eCommerce Sales Forecasting

Dominique R. Grimes

July 21, 2024

Business Problem

This proposal is focused on building a predictive model to forecast retail sales based on historical sales data. Businesses must be one step ahead of the game to beat the competition. ECommerce sales forecasting can position organizations to be proactive instead of reactive. For example, business can effectively manage the number of products in stock to meet forecasted demand. Essentially, sales forecasting identifies business opportunities that would otherwise be missed. It is also utilized to anticipate and mitigate risky situations (Bosze, n.d.). Sales forecast modeling is a tool that supports a proactive sales and marketing approach to rise above the competition and avoid turning belly-up.

Background

It is not uncommon to hear the phase "History repeats itself" in casual conversation. There is more truth to that short simple statement than meets the eye. For centuries humans have inquired how our past influences our future (Nielsen, n.d.). Over time, it was discovered that the truth lies within the numbers through a method called forecasting, a type of time series analysis. "Time series analysis is the endeavor of extracting meaning summary and statistical information from points arranged in chronological order" (Nielsen, n.d.).

Data Explanation

The dataset is from Kaggle and is focused on Amazon sales from the second quarter of 2022 (Anil, 2022). There are twenty-one features and 128,975 sales entries. The raw features are outlined below:

- Category: Type of product.
- Size: Size of the product.
- Date: Date of the sale.

- Status: Status of the sale.
- Fulfilment: Method of fulfilment.
- Style: Style of the product.
- SKU: Stock Keeping Unit.
- ASIN: Amazon Standard Identification Number.
- Courier Status: Status of the courier.
- Qty: Quantity of the product.
- Amount: Amount of the sale.
- B2B: Business to business sale
- Currency: The currency used for the sale.
- Index: Index of the data in the dataset.
- OrderID: Identification number of the order.
- Sales Channel: Amazon or non-Amazon channel.
- Ship Service Level: Expedited or Standard shipping.
- Ship City, Ship State, Ship Country: Shipping address information.
- Promotion IDs: Identifies promotions applied.
- Fulfilled By: Non-Amazon Sales Channels.
- Unnamed: 22: Unknown.

The first step in preparing the data was identifying null values in the dataset. There were ten features with null values; however, I was able to drop four of those features at first glance: currency, ship-country, ship-city, ship-state. Amount and currency were related. There was only one currency (INR) in this dataset; therefore, I was able to drop the feature due to no variability.

Ship-country has a similar scenario that all values were the same. The feature ship-postal-code provides the same information as the ship-city and ship-state columns, so these two were not needed.

Table 1

Features and sum of null values

Courier_Status	6872
currency	7795
Amount	7795
ship-city	33
ship-state	33
ship-postal-code	33
ship-country	33
promotion-ids	49153
fulfilled-by	89698
Unnamed	49050

I observed the values and distribution of each feature. Unnamed only had False and Null values and was not included in the data description. Since I was not able to validate this data, I dropped the feature.

For fulfilled-by, I noticed that it was dependent on Fulfillment. Fulfillment has the values Merchant and Amazon. Fulfilled-by had the values Easy Ship or null. When Fulfillment was Merchant, fulfilled-by was Easy Ship, and when Fulfillment was Amazon, fulfilled-by was null. I made the decision to drop fulfilled-by and made Fulfillment a binary feature with 1 for Amazon and 0 for merchant Easy Ship.

That left null value management for Courier-Status, Amount, ship-postal-code, and promotion-ids features. I approached each feature with the following methods:

- Courier-Status: All zero qty rows were cancelled or null. I made the decision to replace all null Courier-Status values with cancelled status.
- Amount: I replaced the null values with the mean Amount of each category.
- Ship-postal-code: I did not observe any obvious relationships between the null postal codes and the rest of the features. I replaced the 33 null postal codes with the mode zip code. I also updated the data type from float to integer.
- Promotion-ids: This feature had a lot of text that I was unable to interpret. I decided to make this feature binary by changing any values that were not NA to 1 and filling NA with 0. The feature name was changed from Promotion-ids to promotion.

After null value management was complete, I updated several features into binary features, renamed the feature header, and dropped the original feature as shown below:

- B2B → B2b_True
- ship-service-level → ship-service-level_Expedited
- sales_channel → Sales_Channel_Amazon.in

I split the Status column on the "-" and created two new features, Status2 and Status3. Status 3 was transformed into a binary feature labeled Returned. I then filled the NA values with 0 and updated the data type from float to integer.

