
Questão 1

Durante a modelagem de um modelo LightGBM para prever churn de clientes, você identificou a presença de outliers em algumas variáveis numéricas. Considerando que o objetivo é minimizar o impacto desses outliers sem descartar informações relevantes ou distorcer a distribuição dos dados, qual das abordagens a seguir é mais apropriada?

- **(A)** Remover os outliers usando o método do intervalo interquartil (IQR), descartando valores abaixo do percentil 25 e acima do percentil 75.
- **(B)** Permitir que o algoritmo LightGBM lide com os outliers, aproveitando sua robustez na divisão de árvores para minimizar impactos negativos.
- **(C)** Aplicar transformação logarítmica em todas as variáveis para reduzir o impacto dos outliers.
- **(D)** Substituir os outliers por um valor percentil mais próximo (exemplo: percentil 5 ou 95) para evitar impacto na modelagem.

Questão 2

Um modelo de detecção de fraudes foi treinado em um conjunto de dados altamente desbalanceado, onde apenas 0,5% das transações são fraudulentas. Após o treinamento, o modelo apresenta uma acurácia de 99,5%, mas o recall para a classe de fraude é extremamente baixo. Mesmo após o ajuste de pesos das classes, o recall não melhora significativamente. Qual das abordagens abaixo é a mais apropriada para lidar com esse problema e aumentar o recall da classe minoritária?

- **(A)** Ajustar os hiperparâmetros do modelo para maximizar a área sob a curva ROC (AUC-ROC), garantindo uma melhor separação entre as classes.
- **(B)** Aplicar técnicas de clustering na classe minoritária para identificar subgrupos de fraudes e treinar um modelo específico para cada um.
- **(C)** Utilizar a métrica de Precision-Recall Curve (PR-AUC) para otimizar o modelo, focando em encontrar um limiar que priorize o recall para a classe fraudulenta.
- **(D)** Substituir o modelo atual por uma rede neural profunda com múltiplas camadas, pois modelos complexos lidam melhor com dados desbalanceados.

Questão 3

Ao treinar um modelo para prever a progressão de uma doença crônica, um cientista de dados observa que a AUC varia significativamente entre os diferentes folds de uma validação cruzada (Stratified K-Fold). O que o cientista deve investigar primeiro para entender essa inconsistência?

- **(A)** Verificar se os dados possuem dependências temporais, onde a ordem dos registros importa (ex: múltiplos exames de um mesmo paciente em tempos diferentes).
- **(B)** Analisar se o desequilíbrio entre as classes foi mantido em cada fold da validação cruzada.
- **(C)** Examinar o processo de normalização das variáveis para garantir que não houve vazamento de informações (data leakage) entre os folds.
- **(D)** Realizar um teste de concept drift para verificar se a variável alvo mudou sua definição estatística ao longo do tempo.

Questão 4

Um modelo de machine learning foi treinado com dados históricos de vendas de 2021 para prever o volume de pedidos em 2022. No entanto, o desempenho do modelo nos primeiros meses de 2022 foi muito inferior ao esperado, apesar de os dados de validação terem apresentado métricas excelentes. Qual é a razão mais provável para essa queda de desempenho?

- **(A)** O modelo não conseguiu aprender padrões de outliers que ocorreram apenas em 2021.
- **(B)** A normalização dos dados foi aplicada de forma diferente entre os conjuntos de treinamento e de teste.
- **(C)** O modelo sofreu de overfitting severo nos dados de 2021, impedindo a generalização para o ano seguinte.
- **(D)** A relação entre as variáveis preditoras e o volume de vendas mudou devido a fatores externos (Concept Drift).

Questão 5

Em um estudo estatístico sobre a taxa de inadimplência de um banco, um analista calcula um intervalo de confiança de 95% para a média da taxa de inadimplência, resultando em [5%, 7%]. A taxa média observada na amostra foi de 6%. Como esse intervalo deve ser corretamente interpretado?

