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?

   We need to decide on which people we approve from the pending loan request list.

2. What data is needed to inform those decisions?

   We will need information about the new requestors, including demographics such as income, occupation, net worth. We will also need the same information on previous loan requests that have been approved or denied, as well as the result of approval/denial.

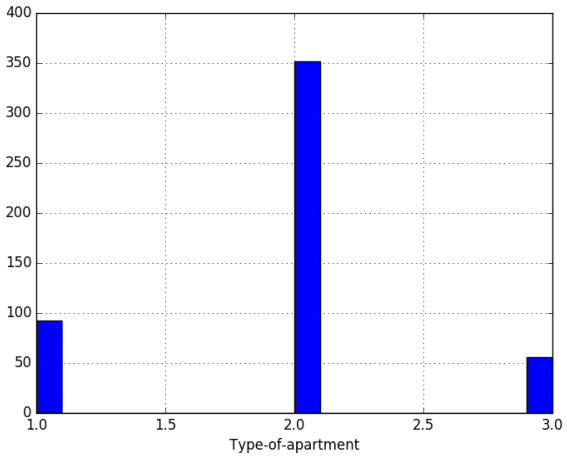
3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

   We should use binary models because we need to decide if we approve or do not approve of loan requests.

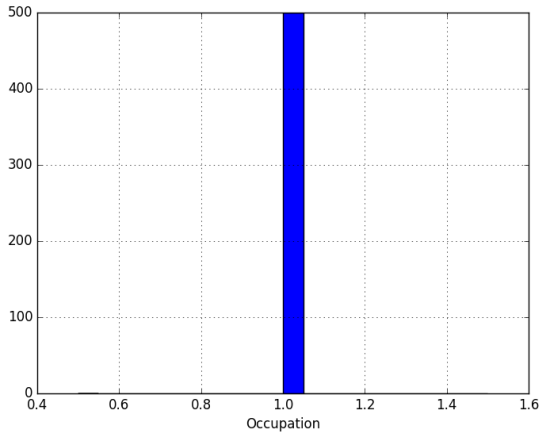
Step 2: Building the Training Set

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

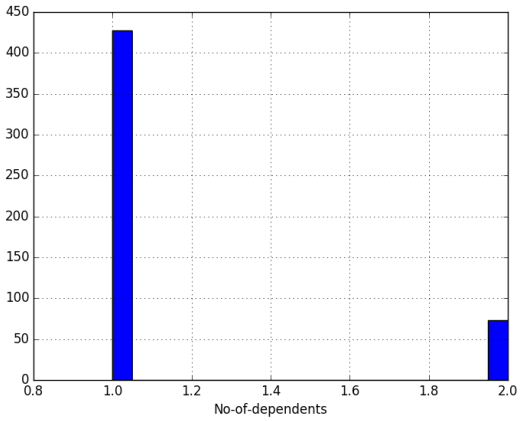
   I removed duration-in-current-address because it only had 156/500 non-missing values.
   I also removed type-of-apartment because it was mostly type 2:



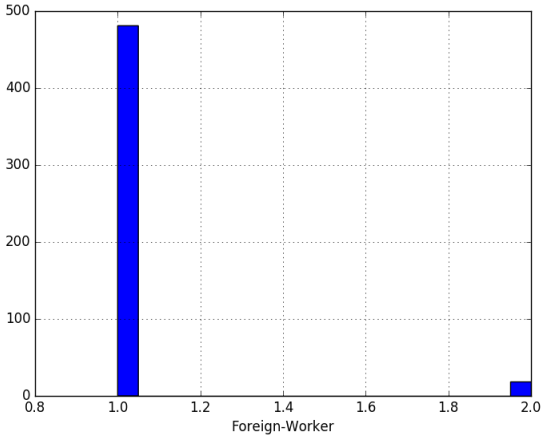
Occupation was also very unbalanced, so I removed it as well:



The same was true for No-of-dependents:

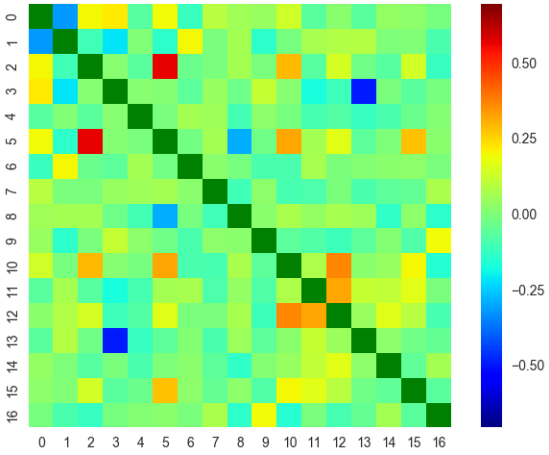


And the same was true for Foreign-Worker:



Concurrent-Credits was only one value, and Gauranters was 90% ‘None’, so I removed both of those fields.

