Project: Creditworthiness

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

1. What decisions needs to be made?
   Which customers are creditworthy to give a loan to?

2. What data is needed to inform those decisions?
   - Data on past applications, including customers' information and loan default status.
   - Data on customers that need to be processed

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

Step 2: Building the Training Set

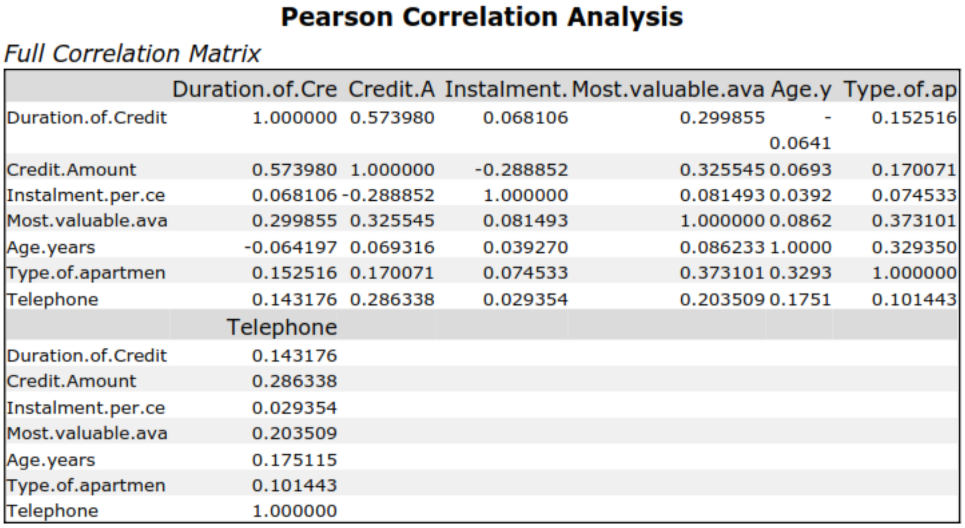
*Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't **need to convert any data fields to the appropriate data types.***

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)*

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.



Pearson Correlation Analysis

Full Correlation Matrix

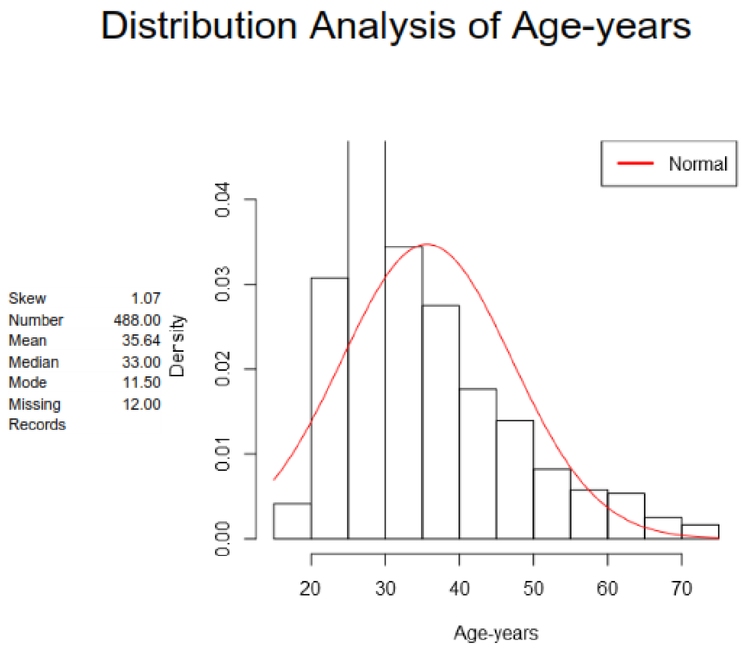
	Duration.of.Cre	Credit.A	Instalment.	Most.valuable.ava	Age.y	Type.of.ap
Duration.of.Credit	1.000000	0.573980	0.068106	0.299855	- 0.0641	0.152516
Credit.Amount	0.573980	1.000000	-0.288852	0.325545	0.0693	0.170071
Instalment.per.ce	0.068106	-0.288852	1.000000	0.081493	0.0392	0.074533
Most.valuable.ava	0.299855	0.325545	0.081493	1.000000	0.0862	0.373101
Age.years	-0.064197	0.069316	0.039270	0.086233	1.0000	0.329350
Type.of.apartmen	0.152516	0.170071	0.074533	0.373101	0.3293	1.000000
Telephone	0.143176	0.286338	0.029354	0.203509	0.1751	0.101443