- **(A)** Existe uma probabilidade de 95% de que a taxa de inadimplência média da população seja exatamente 6%.
- **(B)** Há uma chance de 95% de que a taxa de inadimplência de um novo cliente selecionado aleatoriamente esteja entre 5% e 7%.
- **(C)** Se o banco mantiver suas políticas de crédito, a taxa de inadimplência nos próximos anos estará entre 5% e 7% com 95% de confiança.
- **(D)** Se repetirmos o processo de amostragem e cálculo do intervalo muitas vezes, espera-se que 95% desses intervalos conterão a verdadeira taxa de inadimplência da população.

Questão 6

Você foi designado para monitorar um modelo de recomendação em produção. Após algumas semanas, você percebe que a distribuição de uma das principais variáveis de entrada mudou significativamente em relação ao conjunto de treinamento. Qual técnica é a mais indicada para detectar formalmente se essa mudança (Data Drift) é estatisticamente significativa?

- **(A)** Usar o teste de Kolmogorov-Smirnov (KS) para comparar as distribuições cumulativas da variável no treinamento e na produção.
- **(B)** Comparar a média e o desvio padrão da variável nos dois conjuntos e verificar se a diferença é superior a 10%.
- **(C)** Verificar se a acurácia do modelo caiu abaixo de um limite pré-estabelecido pela equipe de negócios.
- **(D)** Realizar um teste t de Student em todas as variáveis do modelo para identificar quais sofreram alterações.

Questão 7

Uma instituição financeira precisa implementar um modelo para decidir sobre a aprovação de empréstimos pessoais. Devido a regulamentações rigorosas, o banco precisa ser capaz de explicar exatamente por que cada decisão foi tomada (interpretabilidade total). Qual das abordagens abaixo é a mais adequada para atender a esse requisito regulatório?

- **(A)** Utilizar um modelo intrinsecamente interpretável, como uma Regressão Logística ou uma Árvore de Decisão simples.
- **(B)** Empregar um modelo de Boosting (como XGBoost) e utilizar valores SHAP para explicar as previsões individuais.
- **(C)** Treinar uma rede neural e utilizar um modelo de linguagem (LLM) para gerar explicações textuais sobre as decisões.
- **(D)** Remover todas as variáveis complexas e utilizar apenas as três variáveis que possuem a maior correlação linear com o alvo.

Questão 8

Um modelo de risco de crédito está em produção há 12 meses. O monitoramento indica que o desempenho do modelo (AUC) está caindo gradualmente, mas a distribuição das variáveis de entrada (features) permanece estável. Qual é a causa mais provável para esse declínio?

- **(A)** Overfitting ocorrido durante a fase de treinamento, que só se manifestou após um longo período em produção.
- **(B)** Concept Drift: a relação estatística entre as características dos clientes e a probabilidade de inadimplência mudou ao longo do tempo.
- **(C)** Erros sistemáticos na engenharia de features que passaram despercebidos durante os testes de integração.
- **(D)** Data Drift: a distribuição de variáveis como renda e idade dos novos clientes mudou significativamente nos últimos meses.

Questão 9

Ao monitorar um modelo de regressão para previsão de demanda, qual métrica de monitoramento seria a mais eficaz para identificar que o modelo está começando a enviesar suas previsões, permitindo uma investigação mais profunda sobre a causa?

- **(A)** A comparação entre o erro médio absoluto (MAE) da produção e o da validação.
- **(B)** A mudança na distribuição das variáveis preditoras (features) ao longo do tempo.
- **(C)** O monitoramento das variações na distribuição dos resíduos do modelo (diferença entre valor real e previsto).
- **(D)** O acompanhamento da mudança no volume total de previsões geradas pelo modelo diariamente.

Questão 10

Um sistema de detecção de fraudes em tempo real está apresentando alta latência nas respostas, o que prejudica a experiência do usuário. O gargalo foi identificado na etapa de geração de features, que exige cálculos complexos sobre o histórico transacional do cliente no momento da transação. Qual estratégia de engenharia de dados é a mais recomendada para resolver esse problema?

