**Forecasting on Brazilian E-Commerce Dataset**

**Riddhimoy Ghosh**

University of Nebraska at Omaha

**Context**

*Forecasting helps businesses to estimate their costs and revenue based on which they can fix their short-term and long-term strategies. Here we work with an e-commerce dataset which has information of orders made at Olist Store between 2016 to 2018 for multiple marketplaces in Brazil. This project will attempt to forecast sales orders for product categories and their growth in time.*

**Research Questions**

* A three-month forecast of future sales (numbers and figures)
* Determine the three best-selling categories and create a forecast of growth in those categories.
* Determine the fastest-growing category, and create a forecast for its growth.
* Additional insight on overall sales trend and growth over time

**Highlights and Comments on the Data Set Used:**

* The dataset spans across September 2016 to September 2018, but we do not have any orders (only place holder order purchase timestamps) in the initial months. Olist reported about the missing data in November’16 (that may also have affected December’16 data as well) to rolling out a new version of their platform. Similarly, we find the orders tapering off towards the end of August’18 and no orders for September’18. In this context, for a contiguous uninterrupted time frame to base   
     
  Graphical user interface, application

  Description automatically generated

our analysis and forecasting on, we limit our time interval for analysis between Jan’17 to August’18.

* In this project we mainly work with the *orders, order items, products* and *product category name translation* datasets to address the research questions. The orders data set gives the us the order id, order status and the order purchase time stamp; with which we can merge information from the order items dataset to get information how many different kinds of products were included in the order; and finally going one step forward we use the information from the products and product category name translation dataset to get information about the individual products in the order.
* In this context, we define sales as the number of product items ordered (that were not of the status ‘cancelled’ or ‘unavailable’) in a particular time period of interest. If we are interested in a particular product category, we count the number of order ids belonging to the same.  
    
  In this case we see that sales of 2 different products were made through the below order id. 4 units of the yellow highlighted product were sold, and 1 unit of the orange highlighted product was sold in the same order. So for this order, total sales made was 5, 1 for product orange and 4 for product yellow.  
    
    
    
  In technical terms if we map the product ids to their respective product categories and then group by the product category to count the rows ( or number of order ids) we would get the number of sales made for the product category.
* We avoided including the price per unit (*price per unit times quantity* to get the *total sales*) in our analysis and focused more the number of items to be sold, as the price at a future time point maybe different from the current records based on demand, supply and other market conditions. Given a forecast of the quantity, the domain expert can approximate the forecast of the dollar value sales made based on market knowledge.

**Forecasting Models, Intricacies and Evaluation Metric:**

* For RQ1, three-month forecast of future sales, we felt it was necessary to approach this problem as a daily sales forecasting problem. The organization needs to have an idea, of how many orders, in general, are expected to come in the ensuing months at a daily level . This helps prepare for the staffing needs to process those orders and estimate the operational costs. Based on this, short term and long term action plans and budget allocations can be made to meet target goals and define a strategy.
* For RQ2 and RQ3, where we had to focus on individual product categories (top-3 best-selling and the fastest growing respectively) the problems were approached as a monthly forecasting problem. When focusing on individual product categories, we not so much concerned about the day to day running of events. Here we want to have an optimum amount of stock present to meet the monthly sales demands with a general idea of stocks being replenished once or twice a month.
* Since we approach RQ1 as a daily sales forecasting problem, we are able to capture the elements of seasonality present in sales for say, days of week etc. But since we approach the other RQ’s as monthly forecasting problems we are unable to capture the seasonality effects since we have less than 2 years’ worth of working data (January ’17 to August’18). The forecasting models are unable to ascertain if Black Friday Sale in November is an one off event or if it truly impact the sales periodically unless it has seen it has gathered evidence from at least two such events. We try to set our best foot forward by selecting models that are appropriate in these two scenarios, more on which is discussed in the individual subsections.
* Alongside Visually inspecting the individual time series and plotting out the actual and forecasted values, we use [Mean Absolute Percentage Error](https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-percentage-error). The MAPE gives us an idea how much our forecasts differ from the actual value relative to the magnitude of the actual values on average and is not affected by scaling of the target variable. Lower the MAPE better fit is the model.
* We develop our Monthly Forecasting Models (for RQ2 & RQ3) with data from January’17 to July’18 and use August’18 as a hold out set to sanity check our forecasts with the actual sales value of August’18. This gives us an additional layer of check regarding the quality of our forecasts for the ensuing 3 months of September ’18 to November’18.

**RQ1: A three-month forecast of future sales (numbers and figures):**

* We approach RQ1 daily sales forecasting problem. Along with providing reasonably good forecasts, we are also interested in how sales are impacted by the day of the week, month and year so that data driven decisions regarding staffing and operations can be made.

* Keeping the above in mind, our candidate model of choice here is the lmultivariate General Additive Models (GAM) implemented through the [pyGAM](https://github.com/dswah/pyGAM) library.

Graphical user interface, chart

Description automatically generated

* We get a Training set MAPE: 0.44 and reasonably high Pseudo R-squared Value of 0.53.   
  All the variables are significant in forecasting sales at 5% level of significance.

