INSTITUTO DE PESQUISAS TECNOLÓGICAS

COMUNICAÇÃO TÉCNICA

Nº 178937

Machine learning in e-commerce fraud detection: a systematic literature review and comparative analysis of advanced techniques.

Thiago Matos Rocco
Adriano Galindo Leal

Palestra apresentada INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT, 20., CONTECSI, São Paulo. 11 slides.

Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br

Machine Learning in E-Commerce Fraud Detection: A Systematic Literature Review and Comparative Analysis of Advanced Techniques

Thiago Matos Rocco

Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT)

Adriano Galindo Leal

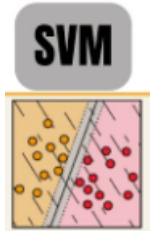
Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT)

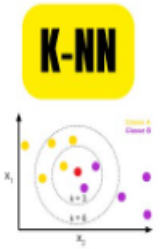
1 - Introdução

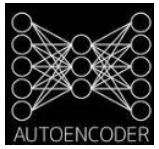
- O Covid-19 trouxe uma nova realidade ao mundo de e-commerce.
- Desafio de combater fraudes nos dias atuais.
- Utilização de aprendizado de máquina como apoio para detecção de fraude e como adapta-las a mudanças.
- A detecção de fraudes no comércio eletrônico é uma tarefa contínua que exige vigilância constante e soluções inovadoras.
- O objetivo desta pesquisa é desenvolver e avaliar modelos de Machine Learning eficazes para a detecção de fraudes em transações de cartão de crédito no e-commerce.

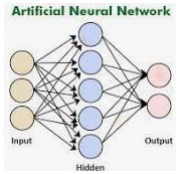
2- Referencial Teórico

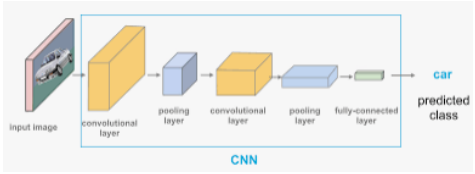
- Tipos de aprendizado de máquina
  - Supervisionados
    - Support Vector Machine
    - Random Forest
    - K-Nearest Neighbors
    - Decision Tree
    - Regressão Logística
  - Não Supervisionados
    - AutoEnconder
  - Redes Neurais
    - Artificial Neural Network
    - Convolutional Neural Network
    - Transformers

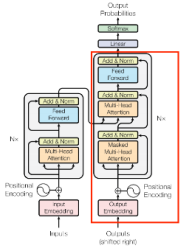










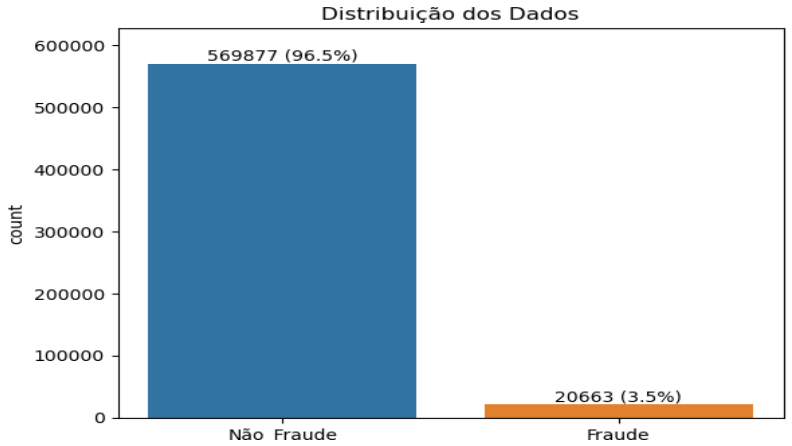


3- Metodologia

- Dataset disponibilizado por IEEE-CIS Fraud Detection, onde temos 434 variáveis e 590.540 mil linhas.

Dados para treino dos modelos		
Nome do dataset	Número de linhas	Número de variáveis
Train_transaction	590.540	394
Train_identity	144.233	41

- Quantidade por tipo de transação (Fraude e Não fraude).

