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

1. What decisions needs to be made?

The decision is whether new customers qualify for loan approvals. Once this decision is made, the bank can then extend loans to these qualified customers. The bank is looking to come up with an better and automatic way to manage the huge increase in credit applications. The short term decision is dealing with the 500 applications. The long term decision is which model to use to automate the process and deal with increased business on a consistent basis.

2. What data is needed to inform those decisions?

The data needed for this decision comprises non-financial information about the applicant such as age, occupation, and dependents, and financial information about the applicant such as assets and stocks. Additional relevant information on all past applications will also be used to build the model.

EA: Awesome: Good answers.

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

A customer can either be approved or not approved. This is a classification problem in which a customer can fall into one of two classes. Therefore, we need to use a binary classification model to make this decision.

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*

Kacper Ksieski

- *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.

After I reviewed the variability of each field using the field summary tool, I decided to remove the Concurrent Credits, Foreign worker, No of depedents, Guarantors, Occupation, and telephone fields due to their low variability, as depicted below.

Kacper Ksieski

Concurrent-Credits

Guarantors

Foreign-Worker

No-of-dependents
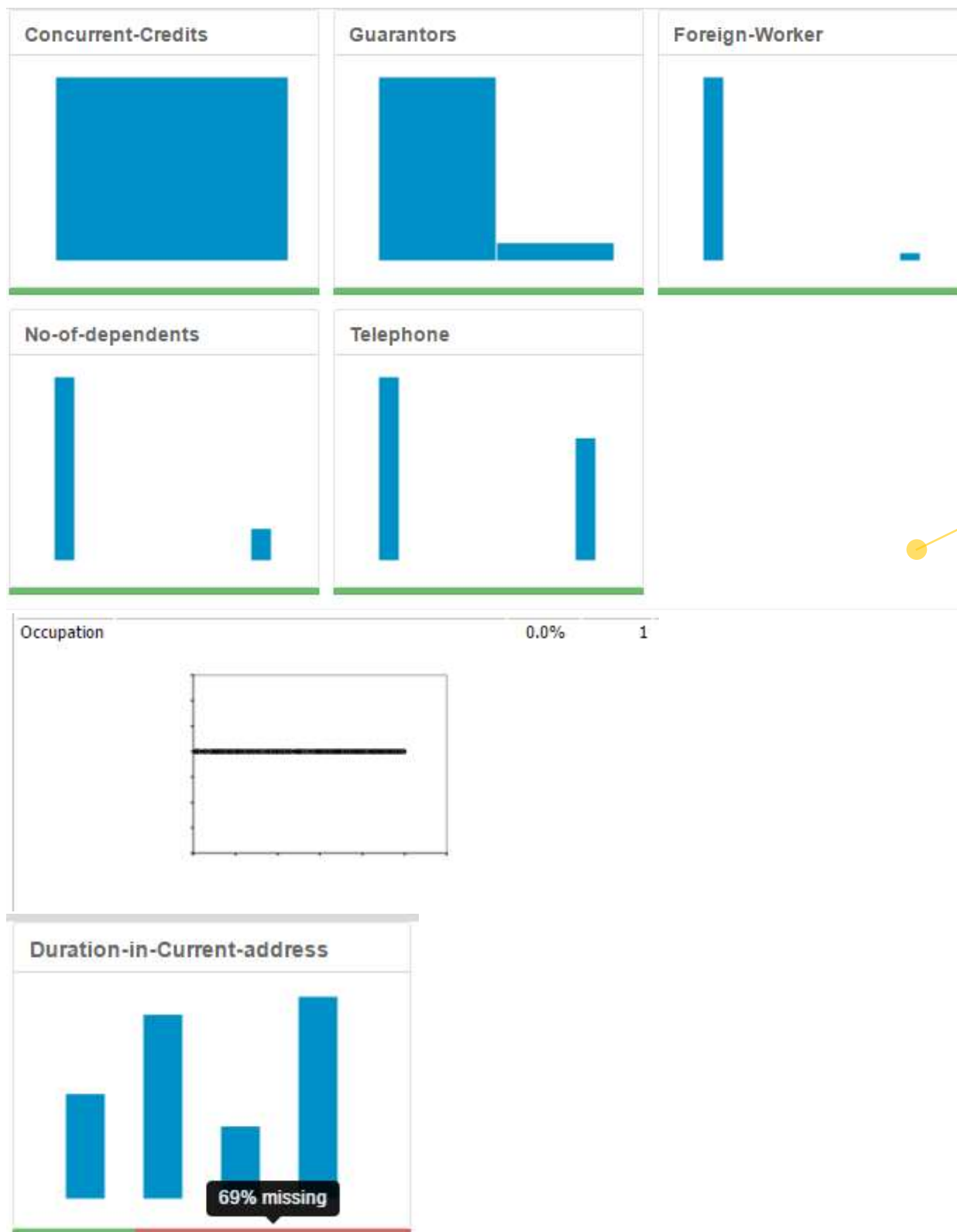
Telephone

EA: Suggestion: Telephone should be removed, but not because of low variability. We remove it based on logic. We don't think it will be helpful to predict creditworthiness.
Good job imputing Age-years. It's acceptable to impute with the average. Since the Age-years data is skewed, imputing with the median is better.

Occupation   0.0%   1

Duration-in-Current-address

69% missing

I decided to remove Duration in Current address because this field is missing 68.8% of its values.

Kacper Ksieski

Finally, I decided to impute Age years because it is only missing 2.4% of its values. I imputed using the average age.

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:*

LOG
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.

Account balance is the most statistically significant variable of the LOG model.

Kacper Ksieski

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	.

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

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?
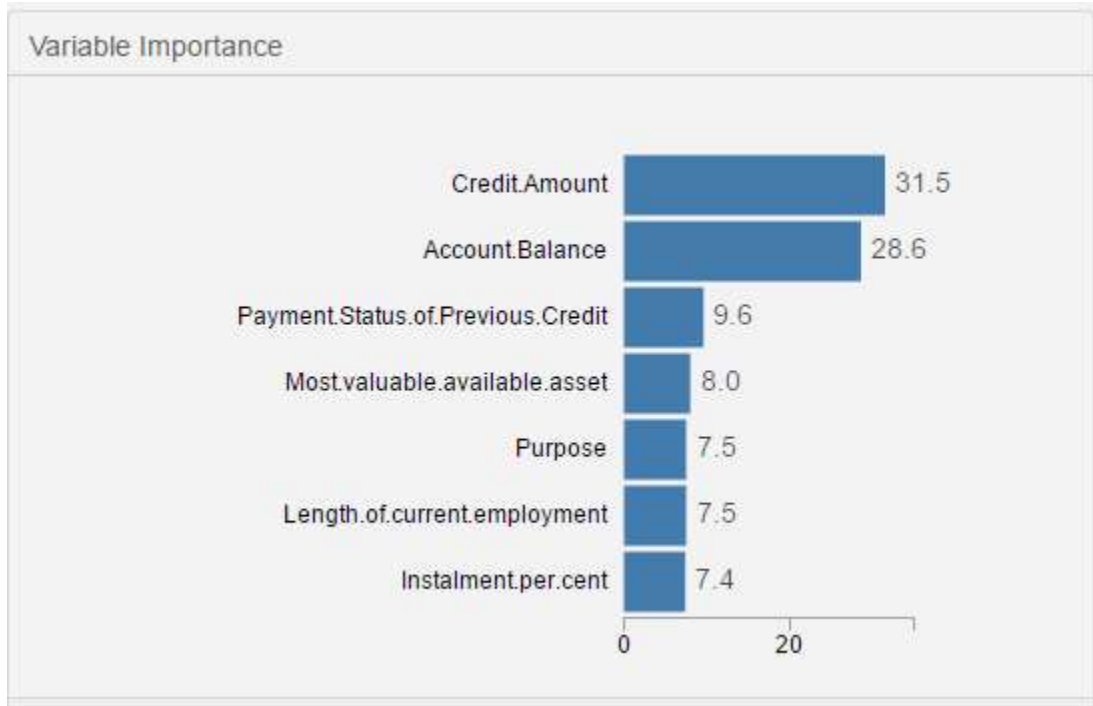
The LOG model is 78% accurate. There is a bias towards Creditworthiness.

Confusion matrix of LOG

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	233	57
Predicted_Non-Creditworthy	20	40

TREE

3. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Credit amount and Account balance are the most important variables of this model.

