Data Analysis Report for Loan Default Prediction

Executive Summary

The data analysis report presents an analysis of loan training data to prepare for model training and loan default prediction. The goal is to help lenders better assess the risk of potential borrowers. The dataset `train_data.csv` contains 145 features and 1827125 number of records. Important features are selected based on correlation among features and targets, PCA, and decision tree, etc.

Data Cleaning and Preprocessing

Before conducting any analysis, data cleaning and preprocessing were performed to ensure data quality for building logistic regression machine learning algorithm. The preprocessing steps before moving onto data exploration include:

1. Removing columns that are over 90% missing data.
2. Creating dummy variable ‘loan_status_dv’ for target variable: charged-off = 0, fully paid or current = 1.

The final dataset went through extra processing steps to remove categorical/object and numerical features. The final dataset train_ML would be used for model building.

Exploratory Data Analysis (EDA)

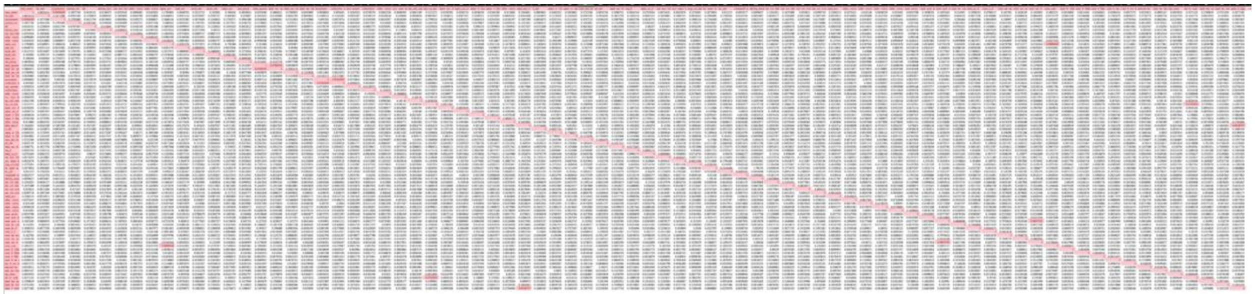
Loan default is highly associated with borrower’s credit score. A few assumptions are considered during EDA. The five major categories for credit score are: payment history(35%), amount owed(30%), length of credit history(15%), credit mix(10%), new credit(10%) (Reference: Wells Fargo). Revolving credit is more impactful than installment. Derogatory and delinquency remarks are considered.

There are 86 numerical features and 23 categorical features. The dataset is highly imbalanced with 92.7% of data are paid off (0). A data profiling report is generated for preprocessed dataset and final EDA’ed dataset. Most data are rightly skewed for numerical variables.

One of the aims of EDA is to reduce as many categorical variables as possible to minimize features used in logistic regression model. The null hypothesis is set to be there is no significant relationship between two categorical variables. Chi-square test is used to test significance and Cramer’s V score is used to determine strength of association. Features with low Cramer’s V score <0.3 and/ or >0.05 p-value are removed. Emp_title and zip_code is removed are also due to large unique values. Among categorical features except time features: term, grade, sub_grade, debt_settlement_flag, emp_length, purpose, title, addr_state, are selected. All time categorical features are deleted.

Among the integer variable, there is high coliniearly among loan_amnt, funded_amnt, and funded_amnt_inv. Investigator(_inv) variable is analognaus to borrowers in this dataset. Correlation between funded_amnt, out_prncp, total_pymnt and funded_amnt _inv, out_prncp_inv, total_pymnt_inv are all above 0.99999.

Correlation matrix (Pearson) of numerical values exported to csv: highlight in red are values >0.9.



41 numerical features (NaNs dropped) were fed into base line logistic regression model to gain insight on feature importance with AUC_score of 0.9936 on validation dataset. Top 10 features are the following:

total_rec_int	6.33E-03	**num_accts_ever_120_pd**	-1.01E-06
loan_amnt	5.67E-03	**mort_acc**	-6.02E-07
recoveries	3.80E-03	**num_tl_op_past_12m**	-4.03E-07
total_rec_late_fee	2.67E-05	**pub_rec_bankruptcies**	-3.22E-07
revol_bal	1.95E-05	**pub_rec**	-2.99E-07

Feature Selection and Engineering

The process of selecting features is:

1. Original train_csv features: 145
2. trn_dropEDA features after EDA: 85
3. trn_dropEDA_time features after time EDA (1 feature added): 75
4. train_ML features after selection from NaNs: 53

53 features are selected for baseline model. With missing values for dti, dti_joint is to replace the missing dti values. Credit history variable 'cr_dur_year' are created to measure credit history last_credit_pull_d - earliest_cr_line.

Conclusion

53 features dataset would serve as the base dataset for machine learning model building. The test data would be processed based on EDA processing procedure. The dataset has highly imbalanced target dataset and large set of irrelevant features. Features may be dropped further for machine learning training.

Reference

Chi-square statistical test: https://towardsdatascience.com/levels-of-measurement-statistics-and-python-implementations-8ff8e7867d0b#:~:text=Chi%2Dsquare&text=It%20is%20a%20test%20to,the%20strength%20of%20the%20correlation.&text=As%20the%20sample%20size%20increases,proportion%20of%20the%20expected%20value.

Loan default Data EDA: https://towardsdatascience.com/machine-learning-predicting-bank-loan-defaults-d48bffb9aee2