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? Since there has been a financial scandal in the city, we as a small bank, have an influx of 500 loan applications to process. The bank's manager would like to evaluate the creditworthiness of the applicants to seize this huge opportunity. The most important decision to be made is to predict whether the applicants are going to default on their loan or not.

- What data is needed to inform those decisions? To accurately predict the creditworthiness of the applications, we need data on all credit approvals from the past loans that our bank has processed so far. A few important variables could be but not limited to the followings:

  - account balance
  - purpose of the load
  - credit amount
  - most valuable asset
  - employment status
  - number of dependent

  We also need a list of new applicants to use to score our model and find out whether they are creditworthy.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions? Our data consists of Continuous, binary, and Non-Binary variables for the past loans. However, our desired outcome (Creditworthiness) is a binary one due to only 2 possible outcomes (Creditworthy or not creditworthy). Therefore it determines the model type that we are dealing with. Classification models such as Logistic Regression, Decision Tree, Forest Model, and Boosted Model are the ones that we need to implement in this project.

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:*

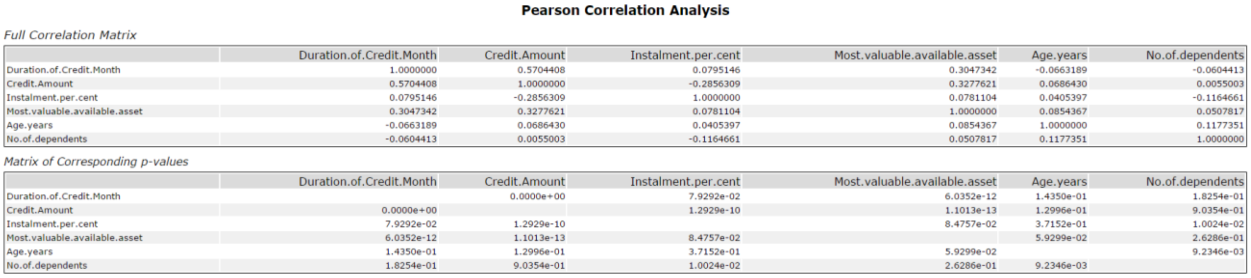
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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:

1. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
   Firstly, I used association analysis to better understand whether there are numerical variables that are highly correlated. The results of Pearson Correlation Analysis is presented below. According to this analysis, there are no highly correlated coefficients. The most highly correlated pair of variables is for the pair of ("Duration.of.Credit.Month" and "Credit.Amount") with a 0.57 value which is still less than 0.7 to be considered highly correlated.



Pearson Correlation Analysis

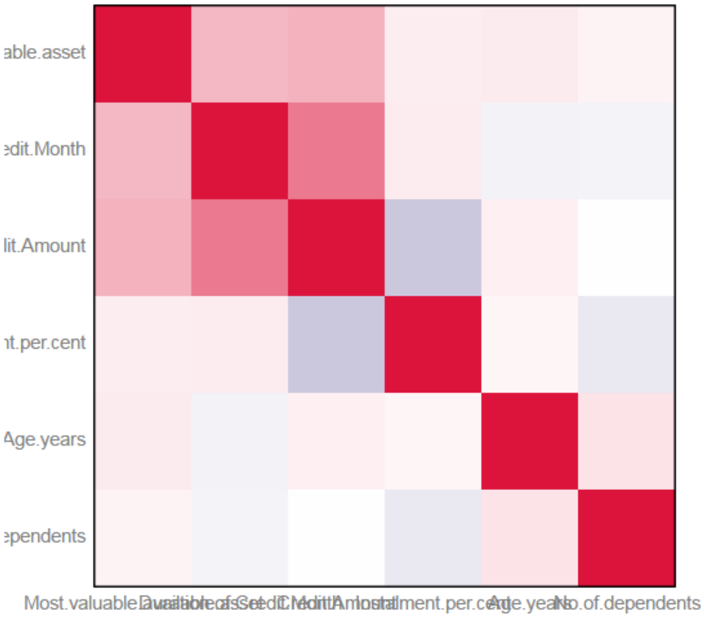
Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	No.of.dependents
Duration.of.Credit.Month	1.0000000	0.5704408	0.0795146	0.3047342	-0.0663189	-0.0604413
Credit.Amount	0.5704408	1.0000000	-0.2856309	0.3277621	0.0686430	0.0055003
Instalment.per.cent	0.0795146	-0.2856309	1.0000000	0.0781104	0.0405397	-0.1164661
Most.valuable.available.asset	0.3047342	0.3277621	0.0781104	1.0000000	0.0854367	0.0507817
Age.years	-0.0663189	0.0686430	0.0405397	0.0854367	1.0000000	0.1177351
No.of.dependents	-0.0604413	0.0055003	-0.1164661	0.0507817	0.1177351	1.0000000

