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.*

Eu prefiro seguir uma metodologia formal, no caso CRISP-DM, a qual é um roteiro de etapas a serem seguidas, isto traz um certo padrão para a solução, já contemplando os possíveis ciclos de reajustes para cada etapa. É importante deixar claro que as etapas de modelagem possivelmente irão variar dependendo do objetivo, já que o fato de ser uma tarefa de classificação, predição, ranqueamento ou outra, irá determinar os dados de entrada para o treinamento do modelo. A modelagem também é iterativa, ou seja, pode ser necessário voltar a alguma etapa anterior para reajustes. Listo abaixo alguns pontos a serem tratados.

- **Entendimento do Negócio:** Onde se têm contato com especialistas da área cliente para o entendimento do problema, determinação dos objetivos, e métricas aceitáveis. *Aqui serão realizadas reuniões tomando anotações das definições;*
- **Obtenção dos Dados**: Os dados podem ser disponibilizados de diversas formas, sejam em arquivos, API, bancos de dados SQL, ou NoSQL. Para cada caso é necessário um estudo diferente, seja das funções a serem utilizadas ou das estruturas das tabelas, de todas as formas deve-se chegar em uma estrutura equivalente a uma tabela única, um dataframe que permita o trabalho. *Ferramentas: Biblioteca Pandas, Requests, e para bases SQL e NoSQL existem pacotes específicos para cada um;*
- **Análise Exploratória de Dados (EDA)**: A anaĺise exploratória visa o entendimento dos dados, sejam na sua composição ou no seu comportamento estatístico, para isto são verificadas a existência de dados não preenchidos (ausentes) e seu comportamento, autliers, variância, correlação entre features, escalas, e outros casos. Normalmente fazendo uso de plotagens e listagens. É importante salientar que as 3 features textuais devem seguir um fluxo distinto, identificando se são ou não categóricas, já que esta informação altera a forma de trabalhar com elas. *Ferramentas: Pandas, Matplotlib, Seaborn, Missingno, SciPy, Statsmodels, NumPy, e outros;*
- **Divisão dos dados entre Treinamento e Teste**: A base deve ser dividida a fim de ter uma separação entre dados de treinamento e de teste, havendo a possibilidade de uma mterceira base de validação. Normalmente a base pode ser dividida em 70% para treinamento e 30% para teste, sendo importante considerar no caso de features categóricas e da feature target (rótulo, ou classe) a sua distribuição, a amostragem deve tentar ser balanceada. Aqui também é necessário separar a feature target (o rótulo, a classe). Já que esta não pode fazer

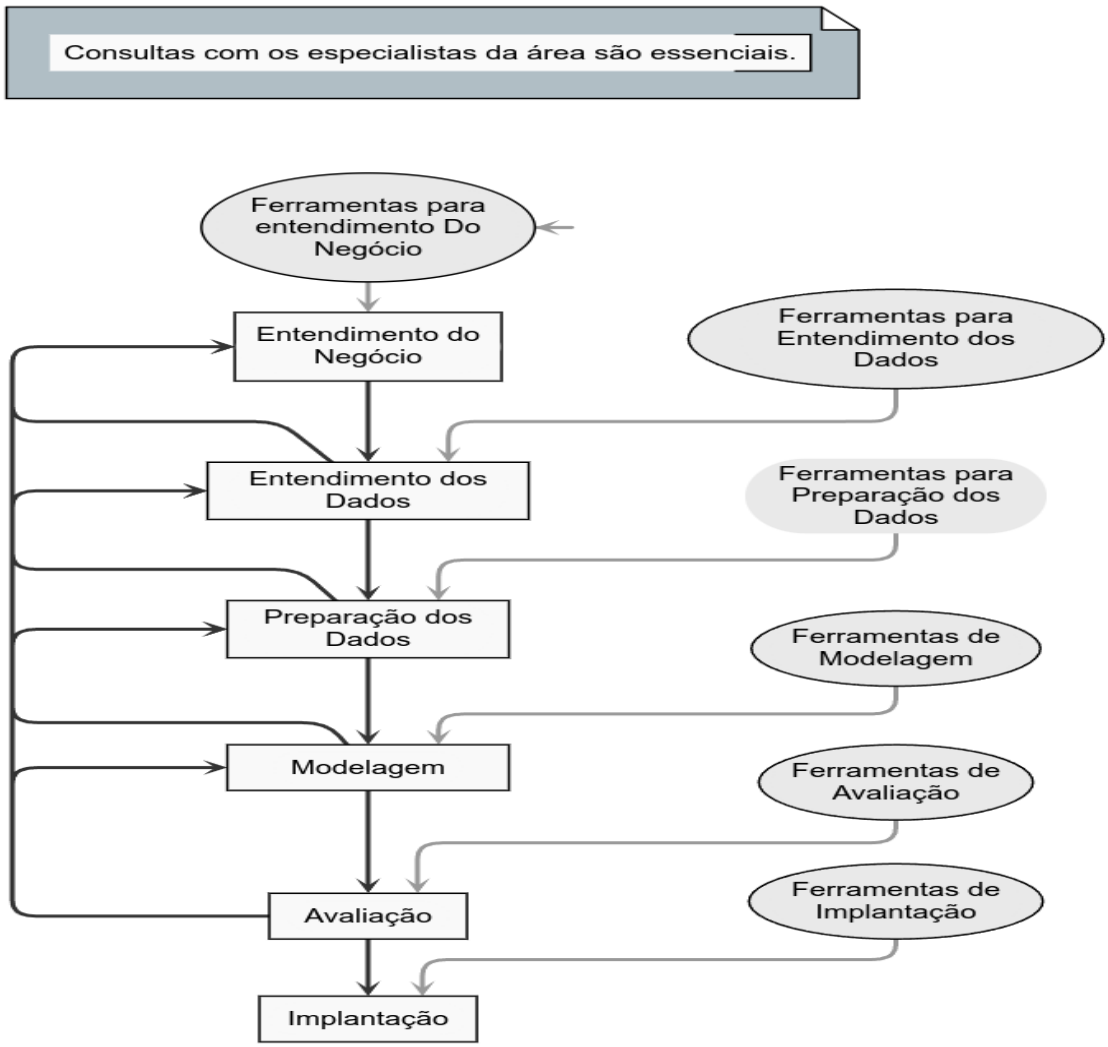
parte do treinamento ou da entrada do teste, sendo usada para tarefas específicas durante o processo de seleção de features e de verificação no escores de desempenho. *Ferramenta: Pandas;*

