DESAFIO CLASSIFICAÇÃO

BOAS VINDAS

Olá! Que bom que você veio!

Este é o terceiro desafio assinado da trilha de desafios que você terá ao longo da jornada. Legal, né?

Em um mundo que muda cada vez mais rápido, é essencial encontrar uma **SOLUÇÃO QUE GARANTA RESKILLING & UPSKILLING CONSTANTES**.

O objetivo dos desafios é cultivar **interface responsiva** entre os avanços tecnológicos, transformações de mercado e mudanças comportamentais dos consumidores, preparando a empresa para emergência de uma nova cultura orientada por dados e preparada para lidar com o novo paradigma trazido pelas tecnologias e metodologias da **Ciência de Dados** e Inteligência Artificial.

ELEMENTOS DOS DESAFIOS

- Contexto de negócio.
- Problema a ser resolvido.
- Macro etapas de uma possível resolução.
- Toolbox (técnicas e algoritmos que podem ser aplicados para resolver o desafio).
- Base de dados.
- Extra (questão extra a ser debatida, por exemplo: estrutura da solução, ética da aplicação, privacidade dos usuários, etc).
- Aula síncrona de resolução e discussão.
- **As soluções elaboradas podem ser utilizadas para enriquecer seus portfólios.**

#FICAADICA

DESAFIO CLASSIFICAÇÃO

Por isso, incentivamos que sempre tente resolver o desafio. Teremos um momento para discussão com a expert sobre os pontos de dificuldades que surgirem, não se preocupe! O importante será participar deste momento!

Caso o desafio aborde um assunto que você não conhece ainda, não se preocupe! Ao longo dos dias você terá aulas com as experts, antes do dia da resolução, que acompanham a complexidade ferramental do case.

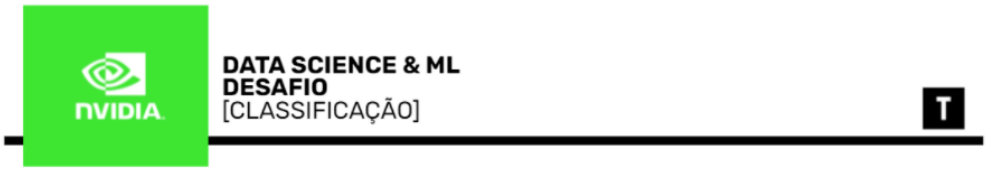
Você tem liberdade para fazer o desafio de forma individual, mas também poderá se juntar com o colega caso queira. Pode sair muita coisa legal daí! :)

Vamos começar?

AVANÇAR

DESAFIO CLASSIFICAÇÃO

POR PATRICIA PAMPANELLI



VÍDEO CONTEXTUAL: https://www.youtube.com/watch?v=I0oOOJZJihA

DESAFIO DE DETECÇÃO E REDUÇÃO DE FRAUDES NO SETOR FINANCEIRO

Embora muito se fale da capacidade de geração de receita das empresas, um dos principais entraves para o sucesso de um negócio está atrelado ao seu custo. A captação de clientes se tornou cada vez mais penosa, assim como a manutenção da carteira. Com isso, a gestão de custo se tornou um fator extremamente estratégico para garantir a margem e o crescimento das empresas.

No entanto, o advento no mundo digital para expandir o negócio ou garantir a sua sobrevivência enfrenta um grande desafio para redução do custo operacional: as fraudes, sobretudo as relacionadas às operações comerciais e financeiras, que impactam, na prática, todos os tipos de indústrias e setores da economia. Do governo, ao banco, do varejista ao microempreendedor, todos precisam conhecer e enfrentar este desafio.

