Summary of Abhay Shah's work

Outline

In this document we summarize the result of an exercise sent to us by Wealthfront that required classification of a loan as a good or bad based on certain attributes associated with it. We begin with outlining the procedure, followed by the methods used, the results of the exercise, and finally the features that play an important role in deciding if the loan is good or bad.

Procedure

The original dataframe had 10,000 rows and 29 columns which did include empty or Not-a-Number (NaN) entries. We start with dropping NaN rows and columns, that is, those rows and columns which have NaN as all of its entries. We also drop those columns that are irrelevant to the problem at hand. Then we fill missing entries for certain columns with either the median of that column or a new value, for example, the column `emp_length` has NaN that are filled with the value `"0_years"`. This is followed by extracting year and month from the `earliest_cr_line` feature. The last step in featurization is to create dummy columns for categorical features and append them to the dataframe.

The most important step is to assign a loan as good or bad. For this we use the `loan_status` attribute, and assign those with values `"Default"` and `"Charged Off"` as **bad**, and the remaining five, `"Fully Paid"`, `"In Grace Period"`, `"Late (16-30 days)"`, `"Late (31-120 days)"`, and `"Current"` as **good**. With this definition, there are 234 out of 9524 loans that are classified as bad ones amounting to 2.45% of total loans.

Methods Used

We start with using the simple logistic regression model by using various train-test splits and normalized/un-normalized data. We find that we can achieve very high accuracy by using un-normalized data using any *reasonable* train-test split ranging from 0.10 to 0.40. As soon as we use normalized data, the accuracy drops significantly.

We also use the Random Forest model and find that the accuracy is not as good as that of the logistic regression with un-normalized data. This certainly suggests that the data we have at hand has more of a linear relationship between the *loan status* and its various attributes. One can continue exploring other various procedures but given the limitation of time, we stop here quite satisfied as it is unusual to achieve such high accuracy using a simple logistic regression model.

Results

One of the metrics we use is the generic accuracy which is defined as the ratio of correctly categorized test data versus total size of the test data. On average, with 500 simulations, the accuracy we attain is **greater than 99%, i.e., 99.7%**. Another metric is the ROC-AUC score, the area under the ROC curve, which is, on average, **0.976**, showing remarkable accuracy in our simplistic logistic regression binary-classification model.

We also look at the True-Positive-Rate or sensitivity, which is defined as the accuracy with which we can predict a loan to be good, and we find that it is, on average, **greater than 99%, closer to 99.9%** to be precise. Most importantly, the True-Negative-Rate or specificity, which is defined as the accuracy with which we predict a loan to be bad, is **greater than 93% and is significantly closer to 94%**.

Finally, we also take a look at the Type-I and Type-II errors. We define the former as the error with which we identify a bad loan as a good one, and vice versa for the latter. We classify about 6% of bad loans as good and almost 0% of good loans as bad. Along with it we look at their counter parts, precision and recall. We outline the results in the following table where we show the metric, the mean with the low and high 90% confidence interval, and the ideal values one would like to compare them with.

Table 1

Metric	Low-90	Average	High-90	Ideal values
Accuracy	99.72%	99.74%	99.76%	100%
ROC-AUC	0.9740	0.9766	0.9791	1.0000
TPR/Sensitivity	99.86%	99.88%	99.90%	100%
TNR/Specificity	93.56%	94.05%	94.56%	100%
Type I error	5.44%	5.94%	6.44%	0%
Type II error	0.10%	0.12%	0.14%	0%
Precision	99.84%	99.85%	99.86%	100%
Recall	94.63%	95.33%	96.03%	100%

Top 4 Features

The top four features that play the most important role in deciding whether a loan is good or bad (given our definition of a good or bad loan outlined in the Procedure-section) are:

Table 2

Features	In favor of	Possible Explanation
`out_prncp`	good loan	Higher the outstanding principal, higher the probability that the loan has just started implying that its status is mostly "Current" which is classified as a good loan.
`total_rec_prncp`	good loan	It means that the borrower has been making regular payments and is nearing the end of its loan period.
`total_rec_int`	bad loan	More the received interest, which is mostly during the first few months, a borrower notices that they have hardly made any payments towards the principal. This might demotivate the borrower towards not making any further payments.
`funded_amnt`	bad loan	Higher the funded loan amount, higher the payments required, and more difficult for the borrower to meet the financial goals.