AUTO LOAN DEFAULT PREDICTION WITH MACHINE LEARNING

Liman Shen

UTEID: ls45929
Computer Science
College of Natural Science
The University of Texas at Austin

ABSTRACT

Auto loans are an important product to both financial institutions and consumers. Recent economic data shows auto loan default rates are on the rise. Meanwhile, financial institutions are always searching for a better predictive model to assist in the loan approval process. This research takes a deep dive into a dataset with auto loan information, loanee information, bureau data and credit history information. The data preprocessing phase includes encoding descriptive data, removing outliers, filling in missing values, and using Synthetic Minority Oversampling Technique (SMOTE) to oversample the imbalanced dataset. Five machine learning models are trained with the dataset: Logistic Regression, Light Gradient-Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Random Forest. Models are evaluated using accuracy, precision, recall, F1 score, confusion matrix and area under the receiver operating characteristic curve (ROC-AUC curve). The analysis results show that LightGBM outperforms other models for auto loan default prediction with the highest accuracy score.

1. INTRODUCTION

The recent growing auto loan default rates have caught the eyes of major financial institutions in the U.S. A machine learning model with better predictive power on whether an auto loan could default will help both financial institutions and consumers to avoid potential losses. This research focuses on finding the proper machine learning model to predict auto loan default rate. The goal is to find the best suited model to perform the binary classification task.

The first and most important step is to find a suitable dataset for our research. After appropriate data cleaning and engineering, five selected machine learning models are trained using the dataset. The models are assessed and compared using classification model evaluation metrics such as accuracy, precision, recall scores, etc. The model with the best accuracy score is determined as the most fitting model for this task.

Research questions are the following:

1) How should the dataset be cleaned? What to do with descriptive data? What to do with missing data? What to do with outliers?
2) How should the machine learning models be trained? Should the data be standardized? Is the data imbalanced? How should the hyper-parameters be tuned for each model?
3) Which machine learning models outperform the others? Which machine learning models have limitations?

2. RESEARCH BACKGROUND

In the last twenty years, U.S. Auto Loan Delinquency Rates have been steady in the range from 2% to 5% (Bord & Nathe, 2022). Since the emergence of COVID, the rates dipped into the lower range as consumers were staying home more often, commuting and traveling less due to COVID restrictive policies and safety concerns. Various stimulus programs were also quickly introduced by the government to assist the families in financial needs (Pramuk, 2020). These stimulus checks were in place over a year, which contributed to the Auto Loan Delinquency Rates staying low in the following COVID recovery period (Nova, 2021). However, the U.S. Auto Loan Delinquency Rates are on the rise most recently. With both high inflation and growing interest rates due to the recent fast paced rate hikes led by the Fed, more families are in a struggle to make monthly auto payments (LeBeau, 2022). More financial institutions are also increasing their provision for credit losses to prepare for the potential economic downturn ahead (Chang, 2022). Therefore, financial institutions are in need of a better machine learning model to improve the accuracy of auto loan default prediction. In May 2022, American Express launched a default prediction competition on Kaggle with a total of $100,000 prize money (Kaggle, 2022).

Financial institutions are generally looking for better models to make default predictions on all financial products not limited to auto loans, but this research will focus on auto loan default prediction using machine learning models.

3. DATA & VARIABLES

3.1 DATASET

The dataset being used for training the machine learning models for this task is from Kaggle (Dhaker, 2018). There are roughly 233,000 records in the "train" dataset, which will be used for this research. The "train" dataset covers the period of August to October 2018. There is also a "test" dataset with 112,000 records, but the "test" dataset does not include the target variable (y), so it will not be used in this research. Below categories of information are available in the "train" dataset:

- Loanee Information
- Loan Information
- Bureau data & history

3.2 VARIABLES

There are a total of forty features in the dataset. Twenty-one are bureau data, including bureau credit score, number of active accounts, number of overdue accounts, status of other loans, credit history etc. Eleven are loan information, including loan disbursed amount, asset cost, loan to asset value, date of disbursement, etc. The rest eight features are loanee information, including date of birth, employment type, identity proof etc. The dependent variable is whether a loan is defaulted ('1') or not ('0'), so this research is a binary classification task. Figure 1 shows the correlation heatmap among all features and the dependent variable before any data engineering.

