

# STAT542 Project 3: Lending Club

Xiaoman Xie; xiaoman2; 658055980

## Data Pre-processing

1. Pre-processing of the features are described in the table below:

Feature	Pre-processing
Loan_amnt	Take the log of orinal value
term	"36 months" -> "36" "60 months" -> "60"
Emp_length	1. Set missing value to "0 years" 2. "< 1 year" -> "0 years" 3. "10+ years" -> "10 years" 4. Convert to numeric scores, eg "x years" -> x
Home_ownership	Replace "ANY" and "NONE" by "OTHER"
Annual_inc	1. Set annual income that are lower than 500 to be 500 2. Take the log of the updated annual income
Loan_status	"Fully paid" -> 0 "Default" or "Charged off" -> 1
dti	Set missing value to 0
Earliest_cr_line	1. Replace Month-Year structure by the year only 2. Convert to numeric
fico	Create a new feature "fico" as the average of "fico_range_low" and "fico_range_high"
Revol_util	Set missing value to 0
Mort_acc	Set missing value to 0
Pub_rec_bankruptcies	Set missing value to 0

2. Exclude features that are not helpful for the loan status prediction, such as "id" and "zip code". Removed features and reason of removal are listed below:

Removed feature	Reason
id	Not helpful in prediction
title	Not helpful in prediction
Emp_title	Not helpful in prediction
Zip_code	Not helpful in prediction
Fico_range_low	Replaced by "fico"
Fico_range_high	Replaced by "fico"
Grade	Can be well represented by "sub_grade"
Application_type	Not helpful in prediction
initial_list_status	Not helpful in prediction
Addr_state	Not helpful in prediction

3. Data frames are converted to numeric matrix in Xgboost model.

### Hyperparameters

Lgm: family = "binomial"

Lgmnet: family = "binomial"

Xgboost: objective = "binary:logistic", eval\_metric = "logloss", eta = 0.15, nrounds = 200

### Run Time

The user time cost by running the script for all three test set is **1360.031** (CPU seconds). The running time for each test is listed below:

	Test 1	Test 2	Test 3
Run Time	7.79	8.26	8.12

### Computer

Macbook Pro, 2.6 GHz, 16 GB memory

### Performance

	Test1	Test2	Test3	Average
Lgm	0.4550	0.4559	0.4551	0.4553
Lgmnet	0.4564	0.4572	0.4566	0.4567
xgboost	<b>0.4471</b>	<b>0.4486</b>	<b>0.4476</b>	<b>0.4478</b>