IE 360 TERM PROJECT

GROUP #10
Hazan Akbaş-2016402051
Berkay Altınel-2015402159
Aslı Atakan-2015402045

Instructor: Mustafa Gökçe Baydoğan

Table of Contents:

1. Introduction

This is the term project for IE360 class which was conducted by the cooperation with Trendyol. The members of the Group #10 are Hazan Akbaş, Berkay Altınel and Aslı Atakan. The task was to predict the daily sales quantity of 8 products based on the historical data which were provided by Trendyol itself and while doing that compete with other groups to obtain the best predictions in terms of the weighted mean absolute percentage error (WMAPE). It should be also mentioned that the latest data provided was 2 days prior to forecast day and the goal was to acquire proximate forecasts throughout the submission period. Daily success was not enough since it was an ongoing project that continued between May 15, 2020 and June 12, 2020. That is why each team had a chance to observe their success and alter their predictions which was quite significant to become successful overall.

The data includes 8 different products along with the daily sales quantity, visit count, favored count, basket count, category sales quantity, brand sales quantity, category visits and overall trendyol visits. There were 3,2,2,1 products in textile, supermarket, electronics and cosmetics categories correspondingly . To approach a solid forecasting method, preliminary analysis was necessary to detect most relevant features for each product.

Success of the project depends mainly on the ability to analyze and interpret the historical data, the overall knowledge about modeling, success in selecting and defining the features to be used, patience to observe and update the forecasting method, and last but not least the collaboration between the members of the group.

At first, the agenda was decided. Clearly it was necessary to build an agile modeling and coding strategy to reach overall success. The strategy was to instigate the model and try distinct methods in the first week of the competition, then verify the method in the following period. There was some simple but important information in the given data. The sales quantity of ‘bikini’ and ‘mont’ were almost zero since the beginning of this year. That is why, instead of focusing on the predictions of these two products, a simple moving average model is applied for them. Concentration was put on the remaining products that have high sold count and the alterations between each day in general for the first step.

In order to decide which approach will be used in the forecasting process, the following methods were tried at first Linear Regression, Moving Average, ETS, ARIMA and Dynamic Linear Regression. Since it is easier to build moving average, ETS and ARIMA models; the first week of the submissions were conducted by using these approaches. However, after the validation process on historical data based on AICc and BIC values, the Dynamic Linear Regression method was selected as the forecast method.

One of the most important parts of the project after selecting Dynamic Linear Regression as the method, was to decide which features to use for each of the products. One of the reasonable features would be lagged sold counts and the initial Dynamic Regression Model was built on top of it. However, with suggestions of Prof. Baydoğan, we decided to use conversion rates as exogenous variables rather than sold count itself.

Correlation analysis was carried out to select the features to use for each product. It was expected that the correlated factors would differ for distinct products which ended up to be accurate. After correlation analysis the exogenous columns that were going to be included in the model were selected for different products.

2. Related Literature

The main resource that is used throughout the project was DataCamp and lecture notes of Prof Baydoğan. The assignments that were conducted as a lecture enhancement was quite helpful in finding the best model that fits the data. The examples used in these assignments were very explanatory and they were very compatible with the data given used for the project. In other words, they were great examples to learn the forecasting methods that are used for the kind of analysis that the project aims to teach us. Moreover, the lectures were also helpful in understanding the path to follow expected in the project. After gathering those class knowledge some other research about "Forecasting methods" were conducted as well; however, the methods that were found are quite similar to the methods which were discussed in the DataCamp and class. ETS and ARIMA models are the ones that all the group members are quite familiar with. On the other hand, Dynamic Linear Regression is a little bit more complex and complicated. The difference between static and dynamic models were clarified by the literature review. It was acknowledged that Dynamic Linear Regression authorizes temporary alterations in the properties of under researched data (Pagoswki et al., 2006). Verifying that the data includes features necessary to conduct Dynamic Linear Regression, we ended up using Dynamic Linear Regression which also had the most accuracy in the historical data.

3. Approach

As a first step, the time frame that was going to be used for making predictions had to be decided. After an exploratory analysis, we decided to use the last 60 days of the data since as we go further to the past than this, we start to observe some inconsistencies. Moreover, we have to remember that 60 days before the first submission period was 12 March, which is also the start of the week that the first Coronavirus cases were observed in Turkey. It was clear that the routine of people and perspective through e-commerce companies started to change since that week. That is also why we chose 60 days as the limiting day.

Since the project required submission throughout a monthly period, deciding on the model to use based on latest data would make it inconsistent throughout the timeline. That is why the model should be rock solid. To reach this, it is argued that the approach might be to try different methods for obtaining the best method. In order to reach that goal different methods such as ETS and ARIMA were utilised at the beginning of the competition, but after some time and gathering the class knowledge together with the raw data, Dynamic Linear Regression was decided as the method that should be used in that project.