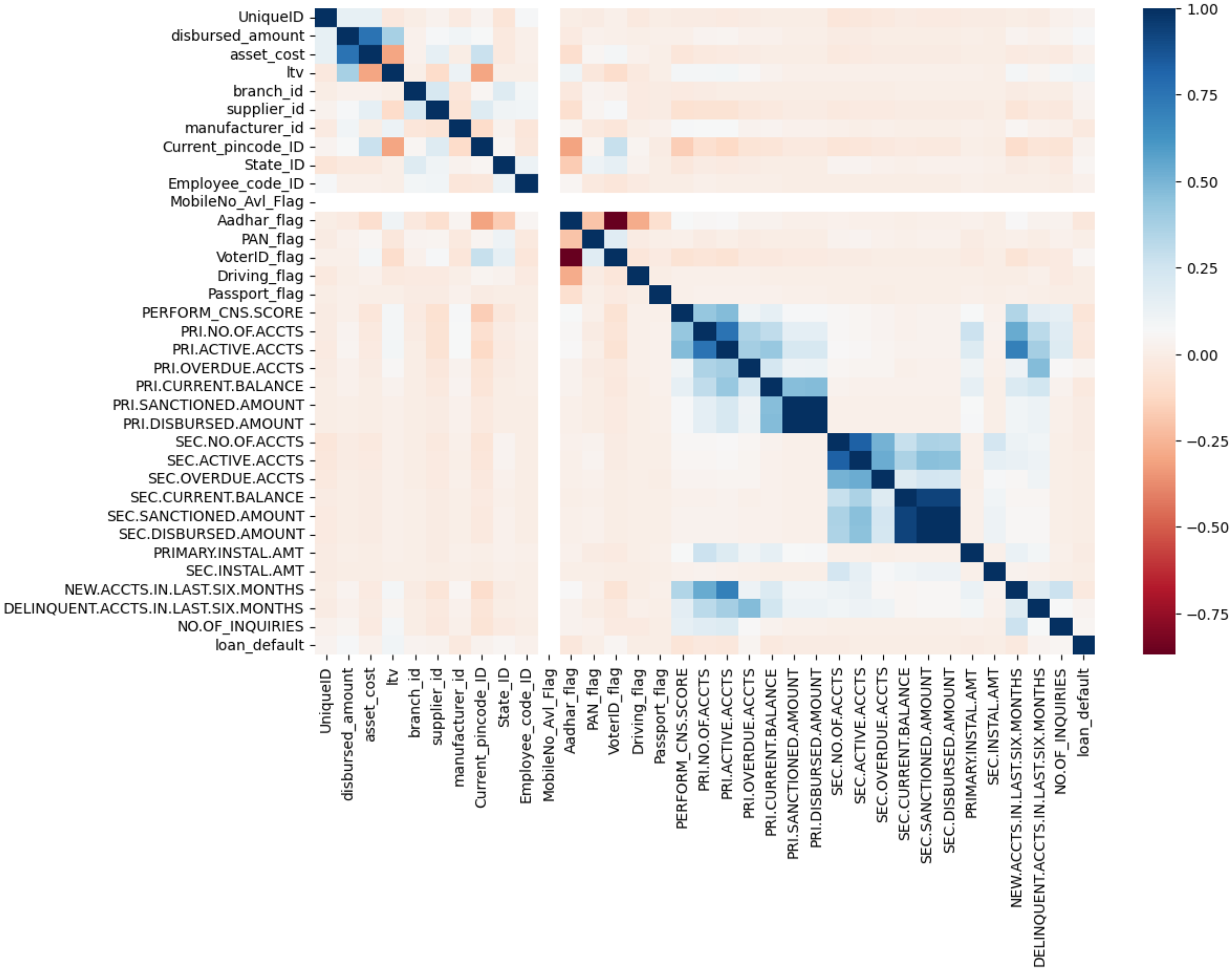


Figure 1. Correlation Heatmap of Features and Dependent Variable

Based on the correlation heatmap and preliminary understanding of the dataset, some columns of data are determined to be immaterial to our research, such as the loan's unique ID, the branch ID where the loan was disbursed, the flag indicating if the loanee provided a mobile number, etc. These irrelevant columns are dropped from the main dataframe. Fortunately, in the entire dataset, only 7,661 employment type data are missing. As the non-missing employment types include only "salaried" and "self-employed", we will interpret the missing employment type as "unemployed". After filling in the missing employment type, the entire column is transformed into categorical values. The columns with date information in string format are converted into numerical values. The bureau credit score and its description are encoded as categorical values. Figure 2 displays the total number of loans disbursed for each credit score category. '0' represents no credit information, '1' represents low credit score (very high credit risk), and ‘5’ represents high credit score (very low credit risk). Figure 3 displays the total number of loans disbursed for each employment type. ‘0’ represents unemployed, ‘1’ represents self-employed, and ‘2’ represents ‘salaried’.

