Predição de Vendas Diárias em Restaurantes do Delivery Center

Anna Beatriz C. de Oliveira[1]

[1]Instituto Federal Goiano - Campus Iporá
Goiás – GO – Brasil

{anna.carlos1}@estudante.ifgoiano.edu.br

Abstract. *This work developed a daily sales forecasting model for Brazilian restaurants using the "Brazilian Delivery Center"dataset from Kaggle. The project, carried out over three weeks, applied supervised learning techniques in Python to predict the total sales value per day and per restaurant. The process included exploratory data analysis (EDA), preprocessing, feature engineering, and regression model training. Two algorithms were tested: Linear Regression ($R^2 = 0.51$) and Random Forest Regressor ($R^2 = 0.73$), with the latter showing a 43% gain in the explanatory power of the data.*
Data cleaning, variable transformation, and hyperparameter optimization techniques were applied to improve predictive performance. The results demonstrate that AI and Data Science tools can effectively support decisions in the delivery sector, enabling revenue forecasting and optimization of operational planning based on historical consumption patterns.

Resumo. *Este trabalho desenvolveu um modelo de previsão de vendas diárias para restaurantes brasileiros usando o dataset "Brazilian Delivery Center"do Kaggle. O projeto, realizado em três semanas, aplicou técnicas de aprendizado supervisionado em Python para prever o valor total de vendas por dia e por restaurante. O processo incluiu análise exploratória de dados (EDA), pré-processamento, engenharia de atributos e treinamento de modelos de regressão. Foram testados dois algoritmos: Regressão Linear ($R^2 = 0,51$) e Random Forest Regressor ($R^2 = 0,73$), sendo que o segundo apresentou ganho de 43% na capacidade explicativa dos dados.*
Técnicas de limpeza de dados, transformação de variáveis e otimização de hiperparâmetros foram aplicadas para melhorar o desempenho preditivo. Os resultados demonstram que ferramentas de IA e Ciência de Dados podem apoiar efetivamente decisões no setor de delivery, possibilitando previsão de receitas e otimização do planejamento operacional baseado em padrões históricos de consumo.Tentar novamente

1. Introdução

A previsão de vendas é uma das aplicações mais relevantes e amplamente exploradas da Inteligência Artificial no contexto empresarial moderno. A capacidade de antecipar a demanda e projetar receitas futuras é essencial para o planejamento estratégico de negócios, principalmente no setor alimentício e de delivery, que depende fortemente de fatores como sazonalidade, comportamento do consumidor e datas específicas.

Nos últimos anos, o crescimento exponencial de plataformas de entrega, como iFood, Uber Eats e Rappi, gerou uma enorme quantidade de dados sobre pedidos, horários, valores e preferências de consumo. No entanto, boa parte desses dados ainda é subutilizada por empresas de pequeno e médio porte, que carecem de ferramentas analíticas para extrair padrões e transformar informações em decisões operacionais.

Neste contexto, o presente projeto teve como objetivo desenvolver um modelo de previsão de vendas diárias que fosse capaz de estimar, com base em registros históricos, o valor total de vendas de cada restaurante em determinado dia. Para isso, utilizou-se o dataset "Brazilian Delivery Center", que contém centenas de milhares de registros de pedidos reais feitos no Brasil.

O trabalho foi desenvolvido ao longo de três semanas, cada uma com foco em uma etapa distinta do pipeline de aprendizado de máquina. Na primeira semana, definiu-se o problema e realizou-se a análise exploratória dos dados (EDA), essencial para compreender o comportamento das variáveis e preparar o conjunto para modelagem. Na segunda semana, foram aplicadas técnicas de pré-processamento e engenharia de atributos, culminando na implementação do modelo base de Regressão Linear. Por fim, na terceira semana, o modelo foi refinado, otimizado e comparado com algoritmos mais complexos, especialmente o Random Forest Regressor, que apresentou o melhor desempenho.

Além de representar um exercício acadêmico, o projeto tem grande relevância prática, uma vez que demonstra o potencial do aprendizado de máquina na previsão de tendências comerciais e na otimização de recursos. A previsão de vendas diárias permite que restaurantes planejem estoques, dimensionem equipes e criem estratégias de marketing mais precisas, reduzindo custos e aumentando a eficiência.

2. Metodologia

A metodologia adotada neste projeto baseou-se no ciclo completo de desenvolvimento de modelos de aprendizado de máquina, seguindo uma abordagem sistemática e iterativa amplamente consolidada na literatura de ciência de dados e engenharia de software. Este ciclo metodológico é composto pelas etapas sequenciais e interdependentes de entendimento do problema, coleta e preparação dos dados, análise exploratória, modelagem, validação e interpretação dos resultados.

