IE 360 Project – Spring 2020

Group 17

Utku Emre Karateke

Simge Bender

Batuhan Tuncel

M.Kubilay Eser

Barış Cebeci

Introduction

Ecommerce shopping is a new trend for our generation. There are many companies serving in this industry. One of them is Trendyol that wants to accurately estimate the daily sales numbers which will give it a competitive edge over its competitors. Hence this company would like to minimize the inventory holding cost and maximize customer satisfaction. The more a company holds inventory, the more it can satisfy customer demand. However, holding excessive inventory causes substantial amount of cost to the company. Online retailers try to find the optimal amount of inventory to be hold to minimize lost sales and inventory holding cost. Our task is to predict the daily sale numbers of a set of products utilizing the data Trendyol offers.

The given datasets of Trendyol consist of 8 different products; La Roche cleanser, Sleepy wet towel, Xiaomi earbuds, Fakir vacuum, Trendyol tights, Oral-B toothbrush, Trendyol bikini and Koton jacket. As of July 6th 2020, there are 432 daily observations for each dataset. For each dataset, there are 11 different variables: the price of the product (price), date (event_date), identification number of the product (product_content_id), the number of sales the product has (sold_count), the number of views the product gets (visit_count), the number of users who favors the product (favored_count), the number of users who add the product to shopping cart (basket_count), the number of products sold in the category of the product (category_sold), the number sales of products of the same brand (category_brand_sold), the number of views the category of the product gets (category_visits), the number of views Trendyol gets (ty_visits).

The quantitative and statistical models were built the come up with the most accurate forecast the future sales. Benchmark methods were conducted to see which model provides the most accurate results.

Literature Overview

Literature review was made to see the similar works done in this area of research. There are several studies conducted on forecasting the sales quantity of a product for an online retail. The first one is “A Study of Models for Forecasting E-Commerce Sales during a Price War in the Medical Product Industry (Pei-Hsuan Hsieh).” This study emphasizes that the integrity of forecasting sales in e-commerce influences merchandise inventory strategies. This case study explores forecasting models that can better inform a medical product company’s inventory plan by using the company's data that includes sales volume, number of views, company ranking and date from Amazon. Selected potential methods to try from this research are the linear trend method, the seasonal variation method and the exponential smoothing method. At the end of the study, the authors decided that it is best to use the seasonal variation method.

The second article is “A survey on retail sales forecasting and prediction in fashion markets” (Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen & Michael Teucke). The article emphasizes that sales forecasting is an indispensable task in consumer-oriented markets. This study shows different strategies models user-generated content and search queries for sales forecasting purposes exponential smoothing, ARIMA, Box & Jenkins model, regression models or Holt-Winters model, neural networks or data mining. It indicates the hybrid models should be implemented by using the above methods.

The third one is “Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach” (Alain Yee Loong, Chong Boying, Li Eric, W.T. Ngai, Eugene Ch'ng, Filbert Lee). This study explores to predict product sales by using valence and volume, free delivery and discounts, and sentiments from user reviews. It is used a big data architecture and Node.JS agents for scraping the Amazon. This study found that although valence and volume, free delivery and discounts, and sentiments from user reviews can all predict product sales, some variables are more important predictors than others and the interplay effects of these variables become more important variables than the individual variables themselves.

Approach

After loading the data into R workspace, they are transformed into time series objects to work with them in a more convenient manner. Then, every variable of every time series object was plotted to have a general idea of the data given. Finally, the models were constructed to find the most accurate forecasts for the future.

The methods to forecast the sales for each product are as follows: naive method, seasonal naive method, Holt's method, simple exponential smoothing, mean, ARIMA (autoregressive integrated moving average), TBATS (exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components), ETS (error, trend, seasonality), multiple linear regression, and multiple linear regression with selected independent variables.

It should be mentioned that it is assumed the length of the test set is 20 days although daily forecasting is 2 days. As long as the train set is the same, 2 days forecasting and 20 days forecasting will give the same result for the first 2 days. 20 days forecasting is used to forecast sales number because the inventory number is unavailable. Therefore 20 days forecasting gives a better idea about how the sales number is going to be and it also gives a better MAPE value most of the time.

The performance of a model is measured by its mean absolute percentage error (MAPE). The formula of MAPE is as follows:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where, M is MAPE, A_t is actual data point and F_t is the forecasted value.

The methods mentioned above were implemented for each product except for two. There was only 1 Koton jacket in the inventory that was sold on April 26th, after that no user ordered that product. It is concluded that naive method serves the best purpose for this case. The second case is Trendyol bikini which had smaller demands because of availability in stock. Inventory quantities are not provided, however after checking Trendyol, it was evident that bikinis were usually not available for purchase, so the models failed to return promising results. Seasonal naive method was chosen for this product, since it gives the minimum MAPE of 100.

