GLOBAL SOLUTION - CYBERSECURITY

Integrantes:
Henrico Nardelli Bela - RM: 95985
Sara Leal - RM: 96302
Emilly Santos - RM: 94437

Architecture review date	**7 de jun de 2022**
Project lead	@ Henrico Bela
On this page	

📋 Overview/Problema

#1 - Resolução de desafio do Kaggle

Vocês deverão resolver o seguinte desafio do Kaggle: https://www.kaggle.com/ealaxi/paysim1 . Trata-se de uma base de dados sintéticos que foi produzida, baseada num cenário real, para predizer uma fraude no setor financeiro.

A entrega será o notebook no Google Colab, devidamente comentado. IMPORTANTE: por ser um desafio do Kaggle, certamente vocês encontrarão resoluções. Faça bom uso, mas também faça as devidas referências e citações! A ideia é que vocês utilizem os códigos da minha e de outras matérias para resolver este exercício, mas fiquem à vontade para buscar outras referências.

🚩 Problemas de Arquitetura

Problema de Arquitetura	Impacto no Negocio	Prioridade	Notas
Identificar quais colunas e dados estavam recebendo as possíveis Fraudes, e onde elas se concentravam	Perda de dinheiro para a empresa.	HIGH	Conseguimos identificar quais colunas e dados estavam concentrados com Fraudes e aplicamos filtros nas colunas para fazer tal analise.

👥 Stakeholders

Nome	
@ Henrico Bela	Desenvolvedor
@EmillyGabrielly	Desenvolvedor
@SaraLeal	Desenvolvedor

