Leveraging Prime Day Data for Targeted Sales

|  |
| --- |
| **Abhinav Adhikari, Olivier Mizero, Nick Oelschlaeger**  University of Nebraska at Omaha  ECON 8310 – Business Forecasting  **Introduction**  The project aims to develop a three-month forecast for Amazon purchase data for Women under the age of 65 who are reportedly the sole account users at a daily level. The data comes from Harvard’s open e-commerce project and contains nearly two million purchase records from 5027 customers. The project followed a five step-development. First, the datasets produced by Harvard were merged and explored for observable trends. Then, additional features were derived from the data. This included isolating Amazon’s annual prime day sales, as well as trying to map products into more generalized categories. Thirdly, three different forecasting methods were developed on the observed records to predict 90 days into the future. Fourth, models were compared, and the best model was determined. Finally, a strategy is proposed for increasing targeted sales based on prime day purchase behavior.  **Data Pre-Processing and Exploration**  The open e-commerce data used in this project contains nearly two million records and 32 features. The data came in two files. A purchase file contained each individual purchase record from the start of 2018 to the middle of August 2024. The second file held survey records for each customer which asked for demographic information and some e-commerce opinion related questions. Purchase quantity and summed purchase prices will be used as key indicators. Date and seasonal trends will be the primary features in each of the time series models.  **Exploration**  Initial data exploration guided the team’s decision on which demographic to model. This exploration was done using both Excel and Python. The first insight derived from this analysis is that prime day has far and away the most sales of any day of the year. This was visually determined from the time-series chart of actual sales, see Figure 1 below. Prime Day sales peaked above Nov-December holiday spending across the entire dataset.  A graph showing a number of data  Description automatically generated with medium confidenceFigure 1: Time-Series chart showing the sum of purchase prices on each day over the dataset.  Data was then analyzed for distributions in subgroups using Excel’s pivot table tool. In the dataset, gender demographics was an even split female to male. We used this demographic as our first level of decision. We then split the data by age groups and filtered to see survey respondents who said they were the only ones who used their account and those who made a purchase on prime day. This left 89,079 records. Of these, 53,254 were transactions made by females, so this became our target. We decided to cut off the 65 and older age group among females because the standard deviation of price for that group was 62.33, $23 higher than the range of the rest of the age groups, which was $27-39. Thus, we continued our project aiming to forecast and market to females ages 18-65 who were the sole account user. We continued to include non-prime day customers so we could leverage comparisons in spending behavior between these two groups.  **Feature Engineering**  The first feature we engineered was a set of dates on which Amazon held their annual Prime Day event. This list was then used to create a binary feature that indicated whether a purchase was made on prime day or not, as well as whether each customer made a prime day purchase at some point in the dataset. The prime day customer feature was used in the subgroup analysis mentioned above.  As part of the process, we attempted to develop a text categorizer to map product titles to Amazon Departments using SpaCy’s textcat\_multilabel model. This approach was unsuccessful due to insufficient computational resources and the complexity of training the model on the large dataset. Future work may address these challenges by leveraging cloud-based tools or pre-trained models to improve feasibility.  To prepare the data for time series modeling, we multiplied purchase price and quantity to create a Revenue feature. We then aggregated the data by day to generate totals of revenue, quantity, prime day customer purchase quantities. From these, we calculated the proportion of sales each day which were made by customers who made a prime day purchase. These values ranged from 0.45 to 1 (Prime Days). We also isolated year, month, day of the month, and day of the week features from our ‘Order Date’ feature. These separated features were used as the time components in the Generalized Additive Model (GAM). Otherwise, the date column was set as the index for the series of daily revenue totals and used to optimize the other model types.  **Predictive Modeling**  With the determined criteria, total revenue was forecast for three months into the future. This forecast predicted daily revenue from December 21 of 2022 for 90 days. The details of the models which were considered are enumerated below.  **Exponential Smoothing**  We initially developed an exponential smoothing model. These models use weighted averages of previous values and optional trends to forecast future models. Our model was built using the version which comes in the ‘statsmodels’ package in Python. We used the built-in optimization feature to determine the level of smoothing, created additive trend and seasonal parameters, and set the seasonal period for 30 days.  The optimized model has a smoothing level of 0.1464, which is the factor for weighing each previous day’s value in the forecast. The additive trend was optimized as nominal, and the seasonal trend was calculated as 0.0328. This value is multiplied by each month’s seasonal value, which is also calculated by the model software.  Initial visual analysis of the model plotted against actual values shows forecasts lie closely to the actual value. It is more likely for the forecasted line to be above the actual values unless Amazon had a peak day in sales. A theory for this is that whenever a peak day happens, the smoothing model increases the forecast through its weighted average and overpredicts the next day. This is especially true during Amazon’s most profitable days such as prime day. The model took several days to recalibrate after these sale injections.  **GAM**  The Generalized Additive Model (GAM) was used to analyze and forecast revenue trends for a specific group of customers. This model is particularly effective at capturing non-linear relationships, making it suitable for complex patterns in factors such as year, month, day, day of the week, and the proportion of Prime customers. These features were modeled using smooth functions, which allow the GAM to adapt flexibly to variations in the data. The model was implemented using the pygam library in Python.  