The R library *GGally* provides a function *ggcorr* which calculates the correlation between every variable in a time series object. It is assumed that, if there is a set of variables that are strongly correlated with the sales number, it might be useful to build a multiple linear regression model using only those variables as predictors. If the correlation between a variable and the sales number is 0.6 or greater, that variable was moved to the set of predictors.

Results

For Fakir vacuum, the mutiple linear model with every variable as predictors turned out the be the best performing model, although sales number is strongly correlated with basket_count, favored_count, visit_count, and category_visits. Its MAPE was 16.11.

For Xiaomi earbuds, multiple linear regression model with 3 predictors consisting of basket_count, favored_count, and visit_count which have a strong correlation with sales number gave the best MAPE which is 10.06.

For Sleepy wet towels, multiple linear regression model with 4 predictors consisting of basket_count, favored_count, visit_count, and category_visits yielded the minimum MAPE of 14.97.

The sales number of La Roche cleanser is best forecasted by multiple linear regression with 5 different predictors: basket_count, favored_count, visit_count, category_visits, and ty_visits. The model has a MAPE of 18.63.

Trendyol tights is best forecasted by multiple linear regression with different 3 different independent variables: basket_count, favored_count, and visit_count. The model has a MAPE of 17.98.

For Oral-B toothbrush, multiple linear regression with 4 different predictors consisting of basket_count, favored_count, visit_count, and category_visits resulted in the smalles MAPE value of 57.30.

The correlation matrices for each time series objects of products can be found in Appendix.

NOTE: You can see the related codes about the project above

https://github.com/BU-IE-360/spring20-utkuemre/blob/master/360-Project.R

Conclusions and Future Work

Most of the methods taught in this course and the other courses at DataCamp deal with time series objects with only one variable other than time index. It is expected to see multiple linear model beats other methods mentioned above since the time series objects in this project have 11 variables some of which have a strong correlation with the sales number

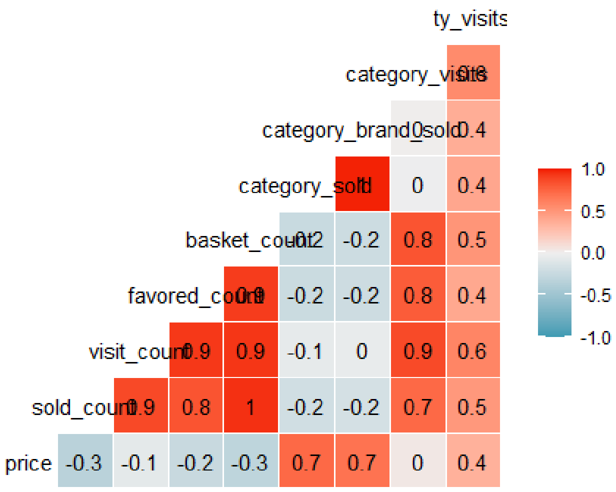
There was no significant promotions or holidays to take into consideration to come up with a better forecasting models. In addition, competitor activities such as discounts and advertising were not available to utilize.

The fact that most of the products were inelastic, their demands did not decrease when the price goes up. It might be explained by Trendyol's personal and general advertising. It is safe to assume pushing a product on the frontpage, or sending a phone notification have a great impact on the sales. The lack of those sort of information makes the forecasts more inaccurate.

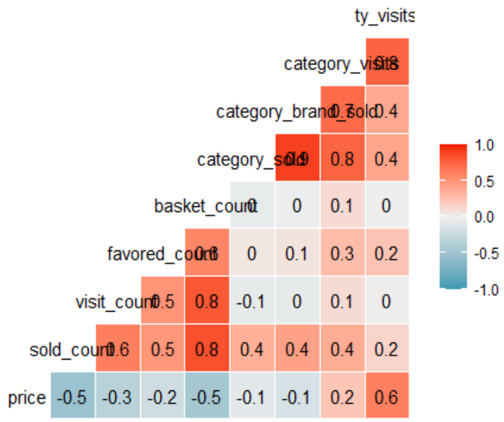
References

1. Hsieh, P.-H. 2019. A Study of Models for Forecasting E-Commerce Sales During a Price War in the Medical Product Industry. In: HCI in Business, Government and Organizations. eCommerce and Consumer Behavior. Springer International Publishing, 3–21.
2. Beheshti-Kashi, S., Karimi, H.R., Thoben, K.-D., Lütjen, M., and Teucke, M. 2014. A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering 3, 1, 154–161.
3. Chong, Alain & Ngai, Eric & Ch'ng, Eugene & Li, Boying & Lee, Filbert. (2015). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. International Journal of Operations & Production Management.

