**1 .Quais/qual Linguagem de programação tem experiência?**

Eu tenho mais de 6 anos de experiência com **Python**, especialmente para aplicações de ciência de dados, automação de processos, machine learning e desenvolvimento de pipelines de dados. Além disso, também tenho experiência com **SQL** para manipulação e consulta de grandes volumes de dados. Já trabalhei com **PySpark** e **JavaScript** (incluindo frameworks como Angular e Node.js) em projetos de backend e microservices. Tenho conhecimento de AWS, Azure, git, github e Airflow.

**2. Você já teve experiência com soluções de forecast/previsão de demanda?**

Sim

**3. (SQL) Considere a tabela 'vendas' com colunas 'id\_venda', 'id\_produto' e 'valor'. Escreva uma consulta que retorne a média móvel dos valores de venda, calculada com uma janela deslizante de tamanho 3, ordenada por 'id\_venda'**

Minha consulta SQL:

SELECT id\_venda, id\_produto,valor,  
AVG(valor) OVER (PARTITION BY id\_produto ORDER BY id\_venda ROWS BETWEEN 2 PRECEDING AND CURRENT ROW) AS media\_movel\_valor  
FROM vendas  
ORDER BY id\_produto, id\_venda;

Comentário: Decidi usar a partição por **id\_produto** porque faz mais sentido entender a média móvel dentro do contexto de cada produto. Isso nos permite analisar o desempenho de vendas de cada item de forma mais detalhada.

**4 . (SQL) Suponha que você tenha uma tabela 'funcionarios' com as colunas 'id\_funcionario', 'nome' e 'salario'. Escreva uma consulta que calcule o percentual do salário de cada funcionário em relação ao salário médio geral. Utilize uma janela (window function) para realizar essa tarefa.**

Minha consulta SQL:

SELECT id\_funcionario, nome, salario,

(salario / AVG(salario) OVER ()) \* 100 AS percentual\_salario

FROM funcionarios;

**5. (SQL) Considere um banco de dados que armazena informações sobre uma biblioteca. O banco de dados possui as seguintes tabelas:**

**tb\_livros**

**id\_livro (chave primária)titulo autor ano\_publicacao**

**tb\_exemplares**

**id\_exemplar (chave primária)id\_livro (chave estrangeira referenciando livros.id\_livro)disponivel (booleano indicando se o exemplar está disponível para empréstimo)**

**tb\_emprestimos**

**id\_emprestimo (chave primária)id\_exemplar (chave estrangeira referenciando exemplares.id\_exemplar) data\_emprestimo data\_devolucao (pode ser NULL se o livro ainda não foi devolvido)**

**Desafio:**

**Escreva consultas SQL para resolver as seguintes tarefas:**

**- Listar os títulos dos livros que estão atualmente emprestados, juntamente com os nomes dos autores.**

**- Encontrar o** **autor com o maior número de livros diferentes emprestados.**

**- Identificar os livros que nunca foram emprestados.**

**- Calcular a** **média de dias que os livros ficam emprestados, considerando apenas os livros que foram devolvidos.**

**Importante:**

**Seja claro em suas consultas e use aliases apropriados para facilitar a leitura.Considere que a data\_devolucao é NULL para exemplares ainda não devolvidos.Você pode criar aliases para colunas conforme necessário.Se necessário, você pode usar funções específicas do SQL do sistema de gerenciamento de banco de dados que está utilizando. Certifique-se de mencionar a tecnologia ou o banco de dados que você está considerando (por exemplo, MySQL, PostgreSQL, SQL Server, databricks ou hive).**

Minhas Consultas SQL

1. Títulos dos livros emprestados e nomes dos autores.

SELECT l.titulo AS titulo\_livro, l.autor AS nome\_autor

FROM tb\_livros l

JOIN tb\_exemplares e ON l.id\_livro = e.id\_livro

JOIN tb\_emprestimos emp ON e.id\_exemplar = emp.id\_exemplar

WHERE e.disponivel = FALSE

AND emp.data\_emprestimo IS NOT NULL

AND emp. data\_devolucaoIS NULL ;

