Universidade Federal Fluminense
Instituto de Ciência e Tecnologia
Departamento de Engenharia
Curso de Graduação em Engenharia de Produção

Jean Nery dos Santos Oliveira Silva
Pedro Costa Ceciliano

ANÁLISE PREDITIVA DA SEGROB NOTLAD

RIO DAS OSTRAS - RJ

2025

SUMÁRIO

1. Introdução....4
2. Fundamentação Teórica....5
3. Metodologia....8
4. Estudo de Caso....12
2.1 Entendimento do Negócio....12
2.1.1 Objetivo Geral do Negócio....12
2.1.2 Contexto....12
2.1.3 Definindo o problema em uma pergunta....12
2.1.4 5W2H....14
2.2 Entendimento dos Dados....14
2.3 Preparação dos Dados....19
2.4 Modelagem dos Dados....20
2.4.1 Modelo Naive....21
2.4.2 Modelo Cumulativo....23
2.4.3 Modelo de Média Móvel....24
2.4.4 Modelo de Suavização Exponencial....26
2.4.5 Regressão Linear Simples e Regressão Linear Dinâmica....28
2.4.6 KNN (K Nearest Neighbors)....30
2.4.7 SVM/SVR (Support Vector Machines/Support Vector Regression)....32
5. Métricas utilizadas....34
3.1 MAPE....34
3.2 RMSE....34
3.3 MAD....34
6. Validação....35
4.1 Metodologia de Validação: Validação Cruzada para Séries Temporais....35
4.2 Outra Metodologia de Validação: Sliding Window Cross-Validation....36
4.3 Qual método devemos utilizar ?....38
4.4 Análise Comparativa dos Resultados (Feed Forward)....39
5. Escolha do Modelo Final e Conclusões....45
6. Referências....47

1. Introdução

A Segrob Notlad, uma marca de fast fashion do Brasil, está embarcando em uma jornada de transformação digital que visa integrar intensivamente inteligência artificial e análise preditiva para aprimorar sua eficiência operacional e a assertividade nas tomadas de decisão de negócio. Dentro desse contexto, a empresa enfrenta o desafio crucial de prever o volume diário de vendas de camisetas para dezembro de 2024, utilizando o histórico de vendas de janeiro de 2022 a novembro de 2024. A acurácia dessa previsão é vital para otimizar o planejamento em um período de alta demanda, minimizando riscos como a falta ou o excesso de estoque, que podem resultar em perdas de receita ou custos desnecessários.

Para abordar este desafio, o projeto adota a metodologia estruturada e cíclica do Cross Industry Standard Process for Data Mining (CRISP-DM), que organiza o processo em seis fases principais: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem dos Dados, Avaliação e Implantação. Esta abordagem garante a organização, rastreabilidade e eficiência na execução do projeto de ciência de dados. A análise exploratória dos dados revelou uma complexa dinâmica de vendas, marcada por uma clara tendência de crescimento ao longo dos anos, sazonalidades anuais e semanais bem definidas, e a presença de outliers associados a eventos comerciais e datas comemorativas. Tais características ressaltam a necessidade de um modelo preditivo capaz de capturar essas nuances para gerar previsões precisas.

Este trabalho está estruturado para apresentar de forma clara e concisa cada etapa da análise. Inicialmente, será detalhada a metodologia CRISP-DM, delineando como suas fases foram aplicadas ao problema da Segrob Notlad. Em seguida, será abordado o entendimento do negócio e dos dados, incluindo a definição do problema, os objetivos e o contexto da Segrob Notlad, além da exploração visual dos dados históricos de vendas. Posteriormente, a seção de preparação dos dados descreverá as etapas de limpeza e organização para a modelagem. A fase de modelagem apresentará os diversos modelos preditivos testados, como Naive, Cumulativo, Média Móvel, Suavização Exponencial, Regressão Linear (Simples e Dinâmica), KNN e SVR, e suas respectivas previsões para o mês de novembro de 2024. A validação dos modelos será realizada por meio de validação cruzada para séries temporais e análise comparativa das métricas de erro (MAD, RMSE e MAPE). Finalmente, serão apresentadas a escolha do modelo final e as conclusões do estudo, com as previsões de vendas para dezembro de 2024 e recomendações para futuras aplicações.

2. Fundamentação Teórica

A compreensão dos modelos de previsão e das métricas de avaliação é essencial para a eficácia de um projeto de análise preditiva. Este trabalho emprega e avalia diversos modelos de séries temporais e de aprendizado de máquina, utilizando métricas cruciais para a validação de suas previsões.

- **Modelo Naive**: Considerado o modelo de previsão mais simples para séries temporais, baseia-se na premissa de que o valor da próxima observação será idêntico ao valor mais recente. Matematicamente, é descrito como $x_t = x_{t-1} + e_t$, onde e_t representa um erro aleatório independente e identicamente distribuído (iid) com média zero ($\mu=0$) e variância constante ($\sigma^2 = V[e]$). A previsão para o período $t+1$ é, portanto, simplesmente o valor atual x_t.
- **Modelo Cumulativo**: Em contraste com o modelo Naive, o modelo Cumulativo acumula a soma dos valores passados, o que auxilia na identificação de tendências de crescimento ou queda ao suavizar flutuações pontuais. Este modelo valoriza mais o histórico da demanda, assumindo que o valor da série temporal no tempo t é uma constante a acrescida de um erro aleatório et. O erro et é iid ($\mu=0, \sigma^2 = V[e]$), garantindo que um erro não influencie o outro e que a dispersão dos erros seja estável. A previsão para o próximo valor ($x\wedge t,t+1$) é calculada como a média aritmética de todos os valores observados até o tempo t.
- **Modelo de Média Móvel**: Este modelo generaliza os modelos Cumulativo e Naive, oferecendo abordagens intermediárias. Ao invés de considerar todas as observações passadas, a Média Móvel calcula a média das últimas M observações. A previsão para o próximo valor ($x\wedge t,t+1$) é a média dos últimos M valores da série. Este modelo é particularmente útil para filtrar ruídos e capturar o comportamento recente da série em cenários sem tendências fortes. A escolha do valor de M é crítica: um M pequeno pode responder rapidamente a ruídos, enquanto um M grande pode obscurecer mudanças importantes de curto prazo.
- **Suavização Exponencial Simples**: Esta técnica de previsão atribui maior peso aos dados mais recentes e uma importância decrescente exponencialmente aos dados mais antigos. Ela pressupõe uma demanda estacionária, sem tendência ou sazonalidade, focando apenas no nível da demanda. A suavização exponencial utiliza uma constante de suavização α, que varia entre 0 e 1, sendo na prática comumente entre 0.1 e 0.3. A previsão para o próximo valor ($x\wedge t+1,t$) é uma combinação ponderada do valor atual

xt e da previsão passada $x\hat{}t-1,t$, sendo expressa pela fórmula $x\hat{}t+1,t=\alpha xt+(1-\alpha)x\hat{}t-1,t$. Um α próximo de 1 indica que a previsão é mais reativa ao valor atual, enquanto um α próximo de 0 resulta em uma previsão mais suave, dando mais peso à previsão passada.

- **Suavização Exponencial Dupla**: Conhecida como Método de Holt, por exemplo, é uma técnica projetada para séries que apresentam uma tendência. Ela estende a suavização simples ao introduzir uma segunda equação para modelar explicitamente a inclinação da série, mantendo e atualizando duas componentes a cada observação: o nível e a tendência. A previsão futura é, então, uma projeção do último nível acrescido da tendência, tornando o método ideal para dados com tendência, mas sem um padrão sazonal claro.
- **Suavização Exponencial Tripla**: Suavização Exponencial Tripla, ou Método de Holt-Winters. Esta abordagem, uma das mais eficazes na previsão estatística, adiciona uma terceira componente para modelar os padrões sazonais, sendo especialmente útil em suas variações aditiva, para sazonalidade com magnitude constante, e multiplicativa, para quando a sazonalidade varia em proporção ao nível da série.
- **Regressão Linear Simples**: É uma técnica estatística amplamente utilizada para modelar a relação entre duas variáveis quantitativas. Busca estimar o valor de uma variável dependente (y) a partir de uma variável independente (x), assumindo que essa relação pode ser representada por uma equação linear. O modelo é matematicamente expresso como $y=\beta0+\beta1x+\epsilon$, onde y é a variável dependente, x é a variável independente, β0 é o intercepto, β1 é o coeficiente angular (indicando a inclinação e direção da relação), e ϵ é o termo de erro aleatório. O principal objetivo é encontrar os valores de β0 e β1 que minimizem a soma dos quadrados dos resíduos, processo conhecido como método dos mínimos quadrados ordinários (OLS). A aplicação da regressão linear simples permite previsões e avaliação do grau de associação entre as variáveis, mas requer que suposições como linearidade, homocedasticidade dos resíduos, normalidade dos erros e independência das observações sejam atendidas para que os resultados sejam confiáveis.
- **Regressão Linear Dinâmica**: Desenvolvida para superar as limitações do modelo simples, esta abordagem enriquece o modelo de regressão com variáveis que representam a dinâmica temporal intrínseca à série de vendas, notadamente os "lags". O processo envolveu a engenharia e seleção de *features*, testando diversas variáveis baseadas em *lags* de vendas. Os *lags* selecionados para compor a versão final do

modelo de Regressão Linear Dinâmica foram: *Lag* de 1 dia (Vendas t-1), fundamental para capturar a correlação de curto prazo; *Lag* de 7 dias (Vendas t-7), impactante para modelar a sazonalidade semanal; e Média Móvel dos Últimos 7 Dias, que ajusta o modelo à tendência local, oferecendo uma visão mais estável. Assim, o modelo prevê as vendas do dia atual com base na tendência geral, no valor de vendas do dia anterior, no valor de vendas do mesmo dia na semana anterior e na média de vendas da última semana.

- **KNN (K-Nearest Neighbors) para Regressão**: O KNN é um algoritmo de aprendizado supervisionado que pode ser aplicado tanto para classificação quanto para regressão. Neste projeto, foi utilizado para prever um valor contínuo (volume de vendas). O modelo realiza a previsão calculando a distância (geralmente euclidiana) entre um novo ponto de dados e os demais pontos existentes no conjunto de dados, identificando os *k* vizinhos mais próximos. A estimativa do valor é então feita a partir da média dos valores desses *k* vizinhos mais próximos. Embora ofereça flexibilidade por ser não-paramétrico e intuitivo em sua lógica , sua principal desvantagem em séries com forte tendência de crescimento é a incapacidade de extrapolar valores, resultando em uma subestimação consistente das vendas futuras. Além disso, sua performance é sensível à engenharia de *features*, podendo ser prejudicada pela inclusão de variáveis irrelevantes.
- **SVM/SVR (Support Vector Machines/Support Vector Regression)**: SVM é um algoritmo de aprendizado de máquina supervisionado para tarefas de classificação e regressão. Na regressão, é conhecido como Support Vector Regression (SVR) e busca ajustar um hiperplano aos dados com uma margem de erro aceitável. Para dados não linearmente separáveis, o SVR emprega o "truque do kernel" (*kernel trick*), transformando os dados em um espaço de dimensão superior onde uma fronteira linear é mais facilmente encontrada. Suas vantagens incluem robustez a *outliers*, o que é benéfico para dados de vendas com picos e valores atípicos, e a capacidade de modelar relações não-lineares complexas através de *kernels* (como o RBF). Contudo, uma desvantagem é a complexidade e o custo computacional na otimização de seus hiperparâmetros, o que pode impactar a performance se não for perfeitamente ajustado ao problema.
- **Random Forest (Árvore de Decisão)**: é um algoritmo de *machine learning* do tipo *ensemble*. Ele opera construindo uma vasta coleção de árvores de decisão independentes e agregando suas previsões para obter um resultado final mais robusto.

Para ser aplicado a problemas de previsão, o algoritmo requer que a série temporal seja transformada em um formato de aprendizado supervisionado, o que é feito através da criação de variáveis (features) como lags de vendas, indicadores de dia da semana, mês, feriados e médias móveis. A previsão final é a média dos resultados de todas as árvores. Embora seja reconhecido pela alta acurácia e pela capacidade de capturar interações complexas e não-lineares, o Random Forest possui como desvantagem notável a dificuldade em extrapolar tendências, pois não consegue prever valores fora do intervalo observado nos dados de treinamento.

- **MAPE (Mean Absolute Percentage Error)**: Calcula a média dos erros absolutos expressos em termos percentuais em relação aos valores reais. Um MAPE menor indica maior precisão do modelo.
- **RMSE (Root Mean Squared Error)**: Avalia a magnitude dos erros, elevando ao quadrado as diferenças entre os valores reais e previstos.
- **MAD (Mean Absolute Deviation)**: Mede a dispersão dos erros, calculando a média das diferenças absolutas entre os valores reais e os valores previstos.
- **Validação Cruzada Sequencial (Forward-Chaining)**: No *Forward-Chaining*, o histórico de dados (2022-2024) é dividido em múltiplos "folds" sequenciais, onde cada *fold* consiste em um período de treino e um período de teste subsequente. O modelo é treinado e avaliado iterativamente; por exemplo, treina-se com os dados dos primeiros 12 meses e testa-se a previsão para o 13º mês, e assim sucessivamente até o final da série. As métricas de erro (MAE, RMSE e MAD) são calculadas em cada iteração, e o desempenho final do modelo é a média desses erros. Essa metodologia garante que as previsões sejam sempre feitas para um período futuro com base apenas em dados passados, simulando exatamente como o modelo será usado na prática e fornecendo uma medida de erro muito mais confiável.
-

3. Metodologia

A metodologia adotada para esse projeto segue os fundamentos do Cross Industry Standard Process for Data Mining (CRISP-DM). O CRISP-DM é uma abordagem estruturada e cíclica composta por seis fases principais, que garantem a organização, rastreabilidade e eficiência na execução de projetos de ciência de dados.

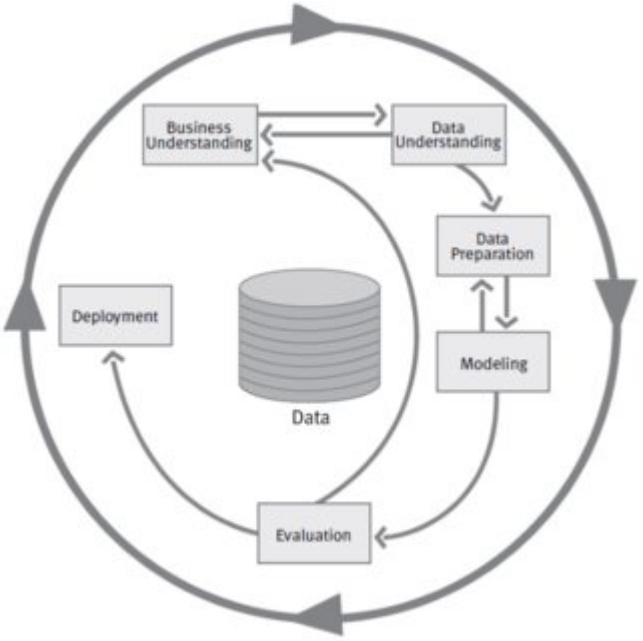


Figura 1: Fases do Modelo CRISP-DM

A primeira fase é o **Entendimento do Negócio**. Nesta fase começa com uma compreensão profunda das necessidades do cliente, incluindo seus objetivos e requisitos do projeto. Essas são algumas tarefas a serem seguidas para essa primeira fase.