- **Pré-processamnto dos Dados**: Fazendo uso do conhecimento adquirido na etapa anterior, passamos a trabalhar os dados, aplicando algoritmos de para o tratamento de dados ausentes, o qual pode variar, indo da eliminação da feature (muitos dados ausentes), até o preenchimento utiliazndo a mediana, média ou utilizando algo mais sofisticado como um KNN Imputer, de todas as formas é necessário "marcar" os dados que forem preenchidos já que são estimados e não reais. Outras ações que cabem nesta etapa são a norrmalização ou . padronização dos dados numéricos, os quais devem ficar na mesma escala. Os dados textuais possuem tratamento separado, onde caso sejam categóricos podem ser codificados, transformando-os em uam representação numérica, para isto existem diversos algoritmos e a escolha irá depender da quantidade de categorías existentes na feature, ou de outro comportamento da mesma, de todas as formas é interessante realizar testes; mas caso as features textuais não sejam categóricas, o tratamento seria a extração de informações usando técnicas de NLP, iniciando com um pré-processamnto de texto, aplicando tokenização, lematização e vetorização (TF-IDF por exemplo). *Ferramentas: category_encoders, Pandas, Gensim, Word2Vec, NLTK, spCy, sklearn, BERT, GPT, HuggingFace Transformers;*
- **Balanceamento de Classes**: Uma base de dados com classes desbalanceadas podem implicar que o modelo tenha uma desproporcionalidade no tratamento dos dados, causando um favorecimento de uma classe sobre outra, um viés, e um erro maior ao tratar a classe minoritária. Para contornar isto a base deve ser balanceada, o que implica em se ter uma quantidade igual ou próxima de registros para cada classe; as formas de tratar isto são a utilização de técnicas como oversampling (exemplo, SMOTE que gera dados sintéticos) para as classes minoritárias ou undersampling para as classes majoritárias, visando equilibrar a distribuição das classes; vale recordar que outra abordagem para a geração de dados sitéticos é o uso de GANs. *Ferramentas: imbalanced-learn, para o uso de GANs seria contruído um modelo específico que pode ser implementado via pyTorch, Tensorflow ou outra lib.*
- **Combinação ou Seleção de Features**: Com a finalidade de reduzir a dimensionalidade pode ser interessante realizar, primeiro, uma seleção de features eliminando features que possuem variância igual ou próxima de 0 (zero), assim como a escolha de apenas uma feature dentro de um conjunto onde a correlação é alta, seja positiva ou negativa; existem também alguns algoritmos específicos para auxiliar na seleção de features como no caso do Recursive Feature Elimination (RFE), Random Forests, testes estatítisticos univariados, e Boruta entre outros. Outra possibilidade é a combinação de features (redução de dimensionalidade), o que implica no uso de PCA, t-SNE ou UMAP entre outras, em alguns casos pode-se combinar algumas delas, por exemplo aplicar PSA e depois t-SNE. Isto irá

ajudar na plotagem (visualização) da dsitribuição dos dados e na execução do treinamento do modelo. *Ferramentas: Pandas, sklearn, Boruta.*