Matrix of Corresponding p-values

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	No.of.dependents
Duration.of.Credit.Month		0.0000e+00	7.9292e-02	6.0352e-12	1.4350e-01	1.8254e-01
Credit.Amount	0.0000e+00		1.2929e-10	1.1013e-13	1.2996e-01	9.0354e-01
Instalment.per.cent	7.9292e-02	1.2929e-10		8.4757e-02	3.7152e-01	1.0024e-02
Most.valuable.available.asset	6.0352e-12	1.1013e-13	8.4757e-02		5.9299e-02	2.6286e-01
Age.years	1.4350e-01	1.2996e-01	3.7152e-01	5.9299e-02		9.2346e-03
No.of.dependents	1.8254e-01	9.0354e-01	1.0024e-02	2.6286e-01	9.2346e-03	

Correlation Matrix with ScatterPlot



In the data cleaning process, I removed the following fields:

a. Duration_in_Current_Address: it has many missing data (69% missing data)
b. Cuncurrent_Credits: It has only one value (low variability), so I have removed it.

c. Occupation: It has only one value (low variability), so I have removed it.
d. Foreign_Workers: It also has a very low variability and the majority of the data was skewed towards “1”.
e. Guarantors: It also has a very low variability and the majority of the data was skewed towards “None”.
f. No_of_Dependents: It has low variability and most of the data was skewed towards 1.
g. Telephone: there is no logical reason for including the variable.
h. There is 2% missing data in “Age_years” field. We decided to impute the Null values with the median value of the field. The median Value is 33 and we replaced the Null values with 33.

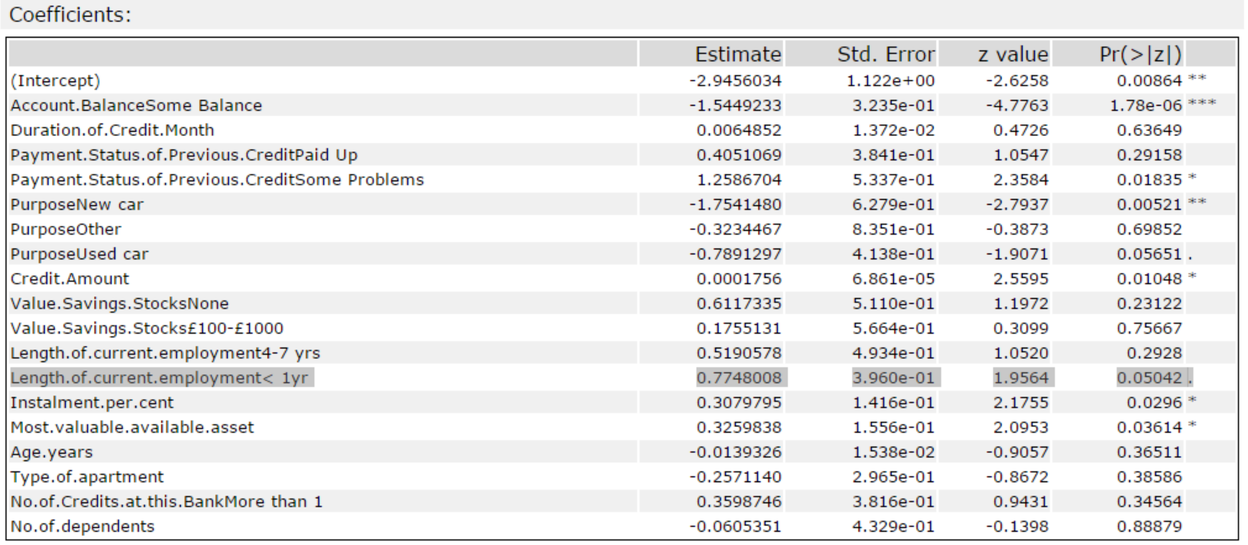
Step 3: Train your Classification Models

