PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

Jéssica Domeneghini de Assis

PREVISÃO DE DETECÇÃO DE FRAUDES EM TRANSAÇÕES FINANCEIRAS UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Belo Horizonte
Agosto/2023

Jéssica Domeneghini de Assis

PREVISÃO DE DETECÇÃO DE FRAUDES EM TRANSAÇÕES FINANCEIRAS UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina, como requisito parcial à obtenção do título de *Especialista*.

Belo Horizonte

Agosto/2023

SUMÁRIO

1. Introdução

Em um cenário em constante evolução, o tema da detecção de fraudes em pagamentos bancários continua a despertar preocupações e ações. A cada ano, a fraude se torna uma indústria que movimenta bilhões de dólares, impactando setores diversos, como comércio eletrônico, saúde e sistemas de pagamento.

A detecção de fraudes se refere à identificação e prevenção de atividades fraudulentas, que envolvem ações enganosas ou ilegais com o objetivo de obter vantagens financeiras, informações pessoais ou outros ganhos indevidos.

Uma abordagem antiga para combater fraudes envolvia o uso de sistemas de regras, que eram definidas manualmente e usadas para tomar decisões sobre a autenticidade de transações, atividades ou comportamentos suspeitos. No entanto, essa abordagem apresenta diversas limitações, entre elas, a falta de adaptabilidade, que permite que novos tipos de fraude passem despercebidos até que as regras sejam atualizadas manualmente. Além disso, há o desafio do aumento do volume e complexidade dos dados, tornando a gestão por meio de regras estáticas uma tarefa difícil, e também vale ressaltar que, a configuração inadequada das regras pode resultar em ocorrências de "falsos positivos" ou "falsos negativos". Essa situação pode causar tanto inconvenientes para os clientes legítimos quanto perdas financeiras para a organização.

Nesse contexto, o potencial do Aprendizado de Máquina emerge como uma ferramenta poderosa para aprimorar a precisão e a eficácia da detecção de fraudes, e é o principal objetivo deste projeto, encontrar um modelo que possa ser capaz de prever se uma transação é fraudulenta ou não. Esses modelos de Inteligência Artificial podem analisar grandes volumes de dados, identificar padrões sutis e aprender com novos exemplos de fraudes à medida que surgem, tornando-se mais adaptáveis e eficazes em comparação com sistemas de regras estáticas, além de reduzir os falsos positivos e negativos, com isso, os bancos podem reduzir riscos financeiros, aumentar a confiabilidade do sistema de pagamento e manter a confiança dos clientes.

2. Descrição do Problema e da Solução Proposta

A detecção de fraude em transações financeiras envolve a identificação de atividades suspeitas que se desviam dos padrões legítimos. Essas atividades podem causar prejuízos significativos para o sistema de pagamentos financeiros.

O desafio deste projeto será explorar um dataset de grande volume de dados e aplicar modelos de aprendizado de máquina que possam identificar padrões e anomalias, e prever a probabilidade de fraudes em novas transações, classificando-as em fraudulentas ou não fraudulentas, reduzindo casos de falsos positivos e negativos, melhorando a eficácia na detecção de transações bancárias fraudulentas, contribuindo para a confiabilidade do sistema bancário e mantendo a confiança da organização.

A solução proposta envolve a aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina para aprimorar a detecção de fraudes em transações financeiras. Serão utilizados algoritmos de Aprendizado de Máquina, como Random Forest e XgBoost (Extreme Gradient Boosting), para cumprir os objetivos estabelecidos. Esses algoritmos de Aprendizado de Máquina são amplamente utilizados para tarefas de detecção de fraude financeira devido à sua capacidade de lidar com dados complexos e não lineares. Eles são considerados métodos ensemble, o que significa que combinam as previsões de vários modelos individuais para melhorar o desempenho geral.

O modelo Random Forest é baseado em árvores de decisão, ele cria múltiplas árvores de decisão durante o treinamento e combina suas previsões para chegar a uma decisão final. Cada árvore é treinada em uma amostra aleatória dos dados, permitindo a diversificação e reduzindo o risco de overfitting. O algoritmo também utiliza o conceito de "bagging" (bootstrap aggregating), que ajuda a reduzir a variância das previsões. Já o modelo XGBoost também é baseado em árvores de decisão, e usa uma abordagem de boosting, que consiste em melhorar o desempenho de modelos de aprendizado fracos, combinando-os em um modelo forte, com isso, pode lidar bem com desequilíbrios entre classes. Ambos os algoritmos são altamente configuráveis, permitindo ajustes para otimizar o desempenho.

