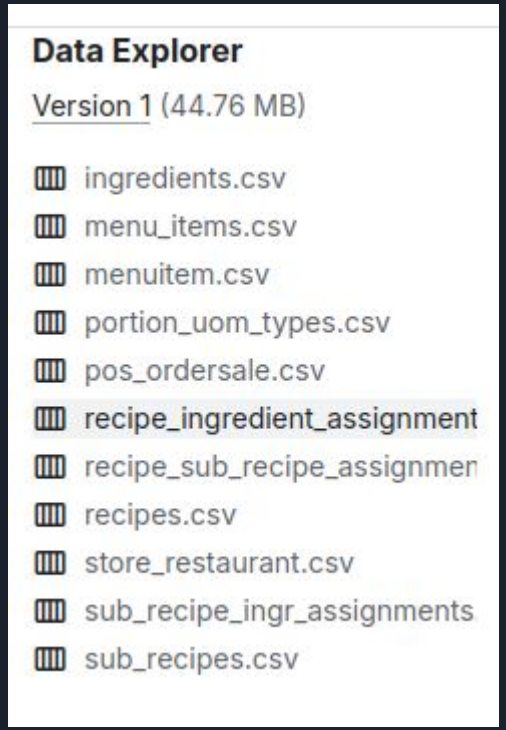
CASE:
Previsão de demanda de ingredientes em uma rede de fast-food

Dados foram retirados do site kaggle link



Nele temos dados gerados artificialmente de compras em quatro restaurantes de uma rede de fast food, além de algumas tabelas normalizadas com informações sobre os pratos, como o nome de cada prato que foi pedido, os ingredientes utilizados em cada prato juntamente com a quantidade de cada ingrediente e sua unidade de medida (gramas, ml, ...).

Além disso para cada compra temos a data em que foi realizada a compra.

Para mais informações sobre as relações entre as tabelas recomendo acessar o link do kaggle que apresenta uma descrição sobre cada tabela individualmente, pois temos muitas tabelas e ocuparia muito espaço nessa apresentação.

Link-https://www.kaggle.com/datasets/rishitsaraf/fast-food-restaurant-chain?select=recipe_ingredient_assignments.csv

Metodologia utilizada: Crisp-DM

1. Entendimento de negócios:

Não é novidade para ninguém que é muito importante fazer a previsão de demanda em qualquer negócio, no contexto de fast food é extremamente necessário para evitar desperdício e ao mesmo tempo sanar a demanda dos clientes, diminuindo os gastos e maximizando os lucros. Uma curiosidade é que segundo uma pesquisa da Mordor Intelligence em torno de 48% da população brasileira consome fast food pelo menos uma vez na semana[1] e é um dos maiores consumidores de fast food do mundo.

Temos dados das vendas de alguns restaurantes de uma rede de fast food com detalhes das compras realizadas nesses restaurantes, além de algumas informações sobre os ingredientes utilizados em cada receita, e seria de grande valor para a rede de fast food fazer a previsão da demanda de cada ingrediente para ter um melhor planejamento, além disso algumas perguntas sobre os dados são pertinentes para o time do fast food, as perguntas são:

- Qual o prato pedido com mais frequência?

- Qual o ingrediente utilizado em maior frequência?

- Quais o ingrediente utilizado em menor frequência?

- Existe diferença no número de vendas entre os restaurante? qual vende mais?

Também foi informado que atualmente eles utilizam a média móvel (com lag = 3) para prever a demanda de cada ingrediente, seria interessante se fosse possível ajustar outros modelos que tenham uma performance melhor.