The first thing to do before starting to use a model is have a general examination of the data. The initial thing to do is look for seasonality in the data. It appears that there is a correlation between some days of the weeks, meaning we can assume weekly seasonality even if it is not a very strong one. Once the weekly seasonality is accepted, it is decided to include some nature of the seasonality into the outlier detection. In order to replace the detected outliers, a method is developed so that its value would be calculated with respect to the previous observation, the next observation and the ratio of the previous weeks' observation. To exemplify, say an observation on a Tuesday is detected to be an outlier. This value will be replaced with the newly calculated value. First, the ratio of last weeks' observation on Monday to last week's observation on Tuesday is calculated. This value is multiplied by this week's observation on Monday and a first prediction for the outlier is created. Then, the ratio of last weeks' observation on Wednesday to last week's observation on Tuesday is calculated. This value is then multiplied by this week's observation on Wednesday and a second prediction for the outlier is created. The final value that the outlier is replaced with is equal to the average of these two values.

The next step after spotting the seasonality is to decompose the data into its components which are Trend, Seasonality and Error. In order to carry out the earlier predictions, the trend component of the data is forecasted with linear regression, the seasonality component has already been spotted and the error component is forecasted using the ARIMA method. These components are then combined and the resulting prediction is submitted. This method will also be used later when the exogenous variables selected for the Dynamic Linear Regression are being forecasted.

Selecting the method is a good start; however, in order to run the model efficiently, the features should be selected carefully as well. One of the strongest sides of the Dynamic Linear Regression is its ability to use more than one exogenous variable in the model. With the development of the model, the initial exogenous variable was set to lagged sold count for simplicity, however, using conversion rates would possibly provide more accurate results.

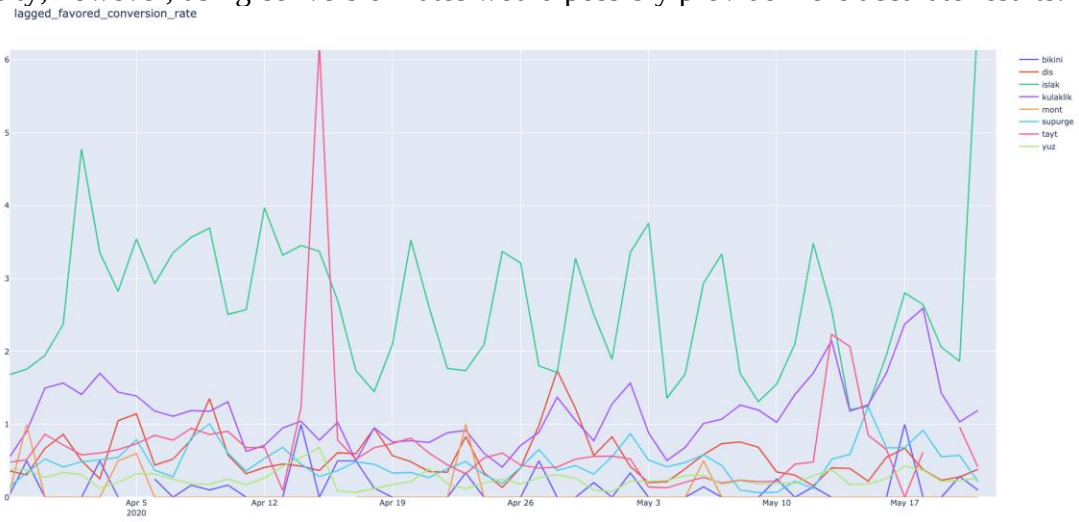

Figure 1: The conversion rate of 2 days lagged favored count to sales quantity

The usage of the selected variables in the Dynamic Linear Regression model can be conducted in two different ways. One way is using the lagged data of that variable for that product which is gathered from 2 days ago. Second way is first forecasting that variable for that day and then using this prediction as an exogenous variable of Dynamic Regression model for sold count. Both ways ended up in similar results but the efficient selection between these two ways is based on small differences which is distinct for each product.

To select good predictors of sold count, correlation analysis was conducted. Based on the correlation coefficients and p-values, exogenous variables for each product are decided. The correlation analysis conducted only accounted for the correlation between sold count and one variable at a time. However, it is possible for some variables to not have a noticeable effect when checked alone but maybe the combination of that variable with another variable has a huge effect on the sold count. And for reaching valid results, further analysis might be necessary.