- **(A)** Utilizar uma Feature Store que suporte o cálculo dinâmico de features no momento da requisição usando processamento paralelo.
- **(B)** Implementar um sistema de cache em memória para armazenar os resultados das transações mais recentes de cada cliente.
- **(C)** Implementar uma Feature Store para armazenar e servir features pré-calculadas em batch, reduzindo o processamento em tempo real.
- **(D)** Migrar todo o pipeline de processamento para um modelo de processamento em lote (batch), eliminando a necessidade de respostas em tempo real.

Questão 11

Na construção de um modelo de churn para uma operadora de telecomunicações, um cientista de dados cria a variável: gasto_total_3_meses / gasto_total_12_meses. Qual é o principal insight ou benefício que essa técnica de engenharia de features busca capturar?

- **(A)** Identificar a tendência de comportamento do cliente, verificando se seus gastos estão aumentando ou diminuindo recentemente.
- **(B)** Garantir que a variável de gasto esteja normalizada entre 0 e 1, facilitando o treinamento de modelos lineares.
- **(C)** Eliminar a necessidade de usar redes neurais recorrentes (RNNs) para capturar padrões temporais nos dados.
- **(D)** Medir a sazonalidade dos gastos do cliente, comparando o trimestre atual com o comportamento anual médio.

Questão 12

Você está projetando um modelo de risco de crédito e deseja capturar mudanças no comportamento dos clientes ao longo do tempo. Qual das seguintes abordagens NÃO é recomendada?

- **(A)** Calcular médias e desvios padrão em janelas de tempo deslizantes para capturar flutuações nos gastos.
- **(B)** Criar features como "média_saldo_últimos_3_meses" / "média_saldo_últimos_12_meses" para medir variações recentes no saldo.
- **(C)** Calcular a categoria de transação mais frequente no último ano para identificar mudanças nos hábitos do cliente.
- **(D)** Incluir a data de cada transação como uma variável numérica no modelo.

Questão 13

Você criou 15.000 features para um modelo de risco de crédito, mas não consegue carregá-las todas de uma vez na memória devido a limitações computacionais. Qual seria a melhor abordagem para selecionar as variáveis mais importantes?

- **(A)** Treinar um modelo menor com subconjuntos aleatórios de features e calcular a importância média das variáveis.
- **(B)** Aplicar Análise de Componentes Principais (PCA) para reduzir a dimensionalidade sem perder informação.
- **(C)** Remover todas as variáveis que possuam mais de 10% de valores ausentes para reduzir o tamanho do conjunto de dados.
- **(D)** Usar eliminação recursiva de features (RFE) com todas as 15.000 variáveis, descartando uma por vez até encontrar as mais relevantes.

Questão 14

Uma instituição financeira processa bilhões de transações por mês, vindas de múltiplos sistemas e armazenadas em um data lake. O objetivo é calcular múltiplas métricas agregadas e identificar padrões de comportamento dos clientes, garantindo que os cálculos sejam rápidos, escaláveis e possam ser atualizados continuamente conforme novos dados chegam. Dado esse cenário, qual é a melhor abordagem?

- **(A)** Utilizar Spark para carregar todos os dados na memória antes de processá-los, garantindo que os cálculos sejam feitos de forma mais rápida.
- **(B)** Utilizar um banco de dados colunar, como MongoDB ou PostgreSQL, para processar consultas SQL otimizadas sobre os dados armazenados.
- **(C)** Usar Spark para processar os dados de forma distribuída e paralela, garantindo escalabilidade e capacidade de atualização contínua.
- **(D)** Aproveitar o desempenho de DuckDB para executar consultas diretamente sobre os arquivos do data lake, sem necessidade de mover ou transformar os dados.

Questão 15

Você está treinando um modelo de detecção de fraudes e precisa unir uma tabela de transações de clientes (300 milhões de registros) com outra contendo detalhes das contas bancárias (1 milhão de registros) no PySpark. A tabela de contas bancárias é relativamente pequena em comparação à de transações. Qual é a melhor abordagem para unir essas tabelas de forma eficiente?