	Telephone
Duration.of.Credit	0.143176
Credit.Amount	0.286338
Instalment.per.ce	0.029354
Most.valuable.ava	0.203509
Age.years	0.175115
Type.of.apartmen	0.101443
Telephone	1.000000

These 2 fields contain null values

1. Age-years (2% missing)
2. Duration-in-Current-address (69% missing).

I decided to remove Duration-in-Current-address. Then, I checked distribution of Age-years. The histogram is right skewed with skewness value of 1.07. So, I decided to impute null with median age.

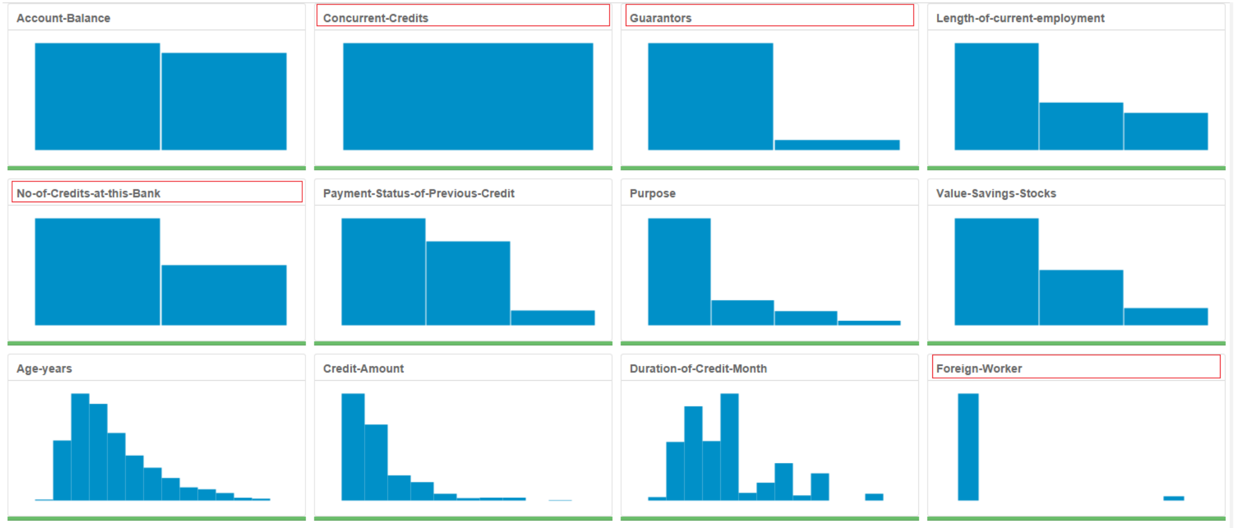


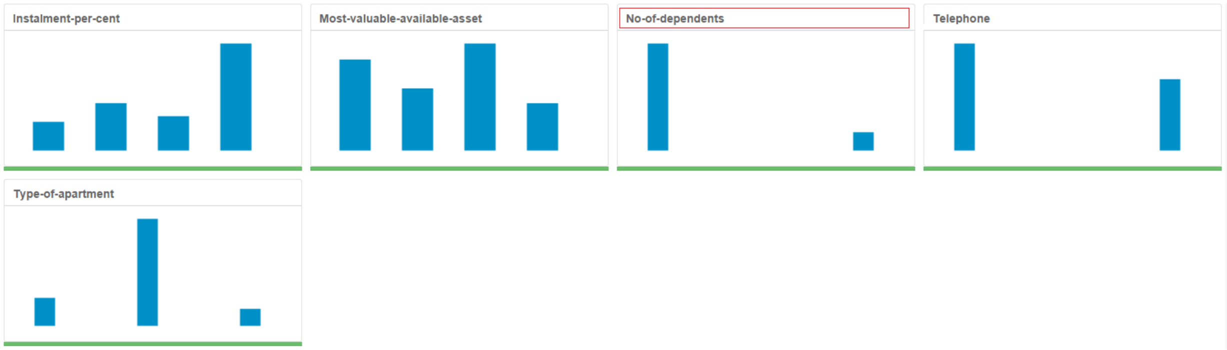
Distribution Analysis of Age-years

6 out of 19 fields had low variability, so I decided to remove these fields, too. They are

1. Concurrent-Credits
2. Guarantors
3. No-of-Credits-at-this-Bank
4. Foreign-Worker
5. No-of-dependents
6. Occupation

My cleaned dataset now have 13 fields with 500 records.