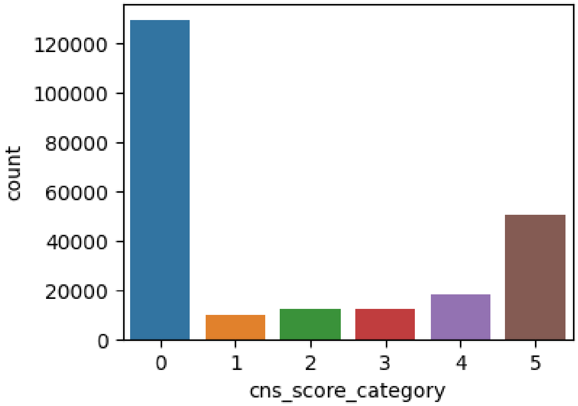


Figure 2. Loans per Credit Score Category

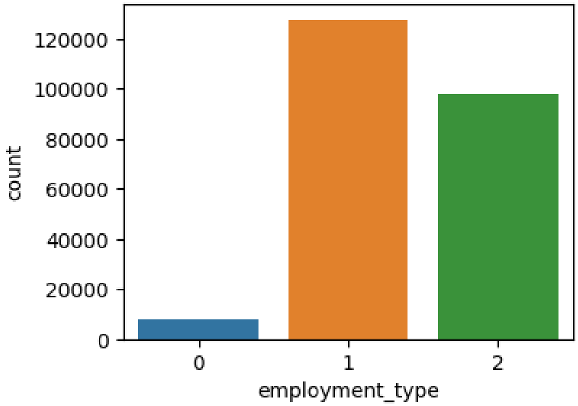


Figure 3. Loans per Employment Type

Bureau data also provide status of other loans which includes both when the loanee is the primary applicant and when the loanee is the secondary applicant, meaning the loanee acts as a co-applicant or a guarantor. The primary and secondary loan information are provided separately so we will consolidate them and include the combined information in our main dataframe.

A few outliers are spotted among the scatter plots of various features. Figure 4 shows the scatter plot between asset cost and disbursed amount, with orange dots being defaulted loans. It is obvious that the majority of the loans are with asset cost under $300,000 and disbursed amount under $200,000. So we will drop the records over this threshold in our main dataframe. Same process is applied to the features exhibited in Figure 5, obvious outliers from these scatter plots are removed from the main dataframe.

