

REGULAMENTO DO DESAFIO – DOCK DATA HACKATHON – Nº 01/2022

Este documento foi elaborado para orientar os participantes com relação ao desafio de projeto que será realizado pelas equipes no dia 20/03/2022.

TEMA:

Fraud Prevention Challenges

Com a pandemia, a digitalização acelerou em todo o mundo e na mesma velocidade os riscos de fraude se multiplicaram para as instituições financeiras. Esse se tornou um dos principais focos das empresas do setor, o que a gente percebe quando olha para os R\$ 5,8 bilhões de perdas evitados pelas frentes de combate às ações fraudulentas em 2021 (Mapa da Fraude).

E é neste cenário super atual e desafiador que a Dock lança seu primeiro Hackathon juntamente com a Microsoft. Estamos colocando na mesa dois problemas reais, com dados reais para que cientistas de dados se aventurem na modelagem de algoritmos de prevenção a fraudes. Vocês poderão escolher para qual deles irão propor a sua solução de machine learning.

CHALLENGES:

A equipe deve escolher somente 01 dos desafios propostos a seguir:

Desafio 1 - Prevenindo a Fraude Ideológica:

O nosso onboarding deve permitir que a Dock e os nossos parceiros construam um processo de cadastramento de clientes seguro e fluido. Porém, sabemos que a todo momento fraudadores tentam se passar por outras pessoas para abrir contas, criar cartões de crédito, emitir boletos, entre outros crimes. Nesse sentido, ter um modelo de score cadastral é um grande aliado na hora de aprovar uma nova proposta. A Dock faz isso lançando mão de uma série de ações, como validação dos dados cadastrais, facematch, fraudbook, OCR etc.

O que esperamos de vocês?

1. Entendimento e tratamento adequado dos dados
2. Construir um modelo classificação de fraude / não fraude (binário)
3. Construir um score de fraude a partir do modelo
4. Apresentar as métricas de validação do modelo (a critério dos times participantes).
5. Insights sobre a exploração dos dados e dos resultados obtidos
6. Criar uma apresentação breve de 06 min

Dataset: Disponível na Conta de armazenamento 'storagehackdia2', container RAW em formato .parquet.

Desafio 2 - Combatendo a Fraude Transacional

O leque de transações bancárias que são processadas pela Dock para seus parceiros é bastante amplo. PIX, TED, Doc, Boletão, pagamentos com cartão etc. Da mesma forma, as medidas preventivas também são muitas. Nelas a Dock procura identificar estas tentativas e bloqueá-las antes mesmo de ocorrerem e, ao fazer isso por meio de Machine Learning, não só prevenimos a fraude, como evitamos a necessidade de um profissional da empresa avaliar cada caso suspeito manualmente, garantindo a escalabilidade e a segurança do negócio.

O que esperamos de vocês?

1. Entendimento e tratamento adequado dos dados
2. Construir um modelo classificação de fraude / não fraude (binário)
3. Apresentar o saving do modelo:
 $saving = VP - (FP + FN)$
VP (verdadeiro positivo) = valor das fraudes previstas que são de fato fraude (R\$)
FP (falso positivo) = valor das fraudes previstas que não são de fato fraude (R\$)
FN (falso negativo) = valor das fraudes não previstas que são de fato fraude (R\$)
4. Construir um score de fraude a partir do modelo
5. Apresentar as métricas de validação do modelo (a critério dos times participantes).
6. Insights sobre a exploração dos dados e dos resultados obtidos
7. Criar uma apresentação breve de 06 min

Dataset: Disponível na Conta de armazenamento 'storagehackdia2', container RAW em formato .parquet.

INFORMAÇÕES ADICIONAIS PARA AUXILIAR NA CONSTRUÇÃO DO PROJETO:

Entendendo o negócio da Dock

Somos a Dock. Uma das líderes em tecnologia para serviços financeiros na América Latina.

Com uma plataforma única e abrangente, oferecemos soluções de Emissão e Processamento de Cartões, Digital Banking, Adquirência e Risk & Compliance. Na prática, essas soluções tecnológicas permitem que as empresas ofereçam aos seus clientes uma série de serviços do universo financeiro, tendo por trás toda a infraestrutura tecnológica da Dock. Estas empresas serão identificadas nas nossas bases de dados dos desafios como Emissores ou Subemissores.