- Determinar os objetivos do negócio: Entender completamente de uma uma visão empresarial, o que o cliente deseja realizar e em seguida definir os critérios de sucesso do negócio.
- Avaliar a situação: Determinar a disponibilidade de recursos e requisitos do projeto, avaliar os riscos e contingências e conduzir uma análise de custo-benefício.
- Determinar metas de mineração de dados: Além de definir os objetivos de negócio, deve-se também definir o que significa sucesso em uma visão técnica de mineração de dados
- Produzir plano de projeto: Selecionar tecnologias e ferramentas e definir planos detalhados para cada fase do projeto.

Por ser a primeira fase do projeto, estabelecer uma forte compreensão do negócio vai servir como fundamentação, ou seja, absolutamente essencial para o andamento do projeto.

A próxima fase é o **Entendimento dos Dados**. Somando com a primeira fase de *Entendimento do Negócio*, essa segunda fase vai direcionar o foco para identificar, coletar e

analisar os conjuntos de dados que podem ajudar a atingir o objetivo do projeto. Suas tarefas são:

- Coletar dados iniciais: Adquirir os dados necessários e (se necessário) inseri-los na ferramenta de análise.
- Descrever os dados: Examinar os dados e documentar suas prioridades de superfície, como formato dos dados, números de registros ou identidades de campo.
- Explorar os dados: Se aprofundar nos dados. Consultar, visualizar e identificar as correlações entre eles.
- Verificar a quantidade dos dados: Qual nível de veracidade dos dados? Documentar qualquer problema de qualidade dos mesmos.

A terceira fase é a **Preparação de Dados**. Normalmente essa fase toma cerca de 70%, ou até mesmo 90% do tempo do projeto. É chamada por alguns de "manipulação de dados", ela serve para preparar os conjuntos de dados finais para modelagem. Ela consiste em cinco tarefas:

- Selecionar dados: Determinar quais conjuntos de dados serão usados e documentar os motivos para inclusão/exclusão.
- Limpar dados: É uma das tarefas mais demoradas. Ela tem o objetivo de corrigir, imputar ou remover valores incorretos.
- Construir dados: Derivar novos atributos que serão úteis. Por exemplo, derivar o índice de massa corporal de alguém a partir dos campos de altura e peso.
- Integrar dados: Criar novos conjuntos de dados combinando dados de várias fontes.
- Formatar dados: Reformatar os dados conforme o necessário. Exemplo: converter valores de string que armazena números em valores numéricos para poder realizar operações matemáticas.

Depois da *Preparação de Dados*, seguimos para a **Modelagem.** Nessa fase diversos modelos serão construídos e avaliados com base em técnicas de modelagem. Esta fase tem quatro tarefas:

- Selecionar técnicas de modelagem: Determinar quais algoritmos utilizar. (por exemplo: regressão).
- Gerar design de Teste: Dependendo da modelagem, talvez seja necessário dividir os dados em conjuntos de treinamento, teste de validação.
- Construir modelo: Construir um código de programação.
- Avaliar o modelo: Interpretar os resultados do modelo com base no conhecimento do domínio, nos critérios de sucesso predefinidos e no design do teste.

O Crisp-DM sugere iterar a construção e a avaliação dos modelos até que se obtenha os melhores modelos. Porém, na prática as equipes continuam iterando até encontrar um modelo "bom o suficiente", prosseguir pelo ciclo de vida do CRISP-DM e, então, melhorar ainda mais os modelos em iterações futuras.

A fase de **Avaliação** analisa de forma mais ampla qual modelo melhor atende aos requisitos do negócio e o que fazer em seguida. Esta fase tem três tarefas:

- Avaliar os resultados: Os modelos atendem aos critérios de sucesso do negócio? Quais devemos aprovar para o negócio?
- Processo de revisão: Revisar todo o trabalho realizado. Algo foi esquecido? Todas as etapas foram executadas corretamente? Resumir as descobertas e corrigir o que for necessário.
- Determinar as próximas etapas: Determinar se deseja prosseguir com a implantação, iterar mais, ou iniciar novos projetos, com base nas duas tarefas anteriores.

"Dependendo dos requisitos, a fase de **Implantação** pode ser tão simples quanto gerar um relatório ou tão complexa quanto implementar um processo de mineração de dados repetível em toda a empresa." - Guia CRISP-DM. Um modelo não é particularmente útil a menos que o cliente possa acessar seus resultados.

- Planejar a implantação: Desenvolver e documentar um plano paraa implantação do modelo.

- Planejar monitoramento e manutenção: Desenvolver um plano para o monitoramento e manutenção para evitar problemas durante a fase operacional (ou fase pós-projeto) de um modelo.
- Produzir relatório final: A equipe do projeto documenta um resumo do projeto, que pode incluir uma apresentação final dos resultados da mineração de dados.
- Revisar projeto: Realizar uma retrospectiva do projeto sobre o que deu certo, o que poderia ter sido melhor e como melhorar no futuro.

O Trabalho talvez não acabe depois dessas seis tarefas. Como estrutura de projeto, o CRISP-DM não define o que fazer após o projeto (também conhecido como operações).

4. Estudo de Caso

2.1 Entendimento do Negócio

2.1.1 Objetivo Geral do Negócio

A Segrob Notlad, uma das maiores marcas brasileiras de fast fashion, está passando por uma fase estratégica de transformação digital. Um dos pilares dessa nova fase é o uso intensivo de inteligência artificial e análise preditiva para melhorar a eficiência operacional e a assertividade nas decisões de negócio. A Segrob Notlad precisa prever o volume diário de vendas de camisetas no mês de dezembro de 2024, com base no histórico de vendas de janeiro de 2022 a novembro de 2024. Isso permitirá à empresas planejar melhor para um período de alta demanda, evitando falta de estoque ou excesso de produtos.

2.1.2 Contexto

A Segrob Notlad é uma marca brasileira de *fast fashion* com uma identidade cosmopolita, fundada no Rio de Janeiro. A empresa se consolidou no mercado por meio de designs versáteis, preços acessíveis e campanhas de marketing arrojadas, direcionadas ao público jovem e urbano. Atualmente, possui mais de 80 lojas no Brasil, com sua base de operações no Rio de Janeiro, além de uma presença emergente na América do Sul e em três lojas conceito na Europa. A marca busca combinar a diversidade brasileira com uma estética minimalista do leste europeu, refletindo a origem de seu fundador. A camiseta básica é um

item-chave, e falhas na previsão de demanda podem gerar rupturas de estoque ou sobras onerosas.

2.1.3 Definindo o problema em uma pergunta

O objetivo central deste projeto é desenvolver um modelo preditivo para responder a uma pergunta específica: **"qual será o volume de camisetas básicas vendidas em cada dia de dezembro de 2024?"**. Para isso, será utilizado o histórico completo de vendas diárias da empresa, compreendido entre janeiro de 2022 e novembro de 2024. Este projeto servirá como uma ferramenta de inteligência para a tomada de decisão estratégica na gestão de suprimentos,

Objetivos Específicos do Projeto de Mineração de Dados:

- Desenvolver um modelo preditivo capaz de estimar com precisão as vendas diárias de camisetas básicas em dezembro de 2024.
- Utilizar dados históricos de vendas (jan/2022 a nov/2024) como base de análise.
- Aumentar a agilidade e a assertividade na tomada de decisão da área de abastecimento e cadeia de suprimentos.
- Servir como um piloto para o uso sistemático de IA na previsão de demanda.

Critérios de Sucesso do Projeto:

- Reduzir Incertezas: Diminuir as dúvidas sobre o volume de vendas futuro, permitindo um planejamento mais assertivo.
- Gerar Vantagem Competitiva: Utilizar a análise preditiva como um pilar da gestão moderna para se destacar no mercado.
- Preparar para o Futuro: Criar uma solução que sirva como preparação real para os desafios de um ambiente corporativo dinâmico, que lida com incerteza e sazonalidade.

Restrições e Riscos:

- Mudanças no escopo do desafio ao longo do tempo (novas variáveis, mudanças na estratégia).
- Qualidade e consistência dos dados históricos.
- Sazonalidade e eventos promocionais (Black Friday, Natal) que podem distorcer padrões históricos.
- Alinhamento entre as equipes de dados e as áreas de negócio.

Recursos e Stakeholders

- Time de análise
- Base de dados de vendas fornecida pela empresa.
- Equipes internas da Segrob Notlad nas áreas de suprimentos, marketing e estratégia digital.

2.1.4 5W2H

Elemento	Resposta
What (O que?)	Prever a demanda diária de camisetas básicas.
Why (Por quê?)	Para otimizar o abastecimento, evitar perdas e melhorar decisões estratégicas.
Who (Quem?)	Time de análise
Where (Onde?)	Todas as lojas que comercializam a camiseta básica, com foco no Brasil.
When (Quando?)	Para o mês de **dezembro de 2024**, com base em dados de jan/2022 a nov/2024.
How (Como?)	Através de modelagem preditiva utilizando técnicas de ciência de dados.
How much (Quanto?)	O valor estimado da demanda diária em unidades por dia.

Tabela 1 : 5W2H

Fonte: Elaboração própria.

2.2 Entendimento dos Dados

A base de dados fornecida pela empresa apresenta o volume de vendas diários para camisetas básicas masculinas entre o período de janeiro de 2022 e novembro de 2024. Com isso, é possível analisar o comportamento desse volume de vendas ao longo de cada mês, como mostra o gráfico abaixo.

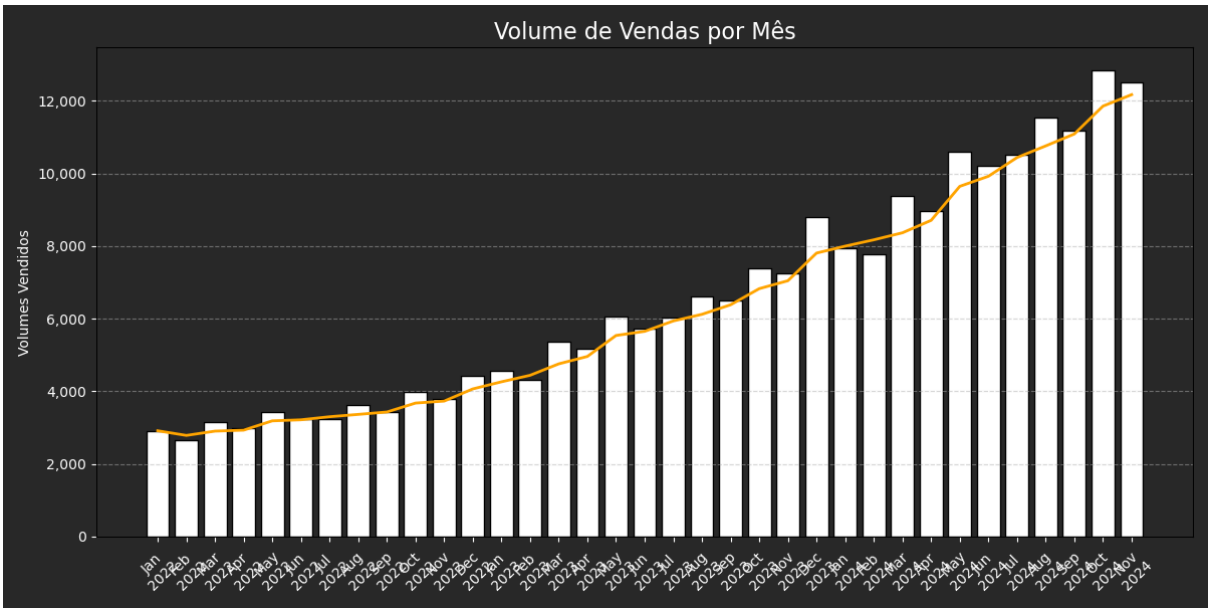


Figura 2: Transformação dos dados fornecidos em gráfico.

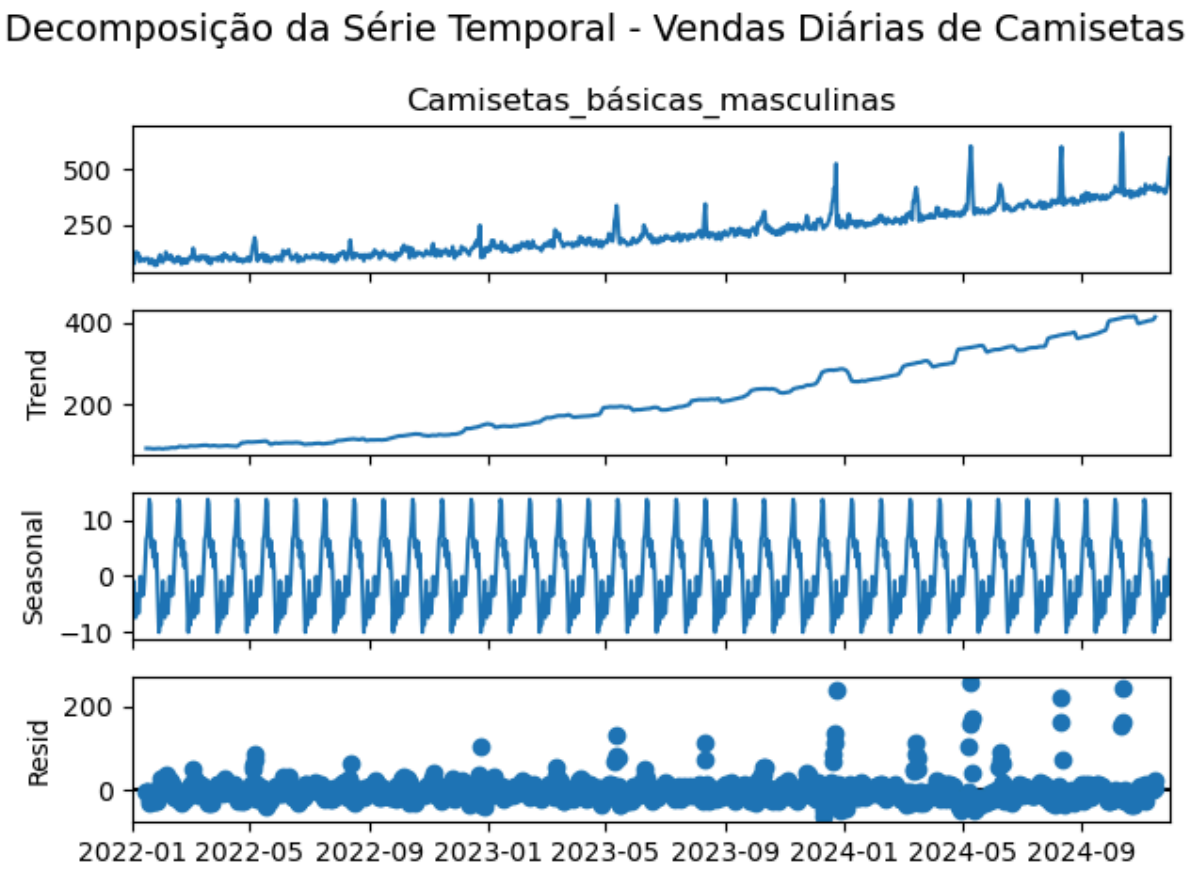


Figura 3: Gráficos para análise de tendência, sazonalidade e nível.