Kacper Ksieski

Variable Importance

Variable	Importance
Credit.Amount	31.5
Account.Balance	28.6
Payment.Status.of.Previous.Credit	9.6
Most.valuable.available.asset	8.0
Purpose	7.5
Length.of.current.employment	7.5
Instalment.per.cent	7.4

4. 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?
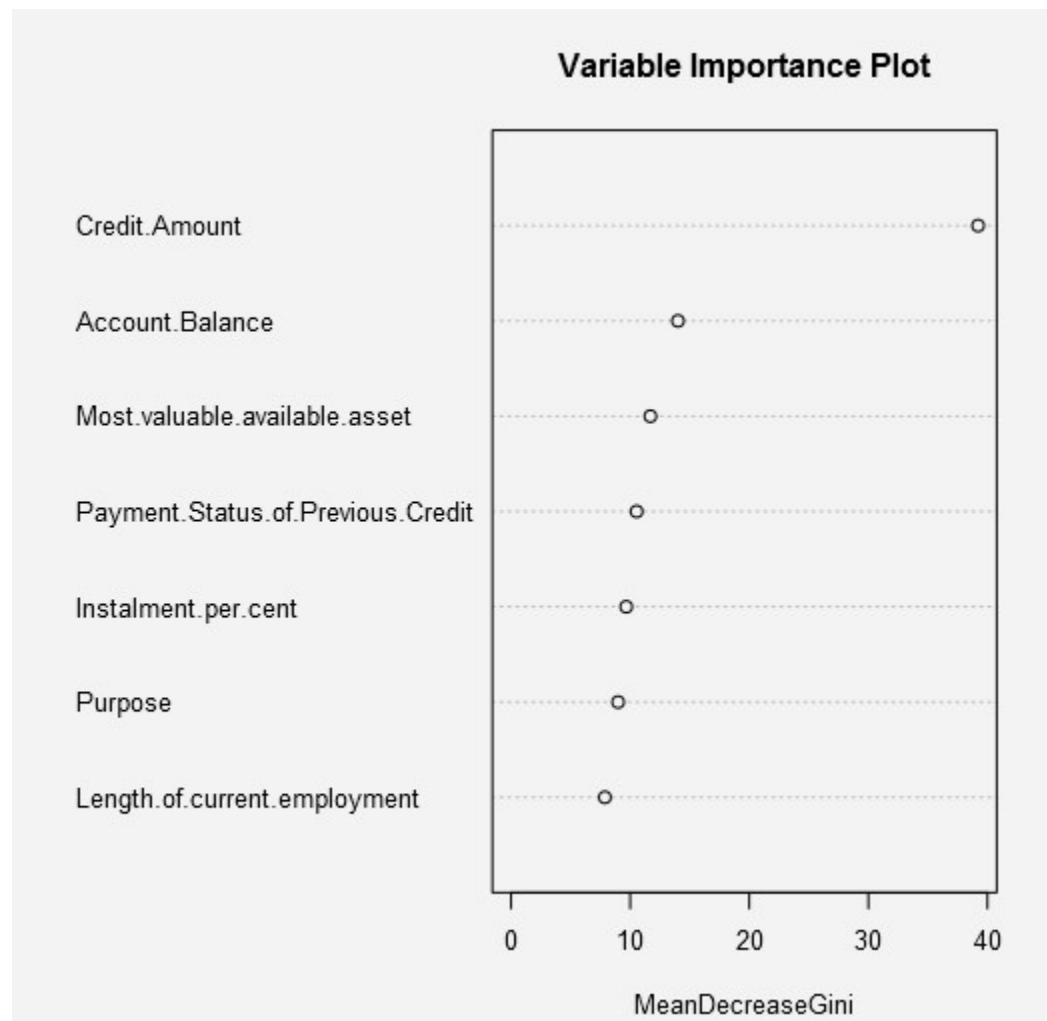
The TREE model is 81.7% accurate. There is a bias towards Creditworthiness.

Confusion matrix of TREE

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	232	43
Predicted_Non-Creditworthy	21	54

FOREST

5. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Credit amount is by far the most important variable of this model.

Kacper Ksieski

Variable Importance Plot

6. 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 FOREST model is 94.3% accurate. There doesn't seem to be much bias in either direction.

