Atividades do Desafio:

1. Para solucionar o problema vamos seguir as seguintes etapas:
    - Conhecimento dos Dados: essa etapa consiste em conhecer como os dados estão distribuídos, inclusive dentro de cada classe, os tipos de dados, valores faltantes e correlações.
    - Pré Processamento: após conhecer os dados essa etapa será responsável por tratar os dados de forma que fique consistente, tratando valores faltantes, eliminando colunas com alta correlação e normalizando os dados, por exemplo.
    - Classificação: nessa etapa serão treinados alguns modelos para abordar o problema, talvez 2 ou 3 modelos com abordagens diferentes para enriquecer a comparação. As principais métricas observadas serão acurácia, precisão e sensibilidade. A matriz de confusão irá permitir calcular qual modelo realmente será mais econômico, visto que temos os custos das classificações e assim podemos ver o modelo que irá obter o melhor desempenho.
    - Otimização: depois dos primeiros testes com os modelos escolhidos é uma boa prática aplicar técnicas de otimização para enriquecer a avaliação dos modelos. Aplicando novas técnicas de tratamento dos dados como também de ajustes dos hiperparâmetros dos modelos.
2. As principais métricas observadas serão matriz de confusão, acurácia, precisão e sensibilidade.
3. A matriz de confusão irá permitir calcular qual modelo realmente será mais econômico usando os valores do problema, substituindo os valores vamos criar um cenário hipotético e assim identificar o modelo com melhor resultado econômico para a empresa.
4. Na matriz de confusão teremos as seguintes possibilidades;
    - Verdadeiro Positivo: quando o veículo está com problema no sistema de ar e o modelo classifica corretamente (US$ 25,00)

- Verdadeiro Negativo: quando o veículo não está com problema no sistema de ar e o modelo classifica corretamente (US$ 0,00)
- Falso Positivo: quando o veículo está com problema no sistema de ar e o modelo classifica incorretamente (US$ 500,00)
- Falso Negativo: quando o veículo não está com problema no sistema de ar e o modelo classifica incorretamente (US$ 10,00)

Assim é possível multiplicar os custos acarretados em cada cenário e analisar qual modelo é, de fato, mais viável para a empresa.

5. Seria interessante entender os principais motivos de perdas para estimar melhor os valores ausentes. A forma como os dados foram obtidos pode sugerir alguma causa de perda mais clara.
6. Uma técnica interessante para reduzir a dimensionalidade desse caso é Análise de Componentes Principais (PCA) que permite identificar os principais componentes dos dados além de poder reduzir a dimensionalidade eliminando os componentes “mais fracos”
7. O PCA citado na resposta anterior permite identificar os principais componentes dos dados
8. Acho que seria válido testar modelos que implementam diferentes abordagens como:
    - Vizinhos mais próximos (KNN)
    - Regressão Logística
    - Classificador Estatístico (Naive Bayes)
    - Árvore de Decisão
9. A matriz de confusão irá permitir calcular qual modelo realmente será mais econômico usando os valores do problema, substituindo os valores vamos criar um cenário hipotético e assim identificar o modelo com melhor resultado econômico para a empresa.
10. Alguns modelos permitem uma visualização melhor das variáveis mais importantes como a Árvore de Decisão, que permite visualizar gerada. Mas também é possível criar visualizações das variáveis nos outros modelos.

11. A matriz de confusão irá permitir calcular qual modelo realmente será mais econômico usando os valores do problema, substituindo os valores vamos criar um cenário hipotético e assim identificar o modelo com melhor resultado econômico para a empresa.
12. Para otimizar os hiperparâmetros seria interessante aplicar a Busca em Grade e a Busca Aleatória
13. É importante deixar claro o comportamento do modelo, principalmente, nos casos que ele classifica como falso positivo, onde a empresa possivelmente perde mais dinheiro.
14. Caso aprovado, o modelo pode ser disponibilizado através de uma API. Isso permitirá mais versatilidade de implementação, como usar uma tarefa que diariamente ou semanalmente, dependendo da necessidade e recursos disponíveis, verifica os veículos que podem estar com problema no sistema de ar.
15. Seria interessante usar alguma ferramenta de monitoramento como ML Flow para registar tanto o histórico de experimentos e métricas daquele modelo
16. Acredito que é importante analisar o comportamento dos dados que são usados pelo modelo, a medida que eles evoluem é interessante estabelecer uma periodicidade de treinamento do modelo.