Dock. Tech your business free.

ETAPAS DE CONSTRUÇÃO

1. Escolha o desafio

Antes de começar, cada grupo deverá escolher qual o desafio seguir. O desafio 1, de fraude ideológica, ou o desafio 2, de fraude transacional. Lembrando que cada desafio possui suas próprias entregas.

2. Explore os dados

Faça uma boa análise exploratória (EDA). Lembre-se, você pode manipular os dados como achar necessário. O importante é que você tenha claro o que está fazendo e documente tudo direitinho. É importante justificar cada decisão, assim saberemos que você tomou os melhores caminhos.

3. Crie seu modelo

É hora de brilhar! Você é livre para propor a solução que achar melhor. Assim como na exploração, é importante que você justifique suas escolhas. Mostre porque seu modelo está redondinho, com as estatísticas e os resultados de cada etapa. Lembre-se, a implementação correta é pré-requisito, então ajude a banca a ver que você fez o dever de casa.

4. Valide seu modelo

Você é quem manda no processo, então cabe a você dividir a sua estratégia de validação. Qual a estratégia de cross validation, se dividirá a base em treino e teste, proporções, critérios etc, você decide, mas não se esqueça de deixar tudo documentado. Eu sei que é repetitivo, mas vamos lá: justificar suas escolhas pode garantir a aprovação do seu modelo. :-)

5. Faça a entrega

Caso utilize Jupyter notebooks, disponibilize no data lake da equipe com toda a modelagem documentada em extensão. pynb

Apresentação: sua equipe terá 05 min para apresentar o trabalho para uma banca julgadora, desde o desenvolvimento até os resultados. A apresentação deverá ser feita a partir do seu notebook, então cuide para que ele fique bem documentado e em um formato entendível. Como alternativa, utilize ferramentas de diagramas para demonstrar a arquitetura, PowerPoint para demonstrar o projeto, etc.

REGRAS GERAIS

Programação

20/03 (Horários -03 UTC) - Hackathon

- 08:45h - Recepção online e testes
- 09:00h - Boas-vindas aos Participantes do Hackathon
- 14:30h - Finalização do desafio
- 14:40h - Início Pitch das equipes
- 16h – Premiação e Encerramento

Equipes e auditoria:

- Os participantes deverão produzir os projetos nas salas do Teams de acordo com a sua equipe.
- A produção dos projetos nas salas do Teams será acompanhada por staff do evento, staff Microsoft e staff Dock a fim de monitorar a evolução e validação comportamental e técnica. Não será permitido a composição / evolução dos projetos em canais paralelos não oficiais, sob risco de desclassificação da equipe.
- O staff não poderá participar da resolução dos projetos, com participação limitada a acompanhamento e dicas pontuais que possam auxiliar na evolução. Somente os participantes da equipe podem operacionalizar as ferramentas de projetos e construção dos desafios.
- Os participantes podem solicitar apoio somente aos MENTORES do evento para entendimento de conceitualização das etapas propostas e tira dúvidas gerais sobre as regras do desafio.

Etapas de construção (obrigatórias):

1. Selecionar um dos challenges.
2. Produzir o desafio no Azure, Grupo de Recursos da sua equipe.
3. Importe os dados do Hub_Hackathon para seu grupo de recursos através de um serviço de Ingestão de dados.
4. Estruture containers de acordo com a camada de dados (raw, silver, gold).
5. Determine o melhor formato para manipulação de dados: parquet, delta, csv, orc, json, etc.

Participação, avaliação e direito a premiação:

Todos os participantes serão avaliados em tempo real pela comitativa de jurados presentes no evento verificando:

1. Colaboração, trabalho em equipe e cordialidade.
2. Foco e disposição.
3. Conclusão e participação ativa no Hackathon, dia 02.
4. Presença nos 02 dias do evento.

As equipes vencedoras do Hackathon terão seus nomes divulgados, bem como dos participantes que atenderem aos critérios de participação aqui mencionados.

A comitativa de jurados avaliadores e staff estarão avaliando em tempo real:

- Chat do evento para verificar interações, dúvidas, colaboração e participação.
- Pontualidade na participação do evento.
- Foco, espírito de equipe e engajamento para resolver os desafios do dia 02.
- Criatividade, qualidade, conclusão e boas práticas para solucionar os desafios.