Cada uma dessas fases foi executada com rigor científico, documentada adequadamente e revisada de forma iterativa para garantir a qualidade e confiabilidade dos resultados obtidos. A adoção desta metodologia estruturada permitiu não apenas o desenvolvimento técnico do modelo preditivo, mas também assegurou que o projeto mantivesse alinhamento constante com os objetivos de negócio estabelecidos inicialmente, facilitando a comunicação dos resultados e a eventual implementação prática da solução desenvolvida.

Todo o processo foi implementado integralmente em Python, versão 3.x, uma linguagem de programação de alto nível reconhecida mundialmente como padrão de facto para projetos de ciência de dados, análise estatística e aprendizado de máquina. A escolha do Python justifica-se por diversos fatores estratégicos: seu rico e maduro ecossistema de bibliotecas especializadas, sua sintaxe clara e expressiva que facilita a leitura e manutenção do código, sua ampla documentação e comunidade ativa, e sua adoção extensiva tanto no ambiente acadêmico quanto no mercado corporativo. Todo o processo foi implementado em Python, utilizando-se as bibliotecas:

– Pandas e NumPy, para manipulação e transformação de dados;

– Matplotlib e Seaborn, para visualizações e análises gráficas;

– Scikit-Learn, para modelagem, ajuste e avaliação dos algoritmos;

– KaggleHub, para a obtenção automatizada do dataset original.

2.1. Coleta e Composição da Base de Dados

O conjunto de dados utilizado foi obtido a partir do repositório "Brazilian Delivery Center", publicado por Nosbielcs (2021) no Kaggle [NOSBIELCS 2021], uma plataforma reconhecida internacionalmente pela qualidade e diversidade de seus datasets voltados para ciência de dados e aprendizado de máquina. Este dataset representa uma fonte valiosa e realista de informações sobre o mercado brasileiro de food delivery, refletindo operações reais de negócios neste setor. Ele contém aproximadamente 370 mil registros de pedidos realizados em diversas cidades do Brasil, distribuídas por diferentes regiões geográficas, estados e perfis socioeconômicos, cobrindo o período entre janeiro e abril de 2021, totalizando 951 restaurantes cadastrados de múltiplos segmentos gastronômicos.

A amplitude temporal do dataset, embora limitada a quatro meses, é suficientemente abrangente para capturar variações sazonais importantes, padrões de consumo ao longo de diferentes dias da semana, e eventos especiais que impactam significativamente o comportamento de compra, como finais de semana prolongados e feriados nacionais. A diversidade geográfica dos dados é particularmente relevante, pois permite que o modelo aprenda padrões de consumo específicos de diferentes regiões do país, considerando variações culturais, econômicas e demográficas que influenciam diretamente as vendas de restaurantes.

Os 951 restaurantes presentes no dataset representam uma amostra heterogênea e representativa do ecossistema de food delivery brasileiro, incluindo estabelecimentos de diferentes portes (desde pequenas lanchonetes até grandes redes), diversos tipos de culinária (brasileira, italiana, japonesa, fast-food, entre outras), variadas faixas de preço, e diferentes níveis de experiência e maturidade no mercado de delivery. Esta diversidade é fundamental para garantir que o modelo desenvolvido possua boa capacidade de generalização e possa ser aplicado a diferentes contextos e perfis de negócio.

Cada um dos aproximadamente 370 mil registros contém informações detalhadas sobre os pedidos, incluindo dados temporais (data e hora do pedido, data e hora de entrega), valores financeiros (valor do pedido, taxas, descontos aplicados), características do estabelecimento, informações sobre o processo de entrega, e outras variáveis relevantes para compreender o comportamento de compra dos consumidores. Esta riqueza de informações possibilita não apenas a construção de modelos preditivos robustos, mas também a realização de análises exploratórias profundas que revelam insights valiosos sobre a dinâmica do mercado de food delivery no contexto brasileiro.

Os principais arquivos utilizados foram:

orders.csv: contém os dados de cada pedido, incluindo identificador (order_id), data de criação (order_moment_created), valor (order_amount) e status (order_status);

stores.csv: reúne informações sobre os restaurantes, como nome (store_name), categoria (store_segment) e localização geográfica.

Após a importação via kagglehub.dataset_download(), verificou-se que os arquivos estavam codificados em “latin-1”, e foi necessário utilizar o parâmetro encoding=”latin-1”na leitura com pd.read_csv() para evitar erros de acentuação.