Step 3: Train your Classification Models

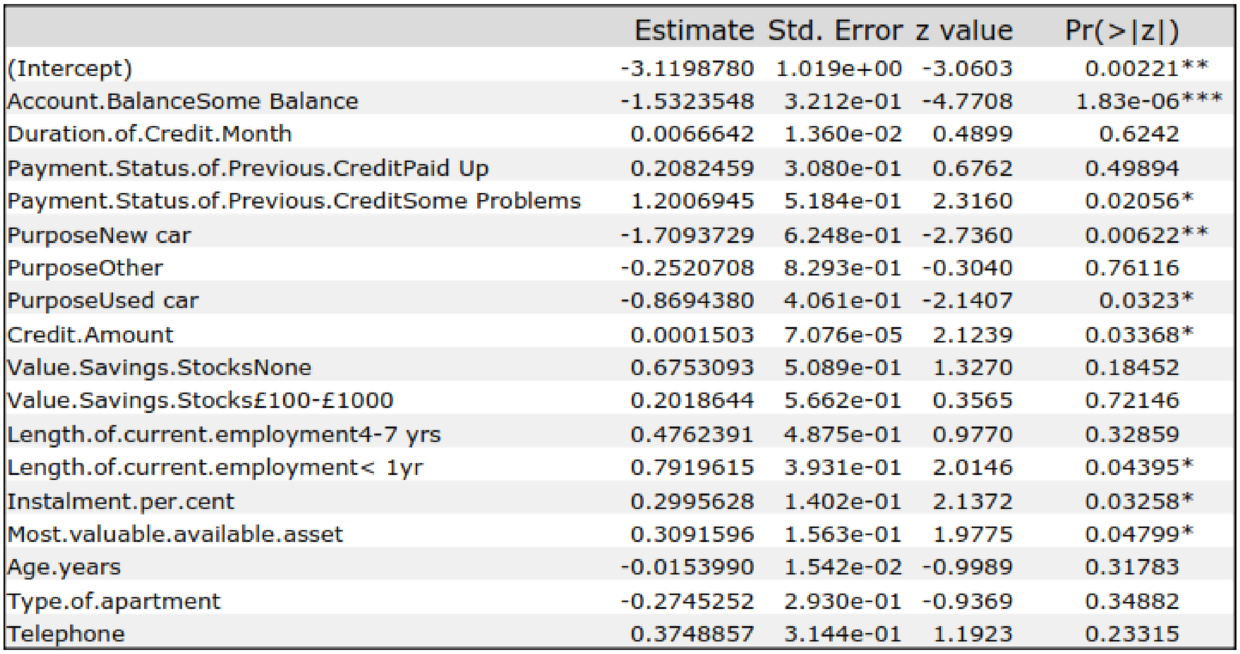
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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.
   I split 70% of my dataset to train model and 30% to validate my models. I trained my data using four models.

   1. Logistic regression



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1198780	1.019e+00	-3.0603	0.00221**
Account.BalanceSome Balance	-1.5323548	3.212e-01	-4.7708	1.83e-06***
Duration.of.Credit.Month	0.0066642	1.360e-02	0.4899	0.6242
Payment.Status.of.Previous.CreditPaid Up	0.2082459	3.080e-01	0.6762	0.49894
Payment.Status.of.Previous.CreditSome Problems	1.2006945	5.184e-01	2.3160	0.02056*
PurposeNew car	-1.7093729	6.248e-01	-2.7360	0.00622**
PurposeOther	-0.2520708	8.293e-01	-0.3040	0.76116
PurposeUsed car	-0.8694380	4.061e-01	-2.1407	0.0323*
Credit.Amount	0.0001503	7.076e-05	2.1239	0.03368*
Value.Savings.StocksNone	0.6753093	5.089e-01	1.3270	0.18452
Value.Savings.Stocks£100-£1000	0.2018644	5.662e-01	0.3565	0.72146
Length.of.current.employment4-7 yrs	0.4762391	4.875e-01	0.9770	0.32859
Length.of.current.employment< 1yr	0.7919615	3.931e-01	2.0146	0.04395*
Instalment.per.cent	0.2995628	1.402e-01	2.1372	0.03258*
Most.valuable.available.asset	0.3091596	1.563e-01	1.9775	0.04799*
Age.years	-0.0153990	1.542e-02	-0.9989	0.31783
Type.of.apartment	-0.2745252	2.930e-01	-0.9369	0.34882
Telephone	0.3748857	3.144e-01	1.1923	0.23315