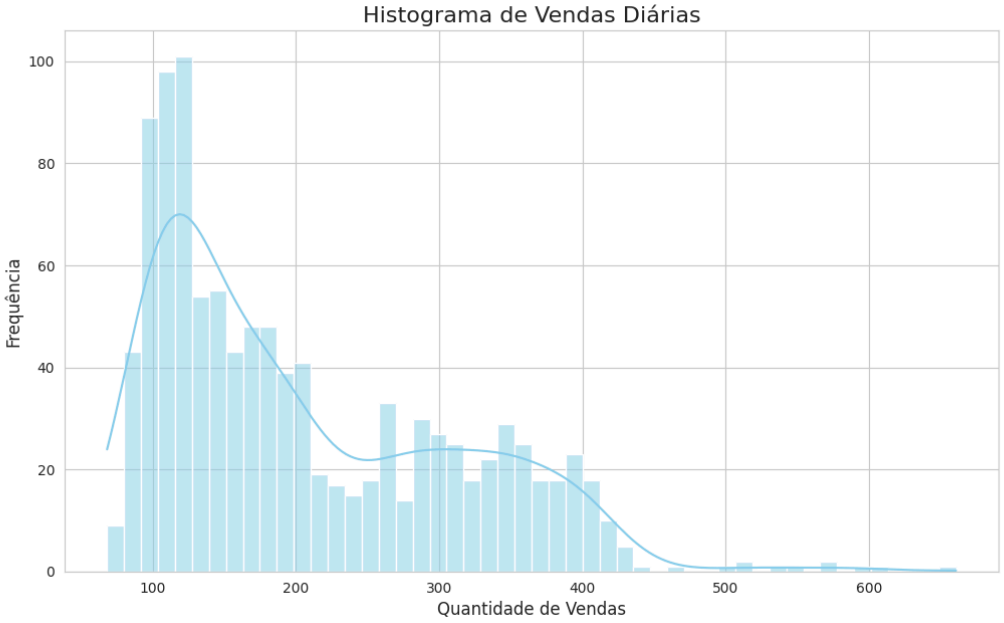


Figura 4: Gráfico Histograma das vendas.

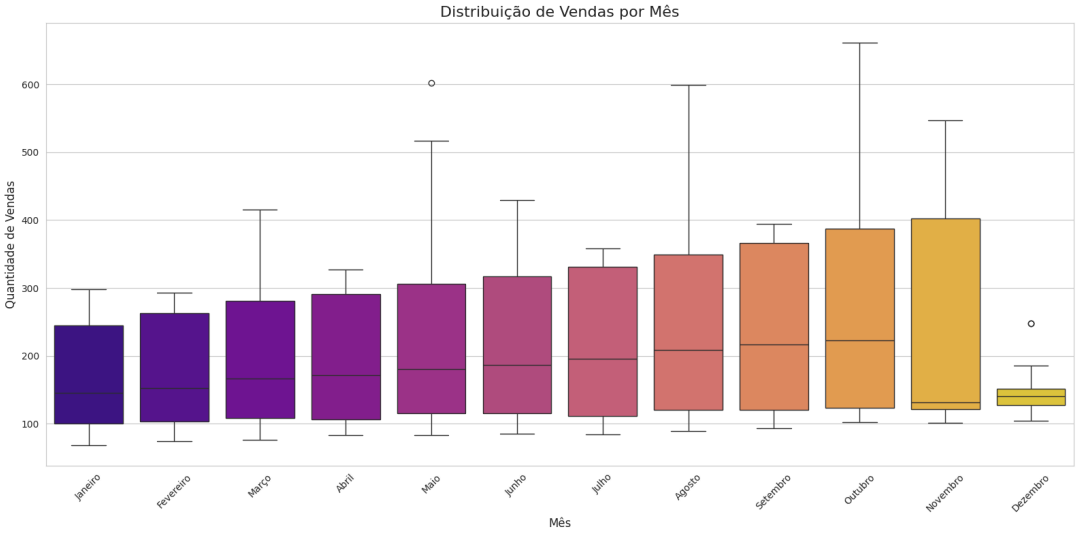


Figura 5: Gráfico Boxplot das vendas por mês

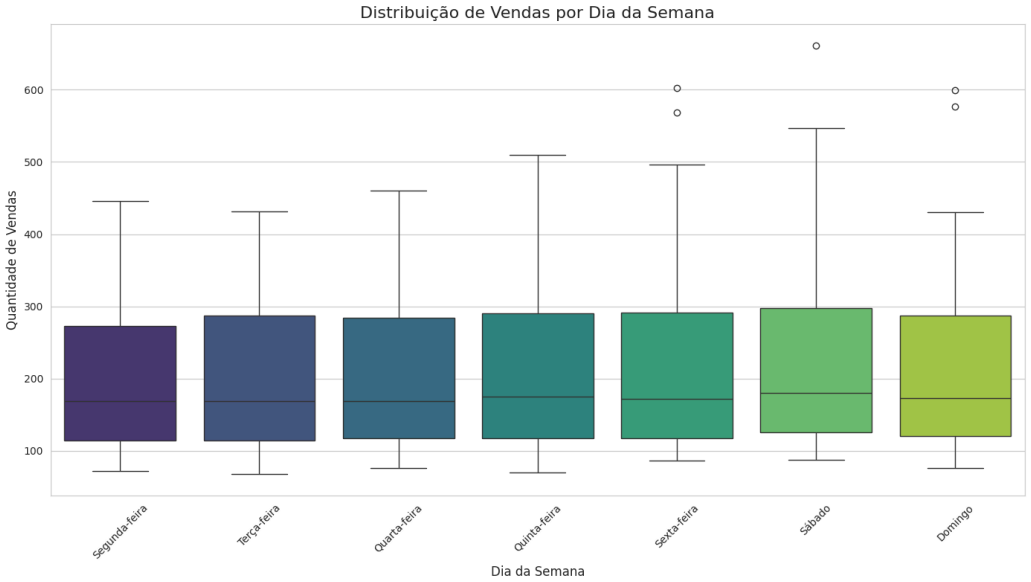


Figura 6: Gráfico Box plot de vendas por dia da semana.

Analisando os gráficos observa-se um comportamento sazonal anual bem definido onde ocorre um aumento das vendas nos meses de novembro e dezembro. Isso pode estar relacionado a datas promocionais em novembro, como a Black Friday, e datas comemorativas em dezembro, como Natal, Ano Novo e recebimento do 13º salário. Também há indícios de sazonalidade semanal em que há um maior volume de vendas nos finais de semana (sexta e sábado) e uma queda aos domingos e segundas-feiras. Há também uma sazonalidade mensal em que as vendas tendem a ser mais altas e com maior variação nos meses de Junho, Julho e Agosto, e também em Novembro e Dezembro, enquanto em Janeiro e Fevereiro apresentam consistentemente os menores volumes de venda. A análise confirma que Dezembro é historicamente um mês de alta, reforçando a importância de uma previsão acurada para este período.

A distribuição das vendas se assemelha a uma distribuição Normal (formato de sino), embora com uma leve assimetria. Na maioria dos dias, as vendas se concentram em torno da média de 192 unidades, com menos dias registrando vendas muito baixas ou muito altas. Pode-se identificar também a presença de outliers, tanto acima quanto abaixo do volume normal de vendas. Existem dias com vendas excepcionalmente altas (acima de 300) e dias com vendas muito baixas. Estes pontos deverão ser investigados e, dependendo do modelo, poderão ser tratados na fase de modelagem.

Em relação a linha de tendência, constata-se uma tendência de crescimento gradual ao longo do tempo, possivelmente ligada à expansão da marca.

O histograma das vendas diárias nos fornece um panorama da frequência dos diferentes volumes de venda ao longo de todo o período. A análise deste gráfico revela:

- Distribuição Assimétrica à Direita: A concentração de dados se encontra na faixa de 100 a 350 unidades vendidas, mas a cauda longa à direita indica a ocorrência de dias com volumes de venda excepcionalmente altos, superando 500 e até 600 unidades.
- Implicação para o Negócio: Essa assimetria é um insight de negócio fundamental. Significa que, embora exista um volume "típico" de vendas, a operação da Segrob Notlad é marcada por picos de demanda expressivos. Um modelo preditivo que se baseie apenas na média ignoraria esses picos, resultando em risco de ruptura de estoque e perda de receita em dias de alta oportunidade. O desafio é prever não apenas o padrão, mas também esses eventos de cauda.

- Múltiplos Picos (Multimodalidade): A presença de mais de um pico sugere que a distribuição de vendas não é estática; ela muda ao longo do tempo. Isso é um forte indicativo da presença de uma tendência de crescimento, onde o "novo normal" de vendas em 2024 é superior ao de 2022.

O Gráfico Box Plot é extremamente rico, pois expõe tanto a tendência de longo prazo quanto a sazonalidade anual:

- Tendência de Crescimento Visível: Ao comparar o mesmo mês em anos diferentes (implícito na dispersão dos pontos), nota-se uma clara tendência de alta. As vendas em 2023 e 2024 são consistentemente superiores às de 2022.
- Padrão Sazonal Anual: Fica evidente um ciclo anual de vendas com dois picos principais: um pico intermediário no meio do ano (destaque para Maio, Junho, Julho e Agosto) e um pico principal, mais forte, no final do ano (Outubro, Novembro e, especialmente, Dezembro). Os meses de Janeiro e Fevereiro representam o período de menor volume de vendas.
- Análise de Outliers e Variabilidade: Os meses de maior volume, como Dezembro, são também os de maior variabilidade (caixas maiores e "bigodes" mais longos), indicando maior incerteza e risco no planejamento. Os outliers (pontos de diamante) são proeminentes em meses-chave:
    - Novembro e Dezembro: Os outliers superiores provavelmente correspondem a eventos como Black Friday e a corrida para as compras de Natal, que são críticos para o faturamento da Segrob Notlad.
    - Maio e Agosto: A presença de outliers nestes meses pode estar ligada a datas comemorativas (como Dia das Mães e Dia dos Pais).

  Analisando o outro Gráfico Box Plot que explora as vendas semanais nos mostra que:

- Ciclo Semanal Claro: As vendas crescem ao longo da semana, atingindo seu pico na Sexta-feira e no Sábado. A mediana de vendas nesses dias é substancialmente maior que a dos demais. Em contrapartida, a Segunda-feira é consistentemente o dia de menor movimento.
- Implicação Operacional: Este padrão deve ditar a estratégia operacional da Segrob Notlad, desde a escala de funcionários nas lojas até a logística de reposição de estoque, que deve ser intensificada para preparar o final de semana.

- Outliers Semanais: A presença de outliers em todos os dias, mas com maior frequência e magnitude nos fins de semana, reforça que os eventos de alta demanda ocorrem preferencialmente nesses dias, potencializando o padrão já existente.

A análise visual dos dados permite concluir que a série temporal de vendas da Segrob Notlad é composta por quatro elementos essenciais que qualquer modelo preditivo deve ser capaz de endereçar:

1. Tendência: Uma forte e clara tendência de crescimento nas vendas ao longo dos anos.
2. Sazonalidade Anual: Um ciclo que se repete todo ano, com picos no meio e no final do ano.
3. Sazonalidade Semanal: Um padrão distinto de vendas para cada dia da semana, com picos nos fins de semana.
4. Outliers e Eventos: Dias de vendas excepcionalmente altas que, em sua maioria, coincidem com os picos sazonais, indicando a influência de feriados e eventos comerciais.

2.3 Preparação dos Dados

Antes que qualquer modelo de análise ou predição seja aplicado, é necessário que os dados estejam devidamente preparados. A qualidade e a consistência dos dados influenciam diretamente os resultados obtidos, em que qualquer discrepância em relação aos dados originais pode comprometer toda a análise. O processo de preparação de dados inclui limpeza, padronização e verificação de integridade para evitar erros ao longo do projeto e assegurar a confiabilidade das conclusões. Esse processo está alinhado com a engenharia de recursos que o faz acontecer por meio do *machine learning*.

Para dar início à preparação dos dados, foi realizada uma verificação inicial para assegurar que todas as colunas estavam devidamente preenchidas e que não havia falta de informações em nenhuma linha da base de dados original, armazenada em formato Excel. Em seguida, foi feita a checagem para identificar e remover possíveis linhas duplicadas, garantindo assim a integridade e a consistência dos dados analisados.

Figura 4: Utilização de um código em python para fazer a verificação de valores inválidos.

2.4 Modelagem dos Dados

Nesta etapa, os dados preparados na fase anterior são analisados com o objetivo de desenvolver modelos preditivos. Essa análise envolve a seleção de algoritmos apropriados, o treinamento dos modelos e a avaliação de seu desempenho, visando à obtenção de resultados confiáveis e relevantes para os objetivos do estudo.

Os modelos foram desenvolvidos com base nos dados coletados no período de janeiro de 2022 a outubro de 2024. Para a etapa subsequente, foi estabelecido que os dados referentes ao mês de novembro serão utilizados na realização dos testes, visando a seleção do modelo mais adequado.

2.4.1 Modelo Naive

O modelo Naive (ou ingênuo) é um dos modelos mais simples de previsão em séries temporais. Ele assume que o valor da próxima observação será igual ao valor mais recente. Ou seja, ele projeta que nada vai mudar.