Confusion matrix of FOREST

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	252	19
Predicted_Non-Creditworthy	1	78

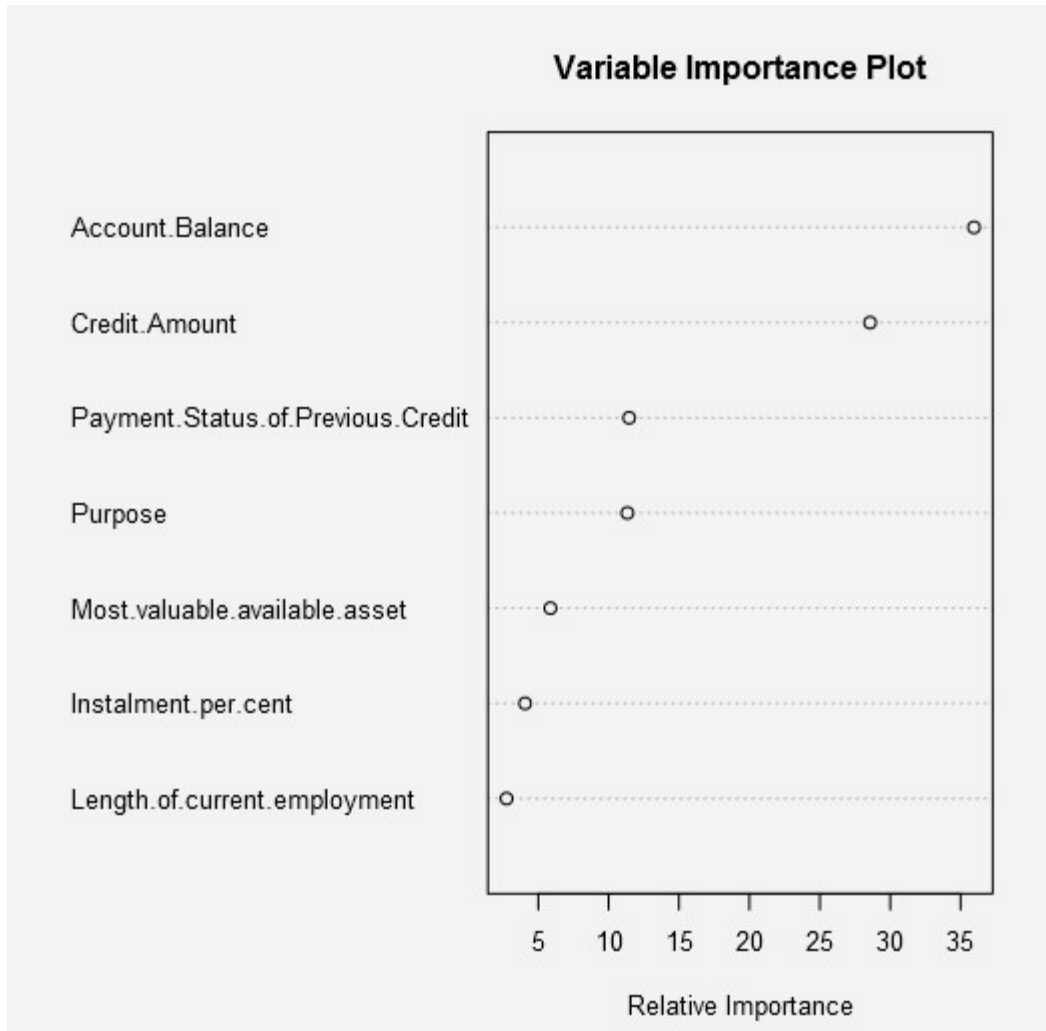EA: 2. Awesome: Good job discussing bias!

EA: 2. Required: The accuracy of the forest model is very high. Have you used a validation set, or is accuracy measured from the training data?

There are too many data points in the confusion matrix, so it looks like the training set is used.

BOOST

7. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

The Account balance and Credit amount variables seem to be the most important.

Kacper Ksieski

Variable Importance Plot

Account.Balance

Credit.Amount

Payment.Status.of.Previous.Credit

Purpose

Most.valuable.available.asset

Instalment.per.cent

Length.of.current.employment

```
    5   10   15   20   25   30   35
```

Relative Importance

8. 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 BOOST model is 78.6% accurate. There does not seem to be a bias in either direction.