O relatório da **ACFE** (https://www.acfe.com/rtm2019/index.html#Learn) (Association of Certified Fraud Examiners) corrobora com esta percepção ao apontar uma previsão de crescimento de 60% nos próximos dois anos nos investimentos em antifraude, uma clara amostra do desafio que as organizações estão enfrentando. No entanto, o mesmo relatório aponta que 58% das empresas declaram não ter níveis, recursos e profissionais suficientes para atuar em ações antifraude, conforme quadro abaixo:

DESAFIO CLASSIFICAÇÃO

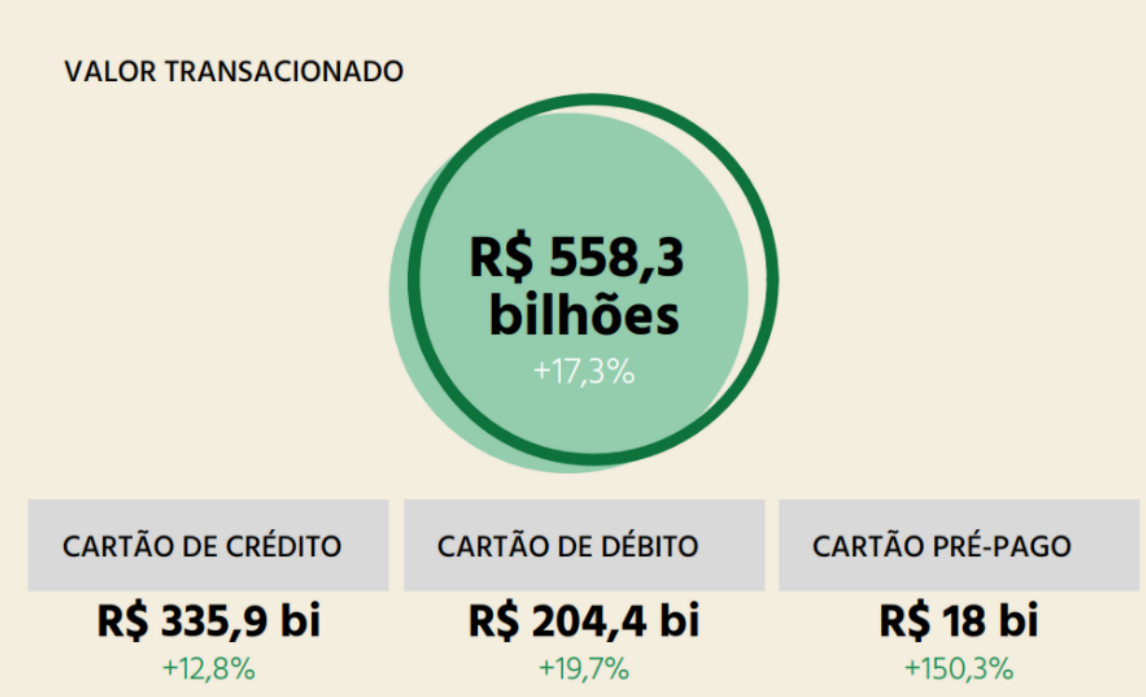
FONTE: ACFE (HTTPS://WWW.ACFE.COM/RTM2019/INDEX.HTML#LEARN)

Considerando que o objetivo dos fraudadores em geral é terem benefícios monetários, torna evidente que o setor financeiro é um dos seus principais alvos. Mesmo o crescente investimento em ações preventivas e de monitoramento não têm sido suficientes para barrar ou frear a escalada dos criminosos. De acordo com a **Psafe** (https://www.psafe.com/), de janeiro a agosto do ano passado foram 920 mil casos somente no Brasil e a **cada minuto, 3,6 fraudes acontecem no país**. Foram detectados, por exemplo, mais de **11 milhões** de tentativas de **fishing bancário**.

A indústria financeira possui grande representatividade no Brasil e no mundo. Para se ter uma ideia, os ativos dos bancos no Brasil somam R$ 7,4 trilhões, superando o próprio PIB do país (Infomoney) (https://www.infomoney.com.br/economia/ativos-de-bancos-somam-r-74-trilhoes-e-superam-pib-brasileiro/), 25-05-2020).

