Capstone Project: VEHICLE LOAN DEFAULT PREDICTION

Thao Tang | April 11, 2023

Problem statement

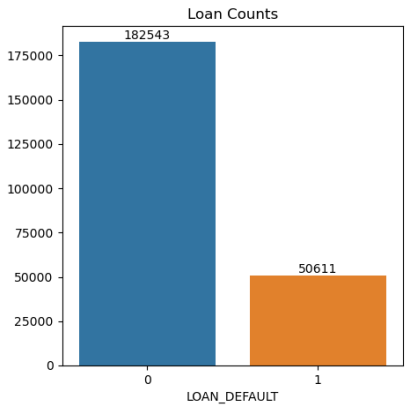
Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting, increased vehicle loan rejection rates and the need for a better credit risk scoring model. This capstone project, therefore, aims to accurately predict borrowers defaulting on a vehicle loan in the first EMI (Equated Monthly Instalments) on the due date, estimate the determinants of vehicle loan default, and differentiate default group from non-default one so that financial institutions can make better decisions on issuing loans, lowering risks and maximizing their profits.

Borrowers' loan applications usually come in large volume, high dimensionality with complex and nonlinear patterns, and predicting defaults or creditworthiness will be greatly valuable to risk management and lending decisions of financial institutions. With machine learning technologies, the adoption of data science will enable new risk-management models with high accuracy and self-learn capability.

Dataset

The dataset is sourced from Kaggle. It was provided by a financial company based in India, named L&T Financial Services, as a part of their recruitment process for Data Scientist position in April 2019[1]. The original data package includes three files: one data dictionary in the form of excel workbook, and two datasets for training and testing purposes in CSV (comma-separated values) format. The test dataset comes without answer key so it was not used in this project.

The training dataset consists of 233,154 loan records with 41 columns including disbursed amount, loan to value of the asset, branch id, date of birth, employment type, information of borrower's primary accounts and secondary accounts, and so on. The target LOAN_DEFAULT column shows the state of each loan record with binary values: 0 as non-default and 1 as default. The



1 Paul, A. (n.d.). *Vehicle Loan Default Prediction*. [online] Kaggle. Available at: https://www.kaggle.com/datasets/avikpaul4u/vehicle-loan-default-prediction?select=train.csv [Accessed 11 Apr. 2023].

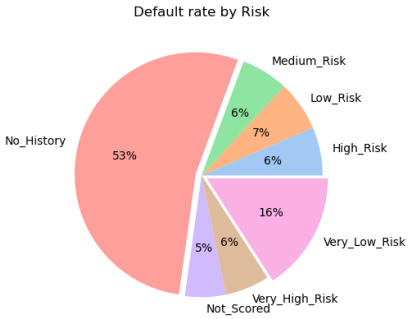
count plot is shown with 50,611 default cases and 182,543 non-default ones.

Data cleaning and preprocessing

The dataset is pretty cleaned without duplicates. However, there are a variety of problems in the dataset, so it still needs data cleaning and transformation before modelling, namely:

1. Filling missing values: EMPLOYMENT_TYPE column carries **7661** missing values, in addition to self-employed and salaried categories. These missing values do not need to be imputed, so I treated them as Unknown.
2. Unit conversion: AVERAGE_ACCT_AGE and CREDIT_HISTORY_LENGTH come with units as the number of years and months, so to make it consistently numeric, I converted them into the number of months.
3. OneHotEncoding: EMPLOYMENT_TYPE and PERFORM_CNS_SCORE_DESCRIPTION
4. Generating features: AGE column was created based on DATE_OF_BIRTH, and DISBURSAL_DATE was extracted to three columms: DISBURSAL_YEAR, DISBURSAL_MONTH, and DISBURSAL_DAY.
5. Dropping features: MOBILENO_AVL_FLAG and DISBURSAL_YEAR were dropped since they contained single value only, **1** for MOBILENO_AVL_FLAG and **2018** for DISBURSAL_YEAR.

After cleaning, exploratory data analysis are conducted showing that there are stark differences in default rate between states, months, and risk groups. In fact, the number of defaults peaked in August, September and October, while some states witnessed high default rate of over 25%, especially Jammu & Kashmir (state id 13) with over 30%. 53% of defaults is caused by Not_Scored borrowers (those with unavailable credit score), and surprisingly, 16% from Very_Low_Risk ones.



Insights, modeling, and results

The dataset was splitted into train and test sets with ratio **3:1**, scaled then fitted to four models, including Logistic Regression, Neural Network, Extreme Gradient Boosting (XGBoost), and Random Forest, in which Logistic Regression acts as a baseline model to compare with the powerful others. Clustering with ground truth is also adopted to picture the default borrowers.

All four models return the highest accuracies of **78.3%** on both train and test sets, meaning that this is the best accuracy that could be learned from this specific dataset. However, this level of accuracy is not ideal at all, hence to improve the accuracy or to more precisely classify the target, the dataset would require additional important features that are highly meaningful to the target. This necessitates domain expertise and is a room for improvement. Regarding the measurement of classification, random forest tends to outperform the other three models, and its threshold **0.2** produced a balanced measurement for both classes, this model, therefore, would be opted.

As a result of Logistic Regression and Random Forest, PERFORM_CNS_SCORE, LTV, and DISBURSED_AMOUNT are of high importance to predict a default borrower. Shuffling values of these features would lower the model accuracies, while a unit increase in PERFORM_CNS_SCORE would increase the odds of default by a factor of **1.62**, and a unit increase in LTV (Loan to Value of the asset) may increase the odds of default by a factor of ~ **1.55**. To put it in a real context, a borrower who has high credit score and apply for a high LTV loan will have higher chance to default than who does not.

	Coefficients	Odds Ratio
PERFORM_CNS_SCORE	0.483877	1.622352
LTV	0.436328	1.547016
ASSET_COST	0.193079	1.212979
PRI_OVERDUE_ACCTS	0.136525	1.146284
AVERAGE_ACCT_AGE	0.114677	1.121511
NO_OF_INQUIRIES	0.106016	1.111840
STATE_ID	0.098081	1.103052
EMPLOYMENT_TYPE_SELF_EMPLOYED	0.089987	1.094160
DELINQUENT_ACCTS_IN_LAST_SIX_MONTHS	0.086775	1.090651
SEC_DISBURSED_AMOUNT	0.081499	1.084912

Clustering with ground truth depicts a default borrower as a person who:

- Applied for high Loan To Value of the asset
- Has high disbursed amount
- Low rank in credit score (PERFORM_CNS_SCORE_DESCRIPTION)
- Fewer loans taken by themselves at the time of disbursement (PRI_NO_OF_ACCTS)
- Purchased from vehicle manufacturers with small ID

Findings and conclusions

Loan to value of the asset and credit score are two utterly important features in predicting a default borrower. However, the accuracy of **78.3%** is not ideal to precisely predict vehicle loan default and estimate its determinants. Therefore, future directions for this project would be to apply domain expertise to generate or add new features that are highly predictive of the target.