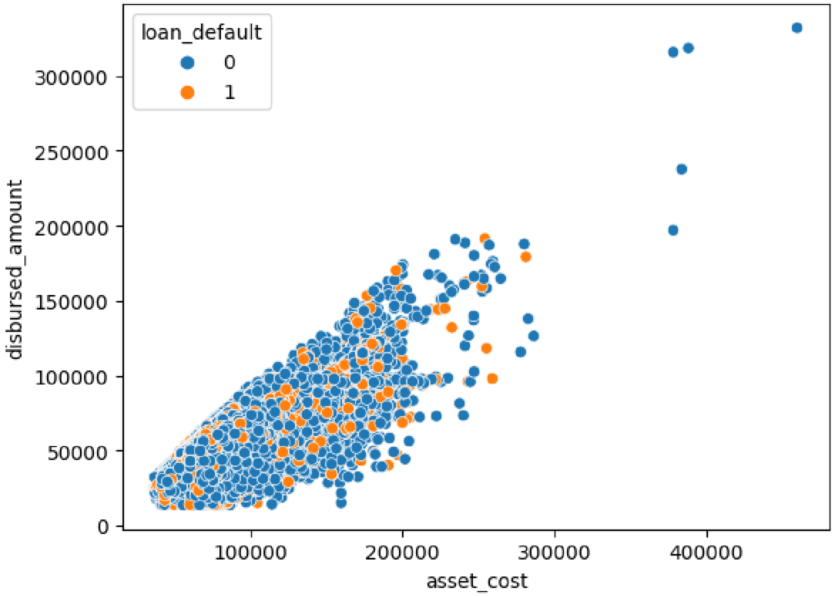


Figure 4. Loan Default Status across Asset Cost and Disbursed Amount

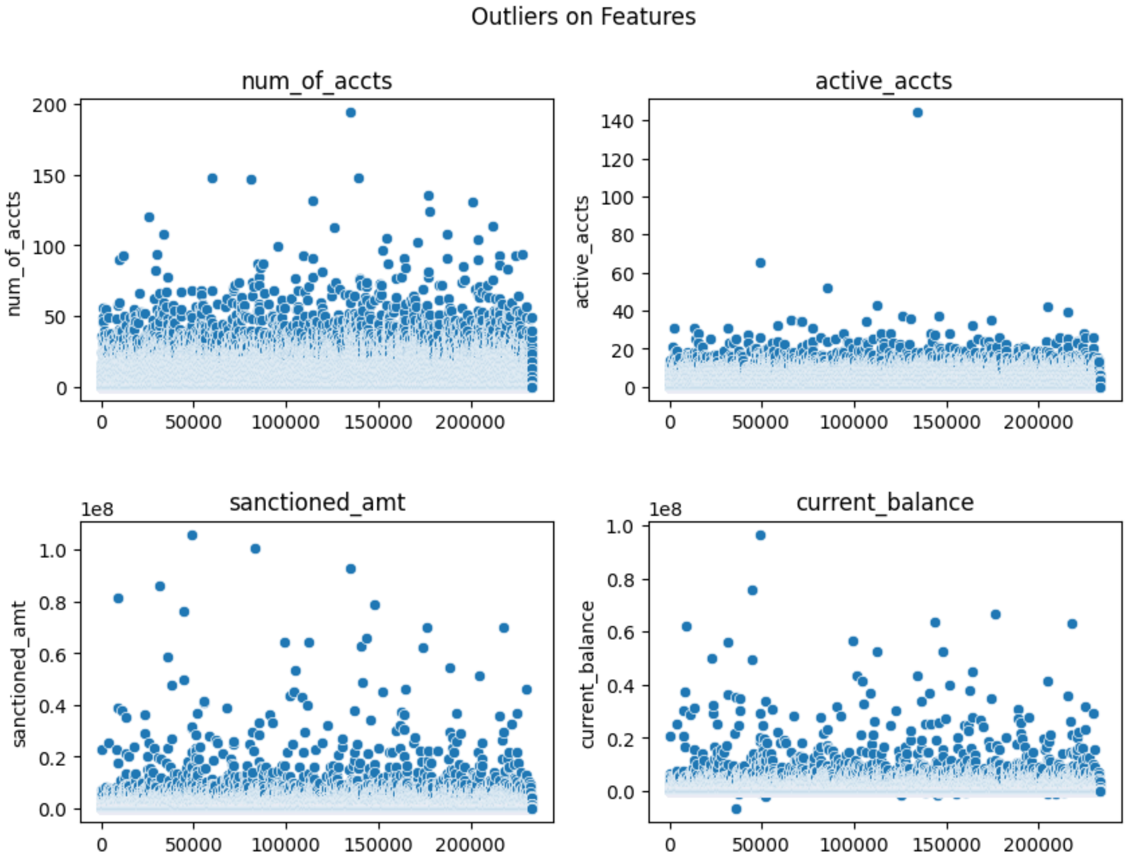


Figure 5. Outlier Observations over Various Features

4. METHODS

As some feature distributions exhibit skewness shown in Figure 6, the dataset must be standardized before being trained by models. Robust scaler is used here to standardize the dataset. Compared to the typical standard scaler, which