De acordo com a **ABECS**(https://www.abecs.org.br/) (Associação Brasileira das Empresas de Cartão de Crédito e Serviços), foram movimentados **R$ 558 bilhões** no primeiro trimestre de 2021 e os **cartões de crédito** representaram **R$ 335 bilhões desse total**. Foram 6,5 bilhões de transações, um crescimento de 11,8%, com destaque para cartão de débito que apresentou aumento de 163% (veja quadros abaixo). Apesar da opulência destes números, o impacto das fraudes é igualmente gritante. Os crimes cibernéticos, considerando apenas transações de cartão de crédito, já projetavam em 2018 um impacto de **6 trilhões de dólares** de perda de receita até 2021 ao redor do mundo.

DESAFIO CLASSIFICAÇÃO



FONTE: ABECS (HTTPS://API.ABECS.ORG.BR/WP-CONTENT/UPLOADS/2021/05/COLETIVA-ABECS-1T21-1.PDF)

Não à toa, no relatório de **novembro de 2020** (**https://api.abecs.org.br/wp-content/uploads/2020/11/Apresentacao-Balanco-3T20.pdf**) é listado entre as prioridades do setor financeiro a

DESAFIO CLASSIFICAÇÃO

Febraban(https://portal.febraban.org.br/) (Federação Brasileira dos Bancos) para mitigar este risco é previsto um gasto anual da ordem de R$ 2 bilhões de reais em TI por ano no Brasil. Porém, o risco é potencializado não somente pela sofisticação dos cyber criminosos, mas também pela insatisfação do consumidor que não só abandona a carteira de clientes, como também se utilizam do próprio código do consumidor, que garante indenização em dobro dos valores cobrados. Tudo isso, sem considerar as ações judiciais por danos morais e materiais que podem ser impetradas pelo consumidor. Segundo o Conselho Nacional de Justiça (CNJ), a projeção é que 28% das ações em trâmite tenham como réus instituições financeiras.

Neste cenário, a inteligência artificial surge como uma ferramenta que dá mais robustez, agilidade e flexibilidade para combater a fraude, trabalhando 24 horas por dia, 7 dias por semana, tornando-se um caminho mais do que possível, necessário para que as instituições financeiras possam combater de forma efetiva os fraudadores. Um exemplo disso é a **American Express**, com mais de 115 milhões de cartões de crédito ativos, que combate às fraudes utilizando inferência e Deep Learning, reduzindo gastos relacionados a fraudes em **U$ 1,2 trilhões**, em processos de detecção que acontecem em milisegundos. Fato que garante à empresa o menor índice de fraudes no mundo por 13 anos consecutivos.

Você como gestor, tem a missão neste desafio de destronar a American Express do posto de melhor instituição no combate à fraude. Para tanto, você precisa propor uma solução para detecção e análise de fraudes que possam reduzir os riscos da empresa e garantir margens saudáveis. Lembre-se, que o resultado do seu trabalho dará ainda mais visibilidade para você e sua área. Seja criterioso, utilize bons argumentos, fatos e justificativas para sua proposta, bem como, claro, faça uma boa execução do seu projeto.

A missão não é fácil, já que o grande volume de informações e legado tecnológico são a realidade da maioria das corporações do setor, mesmo entre empresas líderes. Além disso, iniciativas de inteligência artificial demandam grande capacidade computacional e profissionais que saibam atuar com estas ferramentas. Apesar de ser evidente a importância da inteligência artificial, competências analíticas e manifesto impacto nas receitas, 36% das empresas ainda não percebem com clareza os benefícios das iniciativas de análise de dados para as atividades antifraude e apenas 39% possuem algum programa específico para este fim (veja quadro abaixo). Enquanto muitos possam ver estes números como um risco, você tem a

DESAFIO CLASSIFICAÇÃO

FONTE: ACFE (HTTPS://WWW.ACFE.COM/RTM2019/INDEX.HTML#LEARN)

AVANÇAR

DESAFIO CLASSIFICAÇÃO

BASE DE DADOS

Segue link para o dataset que será utilizado neste desafio. Os dados estão armazenados no formato de csv:

