TESTE 1 -

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.*

A modelagem pode varia de acordo com detalhes específicos do contexto, no entanto, seguem algumas diretrizes para modelar as classes:

- Balancear as classes, podendo:

- Utilizar apenas amostras balanceadas, desde que cada amostra seja estatisticamente representativa.

- Utilizar menos dados da(s) classe(s) maior(es). *Downsampling.*

- Sintetizar dados para as classes menores. *Upsampling.*

Para os dados textuais é importante realizar um pré-processamento específico, como lematização, stemização, remover stop words, etc.

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

Iria explicar o objetivo do projeto. Dizer que na nossa base de dados há classes com muito mais exemplos que outras, então explicar os desafios que isso impõe e assegurar que temos as ferramentas para superá-lo. Por fim é importante informar os riscos e incertezas associadas ao modelo e apresentar estratégias para minimizá-los.

c) *Como faria a validação desse modelo?*

A depender do contexto, algumas estratégias seriam:

- Dividir a base de dados em um grupo de treino e um de validação, geralmente com proporções, respectivamente, de 70% e 30%.

- Dividir a base em algumas partições e treinar várias vezes, a cada vez, utilizando uma dessas partições para validação. Essa técnica é conhecida como k-fold-validation.

Observações:

Se os dados forem adquiridos em tempos diferentes, tomar cuidado para não validar em um dado cujo tempo é anterior ao dado de treinamento.

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

É necessário construir uma estratégia de manutenção do modelo. Pode-se utilizar uma automação que retreina o modelo e compara a melhora, atualizando caso positivo. Adicionalmente pode ser necessário um analisar se o comportamento dos dados não se modificou significativamente, fazendo com que seja preciso atualizar a arquitetura do modelo.

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

Se disponíveis , utilizaria ferramentas de *Cloud* para deploy, podendo ser AWS, Azure ou Google Cloud Platform. Senão seria necessário desenvolver ferramentas equivalentes.

Quais ferramentas utilizar dependeria das restrições de performance do modelo. Por exemplo, para executar em tempo real, pode-se utilizar ferramentas de streaming de dados como o *Amazon Kinesis*, aliado a bancos de dados de baixa latência como Amazon *Dynamo Database.* Para ter escalabilidade utilizaria o modelo implementado em um cluster Kubernetes em Cloud, o qual poderia ser expandido de acordo com a demanda.

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

É importante estudar os modelos do estado da arte, porque a área de I.A. está avançando muito rápido. Atualmente os modelos baseados em florestas demontram os melhores desempenhos para classificação de dados tabulares. Para Visão computacional e NLP há melhor performance em modelos de deep learning.
Também é importante salientar que na prática, a limpeza de dados, preprocessamento e engenharia de features costuma ser o diferencial para a entrega de qualidade, consistindo também como a maior parte do trabalho.

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.

Utilizar agregações pode períodos aliados a estatística descritiva. Por exemplo: média de produtos diários, quantis de produtos mais vendidos, quatis de clientes com mais compras. Fazer isso para quantidade, valor, lucro, etc.

b) Como você responderia as seguintes questões:

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

Calcular, média, média móvel e tendência de crescimento de quantidade/ valor de vendas ao longo do tempo.

ii. Quais são os produtos mais vendidos?

Realizar contagem de vendas agrupando por produto e período e comparar. Utilizando pandas. Polars ou similar. E numpy.

iii. Como as vendas variam por categoria de produtos?

1. **Realizar a contagem de vendas agrupando por categoria e período e comparar. Utilizando pandas. Polars ou similar. E numpy.**

iv. Qual é a distribuição dos valores de venda?

Calcular a densidade de vendas por período, produto, categoria e cliente. Plotando a distribuição de valores.

v. Como os preços dos produtos afetam as vendas?

Utilizar análise causal, por exemplo com controle sintético para confectionar exemplos hipotéticos quando os dados não existem. Utilizaria a biblioteca causalinference.