2. Autor com o maior número de livros diferentes emprestados

SELECT l.autor AS nome\_autor,

COUNT(DISTINCT l.id\_livro) AS numero\_livros\_emprestados

FROM tb\_livros l

JOIN tb\_exemplares e ON l.id\_livro = e.id\_livro

JOIN tb\_emprestimos emp ON e.id\_exemplar = emp.id\_exemplar

WHERE e.disponivel = FALSE

AND emp.data\_emprestimo IS NOT NULL

AND emp. data\_devolucao IS NULL

GROUP BY l.autor

ORDER BY numero\_livros\_emprestados DESC

OFFSET 0 ROWS FETCH NEXT 1 ROWS ONLY;

3. Livros que nunca foram emprestados

SELECT l.titulo AS titulo\_livro

FROM tb\_livros l

LEFT JOIN tb\_exemplares e ON l.id\_livro = e.id\_livro

LEFT JOIN tb\_emprestimos emp ON e.id\_exemplar = emp.id\_exemplar

WHERE emp.id\_emprestimo IS NULL;

4. A média de dias que os livros ficam emprestados, considerando apenas os livros que foram devolvidos.

SELECT AVG(DATEDIFF(DAY, emp.data\_emprestimo, emp.data\_devolucao)) AS media\_dias\_emprestados

FROM tb\_emprestimos emp

WHERE emp.data\_devolucao IS NOT NULL;

Tecnologia Utilizada – SQL Server

**(DATA SCIENCE) O que é normalização de dados no pré-processamento de dados? \***

Uma maneira de realizar a seleção de variáveis

Uma maneira de realizar redução de dimensionalidade

**Uma maneira de dimensionar recursos para uma faixa semelhante de valores**

Uma maneira de detectar e remover valores discrepantes de um conjunto de dados

Uma maneira de transformar dados contínuos em dados categóricos

**(DATA SCIENCE) Como você preveria e enviaria para a área de CRM a probabilidade de venda de perfume na próxima campanha da Natura**  
\*

Com base na experiência que adquiri em projetos realizados em parceria com a equipe de CRM, destaco que tive oportunidade de construir um Modelo de Propensão de Vendas. Esse modelo consegue prever a probabilidade de o cliente realizar uma nova compra dos produtos dentro de uma janela de tempo específica. Entendo que não existe uma fórmula única para modelar problemas complexos que envolvem várias variáveis e eventos externos, mas compartilho um pouco sobre a abordagem que utilizei em outra organização de Varejo e seguirei a mesma referência para a Natura.

Seria interessante criar uma aplicação que gera ofertas personalizadas em um aplicativo ou site. Quando o cliente opta pelo Programa de Fidelidade, no momento da compra, fornecendo seus dados (CPF ou RG por exemplo), conseguiriamos gerar descontos direcionados e ter a rastreabilidade dos dados.

Do ponto de vista técnico, envolveria um esforço com as bases de dados, devido ao alto volume de informações que precisaremos. Trabalharemos junto ao time de engenheiros de dados para direcionar na ingestão dos dados necessários em um Data Lake em núvem, o que facilitaria o processamento do modelo.

De fato, será importante ter o histórico de vendas e de algumas *features* essenciais para o modelo, além de informações sobre clientes, campanhas de CRM, segmentações do CRM e dados externos. Paralelamente, desenvolveriamos camadas de pré-processamento ou ETL dos dados, realizando limpeza, tratamentos, agrupamentos e toda a preparação necessária para as variáveis do modelo.

Para a modelagem, utilizariamos um modelo de classificação, onde a saída era a probabilidade de o cliente comprar um produto, dada uma série de variáveis de entrada. Testariamos diversos modelos e o selecionaria o melhor desempenho (por ex. **RandomForestClassificationModel)**. Esse modelo preveria, a partir de janelas de períodos e de variáveis como ticket médio, recência, períodos sem compras, feriados, entre outras, a propensão de compra para o próximo periodo.

As bases de propensão seriam disponibilizadas para o time de CRM por meio de relatórios iriamos participar semanalmente das discussões de resultados com o time de Marketing/CRM, a fim de identificar oportunidades de melhoria.