- **Seleção e Treinamento do Modelo**: Já que os dados são mistos um bom vaminho a seguir é a escolha de algúm modelo robusto como Random Forest, ou da família Gradient Boosting Machines (GBMs) como XGBoost, LightGBM, CatBoost (onde este último pode ser interessante), ou utilizar uma rede neural. Um caminho interessante é fazer uso de AutoML para selecionar um subconjunto de algoritmos e assim poder fazer uma escolha mais acertiva desde que se aplique ajuste nos hiperparâmetros, lembrando que a experimentação é um ponto importante, assim como o considerar os recursos computacionais disponíveis. Após o treinamento é necessário validar os resultados, verificar métricas de desempenho, neste ponto podemos usar um cross validation como o k-fold para avaliar a generalização do modelo e evitar overfitting. *Ferramentas: sklearn para modelos mais simples, xgboost, lightgbm, catboost para GBMs, tensorflow ou pytorch para redes neurais, e no caso de experimentação com AutoML podemos usar Auto-sklearn, AutoKeras, PyCaret e Autoi-PyTorch;*
- **Avaliação de Desempenho**: Devido o desbalanceamento de classes, devemos focar em métricas como precision, recall, F1-score, e AUC-ROC, além da acurácia, para uma avaliação mais completa. O uso de Matriz de Confusão para a analise da matriz de confusão a fim de entender o desempenho do modelo em cada classe.*Ferramentas: sklearn.metrics;*
- **Ajuste fino e Otimização**: A otimização dos hiperparâmetros é uma atividade importante para o treinamento e produção do modelo, utilizar técnicas como Grid Search ou Random Search é extremamente recomendado. *Ferramentas: sklearn;*
- **Interpretação do Modelo**: Ao ser criado o modelo, é importante fazer uso de algumas ferramentas para entender o quê o leva aos resultados dados (previsão), isto deve ser feito preferencialmente com apóio de especialista da área, já que ele possui um conhecimento aprofundado da importância das features e dos dados. Podem ser utilizadas técnicas como SHAP (SHapley Additive exPlanations) para interpretar modelos complexos. *Ferramentas: shap;*

Segue um diagrama para um melhor entendimento do fluxo do processo macro de modelagem.



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

Para a apresentação perante a área contratante é importante conhecer de antemão os stakeholders que estarão presentes, isto com a finalidade de ajustar a linguagem da apresentação e alinhar com os objetivos pretendidos por cada um, conhecer a formação ou no que atua é fundamental. A apresentação deve ter aspectos visuais como diagramas de processos e de tomadas de decisão, sendo objetivo e sem excesso de informações, mas caso ocorram questionamentos mais informações podem ser apresentadas de forma verbal. A apresentação deve ser iniciada com uma rápida introdução ao problema, apenas para contextualizar e definir o escopo da solução, o que inclui a fonte, lembremos que a área contratante conhece o problema e seus objetivos ao requerer a solução, então não se requer entrar nestes detalhes a não ser que surja alguma discussão.

Feita a introdução, deve-se apresentar os resultados no desempenho, e comparar com o desempenho estipulado como meta, dando esclarecimentos caso a meta não tenha sido alcançada, informando escores de acertos e erros, colocando exemplos práticos. Neste ponto também é importante falar sobre estimativas de impacto, tais como aumento de ganhos ou redução de perdas, ou seja, o impacto da solução no problema da empresa.
É interessante apresentar lições aprendidas e recomendações, medidas coretivas ou melhorias para um aumento do impacto da solução, neste ponto se deve estar aberto a uma pequena discussão onde stakeholders podem fazer seus apontamentos. Podem ser disponibilizados anexos técnicos caso seja solicitado, mas dificilmente ocorre.
Ferramentas: Powerpoint ou equivalente, apresentar um dashboard também pode ser positivo, especialmente se fizer parte da solução mostrando métricas em tempo real.

c) *Como faria a validação desse modelo?*

Irei aqui me aprofundar no que foi descrito no item “a”, aonde algumas ações e ferramentas utilizadas durante a modelagem nos dão maior segurança para uma validação, sendo as etapas necessárias para a validação descritas a seguir:

- **Divisão do Conjunto de Dados**: O conjunto de dados original deve ser dividido em dois ou três subconjuntos, onde 70% dos dados serão para treinamento, 15% ou 30% para teste caso sejam dois subconjuntos, e caso se queira, 15% para validação;
- **Validação Cruzada (Cross Validation)**: Aplicação de algoritmos como o K-fold no conjunto de treinamento a fim de avaliar a estabilidade e confiabilidade, e desde que temos classes desbalanceadas, podemos aplicar uma variante chamada Stratified k-Fold para preservara proporção entre as classes;
- **Avalidação**: Usar métricas como Precisão, Recall e F1-Score para entender o desempenho do modelo, AUC-ROC para avaliar a capacidade do modelo de discriminar entre as classes, e Matriz de Confusão: Para visualizar o desempenho do modelo em cada classe e identificar se há tendências de erros em classes específicas;
- **Análise das Curvas de Aprendizado**: Para isto devemos plotar o desempenho do modelo nos conjuntos de treinamento e validação;
- **Testes de Significância Estatística**: Este teste é realizado para verificar se as diferenças nos desempenhos são significativas ou não;
- **Validação**: Realiar as etapas descrita considerando o conjunto de dados de validação e não mais o de teste;

