

Final Project Proposal

Credit Card Frauds Detection

INFO 7390

Advanced Data Science & Architecture



Professor:

Srikanth Krishnamurthy

Students:

Team 2

Chenlian Xu, Qianli Ma

Background

Fraud means obtaining services/goods and/or money by unethical means, and is a growing problem all over the world nowadays. Fraud deals with cases involving criminal purposes that, mostly, are difficult to identify. Credit cards are one of the most famous targets of fraud but not the only one; fraud can occur with any type of credit products. A critical task to help businesses, and financial institutions including banks is to take steps to prevent fraud and to deal with it efficiently and effectively, when it does happen.

Credit card frauds detection is a classic and popular topic in nowadays data science fields. A lot of companies are looking for professional data analysts to develop fraud detection solution, seeing the hidden pattern behind the data.

The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Goals

1. Accurately predict credit card fraud cases using imbalanced data
2. To protect card holders' and commercial banks' benefits
3. Review and compare different fraud detection techniques and select the best model by trade-offs of different evaluation methods

Data

1. <https://www.kaggle.com/mlg-ulb/creditcardfraud>

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

The datasets contain transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Process Outline

1. Data Cleaning
2. EDA
3. Feature Engineering
4. Model Fitting
 - 5.1 Biased penalty SVM
Resample fraud cases and apply
 - 5.2 Logistic Regression
 - 5.3 Random Forest
 - 5.4 Gradient Boosting Tree
5. Feature Selection
6. Model Selection:
R2, Accuracy Rate, Recall Rate, Precision, AUC
7. Pipeline design
8. Deployment
9. Web API

Milestone

Timeframe	Delivery
Day 1-2	Data Cleaning and EDA
Day 3-4	Feature Engineering, Model Fitting, Feature Selection
Day 5-8	Model Selection
Day 9-10	Pipeline Design, Deployment, Web API