Prompt:

Given access to bank loan data, we have data about all loans asked to the bank, whether the bank decided to grant it and whether the borrower managed to repay it.

1. Come up with a better strategy to grant loans.
2. Describe the impact of the most important variables that lead to the prediction. Focus particularly on the variable “is_employed”.
3. Other variables that I would have like to have in the model.

Challenges:

The first challenge was that information from 2 different data sets needs to be merged.
In Excel, the information was custom sorted by Loan ID for “borrower” and “loan” table. Each table’s information corresponded to an ID. With the same ID and same amount of rows for each table, the information was copied and pasted into a new table called “combined_table”.

The second challenge was that the data has many NA values to deal with. By simply dropping rows with NA values, most of the data would be lost. For each variable, the NA values were dealt with differently. For the “avg_percent_credit_limit_used_last_year” variable, the NA values was treated as 0’s. If there was no credit, there would be no credit average. For “Fully_repaid_previous_loans” and “currently_repaying_other_loans”, the approach was to dummify the variables:

```
get_dummies(s1, dummy_na=True) [1]
   a  b  NaN
0  1  0    0
1  0  1    0
2  0  0    1
```

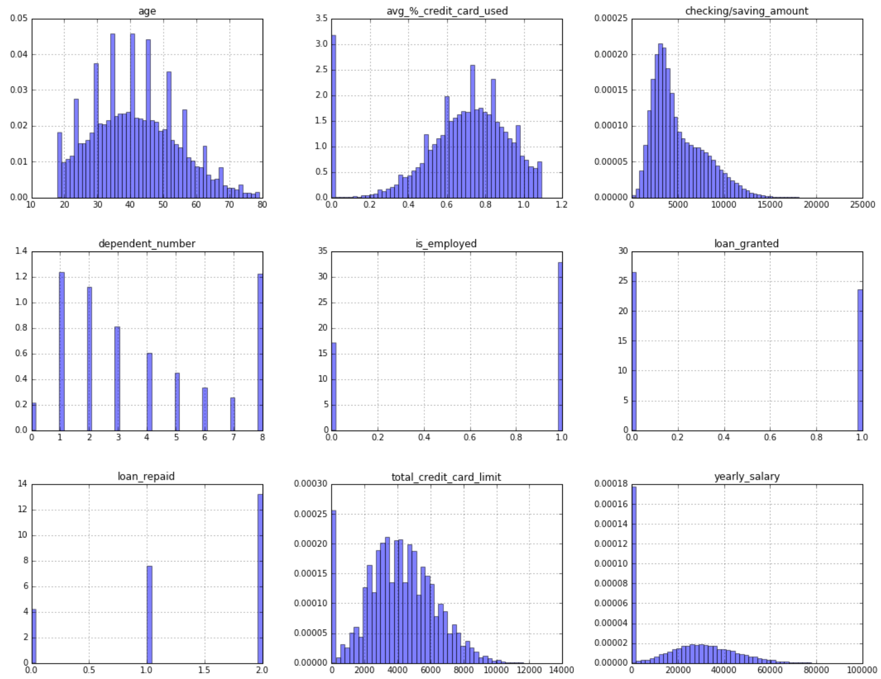
That way, the variables would be retained. For the loan_repaid column, the NA values were converted into a number 2, which stated that the loan was not granted.

The third challenge was the multi-categorical variables for “Loan_Purpose”. The solution was to dummify the variables so each category has its own binary column.

Feature Engineering:

Since check and savings amount were similar, I combined them as one feature for feature reduction.

Data Exploration:



Age, average credit card use, credit card limit, and yearly salary almost seem to follow a normal distribution while checking/savings amount are right tail skewed.

Developing a predictive model

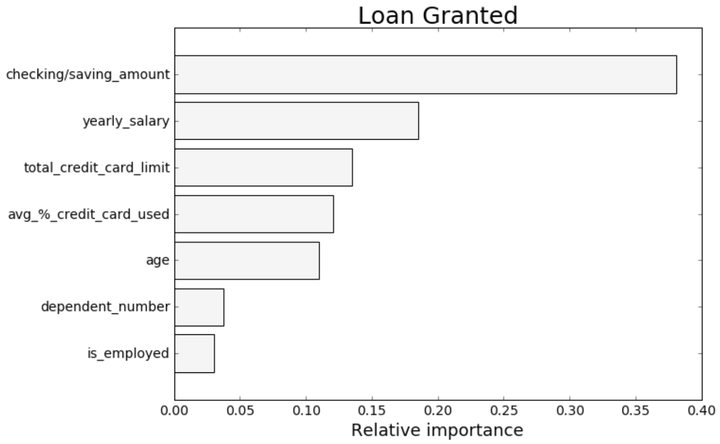
For the model of choice, I chose the Gradient Boosted Classifier. This model was chosen because I wanted to use the gradient to identify "weak learners" or "shortcomings" with each iterative step.

