STAT 542

Project-4: Lending Club Loan Status

1. Introduction:

This project is to build a model to predict the chance of default for a loan using historical loan data issued by Lending Club.

2. Data Pre-processing

Packages used: glmnet","randomForest", "xgboost", "Lime"

The provided dataset contains 30 features with some of the variables having missing values which requires further cleaning and replacing missing values with estimated values. Out of the features with missing values, we do the following steps:

- Variable emp_title is dropped as it has too many levels to have a huge influence in prediction of loan default
- We converted variable emp_length to integer (this has benefits in terms of faster learning with xgboost) and imputed the missing values with mean
- For the remaining continuous variable parameters, we imputed their missing values with their mean values (mean is computed by ignoring NA values)

The other pre-processing steps are as shown below:

- We dropped the features id, grade(this feature is captured by sub-grade), fico_range_high/low (we replace these 2 features with their mean), zipcode (too many levels), title(very generic predictor)
- For the variable term, we converted it into integer. Similarly, we replace the levels any and none for home_ownership with other (again contributes to faster training)
- We do log-transform for the predictors annual_inc and installment as they are highly skewed
- We take only the corresponding year for the variable earliest_cr_line
- We replace the words 'default' and 'charged off' with 1 and 'fully paid' with 0 (binary classification) in loan_status feature

3. Model Selection:

3.1. Linear Regression Model:

The data matrix is converted to one-hot encoded matrix using the function model.matrix in R. We then proceeded to fit a generalized linear model with glm function. We also tried cv.glmnet function to tune the hyper-parameters but it was too slow for this corresponding massive dataset. The glm function gave a reasonable performance while predicting loan_status but the 3-split CV error was not below the requisite threshold of 0.45

3.2. XGBoost Model:

For tuning the parameters of the XGBoost model, we tried tuning the hyper-parameters as given in the following table:

Table 1. Hyper-parameter tuning list

Hyper -Parameter	**Tuning Approach**	**Range Considered**
# of Rounds	Fixed	500
Eta (Learning Rate)	Grid-Search	[0.03-0.3]
Max. Depth	Grid-Search	[2-10]
Row Sampling	Grid-Search	[0.5, 0.75, 1]
Column Sampling	Grid-Search	[0.6, 0.8, 1]

We used watchlist to keep a track of logloss on test dataset to prevent model overfitting. We were able to get 3-split CV error much below the requisite threshold of 0.45 with eta=0.15, Max.depth = 8, # of rounds = 135, Row Sampling= Column Sampling=1. However, we make a compromise from the optimal values to fasten up the model training process by opting for eta to be 0.2475 with a max_depth of 7. We report the values for the compromised set of parameters which also clears the threshold comfortably by giving a 3 split CV of 0.4497

4. Results

Table 1. Output matrix representing performance of each model for each data split

Splits	GLM Model	XGBoost Model	Run Time (secs)
Test-1	0.45537	0.44904	809.08
Test-2	0.45631	0.45072	793.15
Test-3	0.45561	0.4496	782.13
Average	**0.45576**	**0.44979**	**794.79**