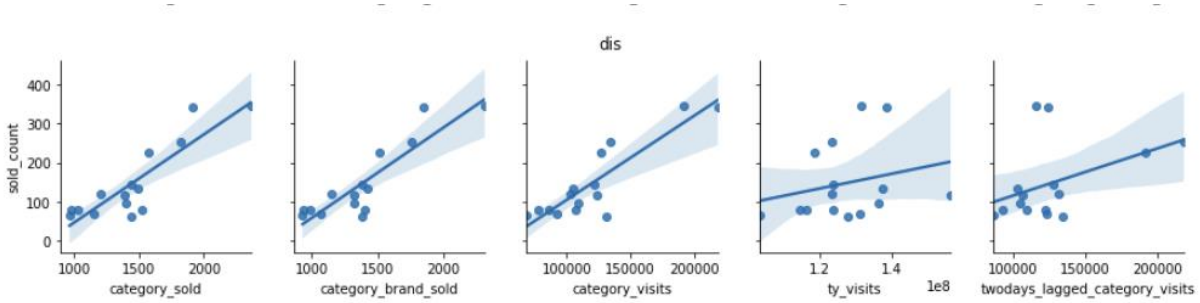

Figure 2: Correlation Analysis for "diş fırçası"

After the development of this model and it being automated, we decided to spend our time searching for better features. In order to do so, we run more than 100 models for each product with combinations of features. The tested features include visit count, favored count, basket count, category visits, category brand sold and price. For example, for the visit count that is forecasted to use as an exogenous feature, several models are run including the permutations of the remaining 5 features next to it. The same procedure is repeated this time trying the lagged version of visit count. The best results in terms of AICc from these permutations were recorded (Appendix A). The model with the features that provided the best result is chosen for each product.

As the final step, based on the suggestions of Prof. Baydoğan, instead of trying to find the best fitting model, we changed our predictions as the average of forecasts from the 3 models with lowest AICc values.

4. Results

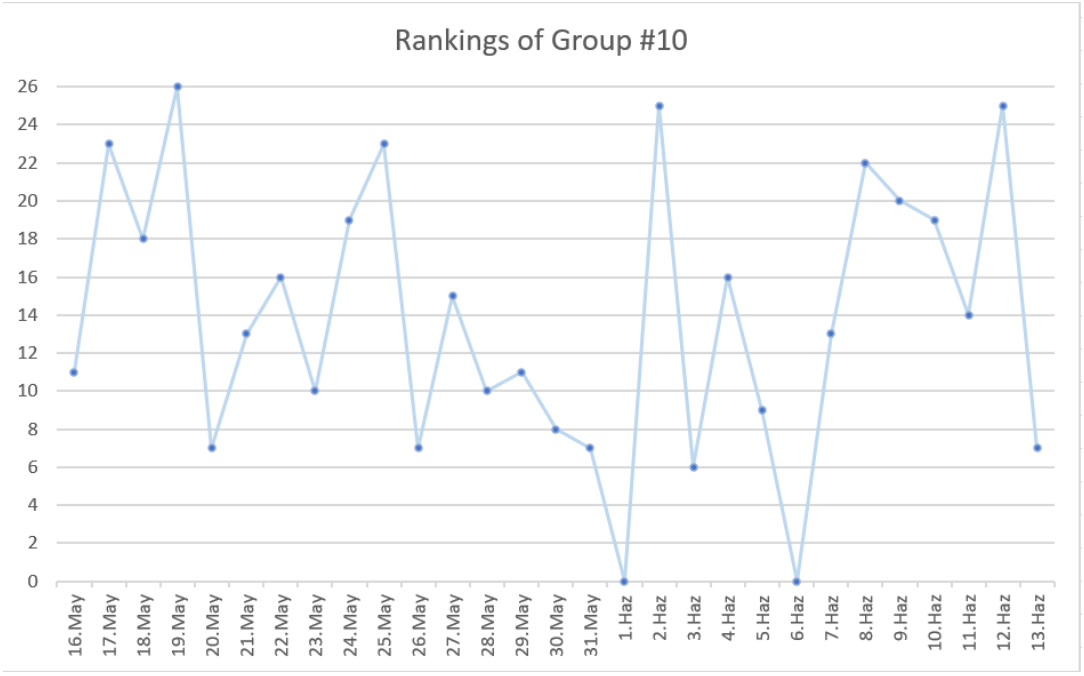

Figure 3: Daily Rankings

At the beginning of the competition, no team has a perfect model and our rankings were okay. They provide somewhat good results, however, the results are volatile. This applies especially to the first week, where we are using the ETS and the ARIMA models to do predictions. Later, we switch to the dynamic

regression model and due to certain problems with the “kulaklık” and “diş fırçası” products, our model started to perform poorly, especially after the 28th of May. As the model continued to perform poorly, we figured out the problem and fixed it on the 1st of June. However, we forgot to send our submission that day so we got a ranking of zero. Another time that a submission is forgotten to be sent is the 6th of June. On the 7th of June, we started to use exogenous variables that were decided from the multiple runs with the permutations of the features. And from then on, until the end of the competition, the finalized model is used.