Critérios de Pontuação:

A pontuação final do Hackathon será estruturada de acordo com os itens a seguir. A equipe vencedora será a que realizar mais pontos somados através da mesa avaliadora. Lembre-se de evidenciar todos os itens a seguir durante a apresentação final do projeto da sua equipe para que a mesa avaliadora classifique a pontuação dos itens.

05 pontos: Técnica:

Responder as perguntas propostas no challenge evidenciando a construção e etapas técnicas através de serviços no Azure.

05 pontos: Ferramentas / Serviços:

Evidenciar a utilização de pelo menos, 02 ferramentas a seguir para conquistar 05 pontos:

- Jupyter Notebooks - Python, SQL ou Scala.
- Databricks - Community / Azure.
- Data Lake no Azure.
- Azure Data Factory.
- Synapse Analytics.
- SQL do Azure.
- Machine Learning Studio.
- Power BI.

10 pontos: Organização e apresentação do projeto:

- Timing de produção e apresentação.
- Trabalho em equipe e colaboração.
- Criatividade e Governança.
- Engenharia de dados: arquitetura, performance, economia.
- Senso crítico e analítico dos resultados.
- A equipe deverá criar uma apresentação (PPT) para demonstrar o projeto à mesa avaliadora ou utilizar o Jupyter notebook do projeto para realizar a apresentação.

REGRAS DO PITCH

Horários e encerramento dos projetos

- Os projetos do hackathon deverão ser finalizados até 20/03, às 14:30h.
- A apresentação dos projetos será iniciada às 14:40h na sala principal do evento em ordem aleatória de equipes.
- Não será permitido a edição de projetos após às 14:30h, sob pena de desclassificação da equipe. Após às 14:30h, o staff técnico fará auditoria em tempo real na assinatura do Azure para verificar qualquer log de edição fora do horário permitido.

Apresentações

- As equipes deverão se organizar antecipadamente para eleger os profissionais que farão a apresentação do projeto.
- O tempo de apresentação de até 05 minutos.
- O formato da apresentação ficará a critério da equipe com possibilidade de demonstração técnica + slides para conceitualização da solução para melhor evidenciar as funcionalidades.
- Projetos não concluídos também poderão ser apresentados.

Mesa avaliadora

Os convidados da mesa avaliadora estarão presentes no evento durante as sessões de pitch (apresentação das equipes) para atribuir nota considerando:

- Pontualidade, organização e qualidade da apresentação final.
- Conclusão do projeto do Hackathon.
- Criatividade, senso de governança, performance e segurança do projeto.
- Menor esforço técnico para chegar aos resultados do desafio.

As notas das avaliações serão somadas, totalizando a nota final para ranking e definição da equipe vencedora.

Premiação

Os participantes da equipe vencedora ganharão um KIT Gamer por participante de acordo com a divulgação no site e regulamento. Os prêmios são oferecidos e entregues pela Dock. Possibilidade de convite para trabalhar na Dock.

Divulgação dos resultados

Após às 16h, durante o evento. Os resultados também serão divulgados no site e mídias sociais.

Possibilidade de convite para trabalhar na Dock.

Os participantes vencedores do Hackathon poderão ser convidados para trabalhar na Dock. Todos os demais participantes do evento que entregarem os laboratórios e demonstrarem participação ativa também poderão ser convidados para trabalharem na Dock.

ANEXO PARA CONSULTA

DICIONÁRIO DOS DATASETS

Desafio 1 - Fraude Ideológica

Variável	Descrição
issuerId	Identificador do Emissor da conta
bureau1	Status do módulo responsável pela validação de dados
bureau1_internalDetails	Informações adicionais para o bureau1
bureau1_reasonCode	Código da Razão da negativa do módulo 1
bureau2	Status do primeiro módulo de biometria facial
bureau2_reasonCode	Código da Razão da negativa do módulo 2
bureau3	Status do segundo módulo de biometria facial
bureau3_internalDetails	Informações adicionais para o bureau3
bureau3_reasonCode	Código da Razão da negativa do módulo 3