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

O melhor caminho é criar uma pipeline, onde uma série de tarefas é executada de forma automatizada, estas etapas vão ser constituídas com base no que foi criado e aprendido durante a

modelagem, adicionando o monitoramento das métricas e uma seleção de parte dos dados processados juntamente com os respectivos resultados do modelo para avaliação humana.
É importante planejar estratégias de “atualização incremental do modelo”, que pode ser um retreinamento periódico do modelo. A avaliação de desvios de conceito é outro item importante, já que alguns conjuntos de dados podem ter sua distribuição alterada durante o tempo, tenhamos como exemplos a mudança nas leis, ou hábitos alimentares ou de consumo, lembrando que modelos são representações matemáticas, estatísticas de um conjunto de dados.
Ao colocar uma pipeline em produção, devemos nos preocupar com a segurança da solução, desde a anonimização de dados, até acesso indevido ou vazamento dos dados. É importante também ter um controle de versionamento como um Git para o código fonte, e o DVC para os dados.
Ferramentas: AirFlow, MLFlow.

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

Após a validação, o ideal é colocar a solução em um ambiente de produção para execução com um cliente piloto, para isto o caminho mais comúm é a criação de um webservice que será acessado via uma interface, uma API que disponibiliza as funções para acesso ao modelo. Isso deve ser encapsulado em um container Docker, que é parecido a uma máquina virtual, o que facilita a sua implantação, no ambiente produção é importante a orquestração dos containers, isto para um balanceamento das solicitações, distribuindo entre vários servidores os chamados e assim evitando gargalos no tempo de resposta.
Faz parte do ambiente de produção, ou do seu planejamento, cuidados com o os recursos conumidos pela solução, monitoramento e loggin, segurança e a Integração Contínua e Entrega Contínua (CI/CD), é importante considerar as possíveis atualizações nos modelos e ferramentas utilizadas no webservice.
Ferramentas: FastAI, Docker, Loggin.

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

A modelagem de soluções de AI é composta de estudos e testes, é uma busca de uma solução de ciência aplicada, não existindo uma “receita de bolo”, mas obviamente não se deve procurar a perfeição, e sim um equilíbrio entre tempo, recursos e o desempenho da dolução. As ferramentas, as técnicas, os algoritmos podem mudar dependendo do objetivo e da composição, situação, dos dados. É importante sim ter uma noção dos possíveis caminhos a seguir, mas sempre exsitirá aprendizado.

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.

Aplicaria Análise Exploratória dos Dados (EDA) para melhor compreensão da sua composição e comportamento estatístico. A análise já pode gerar gráficos para os seguintes pontos:

- **Análise Temporal**: Para desmepenho das vendas dos produtos ao longo do tempo, podensdo identificar sazonalidades;
- **Análise por Categoria de Produto**: Avaliando as categorias e seu impacto na receita;
- **Análise de Clientes**: Identificando segmentos socioeconômicos;
- **Análise de Desempenho Regional**: Comparando diversas lojas ou até com dados disponibilizados por fornecedores, podemos usar algoritmos como ANOVA para identificar diferênças estatísticas significativas;
- **Análise de Tendências e Padrões de Copras**: Para identificar combinações de produtos, para isto usamos análise de cestas de compras com Apriori;
- **Análises de Promoções e Descontos**: Para medir a eficácia da aplicação de descontos nos produtos;
- **Avaliação de Estoque**: Para identificar os produtos com alta e baixa rotatividade (venda);

Ferramentas: Pandas, Matplotlib, Seaborn, sklearn, scipy, geoPandas.

b) Como você responderia as seguintes questões:

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

Para conhecer isso é necessário realizar a agregação dos dados, por dia, mas pode ser agregado por outra medida como semana, ou mês. Com isso podemos visualizar tendências com gráficos de linhas ou criar gráficos de barras para visualizar o desempenho propriamente dito, o qual depende da quantidade de itens vendidos ou até do seu impacto na receita, o que depende do seu valor e margem de lucro. En fim, o desempenho é dado pela sua quantidade de vendas ou impacto na receita. Para uma análise temporal podemo usar a lib Prophete ou algoritmos como ARIMA e SARIMA, e assim tentar prever um desempenho futuro. Na análise de desempenho é necessário cuidar a sazonalidades de produtos. Ferramenta: Pandas, Prophete, ARIMA, SARIMA, ou a própria base de dados (caso seja a fonte de dados), o SQL possibilita realizar as agregações de ftorma fácil e rápida

ii. Quais são os produtos mais vendidos?

Como no item anterior será necessário realizar a agregação das vendas por produto, seja por dia, semana ou mês, mas para este caso vamos relatar apenas os que possuem maior número de vendas desntro do período de tempo solicitado, uma análise comúm é a curva ABC. *Ferramenta: Pandas, SQL*

iii. Como as vendas variam por categoria de produtos?