Item	**Descrição**
Modelo:	Modelo Naive
Se baseia em:	$xt = xt - 1 + et$

Item	Descrição
Onde:	$et \sim iid(\mu = 0, \sigma 2 = V[e])$
Modelo de previsão:	$x\hat{}\, t, t + 1 = xt$

Tabela 2: Descrição do modelo Naive

Fonte: Introdução a Séries Temporais, Dalton Borges

Explicando a tabela:

- O valor atual da série temporal xt é igual ao valor do período anterior xt−1 mais um erro aleatório et. Esse erro representa a variação imprevisível entre um período e outro.
- $et \sim iid(\mu = 0, \sigma 2 = V[e])$
    - iid: erros são independentes e identicamente distribuídos, ou seja, um erro não influencia o outro e todos seguem a mesma distribuição.
    - μ=0: o erro tem média zero, o que quer dizer que, em média, ele não puxa a série nem para cima nem para baixo.
    - σ2=V[e]: o erro tem uma variância constante, indicando que a dispersão dos erros em torno da média é estável.
- $x\hat{}\, t, t + 1 = xt$
    - A previsão para o próximo período (t+1) é simplesmente o valor atual (xt).
    - Isso significa que esperamos que o próximo valor da série seja igual ao último observado, sem tentar ajustar para tendências, sazonalidades ou padrões.

A implementação do modelo, por meio da linguagem Python, na base de dados da empresa,

proporcionou os seguintes resultados para o mês de novembro:

Previsões para Novembro/2024:

	Naive
2024-11-01	412
2024-11-02	412
2024-11-03	412
2024-11-04	412
2024-11-05	412
2024-11-06	412
2024-11-07	412
2024-11-08	412
2024-11-09	412
2024-11-10	412
2024-11-11	412
2024-11-12	412
2024-11-13	412
2024-11-14	412
2024-11-15	412
2024-11-16	412
2024-11-17	412
2024-11-18	412
2024-11-19	412
2024-11-20	412
2024-11-21	412
2024-11-22	412
2024-11-23	412
2024-11-24	412
2024-11-25	412
2024-11-26	412
2024-11-27	412
2024-11-28	412
2024-11-29	412
2024-11-30	412

Figura 5: Previsão de vendas para novembro seguindo modelo naive.

2.4.2 Modelo Cumulativo

Diferente do modelo Naive, o modelo Cumulativo mostra a soma acumulada desses valores. Isso ajuda a identificar tendências de crescimento ou queda de forma mais clara, suavizando as flutuações pontuais. O modelo Cumulativo dá mais importância ao histórico de demanda do que o modelo Naive, em outras palavras, ele valoriza mais o passado.

Item	Descrição
Modelo:	Modelo Cumulativo
Se baseia em:	$xt = a + et$
Onde:	$et \sim iid(\mu = 0, \sigma 2 = V[e])$
Modelo de previsão:	$x\hat{}\, t, t + 1 = (\sum tt - 1\, xi)/t$

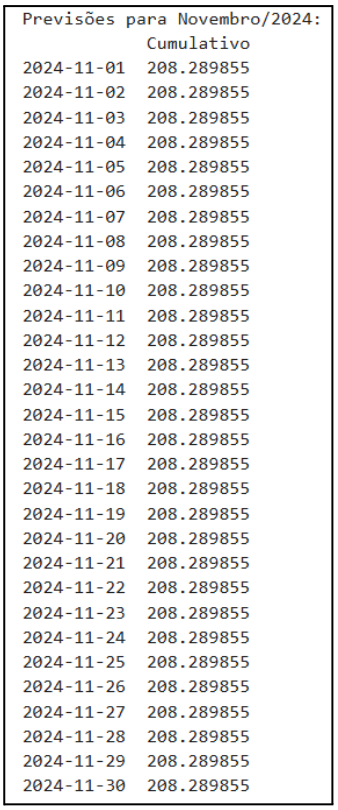
Tabela 3: Descrição do modelo Cumulativo

Fonte:Introdução a Séries Temporais, Dalton Borges

Explicando a tabela:

- Esse modelo assume que o valor da série temporal no tempo t é uma constante α mais um erro aleatório t. A constante aaa representa um nível médio fixo da série.
- $et \sim iid(\mu = 0, \sigma 2 = V[e])$
    - iid: erros são independentes e identicamente distribuídos, ou seja, um erro não influencia o outro e todos seguem a mesma distribuição.
    - μ=0: o erro tem média zero, o que quer dizer que, em média, ele não puxa a série nem para cima nem para baixo.
    - σ2=V[e]: o erro tem uma variância constante, indicando que a dispersão dos erros em torno da média é estável.
- $x\hat{}\, t, t + 1 = (\sum tt - 1\, xi)/t$
    - Essa fórmula representa a média aritmética dos valores observados até o tempo t.
    - É uma forma de suavizar a série, considerando todos os valores anteriores igualmente.
    - Serve para prever o próximo valor assumindo que a tendência média se mantém.

A implementação do modelo, por meio da linguagem Python, na base de dados da empresa, proporcionou os seguintes resultados para os mês de novembro:



Previsões para Novembro/2024:

	Cumulativo
2024-11-01	208.289855
2024-11-02	208.289855
2024-11-03	208.289855
2024-11-04	208.289855
2024-11-05	208.289855
2024-11-06	208.289855
2024-11-07	208.289855
2024-11-08	208.289855
2024-11-09	208.289855
2024-11-10	208.289855
2024-11-11	208.289855
2024-11-12	208.289855
2024-11-13	208.289855
2024-11-14	208.289855
2024-11-15	208.289855
2024-11-16	208.289855
2024-11-17	208.289855
2024-11-18	208.289855
2024-11-19	208.289855
2024-11-20	208.289855
2024-11-21	208.289855
2024-11-22	208.289855
2024-11-23	208.289855
2024-11-24	208.289855
2024-11-25	208.289855
2024-11-26	208.289855
2024-11-27	208.289855
2024-11-28	208.289855
2024-11-29	208.289855
2024-11-30	208.289855

Figura 6: Previsão de vendas para novembro seguindo modelo cumulativo.

2.4.3 Modelo de Média Móvel

A Média Móvel é um modelo que procura generalizar os modelos Cumulativos e Naive e possui abordagens que se situam entre os extremos. Parecido com o modelo Cumulativo, porém aos invés de calcular a média de todas as observações passadas, a Média Móvel tira a média das *M* ultimas observações.

Item	**Descrição**
Modelo:	Média Móvel
Se baseia em:	$xt = a + et$
Onde:	$et \sim iid(\mu = 0, \sigma 2 = V[e])$
Modelo de previsão:	$x\hat{}\, t, t + 1 = (\sum ti = t + 1 - M\hat{}\{X\}i)/\mathrm{M}$

Tabela 4: Descrição do modelo Média Móvel

Fonte:Introdução a Séries Temporais, Dalton Borges

Explicando a tabela:

- $x\hat{}\ t, t + 1 = (\sum ti = t + 1 - M\hat{}\{X\}i)$/M
    - A previsão para o próximo valor (t+1) é a média dos últimos M valores da série.
    - Essa média se move ao longo do tempo, descartando o dado mais antigo e incluindo o mais recente.

Esse modelo é muito útil sem séries de tendências fortes,para filtrar ruídos e captar o comportamento recente da série. Caso o valor de M escolhido for muito pequeno, o modelo responderá rapidamente a ruídos, e se for muito grande, perderá mudanças que não resistem por muito tempo. Normalmente utilizam valores práticos, que reflitam a unidade da escala temporal,como 1 semana, 4 meses, etc.

A implementação do modelo, por meio da linguagem Python, na base de dados da empresa, proporcionou os seguintes resultados para o mês de novembro:

```
Previsões para Novembro/2024:
            Média Móvel
2024-11-01   414.533333
2024-11-02   415.251111
2024-11-03   416.059481
2024-11-04   416.761464
2024-11-05   416.586846
2024-11-06   417.039741
2024-11-07   417.374399
2024-11-08   418.020213
2024-11-09   418.620886
2024-11-10   419.408249
2024-11-11   414.455191
2024-11-12   406.237031
2024-11-13   400.544932
2024-11-14   401.129763
2024-11-15   401.067421
2024-11-16   401.636335
2024-11-17   402.124213
2024-11-18   402.295020
2024-11-19   402.704854
2024-11-20   402.695016
2024-11-21   403.718183
2024-11-22   404.675456
2024-11-23   404.897971
2024-11-24   405.661237
2024-11-25   406.549945
2024-11-26   407.001610
2024-11-27   407.301663
2024-11-28   407.645052
2024-11-29   408.566554
2024-11-30   409.085439
```

Figura 7: Previsão de vendas para novembro seguindo modelo de média móvel.

2.4.4 Modelo de Suavização Exponencial

O modelo de Suavização Exponencial é uma técnica que prevê valores futuros com base em valores passados, dando mais peso para os dados mais recentes e menos peso para os os antigos, por isso o nome “exponencial”, pois a importância dos dados antigos decai exponencialmente.

A suavização exponencial simples implica em :

- Demanda estacionária , sem tendência nem sazonalidade. Considera apenas o nível da demanda.
- O valor das observações diminui com o tempo.
- Utiliza uma constante de suavização α , onde $0 \leq \alpha \geq 1$.
- Na pratica $0{,}1 \leq \alpha \geq 0{,}3$.

Item	**Descrição**
Modelo:	Suavização Exponencial Simples

Se baseia em:	$xt = a + et$
Onde:	$et \sim iid(\mu = 0, \sigma 2 = V[e])$
Modelo de previsão:	$x\hat{}\ t + 1, t = \alpha xt + (1 - \alpha)x\hat{}\ t - 1, t\ com\ 0 \leq \alpha \leq 1$

Tabela 4: Descrição do modelo Suavização Exponencial Simples

Fonte: Suavização Exponencial I, Dalton Borges

Explicando a tabela:

- $x\hat{}t + 1, t = \alpha xt + (1 - \alpha)x\hat{}t - 1, t\ com\ 0 \leq \alpha \leq 1$
  - $x\hat{}t + 1, t$: previsão do próximo valor (tempo t+1) feita no tempo t.
  - α: fator de suavização (entre 0 e 1).
  - Se α≈1: mais peso ao valor atual → previsão mais reativa.
  - Se α≈0: mais peso à previsão passada → previsão mais suave.

A implementação do modelo, por meio da linguagem Python, na base de dados da empresa, proporcionou os seguintes resultados para o mês de novembro:

Previsões para Novembro/2024:

	Suavização Exp.
2024-11-01	398.297011
2024-11-02	395.933658
2024-11-03	397.588005
2024-11-04	396.429962
2024-11-05	397.240592
2024-11-06	396.673151
2024-11-07	397.070360
2024-11-08	396.792314
2024-11-09	396.986946
2024-11-10	396.850704
2024-11-11	396.946073
2024-11-12	396.879315
2024-11-13	396.926046
2024-11-14	396.893334
2024-11-15	396.916232
2024-11-16	396.900203
2024-11-17	396.911424
2024-11-18	396.903569
2024-11-19	396.909067
2024-11-20	396.905219
2024-11-21	396.907913
2024-11-22	396.906027
2024-11-23	396.907347
2024-11-24	396.906423
2024-11-25	396.907070
2024-11-26	396.906617
2024-11-27	396.906934
2024-11-28	396.906712
2024-11-29	396.906867
2024-11-30	396.906759

Figura 8: Previsão de vendas para novembro seguindo modelo de suavização exponencial simples.

2.4.5 Regressão Linear Simples e Regressão Linear Dinâmica

A regressão linear simples é uma técnica estatística amplamente utilizada na análise de dados para modelar a relação entre duas variáveis quantitativas. Trata-se de um método preditivo que busca estimar o valor de uma variável dependente y a partir de uma variável independente x, assumindo que essa relação pode ser representada por uma equação linear (MONTGOMERY; PECK; VINING, 2012).

O modelo de regressão linear simples pode ser expresso matematicamente pela equação:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

onde:

- y representa a variável dependente (ou resposta);
- xé a variável independente (ou explicativa);
- β0 é o intercepto da reta de regressão;
- β1 é o coeficiente angular, que indica a inclinação da reta e a direção da relação entre as variáveis;
- ε é o termo de erro aleatório, que representa os desvios entre os valores observados e os valores ajustados pelo modelo.

O objetivo principal da regressão linear simples é encontrar os valores dos parâmetros β0 e β1 que minimizem a soma dos quadrados dos resíduos, ou seja, as diferenças entre os valores observados e os valores previstos pelo modelo. Esse processo é conhecido como método dos mínimos quadrados ordinários (OLS - *Ordinary Least Squares*).

A aplicação da regressão linear simples permite não apenas realizar previsões, mas também avaliar o grau de associação entre as variáveis envolvidas. Contudo, para que os resultados obtidos por esse modelo sejam confiáveis, é necessário que algumas suposições

sejam atendidas, como a linearidade da relação, a homocedasticidade dos resíduos, a normalidade dos erros e a independência das observações.

Para superar as limitações do modelo simples, evoluiu-se para uma Regressão Linear Dinâmica. O objetivo desta abordagem foi enriquecer o modelo com variáveis que representassem a dinâmica temporal intrínseca à série de vendas, especialmente os "lags".

O processo de construção do modelo dinâmico envolveu a engenharia e seleção de features, onde diversas variáveis baseadas em lags de vendas foram propostas e testadas. Os candidatos iniciais foram:

- Lag de 1 dia (Vendas t-1): Para capturar o efeito de "inércia" ou continuidade das vendas de um dia para o outro.
- Lag de 7 dias (Vendas t-7): Para modelar o forte padrão sazonal dos dias da semana, crucial para negócios de varejo.
- Lag de 14 dias (Vendas t-14): Para reforçar e estabilizar o padrão semanal.
- Média Móvel de 7 dias: Para capturar a tendência local recente, suavizando o ruído diário.

Após a avaliação da performance de diferentes combinações dessas variáveis em um conjunto de validação, foram selecionados os preditores que demonstraram o maior poder explicativo e contribuíram para a redução do erro de previsão.

Os lags selecionados para compor a versão final do modelo de Regressão Linear Dinâmica foram:

1. Lag de 1 dia (Vendas t-1): Demonstrou ser fundamental para capturar a correlação de curto prazo.
2. Lag de 7 dias (Vendas t-7): Foi a variável mais impactante para modelar a sazonalidade semanal, diferenciando fins de semana de dias úteis.
3. Média Móvel dos Últimos 7 Dias: Apresentou grande valor para ajustar o modelo à tendência local, oferecendo uma visão mais estável do que um único lag.

Dessa forma, o modelo final de Regressão Linear Dinâmica foi treinado para prever as vendas do dia atual com base em uma combinação da tendência geral, do valor de vendas do

dia anterior, do valor de vendas do mesmo dia na semana anterior e da média de vendas da última semana.