A tarefa de Aprendizado de Máquina consistirá em treinar esses modelos com um conjunto diversificado de dados de transações financeiras, incluindo informações sobre os clientes e histórico de transações. Ao refinar continuamente o modelo e mantê-lo atualizado com novos dados, as organizações podem ficar um passo à frente dos fraudadores e proteger a si mesmas e seus clientes contra perdas financeiras.

3. Coleta de Dados

Os dados que serão utilizados neste projeto foram obtidos em julho de 2023, na plataforma Kaggle, onde também está disponível um link para acessar o artigo original sobre a geração dos dados do BankSim.

BankSim é um simulador de pagamentos bancários baseado em agentes, ou Agent-based Modeling em inglês, abordagem utilizada para sistemas complexos, nos quais os agentes (entidades individuais) interagem de acordo com regras pré-definidas. Isso permite observar os padrões emergentes resultantes de suas interações. Neste caso, o simulador utiliza dados agregados de transações fornecidos por um banco na Espanha. Seu principal objetivo é gerar dados sintéticos para pesquisa em detecção de fraudes, combinando pagamentos normais e assinaturas de fraudes conhecidas. Foram realizadas análises estatísticas e Análise de Rede Sociais (Social Network Analysis) para desenvolver e calibrar o modelo, analisando as relações entre comerciantes e clientes.

Este conjunto de dados gerado sinteticamente consiste em pagamentos de vários clientes feitos em diferentes períodos de tempo, e com diferentes valores. O dataset possui um total de 594.643 registros, dos quais 587.443 são transações normais e 7.200 são transações fraudulentas. Contém 10 variáveis sendo uma, a variável de classe (target).

Nome do dataset: Synthetic data from a financial payment system **Descrição:** Conjuntos de dados gerados pelo simulador de pagamentos BankSim **Link:** https://www.kaggle.com/datasets/ealaxi/banksim1		
Nome do Atributo	**Descrição**	**Tipo**
Step	Dia do ínicio da simulação, tem 180 steps, referente a 6 meses.	Integer
Customer	ID do cliente	String
Age	Idade do cliente. Categorizada em: 0: <=18, 1: 19-25, 2: 26-35, 3: 36-45, 4: 46-55, 5: 56:65, 6: 65 U: Desconhecido	String

Gender	Gênero do Cliente. E: Pessoa Jurídica, F: Feminino, M: Masculino, U: Desconhecido	String
ZipcodeOri	Código postal de origem	String
Merchant	ID do comerciante	String
zipMerchant	Códido Postal comerciante	String
Category	Categoria da compra. (Será listada juntamente com a análise)	String
Amount	Valor da compra	Float
Fraud	Classe que mostra se a transação é Fraude (1) e Não Fraude (0)	Integer

4. Processamento/Tratamento de Dados

Os dados para o estudo após coletados foram processados para obtenção dos resultados utilizando a ferramenta Jupyter Notebook e a linguagem Python. Inicialmente, foi realizada uma verificação de informações como o tamanho do conjunto de dados e detalhes sobre as colunas, incluindo o tipo de dados e a presença de valores nulos, que, para este dataset não possuem. Além disso, informações descritivas e contagens foram obtidas para algumas variáveis, como idade, categoria e fraude, visando ao tratamento dos dados.

Em seguida, foram executadas as seguintes etapas de pré-processamento:

- Remoção de colunas com apenas um valor (ZipCodeOri e ZipCodeMerchant);
- Eliminação de aspas em valores categóricos;
- Substituição dos valores "U" na coluna idade;
- Conversão dos dados da coluna idade para o tipo Integer;
- Formatação da coluna de categoria.

```
In [7]: # Removendo colunas que possuem um unico valor
        df= df.drop(columns= ['zipcodeOri', 'zipMerchant'])

        # Removendo as aspas dos dados
        df['age'] = df['age'].str.replace("'", "").astype(str)
        df['gender'] = df['gender'].str.replace("'", "").astype(str)
        df['category'] = df['category'].str.replace("'", "").astype(str)
        df['customer'] = df['customer'].str.replace("'", "").astype(str)
        df['merchant'] = df['merchant'].str.replace("'", "").astype(str)

        # Substituindo "U" da coluna Idade
        df['age'] = df['age'].str.replace("U", "7").astype(int)

        # Removendo o 'es' da coluna Categoria
        df['category'] = df['category'].str.replace("es_", "").astype(str)

In [8]: df.head()

Out[8]:
```

	step	customer	age	gender	merchant	category	amount	fraud
0	0	C1093826151	4	M	M348934600	transportation	4.55	0
1	0	C352968107	2	M	M348934600	transportation	39.68	0
2	0	C2054744914	4	F	M1823072687	transportation	26.89	0
3	0	C1760612790	3	M	M348934600	transportation	17.25	0
4	0	C757503768	5	M	M348934600	transportation	35.72	0

Figura 1 - Fonte: Elaborado pela autora (2023).