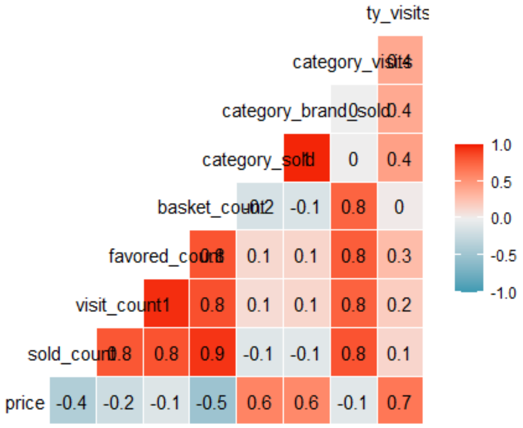
Appendices



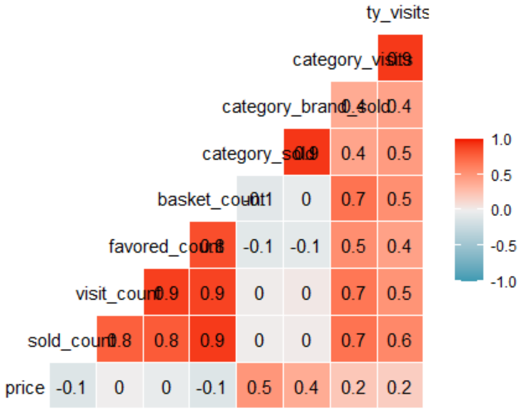
1. Figure 1: Correlation Diagram of Fakir Vacuum



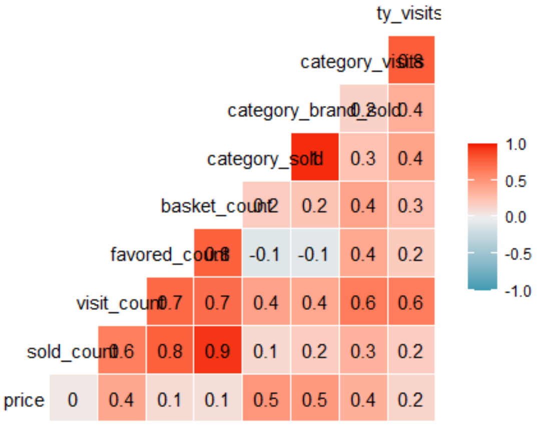
2. Figure 2: Correlation Diagram of Xiaomi Earbuds



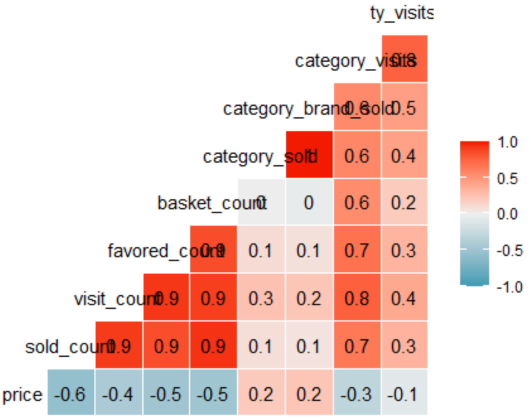
3. Figure 3: Correlation Diagram of Sleepy Wet Towel



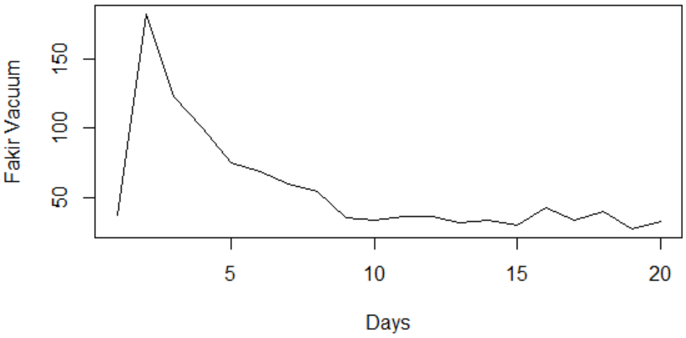
4. Figure 4: Correlation Diagram of La Roche Cleanser



