**Quais passos você tomaria para resolver este problema? Descreva de forma completa e clara todos os passos que você considera essenciais para resolver o problema.**

1. **Entendimento do Problema e Requisitos**:
   * Compreender completamente os requisitos do cliente e os objetivos do projeto.
2. **Coleta e Entendimento dos Dados**:
   * Carregar e explorar os arquivos air\_system\_previous\_years.csv e air\_system\_present\_year.csv.
   * Verificar a qualidade dos dados, tratando valores ausentes (na), outliers e inconsistências.
3. **Preparação dos Dados**:
   * Codificar variáveis categóricas e normalizar variáveis numéricas.
   * Lidar com valores ausentes de maneira apropriada (imputação ou remoção).
4. **Divisão dos Dados**:
   * Dividir os dados em conjuntos de treinamento e teste, garantindo uma boa representação de ambas as classes (pos e neg).
5. **Treinamento de Modelos Preditivos**:
   * Selecionar e treinar vários modelos preditivos, como Regressão Logística, Random Forest e XGBoost.
6. **Avaliação de Modelos**:
   * Avaliar os modelos utilizando métricas apropriadas, como AUC-ROC, precisão, recall e F1-score.
7. **Otimização de Hiperparâmetros**:
   * Utilizar técnicas como Grid Search e Random Search para otimizar hiperparâmetros dos modelos.
8. **Explicabilidade e Interpretação do Modelo**:
   * Analisar a importância das variáveis nos modelos e utilizar técnicas para interpretar os resultados.
9. **Avaliação do Impacto Financeiro**:
   * Calcular o impacto financeiro das predições, considerando custos de manutenção preventiva e corretiva.
10. **Validação e Apresentação dos Resultados**:
    * Validar os resultados no conjunto de dados air\_system\_present\_year.csv e preparar um relatório detalhado para apresentação à diretoria executiva.
11. **Implementação em Produção**:
    * Desenvolver um plano de implementação para colocar o modelo em produção, incluindo monitoramento contínuo e manutenção.
12. **Monitoramento e Reentreinamento**:
    * Estabelecer um processo de monitoramento do desempenho do modelo em produção e definir critérios para reentreinamento.

**2. Qual métrica técnica de ciência de dados você utilizaria para resolver este desafio? Ex: erro absoluto, RMSE, etc.**

Para este desafio, as métricas técnicas mais apropriadas seriam:

* **AUC-ROC**: Para avaliar a capacidade do modelo de distinguir entre caminhões com e sem defeitos no sistema de ar.
* **Precisão e Recall**: Para balancear os falsos positivos e falsos negativos.
* **F1-Score**: Para uma avaliação equilibrada do desempenho do modelo, especialmente com classes desbalanceadas.

**3. Qual métrica de negócio você utilizaria para resolver o desafio?**

A métrica de negócio mais importante seria:

* **Custo Total de Manutenção**: Reduzir o custo total de manutenção considerando inspeções, manutenções preventivas e corretivas.

**4. Como as métricas técnicas se relacionam com as métricas de negócio?**

As métricas técnicas ajudam a avaliar a precisão do modelo preditivo, enquanto a métrica de negócio avalia o impacto financeiro das predições. Um modelo com alta AUC-ROC, precisão e recall indicaria um bom desempenho técnico, que se traduziria em decisões mais acertadas para reduzir custos de manutenção, atendendo assim à métrica de negócio.

**5. Que tipos de análises você gostaria de realizar no banco de dados do cliente?**

* **Análise de Correlação**: Entre variáveis para identificar fatores que influenciam falhas no sistema de ar.
* **Análise de Cluster**: Para segmentar caminhões com base em características de manutenção e uso.
* **Análise de Outliers**: Para identificar dados atípicos que possam influenciar os resultados.

**6. Que técnicas você usaria para reduzir a dimensionalidade do problema?**

* **PCA (Análise de Componentes Principais)**: Para reduzir a dimensionalidade mantendo a variabilidade dos dados.
* **Seleção de Características (Feature Selection)**: Utilizando técnicas como LASSO, Random Forest e Gradient Boosting para selecionar as variáveis mais relevantes.