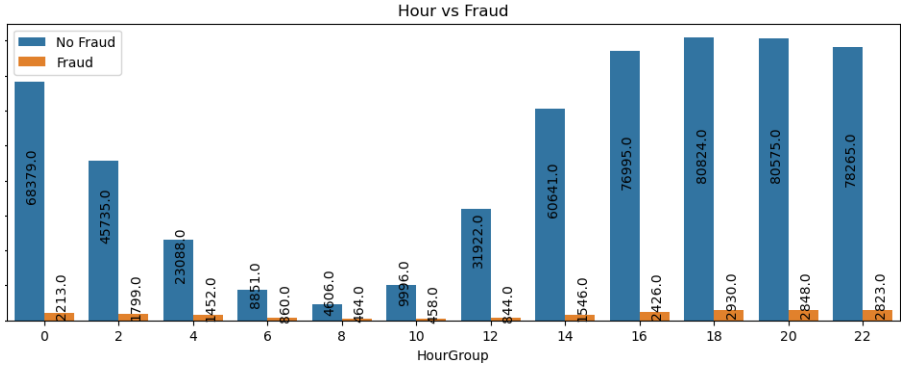


3- Metodologia

Variável	Tipo	Descrição
TransactionID	Inteiro	Número único da transação
isFraud	Booleano	Conteúdo 0 = Não fraude, 1 = fraude
TransactionAMT	*Double*	Valor em dólar da transação
ProductCD	Caracter	Informa o produto comprado apenas por uma sigla, não é possível ver a categoria.
Card 1 até Card 6	*Double* e Caracter	Informações do cartão utilizado.
Addr1 e Addr2	*Double*	Indica o País e região do endereço do comprador
Dist1 e Dist2	Caracter	Campo relacionado a distancia.
P_emaildomain, R_emaildomain	Caracter	E-mail do comprador.
C1 até C14	*Double*	Quantidade de informações diferentes relacionadas ao cartão.
D1 até D15	*Double*	Indica de forma anônima a representação de tempo x compra.
M1 até M9	Booleano e Caracter	Indica se há informações que coincidem com o cartão do comprador.
V1 até V339	Inteiro	São features geradas pela IEEE-CIS, não tem uma descrição clara.
ID_01 até ID_38	Booleano e Caracter	
DeviceType e DeviceInfo	Caracter	Identifica a origem do dispositivo e modelo do dispositivo de compra.

3- Metodologia

- Análise de fraude agrupado de 2 em 2 horas.



- Tratamento de valores NaN.
- Padronização de Colunas.
- Criação de novas features para facilitar a análise dos dados.
- Análise de outros trabalhos realizados para capturar novos insights.
- Utilização de normalização, redução de dimensionalidade, balanceamento e refinamento dos modelos.

3- Metodologia

- O campo IsFraud é a coluna de saída que você precisa prever. Ele indica se uma transação é fraudulenta (1) ou não fraudulenta (0).

- Os dados de saída possuem características de cada transação, ajudando a treinar o modelo ensinando-o o que é fraude ou não.

- Os dados de saída podem ser usados para desenvolver modelos de aprendizado de máquina para prever fraudes.

4- Resultados

- O modelo Random Forest apresentou um desempenho melhor dentro do conjunto de dados. Importante ver que o algoritmo transformers não ficou tão distante.

Modelo	Acurácia
Random Forest	95,5%
Transformer	90,1%
Decision Tree	89,7%
SVM	89,3%
KNN	86,1%
ANN	85,8%
Auto Encoders	84,4%
CNN	83,7%
Logistic Regression	73,3%

Comparação de resultados

4- Resultados

Model	TP	FN	FP	TN
Random Forest	47,9% 109.078	2,1% 4.856	1,8% 3.992	48,3% 110.025
Transformers	41,8% 8.130	8,5% 1.657	32,7% 6.369	17,0% 3.302
Decision Tree	44,3% 101.059	5,6% 12.875	3,6% 8.282	46,4% 105.735
SVM	43,8% 99.843	7,9% 18.094	2,8% 6.326	45,5% 103.689
KNN	44,9% 102.332	5,1% 11.602	1,4% 3.276	48,6% 110.741
ANN	45,7% 104.089	4,3% 9.845	4,9% 11.199	45,1% 102.818
Auto Encoders	82,6% 97.612	13,9% 16.363	1,8% 2.086	1,7% 2.047
CNN	43,3% 98.759	6,7% 15.216	9,7% 22.052	40,3% 91.924
Logistic Regression	41,0% 93.382	9,0% 20.552	17,7% 40.262	32,4% 73.755

Matriz de Confusão

5 - Conclusões

- O destaque principal é a alta performance do algoritmo Random Forest, com precisão de 95,5%, seguido pelo modelo Transformers alcançando 95,0%.
- Métodos de conjunto como Random Forest são efetivos em tarefas de classificação com dados de alta dimensão.
- Características temporais, como a hora da transação, revelaram papel crucial na identificação de fraudes.

6 – Limitações da Pesquisa

- Viés
- Outros Fatores Não Considerados
- Validação Temporal Limitada
- Falta de dados reais
- Recurso computacional