5. Análise e Exploração dos Dados

Continuando com o conjunto de dados, um gráfico foi gerado para uma compreensão mais aprofundada da variável de classe. Tornou-se evidente que os dados apresentam um desbalanceamento, com as ocorrências de fraude representando apenas 1,21% do total. Esse desbalanceamento pode resultar na baixa performance do modelo. Portanto, esses dados desbalanceados serão abordados na fase de feature engineering mais adiante.

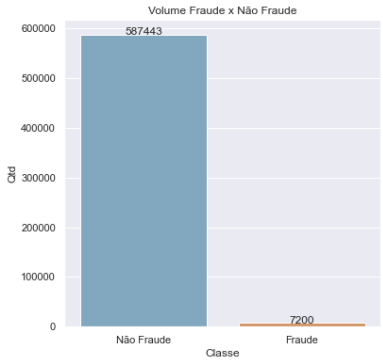

Figura 2 – Fonte: Elaborado pela autora (2023).

A partir desse ponto, o conjunto de dados foi dividido pela coluna de classe em dois dataframes, fraude e não fraude para facilitar a visualização e análise dos dados. Em seguida, foi a realizado a contagem da quantidade de transações para cada categoria, entre fraude e não fraude e combinadas em um outro dataframe. Foram adicionadas duas colunas: uma para o total de transações e outra para calcular o percentual de fraudes em relação ao total de transações por categoria. As observações indicam que, embora a classe de fraudes tenha um número menor de transações, em algumas categorias elas representam uma quantidade significativa. E olhando para quantidade, a categoria de Esportes e Brinquedos (“sportsandtoys”),

foi a categoria com maior transações fraudulentas, seguida da categoria de Sáude (“health”). Por outro lado, em três categorias – Transporte (“transportation”), Comida (“food”) e Conteúdo em geral (“contents”) - não há nenhuma transação de fraude.

```
In [16]: # Contagens de fraude e não fraude por categoria
         fraud_counts = df_fraud['category'].value_counts().reset_index()
         fraud_counts.columns = ['category', 'Fraude']
         not_fraud_counts = df_notfraud['category'].value_counts().reset_index()
         not_fraud_counts.columns = ['category', 'Não Fraude']

In [17]: # Combinando os dataframes
         group_df = pd.merge(fraud_counts, not_fraud_counts, on='category', how='outer')

         # Renomeando coluna
         group_df.rename(columns={'category': 'Categoria'}, inplace=True)

         # Preenchendo valores ausentes com 0
         group_df = group_df.fillna(0)

         # Adicionando coluna Total por Categoria/ Percentual de fraudes em relação ao total por categoria
         group_df['Total'] = group_df['Fraude'] + group_df['Não Fraude']
         group_df['% Fraude'] = round((group_df['Fraude'] / group_df['Total']) * 100, 2)

         # Ordenando o Dataframe pelo valor de Fraude em ordem decrescente
         group_df = group_df.sort_values(by='Fraude', ascending=False)

In [18]: group_df
```

	Categoria	Fraude	Não Fraude	Total	% Fraude
0	sportsandtoys	1982.0	2020	4002.0	49.53
1	health	1696.0	14437	16133.0	10.51
2	wellnessandbeauty	718.0	14368	15086.0	4.76
3	travel	578.0	150	728.0	79.40
4	hotelservices	548.0	1196	1744.0	31.42
5	leisure	474.0	25	499.0	94.99
6	home	302.0	1684	1986.0	15.21
7	hyper	280.0	5818	6098.0	4.59
8	otherservices	228.0	684	912.0	25.00
9	tech	158.0	2212	2370.0	6.67
10	barsandrestaurants	120.0	6253	6373.0	1.88
11	fashion	116.0	6338	6454.0	1.80
12	transportation	0.0	505119	505119.0	0.00
13	food	0.0	26254	26254.0	0.00
14	contents	0.0	885	885.0	0.00

Figura 3 - Fonte: Elaborado pela autora (2023).

O próximo passo envolveu a análise da dispersão de alguns dados categóricos, visando compreender a relação de algumas variáveis entre si, distinguidos por fraude e não fraude.

