Fraudulent Bank Account Prediction:

A Machine Learning Approach to Improve Bank Security

Ava A. Cook

Department of Economics, University of Wisconsin - Madison

ECON 695: Topics in Economic Data Analysis

Professor Harold Chiang

May 12, 2023

Fraudulent Bank Account Prediction: A Machine Learning Approach to Improve Bank Security

American consumers lost $8.8 billion to fraud in 2022 (Vedova, 2023). Fraud takes a variety of forms; one form is fraudulent bank account applications. While less common, bank fraud is malicious. It decreases confidence in financial institutions, thereby hampering economic performance. In relation to this problem, I asked: How can banks use ML (machine learning) to prevent fraud attempts and improve financial security? More specifically, which ML techniques provide the best prediction accuracy, and which methods minimize Type II error (predictions of non fraud for a fraudulent application)?

Method

Data source

This data comes from the Bank Account Fraud Dataset Suite published at the 2022 NeurIPS Conference. The datasets include real world fraud data which has been anonymized and partially cleaned. The Base dataset contains original data, while each Variant has altered statistical properties. I used the Base dataset and Variant V, which has "additional separability bias"; this means that ML techniques can more easily predict fraud and non fraud (Jesus, 2022).

Assessments and Measures

I approached this question as a binary classification problem. The outcome variable of interest is fraud_bool, a Boolean vector describing whether or not an account was fraudulent. I first cleaned the dataset by replacing missing values with NAs. I then created new variables for prev_address_months_count and bank_months_count. Both of these variables were missing many rows but had significant effects on fraud_bool. The datasets are very unbalanced, with only 1.1% of cases being fraudulent. To address this, I resampled the data using an oversampling

method. I then used LASSO for variable selection. I performed classification using LASSO, logistic, LDA, and QDA models. I did this process once for the Base dataset and repeated it for the Variant V dataset.

My goal in this analysis is to identify the best method for predicting fraud. I considered two performance metrics when deciding which method was "best". First, I considered the prediction accuracy, or how often the method correctly predicted fraud or non fraud. I then considered the Type II error rate, which I defined as the percentage of cases predicted as non-fraud which were actually fraud. Both metrics are important in a bank security setting. A high prediction accuracy implies that the bank can trust the model fairly well. However, due to the unbalanced nature of the data, a model could predict all non fraud cases while achieving a high prediction accuracy. Therefore Type II error rate is also important; a low Type II error rate means that few fraudulent accounts will pass through the model without being flagged as potentially fraudulent.

Findings

For both the Base and Variant V analyses, LASSO provided the lowest Type II error rate and Logistic provided the highest prediction accuracy. Prediction accuracy tended to be higher and Type II error rate tended to be lower in Variant V when comparing method performance across datasets.

Base Dataset Analysis

The models in order of descending prediction accuracy are: Logistic, LDA, QDA, LASSO. The models in order of increasing Type II error rate are: LASSO, QDA, LDA, Logistic. We note the tradeoff between low Type II error rate and high prediction accuracy. We find that a

model with high prediction accuracy may predict mostly non fraud at the expense of pinpointing true fraud cases. Figure 2 displays the prediction accuracy, Type II error rate, and test MSE for each model using Base.

Variant V Dataset Analysis

The models in order of descending prediction accuracy are again: Logistic, LDA, QDA, LASSO. The models in order of increasing Type II error rate are: LASSO, QDA, LDA, Logistic. Even with the better separability in the training data, we still note a tradeoff between low Type II error rate and high prediction accuracy. Figure 3 displays the prediction accuracy, Type II error rate, and test MSE for each model using Variant V.

Implications

According to the performance metrics, models with high prediction accuracy tend to have a higher Type II error rate. This implies that a bank could employ a high prediction accuracy model but expose themself to more fraud risk as a result. Instead, I would suggest employing a low Type II error rate model, such as LASS. These models will flag potentially fraudulent account applications. Further security measures could then be imposed on flagged applications to check their legitimacy. These stricter security measures would be costly but could save the bank money that would have been extracted by fraud. Fraud risk measures could inform cost-benefit analyses for security measures, thereby helping banks decide which prediction model to implement.

References

@misc{jesus2022turning,
title={Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation},
author={Sérgio Jesus and José Pombal and Duarte Alves and André Cruz and Pedro Saleiro and Rita P. Ribeiro and João Gama and Pedro Bizarro},
year={2022},
eprint={2211.13358},
archivePrefix={arXiv},
primaryClass={cs.LG}
}

Iacurci, G. (2022, February 22). Consumers lost $5.8 billion to fraud last year - up 70% over 2020. CNBC. Retrieved April 24, 2023, from https://www.cnbc.com/2022/02/22/consumers-lost-5point8-billion-to-fraud-last-year-up-70percent-over-2020.html

Vedova, H. (2023, February 23). New FTC data show consumers reported losing nearly $8.8 billion to scams in 2022. Federal Trade Commission. Retrieved April 24, 2023, from https://www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022

Figure 1

Variable Subset Selected by LASSO

Outcome Variable

```
fraud_bool
```

Explanatory Variables

```
[1] "income"                                "name_email_similarity"
 [3] "prev_address_months_count"            "current_address_months_count"
 [5] "customer_age"                         "days_since_request"
 [7] "intended_balcon_amount"               "payment_typeAB"
 [9] "payment_typeAC"                       "payment_typeAD"
[11] "payment_typeAE"                       "zip_count_4w"
[13] "velocity_6h"                          "velocity_24h"
[15] "velocity_4w"                          "bank_branch_count_8w"
[17] "date_of_birth_distinct_emails_4w"     "employment_statusCB"
[19] "employment_statusCC"                  "employment_statusCD"
[21] "employment_statusCE"                  "employment_statusCF"
[23] "employment_statusCG"                  "credit_risk_score"
[25] "email_is_free"                        "housing_statusBB"
[27] "housing_statusBC"                     "housing_statusBD"
[29] "housing_statusBE"                     "housing_statusBF"
[31] "housing_statusBG"                     "phone_home_valid"
[33] "phone_mobile_valid"                   "bank_months_count"
[35] "has_other_cards"                      "proposed_credit_limit"
[37] "foreign_request"                      "sourceTELEAPP"
[39] "session_length_in_minutes"            "device_osmacintosh"
```

```
[41] "device_osother"                    "device_oswindows"
[43] "device_osx11"                      "keep_alive_session"
[45] "device_distinct_emails_8w"         "month"
[47] "x1"                                "x2"
[49] "prev_address_na"                   "bank_months_na"
```

Figure 2

Prediction Accuracy, Type II Error Rate, and Test MSE for Base Dataset Analysis

Model	**Prediction Accuracy**	**Type II Error Rate**	**Test MSE**
LASSO	0.8076	0.0023	0.9946
Logistic	0.9889	0.0111	0.0111
LDA	0.9882	0.0174	0.9940
QDA	0.8448	0.0064	1.4502

Figure 3

Prediction Accuracy, Type II Error Rate, and Test MSE for Variant V Dataset Analysis

Model	**Prediction Accuracy**	**Type II Error Rate**	**Test MSE**
LASSO	0.8270	0.0034	0.1310
Logistic	0.9900	0.0100	0.0100
LDA	0.9897	0.0090	0.9946
QDA	0.8806	0.0040	1.3427