Previsões para Novembro/2024 (Regressão Linear Simples e Dinâmica)

	Reais	Regressão Linear Simples	Regressão Linear Dinâmica
2024-11-01	398	368.024887	402.354293
2024-11-02	408	368.333344	394.992045
2024-11-03	393	368.641800	389.389527
2024-11-04	389	368.950257	385.143132
2024-11-05	431	369.258714	381.941822
2024-11-06	415	369.567171	379.545899
2024-11-07	409	369.875628	377.770642
2024-11-08	410	370.184085	376.473697
2024-11-09	413	370.492542	375.545358
2024-11-10	424	370.800999	374.901083
2024-11-11	412	371.109455	374.475721
2024-11-12	408	371.417912	374.219061
2024-11-13	408	371.726369	374.092411
2024-11-14	405	372.034826	374.065951
2024-11-15	431	372.343283	374.116703
2024-11-16	411	372.651740	374.226958
2024-11-17	405	372.960197	374.383067
2024-11-18	397	373.268654	374.574514
2024-11-19	416	373.577110	374.793194
2024-11-20	410	373.885567	375.032861
2024-11-21	412	374.194024	375.288701
2024-11-22	403	374.502481	375.557005
2024-11-23	401	374.810938	375.834914
2024-11-24	410	375.119395	376.120225
2024-11-25	390	375.427852	376.411240
2024-11-26	399	375.736309	376.706652
2024-11-27	414	376.044765	377.005451
2024-11-28	433	376.353222	377.306862
2024-11-29	496	376.661679	377.610284
2024-11-30	547	376.970136	377.915256

Figura 9: Previsão de vendas para novembro seguindo modelo de regressão linear simples e dinâmica.

2.4.6 KNN (K Nearest Neighbors)

O KNN (K-Nearest Neighbors) é um algoritmo de aprendizado supervisionado que pode ser usado tanto para classificação quanto para regressão. Neste projeto, como estamos prevendo um valor contínuo (volume de vendas), o KNN será usado para regressão.

Esse modelo realiza a previsão com base no cálculo da distância entre um novo ponto de dados e os demais pontos presentes no conjunto de dados. Para isso, identifica os k vizinhos mais próximos desse novo ponto, considerando, em geral, a distância euclidiana como métrica de similaridade como determina a equação abaixo. A estimativa do valor é feita a partir da média dos valores dos k vizinhos mais próximos.

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Vantagens do Modelo:

- Flexibilidade: Por ser um modelo não-paramétrico, o KNN não faz suposições sobre a distribuição dos dados, o que lhe confere alta flexibilidade para se adaptar a padrões complexos e não-lineares que uma regressão linear não conseguiria capturar.
- Simplicidade Conceitual: A lógica do modelo é intuitiva e fácil de explicar em termos de negócio: "a previsão para hoje é baseada nos dias mais parecidos que já aconteceram", o que o torna um modelo transparente.

Desvantagens no Contexto do Desafio:

- Incapacidade de Extrapolar Tendências: Esta é a principal desvantagem do KNN para o problema da Segrob Notlad. Como o modelo prevê com base em médias de valores passados, ele é estruturalmente incapaz de prever um valor superior ao máximo já visto em seu histórico. Em uma série com forte tendência de crescimento como a da Segrob Notlad, o KNN sempre estará "atrasado", subestimando consistentemente as vendas futuras.
- Sensibilidade à Engenharia de Features: A definição de "proximidade" ou "semelhança" entre os dias depende inteiramente das variáveis (features) utilizadas. Se features irrelevantes forem incluídas, o modelo pode ser levado a encontrar "vizinhos" inadequados, prejudicando a acurácia. A performance mediana do KNN sugere que a combinação de features testada não foi suficiente para superar os modelos mais simples.

Após a implementação do modelo em Python, os resultados obtidos estão apresentados na Figura abaixo.

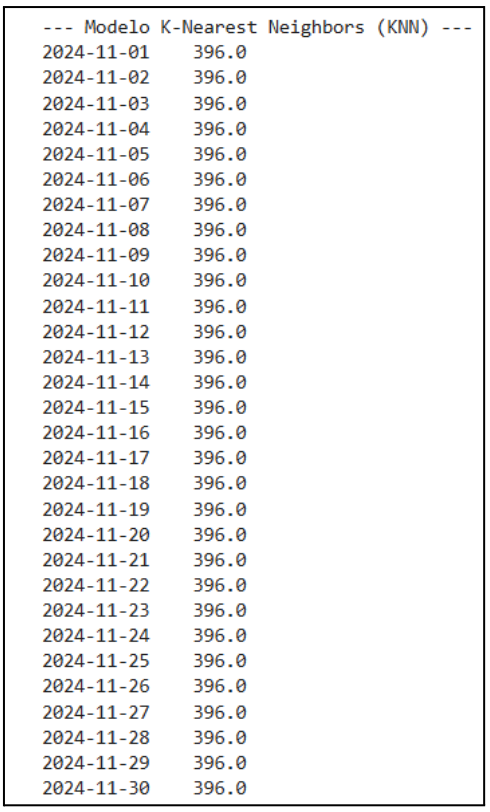


Figura 10: Previsão de vendas para novembro seguindo modelo KNN.

2.4.7 SVM/SVR (Support Vector Machines/Support Vector Regression)

SVM (Support Vector Machine) é um algoritmo de aprendizado de máquina supervisionado usado para tarefas de classificação e regressão. Quando aplicado a problemas de regressão, é conhecido como Support Vector Regression (SVR).

Esse modelo busca definir uma fronteira de decisão (hiperplano) que separe da melhor forma possível os pontos de dados de diferentes classes. Para regressão, o SVR tenta ajustar o hiperplano aos dados o mais próximo possível, considerando uma margem de erro aceitável. Para dados não linearmente separáveis, o SVM/SVR utiliza um "truque do kernel" (kernel trick). Essa técnica transforma os dados em um espaço de dimensão superior, onde se torna mais fácil encontrar uma fronteira linear.

Vantagens:

- Robustez a Outliers: Esta é a maior vantagem teórica do SVR para os dados da Segrob Notlad. A série de vendas possui diversos picos e valores atípicos (outliers), e o SVR é, por natureza, mais robusto a eles do que modelos como a Regressão Linear, o que poderia levar a previsões mais estáveis.

- Capacidade de Modelar Relações Não-Lineares: Com o uso de "kernels" (como o RBF), o SVR é extremamente eficaz em capturar relações complexas e não-lineares entre as variáveis, o que o tornava um candidato promissor para modelar a dinâmica de vendas.

Desvantagem no Contexto do Desafio:

- Complexidade e Custo de Otimização: O desempenho do SVR é altamente dependente da sintonia fina de seus hiperparâmetros. Encontrar a combinação ideal é um processo complexo e computacionalmente caro. A performance obtida, embora razoável, pode indicar que o modelo não foi perfeitamente otimizado para este problema específico.

Após a implementação do modelo em Python, os resultados obtidos estão apresentados na Figura abaixo.

```
--- Modelo Support Vector Regression (SVR) ---
2024-11-01    393.797989
2024-11-02    394.141892
2024-11-03    394.482406
2024-11-04    394.819501
2024-11-05    395.153150
2024-11-06    395.483326
2024-11-07    395.810001
2024-11-08    396.133147
2024-11-09    396.452738
2024-11-10    396.768747
2024-11-11    397.081146
2024-11-12    397.389910
2024-11-13    397.695013
2024-11-14    397.996428
2024-11-15    398.294130
2024-11-16    398.588093
2024-11-17    398.878293
2024-11-18    399.164704
2024-11-19    399.447301
2024-11-20    399.726061
2024-11-21    400.000958
2024-11-22    400.271970
2024-11-23    400.539073
2024-11-24    400.802243
2024-11-25    401.061456
2024-11-26    401.316692
2024-11-27    401.567926
2024-11-28    401.815137
2024-11-29    402.058303
2024-11-30    402.297402
```

Figura 11: Previsão de vendas para novembro seguindo modelo SVR.

2.4.8 Suavização Dupla

O método de Suavização Exponencial Dupla, amplamente conhecido como Método de Holt, é uma técnica de previsão desenvolvida como uma extensão da Suavização Exponencial Simples. Sua principal finalidade é ser aplicada a séries temporais que exibem uma tendência clara, seja de crescimento ou de declínio, uma característica que o modelo simples não consegue capturar.

Enquanto a Suavização Exponencial Simples estima apenas um componente (o "nível" da série), o Método de Holt decompõe a série em dois componentes, que são atualizados a cada nova observação:

1. Nível (lt): Representa o valor médio suavizado da série no tempo *t*. É a linha de base da série.
2. Tendência (bt): Representa a taxa de crescimento (ou inclinação) suavizada da série no tempo *t*.

O modelo utiliza dois parâmetros de suavização distintos para controlar a ponderação dada às observações mais recentes em cada componente:

- Alfa (α): Para o nível. Um valor de alfa próximo de 1 dá mais peso às observações recentes para determinar o nível atual, tornando-o mais reativo.
- Beta (β): Para a tendência. Um valor de beta próximo de 1 faz com que a estimativa da tendência se ajuste rapidamente a mudanças na inclinação da série.

A previsão para h períodos no futuro é calculada somando-se o último nível estimado com a projeção da tendência ao longo desses h períodos (Previsão(t+h)=lt+h×bt). O resultado é uma linha reta inclinada, que segue a tendência observada nos dados.

Características e Aplicabilidade no Desafio:

- Vantagem Principal: A grande força do Método de Holt é sua capacidade de modelar e extrapolar a tendência de crescimento observada nos dados de vendas da Segrob Notlad. Ele é teoricamente superior a modelos como a Média Móvel ou a Suavização Simples, que tendem a ficar "atrasados" em relação a uma série em crescimento.
- Limitação Crítica: A principal desvantagem do modelo é a ausência de um componente para lidar com a sazonalidade. Ele projeta uma linha de crescimento

contínua, sendo incapaz de prever os picos e vales recorrentes que ocorrem em padrões sazonais (como as altas de vendas nos fins de semana ou no mês de dezembro). Na prática, o modelo pode até seguir a tendência geral, mas provavelmente errará sistematicamente para baixo nos picos sazonais e para cima nos vales sazonais.

Após a implementação do modelo em Python, os resultados obtidos estão apresentados na Figura abaixo.

Previsões para Novembro/2024:

	Suavização Exp. Dupla
2024-11-01	409.887810
2024-11-02	410.227288
2024-11-03	410.566765
2024-11-04	410.906242
2024-11-05	411.245720
2024-11-06	411.585197
2024-11-07	411.924674
2024-11-08	412.264151
2024-11-09	412.603629
2024-11-10	412.943106
2024-11-11	413.282583
2024-11-12	413.622061
2024-11-13	413.961538
2024-11-14	414.301015
2024-11-15	414.640492
2024-11-16	414.979970
2024-11-17	415.319447
2024-11-18	415.658924
2024-11-19	415.998401
2024-11-20	416.337879
2024-11-21	416.677356
2024-11-22	417.016833
2024-11-23	417.356311
2024-11-24	417.695788
2024-11-25	418.035265
2024-11-26	418.374742
2024-11-27	418.714220
2024-11-28	419.053697
2024-11-29	419.393174
2024-11-30	419.732652

Figura 12: Previsão de vendas para novembro seguindo modelo de suavização exponencial dupla.

2.4.9 Suavização Tripla

O método de Suavização Exponencial Tripla, mais conhecido como Método de Holt-Winters, é a abordagem mais completa da família de suavização exponencial. Ele foi desenvolvido como uma extensão do método de Holt para incorporar um terceiro componente: a sazonalidade. Isso o torna uma das técnicas estatísticas clássicas mais poderosas e adequadas para séries temporais que, como a da Segrob Notlad, possuem simultaneamente um nível base, uma tendência de crescimento e ciclos sazonais recorrentes.

O Método de Holt-Winters funciona decompondo a série temporal em três componentes, cada um sendo atualizado de forma exponencial a cada nova observação de dado:

1. Nível (lt): O valor médio da série após a remoção do efeito sazonal.
2. Tendência (bt): A taxa de crescimento ou inclinação da série.
3. Sazonalidade (st): O fator de influência do período sazonal específico (por exemplo, o impacto de ser uma segunda-feira vs. um sábado, ou o mês de dezembro vs. fevereiro).

O modelo é governado por três parâmetros de suavização (alfa, α; beta, β; e gama, γ), que controlam a velocidade com que cada um desses componentes se adapta às informações mais recentes.

Uma característica fundamental do método de Holt-Winters é a sua capacidade de modelar a sazonalidade de duas formas distintas:

- Sazonalidade Aditiva: É utilizada quando a magnitude da variação sazonal é relativamente constante ao longo do tempo, independentemente do nível da série. A previsão é calculada como (Nível + Tendência) + Fator Sazonal.
- Sazonalidade Multiplicativa: É aplicada quando a variação sazonal é proporcional ao nível da série, ou seja, ela cresce ou diminui em termos percentuais. A previsão é calculada como (Nível + Tendência) * Fator Sazonal. Este método é frequentemente mais adequado para séries de vendas, onde um aumento de 20% no Natal representa um valor absoluto muito maior em um ano de vendas altas do que em um ano de vendas baixas.

Características e Aplicabilidade no Desafio

- Vantagem Principal: A grande força do método Holt-Winters é sua capacidade de modelar, de forma integrada, todas as principais características identificadas na análise exploratória dos dados da Segrob Notlad: a tendência de crescimento, a sazonalidade semanal (picos nos fins de semana) e a sazonalidade anual (picos no meio e fim de ano).
- Implementação: Para uma aplicação correta, é crucial definir adequadamente o período da sazonalidade (por exemplo, um período de 7 para capturar o padrão

semanal) e escolher o tipo de modelo (aditivo ou multiplicativo) que melhor descreve o comportamento dos dados. Dada a natureza das vendas, o modelo multiplicativo é o candidato teórico mais forte.

- Potencial: Por sua completude, o Holt-Winters é um dos modelos com maior potencial teórico para gerar previsões acuradas para o desafio, superando abordagens que não conseguem lidar com a complexa combinação de tendência e sazonalidade.

Após a implementação do modelo em Python, os resultados obtidos estão apresentados na Figura abaixo.

Previsões para Novembro/2024:

	Suavização Exp. Tripla
2024-11-01	417.598568
2024-11-02	422.825536
2024-11-03	415.543250
2024-11-04	393.982538
2024-11-05	401.574144
2024-11-06	406.665899
2024-11-07	412.709977
2024-11-08	419.811075
2024-11-09	425.038043
2024-11-10	417.755757
2024-11-11	396.195045
2024-11-12	403.786651
2024-11-13	408.878407
2024-11-14	414.922484
2024-11-15	422.023582
2024-11-16	427.250550
2024-11-17	419.968264
2024-11-18	398.407552
2024-11-19	405.999158
2024-11-20	411.090914
2024-11-21	417.134991
2024-11-22	424.236089
2024-11-23	429.463057
2024-11-24	422.180771
2024-11-25	400.620060
2024-11-26	408.211665
2024-11-27	413.303421
2024-11-28	419.347498
2024-11-29	426.448596
2024-11-30	431.675564

Figura 13: Previsão de vendas para novembro seguindo modelo de suavização exponencial tripla.