Primeiramente, foi avaliado o valor de compras por categoria, o que permitiu perceber que os maiores gastos ocorreram em compras fraudulentas, especialmente na categoria de Viagem (“travel”).

Após isso, foi analisado o valor de compras por gênero, evidenciando que as maiores compras fraudulentas foram realizadas pelo gênero feminino. Em seguida, foi verificado o valor de compras por faixa etária, destacando que a faixa de idade de 26 a 35 anos registrou os maiores gastos.

Por último, foi realizado uma visualização da dispersão utilizando o conjunto de dados criado somente com as fraudes. O gráfico apresenta o valor de compras

por faixa etária, distinguindo por gênero. É possível observar que a presença do gênero feminino é mais forte em três faixas etárias, entre 19-25, 26-35 e 56-65 anos.

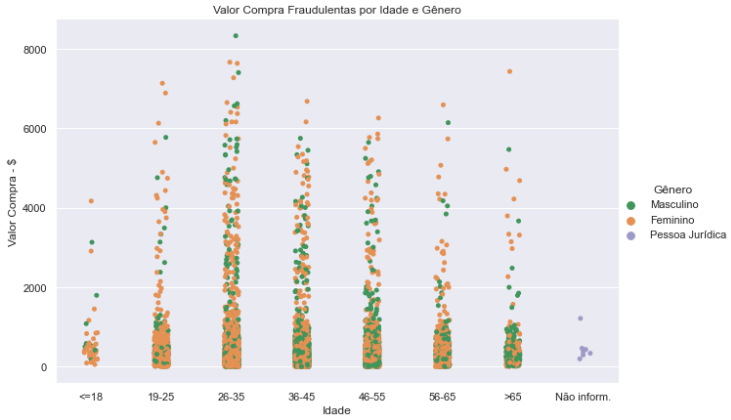

Figura 4 - Fonte: Elaborado pela autora (2023).

A fim de verificar a distribuição dos dados, foi criado um gráfico de barras apresentando a soma total de compras por categoria, separado entre fraude e não fraude. Diferentemente da quantidade de transações fraudulentas por categoria, visto acima na “Figura 3”, observa-se que em relação ao valor, a categoria Viagem (“travel”) se destaca, seguida por Saúde (“health”) e, em seguida, Esportes e Brinquedos (“sportsandtoys”).

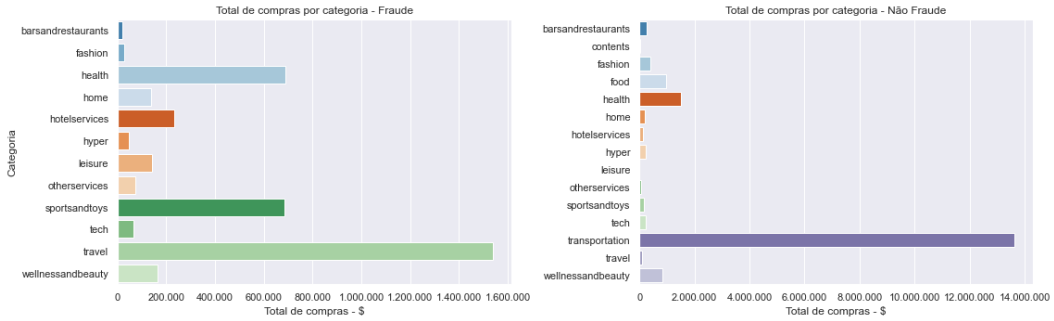

Figura 5 - Fonte: elaborado pela autora (2023).

Após essa etapa, foi criado um gráfico de boxplot dividido em quatro visualizações. As duas primeiras consideram transações de fraude e não fraude. Pode-se observar que a maioria dos valores por transação não fraudulenta não ultrapassou $500 dólares, exceto pela categoria Viagem (“travel”). Ao analisar as transações fraudulentas, a categoria Viagem também apresentou o maior gasto por transação, enquanto as outras categorias não ultrapassaram os $2.000 dólares. Nas outras duas visualizações, a categoria Viagem foi excluída para permitir uma melhor visualização das demais. Tanto para transações normais quanto para transações

fraudulentas, é evidente a presença de muitos valores extremos que não se alinham com a distribuição dos dados.