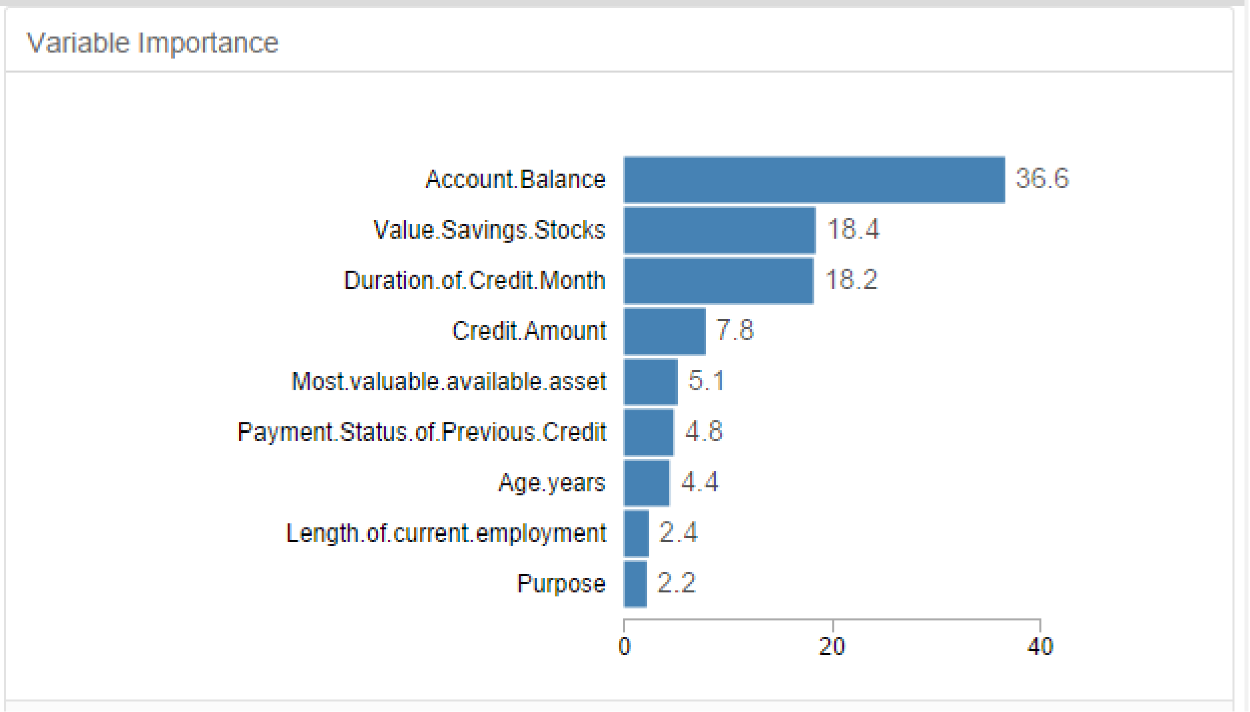
   The following 7 variables are statistically significant for logistic regression model.

   a. Account-Balance
   b. Payment-Status-of-Previous-Credit
   c. Purpose
   d. Credit-Amount
   e. Length-of-current-employment
   f. Installment-per-cent
   g. Most-valuable-available-asset