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

Podemos calcular o n-percentil dos clientes com maior valor comprado em todo o tempo ou em períodos. Depois verificar quais informações são suficientemente frequentes para que se possa considerar como um perfil. Adicionalmente podemos utilizar técnicas de agrupamento para verificar se os principais clientes se distanciam dos demais em termos de perfil.

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

Utilizaria técnicas de agrupamento, provavelmente aliadas a técnicas de redução de dimensionalidade. Também é possível dividir previamente em termos, por exemplo, de faixa de valor em compras, e analisar se e quais dados são correlacionados a cada grupo.

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

Um teste comum nesse case é o teste de hipóteses, ou teste A/B, em que calculamos a probabilidade de probabilidades. Sendo assim consideramos uma hipótese A (por exemplo: A -> Os clientes que mais compram possuem maior faixa de renda). Depois construímos a distribuição das variáveis analisadas (valor em compras X faixa de renda), e calculamos a frequência em que A é verdadeira e A é falsa. Termos um corte de frequência tal que se f < corte rejeitamos a hipótese.

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

Atualmente estou trabalhando em um projeto antifraude com contexto parecido com esse. É muito interessante perceber os detalhes e bottlenecks no dia a dia do desenvolvimento. A engenharia de features é fundamental e não é sabido a princípio se há dados suficientes para responder às perguntas de negócio, sendo que reportar a necessidade de dados, governança, curadoria, etc, é uma das funções do cientista de dados.

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.*

Utilizo a técnica chama item-frequency/ inverse-document-frequency. (TF-IDF). Esse técnica permite mapear cada palavra para uma frequência relacionada a um documento.

Depois de treinar, para um novo documento é possível calcular os valores de TF-IDF, computar uma distância, por exemplo similaridade de cosseno, entre os outros documentos ou classes, e selecionar o documento/ classe mais próximo. Poderia utilizar scikit-learn e nltk.

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

Podemos utilizar uma métrica como a perplexidade, ou as próprias distâncias de cosseno. Se houver rótulos (aprendizado supervisionado), podemos calcular erros e acertos de acordo com estes, criando uma matriz de confusão e calculando métricas de acordo com o problema (acurácia, precisão, recall, etc).

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?

Utilizar os novos textos para retreinar o modelo. Comparar a melhora ou degradação. Escolher se irá atualizar o modelo. Possivelmente criar uma automação para desempenhar esse processo.

TESTE 2 –

1) Como funciona o teste de hipóteses e qual é a sua finalidade na análise estatística?

Consideramos uma hipótese A (por exemplo: A -> Os clientes que mais compram possuem maior faixa de renda). Depois construímos a distribuição das variáveis analisadas (valor em compras X faixa de renda), e calculamos a frequência em que A é verdadeira e A é falsa. Termos um corte de frequência tal que se f < corte rejeitamos a hipótese.

2) O que são redes generativas adversárias (GANs) e quais são os possíveis usos dessas redes?

Redes generativas adversárias consistem, na sua forma tradicional. Em um par de redes, uma gerativa e outra discriminativa. A rede busca fazer com que sua adverária "erre", gerando ruído nos inputs. A rede discriminativa é treinada para acertar, apesar do ruído, de forma que ambas as redes treinam uma a outra, utilizando como função objetivo uma fómula do tipo minmaxloss.

3) O que são modelos de linguagem? Qual a diferença entre LLMs e modelos de linguagem tradicionais?

São modelos de estatística ou aprendizado de máquina que visam classificar, gerar, agrupar ou transformar textos em linguagem natural. LLMs, sigla para large language models se diferenciam por: a grande quantidade de dados e processamento para o treinamento; por, a princípio, utilizarem a arquitetura de rede neural transformers.