Será necessário realizar a agregação não mais por produtos e sim por categorias, a análise é semelhante a anterior, apenas sendo aplicada a categorías, podendo usar a mesma metodologia da curva ABC. Talves seja interessante uma análise por faixas de preços dentro das mesmas categorias, apenas para um melhor detalhamento

iv. Qual é a distribuição dos valores de venda?

A distribuição dos valores de venda pode ter enfoques distintos, pode ser por produtos individuais ou por categorias, da mesma forma que a janela de tempo pode ser por faixas de tempo distintas, podendo mostrar se as vendas são concentradas em uam faixa determinada de valor, e se esta faixa varia conforme alguma sazonalidade. Desta forma podemos visualizar se o maior número de clientes se concentra em alguma faixa específica e quais produtos pertencem a essa faixa, apontando a grupos mais lucrativos. Para isto podemos realizar algumas análises estatísticas e visuais como:

- **Estatística Descritiva**: Gerando estatísticas básicas para ter uma noção da distribuição dos valores de venda dos produtos, valor mínimo e máximo, desvio padrão, variância, media e mediana são um bom começo e podem ser comparados com os gráficos que podem ser gerados;
- **Histogramas**: É uma das melhores maneiras de visualizar a distribuição de dados numéricos, mostrando a frequência de ocorrência de diferentes intervalos de valores de venda. De fácil visualização e interpretação;
- **Boxplot**: Este gráfico nos fornece uma visualização interessante sobre os principais atributos estatísticos dos dados, especialmente se queremos comparar produtos ou grupos de produtos. Podemos também usar o Violinplot o qual nos fornece curvas onde visualmente vemos a variação das concentrações dos dados, e com um plus ao podermos dois conjuntos / clases distintas juntas;
- **Gráficos de Densidade**: O conhecido KDE é uma alternativa ao histograma;
- **Análise de Quartis e IQR**: Quartis e o intervalo interquartil (IQR) pode ajudar a entender melhor a dispersão dos dados e identificar outliers;
- **Segmentação**: A clusterização como também é chamada é um método não supervisionado que pode nos dar alguns insights interessantes, é muito utilizada em análise exploratória.

Ferramentas: Pandas, Seaborn, Matplotlib, KDEplot, Sklearn.

Considerações: É interessante verificar a normalidade dos valores de vendas, assim como verificar os outliers, a sua origem pode ser erros de digitação, fraudes ou riscos e oportunidades. Vale ressaltar que algumas das análises podem substituir outras, nos dando as mesmas informações, porém, ao realizar uma apresentação devemos escolher gráficos de fácil interpretação, já que nem todos sabem como extrair as informações de um Boxplot ou Violinplot por exemplo.

v. Como os preços dos produtos afetam as vendas?

Em uma perspectiva simplista um produto com preço alto irá vender pouco, mas isto possui influência de outras variáveis como é o caso da marca, da qualidade, do preço de venda em concorrentes, da classe econômica do cliente de maior representatividade. Para analisar como os preços dos produtos afetam as vendas, podemos adotar uma abordagem multifacetada que utilise análise estatística, modelagem preditiva e visualizações de dados. Esta análise ajudará a entender a relação entre preço e volume de vendas, identificar pontos de preço ótimos e entender a elasticidade de preço dos produtos. Podemos aplicar as seguintes abordagens:
=

- **Análise Descritiva**: Para entende as faixas de preços e como as vendas são distribuídas nestas faixas;
- **Correlação entre Preço e Venda**: Para identificar se existe correlação entre os valores e a quantidade de produtos vendidos, uma matriz de confusão pode ser interessante, mas as features de maior correlação podem ser listadas agrpando-as em conjuntos;
- **Análise de Regressão**: Para analisar de forma mais detalhada a relação entre preço e vendas;
- **Análise de Elasticidade de Preço**: usada para verificar o impacto nas vendas ao alterar o preço do produto;
- **Clusterizações de Produtos**: Para analisar o impacto do preço sobre os clusters;
- **Teste A/B de Preços**: Este teste pode ser aplicado para verifcar o impacto do preço nas vendas

Existem algumas considerações que devem ser apontadas, como o fato que nem sempre existe uma relação linear entre preço e venda, para isto devemos considerar mais de um modelo de regressão; da mesma forma devemos considerar que o preço não é o único fator que pode afetar as vendas, podendo variar devido ao marketing, região geográfica, sazonalidade, e até cultura.
Ferramentas: Pandas, Matlibplot, Seaborn, statsmodels, scikit-learn.

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

Para analisar o perfil de compra dos clientes podemos fazer uso de RFM (Recência, Frequência, Valor Monetário) a qual, de forma simplificada, consiste em calcular a Recência, Frequencia e o Valor monetário de cada cliente, isto para poder calcular escores que indicarão a classe de importância do cliente. Esa classe de importância são denominações que podemos dar, porém sabemos que escores mais altos indicam maior importância do cliente, mas no final é possível produzir uma matriz onde cada cliente pode ser classificado. Existem outras abordagens, incluindo a clusterização, onde pode ser usado o K-Means ou outro algoritmo.