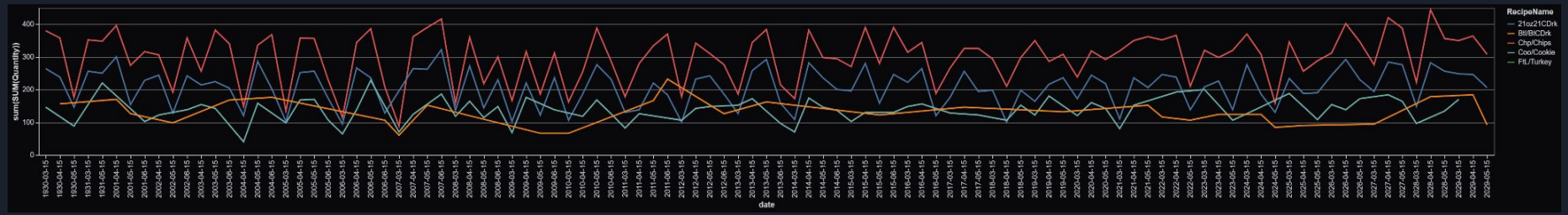
2. Entendimento dos dados

Temos dados que variam ao longo do tempo.
No link de onde foi retirado os dados no kaggle link temos descrição de cada uma das tabelas com informações sobre o significado de cada tabela e suas colunas.
Observando a colunas de datas temos datas entre 1930 e 2029, porém não temos datas entre 1931 e 2001, para a modelagem vamos desconsiderar dados antes de 2001 por estarem muito distantes da situação atual da rede de fast-food.
Detalhes sobre as queries utilizadas podem ser encontradas no notebook analise.ipynb com toda a análise de dados.
Para a contagem da quantidade é utilizado a quantidade de acordo com tabela recipe_ingredient_assignments.csv que tem a unidade de medida utilizada para cada ingrediente especificado na tabela portion_uom_types.csv de acordo com a chave PortionUOMTypeId

2. Entendimento dos dados (perguntas de negócios)

Nesta etapa foram utilizadas algumas queries em SQL (mais fácil de relacionar as tabelas) e funções do pandas (execução mais rápida em alguns casos) para retirar as informações para responder às perguntas de negócios.

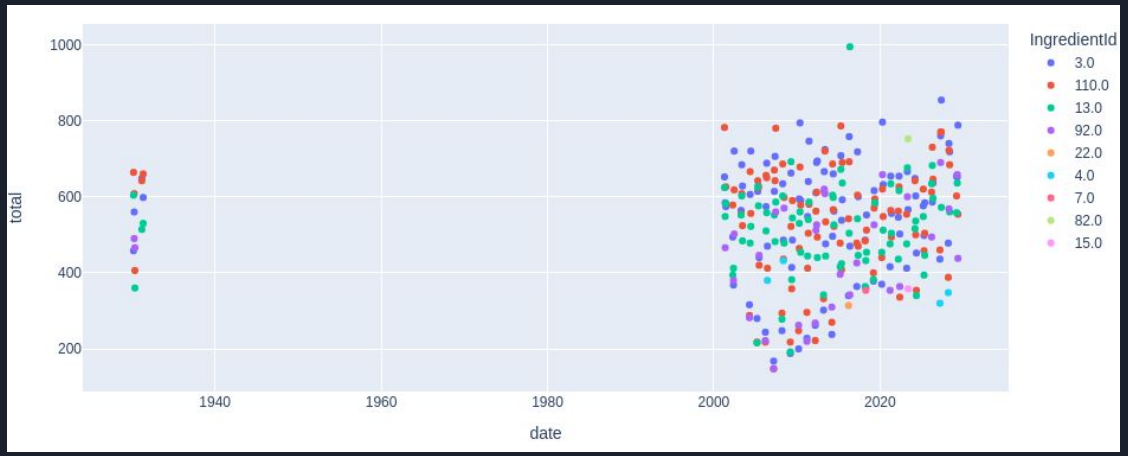
Para responder a primeira pergunta “ Qual o prato vendido com mais frequência? ” foi verificado nos dados que o prato mais vendido varia ao longo do tempo, então na verdade foi selecionado os três pratos vendidos com maior frequência pois de acordo com o intervalo de tempo selecionado o prato mais vendido mudava, na figura abaixo temos essa informação:



Podemos ver a linha em vermelho que representa o Chp/Chips (Chips Chips) é o item mais vendido na maioria das datas, logo em seguida 21oz21CDrk(21oz Carbonated Fountain 21Fnt) que permanece em segundo lugar mas em alguns anos briga com o Btl/BtCDrk(Bottled Carbonated Drink BtlDrk) e o Coo/Cookie(Cookie Cookie), além deles temos o item ftL/Turkey que não aparece no gráfico pois ele ficou entre os 3 primeiros colocados apenas uma vez.

2. Entendimento dos dados

Segunda pergunta de negócio : “Qual o ingrediente utilizado em maior frequência? ” O mesmo problema de ter um determinado ingrediente sendo utilizado com maior frequência em uma data do que outra acontece aqui, então foi selecionado os três ingredientes mais frequentes em cada data.



