John MacDonald
Sanchal Nachappa
Sarvesh Miskin
Suryah Vadivel
Vyshnavi Maringanti

Dataset Analysis: Target e-Commerce Sales

MIS S381N Project Report

Description of Project Goals

Dataset Description

We are investigating the Target e-Commerce Sales dataset, which focuses on Target's operations in Brazil. It covers 100,000 orders placed between 2016 and 2018 and includes detailed information on order status, pricing, payment methods, shipping performance, customer locations, product attributes, and customer reviews. The dataset is provided in eight separate files: customers, sellers, order items, geolocation, payments, orders, and products.

The primary questions we aim to address include:

- Can we forecast daily sales?
- Can we use that forecast to develop better operation plans?
- Can we develop a reproducible model that can be used for various business functions?

Importance of the Problem

This dataset represents a realistic and simplified version of transactional data commonly collected by e-commerce businesses. Analyses such as sales forecasting, customer segmentation, and market basket analysis are high-value industry practices with wide applicability.

The insights we gain from our analysis can be used to:

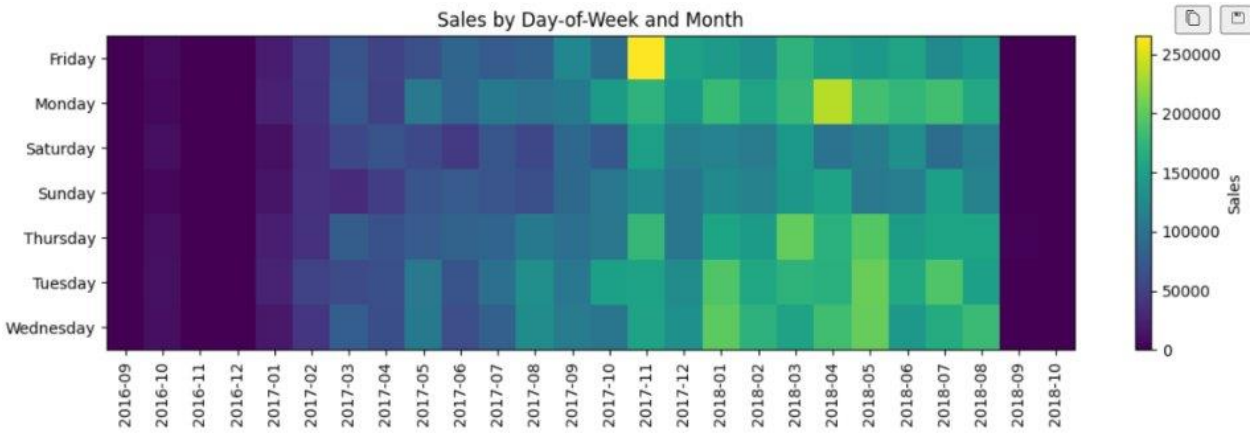
- Drive future profits
- Optimize stocking by region
- Inform targeted digital marketing strategies
- Provide scalable modeling approaches adaptable to larger datasets and additional data sources

Exploratory Analysis

Our exploratory data analysis (EDA) process included:

- Relevance Filtering – Selecting only columns critical to objectives
- Merging – Combining eight feature tables into one dataframe
- Geographic Mapping – Visualizing orders, calculating distances, and analyzing regional buying patterns
- Data Standardization – Renaming columns, fixing date formats, and addressing missing values
- Notable findings from EDA:

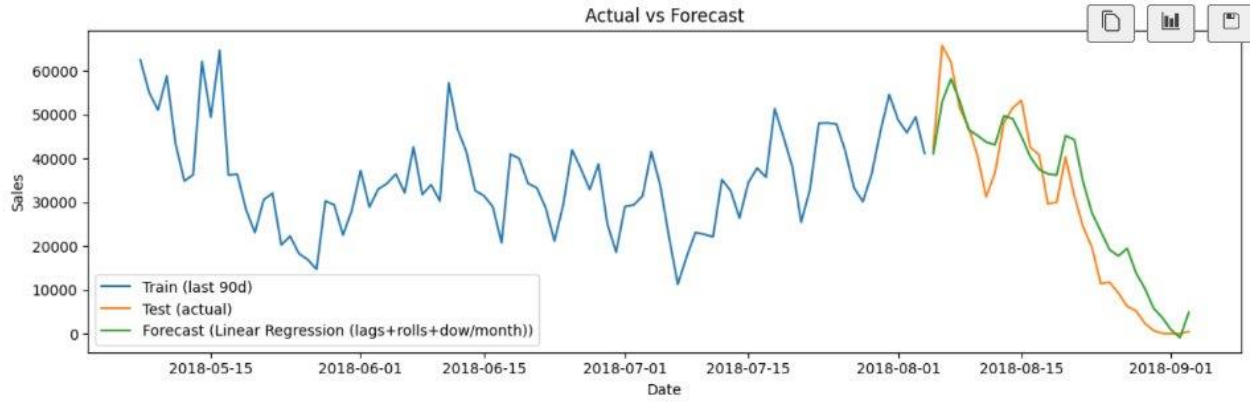
- Clear seasonal peaks in November sales
- Rolling 7-day mean highlighted short-term momentum in sales
- Summary stats: Mean daily sales ≈ 20.7k, Median ≈ 20.5k, Max ≈ 179.2k



Solution and Insights

Metrics Used: MAE, RMSE, sMAPE, WAPE
Models Tested: Naïve Model, Seasonal Naïve, Linear Regression, Gradient Boosting

The best-performing model was the linear regression approach, which achieved a mean absolute error (MAE) of $6,102 and a root mean squared error (RMSE) of $7,265. We found that even with relatively simple feature sets, the model produced significantly better MSE performance than more complex approaches, suggesting that straightforward predictive variables can capture much of the sales variation in this dataset.



Our analysis revealed several important findings. First, there is a strong seasonality effect, with sales peaking sharply in November, likely due to holiday shopping and promotional campaigns. Geographic analysis indicated

opportunities to tailor both stocking and marketing strategies to specific regions, as different areas displayed distinct purchasing patterns. RFM-based customer segmentation highlighted groups with varying recency, frequency, and monetary value, enabling more precise targeting in retention and upsell strategies.

From a forecasting standpoint, the linear regression model proved highly effective for predicting daily sales, providing an actionable tool for operational planning. Based on these results, we recommend implementing regional sales forecasting models with adjustable assumptions tailored to local market conditions. These forecasts can guide inventory allocation, ensuring that high-demand regions are stocked appropriately while avoiding excess in lower-demand areas. Insights from geographic and customer segmentation analyses should also be integrated into marketing strategies to increase relevance and conversion rates.

The predictive modeling capabilities, particularly the linear regression approach, should be embedded into operational planning workflows, with outputs provided in Excel or BI dashboards for accessibility across departments. In the longer term, expanding the modeling framework to incorporate additional features such as customer reviews, web browsing data, or promotional campaign metadata could enhance accuracy and applicability. Finally, the identified bundling opportunities and churn prediction models present avenues for increasing sales and improving customer retention, both of which could contribute meaningfully to profitability.