bureau4	Status do terceiro módulo de biometria facial
bureau4_reasonCode	Código da Razão da negativa do módulo 4
bureau5	Status do módulo de análise documental (OCR)
bureau5_reasonCode	Código da Razão da negativa do módulo 5
bureau5_a_analysis	Informações adicionais para o bureau5
bureau5_b_analysis	Informações adicionais para o bureau5
bureau6	Status do primeiro módulo de Score Cadastral
bureau6_analysis	Informações adicionais para o bureau6
bureau6_reasonCode	Código da Razão da negativa do módulo 6
bureau7	Status do módulo de análise documental (Documentoscopia)
bureau7_analysis	Informações adicionais para o bureau7
bureau7_reasonCode	Código da Razão da negativa do módulo 7
bureau8	Status do segundo módulo de Score Cadastral
bureau8_analysis	Informações adicionais para o bureau8
bureau8_reasonCode	Código da Razão da negativa do módulo 8
fraud_flag	Status de fraude ou não do dataset (0 = não fraude / 1 = fraude)
created	Data de criação do cadastro na Dock

Desafio 2 - Fraude Transacional

Variável	Descrição
issuer_id	Identificador do Emissor da conta
account_id	Identificador da conta
source_id	Identificador do portador da conta
segment	Indica se tem ou não um modelo de pontuação de risco ('Score' = Significa que há modelo 'modelSTUB' = Significa que não há modelo)
depositwithdrawalflag	Especifica se a transação é um depósito ou um saque relacionado a conta. Em caso de transações em contas digitais: D = Depósito, Q = Débito, em caso de transações associadas à cartões relacionados à conta digita: C = Crédito em cartão de crédito da conta, P = Crédito em cartão pré-pago da conta
debitacctcountry	O código do país no qual a conta sendo debitada é mantida

creditacctcountry	O código do país no qual a conta que está sendo creditada é mantida
transaction_value	O valor debitado (debitAmount) expresso em uma moeda constante
transaction_currencycode	O código da moeda para o campo transaction_value
transactiontype	O tipo da transação: A = Ordem de pagamento alterada, D = Pagamento único atrasado / com data antecipada, I = Pagamento único e imediato, L = Lista / pagamento em massa, M = Lista / pagamento em massa, M = Lista atrasada / com pagamento a prazo / S, S = Ordem permanente / pagamento único recorrente, T = Ordem permanente / lista recorrente / pagamento a granel, U = Transferência / pagamento imediato único e urgente
accesschannel	Este campo identifica o tipo de dispositivo através do qual a transação foi feita: A = ATM, B = Telefone (sistema de resposta interativa por voz), C = Computador / Internet, D = Caixa suspensa, F = Fax, G = Pagamento iniciado usando o canal de tokenização, M = Banco móvel pelo navegador da Web, N = Aplicativo móvel, O = Outro, P = Postal (Correio), R = Transferência automática, S = Terminal do ponto de venda (POS), T = Janela do caixa / caixa, U = Quiosque / dispositivo na filial (não ATM), V = Telefone (caixa remoto), Z = Link direto para o back office
processingchannel	Canal de processamento: A = ACH, B = Outro sistema de liquidação líquida diferida, C = Cheque, E = Transferência eletrônica de fundos (não especificado em outro lugar), F = FedWire, G = Outro sistema de liquidação bruta em tempo real, I = Transferência interna, O = Outros, P = Pagamentos mais rápidos, Q = BACS, R = CHAPS, S = SWIFT, W = Dinheiro
onusflag	Especifica se a transação está dentro ou fora de nós. Supõe-se que o customerAcctNumber esteja com a instituição financeira, portanto os seguintes valores se aplicam à outra conta: E = A transação é externa (fora de nós), O = A transação é dentro da marca (por nós), W = A transação é entre marcas (on-we)
reversalindicator	Indicador de reversão de transação: D = Reversão depreciativa, N = Reversão não depreciativa (ou status depreciativo desconhecido), Q = Não é uma reversão
score_adpt_falcon	Pontuação de risco
rule_name	Nome da regra responsável pela decisão
product	Tipo do produto ofertado (P2P / TED / BOLETO / PIX / RECARGA)
fraud_flag	Status de fraude ou não do dataset (0 = não fraude / 1 = fraude)
account_age	Tempo desde a criação da conta em dias
age	Idade
gender	Gênero
marital_status	Estado civil da pessoa
city	Cidade
state	Estado da União Federativa

addresses_types	Tipo do endereço (Residencial / Comercial / Correspondencia / UnitFour)
documents_types	Tipo de documento cadastrado (inscricao_municipal / cnpj / cpf / inscricao_estadual / rg)