É interessante criar classes “presonalizadas” para a empresa em questão, já que depende dos objetivos, da visão da empresa. Como em outros casos de segmentações de clientes e produtos, o enriquecimento co dados sociodemográficos é interessante, já que os padrões de vendas vão variar se o cliente é da classe A, B ou quaqluer outra, e para se ter uma melhor percepção é interessante criar um visualização para ver os percentuais de clientes de cada classe econômica.

Ferramentas: Pandas, Matplotlib, Seaborn, Sklearn.

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

A resposta ao item anterior é válida aqui, com a diferença de que não vamos nos concentrar no perfil dos principais clientes e sim no perfis existentes. As técnicas de RFM e segmentação são as mesma, a análise visual é a mesma, apenas focando em outra informação. Oq ue posso complementar é que se seguimos a solução de clusterização, é interessante analisar as features de maior importância para o resultado, isto nos leva a focar qualquer ação em pontos específicos para uma melhoria na relação com o cliente ou na sua compreenção, existindo uma análise qualitativa

Ferramentas: Pandas, Matplotlib, Seaborn, Sklearn.

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

Para provar uma hipótese referente aos segmentos de clientes, especialmente se a hipótese envolve comparações de médias ou proporções entre os segmentos, podemos utilizar testes estatísticos como o Teste T, ANOVA (Análise de Variância), ou o Teste de Qui-Quadrado, os quais são os comunmente utilizados. É importante verificar a normalidade e homogeinidade dos dados.

Ferramentas: Pandas, Scipy, statsmodels

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

Se possível eu enriqueceria os dados com dados georeferenciados, socioeconômicos, climatológicos, eventos locais (festas, feriados, e outros), dados de mobilidade, e incluso alguma informação financeira importante como o aumento de algum imposto ou a quebra de alguma empresa… Até pensando nisso pode-se realizar alguns estudo utilizando inferência causal. O enriquecimento sempre é uma opção interessante, e alguns dados podem ser obtidos via dados abertos governamentais

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.*

Existem algumas abordagens possíveis, irei descrever superficialmente as principais e depois irei indicar minha sequência de :opçoẽs. Segue

- A manual onde as atividades de NLP são implementadas totalmente pela equipe, incluso a busca. Para este caso são criadas as rotinas de préprocessamento de texto, onde a anonimização é importante; implementa-se a rotina de vetorização, que pode utilizar o TF-IDF (Term Frequency-Inverse Document Frequency) , Word2Vec, ou Doc2Vec (o chamado embeddings de palavras, o BERT também pode ser utilizado), isto para poder gerar uma representação numérica (vetores) a fim calcular a distância. Com isso é possível vetorizar o novo documento e calcular a distância dos diocumentos na base, devolvendo os mais próximos;
- Pode-se utilizar algoritmos como BERT e GPT para uma abordagem mais sofisticadas, no caso de LLMs seria possível aplicar fine-tunig e embedding para realizara tarefa, criando um prompt que possa indicar os processos semelhantes;
- Outra abordagem que pode ser interessante é o uso de Bases de Dados Vetoriais, as quais usam indices vetoriais justamente para esse tipo de problema, o que implica em uma ótima escalabilidade. Um ponto que pode ser interessnate é que este tipo de banco de dados é multimodal, ou seja, permite a inclusão não apenas de textos, mas sim de outros tipos de documentos e dados como é o caso de imagens e incluso vídeo. Um caso de banco vetoria que é FOSS e em Python é o VectorDB da Jina AI.

Eu nunca utilizei um banco de dados vetorial (está nos meus planos), mas acredito ser uma boa opção, que eu escolheria, e com muitas vantagens, como eficiência nas buscas; escalabilidade; precisão, e uma certa flexibilidade (possuem parâmetros para ajustes). Em segundo lugar eu implementaria a tarefa utilizando LLMs, criando um modelo próprio para isso, mas com a desvantagem de que um modelo não é uma base de dados, o que pode implicar em um custo muito elevado na sua manutenção. A vantagem de implementar a solução por inteiro é um controle maior sobre o projeto, e a maior facilidade em incluir um feedback do usuário dando nota para as recomendações, mas é algo que pode ser implementado para todos os casos.

Ferramentas: Spacy, NLTK, Word2Vec, Hugginfaces, VectorDB.

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

A implementação do feedback já mencionado é uma forma de avaliar o sistema, e é a mais confiável; outra abordagem que também envolve o usuário é monitorar quais são os documentos selecionados, de quais faz download, ou se disponibilizada uma visualização do documento, pode-se medir o tempo que passa lendo o documento (o documento aberto e tendo alguma iteração com este). O feedback, ou as outras medidas, poderiam gerar um escore que ajuda em aprimorar as recomendações.