- **(A)** Usar um broadcast join para replicar a tabela menor em todos os nós do cluster e evitar shuffle desnecessário.
- **(B)** Aumentar a memória do Spark e tentar novamente, garantindo que o processamento consiga lidar com o volume de dados.
- **(C)** Converter ambas as tabelas para Pandas e usar o método merge() para realizar a junção localmente.
- **(D)** Aplicar um cross join para garantir que todas as combinações entre transações e contas bancárias sejam testadas.

Questão 16

Uma instituição financeira utiliza um modelo de risco de crédito para aprovar ou recusar solicitações de empréstimo. No entanto, como o modelo foi treinado apenas com dados de clientes aprovados, ele pode não representar corretamente o risco dos clientes rejeitados, resultando em viés nas decisões de crédito. Qual das seguintes alternativas é a melhor abordagem para reduzir esse viés?

- **(A)** Usar inferência de rejeitados para gerar dados sintéticos e treinar o modelo com clientes simulados.
- **(B)** Aplicar técnicas de inferência de rejeitados para estimar o risco dos clientes que não tiveram o crédito aprovado.
- **(C)** Treinar um novo modelo apenas com clientes que não pagaram seus empréstimos, garantindo que ele aprenda a identificar inadimplentes.
- **(D)** Reequilibrar o conjunto de dados removendo parte dos clientes aprovados para tentar compensar a falta de dados de rejeitados.

Questão 17

Um banco deseja melhorar a previsão de risco de crédito incorporando o comportamento financeiro dos clientes. Como um modelo de behavioral scoring difere de um modelo de application scoring?

- **(A)** O behavioral scoring traça um perfil comportamental do cliente com base em características estáveis, ignorando dados transacionais, enquanto o application scoring avalia informações fornecidas no momento da solicitação do crédito.
- **(B)** O application scoring é atualizado mensalmente, enquanto o behavioral scoring é fixo após a aprovação do crédito.
- **(C)** O behavioral scoring usa histórico de transações para avaliar o risco continuamente.
- **(D)** O application scoring considera padrões de gastos ao longo do tempo para prever o risco do cliente.

Questão 18

Um time de risco percebeu que as previsões de inadimplência do seu modelo LightGBM parecem ser "superconfiantes", com muitas previsões acima de 90% ou abaixo de 10%. Eles perguntam se a saída do predict_proba() pode ser interpretada diretamente como a chance real de um cliente não pagar o empréstimo. O que melhor explica esse comportamento?

- **(A)** Sim, predict_proba() representa a probabilidade real de inadimplência, pois foi gerada diretamente pelo modelo.
- **(B)** Modelos de boosting como LightGBM produzem scores que não são calibrados como probabilidades reais, podendo ser ajustados posteriormente.
- **(C)** Esse comportamento ocorre apenas em casos de overfitting severo, onde o modelo distorce a escala das previsões.
- **(D)** Esse problema pode ser resolvido aplicando regressão isotônica diretamente ao modelo, garantindo que os scores sempre representem probabilidades reais.

Questão 19

Você está desenvolvendo um assistente financeiro baseado em LLMs. A equipe de negócios pergunta se o modelo deve ser ajustado via fine-tuning ou se engenharia de prompts é suficiente. O que você deve considerar primeiro?

- **(A)** Fine-tuning é sempre melhor, pois personaliza todas as camadas profundas do modelo, garantindo melhor performance.
- **(B)** Fine-tuning deve ser aplicado apenas se o modelo for pequeno e quando transfer learning não se aplica.
- **(C)** A engenharia de prompts é melhor para tarefas simples onde few-shot learning é suficiente, reduzindo a necessidade de ajustes complexos.
- **(D)** Engenharia de prompts só é útil para processamento de dados estruturados, não para geração de texto.

Questão 20

Um chatbot de consultoria jurídica utiliza um modelo baseado em GPT, mas às vezes gera informações legais incorretas que parecem convincentes. Qual é a melhor forma de mitigar esse problema?