**7. Que técnicas você usaria para selecionar variáveis para o seu modelo preditivo?**

* **Importância das Características (Feature Importance)**: Usando modelos como Random Forest e XGBoost.
* **Métodos de Seleção Automática**: Como Recursive Feature Elimination (RFE).
* **Testes Estatísticos**: Para avaliar a significância de cada variável e métodos como stepwise.

**8. Quais modelos preditivos você usaria ou testaria para este problema? Indique pelo menos 3.**

* **Random Forest**: Devido à sua capacidade de lidar com grandes conjuntos de dados e variáveis irrelevantes.
* **XGBoost**: Pelas suas excelentes capacidades de predição e eficiência.
* **Regressão Logística**: Por sua simplicidade e interpretabilidade.

**9. Como você avaliaria qual dos modelos treinados é o melhor?**

* **AUC-ROC**: Para avaliar a capacidade discriminativa do modelo.
* **Precisão e Recall**: Para balancear os falsos positivos e falsos negativos.
* **F1-Score**: Para uma métrica equilibrada entre precisão e recall.
* **Custo Total de Manutenção**: Simulando o impacto financeiro das predições.

**10. Como você explicaria o resultado do seu modelo? É possível saber quais variáveis são mais importantes?**

Utilizaria técnicas como:

* **Importância das Características**: Fornecida por modelos como Random Forest e XGBoost.
* **SHAP (Shapley Additive Explanations)**: Para explicar a contribuição de cada variável nas predições.
* **LIME (Local Interpretable Model-agnostic Explanations)**: Para explicações locais das predições do modelo.

**11. Como você avaliaria o impacto financeiro do modelo proposto?**

* **Simulação de Cenários**: Calculando os custos de manutenção com e sem o uso do modelo preditivo.
* **Análise de Custo-Benefício**: Comparando os custos de inspeções e manutenções preventivas versus corretivas.
* **ROI (Retorno sobre Investimento)**: Calculando o retorno financeiro da implementação do modelo.

**12. Que técnicas você usaria para realizar a otimização de hiperparâmetros do modelo escolhido?**

* **Grid Search**: Para uma busca exaustiva dos melhores hiperparâmetros.
* **Random Search**: Para uma busca mais eficiente em espaços de hiperparâmetros grandes.
* **Bayesian Optimization**: Para uma abordagem mais inteligente e eficiente na busca de hiperparâmetros ótimos.

**13. Que riscos ou precauções você apresentaria ao cliente antes de colocar este modelo em produção?**

* **Dados Ausentes**: Impacto de dados ausentes na precisão do modelo.
* **Overfitting**: Risco de o modelo se ajustar demais aos dados de treinamento.
* **Mudanças nos Dados**: Necessidade de monitorar mudanças nos padrões de dados ao longo do tempo.
* **Validação Contínua**: Importância de validar continuamente o desempenho do modelo em produção.

**14. Se o seu modelo preditivo for aprovado, como você o colocaria em produção?**

* **Desenvolvimento de API**: Para integrar o modelo com os sistemas existentes da empresa.
* **Pipeline Automatizado**: Para pré-processamento de dados, inferência do modelo e atualização contínua.
* **Monitoramento**: Implementar métricas e alertas para monitorar o desempenho do modelo.

**15. Se o modelo estiver em produção, como você o monitoraria?**

* **Métricas de Desempenho**: Monitorar métricas como precisão, recall e F1-score.
* **Monitoramento de Desvios**: Identificar mudanças nos dados de entrada e no desempenho do modelo.
* **Feedback do Usuário**: Coletar feedback dos usuários para identificar possíveis problemas.

**16. Se o modelo estiver em produção, como você saberia quando reentreiná-lo?**

* **Monitoramento de Desempenho**: Reentreinar o modelo se houver uma degradação significativa no desempenho.
* **Mudanças nos Dados**: Reentreinar o modelo se houver mudanças significativas nos padrões de dados.
* **Períodos Regulares**: Estabelecer um cronograma regular para reentreinamento com base na sazonalidade e na quantidade de novos dados coletados.