uses mean and unit variance to perform the standardization, robust scaler uses the median and the interquartile range to compute the standardization. Interquartile range is the range between the first quartile (25%) and the third quartile (75%). This approach often gives better results if the data includes outliers (Hale, 2019).

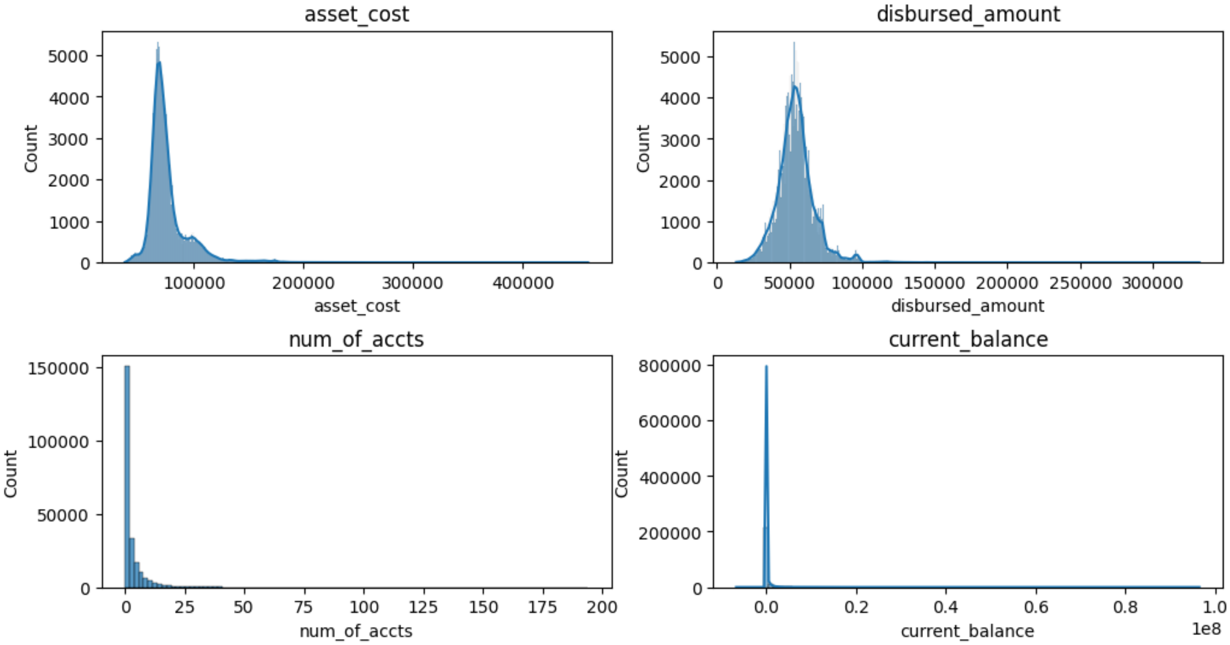
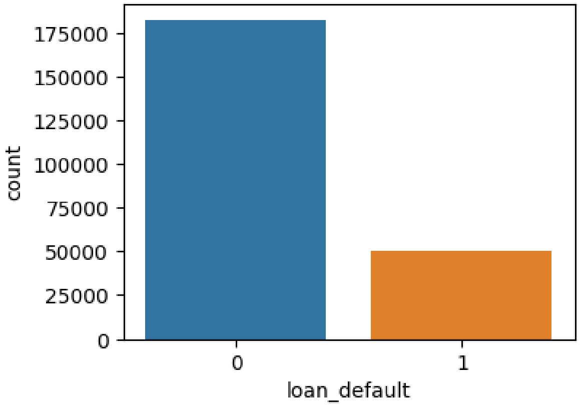
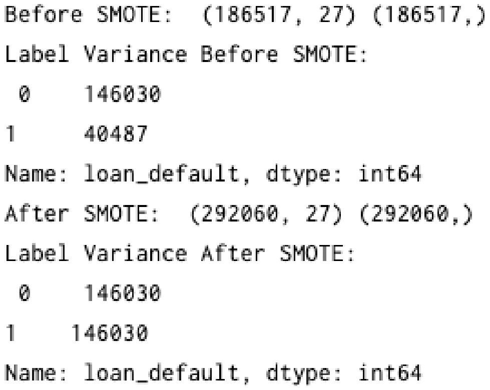


