IE 360

PROJECT REPORT

30.06.2021

Instructor: Mustafa Gökçe Baydoğan

Group 12

Muhammet Enes Üstün

2017402102

Serverhan Demirboğan

2017402009

Table of Contents

Introduction

Term project of the Spring 2021 term of the IE 360 course is about analyzing and forecasting historical sales data. Data is provided by Trendyol. There send the sales data of 9 products and update it daily. Below, there is a list includes information about products in the dataset.

Product Id	Top Hier.	Bottom Hier.	Brand
48740784	Dış Giyim	Mont	ALTINYILDIZ CLASSICS
73318567	Plaj Giyim	Bikini Üstü	TRENDYOLMİLLA
32737302	Plaj Giyim	Bikini Üstü	TRENDYOLMİLLA
31515569	Alt Giyim	Tayt	TRENDYOLMİLLA
6676673	Elektronik Aksesuarlar	Bluetooth Kulaklık	Xiaomi
7061886	Elektrikli Ev Aletleri	Dik Süpürge	Fakir
85004	Cilt Bakım	Yüz Temizleyici	La Roche Posay
4066298	Anne ve Bebek Bakım	Bebek Islak Mendil	Sleepy
32939029	Ağız Bakım	Şarj Edebilir Diş Fırçası	Oral-B

Table 1: Product characteristics

Data starts from 25.05.2020 and still updates every day. Every product has 13 parameters. Besides the "Product Id", "Date", and "Number of products sold"; dataset has parameters as "Price", "Number of users visited the product that day", "How many times this product added to basket that day" etc. There is a sample part of the dataset represents the date "20.06.2020".

event_date	product_content_id	price	sold_count	visit_count	favored_count	basket_count	category_sold	category_brand_sold	category_visits	ty_visits	category_basket	category_favored
2021-06-20	31515569	60,19	387	10957	656	1741	8045	1363	477183	126276503	39043	40585
2021-06-20	32737302	59,99	76	5657	510	414	7149	3292	1173728	126276503	40550	101322
2021-06-20	32939029	143,66	149	4555	596	587	807	743	32256	126276503	2979	2963
2021-06-20	4066298	67,78	892	10038	517	2026	5199	3377	65464	126276503	14368	5018
2021-06-20	48740784	699,98	1	160	13	8	1098	7	253165	126276503	7475	17062
2021-06-20	6676673	106,1	658	22578	1131	1903	5507	1412	283390	126276503	24529	17704
2021-06-20	7061886	294	14	909	72	49	866	129	71231	126276503	3525	5868
2021-06-20	73318567	63,55	15	8673	743	106	7149	3292	1173728	126276503	40550	101322
2021-06-20	85004	86,32	83	3650	1016	388	4591	650	175780	126276503	22889	28599

Table 2: Sample Part of the Dataset

The aim of the project is analyzing the data according to its parameters and forecast the sales (sold_count) data of two days after the last date the dataset has. To create a competitive environment, every group of students submitted their daily forecast to an application programming interface (API). Also, new daily data sent from Trendyol should be requested from the same API. All group's predictions are compared with actual sales data by calculating the weighted mean absolute percentage error (WMAPE).

$$WMAPE_{day} = \frac{\sum_{p \in Products} |y_p - \hat{y_p}|}{\sum_{p \in Products} y_p}$$

Figure 1: WMAPE calculation

After the WMAPE calculations, system rank students according to their predictions' correctness. Best forecast gets the highest point or vice versa. Every day's points collected in the system to create a leaderboard of the competition. Forecast submission started on June 11, 2021, and will end on June 25, 2021, by forecasting the sales data of June 26 (because on the June 25, our latest data is from June 24).

Approach

Firstly, our code combines the new data with the data we already knew using rbind() function. After the code determine the number of days in the dataset and creating the dates and test_dates sequence variables, it starts to analyze every product and generates a forecast for the day after tomorrow according to historical data. For every product, we create two time series objects using xts: train and test objects. We determined dates of these objects in compliance with data we have. For example, some products have not useful data in 2020, so we did not use them for these products. Generally, test period starts by the March 1, 2021, with exceptions.

Before building models, we check correlations between the sales number and auxiliary data such as price or favored_count. We did not check correlations that have missing data such as visit_count. Then we use the helpful data when we build a model ARIMA with regressors for the product. Sometimes we found the best model using auto.arima() function but sometimes we found it by trying. Sometimes we found better model than the model auto.arima() suggests. To make a 2 steps ahead ARIMA forecast, we firstly make a 2 steps ahead forecast for regressors and we used them in our model.