Detaque que na empresa que atuei obtive seguinte resultado: Nos testes do modelo, obtivermos uma acurácia de 0.7 ao final do processo e conseguimos trazer mais de 1MM de retorno financeiro na primeira rodada dentro do aplicativo mas com um imenso espaço para melhorias e alavancas.

**Conhecimentos de Forecast:**

**Acesse o Dataset:**

**M5:**[**https://www.kaggle.com/competitions/m5-forecasting-accuracy/data?select=sales\_train\_evaluation.csv**](https://links.uk.defend.egress.com/Warning?crId=671952d5d53302954327e6ad&Domain=capco.com&Lang=en&Base64Url=eNoVyzEOgCAMBdATQScXE-NRzE-tSEQgtEq8vTq88e1mVUei3rs_EEISz-WkTxWLFktWOge3lSYMtZiDA_PVwA-tMMwqSdgmRRJdrCHmRW6kC__1rPcLOmol-g%3D%3D&@OriginalLink=www.kaggle.com)

**PASSO 1: Escolha 1 categoria desse Dataset**

**Considere:**

**- Queremos prever no nível item/estado/mês.**

**- O horizonte de previsão deve ser de 1 mês;**

**Explique suas escolhas descrevendo o processo de modelagem como se estivesse explicando para o Gerente de Negócio de Planejamento Comercial.**

**(não é necessário desenvolvimento de código)**

Ao analisar as categorias do dataset, eu escolheria a categoria HOBBIES para o modelo de previsão porque é uma área com bastante variação e forte impacto de tendências. Diferentemente da categoria de alimentos, que possui um padrão de compra mais constante, produtos relacionados a hobbies tendem a ter uma demanda que muda significativamente conforme a época do ano, lançamentos ou modas. Seria um desafio interessante no meu ponto de vista.

Observando o dataset, podemos notar que as vendas diárias estão organizadas em formato colunar, com o histórico diário registrado em cada coluna. Além disso, temos dois outros dataframes que trazem informações adicionais, incluindo características sobre o calendário, feriados, eventos, além do histórico de preços e lançamentos dos produtos. Investir um tempo na análise e compreensão dos dados é de extrema importância, pois isso nos permite verificar possíveis diferenças entre lojas, departamentos e categorias.

A granularidade dos dados abrange informações sobre lojas, departamentos e categorias, chegando até o ID do produto e a data específica. Para realizar previsões mensais, será necessário elevar a granularidade dos dados. Utilizaria o pacote Pandas como ferramenta para transformar o histórico de vendas de um formato colunar para um formato de linha, facilitando assim o processo de agregação.

Após essa transformação, agregaria os dados mensalmente, incorporando variáveis e recursos relevantes, como preço médio, desvio padrão e o preço do mês anterior. Também incorporaria informações sobre eventos ao longo do ano, contabilizando a quantidade de eventos por mês. Com esse tratamento, a base de dados seria enriquecida com variáveis que auxiliarão nas previsões, tornando-a mais robusta e informativa.

Depois de realizar essa parte, que chamamos de feature engineering, poderíamos pensar em treinar modelos Poderíamos optar por montar um modelo por loja ou por departamento. Neste caso, eu escolheria dividir por loja, sabendo que existe uma diferença significativa entre elas.

Dividir a base em conjuntos de treino e teste é crucial. Eu selecionaria o último mês de vendas como meu conjunto de teste, enquanto os meses anteriores serviriam como janela de treino. Começaria com abordagens simples, como previsões usando média móvel ou modelos estatísticos clássicos como ARIMA. Poderia até optar por uma abordagem mais simplista, replicando o valor do próximo mês igual ao valor atual. Isso me daria um ponto de partida para métricas de regressão como MAPE e MAE.

Após isso, seria hora de utilizar modelos de Machine Learning, como o LightGBM, que particularmente eu gosto bastante. Em seguida, explicarei melhor sobre esse modelo, que é muito valorizado em competições. Também poderíamos aplicar modelos de Deep Learning, utilizando o TensorFlow com um modelo sequencial LSTM, o que certamente melhoraria as métricas de avaliação.