I also made a heatmap of the training data (other than independent variables with too many NANs or only one unique value). Anything with a value greater than 0.7 or less than -0.7 will be bright green (like the diagonal, which is 1), so we can see nothing is strongly correlated:



The key to the heatmap axis labels is below:

0 : Credit-Application-Result
1 : Account-Balance
2 : Duration-of-Credit-Month
3 : Payment-Status-of-Previous-Credit
4 : Purpose
5 : Credit-Amount
6 : Value-Savings-Stocks
7 : Length-of-current-employment
8 : Instalment-per-cent
9 : Guarantors
10 : Most-valuable-available-asset
11 : Age-years
12 : Type-of-apartment
13 : No-of-Credits-at-this-Bank
14 : No-of-dependents
15 : Telephone
16 : Foreign-Worker

Step 3: Train your Classification Models

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.

For the logistic regression model, Credit-Amount seemed to be the only meaningful variable by way of t-test (p-value < 0.05). The next two most important were Instalment-per-cent and Most-valuable-available-asset.

The categorical variables were dummied, and so they are linearly dependent on their dummies, so we can't get p-values for those.

Logit Regression Results

Dep. Variable:	y	No. Observations:	350
Model:	Logit	Df Residuals:	332
Method:	MLE	Df Model:	17
Date:	Mon, 19 Dec 2016	Pseudo R-squ.:	0.2095
Time:	01:18:08	Log-Likelihood:	-168.33
converged:	False	LL-Null:	-212.95
		LLR p-value:	8.413e-12

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
const	-5.7053	nan	nan	nan	nan	nan
x1	-0.0143	0.014	-1.051	0.293	-0.041	0.012
x2	-0.0002	6.68e-05	-2.276	0.023	-0.000	-2.11e-05
x3	-0.2542	0.142	-1.788	0.074	-0.533	0.024
x4	-0.2675	0.150	-1.786	0.074	-0.561	0.026
x5	0.0163	0.015	1.102	0.270	-0.013	0.045
x6	0.1625	0.301	0.539	0.590	-0.428	0.753
x7	7.0878	nan	nan	nan	nan	nan
x8	8.4902	nan	nan	nan	nan	nan
x9	0.7981	8.25e+06	9.67e-08	1.000	-1.62e+07	1.62e+07
x10	0.1640	8.14e+06	2.02e-08	1.000	-1.59e+07	1.59e+07
x11	-0.4332	8.31e+06	-5.22e-08	1.000	-1.63e+07	1.63e+07
x12	-2.3467	nan	nan	nan	nan	nan
x13	-0.8558	nan	nan	nan	nan	nan
x14	-2.7140	nan	nan	nan	nan	nan
x15	-1.9523	nan	nan	nan	nan	nan
x16	2.0830	nan	nan	nan	nan	nan
x17	1.3020	nan	nan	nan	nan	nan
x18	2.1867	nan	nan	nan	nan	nan
x19	0.6866	nan	nan	nan	nan	nan
x20	0.1305	nan	nan	nan	nan	nan
x21	0.0725	nan	nan	nan	nan	nan
x22	0.4936	nan	nan	nan	nan	nan
x23	0.3961	nan	nan	nan	nan	nan

Variables key:

x1 : Duration-of-Credit-Month
x2 : Credit-Amount
x3 : Instalment-per-cent
x4 : Most-valuable-available-asset
x5 : Age-years
x6 : Telephone
x7 : Account-Balance_No Account
x8 : Account-Balance_Some Balance
x9 : Payment-Status-of-Previous-Credit_No Problems (in this bank)
x10 : Payment-Status-of-Previous-Credit_Paid Up
x11 : Payment-Status-of-Previous-Credit_Some Problems
x12 : Purpose_Home Related
x13 : Purpose_New car
x14 : Purpose_Other
x15 : Purpose_Used car
x16 : Value-Savings-Stocks_< £100
x17 : Value-Savings-Stocks_None
x18 : Value-Savings-Stocks_£100-£1000

x19 : Length-of-current-employment_1-4 yrs
x20 : Length-of-current-employment_4-7 yrs
x21 : Length-of-current-employment_< 1yr
x22 : No-of-Credits-at-this-Bank_1
x23 : No-of-Credits-at-this-Bank_More than 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?