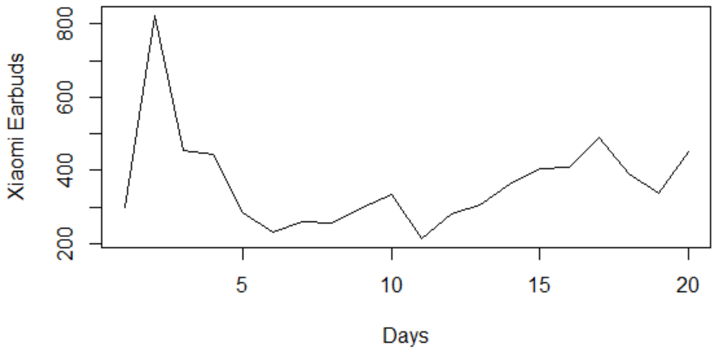
5. Figure 5: Correlation Diagram of Trendyol Tights



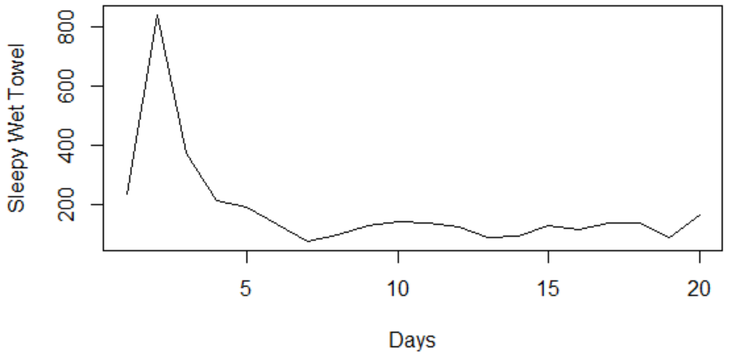
6. Figure 6: Correlation Diagram of Oral-B Toothbrush



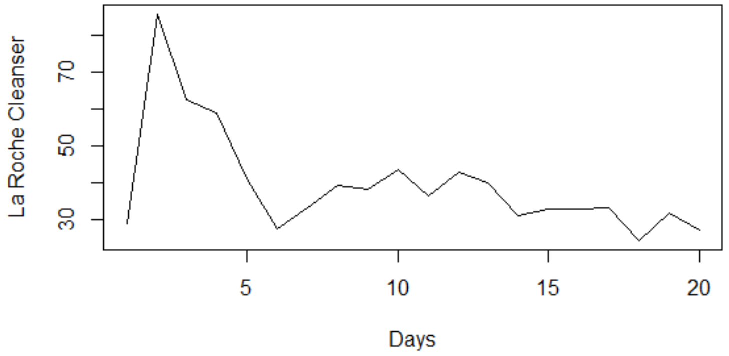
7. Figure 7: Plot of Fakir Vacuum



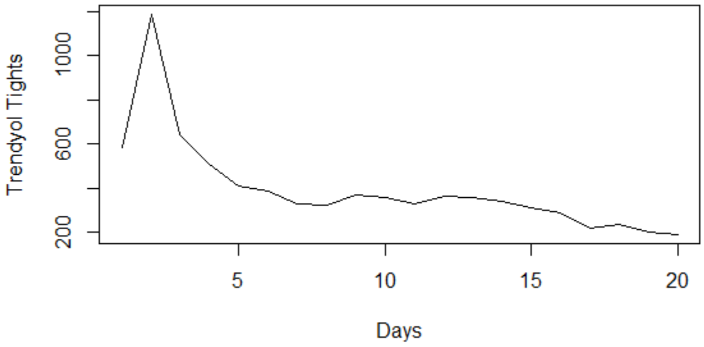
8. Figure 8: Plot of Xiaomi Earbuds



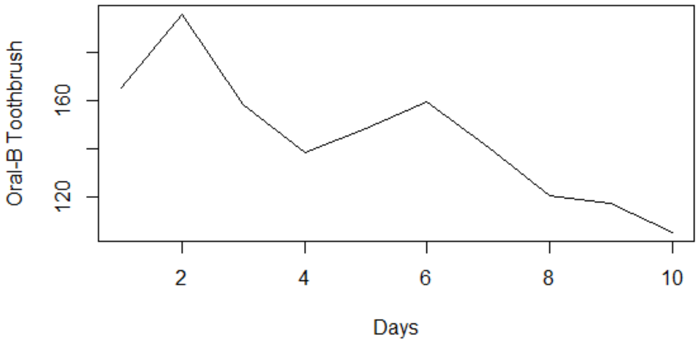
9. Figure 9: Plot of Sleepy Wet Towel



10. Figure 10: Plot of La Roche Cleanser



11. Figure 11: Plot of Trendyol Tights



12. Figure 12: Plot of Oral-B Toothbrush