Figure 6. Skewed Distribution of Features

Synthetic Minority Oversampling Technique (SMOTE) is used to oversample the dataset as Figure 7 shows our dataset is imbalanced with 78% of the loans not defaulted and 22% defaulted. SMOTE will pick a random observation from the minority class, and compute the k-nearest neighbors for this observation. Then synthetic data points will be created between the original observation and its neighbors (Brownlee, 2020). After using SMOTE to resample the train dataset, we have a balanced distribution of the dependent variable. Note that SMOTE is only applied to the train dataset, but not to the test dataset.





```
Before SMOTE:  (186517, 27) (186517,)
Label Variance Before SMOTE:
 0    146030
1     40487
Name: loan_default, dtype: int64
After SMOTE:  (292060, 27) (292060,)
Label Variance After SMOTE:
 0    146030
1    146030
Name: loan_default, dtype: int64
```

Figure 7. Data Imbalance

Five machine learning models are selected to train the data for this binary classification task: Logistic Regression, Light Gradient-Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Random Forest. There are many advantages of LightGBM when it is compared to the robust XGBoost, such as faster training speed, better efficiency, and lower memory use. However, it is possible LightGBM will overfit the data without proper parameter tuning (Kasturi, 2019). For Logistic Regression, we used Newton method solver as other methods have potential convergence issues. When training with LightGBM, several hyper-parameters were tuned, such as setting max depth to 10 to limit the tree depth so we can avoid over-fitting, setting the learning rate to 0.1 and num iterations to 1,500 to increase accuracy. Training with XGBoost took the longest time among the five models, we used similar hyper-parameter tuning as LightGBM, with the additional parameter of gamma, alpha, subsample for better accuracy. For KNN model training, we selected 15 neighbors and used ‘distance’ weighting so that closer neighbors have a greater influence. For Random Forest, we limited the max depth, increased the min sample split and min sample leaf to make sure we don’t have a tree model that is too deep.

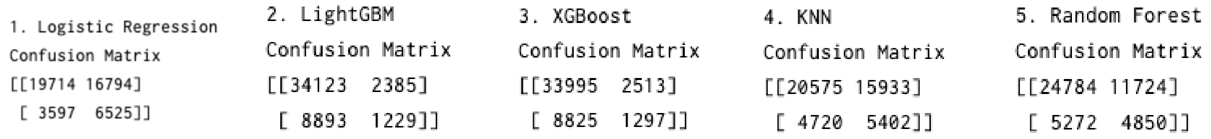
5. RESULTS

We use the metrics accuracy, precision, recall, F1 score, confusion matrix and area under the receiver operating characteristic curve (ROC-AUC curve) to evaluate prediction performance for the five models. However, as our primary goal is to correctly predict whether a loan could default, the most important metrics for our research task is accuracy score.

Table 1 summarizes the evaluation metrics from all models trained. Figure 8 displays the confusion matrix for each model. Logistic Regression and KNN models have similar results with more balanced value across all metrics. Between the two models, Logistic Regression scores higher than KNN on all metrics. However, both models underperform on accuracy scores (below 60%). LightGBM and XGBoost share similar results with higher accuracy and ROC-AUC scores, but they both exhibit lower recall scores. Random Forest model brings results in the middle of the pack, with decent scores on both accuracy rate and recall rate.