After reviewing all the data types of the features, I noticed Date needed to be adjusted from object to datetime. I proceeded in splitting the Date into separate Month, Day, year, week and weekday features. I dropped the original Date. I also eliminated features that had little to no variability or were unique. This allowed me to drop year, Order_ID, index, and SKU.

Analysis

I analyzed time related features through value counts and histograms. I observed that most sales take place at the beginning of the month with additional peaks at the middle and end of the month. I could interpret that this is due to common paychecks falling around these times of the month. As for day of the week, the most sales are on Saturdays and the least are on Wednesdays. When looking at the week of the year, there is a bimodal distribution with more sales in the first half of the quarter and less sales in the second half of the quarter. This dataset is specific to the second quarter which includes months April, May, and June. One could assume that potentially there is more clothes shopping in April to update wardrobes for hotter weather. As for outliers, there were no obvious outliers for these features.

Figure 1

Histograms of Date, Day of Week, and Week of Year

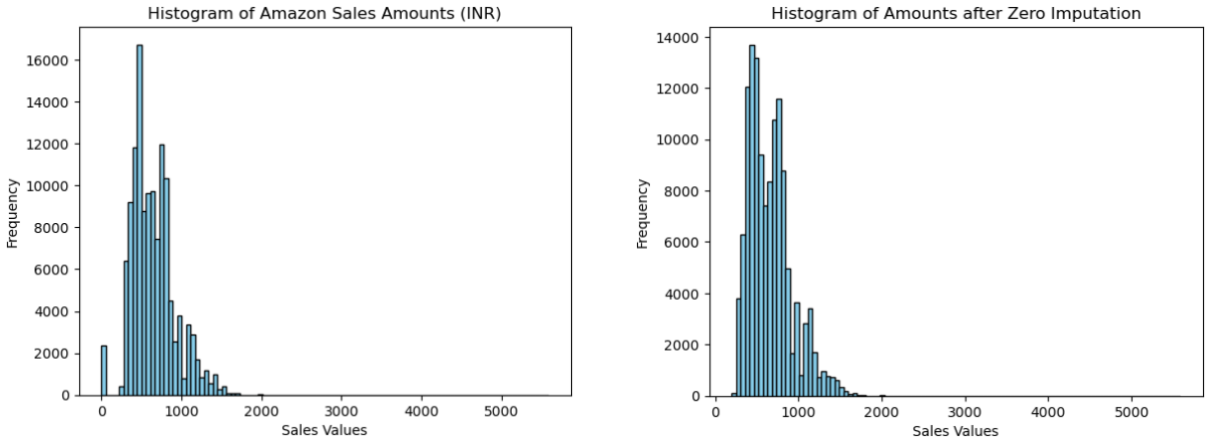
The only continuous feature was Amount. I reviewed summary statistics including count, mean, standard deviation, range and quartiles of the feature. I then plotted a histogram of sales and noticed values that were 0. I assumed this was due to cancellations and investigated further. When filtering the data frame for cancellations, there were no 0 Amounts. I filtered the data frame for zero amounts and did not observe any obvious reasons why the amount should be zero.

I decided to replace all zero Amounts with the mean values per Category like I did for the null value imputation. I also reviewed the mean values after replacing the 0 amounts. As anticipated, the values increased slightly since it was no longer skewed by the 0 amounts.

The distribution of Amount is right skewed with outliers. There are several outliers that start around 1,250 thousand dollars with a clear break at 3,000 and above. Since these are valid purchases, I am choosing not to remove the outliers currently.

Figure 2

Histogram of Amounts Before and After 0 Value Imputation



For bivariate analysis I plotted the daily total sales in date order. It generally follows the same bimodal distribution as the Week per Year histogram.

I completed CramersV correlations between several different combinations of categorical features. I have outlined the Cramer's V correlation values below:

- Status2 and Courier_Status: .73
- Style and ASIN: .99
- Style and Category: 1.0

These are all strong correlations; therefore, I dropped features Status, Status2, ASIN, and Style.

I used decomposition to look for trends and seasonality in the time series data. There was a slight downward trend. I applied the Dicky-Fuller to verify that the data was stationery with a p-value significantly less tan .05. Therefore, differencing was not needed.