The GAM achieved a Mean Absolute Error (MAE) of $1,018.64, meaning the average prediction error was approximately $1,018. The Root Mean Square Error (RMSE) was $1,793.36, which reflects the influence of larger errors on the overall accuracy. While the GAM performed better than Exponential Smoothing, its error metrics were higher than SARIMAX, suggesting it was less precise for this dataset. Nevertheless, the GAM provided valuable insights into the factors driving revenue trends, such as recurring seasonal variations and shifts in customer behavior. By identifying these patterns, the GAM offered a detailed understanding of revenue dynamics, although its predictive performance could be improved with further feature tuning and data updates.  **SARIMAX**  To forecast revenue over the next three months, we also implemented the SARIMAX (Seasonal Autoregressive Integrated Moving Average Model with Exogenous Regressors) model, which combines time series patterns such as seasonality and trends with external influences from other variables. The model included three key exogenous variables: the proportion of purchases made by Prime customers (prime\_proportion), the absolute number of Prime-related purchases (Prime Purchase), and the number of Prime customers contributing to daily revenue (Prime Customer). These variables were chosen because they reflect customer behavior, particularly around events like Prime Day and holidays, which significantly impact revenue.  The model was configured to account for both short-term relationships and weekly seasonality since purchasing behavior often follows recurring weekly patterns, with notably differences between weekdays and weekends. After fitting the model to the data its performance was evaluated to ensure it captured the patterns in revenue effectively. To forecast revenue for the next three months, the model utilized exogenous variables for the same period. By extending the Prime-related data into the forecast period, SARIMAX was able to predict revenue with a greater degree of accuracy, particularly for periods with heightened activity caused by Prime customers. The model’s predictions provided both the expected revenue values and a range of confidence intervals to account for any uncertainty in the estimates.  **Interpretation of Results**  The three models namely, SARIMAX, Exponential Smoothing, and GAM, were assessed for accuracy where each model’s ability to forecast revenue for the three-month period was evaluated based on the alignment of predicted values with observed trends and their respective errors.  The SARIMAX model demonstrated superior accuracy among the three models (Figure 2). Its ability to include external variables related to Prime customer behavior gave it a distinct advantage, especially during periods of increased purchasing activity such as holidays or Prime Day. For instance, the proportion of Prime customers and the absolute number of Prime-related purchases allowed SARIMAX to anticipate sharp increases in revenue that other models could not capture. The error metrics confirmed its performance, as SARIMAX achieved the lowest prediction error with a Mean Absolute Error (MAE) of 619.15 and a Root Mean Squared Error (RMSE) of 894.41, indicating its reliability. Additionally, the residuals for this model were evenly distributed around zero, with no significant patterns, showing that the model effectively captured both historical revenue trends and external influences.    Figure 2: Comparing 3-month revenue forecasts for exponential smoothing, SARIMAX, and GAM models.  By comparison, the Exponential Smoothing model underperformed because it does not account for external variables. This limitation caused the model to systematically underestimate revenue during periods of heightened activity, such as holidays or promotions. While Exponential Smoothing can model trends and seasonality, it failed to reflect the influence of Prime-related factors, which are crucial drivers of revenue. This resulted in large errors (MAE of 1019.10 and RMSE of 2162.65) and residuals that showed a consistent pattern of underprediction during revenue spikes.  The GAM model performed better than Exponential Smoothing by capturing smooth seasonal trends. However, it struggled to balance accuracy across different periods. While GAM performed well during regular revenue periods, it tended to overpredict revenue when activity was stable and underpredict it during periods of sharp increases because it overweighed year or month. This overfitting to seasonal effects caused the model to lose precision, as it failed to incorporate external influences that drive revenue fluctuations. The error metrics (MAE of 1018.64 and RMSE of 1793.36) and residual analysis for GAM revealed larger deviations compared to SARIMAX, especially during periods of peak activity.  The residual analysis of all three models highlights the key differences in their performance (Figure 3). SARIMAX residuals were centered tightly around zero, indicating a well-fitted model. In contrast, Exponential Smoothing showed underestimation, particularly during revenue peaks, while GAM displayed a mix of overprediction and underprediction due to its overreliance on seasonal trends.  Figure 3: Residual analysis of the three models.  The SARIMAX model delivered the most accurate and reliable revenue forecasts over the three-month period. By including external factors such as Prime customer behavior, it successfully captured both the seasonal and external influences on revenue, outperforming both Exponential Smoothing and GAM.  **Recommendations**  The SARIMAX forecasted values added up to a total revenue of $611,935. After recognizing the outsized spending behavior during our initial data exploration, our recommendation is to try to leverage prime day purchase targets during the forecasted periods, especially since January-March are historically low months for sales and do not compete with the actual Prime Day. We isolated 110 target product categories by calculating which products had a higher average price point on Prime Day. We narrowed the results to include only products who had at least 3.5% of all their purchase records on prime day. That is, we wanted to get rid of frequently purchased items and focus on those which users may be more likely to abnormally buy.  The goal of these targeted products would be to have a soft marketing campaign to alert customers that prime day like deals are happening on select products. To estimate a return on this campaign, we forecasted purchase quantities for products in one of these categories. This returned an expected 866 sales. We then conservatively resampled all purchases in these categories at a 3.5 and 10% increase in quantity ten times, as well as aggressively sampling with replacement from just prime day purchases compared to all purchases. The conservative estimates returned average revenue increase of $1,500 and $4,200 respectively to the 3.5 and 10% increase. The aggressive estimates returned $23,000 and $27,000 average expected returns. These indicate single digit returns on the overall revenue generation. But that is the strength of the strategy. These products account for 6.6% of revenue from the observed data and less than 10% of available categories. So, it is a marginal target for marginal change, but one which is derived from outsized observational difference over five years of spending patterns.  **References**  Amazon. (n.d.). *The history of Prime Day*. About Amazon. Retrieved December 18, 2024, from <https://www.aboutamazon.com/news/retail/the-history-of-prime-day> |
|  |