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:*

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
   Here I have brought the predictor variables for each Model followed by the P-Value and Variable Importance chart:
   1. Logistic Regression Model. The model report output is attached below:



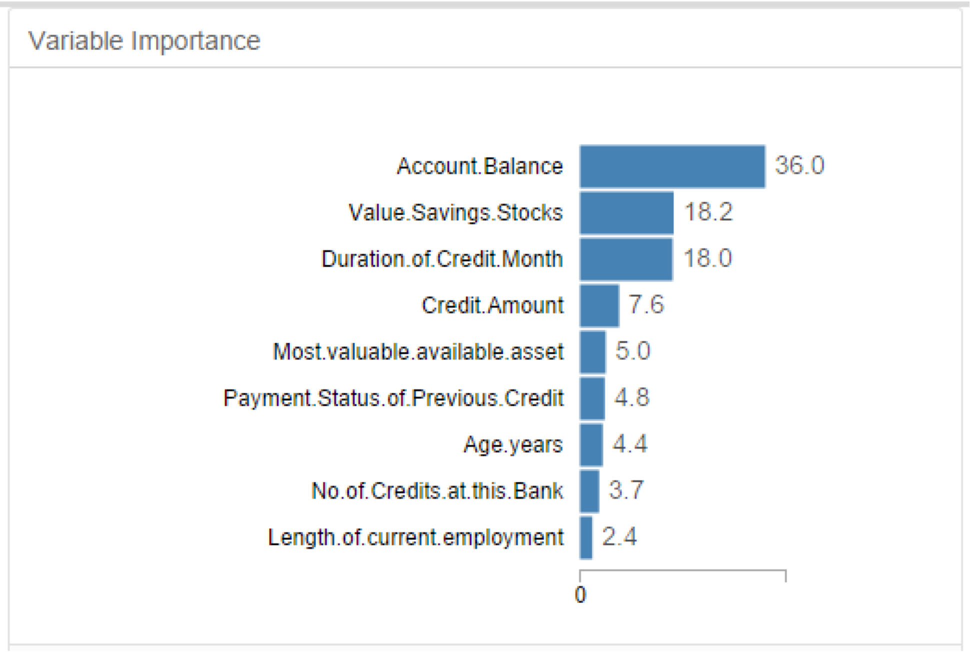
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9456034	1.122e+00	-2.6258	0.00864	**
Account.BalanceSome Balance	-1.5449233	3.235e-01	-4.7763	1.78e-06	***
Duration.of.Credit.Month	0.0064852	1.372e-02	0.4726	0.63649	
Payment.Status.of.Previous.CreditPaid Up	0.4051069	3.841e-01	1.0547	0.29158	
Payment.Status.of.Previous.CreditSome Problems	1.2586704	5.337e-01	2.3584	0.01835	*
PurposeNew car	-1.7541480	6.279e-01	-2.7937	0.00521	**
PurposeOther	-0.3234467	8.351e-01	-0.3873	0.69852	
PurposeUsed car	-0.7891297	4.138e-01	-1.9071	0.05651	.
Credit.Amount	0.0001756	6.861e-05	2.5595	0.01048	*
Value.Savings.StocksNone	0.6117335	5.110e-01	1.1972	0.23122	
Value.Savings.Stocks£100-£1000	0.1755131	5.664e-01	0.3099	0.75667	
Length.of.current.employment4-7 yrs	0.5190578	4.934e-01	1.0520	0.2928	
Length.of.current.employment< 1yr	0.7748008	3.960e-01	1.9564	0.05042	.
Instalment.per.cent	0.3079795	1.416e-01	2.1755	0.0296	*
Most.valuable.available.asset	0.3259838	1.556e-01	2.0953	0.03614	*
Age.years	-0.0139326	1.538e-02	-0.9057	0.36511	
Type.of.apartment	-0.2571140	2.965e-01	-0.8672	0.38586	
No.of.Credits.at.this.BankMore than 1	0.3598746	3.816e-01	0.9431	0.34564	
No.of.dependents	-0.0605351	4.329e-01	-0.1398	0.88879	