Contraintuitivamente neste caso é melhor visualizar com o scatter plot, pois temos algumas datas faltantes, e em alguns dias específicos os 3 ingredientes mais utilizados são diferentes, então se utilizarmos o line plot teriamos saltos grandes entres os pontos. Como não temos um número grande de ingredientes podemos ver na tabela de descrição dos ingredientes o que cada um significa. Podemos verificar o valor da frequencia de cada ingrediente no eixo y.

Ao contrário do que se esperava os ingredientes mais utilizados variam bastante ao longo do tempo, como tinhamos pouca variabilidade de receitas/pratos que são pedidos no restaurante faria sentido se esperar que os ingredientes não variassem tanto assim já que os ingredientes utilizados em determinado prato teoricamente seria o mesmo. Uma explicação possível para isso é que alguns ingredientes podem ser equivalentes e modificamos por razões de custo, ou que ao longo do tempo os ingredientes foram mudados para agradar o paladar dos clientes.

2. Entendimento dos dados

Continuando na segunda pergunta de negócio temos um tabela com a descrição dos ingredientes mais utilizados de acordo com o id:

IngredientId	IngredientName	IngredientShortDescription
3	Chicken Strips ...	Chicken Strips
110	Fountain Beverage syrup ...	Fountain Beverage syrup
13	Turkey ...	Turkey
92	Marinara Sauce ...	Marinara Sauce
22	Chips ...	Chips
4	Chicken. oven roasted patty ...	Chicken, single piece
7	Ham ...	Ham
82	Pastrami, beef ...	Pastrami, beef
15	Cookies ...	Cookies

2. Entendimento dos dados

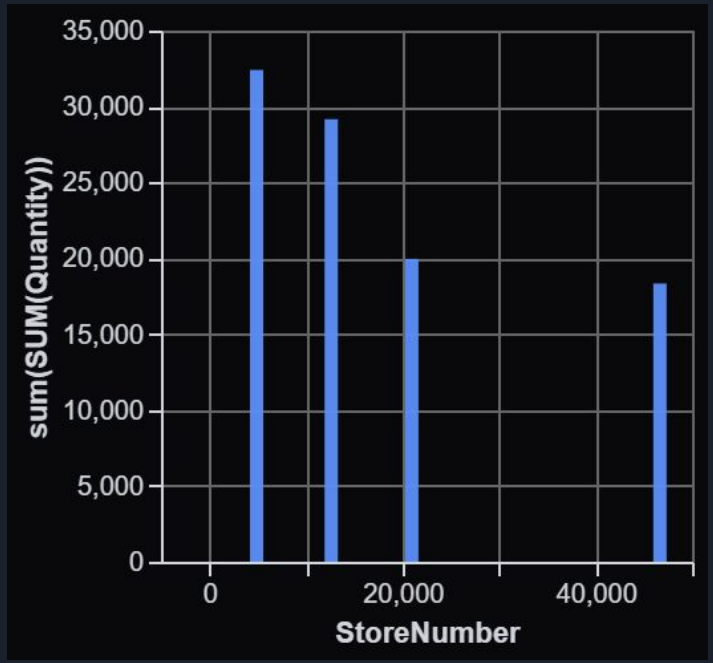
Terceira pergunta de negócio: “Qual o ingrediente utilizado em menor frequência?” temos o mesmo problema de termos uma variação em relação ao ingrediente menos utilizado de acordo com a data, neste caso variava mais do que no caso dos ingredientes mais utilizados então os dados são apresentados em forma de tabela com a descrição dos ingredientes.

	IngredientId	IngredientName	IngredientShortDescription
0	25	Green Peppers ...	Green Peppers
1	17	Olives ...	Olives
2	172	Coffee ...	Coffee, filter style gro
3	528	House Sandwich Sauce ...	House Sandwich Sauce
4	163	Beverage, Juice Box ...	Juice Box
5	44	Milk, bottle ...	Milk, bottle
6	274	Sweetener packets ...	Sweetener packets
7	30	Cheese, Shredded Monterey Cheddar ...	Cheese, Shredded Montere
8	160	Cup, Cold 30 oz. ...	Cup, Cold 30 oz.
9	68	Cup, Hot Cup, 16 oz. ...	Cup, Hot Cup, 16 oz
10	103	Pizza, pre-made ...	Pizza, Pre-made 8" chees "
11	26	Onions ...	Onions
12	21	Apples ...	Apple slices
13	43	Vinegar ...	Vinegar, Red Wine
14	122	Olive Oil Blend, Sauce ...	Olive Oil Blend, Sauce
15	181	Toy, Kids Meal Premium ...	Toy, Kids Meal Premium
16	225	SUBWAY Card ...	SUBWAY Card
17	81	Juice, Bottle ...	Juice, Bottle
18	265	Juice, Chilled Orange ...	Juice, Chilled Orange Ju
19	115	Cups, Hot Cup 12 oz. ...	Cups, Hot Cup 12 oz.