   So I decided to train my logistic regression model with these 7 variables

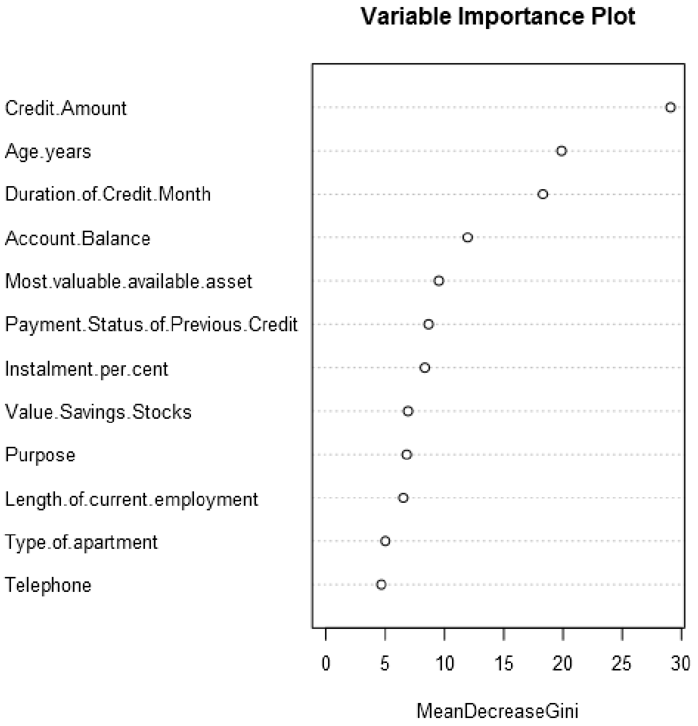
2. Decision Tree

These 10 variables are most important for decision tree model



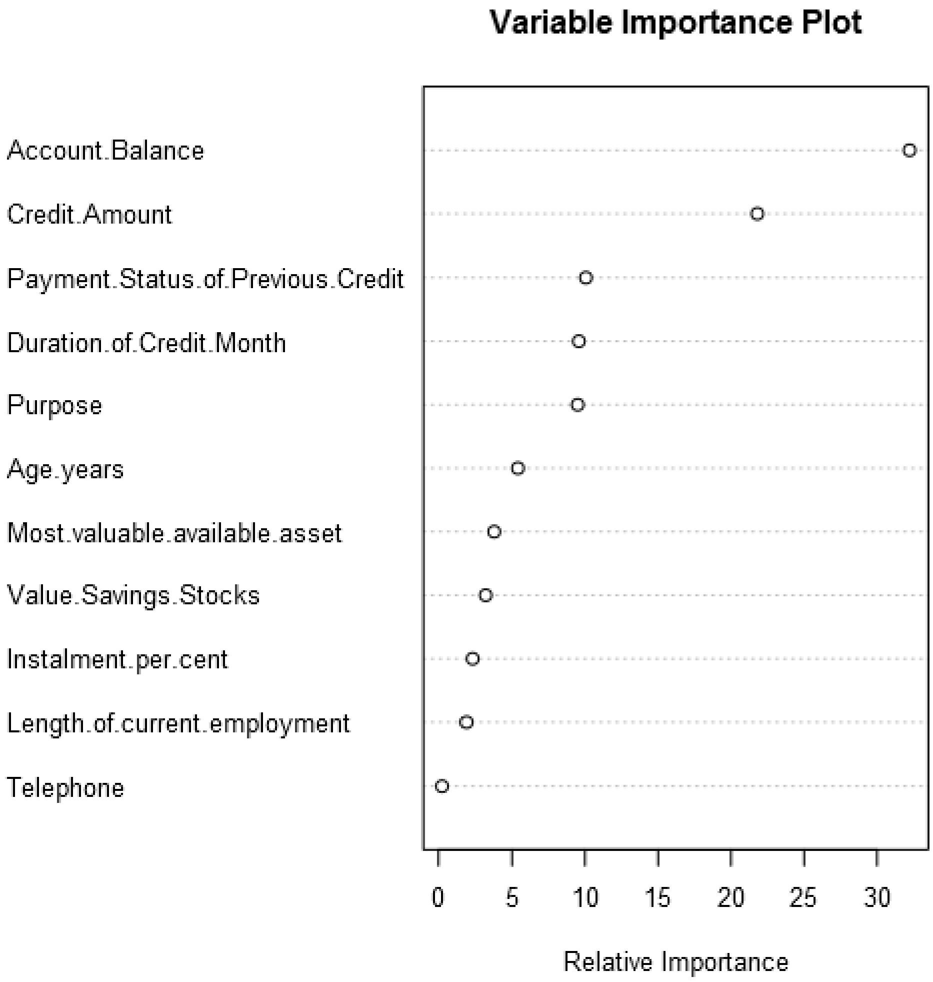
3. Forest Model

For forest model, these are important variables.