2.4.10 Random Forest (Árvore de Decisão)

O Random Forest é um algoritmo de *machine learning* do tipo *ensemble* (conjunto), considerado um dos métodos mais poderosos e versáteis para tarefas de regressão e classificação. Em vez de construir um único modelo preditivo, sua estratégia consiste em criar uma "floresta" com centenas ou milhares de árvores de decisão e, em seguida, agregar os resultados de todas elas para gerar uma previsão final mais precisa, estável e robusta.

O funcionamento do Random Forest se baseia em duas técnicas principais que garantem a diversidade entre as árvores, o que é crucial para a eficácia do método:

1. Bagging (Bootstrap Aggregating): Para cada árvore que será construída na floresta, é criada uma amostra de dados de treinamento por meio de um processo de amostragem aleatória com reposição (bootstrap). Isso significa que cada árvore é treinada em uma versão ligeiramente diferente do conjunto de dados original.
2. Aleatoriedade de Features: Em cada nó de uma árvore de decisão, quando o algoritmo precisa escolher a melhor variável para fazer uma divisão, ele não considera todas as variáveis disponíveis. Em vez disso, ele seleciona um subconjunto aleatório de features, forçando a árvore a encontrar a melhor divisão possível apenas dentro daquele subconjunto.

Essa dupla aleatoriedade (nos dados e nas features) resulta em árvores altamente descorrelacionadas. Para uma tarefa de previsão (regressão), a predição final do Random Forest é simplesmente a média das previsões de todas as árvores individuais. Esse processo de agregação reduz a variância e a tendência de *overfitting* (sobreajuste) que uma única árvore de decisão teria.

O Random Forest não compreende o tempo de forma nativa. Para aplicá-lo a um problema de previsão como o da Segrob Notlad, é necessário transformar a série temporal em um formato de aprendizado supervisionado. Isso é feito através da engenharia de features, onde o objetivo é criar variáveis preditoras (X) para prever o valor futuro da série (Y). As features criadas para este projeto incluiriam:

- Lags de Vendas: O valor das vendas do dia anterior (t-1), do mesmo dia na semana anterior (t-7), etc.
- Features de Calendário: Dia da semana, semana do ano, mês, ano.
- Estatísticas de Janela Móvel: Média, mediana ou desvio padrão das vendas nos últimos 7 ou 30 dias.

Características e Aplicabilidade no Desafio

- Vantagem Principal: Sua maior força é a capacidade de capturar relações não-lineares e interações complexas entre as variáveis sem a necessidade de especificá-las previamente. Por exemplo, ele pode aprender que o impacto das vendas de uma

sexta-feira (lag t-1) é diferente se essa sexta-feira cair em dezembro ou em março. Além disso, ele oferece como subproduto a importância das features, que pode gerar insights de negócio valiosos para a Segrob Notlad.

- Limitação Crítica: A desvantagem mais significativa para séries com forte tendência, como a de vendas da Segrob Notlad, é a sua incapacidade de extrapolar. Um modelo baseado em árvores não consegue prever valores que estejam fora do intervalo observado em seu conjunto de treinamento. Se as vendas estão em constante crescimento, o Random Forest atingirá um "teto" correspondente ao valor máximo que ele já viu, sendo incapaz de projetar o crescimento futuro. Isso explica por que, muitas vezes, modelos mais simples que capturam a tendência de forma explícita podem superá-lo em cenários de forte crescimento.

Após a implementação do modelo em Python, os resultados obtidos estão apresentados na Figura abaixo.

```
--- Previsões de Vendas para Novembro de 2024 (Random Forest) ---
            Previsão_Vendas
Data
2024-11-01           399.67
2024-11-02           403.64
2024-11-03           400.39
2024-11-04           395.34
2024-11-05           399.77
2024-11-06           411.67
2024-11-07           410.52
2024-11-08           412.51
2024-11-09           412.95
2024-11-10           421.10
2024-11-11           408.86
2024-11-12           408.65
2024-11-13           408.31
2024-11-14           409.40
2024-11-15           422.76
2024-11-16           414.22
2024-11-17           411.82
2024-11-18           402.31
2024-11-19           410.24
2024-11-20           410.65
2024-11-21           410.04
2024-11-22           412.07
2024-11-23           403.70
2024-11-24           408.42
2024-11-25           404.80
2024-11-26           407.97
2024-11-27           414.76
2024-11-28           442.31
2024-11-29           472.74
2024-11-30           469.58
```

Figura 13: Previsão de vendas para novembro seguindo o Random Forest.

5. Métricas utilizadas

Nessa etapa do projeto, a análise e avaliação de modelos preditivos é essencial para compreender as métricas utilizadas para mensurar a precisão das previsões, bem como os métodos de validação empregados para garantir sua confiabilidade.

A qualidade das previsões de um modelo pode ser mensurada por meio de diversas métricas estatísticas, tais como o MAPE (Mean Absolute Percentage Error – Erro Percentual Médio Absoluto), o RMSE (Root Mean Squared Error – Raiz do Erro Quadrático Médio) e o MAD (Mean Absolute Deviation – Desvio Absoluto Médio). > Para a aplicação dessas métricas, serão comparados os valores referentes ao mês de novembro, tomando como base os valores reais e os valores previstos, possibilitando uma análise precisa do desempenho do modelo.

3.1 MAPE

Essa métrica faz o cálculo da média dos erros absolutos expressos em termos percentuais em relação aos valores reais. Quanto menor o MAPE, melhor a precisão do modelo.

$$MAPE = \frac{\Sigma(\frac{|y_i - \widehat{y_i}|}{y_i})}{n}$$

- y_i - valor real;
- $\widehat{y_i}$ - valor previsto;
- n - total de observações.

3.2 RMSE

Essa métrica avalia a magnitude dos erros elevando ao quadrado as diferenças entre os valores reais e previstos.

$$RMSE = \sqrt{\frac{1}{n}\Sigma(y_i - \widehat{y_i})^2}$$

3.3 MAD

Essa métrica mede a dispersão dos erros, calculando a média das diferenças absolutas entre os valores reais e os valores previstos.

$$MAD = \frac{1}{n}\Sigma|y_i - \widehat{y_i}|$$

6. Validação

Após a aplicação inicial dos modelos, a etapa de validação é crucial para garantir que a escolha da abordagem final não seja baseada em acaso, mas sim em uma performance robusta e generalizável. Nesta seção, descrevemos a metodologia de validação empregada e comparamos os modelos quantitativamente para selecionar o mais adequado para o desafio de negócio da Segrob Notlad.

A validação dos modelos de previsão é uma etapa essencial para garantir a confiabilidade das estimativas realizadas. Para isso, foram calculadas métricas estatísticas por meio da linguagem de programação Python, que avaliam o desempenho de cada abordagem utilizada para prever os valores referentes ao mês de novembro de 2024.

4.1 Metodologia de Validação: Validação Cruzada para Séries Temporais

Para obter uma estimativa realista do desempenho de cada modelo, utilizamos uma técnica de Validação Cruzada (Cross-Validation). No entanto, a validação cruzada padrão (k-Fold), que divide os dados aleatoriamente, não pode ser aplicada a este problema. A razão é que ela destruiria a ordem cronológica dos dados de vendas, permitindo que o modelo fosse treinado com informações do futuro para prever o passado, o que geraria uma avaliação de performance irrealista.

Para o caso da Segrob Notlad, aplicamos a abordagem correta para dados temporais, conhecida como Validação Cruzada Sequencial (ou *Forward-Chaining*). O processo funciona da seguinte forma:

1. Criação de "Folds" Temporais: O histórico de dados (2022-2024) é dividido em múltiplos "folds" (dobras) sequenciais. Cada fold consiste em um período de treino e um período de teste subsequente.
2. Treinamento e Teste Iterativo: O modelo é treinado e avaliado múltiplas vezes. Por exemplo:
    - Iteração 1: Treina-se com os dados dos primeiros 12 meses e testa-se a previsão para o 13º mês.

    - Iteração 2: Treina-se com os dados dos primeiros 13 meses e testa-se a previsão para o 14º mês.
    - ... e assim por diante, até o final da série.
3. Cálculo da Média de Erro: As métricas de erro (MAE, RMSE e MAD) são calculadas em cada iteração. O desempenho final de um modelo é a média desses erros.

Alguma vantagens e desvantagens desta metodologia:

Vantagens:

- Maximiza o uso dos dados: O modelo tem a oportunidade de aprender com um volume crescente de informações, o que é ideal para capturar tendências de longo prazo e padrões complexos que exigem uma base de dados maior.
- Robustez em processos estáveis: Se o comportamento de compra do consumidor da Segrob Notlad for relativamente estável ao longo dos anos, este método tende a gerar modelos mais robustos, pois se beneficia de toda a história disponível.

Desvantagens:

- Vulnerabilidade a "Concept Drift": Se o comportamento do consumidor mudar ao longo do tempo (por exemplo, devido a novas campanhas de marketing, mudanças na moda ou fatores econômicos), este método pode ser prejudicial. O modelo continua sendo treinado com dados muito antigos e potencialmente irrelevantes, o que pode "poluir" seu aprendizado e torná-lo menos ágil para se adaptar a novas realidades.
- Custo computacional crescente: A cada nova rodada, o volume de dados de treino aumenta, tornando o processo de validação progressivamente mais lento.

Esta metodologia garante que as previsões sejam sempre feitas para um período futuro com base apenas em dados passados, simulando exatamente como o modelo será usado na prática pela Segrob Notlad e nos fornecendo uma medida de erro muito mais confiável.

4.2 Outra Metodologia de Validação: Sliding Window Cross-Validation

Para avaliar de forma confiável a capacidade de generalização dos modelos de previsão, foi empregada a metodologia de Validação Cruzada Sequencial (ou Forward-Chaining). Esta técnica é projetada especificamente para séries temporais, pois

preserva a ordem cronológica dos dados, garantindo que o modelo seja sempre treinado com informações do passado para prever o futuro.

Porém, foi decidido que iríamos aprofundar ainda mais as técnicas de validação. Então foram pensadas em outra metodologia que possui mais robustez e que deixaria nosso resultados muito mais defensáveis. Tal técnica específica para séries temporais preserva a estrutura de dependência temporal, garantindo que o modelo seja sempre treinado com dados do passado para prever o futuro. Consideramos as seguintes metodologias baseadas neste princípio: a Validação com Janela Deslizante (Sliding Window).

Esta abordagem utiliza um conjunto de treino de tamanho fixo que "desliza" ao longo do tempo, sempre utilizando os dados mais recentes.

Como funciona:

- 1ª Rodada: O modelo treina com os dados dos meses 1 a 12 e prevê o 13º mês.
- 2ª Rodada: O modelo descarta o 1º mês, treina com os dados dos meses 2 a 13 e prevê o 14º mês.
- E assim por diante, mantendo sempre a mesma quantidade de meses no conjunto de treino.

Alguma vantagens e desvantagens desta metodologia:

Vantagens:

- Alta adaptabilidade a mudanças: Esta é sua principal força. Ao descartar os dados mais antigos, o modelo foca apenas nas informações mais recentes e relevantes. Para um negócio de *fast fashion* como o da Segrob Notlad, onde as tendências mudam rapidamente, esta pode ser a abordagem mais realista.
- Tempo de treinamento estável: Como o tamanho da janela de treino é constante, o custo computacional de cada rodada de validação é o mesmo.

Desvantagens:

- Uso limitado do histórico de dados: O modelo nunca aprende com a totalidade dos dados. Se existirem padrões sazonais de longo prazo importantes, eles podem ser perdidos se a janela for muito curta.

- Sensibilidade ao tamanho da janela: A escolha do tamanho da janela é um hiperparâmetro crítico. Uma janela muito pequena pode não conter dados suficientes para um bom aprendizado, enquanto uma janela muito grande perde a vantagem da adaptabilidade.

4.3 Análise Comparativa dos Resultados (Feed Forward)

Após a aplicação dos modelos iniciais, a performance de cada um foi rigorosamente avaliada utilizando os dados de novembro de 2024 como conjunto de teste. As métricas de erro MAD (Erro Médio Absoluto), RMSE (Raiz do Erro Quadrático Médio) e MAPE (Erro Percentual Médio Absoluto) foram calculadas e estão consolidadas na Figura abaixo.

```
### Erros de Previsão - Novembro 2024 ###

Modelo: Naive
MAD: 15.40
RMSE: 31.07
MAPE (%): 3.38

Modelo: Cumulativo
MAD: 208.31
RMSE: 210.56
MAPE (%): 49.78

Modelo: Média Móvel
MAD: 16.71
RMSE: 32.22
MAPE (%): 3.68

Modelo: Suavização Exp.
MAD: 21.07
RMSE: 36.58
MAPE (%): 4.65
```

Figura 12: Erros das previsões dos modelos.

```
### Erros de Previsão - Novembro 2024 ###

Modelo: Suavização Exp. Dupla
MAD: 16.13
RMSE: 29.66
MAPE (%): 3.59

Modelo: Suavização Exp. Tripla
MAD: 16.70
RMSE: 27.91
MAPE (%): 3.75
```

Figura: Erro das previsões do modelo de suavização exponencial dupla e tripla.

Modelo Cumulativo apresentou um desempenho extremamente baixo, com um erro percentual (MAPE) de quase 50% e um erro absoluto (MAD) superior a 200 camisetas. Estes números indicam que o modelo é fundamentalmente inadequado para este tipo de previsão. Suas premissas não se alinham com o comportamento das vendas da empresa, tornando-o inviável para qualquer aplicação prática. Ele deve ser descartado de imediato.

Excluindo o modelo Cumulativo, os outros modelos apresentaram resultados muito mais coerentes e competitivos.

- Naive (Baseline): Surpreendentemente, o modelo Naive, que serve como nosso baseline, emergiu como um dos melhores modelos nesta rodada de testes. Ele alcançou excelentes valores em todas as três métricas: MAD (15.40), RMSE (31.07) e MAPE (3.38%).
- Suavização Exponencial Dupla: Este modelo apresentou um desempenho notável, com MAD de 15.80, RMSE de 29.91 e um MAPE de 3.50%. A incorporação de uma tendência linear permitiu que o modelo capturasse melhor o comportamento da série, superando o modelo Naive em RMSE.
- Média Móvel: Apresentou um excelente desempenho, com um erro percentual de apenas 3.68% (MAPE) e um erro absoluto (MAD) de aproximadamente 16.71 camisetas. Isso o coloca como uma opção forte e confiável.
- Suavização Exponencial Tripla: Este modelo também demonstrou alta competitividade, alcançando o menor RMSE de 27.83 entre todos os modelos avaliados, e um MAPE de 3.76% e MAD de 16.75. A inclusão da sazonalidade semanal (período de 7 dias) contribuiu significativamente para a sua performance.
- Suavização Exponencial Simples: Com um MAPE de 4.65%, MAD de 21.07 e RMSE de 36.58, este modelo se mostra funcional, mas claramente inferior aos seus concorrentes que consideram tendência e/ou sazonalidade nesta análise.