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?*

Para o caso do Banco de Dados Vetorial, seria apenas feita uma pipeline onde se realiza o preprocessamento e vetorização do documento, para logo em seguida salvar o vetor na base. Não é necessário nenhum treinamento extra, ao meu ver. A função de feedback continuará ativa e já incorporará os novos documentos.

TESTE 2 –

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

O teste de hipóteses é um método estatístico que permite verificar se uma suposição (hipótese) sobre um parâmetro populacional (hipótese nula) é verdadeira, com base em dados amostrais. A finalidade é avaliar a validade de afirmações ou premissas, determinando se as evidências dos dados são suficientes para rejeitar a hipótese nula em favor de uma hipótese alternativa, dentro de um nível de confiança estabelecido, geralmente através do cálculo do p-valor. Normalmente utilizamos p-valor igual ou menor que 0,05 para rejeitar a hipótese nula, já que o p-valor é o menor índice de significância para rejeitar a hipótese nula. Consideremos que o teste de

hipóteses é como um experimento para verificar uma suposição. Imaginemos que uma moeda é lançada várias vezes para descobrir se a moeda é justa (para isto deve dar o mesmo número de caras e coroas, aproximadamente), anotando os resultados, se no final o número de uma das faces é muito superior ao da outra, isto implica que podemos duvidar que a moeda seja justa.

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

Redes Generativas Adversárias (GANs) são um tipo de inteligência artificial, uma arquitetura, composta de um gerador e um descriminador, existindo uma “competição” entre os dois. O papel do gerador é criar dados falsos que se assemelhem a dados reais, estes dados podem ser textos; imagens; sons; por outro lado o papel do descriminador é identificar se o dado criado é verdadeiro ou falso, iniciando assim uma competição entre as duas entidades. As GANs são utilizadas na geração de textos, imagens realistas ou hiperrealistas, vídeos, música, questoões relacionadas com arte. O motivo dass alucinações que muitas vezes se manifestam em LLMs pode ser um fator de melhoria para questões artisticas, existe uma discussão sobre o assunto de pensamento ráido e lento.

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

Modelos de linguagem são sistemas de AI que aprendem a compreender e gerar texto imitando a linguagem humana, ou a extrair informações desses textos. A principal diferença entre os Modelos de Linguagem de Grande Escala (LLMs, como GPT-3) e modelos tradicionais está no tamanho do “corpus” de treinamento e na capacidade: LLMs são treinados com quantidades massivas de dados (corpus é o conjunto de textos, que são os dados), permitindo-lhes entender e gerar texto com uma complexidade e nuance muito maiores, tornando-os mais versáteis em uma ampla gama de tarefas de linguagem natural.

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

Utilizaria a ANOVA, ele testa se as médias entre os grupos diferem significativamente. Se o resultado indicar diferenças, testes pós-hoc, como Tukey ou Bonferroni, podem ser aplicados para identificar quais grupos diferem especificamente. Também pode ser aplicado algum algoritmo de clusterização, visualmente irá mostrar padrões que podem nos dar insights, incluso podemos extrair a distância entre elementos ou dos grupos em si, mesmo não dando informaçÕes estatísticas pode ser uinteressante. Incluso recomendo em questão anterior.

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

É extremamente importante, já que é nesta etapa que se realiza a limpeza e normalização do texto, isto impacta na eficiência dos modelos. As etapas podem variar dependendo do tipo de texto e a finalidade do modelo, irei relacionar as mais comúns: Converter todo o texto para minúsculas; remoção de stopwords, pontuações e caracteres não relevantes; tokenização (dividir o texto em palavras ou frases), lematização ou stemming; vetorização. Na limpeza pode ser necessário eliminar links http e realizar anonimização.

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.

A vetorização de texto converte palavras ou documentos em vetores numéricos (uma lista de números, onde cada número representa uma característica do texto), permitindo que modelos de machine learning os processem. O TF-IDF destaca palavras importantes em documentos, ponderando a frequência da palavra (contagem da palavra no texto) pelo inverso de sua frequência nos documentos. O Word2Vec gera vetores que capturam relações semânticas entre palavras, treinando redes neurais em contextos de palavras. Ambos transformam texto em formatos numéricos que refletem significados ou importâncias, facilitando tarefas como classificação ou busca de documentos.

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?

A análise de sentimentos classifica emoções, ou opiniões no texto, a classificação pode variar conforme a necessidade, mas a mais simples é composta por três classes que são a positiva, neutra e negativa. A grosso modo são utiliadas palavras cuja polaridade (positiva ou negativa) é predeterminada, realizando uma contagem destas palavras existentes no texto a analisar, sendo o resultado da soma das polaridades o valor final do sentimento. A abordagem mais sofisticada envolve redes neurais, modelos como BERT para realizar a classificção dos sentimentos, ao quais podem ser associados a uma ou mais entidades contidas no texto analizado. Para avaliar a eficácia do modelo podemos considerar um conjunto de dados de teste já rotulado com as classes de sentimentos desejados, cujos rótulos serão utilizados na comparação com as classificações geradas pelo modelo sobre a mesma base, para isto mede-se a precisão, recall, F1-score e, às vezes, a AUC-ROC.