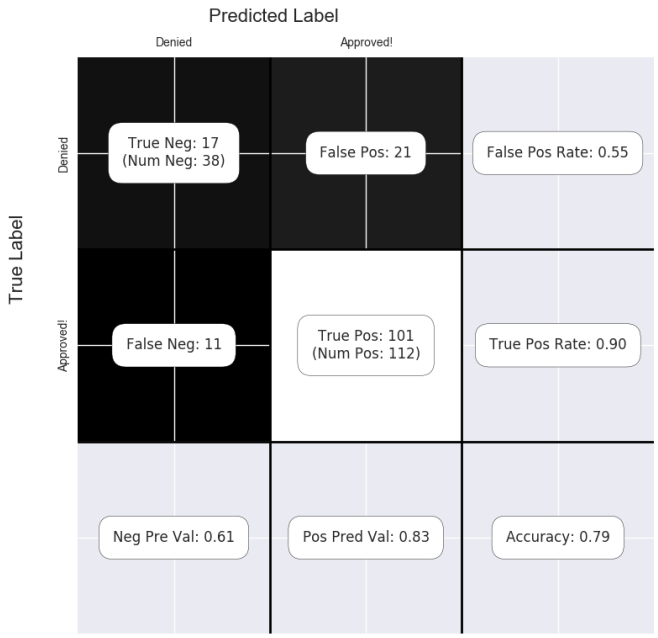
   I don't see accuracy as being useful here because we have to choose a probability threshold for classification as approved or not...nevertheless, with the default 0.5 threshold for classification of approval, the accuracies are:

logistic regression	0.797
decision tree	0.760
random forest	0.780
gradient forest	0.800

   The AUC scores (more useful for binary classification like this) are:

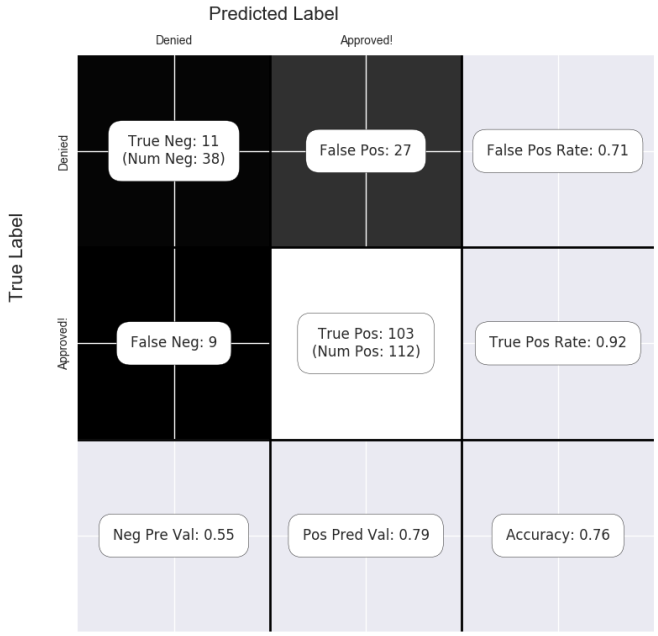
logistic regression	0.751
decision tree	0.713
random forest	0.748
gradient trees	0.759