Confusion matrix of BOOST

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	243	65
Predicted_Non-Creditworthy	10	32

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"

Kacper Ksieski

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

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LOG	0.7800	0.8582	0.7985	0.8034	0.6667
TREE	0.8171	0.8788	0.8071	0.8436	0.7200
FOREST	0.9429	0.9618	0.9920	0.9299	0.9873
BOOST	0.7857	0.8663	0.8266	0.7890	0.7619

EA: Required: We have to compare metrics from a validation set.

The chart above demonstrates that the FOREST model has a much higher accuracy rate than the other models. This means that the predictive power of the model is strong.

Kacper Ksieski

ROC curve
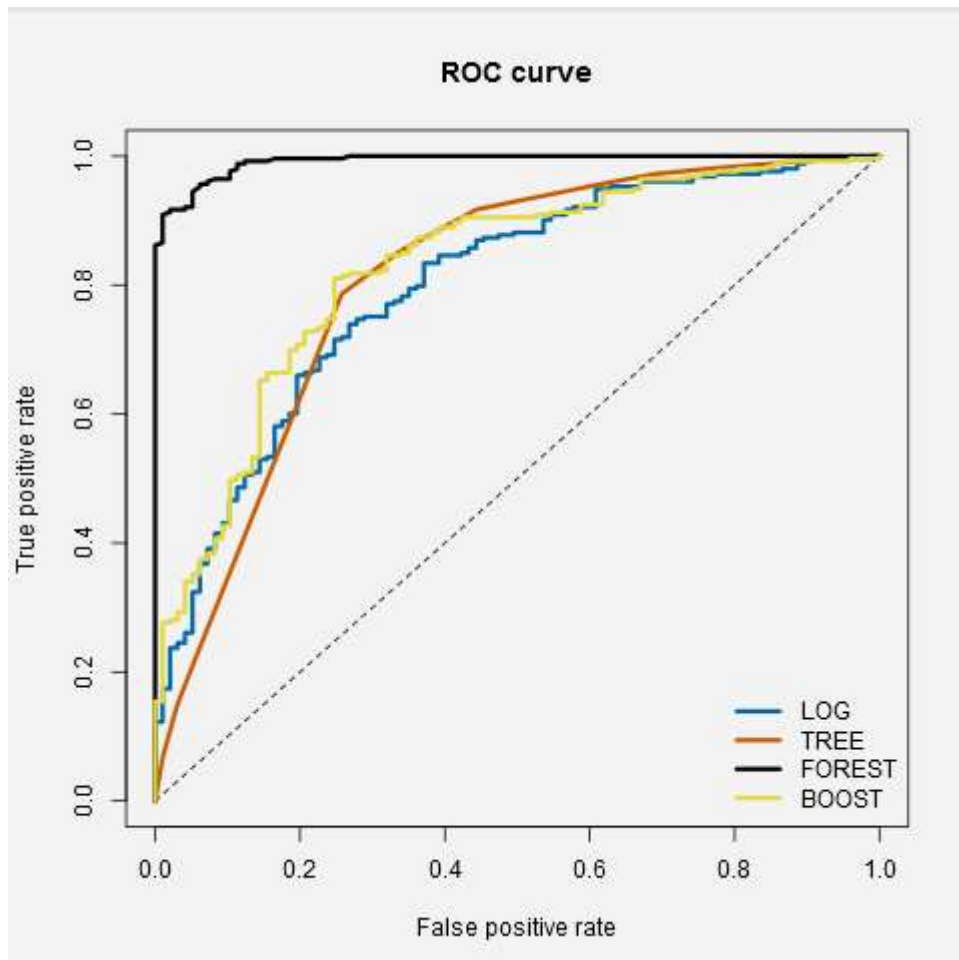
According to the ROC graph above, the FOREST model performs extremely well in general and in comparison to the other graphs.

Confusion matrix of FOREST

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	252	19
Predicted_Non-Creditworthy	1	78

Per the confusion matrix of the FOREST model above, there is a slight bias towards creditworthiness.

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

    2.   How many individuals are creditworthy?
396 individuals are creditworthy out of the 500 new applicants.

EA: The final prediction of creditworthy loan applicants is a bit low.

Kacper Ksieski

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.

Kacper Ksieski