As primeiras linhas dos dados foram inspecionadas com pedidos.head() e lojas.head(), o que confirmou a integridade dos registros e permitiu identificar as colunas relevantes para a análise.

2.2. Limpeza e Tratamento dos Dados

Durante a inspeção inicial dos dados, realizada como parte da etapa de compreensão e validação da estrutura do dataset, constatou-se que a coluna de data identificada como ”order_moment_created”, que registra o momento exato em que cada pedido foi criado no sistema, estava armazenada originalmente no formato de texto (string). Esta representação textual, embora legível para humanos, não é adequada para operações analíticas e computacionais que envolvem manipulação temporal, como extração de componentes de data (dia, mês, ano, dia da semana), cálculos de diferenças temporais, ordenação cronológica, e agregações por períodos específicos.

Além disso, manter datas como strings impossibilita a utilização de funcionalidades nativas do Pandas especificamente projetadas para séries temporais, comprometendo tanto a eficiência computacional quanto a precisão das análises subsequentes.Reconhecendo a criticidade desta variável temporal para o objetivo central do projeto, a previsão de vendas diárias, tornou-se imperativa a conversão desta coluna para o tipo de dado datetime, um formato especializado que representa instantes temporais de forma estruturada e permite operações sofisticadas sobre datas e horários.

O tipo datetime no Pandas encapsula informações completas sobre ano, mês, dia, hora, minuto, segundo e até frações de segundo, além de oferecer suporte nativo para fusos horários, facilitando enormemente a manipulação e análise de dados temporais. Esta conversão foi executada utilizando a função ”pd.to_datetime()”do Pandas, uma ferramenta robusta e flexível capaz de interpretar múltiplos formatos de representação textual de datas e convertê-los automaticamente para o tipo datetime apropriado.

Código: pedidos[”order_moment_created”] = pd.to_datetime(pedidos[”order_moment_created”], errors=”coerce”)

A função errors=”coerce”foi empregada para forçar a conversão e transformar valores inválidos em NaT, que depois foram removidos com dropna().

Em seguida, criou-se um conjunto de novas colunas derivadas a partir da data, como:

– data_pedido (apenas a data sem hora),

– semana_pedido (número da semana no ano),

– dia_semana (nome do dia da semana),

– mes (nome do mês).

Essas variáveis foram fundamentais para capturar padrões temporais e sazonais de vendas. Também foram removidos os registros com valores nulos ou iguais a zero em order_amount, uma vez que representam pedidos inválidos ou cancelados.

Após o tratamento, os dados de pedidos e de restaurantes foram mesclados pela chave store_id, utilizando o comando:

Codígo: dados = pd.merge(pedidos, lojas, on=”store_id”, how=”inner”)

Essa junção permitiu relacionar cada pedido ao restaurante correspondente e ao seu segmento de atuação.

2.3. Análise Exploratória de Dados (EDA)

A análise exploratória teve papel central no projeto, pois possibilitou compreender a estrutura, o comportamento e as tendências do conjunto de dados antes da modelagem.

Inicialmente, o comando dados.describe() foi utilizado para gerar estatísticas descritivas. As saídas revelaram que o valor médio dos pedidos era de aproximadamente R$ 60,00, com um valor máximo acima de R$ 2.000,00, evidenciando a existência de outliers significativos.

Para visualizar a distribuição dos valores, utilizou-se o gráfico de histograma com sns.histplot(dados[”order_amount”]). A forma assimétrica da curva indicou concentração de pedidos de baixo valor e poucos pedidos muito altos, característica típica de transações de delivery. Esse comportamento motivou a aplicação da transformação logarítmica (np.log1p) sobre a variável de vendas, suavizando a influência dos valores extremos.

Outro aspecto importante explorado foi o comportamento das vendas ao longo dos dias da semana. O agrupamento foi feito com:

Código: vendas_dia_semana = dados.groupby(”dia_semana”)[”order_amount”].sum()
vendas_dia_semana.plot(kind=”bar”)

O gráfico mostrou um padrão nítido: as vendas aumentam de forma consistente entre quinta e sábado, com picos nas sextas-feiras, e caem significativamente nas segundas e terças. Essa constatação foi crucial para incluir o dia da semana como variável preditora nos modelos.

A análise por segmento (store_segment) revelou diferenças marcantes entre categorias. O segmento “food” apresentou o maior volume de vendas totais, enquanto o segmento “beverages” registrou o maior ticket médio. Esses achados reforçam a necessidade de incluir o tipo de restaurante como variável categórica no modelo.