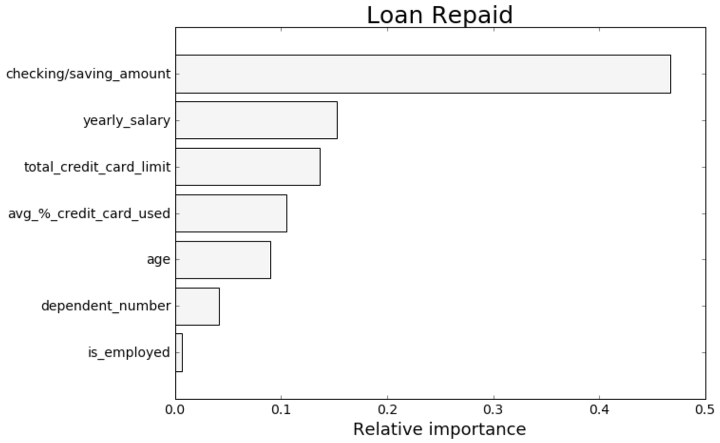
For the loan granting strategy, I propose 2 models:

- the first to predict whether the person get can a loan
- second then whether that same person can pay the loan.

With the label "loan_granted", there was a 78% accuracy in identifying people who received a loan. The ability to pay the loan after the loan was granted was much easier to predict with approximately a 89% accuracy. I believe that this two model system is necessary because the first thing is to see if the person can get a loan and then see if the same person can pay it back.

Most important variables for prediction:



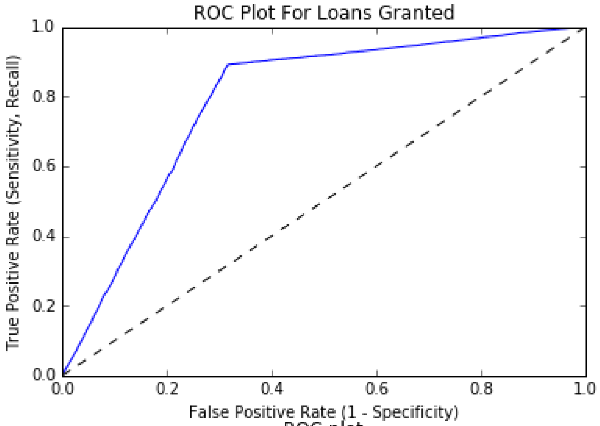


Repaid previous loans, loan purpose, repaying other loans, and first loan variables were dropped because these had little to no impact on granting a loan. Nor did it impact the ability to repay the loan once it was granted.

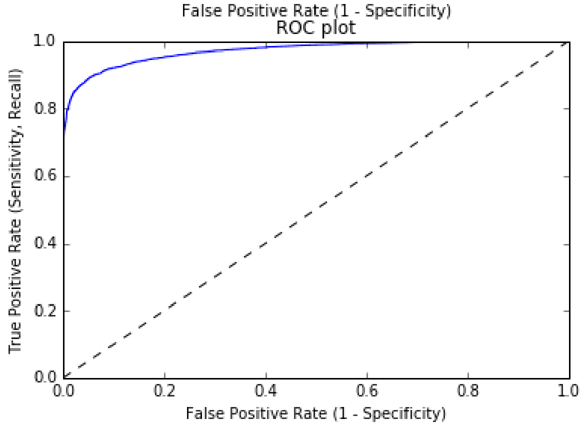
The top most decisive factors in determining if someone was granted a loan was checking/savings amount, yearly salary, credit card limit, and credit average percent of card limit.

The "is_employed" variable appears to have little impact on the outcome of whether a loan was granted. Nor was it a deciding factor whether the person repaid the loan. The most important variable appears to be the amount of savings a person has in order to receive and pay back a loan.

Metrics:



Loan Granted Model:
precision: 0.70
recall: 0.89
F1 score 0.78



Loan Repaid Model:
precision: 0.93
recall: 0.90
F1 score 0.92

Metrics confirm that after the data was segmented into loans granted, who can pay back the loan was much easier to predict.

Other Variables I would've liked to see:

- Education: Does education affect the ability for someone to get a loan and pay it back?
- Job Type:Does the type of person's job affect the loan outcome?
- Marital Status :Does being married or not affect the outcomes for loan?
- Location: Do certain people in certain parts of the country do better with loans than others?
- Loan Amount : Does the loan amount influence the ability to pay it back?

Conclusion:

The 2 part classification process would definitely help in developing a better strategy for loans, which will increase the profitability for the bank.

References:

1. **PANDAS.GET_DUMMIES — PANDAS 0.17.0 DOCUMENTATION** "Pandas.Get_Dummies — Pandas 0.17.0 Documentation". *Pandas.pydata.org*. N.p., 2016. Web. 27 Dec. 2016