Advanced Fraud Prevention System :

Machine Learning in Credit Card Fraud Detection

Rahul Raj PR

APDS - IIIT Bangalore & Upgrad

17th - Jan 2023

Agenda

- Objective
- Background
- Key Insights
- Cost Benefit Analysis
- Appendix :

    Data Source

    Data Methodology

    Attached Files

Objective

To introduce the implementation of a machine learning-based credit card fraud detection system, which aims to significantly reduce the costs incurred from fraudulent transactions.

Background

- By utilizing machine learning techniques, we have developed a model that detects fraudulent activity at an early stage, minimizing financial losses.
- A thorough cost-benefit analysis has been conducted, demonstrating the cost-effectiveness of deploying the machine learning-based fraud detection system.

Key Insights

The top five features in predicting fraud transactions appear to be the transactional amount, hour of transaction, and categories such as gas transport, grocery point-of-sale, and online shopping.

SN	Features	Importances
1	amt	0.171255
2	category_gas_transport	0.140089
3	trans_hour	0.106834
4	category_grocery_pos	0.066731
5	category_shopping_net	0.059787
6	category_food_dining	0.052406
7	category_home	0.049857
8	trans_dayofweek	0.046153
9	category_shopping_pos	0.037699
10	category_entertainment	0.033398
11	trans_month	0.032227
12	timedelta_last_trans	0.031129
13	category_misc_net	0.029177
14	state	0.027372
15	category_personal_care	0.026281
16	category_travel	0.020758
17	category_misc_pos	0.017289
18	category_health_fitness	0.015217
19	category_kids_pets	0.010693
20	cust_age	0.008693
21	gender_f	0.007082
22	city_pop	0.003499
23	lat_dist_cust_merch	0.001698
24	lat_dist_prev_merch	0.001655
25	long_dist_prev_merch	0.001534
26	long_dist_cust_merch	0.001486

Current Incurred Losses

- 77,183 credit card transactions were processed on an average in a month
- 402 cases of fraud were identified as fraudulent transactions
- Average transaction amount for fraudulent transactions was $530
- Total cost incurred due to fraud was $213,392

After New Model Deployment

- The model detected 325 out of 402 total fraudulent transactions per month (approx. 81%)
- The cost of providing customer support for these transactions was $488
- The model was not able to detect 77 fraudulent transactions, resulting in a loss of $40,758
- The total cost incurred after deploying the new model was $41,246
- The new model resulted in a savings of $172,146, a reduction in losses of approximately 80.67%.

Appendix: Data Attributes

variable	description
trans_date_trans_time	A combination of date and time of the transaction.
cc_num	Credit card number used for the transaction.
merchant	The name of the merchant where the transaction took place.
category	The category of goods or services purchased.
amt	The amount of the transaction.
first	The first name of the cardholder.
last	The last name of the cardholder.
gender	The gender of the cardholder.
street	The street address of the cardholder.
city	The city of the cardholder.
state	The state of the cardholder.
zip	The zip code of the cardholder.
lat	The latitude of the cardholder address.
long	The longitude of the cardholder address.
city_pop	The population of the city of the cardholder.
job	The job title of the cardholder.
dob	The date of birth of the cardholder.
trans_num	The unique transaction number.
unix_time	The time of the transaction in UNIX time format.
merch_lat	The latitude of the merchant location.
merch_long	The longitude of the merchant location.
is_fraud	A binary variable indicating whether the transaction is a fraud or not.

Appendix: Data Methodology

- The dataset used in this study was a simulated dataset from Kaggle, which was characterized by its high degree of imbalance and large size.
- To address this, various resampling methods were employed, including sklearn resample, SMOTE, and ADASYN.
- A total of 12 models were trained using decision trees, random forests, XGBoost, and hybrid ensemble chain models. Hyperparameter optimization techniques such as Halving Random Search CV and Bayesian search CV were used to optimize the models.
- The performance of the models was evaluated using ROC-AUC as the primary metric, with a focus on achieving a balance between precision and recall.
- XGBoost with Bayesian hyperparameter optimization was found to be the best algorithm for the dataset, and performance was further improved through the use of libraries such as feature_engine for feature encoding and transformation.

Attached Files

- RootCauseAnalysis.pdf
- Cost Benefit Analysis.xlsx
- CreditCardFraud_detection.ipynb