É importante lembrar que, em competições, frequentemente utilizamos 2 ou 3 modelos e os combinamos para tomar decisões. Por exemplo, um modelo pode prever por loja/mês, outro por departamento ou produto, e, ao final, podemos otimizar as previsões de forma a respeitar os limites possíveis. Apesar disso, na prática, esse nível de complexidade não é sempre necessário. Tendo uma previsão estável dentro dos limites, já conseguimos realizar excelentes ações dentro da empresa.

Com certeza, teríamos um trabalho de calibração do modelo, acompanhamento de performance e ajustes pontuais nos dados. No entanto, conseguiríamos alcançar um resultado satisfatório inicial, cientes de que ainda teremos muito esforço pela frente.

**Conhecimentos de Forecast:**

**2- Qual modelo (clássico ou Machine Learning) de forecast você escolheria para previsão.**

Eu começaria com uma abordagem gradual, optando por técnicas mais simples no começo. Para estabelecer um baseline nas métricas de avaliação, usaria um modelo simplista, como a média móvel ou o método naive. Esses modelos ajudam a entender o padrão básico dos dados sem a complexidade adicional de métodos mais sofisticados.

Depois, passaria para modelos estatísticos de séries temporais, como ARIMA. Esses modelos lidam bem com dados temporais, capturando tendências, sazonalidades e padrões de autocorrelação. O lado bom de usar ARIMA é que ele consegue explicar os dados, permitindo ajustes e uma melhor compreensão do comportamento das vendas ao longo do tempo.

Por fim, usaria técnicas de Machine Learning e Deep Learning, escolhendo o LightGBM ou um modelo LSTM (Deep Learning). O LightGBM é eficiente para grandes volumes de dados, capturando interações complexas entre variáveis, enquanto os LSTMs são úteis para sequências de dados, pois capturam dependências temporais de longo prazo. Não menos importante, conseguimos ganhar mais performance tunando nosso modelo e encontrando valores ótimos de hyperparametros.

**3- Qual metodologia de backtesting você usaria e por que essa metodologia escolhida?**

A metodologia de backtesting em projetos de previsão de séries temporais que eu escolheria seria a **Janela Deslizante (Rolling Window)**. Essa técnica é especialmente eficaz porque permite avaliar a performance do modelo em diferentes períodos do tempo, mantendo a ordem temporal dos dados. A abordagem da janela deslizante envolve treinar o modelo em um conjunto fixo de dados e, depois, testá-lo em outros período futuros.

**4- Qual/Quais métricas você avaliaria para sinalizar se geramos estoque ou falta de produtos?**

A previsão de demanda já é uma métrica importante para acompanhar via Dashboard, pois fornece uma base sólida para o gerenciamento de estoque. Além disso, compreender o **ciclo de vida do produto** é ótimo, pois isso nos ajuda a identificar quando um produto pode estar próximo de sua obsolescência e ajustar a estratégia de reabastecimento.

**A taxa de ruptura do estoque** também é, pois uma alta taxa pode sinalizar que estamos perdendo vendas e deixando os clientes sem produto. Por fim, o **lead time de reposição** nos ajudaria a saber quanto tempo levamos para reestabelecer o estoque.

Embora não seja especialista em logística, quero contribuir com métricas e insights que ajudem a otimizar nossa gestão de estoque.

**5- Qual/Quais métricas você avaliaria a assertividade da previsão de demanda?**

Para avaliar a assertividade da previsão de demanda, eu consideraria principalmente duas métricas: Mean Absolute Error (MAE) e Mean Absolute Percentage Error (MAPE).

O MAE calcula a média das diferenças absolutas entre as previsões e os valores reais, fornecendo uma indicação clara de quão próximo o modelo está de fazer previsões precisas. E o MAPE calcula o erro médio em termos percentuais, permitindo uma avaliação da precisão relativa do modelo em relação ao tamanho da demanda.

Além dessas, podemos medir pelo Mean Squared Error (MSE) que calcula a média dos erros quadráticos entre as previsões e os valores reais, penalizando erros maiores de forma mais severa do que o MAE pois eleva o resíduo ao quadrado. Com essas três métricas teremos uma avaliação robusta.