5. Conclusions and Future Work

During the project there was a strong competition between teams. This is one of the main drivers for each group to search for improved methods to make better predictions. Even though the results were great that day for a group, they still need to conduct some improvements to stay in the competition since there is always a chance for other groups to improve their solutions and get better results. That is why this project is basically an ongoing advancement process and this end up in reaching the best method.

The project continued between 15 May and 12 June and it is quite important to make submissions each day to reach higher places in the competition. The submissions were forgotten for 2 days, that is why we believed that we would be in a much better position if the submissions were conducted every day. Furthermore, it is also significant to obtain new data each day to reach more accurate prediction. This was done successfully by our team.

The timeline of the project made it harder to make use of long term data. Due to the pandemic conditions, making predictions based on the past data has some difficulties. Trying to guess the reactions that people give to the current conditions in the e-commerce environment is complicated. So even though there are data, these data are not 100% helpful to reach accurate predictions. There is a lockdown side that need to be considered as well. During the weekends and special holidays there were not any delivery in some of the cities. That might have a negative effect in the predictions. That is why in order to make accurate predictions, a lot of effort needed to be done. Analyzing the data is not enough. Being up to date and following the current situations are other important points to reaching the accurate conclusions.

One self-criticism should be made based on the research side of the team. Ongoing conditions could be analyzed more carefully, and the effects of these conditions into the e-commerce side should have been predicted with caution. Furthermore, since the holidays are quite challenging for making predictions as well, those days could have been examined more deeply.

To sum up, this project has been quite challenging but also quite improving. For the future work, the suggestion might be to move faster in making predictions. Trying different approaches is necessary but still moving from one method to another takes time. If those improvements could be conducted faster, it takes less time to reach the best solution.

6. References

Hyndman, Rob J. “Forecasting in R.” *DataCamp*, learn.datacamp.com/courses/forecasting-in-r.

Matteson, David S. “Time Series Analysis in R.” *DataCamp*, learn.datacamp.com/courses/time-series-analysis-in-r.

Pagowski, M., et al. "Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts." *Atmospheric Environment* 40.18 (2006): 3240-3250.

7. Appendix

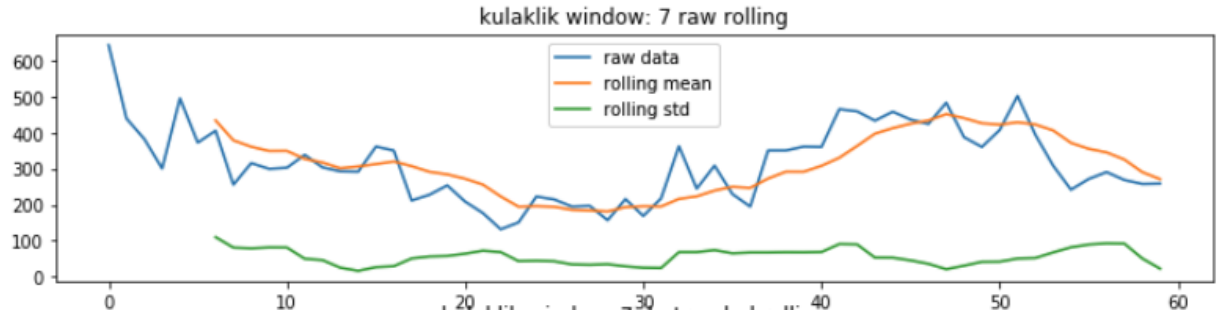

Figure 4 Rolling window with 7 days

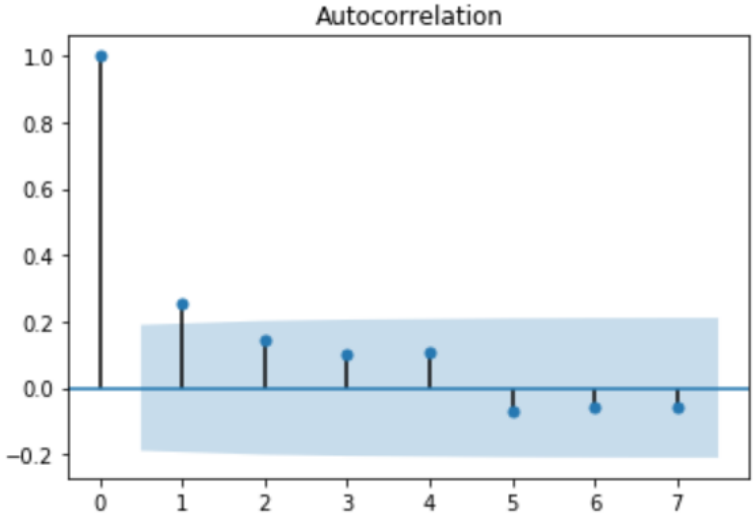

Figure 5: Autocorrelation with lags = 7