After the ARIMA model, we built a linear regression model using the same correlations. We used the modes that have the lowest residual standard errors and the highest r-squared values.

At the end, we calculated MAPE or MAE values for each model to decide which of them we will use for prediction submissions. Usually, we pick both ARIMA and Linear forecast and sent their mean.

Let's look at our approach to the products one by one.

1. 48740784 - Dış Giyim Mont

We created our train times series object started from 29.09.2020 until 28.02.2021, and the test data starts from March 1st, 2021, to today. There was no useful data before September 29, 2020.

When we check the correlations between sales data and price, favored_count, basket_count, and category_sold data; we found high correlation with price, basket_count, and category_sold; so, we used them in our models.

We created an ARIMA model by order (1,0,0) with the lowest AIC value equals to 270.7151. The prediction sketch for the test dates of the ARIMA model is below:

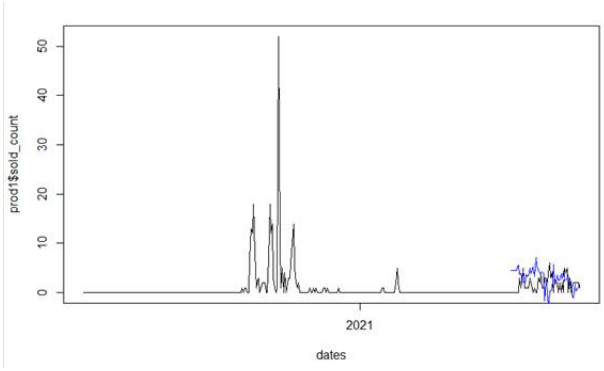

Chart 1: ARIMA predictions of the product 1

After the ARIMA, we built a linear model to forecast the data using the same correlations without the price da ta. The prediction graph for the test dates is below:

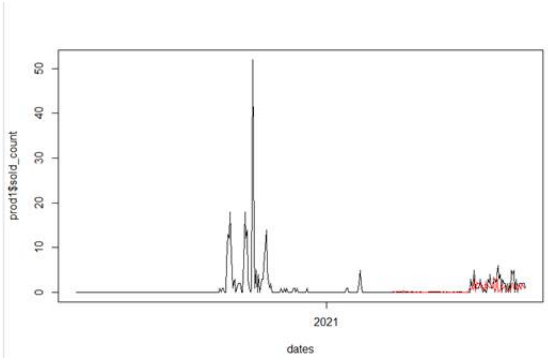

Chart 2: Linear regression model predictions of the product 1

Since we found these graphs acceptable, we submitted the average of the two models' forecasts as our prediction.

The number of sales for this product change between 0 to 1 usually according to stock problems because the product is a coat is used in winters and we are in summer now. Our predictions usually in the range of 0 to 2 is good number actually, but it might give a result as a high WMAPE value.

2. 32939029 - Ağız Bakım Şarj Edebilir Diş Fırçası

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 30.04.2021, and the test data starts from May 1st, 2021, until today.

When we check the correlations between sales data and price, category_favored, basket_count, and category_sold data; we found high correlation with category_favored, basket_count, and category_sold; so, we used them in our models.

We created an ARIMA model by order (2,0,1) with the lowest AIC value equals to 3027.149. The prediction sketch for the test dates of the ARIMA model is below:

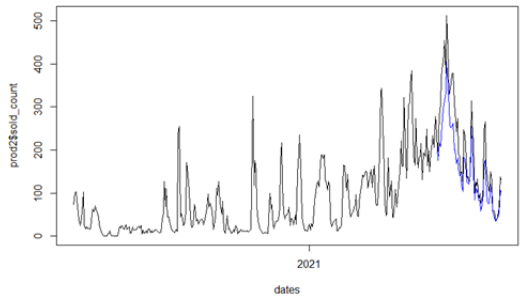

Chart 3: ARIMA predictions of the product 2

We have the linear regression model forecast for the test period:

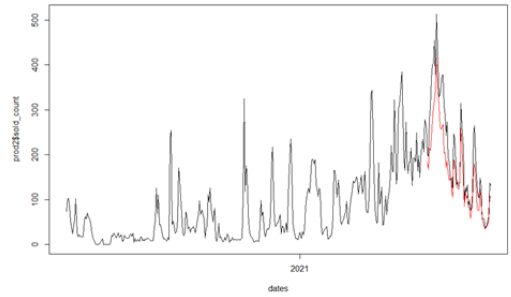

Chart 4: Linear regression model predictions of the product 2

Both our models seemed like they had good and similar predictions. We calculated MAPE for both models. ARIMA model had the value 0.1993236 and linear regression had 0.1904894. Because of the close values of MAPE, we used both models to make our prediction.