[https://drive.google.com/file/d/1zjK8zQK5zvhR_r2chWI5dCjeOwASlPfb/view?usp=sharing]
(https://drive.google.com/file/d/1zjK8zQK5zvhR_r2chWI5dCjeOwASlPfb/view?usp=sharing)

A base de dados que será utilizada para o desenvolvimento deste desafio contém aproximadamente 6,3 milhões de transações. Os dados contidos nesta base de dados são simulados e compreendem 30 dias. Estas transações podem ser de diferentes tipo e foram executadas via mobile devices. As features contidas nesta base de dados são:

- step - Passo: representa o total de horas transcorrido desde o início da simulação. Esta feature vai variar entre 1 e 744 (30 dias);

- type - Tipo: tipo de transação (depósito, saque, débito, pagamento e transferência);

- amount - Quantia: total que foi transacionado;

- nameOrig - ClienteOrigem: cliente que iniciou a transação

- oldbalanceOrg - SaldoInicialOrigem: saldo da conta de origem antes da transação;

- newbalanceOrig - SaldoFinalOrigem: saldo da conta de origem após a transação;

- nameDest - ClienteDestino: cliente de destino da transação;

- oldbalanceDest - SaldoInicialDestino: saldo da conta de destino antes da transação;

DESAFIO CLASSIFICAÇÃO

- isFraud - ÉFraude: flag que define se a transação é fraudulenta ou não. Nesta simulação o objetivo da fraude é assumir a conta do usuário, esvaziá-la transferindo para outra conta e então sacando o dinheiro.

- isFlaggedFraud - SinalizadaComoFraude: automaticamente marcadas pelo banco como fraude por tentarem transferir mais de 200.000 em uma única transação.

AVANÇAR

DESAFIO CLASSIFICAÇÃO

ROTEIRO PARA DESAFIO

TRATAMENTO DE DADOS

- Faça o tratamento das features do dataset que são textuais (tipo de transação, origem, destino e categorização).

- Faça o tratamento das variáveis numéricas.

ANÁLISE EXPLORATÓRIA

- Calcule a correlação entre as variáveis que compõem o dataset. Existem variáveis que são altamente correlacionadas? Existe a necessidade de utilizar todas as variáveis que estão no dataset?

MODELAGEM

- É interessante criar novas features para a modelagem do problema (Ex: número de transações que foram feitas a partir de uma conta). Seja criativa(o)!

- Experimente modelos diferentes para a resolução de problemas de classificação (Ex: RandomForestClassifier, XGBoost)

ANÁLISE DOS RESULTADOS

- Faça o cálculo das classificações para avaliar os resultados que foram obtidos com os modelos avaliados. Qual modelo obteve o melhor resultado?

- Dada a matriz de confusão e o limiar que vocês escolheram, qual os possíveis impactos no negócio vocês identificam nesta solução? Quais observações vocês podem fazer com relação aos impactos dos falsos positivos e falsos negativos?

- Faça um comparativo entre o desempenho computacional do treinamento de cada modelo. O tempo de treinamento do modelo é fundamental para o desenvolvimento das soluções de Data Science, especialmente quando estamos trabalhando com datasets e problemas mais complexos. Qual o tempo de treinamento utilizando cada uma das abordagens?

DESAFIO CLASSIFICAÇÃO

que você criou tiveram relevância para o resultado final do modelo?

ANÁLISE DA SOLUÇÃO

- Qual o desempenho do treinamento utilizando uma infraestrutura com GPU? Qual o ganho percentual de desempenho com a utilização desta infraestrutura?

- Discuta os resultados obtidos com os resultados da matriz de confusão.

- Quais outras soluções vocês poderiam propor para este desafio?

- Quais variáveis que não estavam presentes no dataset que vocês entendem que seriam relevantes para a solução do problema?

- Quais variáveis que não estavam presentes no dataset que vocês entendem que seriam relevantes para a solução do problema?