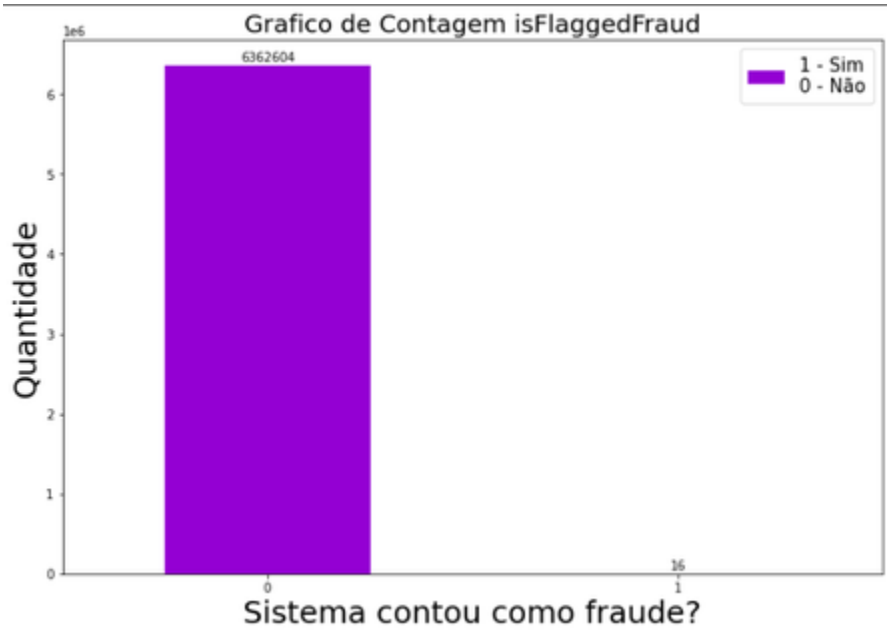
✏️ Atributos Importantes

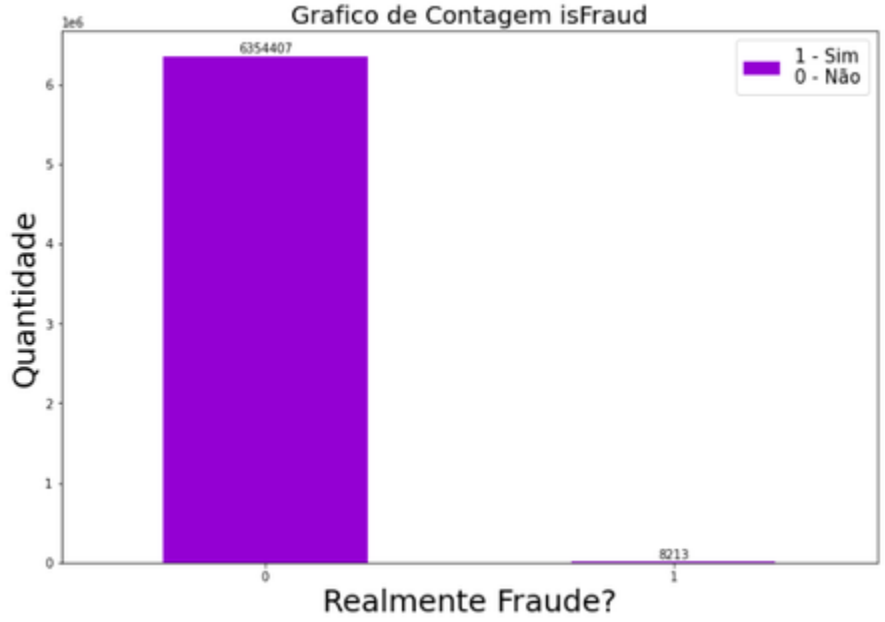
Atributos categóricos e numéricos

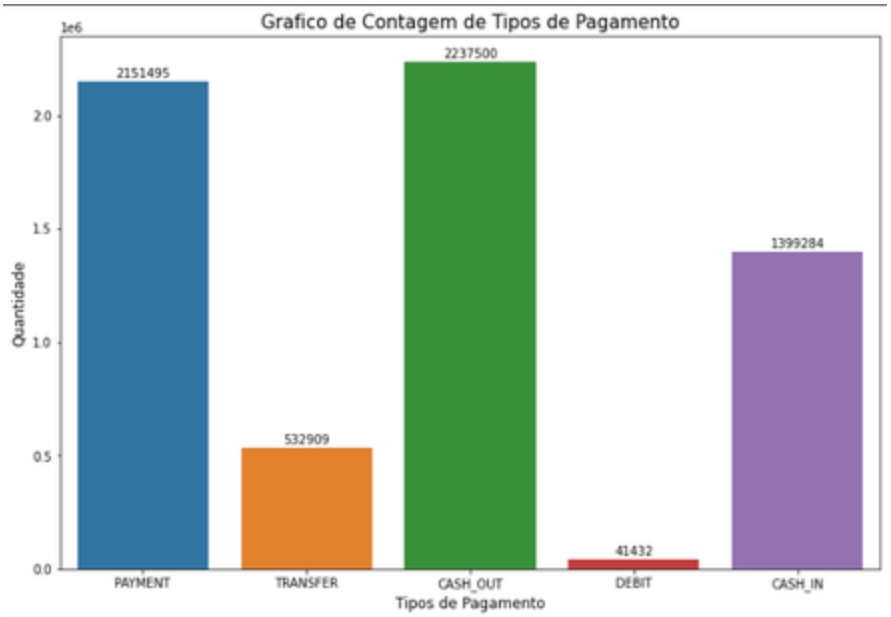
	Definição	Atributo é importante? Porque?	Notas

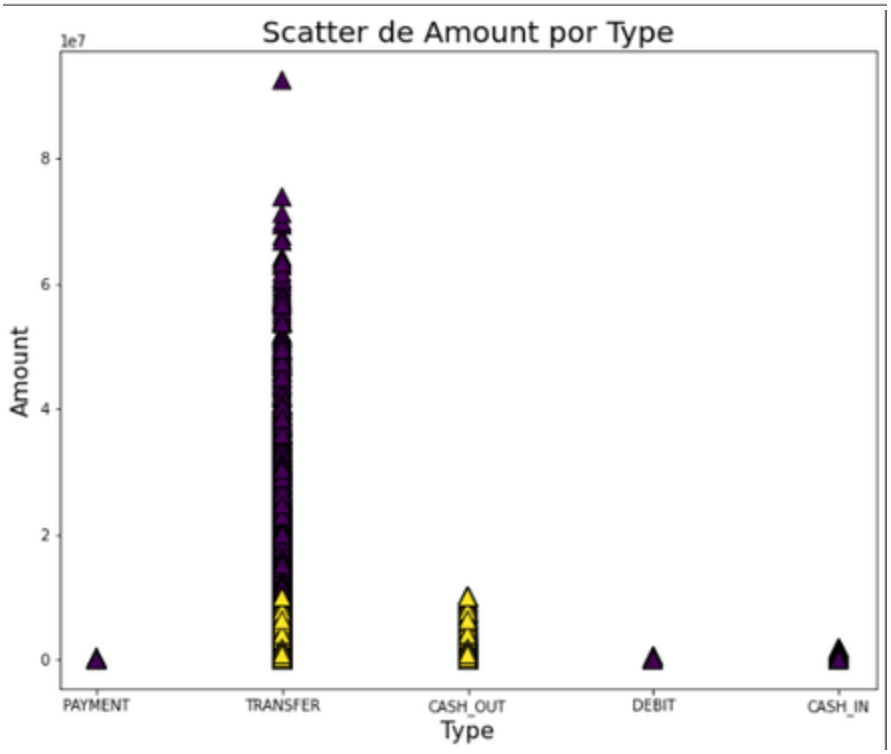
Atributos	type, amount, oldBalanceOrig, newBalanceOrig	Sim, os atributos type, amount, oldBalanceOrig e newBalanceOrig são as principais variáveis para se definir se uma transação foi Fraudulenta ou não.	Conseguimos identificar que essas variáveis são importantes de acordo com os gráficos produzidos e mostrados abaixo.

