## **Teste Data Scientist Jr – Bix Tecnologia – 13/07/24 – 20:23**

**Candidato – Bruno Martins Bartolomeu**

## **Challenge Activities**

To solve this problem we want you to answer the following questions:

1. What steps would you take to solve this problem? Please describe as completely and clearly as possible all the steps that you see as essential for solving the problem.
2. Which **technical** data science metric would you use to solve this challenge? Ex: absolute error, rmse, etc.
3. Which business metric  ***would*** you use to solve the challenge?
4. How do technical metrics relate to the business metrics?
5. What types of analyzes would you like to perform on the customer database?
6. What techniques would you use to reduce the dimensionality of the problem?
7. What techniques would you use to select variables for your predictive model?
8. What predictive models would you use or test for this problem? Please indicate at least 3.
9. How would you rate which of the trained models is the best?
10. How would you explain the result of your model? Is it possible to know which variables are most important?
11. How would you assess the financial impact of the proposed model?
12. What techniques would you use to perform the hyperparameter optimization of the chosen model?
13. What risks or precautions would you present to the customer before putting this model into production?
14. If your predictive model is approved, how would you put it into production?
15. If the model is in production, how would you monitor it?
16. If the model is in production, how would you know when to retrain it?

Segue abaixo respostas:

**1. Medidas para Resolver o Problema**

1. **Exploração e Limpeza dos Dados**:
   * Carregar os arquivos CSV e examinar a estrutura e o conteúdo.
   * Tratar valores ausentes e possíveis outliers.
   * Decodificar as colunas codificadas, se possível, ou trabalhar com a equipe técnica para entender o significado de cada coluna.
2. **Análise Exploratória dos Dados (EDA)**:
   * Realizar análises estatísticas e visualizações para entender a distribuição dos dados.
   * Identificar correlações entre as variáveis e a classe alvo (“pos” e “neg”).
   * Examinar a frequência de falhas no sistema de ar ao longo do tempo e em diferentes condições.
3. **Pré-processamento dos Dados**:
   * Normalizar ou padronizar as variáveis conforme necessário.
   * Tratar colunas categóricas através de técnicas como One-Hot Encoding.
   * Dividir o dataset em conjuntos de treinamento e teste.
4. **Redução de Dimensionalidade**:
   * Aplicar técnicas como PCA (Análise de Componentes Principais) ou LDA (Análise Discriminante Linear) para reduzir o número de variáveis, mantendo a variabilidade dos dados.
5. **Seleção de Variáveis**:
   * Utilizar métodos como Seleção por Filtro (testes estatísticos), Wrapper (RFE - Recursive Feature Elimination) e Embedded (Árvores de Decisão, Random Forest) para identificar as variáveis mais relevantes.
6. **Treinamento e Avaliação de Modelos Preditivos**:
   * Treinar múltiplos modelos (ex.: Regressão Logística, Random Forest, Gradient Boosting, Redes Neurais).
   * Avaliar os modelos utilizando métricas apropriadas e escolher o melhor com base na performance.
7. **Interpretação e Explicabilidade do Modelo**:
   * Utilizar técnicas como SHAP (SHapley Additive exPlanations) ou LIME (Local Interpretable Model-agnostic Explanations) para entender a importância das variáveis.
8. **Avaliação Financeira**:
   * Calcular a economia potencial com a redução de manutenções corretivas.
   * Comparar custos e benefícios de diferentes estratégias de manutenção.

**2. Métrica Técnica de Ciência de Dados**

* **Acurácia, Precisão, Recall, F1-Score**: Para avaliar a performance do modelo de classificação.
* **ROC AUC**: Para entender a capacidade do modelo em distinguir entre as classes.

**3. Métrica de Negócios**

* **Custo Total de Manutenção**: A principal métrica de negócios será a redução dos custos totais de manutenção, considerando manutenções preventivas e corretivas.

**4. Relacionamento entre Métricas Técnicas e de Negócios**

* As métricas técnicas de desempenho do modelo (como precisão e recall) diretamente impactam a métrica de negócios, pois um modelo mais preciso e sensível pode identificar caminhões que realmente precisam de manutenção preventiva, reduzindo custos com manutenções corretivas.

**5. Tipos de Análises no Banco de Dados**

* **Análise Temporal**: Para verificar tendências ao longo do tempo.
* **Análise de Correlação**: Para identificar variáveis que têm forte correlação com falhas no sistema de ar.
* **Análise de Distribuição**: Para entender a distribuição das variáveis e identificar possíveis padrões.

**6. Técnicas para Reduzir a Dimensionalidade**

* **PCA (Principal Component Analysis)**: Para transformar as variáveis em componentes principais que explicam a maior parte da variância.
* **LDA (Linear Discriminant Analysis)**: Para maximizar a separação entre as classes.

**7. Técnicas de Seleção de Variáveis**

* **Feature Importance (Random Forest)**: Para identificar as variáveis mais importantes.
* **RFE (Recursive Feature Elimination)**: Para selecionar iterativamente variáveis que melhoram a performance do modelo.
* **Testes Estatísticos (Chi-square, ANOVA)**: Para avaliar a relevância das variáveis.

**8. Modelos Preditivos para Testar**

* **Regressão Logística**
* **Random Forest**
* **Gradient Boosting Machines (GBM)**
* **Redes Neurais**

**9. Critério para Classificar o Melhor Modelo**

* **Acurácia e F1-Score**: Balanceia precisão e recall.
* **ROC AUC**: Avalia a capacidade do modelo de separar as classes.
* **Curva de Custo-Benefício**: Avalia o impacto financeiro direto das predições.

**10. Explicação do Resultado do Modelo**

* **SHAP e LIME**: Para identificar quais variáveis têm mais impacto nas predições e como influenciam as decisões do modelo.

**11. Avaliação do Impacto Financeiro**

* **Análise de Cenários**: Comparar custos de diferentes estratégias de manutenção.
* **ROI (Return on Investment)**: Calcular o retorno financeiro da implementação do modelo.

**12. Otimização de Hiper parâmetros**

* **Grid Search e Random Search**: Para testar diferentes combinações de hiperparâmetros.
* **Bayesian Optimization**: Para uma busca mais eficiente dos hiperparâmetros.

**13. Riscos e Precauções**

* **Qualidade dos Dados**: Garantir que os dados estejam completos e corretos.
* **Interpretação dos Resultados**: Certificar-se de que as predições são compreendidas corretamente.
* **Monitoramento Contínuo**: Estabelecer processos para monitorar a performance do modelo ao longo do tempo.

**14. Colocando o Modelo em Produção**

* **Pipeline de Produção**: Automatizar o processo de coleta de dados, predição e monitoramento.
* **Integração com Sistemas Existentes**: Garantir que o modelo se integre bem aos sistemas de manutenção atuais.

**15. Monitoramento do Modelo em Produção**

* **Dashboards e Alertas**: Para monitorar a performance do modelo em tempo real.
* **Métricas de Performance Contínuas**: Acompanhar a acurácia e outros indicadores-chave do modelo.

**16. Training do Modelo**

* **Monitoramento de Drift**: Detectar mudanças na distribuição dos dados que possam afetar a performance do modelo.
* **Atualização Regular**: Treinar o modelo periodicamente com dados novos para manter a precisão.