3. 4066298 - Anne ve Bebek Bakım Bebek Islak Mendil

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 28.02.2021, and the test data starts from March 1st, 2021, until today.

When we check the correlations between sales data and price, category_favored, basket_count, and category_sold data; we found high correlation with we found high correlation with all four of them (range between 0.565 to 0.895); so, we used them in our models.

We created an ARIMA model by order (1,0,0) with the lowest AIC value equals to 3394.042. The prediction sketch for the test dates of the ARIMA model is below:

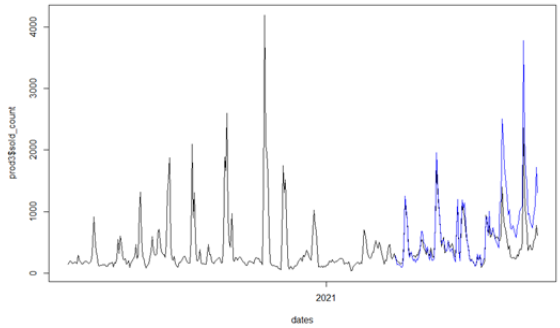

Chart 5: ARIMA predictions of the product 3

We have the linear regression model forecast for the test period:

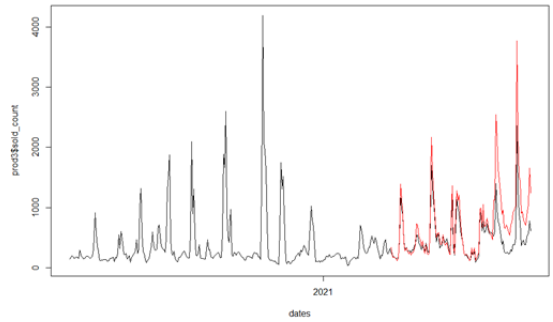

Chart 6: Linear regression model predictions of the product 3

Both models had worse predictions than other products' models. ARIMA model got MAPE value equals to 0.419247, linear model got 0.4524202. These values are lowered ones after our interventions to the models. Since they are close to each other, we used both in our calculations.

4. 85004 - Cilt Bakım Yüz Temizleyici

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 28.02.2021, and the test data starts from March 1st, 2021, until today.

When we check the correlations between sales data and price, category_favored, basket_count, and category_sold data. We found correlation with basket_count (0.816) and category favored (0.672); so, we used them in our models.

We created an ARIMA model by using auto.arima() function. AIC value equals to 2609.712. The prediction sketch for the test dates of the ARIMA model is below:

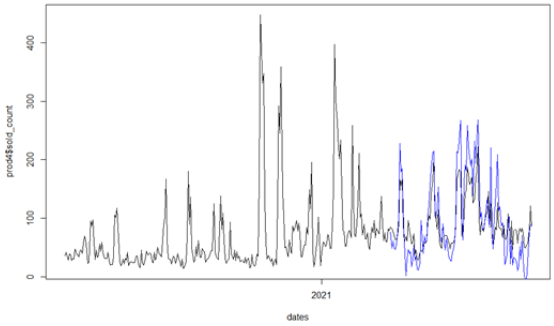

Chart 7: ARIMA predictions of the product 4

We have the linear regression model forecast for the test period:

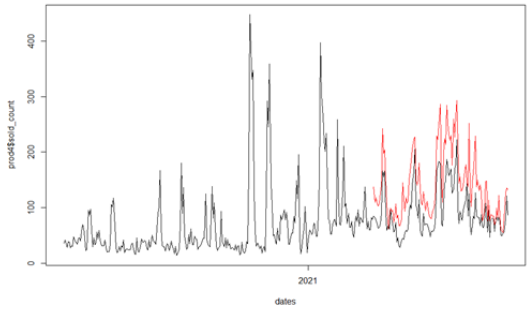

Chart 8: Linear regression model predictions of the product 4

The ARIMA model gave the MAPE value 0.348545, and the linear regression model's value equals to 0.5755782. Although the ARIMA model was more accurate, we used the mean of both models' predictions because predictions of ARIMA model was lower than the real data, linear regression model's higher.

5. 6676673 - Elektronik Aksesuarlar Bluetooth Kulaklık

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 28.02.2021, and the test data starts from March 1st, 2021, until today.

When we check the correlations between sales data and price, favored_count, category_favored, basket_count, and category_sold data; we found high correlation just with basket count that has value 0.863; so, we used it in our models.

