PROJECT 3: Lending Club Loan Status

We have been provided with the historical loan data from Lending club . The dataset recorded categorical target variable hence posed a Supervised Classification problem. We worked on the filtered dataset that was provided to us from the staff members that contain only 30 features and 3 target classes(Default, Charged Off, Fully Paid).

Following are the steps that I followed:

1. We start by reading in the train and test data into memory and setting the id field as index as it was used to uniquely identify every observation.
2. We treat **Default** and **Charged Off** as one class , hence I replaced all occurrence of Default with Charged Off such that we now have only 2 classes in target variable.
3. Next I created a function to keep track of various types of variables that are there in our dataset (Categorical, ordinal, nominal, discrete, continuous).
4. Next I imputed the missing value in the dataset with mean for continuous variable and mode for categorical variable.
5. Then I went ahead doing some transformation in the features like extracting the integer part from **term** , **emp_length.** I created an ordinal encoding dictionary for **Grade** and **Sub Grade.**
6. I decided to drop **'title', 'zip_code', 'emp_title'** as it contained many levels and an one hot encoding would just blow up the design matrix.
7. I extracted year and month from the **earliest_cr_line** and used them as additional features.
8. Next for all the categorical variables left in the dataset , I did one hot encoding.
9. I used inter quartile range for clipping extreme values in continuous variables.
10. Next I encoded my target classes as 0 for Fully paid and 1 for Charged Off.
11. I used standard scaler to scale all my variables.
12. Next I decided to build 3 models out of which my first model was xgboost where I set the boosting rounds to be 70, objective function to be binary logistic regression, maximum dept of tree to be 7 and learning rate to be 0.27. These parameters were chosen by going Grid search and Cross validation on train data.
13. My second model was random forest with tree depth equal to 18 and no of trees equal to 80, again the parameters were chosen by doing Grid search and Cross validation on train data.

14. My third model was Logistic regression with default params.
15. Out of the 3 models , Model 1(Xgboost) seemed to work best and gave the lowest log loss.
16. Predictions from every model is saved to disk as per the format asked in instructions.

Model Evaluation Results:

Index	model1	model2	model3
Test-Set1	0.4478	0.4538	0.4582
Test-Set2	0.449	0.4548	0.4586
Test-Set3	0.4479	0.4537	0.4583

Average across 3 train-test splits:

model1	0.448233
model2	0.454100
model3	0.458367

Running Time:

Script executed successfully in **7 minutes**.

System Specifications:

Model Name:	MacBook Air
Model Identifier:	MacBookAir7,2
Processor Name:	Intel Core i5
Processor Speed:	1.6 GHz
Number of Processors:	1
Total Number of Cores:	2
L2 Cache (per Core):	256 KB
L3 Cache:	3 MB
Memory:	8 GB

Acknowledgements:

We had referenced various Kaggle kernels and some stack overflow discussions on some transformation and approach.