STAT 542 Project 3 Report

Ankush Agrawal (NetID : ankusha2)

To predict the lending club loan status, I followed couple of preprocessing steps. These were done to ensure a smooth run of the models and to build a robust model. The steps are as follows:

1. Considering there were a number of factor/categorical variable in the data, I explored each one in detail to understand their values. I figured some of them had too many levels to them and hence they might not contribute to the overall model building exercise. For example, emp_title, title and zip_code had factor levels running in tens of thousands and I removed them. I also removed earliest_cr_line as it was a date column and it was misleading to keep it as a factor column. I thought about calculating the time through this column but as I developed a model I felt this was not necessary. I also noticed that sub_grade has more granular data than grade so I removed grade and kept sub_grade. I also noticed that fico_range_low and fico_range_high has an almost perfect correlation (0.9999), so I decided to keep one of them and remove the other. So in essence I have removed the following variable : earliest_cr_line, emp_title, grade, title, zip_code and fico_range_low.
2. Once I removed the variables, I explored all the other variables to explore the remaining issues in it. I noticed that I needed to treat a few more categorical variables and treat missing values in the data. First, I noticed that emp_length and term_num can be converted into numerical data and they are wrongly kept as categorical data. So I changed them into their numerical order. Next I also removed a few redundant and less frequently occurring categories in home_ownership. So I merged all the instances of “ANY” and “NONE” to “OTHER” in home_ownership.
3. There is no easy way to deal with missing values. In the past I have replaced the missing values as average or median values, but I was never satisfied with the statistical or theoretical basis for it. In fact, there does not exist a strong literature and reasoning behind it (if there does, I would be more than glad to read it). Since, any imputation will bias my model to prefer the values in the favour of what I impute with, I decided to drop all the observations with missing values. I believe I am losing roughly 6-7% of my overall data but I think it is better than biasing your model and skewing it for particular values.
4. Next, I one hot encoded all my categorical data to let it smoothly run into my model
5. For the model, I used a logistic regression model using 10fold cross validation glmnet. I chose the lambda.min (Which gives the lowest misclassification error) to get my predictions.
   My predictions on the 3 train and test splits showed me the following logLoss (based on evaluation script in project3.html):

	Test1	Test2	Test3	Average
glmnet	0.4505	0.4521	0.4517	0.4514

I am glad I could meet the criteria set forth for the project and I ran multiple times to see if the results are not arbitrary and they do not change a lot in multiple runs.

My run time for the code was 26.2 mins and I feel that is a lot but I think with the amount of data and 10 fold cross validation is bound to take up that much time.

Here are the details of my session info :

```
R version 3.5.1 (2018-07-02)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)
```

I used the following packages :

1. glmnet_2.0-16
2. caret_6.0-80