TESTE 1 – RESPOSTAS

1) Exercício 1

Suponha que você possui uma base de dados rotulada com 10 classes não balanceadas, essa base é formada por 40 features de metadados e mais 3 de dados textuais abertos.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

a) *Descreva como faria a modelagem dessas classes.*

Diante da situação problema apresentada, a modelagem iniciaria com a análise exploratória dos dados, em que os dados seriam carregados para entender sua estrutura, analisar a distribuição, o desbalanceamento e explorar as features de metadados e os dados textuais. Para isso, seria utilizado as bibliotecas pandas, matplotlib, seaborn. Em seguida, seria iniciado o processo de pré-processamento tanto dos dados de metadados como dos dados textuais, incluindo a formatação dos dados. Em relação aos dados textuais, seria utilizados as bibliotecas NLTK/spaCy para o processamento da linguagem natural, TF-IDF/Word2Vec para tokenização e vetorização dos dados textuais e para os metadados, utilizaria a biblioteca scikit-learn para normalizar os dados. Feito assim, iniciaria a etapa de balanceamento dos dados com o pacote imbalanced-learn. Na etapa de modelagem ou de definição do modelo, gosto de utilizar mais de um modelo, isso inclui modelos lineares e modelos baseados em árvore, como Logistic Regression, SVC, Random Forest, Gradient Boosting. Após implementados, as métricas de todos são avaliadas como: Accuracy, Precision, Recall, F1-Score e AUC-ROC, matriz de confusão, validação cruzada. Isso permite, encontrar o modelo que melhor se adequa ao problema. Caso nenhum, apresente boas métricas ou não atenda ao modelo, modelos de redes neurais ou combinação de modelos serão avaliados e implementados e terão suas métricas avaliadas. Definido o melhor modelo, ele será implementado para treinamento e fazer previsões. Também seria criados testes automatizados para validação do modelo.

b) *Ao finalizar essa modelagem, como iria apresentar essa modelagem para a área contratante?*

Seria feita uma apresentação contendo a descrição do problema, a contextualização e os objetivos, uma breve descrição sobre os dados utilizados. Descrição das etapas de análise exploratória e pré-processamento, falaria sobre os modelos implementados e suas validações, apresentando uma tabela as métricas encontradas, os gráficos das matrizes de confusão e curva AUC-ROC e mais gráficos relevantes a situação de problema, de forma a justificar o modelo escolhido. E por fim, apresentaria o desempenho do modelo final, explicaria a implicação das métricas no contexto do negócio, gráficos da matriz de confusão, curva AUC-ROC e outros relevantes e outras informações relevantes observadas durante o treinamento do modelo.

c) *Como faria a validação desse modelo?*

Para validação do modelo como se trata de uma classe desbalanceada e dados mistos, o conjunto de dados seria dividido em dados de treino, dados de validação e dados de teste. Como parte do processo, seria analisado a validação cruzada, as métricas

de avaliação: accuracy, precision, recall, F1-Score, curva AUC-ROC, matriz de confusão, também seria feito ajustes de parâmetros para validar o desempenho do modelo. E por fim, validar o modelo com os dados de teste. Durante a etapa de validação seria criado um documento com o registro dos parâmetros utilizados. Aqui, as bibliotecas a serem utilizadas são as mesmas que constam no item 'a'.

d) *Supondo que esses dados são recebidos diariamente, como iria trabalhar com esse desafio?*

Se os dados são recebidos diariamente, seria necessário automatizar o processo de coleta de dados e validação, podendo utilizar a ferramenta Apache Airflow. O modelo deveria ser treinado de forma contínua, uma vez, que novos modelos estão sempre chegando. Utilizar banco de dados eficientes e robustos para armazenamento dos dados, definir um pipeline de desenvolvimento e de processamento de dados. Monitoramento contínuo do modelo e dos pipelines definidos, análise sobre a escalabilidade da infraestrutura, devido ao volume de dados e a carga de processamento e atualização contínua. E definição do uso de práticas e ferramentas de CI/CD.

e) *Como levaria esse projeto para um ambiente produtivo?*

Para levar o modelo para produção seria desenvolvido APIs para integração com outras aplicações, caso seja necessário fazer integração. Conteinerização do modelo de forma a facilitar a implantação e escalabilidade, poderia utilizar o Docker para esse papel.