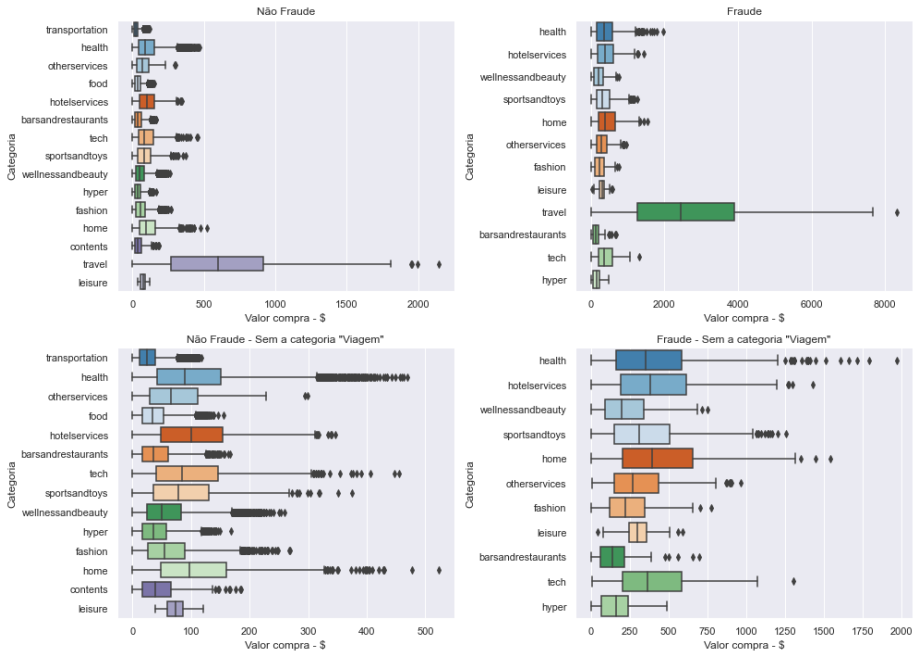

Figura 6 - Fonte: elaborado pelo autor (2023)

Os passos seguintes da análise focaram na exploração mais detalhada da distribuição das frequências de cada variável categórica, usando gráficos de barras separados entre transações fraudulentas e transações normais. É notável que aparentemente existe uma relação entre a quantidade de transações e o total de compras. Observando-que, por faixa etária, a maior quantidade de transações continuou na faixa de 26 a 35 anos. Em relação ao gênero, o feminino também se destacou como o mais frequente.

Por último, realizou-se uma análise da frequência de transações por comerciantes. A maior quantidade de transações de fraude por comerciante foi de 1.634 transações, o que representa aproximadamente 23% do total de transações fraudulentas, e quando comparado ao total de transações realizadas por este comerciante, equivalem a aproximadamente 47%.

Concluindo esta etapa de análise para identificar relações e padrões, foram obtidos os seguintes insights:

- Desbalanceamento do Dataset: O desequilíbrio nos dados será abordado na próxima etapa usando métodos de balanceamento afim de melhorar o desempenho do modelo.

- Variável de Categoria: As categorias de transações parecem estar correlacionadas com as fraudes. Algumas categorias podem ter maior propensão a fraudes do que outras.
- Padrões entre Transações: As compras fraudulentas na categoria Viagem (“travel”) apresentam uma incidência particularmente alta. O gênero feminino parece estar associado a compras fraudulentas de maior valor, e a faixa etária de 26 a 35 anos tem os maiores gastos. E exceto em categoria, esses padrões também se refletem quando visto pela quantidade de transações.
- Distribuição de Gastos: Muitas transações fraudulentas têm valores excepcionalmente elevados em comparação com transações normais. Isso é um indicativo de atividades suspeitas e/ou anomalias.

Essas descobertas serão abordadas nas próximas etapas de modelagem e análise preditiva, onde métodos serão utilizados para lidar com o desequilíbrio entre classes, além disso, explorar ainda mais os padrões identificados, utilizando técnicas de codificação de variáveis para converter dados catégoricos em formatos númericos, permitindo a exploração da correlação entre os dados e por fim, desenvolver um modelo preditivo eficaz para detecção de fraudes.

6. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Em elaboração.

7. Aplicação de Modelos de Aprendizado de Máquina

Em elaboração.

8. Avaliação dos Modelos de Aprendizado de Máquina e Discussão dos Resultados

Em elaboração.

9. Conclusão

Em elaboração.

10. Links

- Código fonte - https://github.com/jdomeneghini/ProjetoIntegrado/

11. Referências

- https://www.researchgate.net/publication/265736405_BankSim_A_Bank_Payment_Simulation_for_Fraud_Detection_Research
- https://www.insper.edu.br/noticias/crescimento-de-golpes-no-comercio-eletronico-impulsiona-mercado-de-prevencao/
- https://resources.sift.com/ebook/machine-learning-fraud-prevention-whats-next/