Amazon Sales Analysis

Victor Azih, Soren Basnet, Vincent Buchner
October 30, 2024

Abstract

This paper analyzes the "E-commerce Sales Dataset" from Kaggle to identify optimal regression models for predicting consumer spending (Amount) in Indian Rupees (INR). Of the six regression algorithms evaluated, Random Forest, Ridge, and SGD Regression models were found to perform best in predicting the Amount spent. The dataset includes 117,123 rows and 23 unique columns, with a correlation analysis revealing relationships between certain variables, such as Quantity (items per purchase order) and Fulfillment type (Amazon or Merchant), which showed a weak but positive correlation of 0.316 indicating Amazon-fulfilled orders have a higher quantity of orders. Additionally, promotion IDs and Quantity had a weaker correlation of 0.278.

Ridge Regression and SGD Regressor yielded low Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R-squared value in predicting the Amount. Although these models performed well, further exploration and parameter optimization, such as hyperparameter tuning, could enhance predictive accuracy.

Introduction

The dataset used for this analysis was sourced from Kaggle, credited to the username Anil, and represents real-world e-commerce transactions. The column names on the original dataset is as follows :

Column	Non-Null Count	Dtype
index	128975 non-null	int64
Order ID	128975 non-null	object
Date	128975 non-null	object
Status	128975 non-null	object
Fulfillment	128975 non-null	object
Sales Channel	128975 non-null	object
ship-service-level	128975 non-null	object
Style	128975 non-null	object
SKU	128975 non-null	object
Category	128975 non-null	object
Size	128975 non-null	object
ASIN	128975 non-null	object

Column	Non-Null Count	Dtype
Courier Status	128975 non-null	object
Qty	128975 non-null	int64
currency	128975 non-null	object
Amount	121180 non-null	float64
ship-city	128975 non-null	object
ship-state	128975 non-null	object
ship-postal-code	121180 non-null	float64
ship-country	128975 non-null	object
promotion-ids	128975 non-null	object
B2B	128975 non-null	bool
fulfilled-by	128975 non-null	object
Unnamed: 22	128975 non-null	object

The analysis follows a structured approach encompassing data exploration, preparation, modeling, and evaluation. Key columns in the dataset include transaction Amount, promotion IDs, fulfillment method, and various product categories. This dataset provides insight into consumer spending behavior and the factors that influence online purchases.

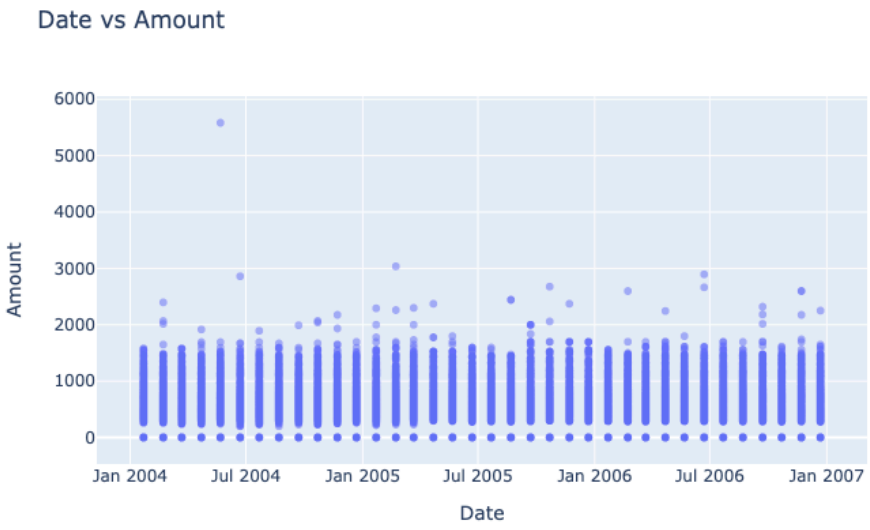
Exploratory Data Analysis

Several columns, such as : index, Order ID, Sales Channel,Style, SKU, ASIN, ship-city, ship-state, ship-postal-code, ship-country, fulfilled-by and Unnamed: 22 were excluded from analysis as they were irrelevant to the response variable (Amount). Categorical columns were encoded to streamline analysis: shipping statuses were encoded numerically (e.g., "Shipped" as 1, "Shipped - Delivered" as 2, "Canceled" as 0), fulfillment types were encoded (Amazon as 0, Merchant as 1), and shipping levels were encoded (Expedited as 0, Standard as 1).