A screenshot of a computer screen

Description automatically generated with medium confidence

* Now analyzing the below partial dependency plots we can can see how each of these time covariates affect the daily sales.
  + We find in that sales have been increasing as we went through the months of 2017.
  + November had a significant influence to the sales owing mainly to the Black Friday Sales.
  + Day of the Week has significant impact to the daily sales as well. Here we observe that day 0 (Monday) contributes the highest to the daily sales, and as we progress along the week the positive impact of the day of the week on daily sales declines and then starts to impact negatively on Saturdays and Sundays (day 6), hitting the lowest on day 5 (Saturday)

Graphical user interface, chart, application

Description automatically generated

**RQ2: Determine the three best-selling categories and create a forecast of growth in those categories:**

* We approach RQ2 as a monthly sales forecasting problem, as the objective of the forecast would likely address the stock levels required to meet the order requirements for these product categories and would not be directly related to the day to day staffing and operations planning.
* First we find the top-3 product categories based on the total orders over our entire working time frame of January ’17 to August’18 and plot out the monthly trends as below. Bed/Bath/Table, Health/Beauty and Sports/Leisure are our top3 categories of interest.

Chart, line chart

Description automatically generated

* Here our forecasting model of choice is the Exponential Smoothing, since we a relatively small number of monthly observations of less than two full years of each product category of interest. The assumption free Exponential Smoothing model is the go-to model in these kind of scenarios where we start of with a relatively simpler model for a short time interval into the future when we have less datapoints, and then continue to build up on our base model as we continue to gather more data points.
* As noted before, we use August’18 as a holdout set to sanity check our forecasts. We also acknowledge that at this point that this is a short term forecast into the future and we are unable to capture the seasonality as we donot have at least two full years’ worth of data, but as we gather data points in the future, we will be able to build up on this model.
* **Top Selling Product Category: Bed/Bath/Table**  
  Chart, line chart

  Description automatically generated
  + We see some upward movement in the sales peaking in November ’17 after which we see a downward patter. Our exponential smoothing model puts more weight on the recent observations as be visible with the model fit.
  + We do reasonably well with our model fits as well. Individual monthly forecasts are available in the interactive plot outputs.  
      
    Text

    Description automatically generated
* **2nd Top Selling Product Category: Health/Beauty**

Chart, line chart

Description automatically generated

* + We notice a sharp increasing trend in our data. We try to capture this pattern with an additive trend component in our forecasting modelText

    Description automatically generated
* **3rd Top Selling Product Category: Sports/Leisure**

Chart, line chart

Description automatically generated

* + We find our model to be reasonably well fit based on following the highs and lows of the actual data. Based on our sanity check on August ’18 our model should be able to generalize well with its forecasts for the 3 months in the future.

Text

Description automatically generated

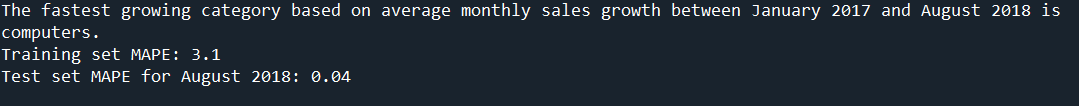
**RQ3: Determine the fastest-growing category, and create a forecast for its growth:**

* We approach RQ3 as a monthly sales forecasting problem. We calculated the month over month sales percentage changes for each product category and come up with a metric to determine the fastest growing one.
* In this context we tried metrices like the **‘median percentage change’** and the **‘most number of positive percentage changes’** and come with the top product category, but all these metrices required additional assumptions for breaking ties or set a minimum of number of monthly records filters etc. So in this case we choose the fastest-growing product category as the one with the highest **‘mean percentage change’.** *We acknowledge that this metric would be susceptible to outliers but have to proceed with it now for project time constraints.**We would strive to come with a better approach as a takeaway learning from the project.*
* *We note that ‘computers’ is the fastest growing product category based on highest* **‘mean month over month percentage change’** metric**.**

Chart

Description automatically generated

* Here also, our forecasting model of choice is the Exponential Smoothing, since we a relatively small number of monthly observations of less than two full years of each product category of interest.
* Given the high fluctuations in the sales data, our exponential smoothing model performs reasonably well based on the accuracy metric.



Chart, line chart

Description automatically generated

**Challenges, Learnings and Takeaways:**

* I felt the toughest decision needed to be made tackling forecasting problems in a time series context is defining the level of granularity of the problem. I faced dilemmas in whether to approach individual problems as daily forecasting or monthly forecasting problems. It could also have the case that all the problems would have been better approached by making them weekly prediction problems. Business Contexts and requirements would be essential here. But then also I believe how we set up the problem could make or break the results and direct the overall acceptance of the approach to the stakeholders. I hope to improve on this as I complete more projects and brainstorm about the results delivered.
* The other important aspect is defining the right metric to measure the performance of a product to answer a business question. In our case to find out the fastest growing product could define metrics like ‘most number of positive percentage changes’, ‘highest median percentage change’ and ‘highest mean percentage change’. Some of them made more sense and were more robust but gave rise to other decisions that were needed to be made. In the end we had to go for the most obvious one in spite of its flaws due to time constraints.

**References:**

* “Brazilian E-Commerce Public Dataset by Olist",. 2022. *Kaggle.com*,  
   (available at https://www.kaggle.com/olistbr/brazilian-ecommerce)