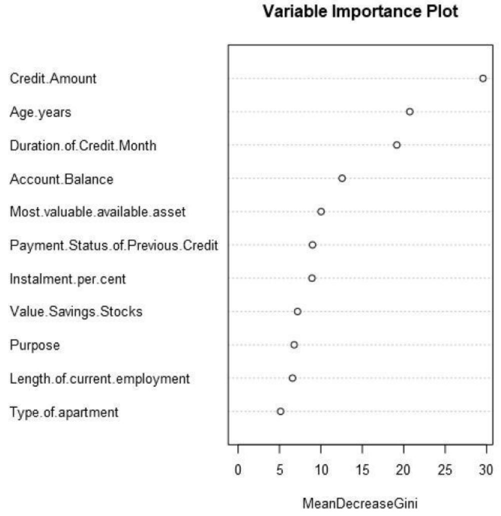
The significant variables are:

a. Account balance
b. Duration of Credit Month
c. Payment Status of previous, (Credit Some Problems is the most significant among the variations)
d. Purpose (New Car, Used Car are the most significant ones among the variations)
e. Credit Amount
f. Length of current employment (<1yr is significant)
g. Instalment Per Cent
h. Most valuable available asset

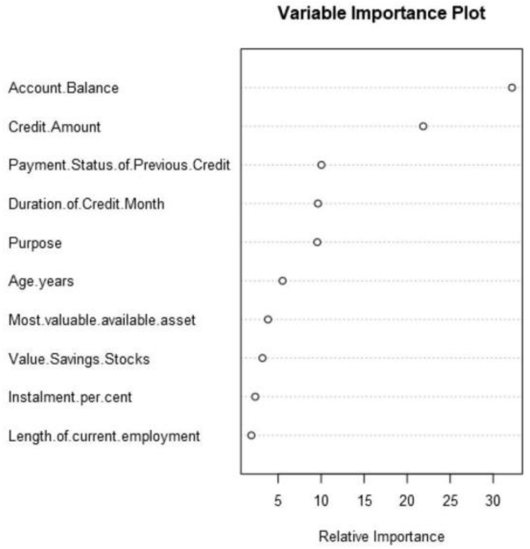
2. Decision Tree Model. The model report output is attached below:



3. Forest Model. Variable Importance Plot is attached below:

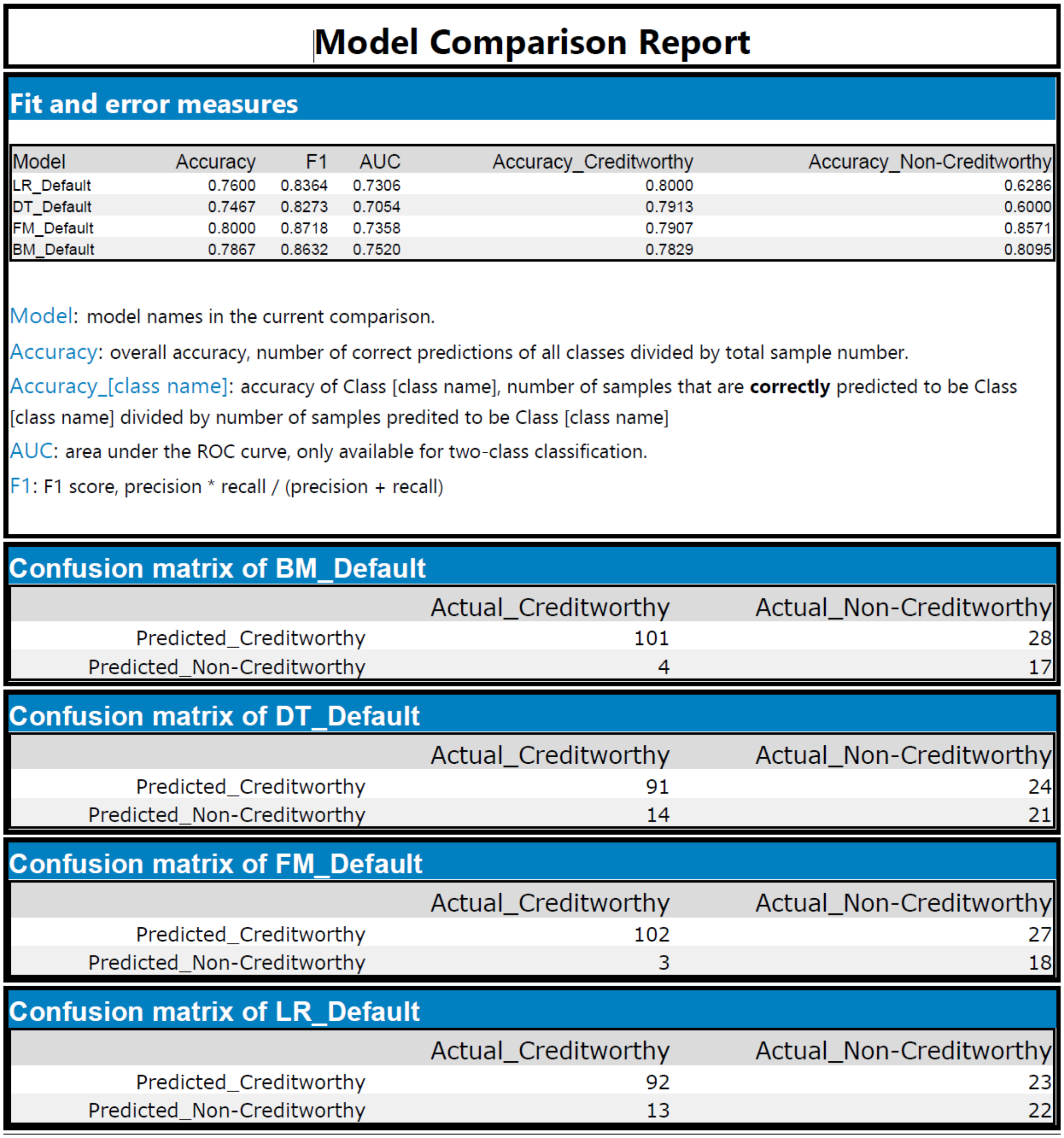


4. Boosted Model. Variable Importance Plot is attached below:



2. 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?
   The results for all the models are brought here. As you can see, the overall accuracy for the models are as below:
   - Forest Model: 0.8. For this model we see a Bias towards Creditworthy as it produced more Creditworthy that it should
   - Boosted Model: 0.7867
   - Logistic Regression Model: 0.76
   - Decision Tree Model: 0.7467

   Considering the confusion matrices for the all four models, it seems that there are biases toward Creditworthiness prediction, because they all have predicted more "Creditworthy" than the actual number of creditworthy individuals.
   In our original dataset, there are 105 Actual creditworthy individuals and only 45 non-creditworthy ones. In summary, Boosted model has predicted (129 Creditworthy and 21 Non-Creditworthy), Decision Tree model has predicted (115 Creditworthy and 35 Non-Creditworthy), Forest model has predicted (129 Creditworthy and 21 Non-Creditworthy), and finally, Logistic Regression model has predicted (115 Creditworthy and 35 Non-Creditworthy). Therefore, it is evident that all four models have overestimated the Creditworthiness and underestimated the number of non-creditworthy individuals. There appears to be a bias toward Creditworthiness.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Default	0.7600	0.8364	0.7306	0.8000	0.6286
DT_Default	0.7467	0.8273	0.7054	0.7913	0.6000
FM_Default	0.8000	0.8718	0.7358	0.7907	0.8571
BM_Default	0.7867	0.8632	0.7520	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)