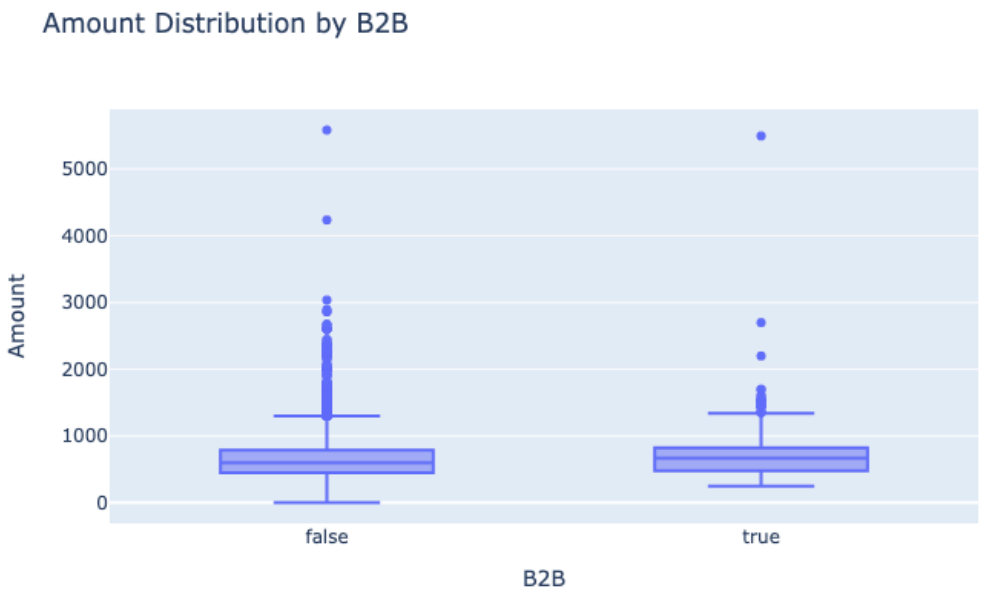
Analysis of the promotion-IDs column showed that various promotions influence pricing differently but do not necessarily correlate with significant variations in the Amount. Scatter plots reveal dense clustering around lower amounts with some outliers, suggesting that while promotions impact transaction values, they are not primary drivers of variability.

With the remaining columns the following significant exploratory plots were crafted:

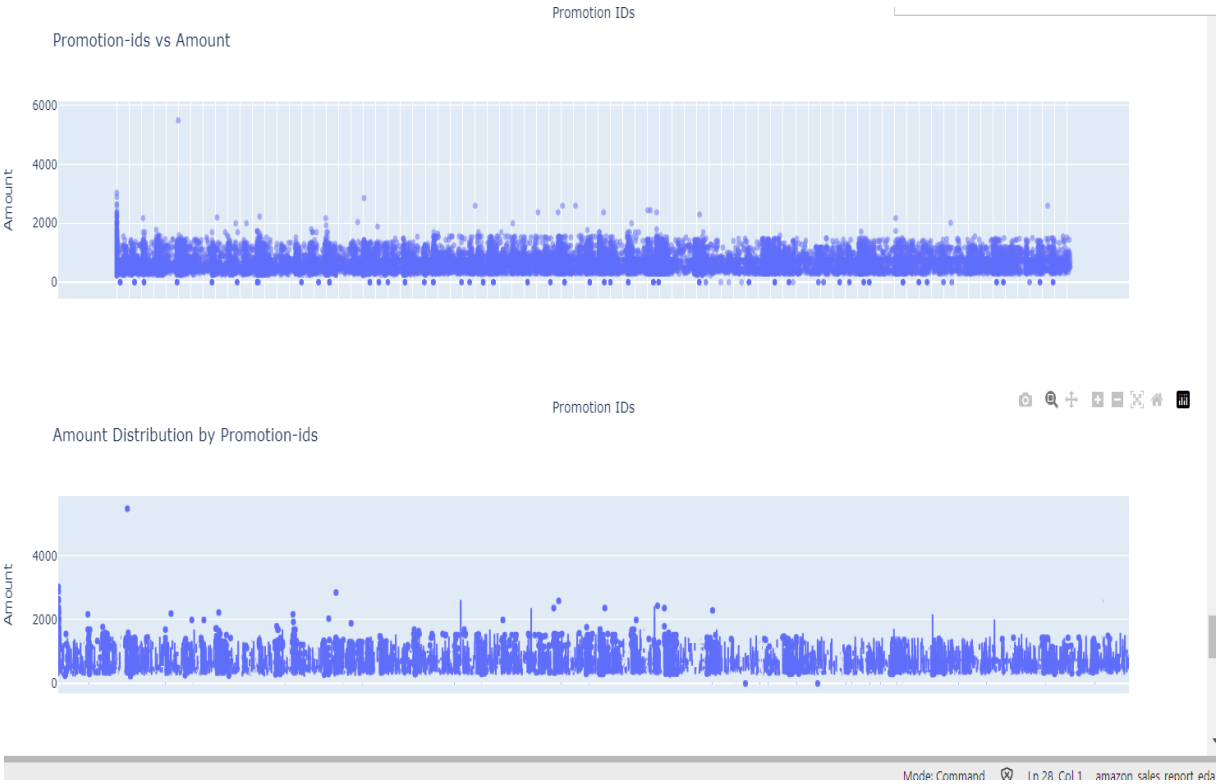
Amount spent and Date of the purchase:



Date vs. Amount (Scatter Plot): The time series scatter plot, which covers the period from January 2004 to January 2007, displays no discernible seasonal or temporal trends and a regular pattern of transaction amounts grouped within 0-2000 with sporadic outliers. This graphic is important to our research since it suggests that time may not be a reliable indicator of amount because the distribution stays mostly constant over time, indicating that our price prediction model may not need to take temporal effects into account.



The box plot, which shows median values of about 599 for non-B2B transactions with quartiles (Q1: 449, Q3: 788) and outliers up to 5584, offers statistical measurements of the amount distribution by B2B. The spread and central tendency of amounts in each category are displayed in this picture, which helps us determine whether B2B status has a substantial impact on pricing patterns and how we would need to handle outliers in our predictive modeling.



The scatter plot shows a continuous trend of dense clustering between 0-2000 with some outliers up to 5000, even when promotion IDs cause some variance in transaction amounts. It implies that promotion IDs could not have strong predictive ability because they don't show significant variations in their influence on pricing.

With constant box sizes, a large number of outliers above the whiskers, and comparable distributions and medians across promotion IDs, the box plot supports this uniform trend. This is significant since it provides more evidence that, although promotions can be incorporated as a predictor, our model probably won't use them as the main source of price fluctuations.

Analysis

Six machine learning models were tested to predict the Amount, with evaluations based on MAE, RMSE, and R-Squared scores:Through hyperparameter tuning with grid search, optimized parameters were identified for key models:

- SGDRegressor: Tuned parameters included `alpha`, `eta0`, and `learning_rate` set to 0.0001, 0.001, and invscaling, respectively.
    - Final results were: MAE = 214.78, RMSE = 279.75, R2 = 0.72.
- Ridge Regressor: Optimal parameters included `alpha = 0.1`, `fit_intercept=True`, `solver=sag`, and `tol=0.01`.
    - Final results were: MAE = 214.82, RMSE = 279.76, R2 = 0.72.
- Random Forest: Best parameters included `max_depth=10`, `min_samples_leaf=1`, `min_samples_split=2`, and `n_estimators=200`.
    - Final results were: MAE = 202.62, RMSE = 248.50, R2 = 0.75.

Conclusion

The analysis suggests that Ridge, SGD Regressors and Random Forest Regressor are strong candidates for predicting online consumer spending, as they produced low MAE and RMSE. However, the Random Forest model showed slightly improved performance after tuning. Further optimization, especially with ensemble models and additional feature engineering, could enhance the model's predictive accuracy.

Model	MAE	RMSE	R-Squared
Linear Regression	214.82	279.76	0.72
Ridge	214.82	279.76	0.72
SGDRegressor	214.8	279.97	0.72
K-Neighbors Regressor	233.77	304.02	0.67
Decision Tree Regressor	227.22	297.18	0.68
Random Forest Regressor	214.57	280.3	0.72