Methods

I had approximately 90 days of data and decided to split the train and test data on June 15th, 2022. This left approximately two weeks of sales data in the test dataset.

For modeling, I used the features month, day, dayofweek, dayofyear, weekofyear, and Amount. I applied auto ARIMA, Random Forest Regression, Linear Regression models, and XGBoost Regression.

I evaluated the models using root mean squared error (rmse) and Weighted Mean Absolute % Error (wmape). The evaluation metric results varied from rmse of 126,488 to 211,753 and wmape from 14.0% to 24.3%. While the rmse of the auto ARIMA model is slightly higher than the Liner Regression, the wmape is more than 8% lower. Therefore, the preferred model is the auto ARIMA.

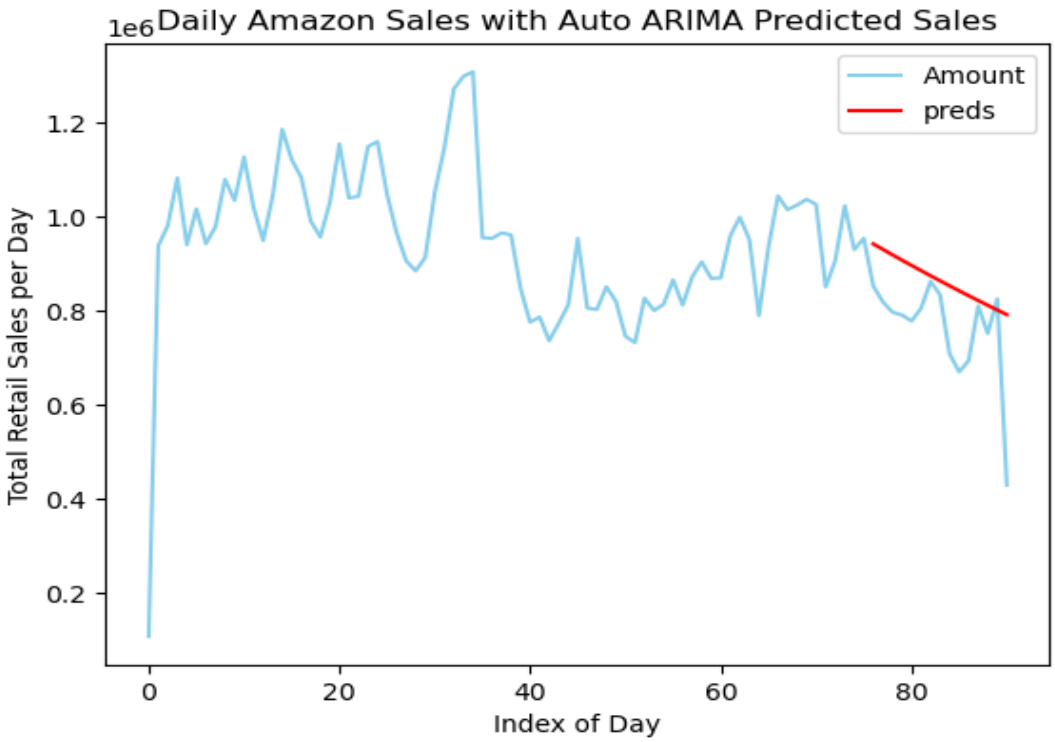
Table 2

Evaluation metric results

Model	Root Mean Squared Error	Weighted Mean Absolute % Error
Auto ARIMA	135,386	14.0%
Random Forest Regression	211,753	24.3%
Linear Regression	126,488	22.4%
XGBoost Regression	207,064	22.7%

Figure 3

Time Series Plot of Total Sales per Day



Conclusion

With the power of time series modeling, data science allows us to learn from historical patterns to forecast outcomes within a certain confidence. In this case, the second quarter 2022 Amazon sales data form India showed a downward trend with no seasonality over 90 days. This was evaluated through decomposition as well as the Dickey-Fuller test and determined to be stable. Auto ARIMA, Random Forest Regressor, Linear Regression, and XGBoost Regression were the models fit. While the models need further enhancements and tuning, Auto ARIMA was the strongest to date based on the evaluation metric wmape.

