Project 3 report

- Data preprocessing

Seven variables with missing values:
Drop the variable emp_title, title, as it has missing values and too many categories.
Convert emp_length to integers and impute mean for missing values.
For dti, revol_util, impute mean for missing values.
For mort_acc, pub_rec_bankruptcies, impute mode for missing values.

The rest variables:
For term, convert term to integers.
For grade, drop grade variable as it is implied by subgrade.
For home_ownership, replace 'any' and 'none' with 'other'.
For annual_int, it is highly skewed, do log transform.
For loan_status, replace 'default' with 'charged off' and code 'charged off' as 1, 'fully paid' as 0.
For zip_code, too many categories and implied by state, drop zip_code.
For earliest_cr_line, extract year for simplicity.
For fico_range_low and fico_range_high, highly correlated, use the average of them as a new variable.
For revol_bal, highly skewed, do log transform.
For id, will be dropped later.

Do one hot encoding for factors.
Do center and scale for numerical variables.

Transform the matrix into Dmatrix for xgboost.

- GBM hyperparameters

Random search the combination of the hyperparameters, after 100 iterations, choose the best combination: {objective = 'binary:logistic', eval_metric = 'logloss', max_depth = 7, eta = 0.2475, gamma = 0, subsample = 1, colsample_bytree = 1, min_child_weight =4}, seed = 419, nrounds = 131.

- Run time

First: 16.17 mins

Second: 16.69 mins
Third: 16.28 mins

- Performance summary

	Test1	Test2	Test3	Average
xgboost	0.446771453463463	0.448288977859764	0.447035494670202	0.4473653

- Computer system

Macbook Pro, 2.3GHz, i5, 8GB memory.