Table 1. Model Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
1. Logistic Regression	0.5627	0.2798	0.6446	0.3902	0.6286
2. LightGBM	0.7581	0.3401	0.1214	0.1789	0.6190
3. XGBoost	0.7569	0.3404	0.1281	0.1862	0.6247
4. KNN	0.5571	0.2532	0.5337	0.3435	0.5655
5. Random Forest	0.6353	0.2926	0.4792	0.3634	0.6227



```
1. Logistic Regression
Confusion Matrix
[[19714 16794]
 [ 3597  6525]]
```

```
2. LightGBM
Confusion Matrix
[[34123  2385]
 [ 8893  1229]]
```

```
3. XGBoost
Confusion Matrix
[[33995  2513]
 [ 8825  1297]]
```

```
4. KNN
Confusion Matrix
[[20575 15933]
 [ 4720  5402]]
```

```
5. Random Forest
Confusion Matrix
[[24784 11724]
 [ 5272  4850]]
```

Figure 8. Confusion Matrix for Each Model

As the goal of our research is to find the model with best predictive power on auto loan default rate, achieving higher accuracy is our primary target. Therefore, among the five machine learning models we trained for this task, LightGBM takes the crown. While training the models, LightGBM also outperforms the XGBoost, KNN and Random Forest models on speed. Below figures present further information on the winning LightGBM model. Figure 9 shows the ROC-AUC curve. Figure 10 shows the precision-recall curve. Figure 11 displays the importance of each feature for the LightGBM model. It is not surprising to see among the most significant features, the top three features being loan specific information, and the next two features involving loanee's year of birth and credit score.