- **(A)** Fazer fine-tuning do modelo com mais exemplos de respostas jurídicas corretas, garantindo maior aderência às diretrizes legais.
- **(B)** Aplicar Engenharia de Prompt para estruturar melhor as perguntas e garantir que o modelo consulte fontes externas confiáveis ao formular a resposta.
- **(C)** Usar Retrieval-Augmented Generation (RAG) para fornecer referências factuais e reduzir alucinações.
- **(D)** Reduzir a temperatura para 0, tornando as respostas mais controladas e eliminando a geração de informações incorretas.

Questão 21

Uma empresa de varejo utiliza Gradient Boosting para prever a demanda de produtos, mas o modelo está apresentando overfitting nos dados de treinamento. Qual hiperparâmetro deve ser ajustado primeiro?

- **(A)** Reduzir a taxa de aprendizado para evitar que o modelo se ajuste demais aos dados de treino.
- **(B)** Aumentar o número de iterações do boosting para permitir que o modelo aprenda melhor os padrões da demanda.
- **(C)** Aumentar a quantidade de árvores para reduzir a variância das previsões e evitar overfitting.
- **(D)** Reduzir a profundidade das árvores para limitar a complexidade do modelo e melhorar a generalização.

Questão 22

Um modelo de manutenção preditiva para equipamentos industriais foi treinado e testado normalmente, atingindo 99,9% de AUC no conjunto de teste. Em uma validação out-of-time, a métrica permaneceu alta. No entanto, ao ser implantado na produção, o modelo falha completamente e suas previsões se tornam inúteis. Qual das opções abaixo explica a causa mais provável desse problema?

- **(A)** O modelo apresenta overfitting severo, o que fez com que a AUC fosse superestimada nos testes.
- **(B)** Uma variável no treinamento contém informações que indiretamente antecipam a falha, mas essa informação não está disponível na produção (Data Leakage).
- **(C)** As distribuições das variáveis utilizadas pelo modelo mudaram entre o treinamento e a produção, impactando sua capacidade de generalização.
- **(D)** O modelo sofreu de data leakage, pois foi treinado com variáveis altamente correlacionadas, levando a previsões enviesadas.

Questão 23

Um banco que oferece cartões de crédito utiliza um modelo de machine learning para atribuir scores de crédito a novos clientes que não possuem histórico bancário na instituição. Dois clientes receberam um score de 900, e o banco agora precisa decidir o limite do cartão de crédito para cada um. Com base nessa informação, qual das conclusões a seguir é mais correta em relação ao limite de crédito que deve ser oferecido?

- **(A)** Como o modelo atribuiu o mesmo score para ambos, o banco deve oferecer o mesmo limite de crédito, garantindo um critério justo e padronizado.
- **(B)** Ambos apresentam um nível de risco semelhante, mas podem receber limites diferentes dependendo de fatores como renda e capacidade de pagamento.
- **(C)** Um score de 900 indica uma baixa probabilidade de inadimplência, portanto ambos devem receber um limite elevado para maximizar o lucro da instituição.
- **(D)** Clientes com score alto também possuem maior renda, então o modelo pode prever não apenas o risco, mas também a capacidade de pagamento, permitindo um limite adequado.

Questão 24

Você precisa criar um conjunto de features para um modelo de previsão de risco em transações financeiras. Seu conjunto de dados não é muito grande, mas contém milhões de registros armazenados em formato Parquet. Além disso, você precisa de uma solução que permita processar agregações complexas e cálculos de janelas temporais com alta velocidade. Qual é a melhor abordagem para esse cenário?

- **(A)** Utilizar Spark para processar os dados em paralelo e garantir escalabilidade ao carregar os arquivos Parquet em DataFrames distribuídos.
- **(B)** Empregar Featuretools para gerar automaticamente novas features e simplificar o pipeline de engenharia de features.
- **(C)** Escrever consultas SQL otimizadas para um banco de dados relacional tradicional, garantindo consistência nas features criadas.
- **(D)** Aproveitar DuckDB para processar agregações diretamente sobre arquivos Parquet, evitando a necessidade de carregar os dados na memória e maximizando a velocidade das consultas.
