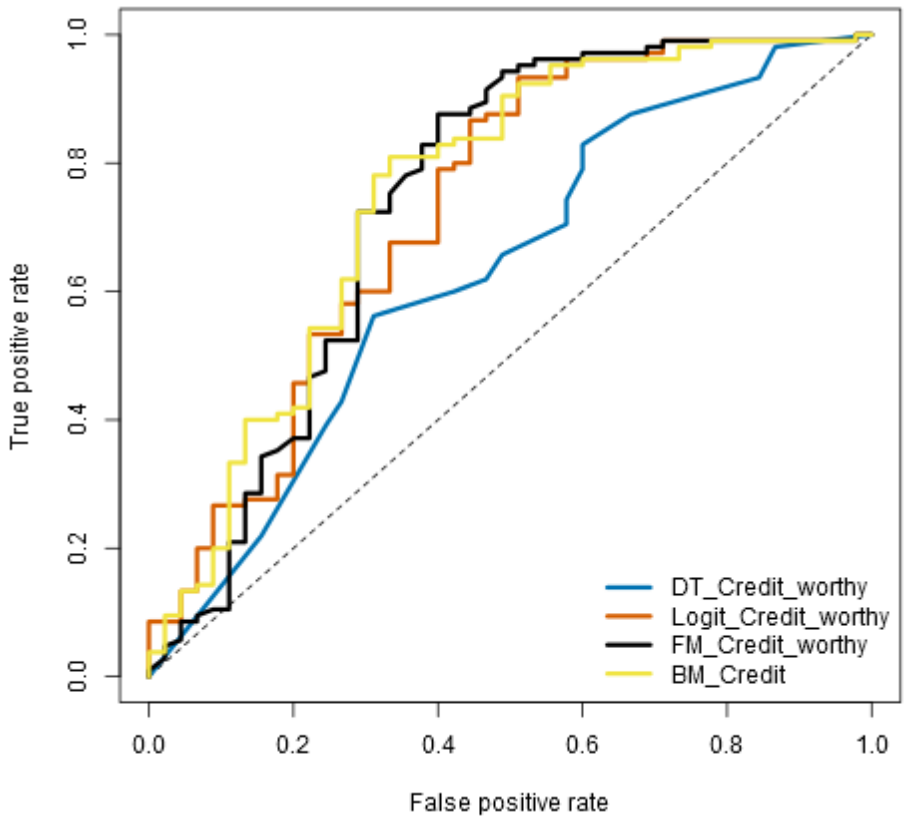


Fig. 3 ROC curve

Fig. 3 above shows that FM gives more of true positive result.

Confusion matrix of FM_Credit_worthy

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

So the actual creditworthy is around 96%.,which gives very less room for bias.

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

- How many individuals are creditworthy?
  415 individuals are creditworthy.

Models which have bias towards credit worthy or non-creditworthy for business implication is not good since that model will always have the tendency to predict more false positive values and less of false negative values.

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.

References

Udacity forums for understanding some issue.