ETM 58D

Business Analytics

Project Report

Prepared By

Ceren Demirkol

Okan Güven

Sevgican Varol

I. Introduction

In this project, our task was to predict the sales quantity of 8 different products on Trendyol. We were given historical data for the products such as sold count, visit count, price and other categorical counts. We tried to fit a prediction model to forecast the next day's sold count values with historical data until the day before, in other words we made predictions two days in advance.

Products were:

- TrendyolMilla - Tights
- Oral-B - Rechargeable Toothbrush
- Koton - Coat
- Sleepy - Wet Wipes
- TrendyolMilla - Bikini Top
- Xiaomi - Bluetooth Headphones
- Fakir - Vacuum Cleaner
- La Roche Posay - Facial Cleanser

II. Approach

We read the data with the code provided; one important point to be handled before modeling was NA values for the days without sales. When a product had 0 sales; price, category_visits, ty_visits, category_sold and category_brand_sold variables were marked as NaN. We didn't want to ignore those days and remove NA values; having 0 sales on some days is also valuable to improve the model. NA values appeared as -1 when we first read the data. We turned these -1 values into NA; then performed imputation using zoo package, na.locf function. This function fills in the NA values with the closest neighbour in the same column.

At first, we analyzed the data by looking at the correlation between the features. Constructing some plots, we saw a strong relation between the sold count and features like basket count, visit count, category visit and price.

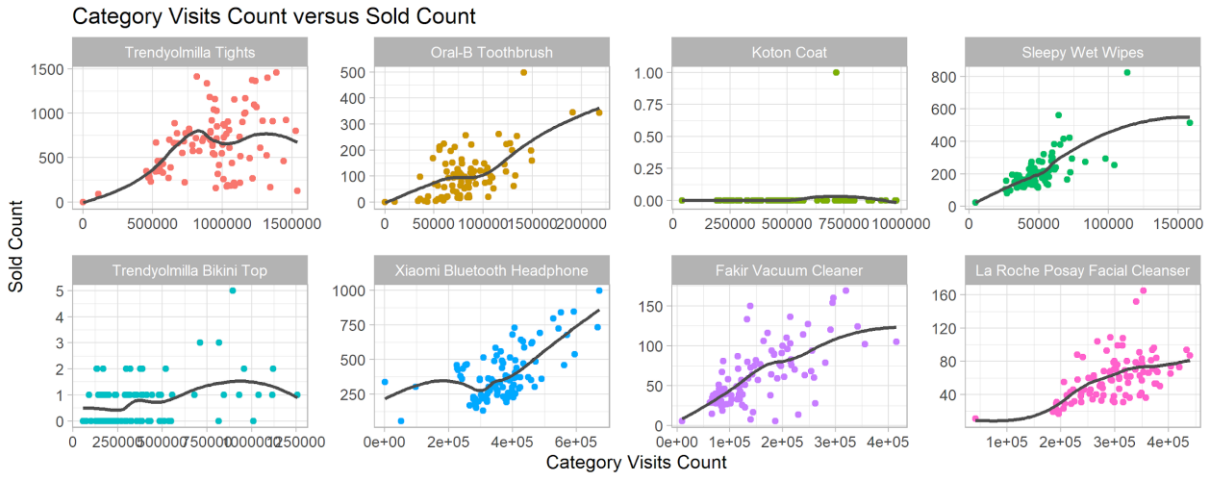

Figure 1: Category Visit Count versus Sold Count Plot

As we can see from the plots, there is a fairly straight proportionality between the sold count and the visits as well as the basket count. However, we have to keep in mind the availability (stock) of the products also affect the plots. Not just the Koton coat which was almost never in stock and therefore always zero, but the products which are temporarily out of stock as well. For the products like tights and bikini tops, which are size dependent and therefore can have a disturbed relation due to having some sizes in stock and some out of stock, they still have somewhat of a direct relation but not as much as other consumer products that can be bought anytime by anyone without the hesitation of size.