Confusion matrices: logistic regression

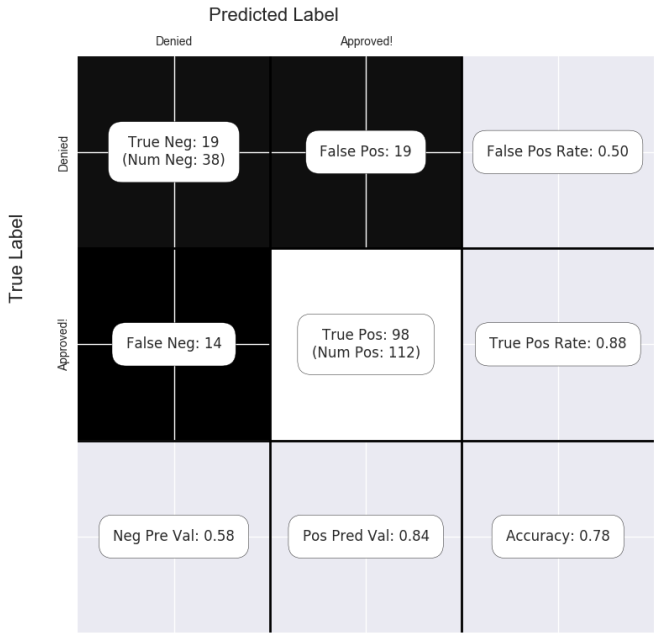


decision tree

false positives increased a lot, worst one overall

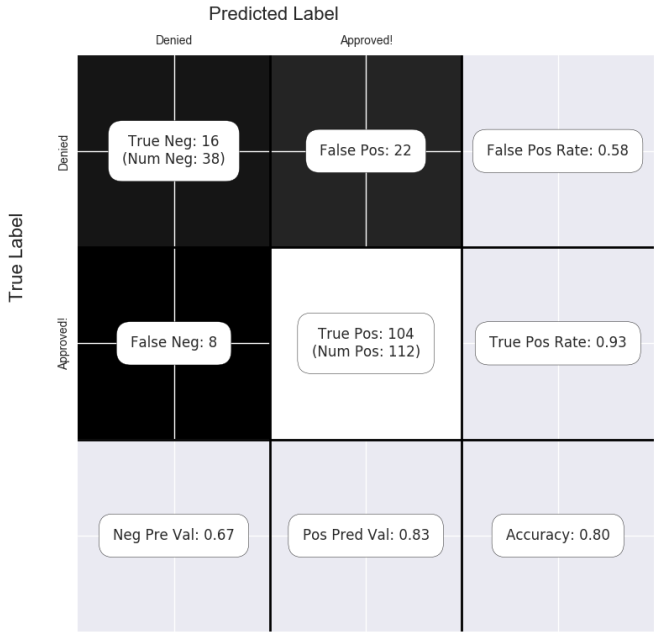


random forest



gradient boosted forest

best overall



I picked the gradient boosted model because it had the highest AUC and accuracy (best performance), as it often does compared with the other models I tried.

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:

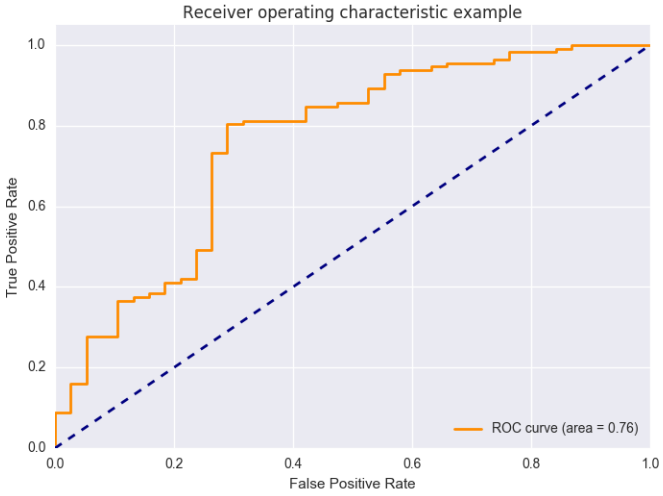
1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments

c. ROC graph
d. Bias in the Confusion Matrices

The gradient booster had the highest accuracy and area under the ROC curve (AUC score).

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

I chose the gradient boosted forest model because it had the highest accuracy at a default 0.5 level, and the highest AUC score. It also had the thrid-best false positive rate, and second-best false negative rate. I'm not sure what's better to have – false positives or false negatives (depends on how much risk you want to take, tune it for more false negatives for less risk, and vice-versa). The model didn't seem biased towards predicting more false positives or false negatives, compared with the other models. The ROC curve looks like this:



It seems like a false positive rate of around 0.3 might be a good choice.

2. How many individuals are creditworthy?

386 out of 500 total applications.