A frequência dos ingredientes variam entre zero a 4 unidades utilizadas de cada ingrediente, de acorco com sua unidade de medida na tabela

2. Entendimento dos dados

Quarta pergunda de negócios : Existe diferença no número de vendas entre os restaurante? qual vende mais?



Podemos perceber que temos uma diferença no número de vendas para cada loja.
O eixo X representa o Id de cada loja e o eixo y a quantidade de vendas para cada loja.

2 e 3. Entendimento e preparação dos dados

Como temos dados que variam ao longo do tempo e vamos ajustar modelos para prever a demanda de vários ingredientes é importante verificarmos algumas características sobre a frequência de cada ingrediente, pois assim podemos escolher modelos que se adequem melhor a nossa série temporal e fazermos experimentos para escolher o modelo mais adequado para nossos dados.

Vamos verificar se existe tendências, sazonalidade e estacionáriedade.

Antes foi aplicado algumas transformações nos dados para pegar a frequencia de cada ingrediente para cada loja, e deixamos os dados no formato de dicionário:

{'StoreNumber' : { 'ingrediente': { 'data' : [datas], 'frequencia': [] } } }

Para verificar autocorrelação, o que implica sazonalidade, nas séries temporais foi utilizado o teste de hipóteses Ljung-Box que tem como hipótese nula que o dado é independentemente distribuído(não possue autocorrelação) para um certo lag, um lag é uma unidade da unidade de tempo que estamos utilzando se tivermos ano por exemplo 1 lag seria 1 ano, com esse teste é possível verificar se existe autocorrelação, mas não o tipo da autocorrelação. Caso existir autocorrelação para verificar se a autocorrelação é linear ou não podemos plotar o gráfico ACF que se baseia na correlação linear de Pearson.

2 e 3. Entendimento e preparação dos dados

Resultados dos testes de autocorrelação:

total de ingredientes que rejeitaram a hipótese de autocorrelação 62

total de ingredientes 276

A maior parte dos ingredientes não possuem autocorrelação em sua série temporal com lag 1.

2 e 3. Entendimento e preparação dos dados

Para verificar estacionariedade (se a série possui média e variância constantes) foi utilizado o teste (ADF) Augmented Dickey-Fuller que assume na hipótese nula é que a série não é estacionária, quando rejeitamos a hipótese nula temos que a série é estacionária.

Temos como resultado:

total de ingredientes que rejeitaram a hipótese de estacionariedade 228

total de ingredientes 276

A maioria dos ingredientes não possuem série temporal estacionária, ou seja, não possuem média e variância constante.

2 e 3. Entendimento e preparação dos dados

Teste Mann-Kendall verifica se existe uma tendência positiva ou negativa nos dados.

Hipótese nula : Não há tendência nos dados

Resultado:

total de ingredientes que rejeitaram a hipótese de não há nenhuma tendencia na serie temporal 3

total de ingredientes 276

Ou seja a maioria das séries temporais possuem um tendência, positiva ou negativa, e além disso a maioria não possui comportamento estacionário (o que é coerente com o teste de que há tendência) e a maioria deles não apresenta autocorrelação. Essas informações são cruciais para ajustar os modelos de series temporais corretamente.

4. Modelagem

Considerando essas características dos dados podemos ajustar os modelos:

- Exponential Smoothing Models que pode ser ajustado em dados que possuem tendências, sazonalidade e pode ser ajustado mesmo se a series não for estacionária.

- State Space Models que pode ser ajustado em dados não estacionários e com tendências.

- Além desses modelos de séries temporais podemos utilizar modelos de aprendizado de máquina como o random forest e xgboost.