Confusion matrix of BM_Default

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

Confusion matrix of DT_Default

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

Confusion matrix of FM_Default

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of LR_Default

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

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

Step 4: Writeup

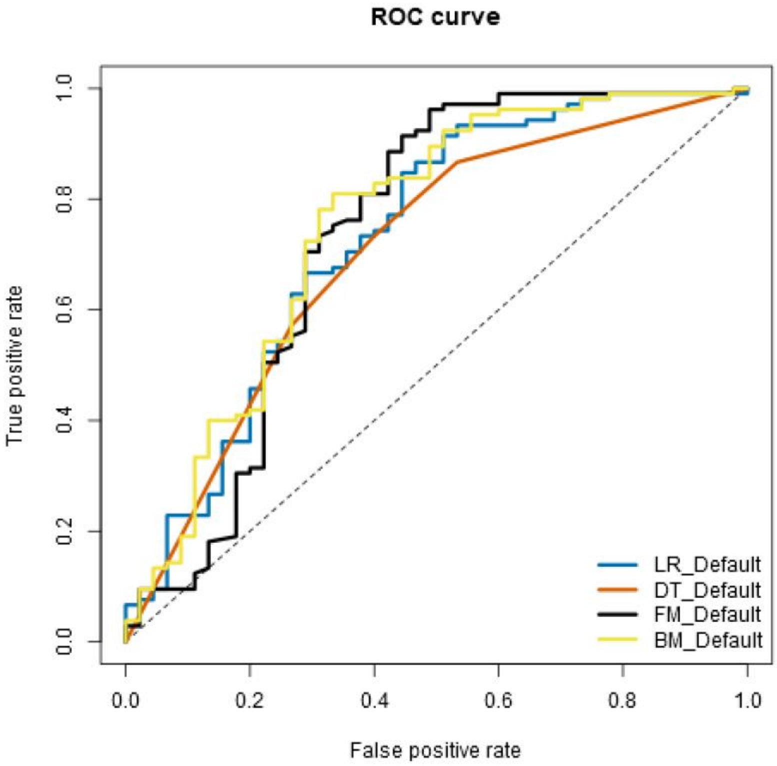
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as “Creditworthy”

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)

For the missing data in “Age_Year” field, I used the median value of the filed. After the data was ready and cleaned, I stored 30% of data as Validation data and the remaining 70% as the Estimation data. I then, created 4 different models using Logistic regression, Decision tree, Forest Model, and Boosted Model. After that, we have used a model comparison tool to compare the results from the various models side by side and see which one has produced the most accurate predictions. According to the table presented earlier, the Forest Model has the best Overall Accuracy as well as the best Non-Creditworthy accuracy. However, it is has shown a slightly less accurate result in terms of Creditworthy accuracy compared to Logistic regression and Decision Tree models with an accuracy of 0.8 and 0.7913 respectively. Overall, Forest model is chosen as the best predicting model to score the data.

Answer these questions:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set: Our chosen model which was Forest Model showed an overall accuracy of 0.8 which is the best compared to the other 3. The result for the other 3 models were (0.7867 for Boosted Model), (0.7467 for Decision Tree Model), (0.76 for Logistic Regression Model).
    b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments: The accuracy within “Creditworthy segment was 0.7907 with 102 correctly predicted and only 3 incorrect predictions. The accuracy result was the third best after Logistic Regression model and Decision Tree model with Creditworthy accuracy of 0.8 and 0.7913 respectively.
    Accuracy within “Non-Creditworthy” segment for Forest model (our chosen model) was 0.8571 which was the highest among the 4 models, followed by Boosted model with an accuracy of 0.8095. The chosen model could predict 27 Non-Creditworthy applicants and only 18 incorrect predictions.
    c. ROC graph: A model with high discrimination ability will have high sensitivity and specificity simultaneously, leading to an ROC curve which goes close to the top left corner of the plot. A model with no discrimination ability will have an ROC curve which is the 45-degree diagonal line. According to the following graph of ROC (Receiver Operating Characteristic) and the values presented as AUC (Area Under Curve of ROC) we see that our chosen model (Forest model) has the second highest value (0.7358) after Boosted model with a AUC of 0.7520. The AUC gives the probability that the model correctly predicts Positive rates.



d. Bias in the Confusion Matrices: The accuracy within “Creditworthy segment was 0.7909 with 102 correctly predicted and only 3 incorrect predictions. Based on the results shown above, the model seems to be a little biased towards Creditworthy. For Non-Creditworthy, it has 27 correctly predicted results and 18 incorrectly predicted.

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

2. How many individuals are creditworthy? 415 individuals were predicted as “Creditworthy” using our Forest Model.

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.