Assumptions

With this slice of data, the Dicky Fuller test determined that the data was stationary. I recommend at least one year of data to determine if the sales forecast data is truly stationary. This assumption can greatly impact model performance, especially since auto ARIMA assumes stationarity. Invertibility is one other assumption of the auto ARIMA model. Invertibility means the model's error terms can be expressed as a linear combination of current and past forecast errors (Tamplin, 2023).

Some sales amounts were imputed with the mean of the category. It was assumed that the 0 values were invalid since none of the items were cancelled. There could potentially be other reasons why there were zero Amounts such as comped purchases or certain promotions that were not captured.

Limitations

Since the data available was only one quarter of data, this limited the ability to account for seasonality in the sales. The sales data is also very specific to India sales. Regarding the model, auto ARIMA is limited to one variable for model forecasting.

Challenges/Issues

A challenge was that ten of the features had null values that had to be managed. Fortunately, the impacts were minimal since some features were dropped, others were changed to binary, and the remaining made up only a small percentage that had to be imputed.

Time series is very susceptible to outliers and anomalies. Noise in the data can have an impact on model effectiveness, making it difficult to identify the true pattern within the data.

There were no negative sale amounts for returns. This is something that was not accounted for in the model; therefore, actual sale values may be inflated.

Additional Applications

An additional way that this data can be leveraged is to predict cancellations. By predicting cancellations, sales can be targeted more accurately. This type of model would also allow to observe causes of cancellations to provide the ability to reduce cancellations. One other way is to predict the quantity of items needed in stock to support sales demands. Forecasting in general is used in many ways. This type of model can be applied to areas such as politics, weather, and sports to name a few (see Appendix for additional use cases).

Recommendations

To improve the model further, I suggest including cross validation. I'd also explore removing outliers. For cancellations these could be removed or make into negative Amounts. Additional analysis and tweaks can be made to the model through hyperparameter tuning as well.

Implementation Plan

For sales forecasting at a large corporation like Amazon, a streaming ingestion process is recommended. This allows for live observation and identification of potential forecasted risk in sales. I recommend implementing reinforcement learning on the model so that it continues to learn and improve. Reinforcement learning can help capture temporal dependencies and patterns that impact the model. Root causes for increases and decreases in sales will need to be explored to mitigate risk.

Sales forecasting results are suggested to be displayed and communicated in different ways based on the teammate's role. For example, senior leaders may need a live streaming dashboard with alerts, and front-line workers may see projections on a report at a quarterly

meeting or email communication. Disclaimers should be clear on the expected error of the forecasted results.

Ethical Assessment

Sales forecasting is foundational to any business decision that is made. This can get messy quickly. For example, careers may be at stake based on estimated forecasting that has not happened yet.

Another ethical scenario to take into consideration is eliminating bias as much as possible in the model itself. It's possible to skew the model by including or excluding certain data in modeling to sway the results. It is essential data science initiatives are not intentionally or unintentionally be steered by political views (Nau, n.d.).

References

Anil. (2022). *E-Commerce Sales Dataset.* Kaggle. https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data

Bosze, A. (n.d.). *What is an eCommerce Sales Forecast? (+ 3 Ways to Calculate It).* DOOFINDER. https://www.doofinder.com/en/blog/sales-forecast-e-commerce

Nau, Robert. (n.d.). *Statistical forecasting: notes on regression and time series.* Fuqua School of Business, Duke University. https://people.duke.edu/~rnau/ethics.htm

Nielson, Aileen. (n.d.). *Practical Time Series Analysis.* O'Reilly. https://www.oreilly.com/library/view/practical-time-series/9781492041641/ch01.html

Peixeiro, Marco. (2023). *The Complete Guide to Time Series Models.* Builtin. https://builtin.com/data-science/time-series-model

Tamplin, True. (2023). *Autoregressive Integrated Moving Average (ARIMA).* Finance Strategists. https://www.financestrategists.com/wealth-management/fundamental-vs-technical-analysis/autoregressive-integrated-moving-average-arima/

Appendix

- Forecasting number of patients
- Forecasting website visitors
- Transportation Forecasting
- Forecasting Stocks
- Forecasting Market Share
- Demand Forecasting
- Financial Forecasting
- Climate Forecasting
- Economic Forecasting
- Healthcare Forecasting
- Environmental Studies Forecasting
- Social Studies Forecasting