5. Validação dos modelos

Para validar os modelos foi utilizado as métricas

Mean Square Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Square Error (RMSE) é igual a raiz quadrada do MSE

Além disso adicionei uma medida que é igual a contagem das vezes em que o modelo subestimou a quantidade de ingredientes, ou seja, previu a quantidade menor do que ela realmente é.

5. Validaçao dos modelos

Foi utilizado um processo parecido com o k-fold, mas ao invés de aleatoriamente criar os folds vamos criar o treino de tal forma que no treino tenhamos apenas observações passadas em relação ao conjunto de teste e no conjunto de teste vamos ter apenas observações no futuro em relação ao conjunto de treino.

Além disso vamos selecionar apenas observações dos anos dois mil para frente, pois temos poucos dados antes dos anos dois mil e isso pode atrapalhar o nosso modelos pois são dados muito antigos e distantes do atual, além de serem em menor quantidade.

5. Validação dos modelos

Ilustração do k-fold time series split, com k = 2.

5. Validação dos modelos

Resumo:

Foi separado a frequência de cada ingrediente agrupado de acordo com cada loja e tivemos dados nesse formato: {'StoreNumber' : { 'ingrediente': { 'data' : [datas], 'frequencia': [frequencias] } } }

Então para cada ingrediente em cada loja foi feita a divisão de treino e teste time series k-fold com k = 2.

Então foi ajustado os modelos Exponential Smoothing Model, State Space Model, random forest e XGBoost.

Calculado as métrica SME, RSME e a contagem de estimativas abaixo do valor real.

Então foi escolhido o modelo com menor RSME, como sabemos que antes a rede estava utilizando o modelo média móvel ele também foi ajustado e foi comparada a performance do modelo escolhido com menor RSME e o modelo média móvel com lag = 3 utilizado anteriormente.

5. Validação dos modelos

Resultados da comparação da performance do modelo gerado no experimento com menor RSME e o RSME do modelo média móvel com lag = 3 utilizado anteriormente, considerando todos os ingredientes de todas as lojas:

Média do RSME dos modelos de média móvel: 16

Média do número de predições abaixo do valor real para a média móvel: 32

Média do RSME dos modelos com menor RMSE: 14

Média do número de predições abaixo da média para o modelo com RSME mínimo: 35

média das diferenças entre o rsme da média móvel e dos modelos com menor RSME: 1

Desvio padrão das diferenças entre o rsme da média móvel e dos modelos com menor rsme: 1

5. Validação dos modelos

Por mais que tenhamos ingredientes que possuem comportamentos e demandas diferentes ao tirar uma média entre o RSME dos modelos com menor RSME e o RSME do modelo média móvel (com lag = 3), juntamente com o desvio padrão dessas diferenças para todos os modelos que é igual a 1, podemos ter uma noção de como está performando os modelos de nosso experimento, e de acordo com os resultados temos uma melhora entre os valores preditos com os modelos que escolhemos em nosso experimento em relação com a média móvel, na média temos uma redução de 12,5% nos valores do RSME, o que indica que as previsões feitas pelos modelos do nosso experimento erram menos. Além disso com os modelos de nosso experimentos tendemos a errar um pouco mais de tal forma que subestimamos a demanda do ingrediente, ou seja, estimamos uma quantidade um pouco menor do que o real valor, não temos uma diferença muito grande em relação a essa métrica quando comparamos com a mesma métrica no caso da média móvel, tendo um aumento de 9% nessa métrica com os modelos do nosso experimento, neste caso teria de ser avaliado o real custo de errar a estimativa para mais ou menos e então escolher o modelo de maneira a minimizar os custos em um caso real teria de ser avaliado individualmente esse custo para cada ingrediente.

6. Deployment

Todos os experimentos e modelos foram ajustados e salvos utilizando o MLFlow, o MLFlow é uma ferramenta para fazer o gerenciamento de modelos.

No MLFlow temos uma funcionalidade muito interessante onde é gerada uma API para acessar os modelos, verificar suas métricas de validação, fazer predições com os modelos ajustados verificar dados sobre os experimentos que foram realizados entre outras coisas.

Então podemos disponibilizar o modelos para uso utilizando a API de modelos do MLFlow, poderíamos treinar os modelos em um container docker e disponibilizar eles em nuvem por exemplo e de tempos em tempos fazer calibragem dos modelos.