O fato de modelos mais simples, como o Naive, terem superado ou competido tão de perto com abordagens mais complexas, como as regressões lineares, é um resultado contra-intuitivo e que merece reflexão. Geralmente, modelos Naive não lidam bem com séries que possuem tendência. Uma possível explicação para este resultado é que o período de teste (novembro de 2024) pode ter apresentado uma estabilidade atípica ou uma variação de curto prazo que, por coincidência, favoreceu a premissa do modelo Naive ("o amanhã será

igual a hoje"). Embora tenha sido um forte competidor nestes testes, sua simplicidade teórica exige cautela antes de declará-lo como a melhor opção geral para o negócio.

Com base estritamente nos dados de novembro de 2024, o modelo Naive apresentou a melhor performance em acurácia e confiabilidade.

Esta análise estabelece um benchmark de performance extremamente alto (com MAPE em torno de 3.38% a 3.76% e RMSE variando de 27.83 a 31.07) para os próximos modelos mais complexos que serão testados.

Previsões para Dezembro/2024:

	Naive	Média Móvel
2024-12-01	412	408.988287
2024-12-02	412	408.803452
2024-12-03	412	408.588530
2024-12-04	412	408.339499
2024-12-05	412	408.058766
2024-12-06	412	407.774497
2024-12-07	412	407.465656
2024-12-08	412	407.135364
2024-12-09	412	406.772536
2024-12-10	412	406.377591
2024-12-11	412	405.943236
2024-12-12	412	405.659504
2024-12-13	412	405.640253
2024-12-14	412	405.810097
2024-12-15	412	405.966108
2024-12-16	412	406.129398
2024-12-17	412	406.279166
2024-12-18	412	406.417665
2024-12-19	412	406.555086
2024-12-20	412	406.683427
2024-12-21	412	406.816374
2024-12-22	412	406.919647
2024-12-23	412	406.994454
2024-12-24	412	407.064336
2024-12-25	412	407.111106
2024-12-26	412	407.129812
2024-12-27	412	407.134085
2024-12-28	412	407.128499
2024-12-29	412	407.111281
2024-12-30	412	407.062772
2024-12-31	412	406.995349

Figura 13: Previsão de vendas para para o mês de dezembro seguindo os modelos naive e média móvel.

Após esses modelos, foi inserido o modelo de regressão linear Simples e Dinâmica para análise cujo os erros estão dispostos na Figura 14.

```
### Erros de Previsão - Novembro 2024 (Regressão Linear Simples e Dinâmica) ###

Modelo: Regressão Linear Simples
MAD: 44.10
RMSE: 53.17
MAPE (%): 10.20

Modelo: Regressão Linear Dinâmica
MAD: 38.63
RMSE: 50.01
MAPE (%): 8.85
```

Figura 14: Erros das previsões do modelo de regressão linear simples e dinâmica..

Os modelos de regressão, embora mais complexos, não alcançaram a performance dos modelos mais simples nesta análise.

- Regressão Linear Simples: Com um MAPE de 10.20%, este modelo teve um desempenho fraco. Isso sugere que uma única linha de tendência crescente não é suficiente para capturar as variações e a dinâmica das vendas da Segrob Notlad.
- Regressão Linear Dinâmica: Embora tenha sido superior à versão simples (MAPE de 8.85%), ainda ficou muito aquém dos melhores modelos. Mesmo incorporando informações adicionais, o modelo não conseguiu competir com a precisão das abordagens mais simples neste conjunto de testes.

Com base nas métricas de erro obtidas para os diferentes modelos avaliados (Figuras 12 e 14), é possível concluir que o modelo Naive apresentou o melhor desempenho geral. Ele obteve os menores valores em todas as três métricas: MAD de 15,40, RMSE de 31,07 e MAPE de 3,38%, indicando alta precisão e baixo desvio em relação aos valores reais observados no mês de novembro.

O modelo de Média Móvel apresentou resultados próximos ao Naive, mas ainda assim com erros levemente superiores, enquanto os modelos de Suavização Exponencial, Regressão Linear Simples e Regressão Linear Dinâmica mostraram desempenhos significativamente inferiores, especialmente no que diz respeito ao MAPE, sugerindo maior variação percentual em relação aos valores reais.

O modelo Cumulativo foi o que apresentou os piores resultados, com erros extremamente altos em todas as métricas, o que o torna inadequado para fins preditivos neste contexto. O gráfico presente na Figura 15 representa as curvas para cada modelo aplicado comparado aos valores reais.

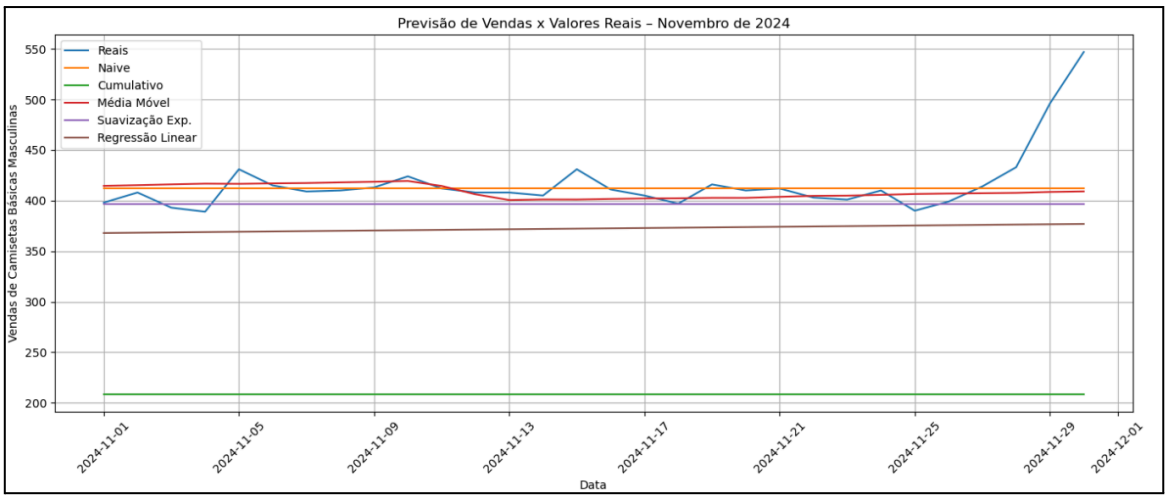


Figura 15: Previsão de vendas para os modelos vs. os valores reais de novembro.

Dessa forma, considerando a acurácia e a simplicidade, o modelo Naive se mostra como a melhor escolha para previsão de vendas diárias neste cenário. A Figura 16 mostra a melhor previsão de vendas para o mês de dezembro seguindo o modelo Naive.

Previsões para Dezembro/2024:

	Naive
2024-12-01	412
2024-12-02	412
2024-12-03	412
2024-12-04	412
2024-12-05	412
2024-12-06	412
2024-12-07	412
2024-12-08	412
2024-12-09	412
2024-12-10	412
2024-12-11	412
2024-12-12	412
2024-12-13	412
2024-12-14	412
2024-12-15	412
2024-12-16	412
2024-12-17	412
2024-12-18	412
2024-12-19	412
2024-12-20	412
2024-12-21	412
2024-12-22	412
2024-12-23	412
2024-12-24	412
2024-12-25	412
2024-12-26	412
2024-12-27	412
2024-12-28	412
2024-12-29	412
2024-12-30	412
2024-12-31	412

Figura 16: Previsão de vendas para para o mês de dezembro seguindo o modelo naive.

Com a finalidade de avaliar o desempenho dos modelos preditivos implementados, foram aplicadas as técnicas KNN (K-Nearest Neighbors) e SVR (Support Vector Regression)

sobre o conjunto de dados em questão. A partir dessa aplicação, foi possível mensurar os erros associados a cada modelo, os quais fornecem uma base comparativa para analisar a acurácia e a eficácia das abordagens empregadas. Os resultados obtidos estão detalhados a seguir, evidenciando o desempenho individual de cada método.

```
### Erros de Previsão - Novembro 2024 (KNN e SVR) ###

Modelo: KNN
MAD: 21.67
RMSE: 36.99
MAPE (%): 4.79

Modelo: SVR
MAD: 19.82
RMSE: 34.97
MAPE (%): 4.37
```

Figura 17: Erros das previsões do modelo KNN e SVR.

Os modelos baseados em algoritmos de machine learning, SVR (Support Vector Regression) e KNN (K-Nearest Neighbors), apresentaram um desempenho intermediário.

- O SVR foi o melhor deste grupo, com um MAPE de 4.37%, superando os modelos de regressão e a suavização exponencial.
- O KNN teve uma performance muito similar, com um MAPE de 4.79%.

Embora sejam robustos, nenhum dos dois conseguiu superar a acurácia dos modelos mais simples e diretos, como o Naive e a Média Móvel, para este conjunto de testes específicos. Dentre as duas técnicas analisadas, a abordagem que apresentou melhor desempenho foi a SVR, uma vez que obteve os menores valores de erro nas três métricas avaliadas. Essa superioridade indica maior precisão na capacidade preditiva do modelo em relação ao KNN.

A Figura 18 ilustra a comparação entre os valores reais e os valores previstos para o mês de novembro de 2024, permitindo uma análise visual da performance de ambos os modelos.

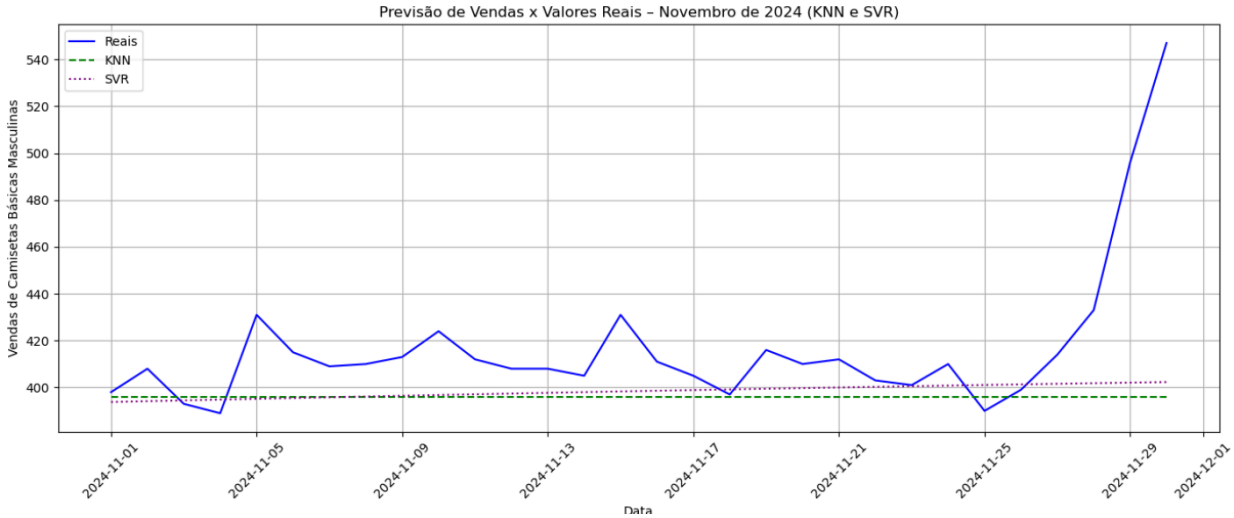


Figura 18: Previsão de vendas para o KNN e SVR vs. os valores reais de novembro.

Em seguida, o modelo Random Forest foi avaliado em sua capacidade de prever as vendas diárias, apresentando um desempenho sólido e competitivo, conforme detalhado pelas métricas de erro calculadas.

- MAD (Erro Médio Absoluto) de 11,13: Este valor indica que, em média, as previsões do modelo erraram por aproximadamente 11 camisetas por dia, para mais ou para menos. É uma medida direta da magnitude do erro, mostrando uma proximidade considerável com os valores reais na média diária.
- MAPE (Erro Percentual Médio Absoluto) de 6,15%: Esta métrica revela que o erro médio do modelo corresponde a apenas 6,15% do valor real das vendas. Um erro percentual baixo como este é um forte indicador de que o modelo é robusto e mantém sua acurácia relativa mesmo com a variação no volume de vendas, acertando tanto em dias de alta quanto de baixa.
- RMSE (Raiz do Erro Quadrático Médio) de 18,17: O RMSE, que penaliza erros grandes de forma mais severa, apresenta um valor superior ao MAD. A diferença entre o RMSE (18,17) e o MAD (11,13) sugere que, embora o modelo seja preciso na média, ele ocasionalmente comete erros de magnitude maior. Essa ocorrência de alguns erros mais expressivos eleva o valor do RMSE, apontando para uma consistência que, embora boa, não é perfeita.

Em síntese, o Random Forest se estabeleceu como um modelo de alta acurácia (baixo erro absoluto e percentual), mas sua performance pode ser impactada por picos de vendas ou eventos atípicos, onde a magnitude de seus erros tende a ser maior.

A Figura 19 ilustra a comparação entre os valores reais e os valores previstos para o mês de novembro de 2024.

