Machine Learning Report for Loan Default Prediction

Executive Summary

The goal of building machine learning model for loan default prediction is to predict loan status to be charged-off (encoded as 1) or current/fully paid (encoded as 0). The logistical regression model is used for test data prediction. The best ROCAUC score obtained is 0.77 using training dataset from exploratory data analysis (EDA) process. The test dataset ‘updated_test_data_20200728.csv’ contains 146 features and 211627 number of records.

Data Cleaning and Preprocessing via EDA

Test data are transformed per EDA procedure:

1. Removing columns that are over 90% missing data.
2. Creating dummy variable for target variable: charged-off = 0, fully paid or current = 1.
3. Removing features after preliminary EDA based on high correlation, Cramer’s V scores, and p-values.
4. Investigate time associated features (months/years/days) and create a new calculated feature as credit duration year ‘cr_dur_year’. Removed the rest of time-associated features for model building.
5. Select and remove features with NaNs.

‘index’ and data with more than 90% missing rate are removed. ‘collection_recovery_fee’, ‘out_prncp’, ‘out_prncp_inv’, ‘recoveries’ are all highly correlated with targe varaiable and are missing. The training dataset process from EDA train_ML would be modified to match the data dimension of test dataset. Imputation is conducted for dti and inq_last_6mths.

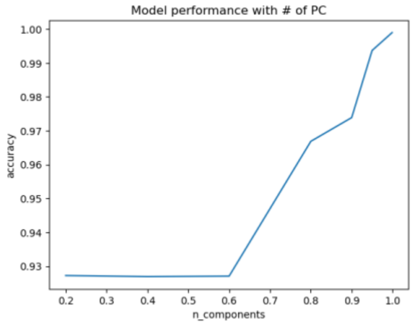
Modeling

Supervised binary classification logistic regression model is used to predict whether borrow would be charged off or stay current. The process flow of logistic regression is the following:

1. Helper functions for one-hot encoding, scoring, plotting ROCAUC
    a. df_trn(df_train, df_test, features_drop=[]):
    b. score(lr, v_val, v_pred): score summary and confusion matrices
    c. plot_AUCROC(lr, v_val, v_pred, v_pred_prob, Name): plot ROCAUC graphs
2. Prepare features and target
3. Split training dataset into training and validation set. Use all test data for prediction
4. Normalization
5. Fit various logistic models
    a. Solver: ‘lbfgs', ‘'newton-cholesky', ‘saga’

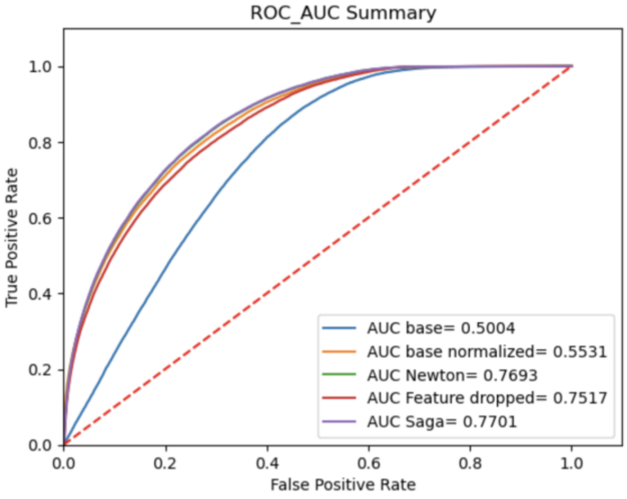
6. Predict use trained model
7. Extract test data result.

PCA analyses show that increasing features would not help with model performance improvement. Using the newton solver model for reduced features, about 0.018 ROCAUC score dropped.



Discussion and Conclusion

Saga solver is the best among all solvers. It presents a huge flexibility in adapting different penalty methods. However, it took the longest time to train and may not be economically efficient. Saga model has low precision and high recall, which is good for detection, however false positives occurrence may be high. Due to the importance of detecting default borrower rather than relying solely on accuracy. Saga model is optimal among the models tested.



In the future, under-sampling can be incorporated due to highly imbalanced dataset due to sufficient quality of data. Categorical features can be investigated further, such as geographic data (zip code, state) and borrower's title. Coefficients from logistic model may contribute to the iterative EDA process of feature selection. Logistic model can be further improved using gridsearch cross validation or using XGBoost to determine result from multiple model results. Interviews should be conducted cross-functionally in finance and business team to understand whether company focuses more are making profit or reducing risk at loosing revenue. Threshold of predication can also be set depending on lender's ability to tolerate risk.

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