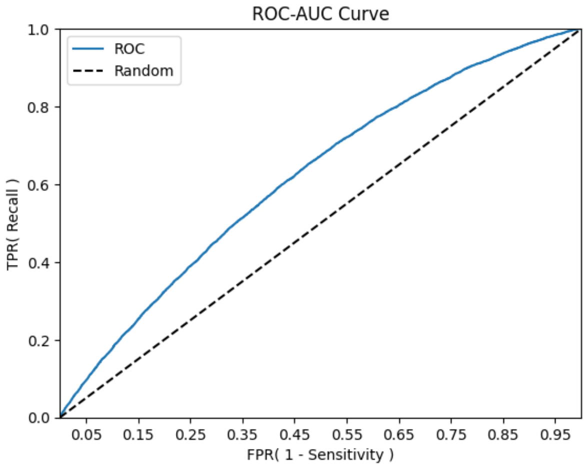


Figure 9. LightGBM ROC-AUC Curve

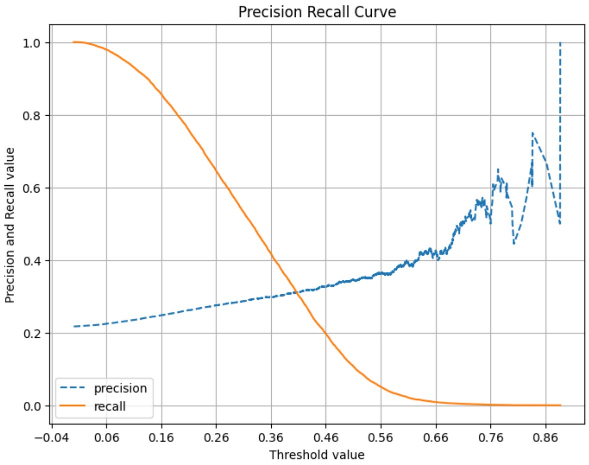


Figure 10. LightGBM Precision-Recall Curve

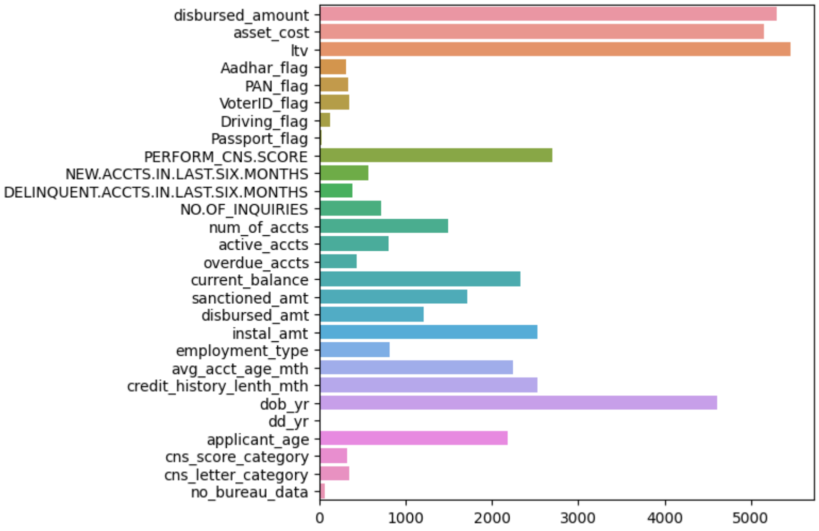


Figure 11. LightGBM Feature Importance

6. CONCLUSION

This research performed heavy data engineering in the preprocessing phase to ensure the dataset is clean before being trained with the machine learning models. These data cleaning steps addressed the initial research questions posed in the introduction section with regards to understanding the dataset and extracting important information from the dataset. As the original dataset is skewed and imbalanced, we used the robust scaler to standardize it and an oversampling technique to rebalance the training dataset. Hyper-parameters were tuned for each model specifically to achieve better accuracy and avoid over-fitting.

Although LightGBM is determined to be the best suited model for this research task due to higher accuracy score compared to other models, it is still disappointing that none of the accuracy scores from the five models exceed 80%. The winning LightGBM model also shows a much lower recall score, and a below 50% precision score. These concerns raise the question that maybe more hyper-parameter tunings could be done to improve the evaluation metrics further. In addition, there could be better fitted classification models that are not evaluated in this research. We will leave these questions and possibilities for future research to uncover.

REFERENCES

Bord, V. M., & Nathe, L. M. (2022, 2 11). *Delinquency Rates and the "Missing Originations" in the Auto Loan Market*. FEDS Notes. Washington: Board of Governors of the Federal Reserve System. https://doi.org/10.17016/2380-7172.3053

Brownlee, J. (2020, 1 17). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Chang, E. (2022, 10 21). Banks Prepare for Consumers Unable to Pay Back Loans. *TheStreet*. https://www.thestreet.com/investing/banks-prepare-for-consumers-unable-to-pay-back-loans

Dhaker, M. (2018). *L&T Vehicle Loan Default Prediction*. Kaggle. https://www.kaggle.com/datasets/mamtadhaker/lt-vehicle-loan-default-prediction

Hale, J. (2019, 3 4). *Scale, Standardize, or Normalize with Scikit-Learn*. Towards Data Science. https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

Kaggle. (2022, 5 25). *American Express - Default Prediction*. Kaggle. https://www.kaggle.com/competitions/amex-default-prediction/

Kasturi, S. N. (2019, 7 11). *XGBOOST vs LightGBM: Which algorithm wins the race*. Towards Data Science. https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d

LeBeau, P. (2022, 11 8). Auto loan delinquencies rise as loan-accommodation programs end. *CNBC*. https://www.cnbc.com/2022/11/08/auto-loan-delinquencies-rise-as-loan-accommodation-programs-end-.html

Nova, A. (2021, 8 6). Pandemic-era relief is drying up. But families still have options. *CNBC*. https://www.cnbc.com/2021/08/06/unemployment-benefits-stimulus-checks-pandemic-era-aid-is-ending-.html

Pramuk, J. (2020, 3 25). Here's what's in the $2 trillion coronavirus stimulus bill. *CNBC*. https://www.cnbc.com/2020/03/25/coronavirus-stimulus-bill-updates-whats-in-the-2-trillion-relief-plan.html