Gráficos









Objetivo

Resolver o desafio do Kaggle, utilizando inteligência artificial, para identificar transações fraudulentas.

Base de dados

https://www.kaggle.com/ealaxi/paysim1

A base de dados do Kaggle, foi criada para a realização de um desafio, que solicita uma analise de dados descritiva, e seguindo o ciclo de vida de machine learning.

Bibliotecas

1. Pandas

2. Seaborn
3. Matplotlib
4. Sklearn
    a. Preprocessing - Standard Scaler
    b. Model Selection - Train Test Split
    c. Linear Model - Logistic Regression
    d. Neighbors - K Neighborns Classifier
    e. Metrics - Confusion Matrix, Classification Report, Average Precision Score
5. Warnings

🤖 Modelagem

Para a modelagem, utilizamos o ciclo de vida de Machine Learning:

Começando com:

1. Business Understanding
2. Data Mining
3. Data Cleaning
4. Data Exploration
5. Feature Engineering
6. Preditive Modeling
7. Data Visualization

📝 Algoritmos utilizados

Segue o link para o google colab!

https://colab.research.google.com/drive/1wYpScvUYQPvAwapQ4lZNkPq9T4fmmY2S?usp=sharing

📈 Métricas

```
------------------------------
Metricas - Logistic Regression - Sem dados Padronizados
------------------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1270904
           1       0.36      0.41      0.38      1620

    accuracy                           1.00   1272524
   macro avg       0.68      0.71      0.69   1272524
weighted avg       1.00      1.00      1.00   1272524
```

```
------------------------------
Metricas - Logistic Regression - Com dados Padronizados
------------------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1270904
           1       0.90      0.42      0.57      1620

    accuracy                           1.00   1272524
   macro avg       0.95      0.71      0.79   1272524
weighted avg       1.00      1.00      1.00   1272524
```

```
----------------------------
Metricas - KNeighborsClassifier
----------------------------
Score: 0.9995
              precision    recall  f1-score   support

           0       1.00      1.00      1.00   1270904
           1       0.90      0.42      0.57      1620

    accuracy                           1.00   1272524
   macro avg       0.95      0.71      0.79   1272524
weighted avg       1.00      1.00      1.00   1272524
```

📋 Avaliação

Ao avaliar as métricas apresentadas acima, podemos identificar que a melhor métrica a ser utilizada é **Precisão**. Pois com a Precisão podemos avaliar exatamente quais as transações Fraudulentas, e quais não são.

Por exemplo, ao classificar uma ação como um bom investimento, é necessário que o modelo esteja correto, mesmo que acabe classificando bons investimentos como maus investimentos (situação de Falso Negativo) no processo. Ou seja, o modelo deve ser preciso em suas classificações, pois a partir do momento que consideramos um investimento bom quando na verdade ele não é, uma grande perda de dinheiro pode acontecer.

👣 Próximos passos

	Objetivo	Descrição	Estimativa	Documentação
1	Melhor entendimento das Features do Dataset	Mais estudos em cima de Feature Engineering	Colocaremos uma estimativa de 6 meses para melhor avaliação do modelo e aumentar a assertividade do modelo.	Comentar os progressos obtidos.

🤝 Referencias

- Sklearn documentation
- Pandas documentation
- Seaborn documentation
- Matplotlib documentation
- Para melhor entendimento do data set: https://www.kaggle.com/code/jmbebon/predicting-fraud-in-financial-payment-ser-c2f5e8