EXTRA - Existe mais algo que gostaria de relatar sobre esse caso?

As técnicas para balanceamento das classes, tem papel crucial na qualidade do modelo. Caso, não seja feita corretamente irá afetar o desempenho do modelo, podendo trazer resultados errôneos.

2) Exercício 2:

Suponha que você tenha uma base de dados de vendas de uma loja de varejo que inclui informações sobre produtos, clientes, datas de compra e valores das vendas. A base de dados possui, em média, 10.000 registros diários.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

a) Como você iria explorar os dados para obter insights sobre o desempenho das vendas.

Iniciaria com análise exploratória dos dados para entender como os dados estão dispostos, utilizaria a biblioteca pandas para carregamento dos dados, verificaria se cada coluna está com o tipo adequado dos seus dados, se é numérico, categórico, data. Faria a limpeza e preparação dos dados: avaliaria se há valores faltantes, faria tratamento dos outliers. Em seguida, faria uma análise sobre a tendências das vendas em

relação ao tempo, como, por exemplo, vendas por dia/semana/mês/período; analisaria quais produtos/categorias são mais vendidos ou menos vendidos, buscaria por padrões em relação ao tipo de cliente e avaliaria a relação entre as variáveis. Para isso, utilizaria a biblioteca numpy e as bibliotecas matplotlib/seaborn para gerar visualização gráfica, como gráfico de linha, histogramas, boxplots, entre outros. Também utilizaria bibliotecas estatísticas para análise do problema, como scipy, statsmodels, fazer cálculo da média, mediana, análise de correlação entre as variáveis, testes de hipóteses nas vendas entre grupos significativos. Se necessário, fazer uso de machine learning para fazer previsões ou encontrar padrões. Com essas informações, poderia ter insights sobre o período que um x produto/categoria é mais vendido, o que não tem saída, qual é o tipo de cliente alvo, de forma a melhorar o desempenho das vendas.

b) Como você responderia as seguintes questões:

i. Qual é o desempenho de vendas ao longo do tempo?

Considerando que a etapa de pré-processamento/limpeza dos dados já foi realizada, aplicaria a técnica de séries temporais para fazer agrupamento por períodos (diário/semanal/mensal/bimestral etc.) e analisar as tendências. Calcularia as vendas totais para cada período, compararia com os períodos anteriores e tentaria identificar a sazonalidade (por exemplos, feriados com datas comemorativas). Geraria gráficos de linha, barras e o tipo heatmap (permite visualizar a intensidade das vendas para dados diários e semanais) e com essas informações, encontrar padrões e retornar o desempenho das vendas ao longo do tempo.

ii. Quais são os produtos mais vendidos?

Seguindo a análise do item anterior, faria uma análise agregada para agrupar as vendas por produto e somaria as quantidades para identificar os produtos mais vendidos e menos vendidos.

iii. Como as vendas variam por categoria de produtos?

Seguindo a análise do item 'i', agruparia os dados por categoria de produto e faria comparações em relação o total de vendas ou receita gerada.

iv. Qual é a distribuição dos valores de venda?

Seguindo a análise do item 'i', a partir dos gráficos gerados como histograma, boxplots, média, desvio padrão poderia identificar a distribuição dos valores de venda.

v. Como os preços dos produtos afetam as vendas?

Aqui seria aplicado a análise de correlação entre o preço do produto e quantidade vendida.

vi. Qual é o perfil dos principais clientes em termos de compras?

Realizar segmentação do cliente com base na análise de perfil, utilizando dados estatísticos como frequência, gastos etc.

c) Como você faria para identificar grupos de clientes nessa base de dados?