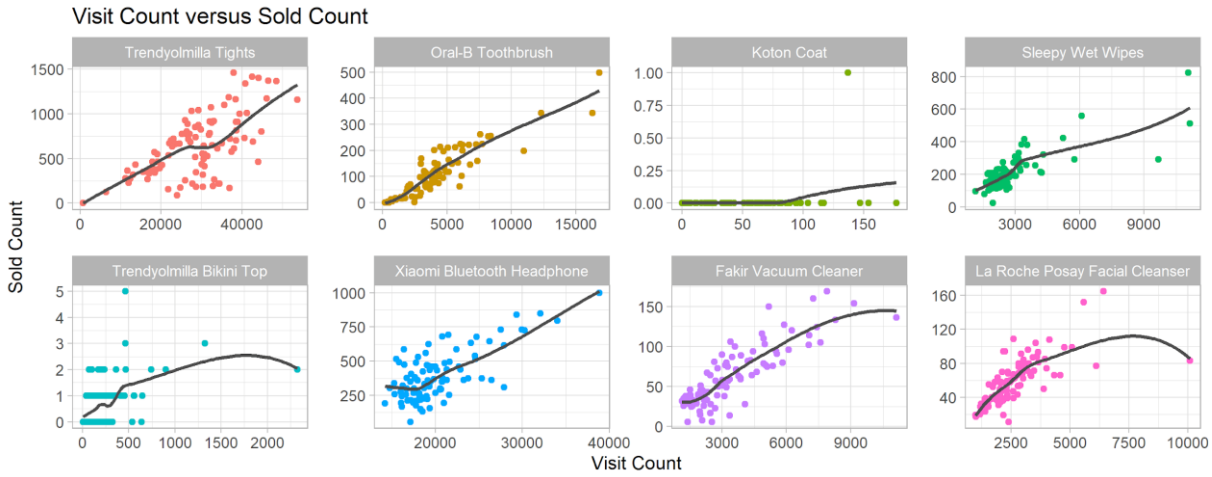

Figure 2: Visit Count versus Sold Count Plot

Even though visit count has a better relation than category visit, we can see that basket count has a sharper correlation than both of them.

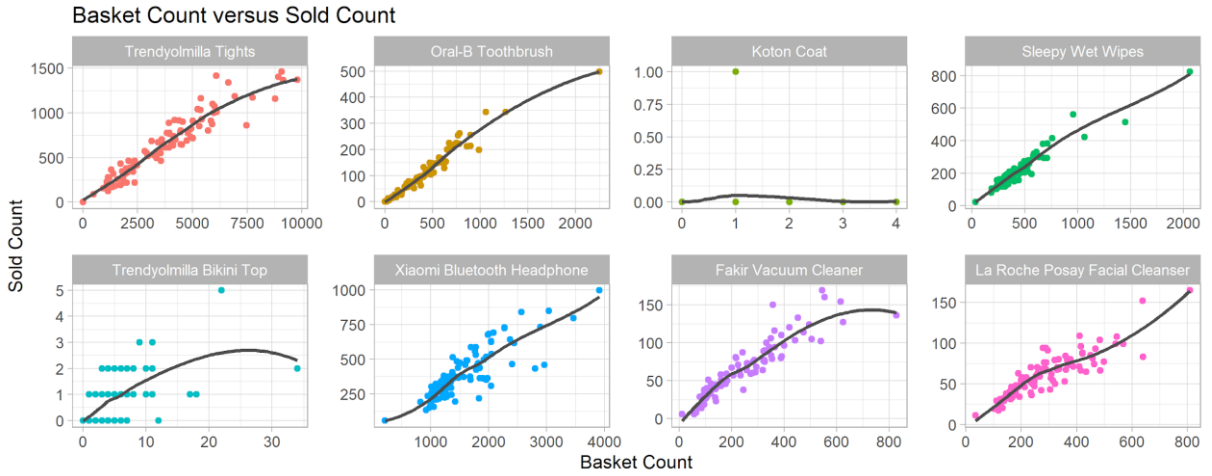

Figure 3: Basket Count versus Sold Count Plot

When looking at the price and sold count correlation, as expected we can see that it's inversely proportional. Again, excluding the Koton coat, we can still say that the decay in sold count increases as price increases. Because tights and bikini tops are seasonal products and we have few data for them, it's difficult to comment on their relation between price and sold count.