Também foram gerados gráficos de série temporal com sns.lineplot, mostrando o comportamento diário das vendas ao longo dos meses. Foi observada uma flutuação periódica, com tendência a picos em fins de semana, caracterizando um comportamento sazonal.

Por fim, o gráfico de dispersão entre número de pedidos e valor total de vendas por restaurante mostrou uma alta concentração em poucos estabelecimentos, confirmando a Lei de Pareto (80/20): aproximadamente 20% dos restaurantes respondiam por 80% das receitas.

Em resumo, a EDA revelou um conjunto de dados fortemente sazonal, com relações não lineares entre as variáveis, presença de outliers e forte influência de fatores temporais e categóricos. Essas conclusões embasaram as próximas etapas do projeto.

3. Resultados

Após a análise e limpeza dos dados, iniciou-se a etapa de modelagem. O conjunto de dados foi dividido em 80% para treinamento e 20% para teste, utilizando a função train_test_split do Scikit-Learn.

O modelo base escolhido foi a Regressão Linear, implementada com o objeto LinearRegression(). Esse modelo serviu como referência para comparação com métodos mais avançados.

As métricas obtidas foram:

– MAE (Erro Médio Absoluto): 0,79

– RMSE (Raiz do Erro Quadrático Médio): 1,02

– R^2 (Coeficiente de Determinação): 0,51

Esses resultados indicaram que o modelo capturou parte da tendência geral das vendas, mas teve dificuldades em prever variações bruscas, principalmente em períodos de alta demanda.

Em seguida, foram testados modelos regularizados (Ridge e Lasso Regression) e, posteriormente, o Random Forest Regressor, um algoritmo de aprendizado em conjunto que combina diversas árvores de decisão.

A validação cruzada com K=5 foi aplicada para avaliar a estabilidade dos resultados, e o ajuste de parâmetros foi realizado com GridSearchCV.

Os resultados comparativos foram:

Modelo – MAE – RMSE – R^2

Regressão Linear – 0.79 – 1.02 – 0.51

Ridge Regression – 0.75 – 0.98 – 0.56

Random Forest Regressor – 0.63 – 0.84 – 0.73

O modelo Random Forest superou os demais, atingindo $R^2 = 0{,}73$, ou seja, foi capaz de explicar 73% da variação total das vendas diárias. Essa melhora decorre da capacidade do algoritmo de lidar com relações não lineares, interações complexas entre variáveis e presença de outliers, o que a regressão linear não consegue capturar.

A análise de importância das variáveis revelou que as quantidades de pedidos diários e o mês do ano foram os fatores mais influentes, seguidos pelo dia da semana e pelo segmento do restaurante. Esses resultados confirmam a forte influência da sazonalidade e do comportamento do consumidor sobre o desempenho das vendas.

Gráficos de dispersão entre valores reais e previstos mostraram uma linha de tendência próxima ao ideal (45°), e os resíduos apresentaram distribuição aleatória, sugerindo que o modelo não sofre de sobreajuste e generaliza bem os padrões aprendidos.

4. Conclusão

O projeto “Previsão de Vendas Diárias” alcançou com sucesso seus objetivos, implementando todas as etapas fundamentais de um pipeline de aprendizado de máquina: desde o entendimento e a análise dos dados até a comparação e validação de modelos preditivos.

A etapa de análise exploratória foi crucial, permitindo compreender o comportamento temporal e segmentado das vendas e revelando a presença de outliers e padrões sazonais. O modelo Random Forest Regressor apresentou desempenho superior, com R^2 = 0,73, superando em 43% o modelo base linear, e reduzindo significativamente os erros médios de previsão.

Esses resultados demonstram que a utilização de técnicas de aprendizado de máquina é eficaz para previsão de vendas no setor de delivery, podendo ser aplicada em contextos reais de gestão comercial.

Como perspectivas futuras, recomenda-se expandir o escopo do projeto incluindo variáveis externas — como condições climáticas, feriados e eventos regionais — e testar modelos de séries temporais (ARIMA, Prophet, LSTM) para capturar dependências temporais mais complexas.

Em síntese, o trabalho evidencia a importância do uso da Inteligência Artificial para a tomada de decisão orientada por dados, e consolida o domínio técnico e analítico dos autores no uso de ferramentas modernas de Ciência de Dados.

Referências

NOSBIELCS (2021). Delivery center: Food & goods orders in brazil. Kaggle.