AVANÇAR

DESAFIO CLASSIFICAÇÃO

TOOLBOX

GOOGLE COLAB

- Vídeo sobre Google Colab: **Jupyter on the web with Colab - YouTube**(https://www.youtube.com/watch?v=yElc9z-Ad3k)

- Como importar datasets armazenados no Google Drive: **How to Import and Export Datasets in Google Colab | by Mohammad Masum | Towards Data Science** (https://towardsdatascience.com/google-colab-import-and-export-datasets-eccf801e2971)

RAPIDS

Hoje em dia, a ciência de dados e o machine learning se tornaram o maior segmento de computação do mundo. As melhorias na precisão dos modelos analíticos se traduzem em bilhões no resultado final. Para construir os melhores modelos, os cientistas de dados precisam trabalhar duro para treinar, avaliar, iterar e retreinar para resultados e modelos de desempenho altamente precisos. Com o RAPIDS™, processos que levariam dias são feitos em minutos, facilitando e agilizando a construção e a implantação de modelos de geração de valor.

- Instalação do pacote do RAPIDS no Google Colab: [rapids-colab.ipynb - Colaboratory (google.com)] (https://colab.research.google.com/drive/1rY7Ln6rEE1p0IfSHCY0Vaqt80vD035J0#forceEdit=true&sandboxMode=true&scrollTo=m0jdXBRiDSzj)

- Exemplos de notebooks utilizando RAPIDS para classificação: [https://github.com/rapidsai/cuml/tree/8cccaf281d3a32115ee1e5fcea7fb8b7eabb4927/notebooks] (https://github.com/rapidsai/cuml/tree/8cccaf281d3a32115ee1e5fcea7fb8b7eabb4927/notebooks)

- Exemplos de notebooks utilizando o classificador XGBoost: [https://github.com/rapidsai-community/notebooks-contrib/blob/main/getting_started_materials/intro_tutorials_and_g

DESAFIO CLASSIFICAÇÃO

contrib/blob/main/getting_started_materials/intro_tutorials_and_guides/07_Introduction_to_XGBoost.ipynb)

- Medida de Tempo de Execução no Notebook:

Para avaliar o desempenho utilizando o pacote do RAPIDS é necessário realizar a medição das execuções das células do notebook. Um exercício interessante é comparar o desempenho dos pacotes que compõem o RAPIDS com o Pandas e o Sklearn. A medida pode ser feita incluindo o %%time no início da célula do notebook:

```
%%time

from time import sleep

for i in range(3):
    print(i, end=' ')
    sleep(0.1)
```

O output esperado é:

```
0 1 2
CPU times: user 5.69 ms, sys: 118 µs, total: 5.81 ms
Wall time: 304 ms
```

CPU times: de forma simplificada, é o tempo total de execução do código contido nesta célula do notebook.

Wall time: é o tempo transcorrido desde que o código foi submetido para execução pelo sistema até que o processo seja concluído.

DESAFIO CLASSIFICAÇÃO

DESAFIO CLASSIFICAÇÃO

A NVIDIA

Ao longo das duas últimas décadas a NVIDIA tem se destacado como um dos maiores provedores de solução de IA para diversos segmentos da indústria, mas também disponibilizado um ecossistema como [NGC] (**https://www.nvidia.com/pt-br/gpu-cloud/**) e treinamentos, como [DLI] (**https://www.nvidia.com/pt-br/training/**) (Deep Learning Institute), para formar e capacitar profissionais para atuarem com Inteligência Artificial.

A **NVIDIA** fica à sua disposição para que você possa tirar dúvidas técnicas ou saber mais informações sobre nossos produtos e soluções.

Patricia Pampanelli

https://www.linkedin.com/in/patricia-pampanelli/ (Deep Learning Solutions Architect) ppampanelli@nvidia.com

Marcel Saraiva

https://www.linkedin.com/in/saraivamarcel/ (Enterprise Sales Manager) msaraiva@nvidia.com

AULA AVALIADA COM SUCESSO!

Obrigada pelo seu feedback!

REFAZER

DESAFIO CLASSIFICAÇÃO