Considerando que as etapas de carregamento dos dados/pré-processamento/ limpeza dos dados foi realizada, poderíamos adotar um algoritmo de clusterização como o K-means ou DBSCAN, o número de clusters é um parâmetro que deve ser ajustado e depois das definições realizadas, seria avaliado as características de cada clusters para entender os segmentos de clientes e dessa forma, obter insights para agrupamento dos clientes.

d) Qual teste estatístico você usaria para provar uma hipótese referente aos segmentos de clientes? e como iria aplicá-lo?

Poderia utilizar o Teste Estatístico Qui-Quadrado de Independência, pois ele é usado para determinar se há um relacionamento entre duas variáveis categóricas, como o segmento e categoria dos produtos ou frequência de clientes recorrentes e segmento.

Extra - Pensando nos dados acima, seria possível fazer mais algum tipo de análise?

Sim, poderia explorar a análise de séries temporais usando modelos como ARIMA e SARIMA, que poderia auxiliar na previsão de estoque, planejamento de vendas e marketing.

3) Exercício 3

Suponha que você tenha uma base de dados contendo textos jurídicos, como decisões judiciais, petições e documentos legais. A base de dados inclui informações sobre o conteúdo

do texto, data, jurisdição e outras informações relevantes. Seu objetivo é criar um sistema de recomendação que sugira textos jurídicos semelhantes a um texto de referência.

Para todos os itens: Informe as bibliotecas usadas, se necessário, o motivo de cada decisão, explore as possibilidades.

a) *Descreva como você desenvolveria o sistema de recomendação que recebe um texto de referência e sugere os textos mais semelhantes a ele na base de dados.*

Começaria com a coleta da base de dados de textos jurídicos, carregamento dos dados com a biblioteca pandas, faria uma análise exploratória dos dados para entender como os mesmos estão dispostos, depois faria a limpeza dos dados para remover caracteres especiais, padronizaria tudo em letras minúsculas. Em seguida começaria a etapa de pré-processamento em que seria a tokenização, normalização (stemming/lemmatization) e a remoção de stop words, utilizando as bibliotecas NLTK/spaCy, faria a vetorização dos textos com TF-IDF/Word2Vec, aqui abordaria também o cálculo de similaridade de cosseno, importante para analisar a similaridade entre dois documentos, que seria uma abordagem de filtragem baseada em conteúdo. Esse treinamento permitiria escolher os textos com maior similaridade para recomendação. O conjunto de dados seria dividido em dados de treino, dados de validação e dados de teste, após o treinamento do modelo, o sistema seria avaliado sobre o conjunto de dados de testes e faria refinamentos nos parâmetros do modelo a fim de melhorar o desempenho.

b) *Como você avaliaria esse sistema de recomendação?*

A avaliação seria feita com base na divisão do conjunto de dados em dados de treino, dados de validação e dados de teste. Os dados de treino são usados para ajustar o modelo, os dados de validação são usados para ajustar os hiperparâmetros e os dados de teste para avaliação final do modelo. Avaliaria as métricas como: precision – que permite dizer que os documentos recomendados são semelhantes aos documentos de referência; recall - se o sistema é capaz de recomendar a maioria dos documentos relevantes; F1-Score – útil quando se faz necessário um equilíbrio entre precision e recall. Aqui seria importante contar com a avaliação humana (especialistas jurídicos) sobre a relevância dos documentos recomendados.

c) *Suponha que novos textos jurídicos sejam adicionados diariamente. Como você manteria o sistema de recomendação atualizado e garantiria que ele continue a fornece recomendações relevantes?*

Sistema automatizado para coleta de dados e validação com atualização periódica dos novos dados, abordagem de machine learning para suportar treinamento incremental, para evitar que o modelo toda vez tenha que ser re-treinado do zero, uma vez que exige escalabilidade de infraestrutura e alto poder de processamento. Acompanhamento das métricas de avaliação e dos feedbacks dos usuários, atualização de palavras-chave/conceitos de acordo com os dados de entrada. Criação de testes automatizados como testes unitários, testes de regressão, testes de performance para garantir que o sistema continua funcionando mesmo com o aumento dos dados. Adoção de pipeline de desenvolvimento juntamente com as práticas de CI/CD.