4) Suponha que você tenha um conjunto de dados com três ou mais grupos para comparar e deseja determinar se há diferenças significativas entre eles. Descreva como você escolheria entre o teste ou outras técnicas estatísticas

Poderia utilizar estatística descritiva, técnicas de redução de dimensionalidade aliadas a métricas de distâncias. Algoritmos de clusterização. Poderia utilizar as bibliotecas scikit-learn, scipy, numpy.

5) Qual é a importância do pré-processamento de texto em tarefas de NLP? Quais são as etapas comuns no pré-processamento de texto?

O pré-processamento é importante para eliminar ruídos e padronizar os dados. Por exemplo: palavras como correr, corri, correu, são praticamente a mesma mas precisam ser preprocessadas para que isso seja levado em consideração computacionalmente, processo que é conhecido como Lematização. Outros processos são:

Padronização de minúsculas
Remoção de acentos
Stemização (parecido com lematização, porém busca gerar tokens, nem sempre é necessário).

Radicalização: Substituir cada palavra por seu radical linguístico.

6) Descreva o processo de vetorização de texto e como modelos de linguagem como o **Word2Vec ou o TF-IDF podem ser usados para representar palavras e documentos.**

Vetorização consiste em transformar dados não numéricos (por exemplo, texto) em dados numéricos.
Na técnica word2vec, por exemplo a arquitetura CBOW, utilizamos uma rede neural e preprocessamos os inputs tal que cada grupo de texto terá n palavras, sendo que a palavra central é aprendida pela rede e as outras são o input. Dessa forma, se n = 5, para o texto "É importante estudar aprendizado de máquina", alguns inputs seriam:

E importante _ aprendizado de
Importante estudar _ de maquina
Etc..

Utilizo a técnica chama item-frequency/ inverse-document-frequency. (TF-IDF). Esse técnica permite mapear cada palavra para uma frequência relacionada a um documento.
Depois de treinar, para um novo documento é possível calcular os valores de TF-IDF, computar uma distância, por exemplo similaridade de cosseno, entre os outros documentos ou classes, e selecionar o documento/ classe mais próximo.

7) O que é a análise de sentimento em NLP e quais são os principais métodos para realizar essa tarefa? Como você avaliaria a eficácia de um modelo de análise de sentimento?

É analisar um sentimento relacionado a um texto. Um exemplo comum é analisar se comentários sobre um produto são positivos ou negativos. Geralmente é modelado como aprendizado supervisionado com classificação binária.
Algoritmos de análise de sentimento incluem: Análise de frequência de palavras em classes, análise bayesiana.

8) Qual é a diferença entre a classificação de texto e o agrupamento (clustering) de texto em NLP? Em que situações cada um é mais apropriado?

Classificação de texto geralmente se refere a uma estratégia de aprendizado supervisionado em que temos as "respostas" para realizar o treinamento. Agrupamento consiste em técnicas para agrupar entidades de dados e não está relacionado a termos "respostas" previamente.

Agrupamento é a estratégia possível quando não temos acesso aos rótulos. Classificação em geral está relacionada a uma melhor resposta quando há dados suficientes para treiná-la. É comum combinar ambas as técnicas.

9) Explique o conceito de reconhecimento de entidades nomeadas (NER) em NLP e suas aplicações práticas.

Esse conceito consiste em reconhecer padrões semânticos em um texto, de acordo com anotações de cada entidade. Por exemplo, reconhecer nomes próprios. Muito utilizado para detecção de padrões e busca.

10) Como você lidaria com problemas de desequilíbrio de classe em tarefas de classificação de texto em NLP? Quais estratégias seriam eficazes?

<u>Algumas abordagens</u>:
- Upsampling: gerar dados sinteticamente, para aumentar a classe menor. Pode utilizar técnicas como interpolação ou I.A. generativa. Requer cautela pois introduz viés.
- Downsampling: eliminar dados da classe maior. Requer cautela pois introduz viés. Além disso a redução de dados pode causar underfitting. Pode ser realizar amostragens e avalia-las estatisticamente de forma a computar o viés.