Case para Cientista de Dados Sênior - Logcomex

Dados os seguintes arquivos em anexo:

- historico_202001_202405.csv
- base_teste_202405.csv
- base_validacao_mes.csv

- O arquivo histórico contém a quantidade de embarque vindos da China, com total de valor e peso das mercadorias de um filtro de 10 diferentes hscodes para os estados de SP, RJ, MG e RS, no período entre 01/2020 até 05/2024.

Dicionário de Dados:

1. anomes: mês de referência do embarque no formato AAAAMM
2. cod_ncm: código da mercadoria com 8 dígitos
3. hscode: código da mercadoria com 4 dígitos
4. cod_pais_origem: país de origem da carga (160 – China)
5. urf_cod: código do porto, aeroporto ou unidade alfandegária onde é registrada a entrada da carga.
6. via_transp_cod: tipo de transporte de entrada da carga: 01 – Marítimo / 04 – Aéreo / 07 – Rodoviário, etc
7. sgl_uf_import: UF do importador da carga
8. city_cod: Código da Cidade
9. cidade_import: Cidade do importador da carga
10. qtd_emb: Quantidade de embarques na combinação: anomes / cod_ncm / hscode / cod_pais_origem / urf_cod / via_transp_cod / sgl_uf_import / cidade_import
11. qtd_imp: Quantidade de importadores no mês
12. qtd_exp: Quantidade de exportadores do país origem no mês
13. tot_valor: Valor total da mercadoria em dólares na combinação
14. tot_peso: Peso total da mercadoria em kilos na combinação

Com base no arquivo de histórico fornecido, efetuar:

1. Uma análise exploratória da base, avaliando sazonalidades, outliers, etc.
2. Criar um modelo de ML que possa prever se uma dada combinação cod_ncm / hscode / cod_pais_origem / urf_cod / via_transp_cod / sgl_uf_import / cidade_import pode ocorrer em um dado mês ou não.
3. Utilizar até o mês 04/2024 como **treinamento** e o mês 05/2024 como **teste**.
4. Criar quaisquer variáveis acessórias que possa achar necessário para ajudar no modelo.
5. Levar em conta a possibilidade de ocorre sazonalidade das combinações ao longo do tempo.

6. Utilizar o arquivo base_teste_202405.csv para prever o campo combinação_202405, onde 0 indica que a combinação não ocorreu em 05/2024 e 1 indica que a combinação ocorreu em 05/2024.
7. O objetivo do modelo será maximizar os TP (Verdadeiros Positivos) e minimizar os FP (Falsos Positivos), isto é, se a combinação ocorre em 05/2024 (1), e o modelo prever que ocorre (1), será considerado um Verdadeiro Positivo, porém se a combinação não ocorre no mês (0), e o modelo prever que ocorre (1), será considerado um Falso Positivo.
8. Utilizar o arquivo base_validacao.csv para prever se as combinações ocorrem no mês 06/2024 e no mês 07/2024. Aqui será avaliado posteriormente se o modelo criado não apresenta overfitting.

Como resultado, espera-se os códigos em python (scripts ou jupyter notebook), as análises efetuadas, e resultados obtidos: acurácia e precisão na parte da base utilizada para teste, tempo de execução, matriz de confusão, etc. Será considerado na avaliação a documentação e os pareceres realizados.