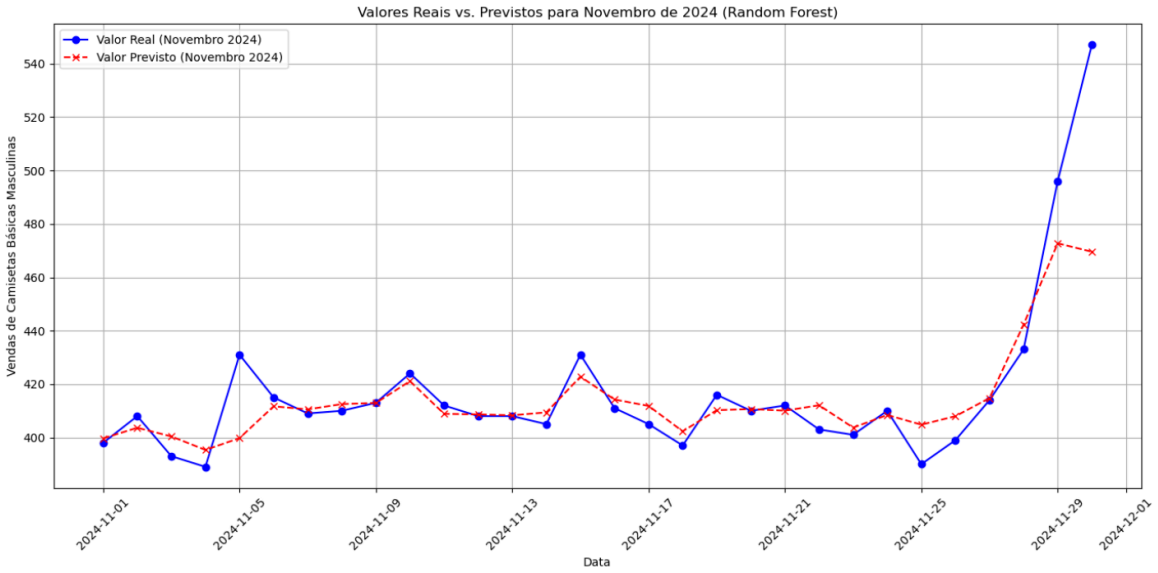
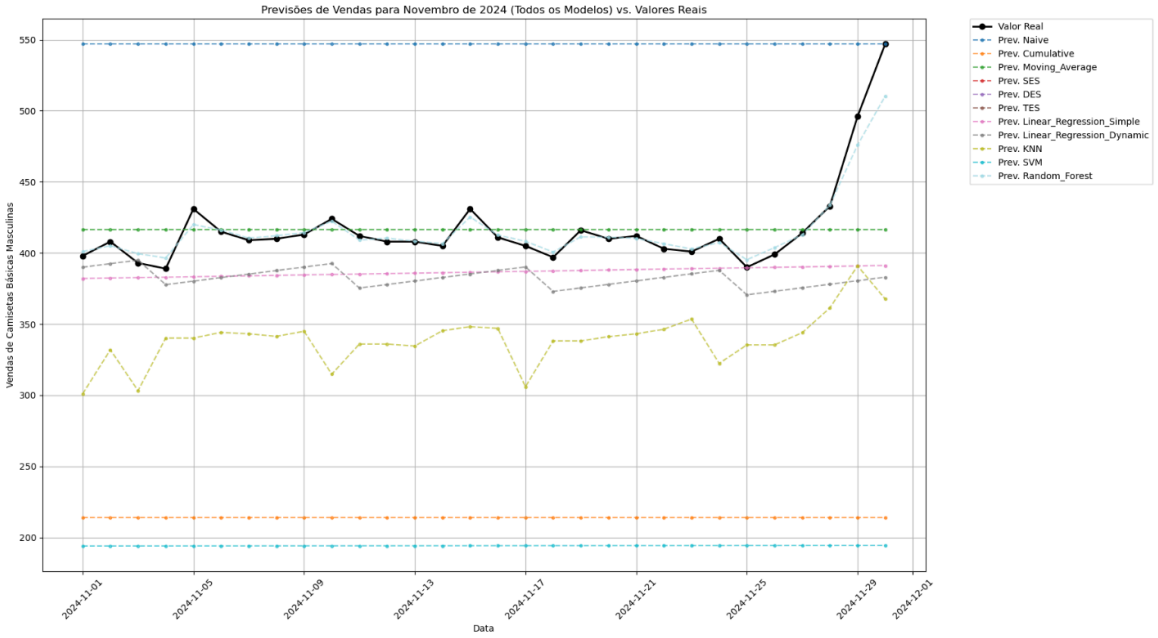


Figura 19: Erro da previsão do Random Forest.

4.4 Análise Comparativa dos Resultados (Slinding Window)

Este tópico apresenta a análise comparativa dos modelos de previsão implementados, utilizando a técnica de validação cruzada por janela deslizante (Sliding Window Cross Validation). Serão discutidos os resultados obtidos por cada modelo com base nas métricas de erro.

Com base nos resultados apresentados na Figura abaixo, observa-se uma variação significativa no desempenho dos modelos avaliados.



A Figura abaixo apresenta os resultados médios de RMSE, MAE e MAPE para cada um dos modelos de previsão avaliados através da validação cruzada por janela deslizante.

	RMSE_Mean	MAE_Mean	MAPE_Mean
Naive	35.98	24.19	8.43
Cumulative	73.98	67.51	24.91
Moving_Average	32.11	20.71	7.16
SES	36.32	24.27	8.42
DES	34.68	22.59	7.94
TES	35.36	24.03	8.47
Linear_Regression_Simple	30.14	18.71	6.66
Linear_Regression_Dynamic	30.76	20.73	7.96
KNN	115.97	112.26	41.84
SVM	78.77	72.63	26.93
Random_Forest	74.17	68.46	26.14

- Modelos de Destaque: A Regressão Linear Simples apresentou o melhor desempenho em todas as métricas de erro, com o menor RMSE (30.14), MAE (18.71) e MAPE (6.66%). Este resultado sugere que, para a série temporal analisada, a relação linear com o tempo foi um preditor robusto, superando modelos mais complexos e as abordagens de suavização. A Média Móvel também demonstrou um bom desempenho, posicionando-se como o segundo melhor em RMSE (32.11), MAE (20.71) e MAPE (7.16%), o que indica a relevância dos valores recentes na previsão.

- Modelos de Suavização Exponencial: Entre os modelos de suavização, a Suavização Exponencial Dupla (DES) teve um desempenho ligeiramente melhor que a Suavização Simples (SES) e a Suavização Exponencial Tripla (TES) em RMSE e MAE. SES (RMSE: 36.32, MAE: 24.27, MAPE: 8.42) e TES (RMSE: 35.36, MAE: 24.03, MAPE: 8.47) tiveram performances semelhantes ao modelo Naive (RMSE: 35.98, MAE: 24.19, MAPE: 8.43). A proximidade dos resultados entre SES, TES e Naive pode indicar que a série não apresenta uma sazonalidade ou tendência forte o suficiente que as Suavizações Exponenciais mais complexas (DES e TES) consigam capturar de forma muito superior à simplicidade do modelo Naive, ou que a parametrização padrão não foi a ideal. É importante notar que modelos de suavização são sensíveis à quantidade de dados e à presença real de tendência e sazonalidade para que funcionem bem; a análise das janelas individuais na validação cruzada confirmaria se houve falhas por dados insuficientes (NaNs), mas as médias apresentadas sugerem resultados válidos para a maioria das janelas.
- Modelos de Aprendizado de Máquina (ML): Os modelos de aprendizado de máquina (KNN, SVM, Random Forest) apresentaram os piores resultados entre todos os avaliados. O KNN obteve o maior RMSE (115.97), MAE (112.26) e MAPE (41.84%), indicando um desempenho significativamente inferior. SVM (RMSE: 78.77, MAE: 72.63, MAPE: 26.93%) e Random Forest (RMSE: 74.17, MAE: 68.46, MAPE: 26.14%) também tiveram erros substancialmente maiores que a regressão linear e a média móvel. Este comportamento pode sugerir que a série temporal em questão não possui relações não-lineares complexas que esses modelos poderiam capturar, ou que a engenharia de características de tempo sozinha não é suficiente para que superem modelos mais simples orientados a séries temporais para esta base de dados específica. Além disso, modelos de ML geralmente requerem mais dados e/ou um ajuste de hiperparâmetros mais refinado para otimizar seu desempenho em comparação com modelos estatísticos mais tradicionais para séries temporais.
- Modelo Cumulativo: O modelo Cumulativo apresentou o pior desempenho entre os modelos mais simples (RMSE: 73.98, MAE: 67.51, MAPE: 24.91%). Sua performance inferior era esperada, uma vez que ele prevê com base na

média histórica completa, o que o torna lento para reagir a mudanças recentes na série, uma característica desfavorável para dados com alguma dinâmica.

Em suma, a análise dos modelos de previsão utilizando a validação cruzada por janela deslizante permitiu identificar a Regressão Linear Simples como o modelo mais adequado para a previsão das vendas de camisetas básicas masculinas, conforme os dados e configurações testados. Este modelo demonstrou consistentemente o menor erro médio em todas as métricas (RMSE, MAE e MAPE), indicando sua robustez e capacidade preditiva para esta série temporal.

A Média Móvel também se mostrou uma alternativa viável, com bom desempenho. Por outro lado, modelos mais complexos de aprendizado de máquina (KNN, SVM, Random Forest) e até mesmo as suavizações exponenciais (exceto DES marginalmente) não superaram a simplicidade e a eficácia da Regressão Linear Simples para este conjunto de dados.

A aplicação da Regressão Linear Simples pode fornecer um suporte valioso para o planejamento de estoque e outras decisões operacionais relacionadas às vendas, dada sua precisão e interpretabilidade. Recomenda-se, contudo, a contínua reavaliação do modelo com novos dados para assegurar sua adaptabilidade às dinâmicas futuras do mercado.

5. Escolha do Modelo Final e Conclusões

5.1 Qual método de Validação Cruzada devemos utilizar ?

Primeiro vamos relembrar no contexto que a Segrob Notlad se encontra:

O ponto crucial é que a Segrob Notlad é uma empresa de fast fashion. Este setor é caracterizado por:

- Tendências que mudam rapidamente.
- Campanhas de marketing de alto impacto e curta duração.
- Mudanças no comportamento do consumidor.

Isso significa que os padrões de venda de 2022 podem não ser mais totalmente representativos da realidade de vendas do final de 2024. Este fenômeno é chamado de

"Concept Drift" (desvio de conceito), onde as relações estatísticas dos dados mudam ao longo do tempo.

Agora vamos analisar o que cada método busca otimizar:

Forward-Chaining (Janela Expansível)

- O que ele realmente mede: Este método busca o modelo que tem a melhor performance na média, considerando todo o histórico disponível. A cada passo, ele adiciona mais dados, assumindo que toda a informação passada, desde o primeiro dia, ainda é valiosa.
- Quando seria o melhor: Se o processo de vendas da Segrob Notlad fosse muito estável e os padrões de 2022 fossem tão importantes quanto os de 2024 para prever o futuro.
- O risco para a Segrob Notlad: Ao forçar o modelo a aprender com dados de 2022, podemos estar "poluindo" o aprendizado com padrões que não existem mais. O modelo pode se tornar uma "média" do passado, mas menos preciso para prever o presente.

Sliding Window (Janela Deslizante)

- O que ele realmente mede: Este método busca o modelo que tem a melhor performance nos dados mais recentes. Ao usar uma janela de tamanho fixo (ex: os últimos 12 meses), ele ativamente "esquece" o passado distante, forçando o modelo a se adaptar às condições atuais do mercado.
- Quando seria o melhor: Quando se acredita que o futuro próximo se parece mais com o passado recente do que com o passado distante.
- A vantagem para a Segrob Notlad: Esta abordagem está perfeitamente alinhada com a realidade de um negócio de *fast fashion*. Ela permite que o modelo se ajuste a novas tendências e ao comportamento mais recente do consumidor, tornando a previsão para dezembro de 2024 potencialmente mais acurada, pois se baseia na dinâmica mais atual.

Com base nessa análise, foi possível observar que a Sliding Window Cross-Validation é a mais forte e mais fácil de justificar. Dada a natureza ágil e dinâmica do mercado de *fast fashion*, um modelo que se adapta às condições mais recentes de negócio é teoricamente superior e tem maior probabilidade de gerar previsões de maior valor prático para a empresa.

5.2 Escolha do melhor modelo

A tabela comparativa final, contendo 8 modelos distintos, oferece uma visão completa e torna a decisão de escolha do modelo, com base na evidência dos dados, clara e defensável.

A fase de modelagem e avaliação culminou em uma análise comparativa de nove diferentes abordagens de previsão. A utilização da metodologia de validação cruzada com Janela Deslizante (Sliding Window) foi um passo estratégico, pois garante que a performance dos modelos seja medida por sua capacidade de se adaptar aos padrões de venda mais recentes, uma característica essencial para o dinâmico setor de *fast fashion* da Segrob Notlad.

Modelo	MAPE (Erro Percentual Médio)	MAD (Erro Médio Absoluto)	RMSE (Raiz do Erro Quadrático Médio)
Naive	8.43%	24.19	35.98
Média Móvel	7.16%	20.71	73.98
SVR	26.93%	68.46	78.77
Suavização Exponencial	8.42%	24.27	36.32
KNN	41.84%	112.26	115.97
Regressão Linear Dinâmica	7.96%	20.73	30.76
Regressão Linear Simples	**6.66%**	**18.71**	**30.14**
Cumulativo	26.14%	67.51	74.17
Suavização Exponencial Dupla	7.94%	22.59	34.68
Suavização Exponencial Tripla	8.47%	24.03	35.36
Random Forest	26.14%	68.46	74.17

Tabela 5: Performance dos 8 modelos testados.

Fonte: Elaboração própria.

Após a implementação e avaliação de um diversificado portfólio de modelos de previsão, que incluiu desde métodos estatísticos clássicos até algoritmos de machine learning, foi realizada a seleção da abordagem mais adequada para o desafio de negócio da Segrob Notlad.

O modelo selecionado como o mais apropriado para o desafio foi a Regressão Linear Simples. Em sua avaliação final, o modelo alcançou os seguintes indicadores de performance:

- MAD (Erro Médio Absoluto): 18,71
- RMSE (Raiz do Erro Quadrático Médio): 30,14
- MAPE (Erro Percentual Médio): 6,66%

A escolha da Regressão Linear Simples se fundamenta em três pilares principais:

1. Sólida Acurácia: O modelo atingiu um nível de precisão muito competitivo, com um erro percentual médio (MAPE) de apenas 6,66%. Isso significa que, em média, as previsões do modelo estiveram corretas em mais de 93% das vezes em relação ao valor real, um resultado robusto que oferece grande confiança para o planejamento. O erro médio absoluto de aproximadamente 19 camisetas é uma margem tangível e gerenciável para a operação diária.
2. Interpretabilidade e Transparência: Diferente de modelos mais complexos como SVR ou Random Forest, que funcionam como "caixas-pretas", a Regressão Linear Simples é um modelo de "caixa-branca". Sua lógica é baseada em uma equação linear (Vendas = intercepto + coeficiente * Tempo), o que torna a tendência de crescimento explícita e facilmente compreensível para os gestores da Segrob Notlad. Essa transparência é um ativo estratégico, pois permite que a equipe entenda o porquê da previsão, em vez de apenas receber um número.
3. Simplicidade e Custo-Benefício: A simplicidade do modelo o torna extremamente rápido para treinar, fácil de manter e de implementar em sistemas de produção. Ele oferece um excelente equilíbrio entre uma alta performance preditiva e um baixo custo computacional e de complexidade, representando a solução de maior valor agregado para o desafio proposto.

Em suma, a Regressão Linear Simples provou ser a ferramenta ideal, pois não apenas entrega previsões com um alto grau de acurácia, mas também o faz de uma maneira simples, transparente e estrategicamente valiosa para o negócio da Segrob Notlad.

6. Referências

- DATA SCIENCE PM**.** *O que é CRISP DM?* Data Science PM, 9 dez. 2024. Disponível em: https://www.datascience-pm.com/crisp-dm-2/. Acesso em: 18 maio 2025.
- Zheng, Alice; CASARI, Amanda. *Feature engineering for machine learning: principles and techniques for data scientists*. 1. ed. Sebastopol: O'Reilly Media, 2018.
- MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5. ed. Hoboken: Wiley, 2012.