4. Boosted Model

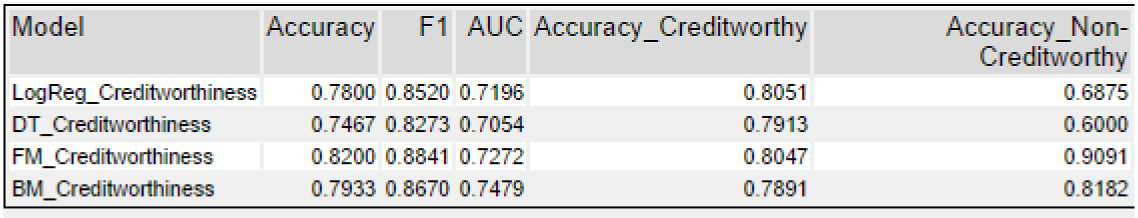
For boosted model, these variables are important.



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?

I have validated my models against the Validation set and overall accuracy are

- 78% for Logistic Regression Model
- 75% for Decision Tree Model
- 82% for Forest Model and
- 79% for Boosted Model.



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogReg_Creditworthiness	0.7800	0.8520	0.7196	0.8051	0.6875
DT_Creditworthiness	0.7467	0.8273	0.7054	0.7913	0.6000
FM_Creditworthiness	0.8200	0.8841	0.7272	0.8047	0.9091
BM_Creditworthiness	0.7933	0.8670	0.7479	0.7891	0.8182

Confusion Matrix for each model

Logistic Regression	Actual_Creditworthy	Actual_Non-Creditworthy	Sum	Accuracy
Predicted_Creditworthy	95	23	118	0.81
Predicted_Non-creditworthy	10	22	32	0.69
Sum	105	45	150.00	0.78
Sensitivity and Specificity	0.90	0.49		

Decision Tree	Actual_Creditworthy	Actual_Non-Creditworthy	Sum	Accuracy
Predicted_Creditworthy	91	24	115	0.79
Predicted_Non-creditworthy	14	21	35	0.60
Sum	105	45	150.00	0.75
Sensitivity and Specificity	0.87	0.47		

Forest Model	Actual_Creditworthy	Actual_Non-Creditworthy	Sum	Accuracy
Predicted_Creditworthy	103	25	128	0.80
Predicted_Non-creditworthy	2	20	22	0.91
Sum	105	45	150.00	0.82
Sensitivity and Specificity	0.98	0.44		

Boosted Model	Actual_Creditworthy	Actual_Non-Creditworthy	Sum	Accuracy
Predicted_Creditworthy	101	27	128	0.79
Predicted_Non-creditworthy	4	18	22	0.82
Sum	105	45	150.00	0.79
Sensitivity and Specificity	0.96	0.40		

Yeah. There is some bias in Creditworthy over Non-creditworthy in Logistic Regression Model and Decision Tree model, and some bias in Non-creditworthy over Creditworthy in Forest Model.

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

Step 4: Writeup

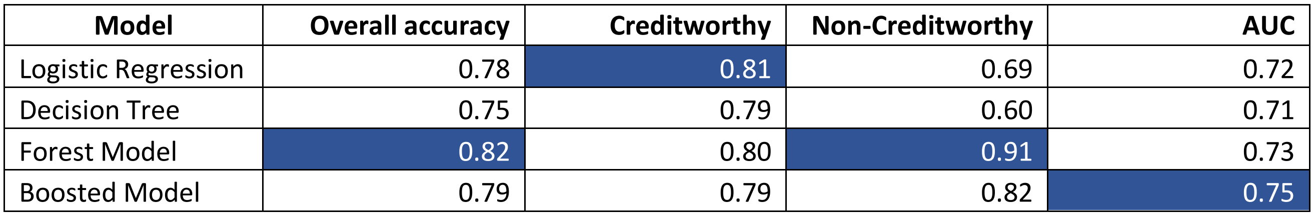
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    c. ROC graph
    d. Bias in the Confusion Matrices

Comparing 4 models



Model	Overall accuracy	Creditworthy	Non-Creditworthy	AUC
Logistic Regression	0.78	0.81	0.69	0.72
Decision Tree	0.75	0.79	0.60	0.71
Forest Model	0.82	0.80	0.91	0.73
Boosted Model	0.79	0.79	0.82	0.75

Forest Model has best overall accuracy.
Logistic Regression Model has best Creditworthy accuracy.
Forest Model has best Non-Creditworthy accuracy.
As for ROC curve, Boosted Model looks like the best fit.

I decided to select Forest Model to score customers' application data, since it has best overall accuracy and best Non-creditworthy accuracy, and it also has second best accuracy in Creditworthy and AUC value.

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

2. How many individuals are creditworthy?
   410 customers are 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.