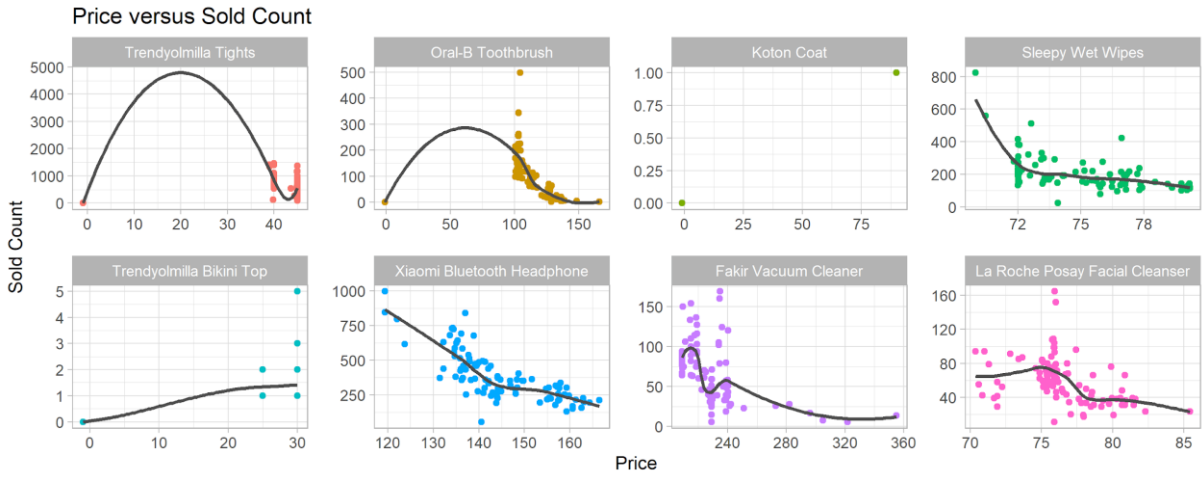

Figure 4: Price versus Sold Count Plot

At the beginning we submitted our predictions with naive approaches. We basically made our predictions based on weekly changes and took the values from the last 14 days, calculated the mean sold count value for the first and the second half, then calculated the change in terms of percentage. Then we assumed the sold count would change the same amount from the sold count value of six days ago. This method can be quite correct for seasonal products. But we didn't continue with that prediction because it was too naive and we tried different methods.

We tried to use a linear prediction method for forecasting but since there were too many parameters it was requiring preliminary prediction for other parameters before predicting the sold count values. While there are too many parameters that change the sold count of any product on e-commerce, we thought that linear prediction methods would increase the error rate.

For better estimation we wrote a code to read price value from a web site. We used the price value to see whether there is a special discount or not. If there was a discount we adjusted our sold count predictions.

In the final model, we used the forecast function from the forecast package. To use this function, we first transformed the data into time series data. Forecast function identifies the patterns in the data and makes predictions from previous patterns. Working with time series types of inputs and different models; forecast function produces forecasts accordingly.

In our case, we didn't specify the model. If the model=NULL, forecast makes forecasts using exponential smoothing state space model (in the data are non-seasonal or the seasonal period is 12 or less) or stlf (if the seasonal period is 13 or more). Stlf decomposes the time series into seasonal, trend and irregular components using Loess forecasting model.

While using the forecast package, we used two different approaches: 30-day forecast model and 90-day forecast model. Our prediction period was after Covid-19 period and into the normalization dates; we saw from our preliminary analysis that the behavior during this period was very different from the previous dates. So we looked at 2 different periods; we formed forecasts with 30-day data and 90-day data. For example for tooth brush or face cleaning products, we used a 90-day forecast but for seasonal products like bikini, we used a 30-day forecast.

III. Results

Using 30-day forecasts and 90-day forecasts, we produced our results; our predictions are combinations of two approaches. Here are our rankings:

Predicted Date	Ranking Points	Predicted Date	Ranking Points	Predicted Date	Ranking Points
16-06-2020	8	23-06-2020	7	30-06-2020	7
17-06-2020	4	24-06-2020	6	01-07-2020	7
18-06-2020	0	25-06-2020	7	02-07-2020	5
19-06-2020	8	26-06-2020	8	03-07-2020	6
20-06-2020	5	27-06-2020	4	04-07-2020	4
21-06-2020	5	28-06-2020	7	05-07-2020	6
22-06-2020	5	29-06-2020	4	06-07-2020	8

Figure 5: Prediction Rankings

While performing these analyses and predictions, we faced some challenges such as:

- We spotted that the discounts applicable in the basket are not reflected in the dataset provided, which skewed the predictions in several cases. (Ex: Bikini)
- We could not observe the price changes during the day and the discounts. We didn't know which days were the discount days other than the 'Black Friday' period. Some 'surprise' discount periods affected the results.
- Stock availability was not provided in the dataset, so we included those factors manually when necessary.
- Prediction period was after the Covid-19 period. Customer behavior changed drastically, so we didn't have the chance to interpret seasonality in a larger sense.
- As many other industries, Covid-19 period took its toll on e-commerce; increasing traffic and sales during this period affected the models and predictions as well.

IV. Conclusion and Future Work

While some products' sales were successfully predicted with the models; some products were very hard to predict or put into a trend. (earphones vs. tights)

Price sensitivities of different product categories were easily observable with the sales trends. Earphones could be an example of this, when the price of earphones dropped below 140 TL, we saw significant increases in sales.

We haven't included external data in our models; we scraped prices from the website daily but missed the price fluctuations that occured during the day. We didn't feel the need to include other external data such as temperature; our product set and circumstances would not benefit from such external data. Other products could benefit from including such external factors in the models.

Hourly sales, clickstream data, number of sellers and pricings, stock availability, discount periods, hourly stock and price changes could be included in the model to improve model accuracy.

We were able to form some solid models for sales forecasting with limited data; with additional data, more sophisticated models can be formed.