Proposta de Mecanismo para Detecção de Transações de Cartão de Crédito Fraudulentas.

Douglas A. Clementino[1], **Rafael de Paulo Dias**[1]

[1]Departamento de Informática
Universidade Federal do Paraná (UFPR) – Curitiba – Brasil

[1]{dac17,rpd17}@inf.ufpr.br

1. Introdução

1.1. Descrição

A partir da dataset 'Credit Card Fraud Detection' disponibilizado na plataforma Kaggle [ULB 2021], este trabalho buscará descrever processo de construção de uma ferramenta para detecção de transações de crédito fraudulentas, possibilitando a agilização das respostas a esse ataque por parte das credoras.

1.2. Motivação

- Detectar operações de crédito fraudulentas.
- Impedir reincidência em aprovação de crédito fraudulenta.
- Propor modelo base para detecção de operações fraudulentas de crédito, que futuramente pode ser aprimorado para senários reais, propensos a maiores diversidades de ataques (resultando em *Concept Drift*).

2. Visão Geral

2.1. Coleta e Pré-Processamento

A base de dados foi coletada diretamente da plataforma Kaggle, ela possui formato CSV e possui as seguintes variáveis:

- **Time:** Diferença de tempo decorrido desde a primeira amostra de dataset.
- **Amount:** Quantia de crédito solicitada.
- **Class:** Classe da transação, sendo '0' o valor para transações válidas e '1' o valor para transações fraudulentas.
- **v1-v28:** Variáveis escalares que, segundo o próprio fornecedor da base, seriam decorrentes do processo de redução de dimensionalidade a través da técnica PCA (*Principal Component Analysis*). Dessa forma, possibilitando a distribuição dos dados sem comprometer as informações pessoais dos clientes.

Dessa forma, exceto pelas 3 primeiras variáveis, não é possível efetuar combinação de variáveis a fim de enriquecer a base de dados final.

2.2. Distribuição das Classes

Como descrito anteriormente, a base de dados está dividida entre transações válidas, indicadas pelo valor '0' em varável 'Class' (sendo elas 284.315, representando 99,82% da base total) e as transações fraudulentas, indicadas pelo valor '1' (sendo elas 492, representando 0,17% da base total). Dessa forma, está claramente evidenciado um desbalanceamento do dataset.

3. Planejamento Futuro

A seguir, serão descritas os pontos que pretendem ser abordados a fim de melhorar a performance do modelo gerado dado o cenário proposto.

3.1. Dados Desbalanceados

A fim de tratar do desbalanceamento na presença de classes, serão propostas as seguintes abordagens:

3.1.1. Balanceamento

A fim de enriquecer a base de treino e melhorara a qualidade do modelo final, pretende-se aplicar a técnica de SMOTE (*Synthetic Minority Over-sampling Technique*) a fim de efetuar a geração sintética de entradas de classe minoritária, ou *Oversampling* e *Undersampling*, a fim de obter balanço entre classes em base de treino.

3.1.2. Algorítmos

Buscará utilizar algoritmos para classificação que obtém desempenho satisfatório em situações de base de dados desbalanceada, como *Decision Tree* e *Adaptive Random Forest*.

3.2. Remoção de *Outliers*

Planeja-se efetuar a remoção de entradas *Outlier* de dataset. Porem, dada a disparidade na representatividade de classes, espera-se a obtenção de bons resultados na remoção de *outliers* apenas após a aplicação de técnicas para balanceamento (SMOTE, *Oversampling* e *Undersampling*), caso contrário, espera-se uma disparidade ainda maior na representação das classes, prejudicando o processo de aprendizagem.

4. Resultado Esperado

Ao final do processo, espera-se definir um modelo que sirva de ferramenta para a identificação de transações fraudulentas em ambientes estáticos, e que possa servir de referência para definição de modelos que sejam executados em ambientes reais, que são dinâmicos e propensos à *Concept Drift*.

Referências

ULB, M. L. G. (2021). Credit card fraud detection. `https://www.kaggle.com/mlg-ulb/creditcardfraud/`. Acessado em: 09/11/2021.