5. System Information and Run Time

Processor: Intel i5-2500 CPU @ 3.30 GHz

Installed Memory: 12 GB

System Type: 64-bit Operating System, x64-based processor

Operating System: Windows 10

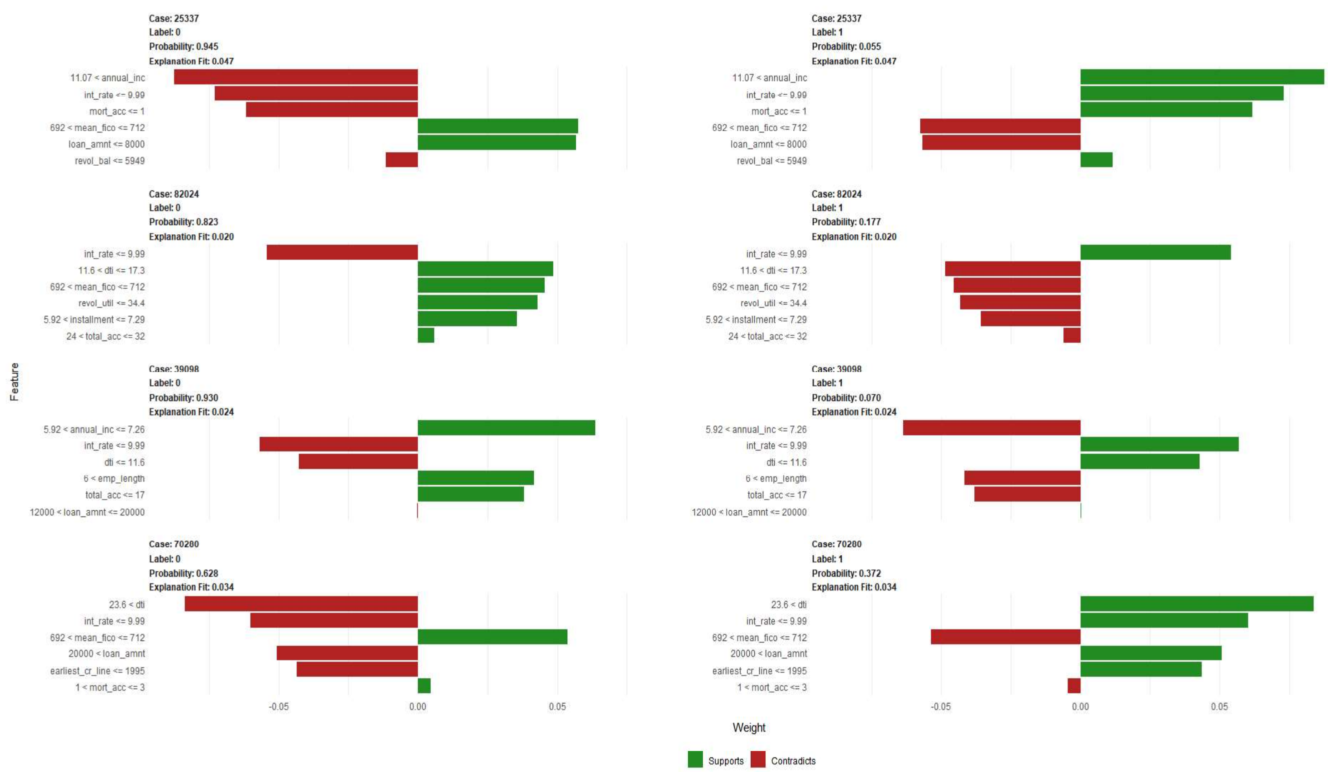
Total Run Time: 2384.36 sec (≈ 39 minutes) [for 3 splits]

6. Performance on New data 2018Q3 and 2018Q4

We had to do a lot of pre-processing for these datasets first by removing columns that weren't present in the original training data. Further, we carried out each of the pre-processing steps mentioned in Section-2 (Data Pre-processing). We are submitting our predictions in 2 files. We report a log-loss of **1.9524 on 2018Q3 data and 1.7151 on 2018Q4** data with the models trained on historical data

7. Model Explanation

We used LIME package to come with feature explanation for randomly sampled 4 cases. The plots are as shown below:



The above Lime plot shows the feature explanation for 2018Q3 on randomly sampled cases. The columns on the left is the feature explicability for label 0 (fully paid) and columns on the right is the feature explicability for label 1. We see that for case 25337, the log of annual income greater than 11.07, interest rate less than 9.99% and mort account being less than 1 contradicts the assignment of label 0 for that particular case while mean_fico_score being between 692 anmd 712, loan amount being less than or equal to 8000 is supporting the assignment of label 0 to that particular tuple. Other details can be interpreted from the plot such as the probability of label 0 for that case is 0.945 and the explanation fit for these 6 factors is 0.047 (which is pretty low for drawing strong conclusions). Similar interpretations can be made for other cases. The plot below shows the LIME plot for 6 features and 2 labels for 2018Q4 data. (We have attached plots with our submission which you can refer to in case of brevity)