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 é um processo supervisionado que atribui categorias pré-definidas a textos com base em seu conteúdo, ideal para tarefas como filtragem de spam, classificar tipos de documentos, ou na análise de sentimento, todos casos onde as categorias são conhecidas. Clusterização é um processo não supervisionado que agrupa textos semelhantes sem categorias pré-definidas, útil para explorar dados e descobrir padrões ou temas recorrentes, como na organização de documentos ou na sumarização de informações. Existe um caso específico, que não é uma clusterização, apesar de ter um funcionamento parecido, que é a Modelagem de Tópicos (Topic Modelling), o qual agrupa documentos com base na probabilidade de cada toquem existir no documento, esta abordagem nos possibilita identificar os termos mais importantes e com isto chegar no tópico tratado em cada grupo. Para isso usamos LDA, NMF e BERTtopic entre outros.

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

O reconhecimento de entidades nomeadas (NER) é um processo de NLP que identifica e classifica entidades (elementos-chave) em textos, como nomes de pessoas, cidades, países, organizações, ruas, datas, valores monetários, doenças, espécies animais, entre outros. A base de dados (textos) deve ser rotulada para cada entidade no texto, indicando o rótulo e a localização da entidade (inicio e fim), existem bases específicas para diversas áreas de conhecimento como a biomedicina, e o direito. As aplicações práticas incluem extração de informações para alimentar bases de dados, extração para o enriquecimento de conteúdo para sistemas de recomendação, análise de sentimentos direcionada a entidades, melhora na eficácia de bots e assistentes virtuais. Uma área que têm ganho atenção é a mineração de dados para a criação de grafos de conhecimento que auxiliam LLMs.

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

A resposta dada em item anterior também é válida para textos, e consiste em reamostragem dos dados para equilibrar as classes, seja por oversampling da classe minoritária ou undersampling da classe majoritária; uso de pesos de classe para ajustar a importância das classes no treinamento do modelo; para o caso das classes com pouca representatividade pode-se tentar gerar dados sintéticos atraves de GANs. As métricas de avaliação apropriadas, seriam F1-score ou AUC-ROC, que são menos sensíveis ao desequilíbrio de classes. Além disso, técnicas avançadas como o uso de algoritmos de aprendizado que são intrinsecamente mais robustos a desequilíbrios, como árvores de decisão, que também podem ser consideradas para a classificação, assim como o CatBoost que apesar de ser sensível ao desbalanceamento, ele oferece recursos para lidar com o problema, como a possibilidade de definir pesos para as classes ou usar o parâmetro “auto_class_weights='Balanced'” na configuração do modelo, o que ajuda a ajustar automaticamente os pesos das classes com base em suas frequências..

TESTE 3 - CASE

Exercício 3 respondido em arquivos separados.

Amostra:

Contextualização:

O Base de dados canada_amostra em formato CSV representa um conjunto de empresas do Canadá com a respectiva descrição de seus produtos, dados econômicos e localização.

Assim, podemos caracterizar cada variável:

name: nome da empresa;

description: descrição do produto da empresa;

employees: número de empregados da empresa;

total_funding: Total de investimento já recebido pela empresa;

city: cidade;

subcountry: estado;

lat: latitude da cidade;

lng: Longitude da cidade.

1) Problema:

Deseja-se prospectar empresas que possuam soluções em **tratamento de água** , principalmente, elativas à : **solutions on waste and water, Improve water quality and water efficiency use, water contamination, water for human consumption, water resources** .

a) EXERCÍCIO 1 - Aplique um algoritmo de ML (ou um conjunto deles) capaz de selecionar as principais empresas indicadas para desenvolver a solução de acordo com seu alinhamento com o tema (Justifique a escolha do algoritmo).

b) EXERCÍCIO 2 - Faça uma análise exploratória dos resultados acrescentando as demais variáveis contidas no dataset. Quais insights você pode obter a partir desses dados? Quais são as principais cidades (pólos de desenvolvimento) para essa solução?

c) EXERCÍCIO 3 - EXTRA - Se você terminou o desafio de forma rápida, temos mais algumas perguntas para serem respondidas. Elas, como dito, não são obrigatórias, então sinta-se à vontade em não as responder ou até mesmo respondê-las parcialmente. Essa parte visa observar seu entendimento de um ambiente real de produção.

i) a) organize seus códigos em pacotes garantindo seu versionamento e documentação (bibliotecas auxiliares, etc.).

ii) b) construa testes automatizados para validação do seu pacote.

iii) c) crie uma imagem Docker capaz de executar suas análises em um ambiente de produção.

iv) d) crie um GitHub público e suba todo o código do Teste 3, e disponibilize para avaliação.