2019

Capstone Project I Proposal

LOAN DEFAULT PREDICTION

NGOC PHAN

Problem Statement

Loan default occurs when the borrower fails to make payments on the loan for an extended period of time (LendingClub). According to an article in The Washington Post, “seven million Americans are 90 days or more behind on their auto loan payments in 2018” (Long, Heather). The article, U.S. Debt Statistics, also mentioned that, “every quarter 250,000 borrowers default on federal student loans” (DebtCleanse).

Mission Statement

The purpose of this project is to develop machine learning models that 1) identify the patterns of loan default and the characteristics of defaulter, 2) predict the likelihood of default, 3) minimize loan default rate, and 4) enable better applicant qualification screening process.

Audience

The project would benefit the following audiences: federal government and financial institutions including banks, credit unions, and finance firms, etc.

Dataset

The loan datasets are collected on LendingClub website at https://www.lendingclub.com/info/download-data.action that include 2018 loan data by quarter. Each dataset is in csv format and stored in a zip file. There are 144 columns in the dataset. The list below shows the name of each column in the dataset.

['id', 'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'issue_d', 'loan_status', 'pymnt_plan', 'url', 'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'dti', 'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths', 'mths_since_last_delinq', 'mths_since_last_record', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'initial_list_status', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d',

```
'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d',
'collections_12_mths_ex_med', 'mths_since_last_major_derog', 'policy_code',
'application_type', 'annual_inc_joint', 'dti_joint', 'verification_status_joint',
'acc_now_delinq', 'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_act_il',
'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il', 'il_util',
'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util', 'total_rev_hi_lim', 'inq_fi',
'total_cu_tl', 'inq_last_12m', 'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy',
'bc_util', 'chargeoff_within_12_mths', 'delinq_amnt', 'mo_sin_old_il_acct',
'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op', 'mo_sin_rcnt_tl', 'mort_acc',
'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq',
'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd', 'num_actv_bc_tl',
'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl', 'num_il_tl', 'num_op_rev_tl',
'num_rev_accts', 'num_rev_tl_bal_gt_0', 'num_sats', 'num_tl_120dpd_2m',
'num_tl_30dpd', 'num_tl_90g_dpd_24m', 'num_tl_op_past_12m', 'pct_tl_nvr_dlq',
'percent_bc_gt_75', 'pub_rec_bankruptcies', 'tax_liens', 'tot_hi_cred_lim',
'total_bal_ex_mort', 'total_bc_limit', 'total_il_high_credit_limit', 'revol_bal_joint',
'sec_app_earliest_cr_line', 'sec_app_inq_last_6mths', 'sec_app_mort_acc',
'sec_app_open_acc', 'sec_app_revol_util', 'sec_app_open_act_il',
'sec_app_num_rev_accts', 'sec_app_chargeoff_within_12_mths',
'sec_app_collections_12_mths_ex_med', 'sec_app_mths_since_last_major_derog',
'hardship_flag', 'hardship_type', 'hardship_reason', 'hardship_status', 'deferral_term',
'hardship_amount', 'hardship_start_date', 'hardship_end_date',
'payment_plan_start_date', 'hardship_length', 'hardship_dpd', 'hardship_loan_status',
'orig_projected_additional_accrued_interest', 'hardship_payoff_balance_amount',
'hardship_last_payment_amount', 'debt_settlement_flag', 'debt_settlement_flag_date',
'settlement_status', 'settlement_date', 'settlement_amount', 'settlement_percentage',
'settlement_term']
```

Step-by-Step Problem-Solving Approach

1. Import and clean data
2. Perform exploratory data analysis to identify patterns of loan default and characteristics of defaulter
3. Tell story with data
4. Develop machine learning models
    a. Perform feature extractions using feature engineering
    b. Select appropriate models based on the analysis on step 2
    c. Perform hyperparameter tuning on the models
    d. Evaluate models' performance
    e. Select the best model(s)

Deliverables

The project will be delivered as follows:

1. Code
2. A paper or a blog

References

1. LendingClub. *What is the difference between a loan that is in "default" and a loan that has been "charged off"?* Retrieved August 17, 2019 from https://help.lendingclub.com/hc/en-us/articles/216127747
2. Long, Heather. (2019, February 12). *A record 7 million Americans are 3 months behind on their car payments, a red flag for the economy*. Retrieved August 18, 2019 from https://www.washingtonpost.com/business/2019/02/12/record-million-americans-are-months-behind-their-car-payments-red-flag-economy/?noredirect=on
3. DebtCleanse. *U.S. Debt Statistics*. Retrieved August 17, 2019 from https://debtcleanse.com/debt-statistics/