We created an ARIMA model by order (0,1,4) with the lowest AIC value equals to 3313.264. The prediction sketch for the test dates of the ARIMA model is below:

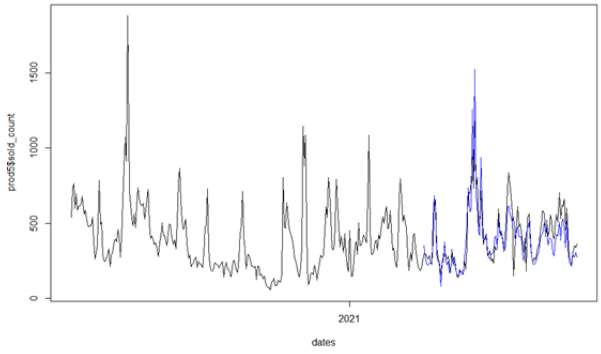

Chart 9: ARIMA predictions of the product 5

We have the linear regression model forecast for the test period:

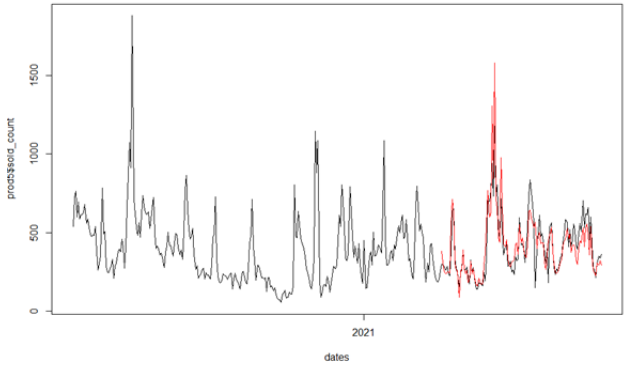

Chart 10: Linear regression model predictions of the product 5

Both models' forecast was not bad. Their MAPE's were around 17-18 percent. So, we made our prediction by using both models' mean.

6. 7061886 - Elektrikli Ev Aletleri Dik Süpürge

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 28.02.2021, and the test data starts from March 1st, 2021, until today.

When we check the correlations between sales data and price, favored_count, basket_count, and category_sold data. we found high correlation just with basket count that has value 0.87; so, we used it in our models.

We created an ARIMA model by using auto.arima() function. AIC value equals to 2497.014. The prediction sketch for the test dates of the ARIMA model is below:

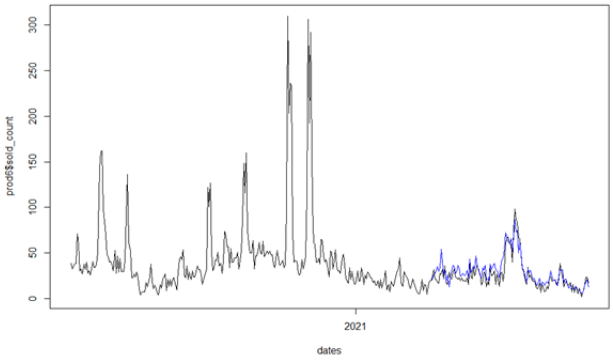

Chart 11: ARIMA predictions of the product 6

We have the linear regression model forecast for the test period:

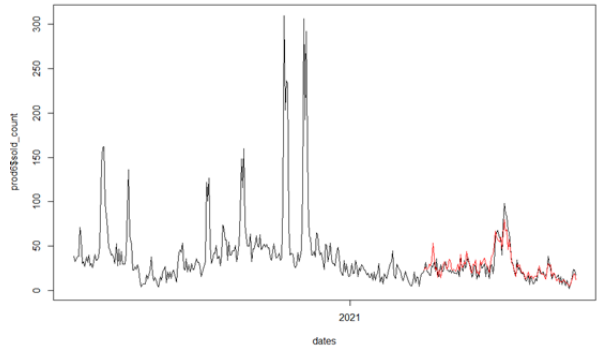

Chart 12: Linear regression model predictions of the product 6

We used both model in our predictions, again. ARIMA had error about 24%, linear model has %30. Since the sales numbers for this product were low, these error differences represented decimal numbers. So, we submitted the average of the models.

7. 31515569 - Alt Giyim Tayt

We created our train times series object started from on the beginning of our full data, 25.05.2020 until 28.02.2021, and the test data starts from March 1st, 2021, until today.

When we check the correlations between sales data and price, favored_count, basket_count, and category_sold data; we found high correlation just with basket count that has value 0.84; so, we used it in our models.

We created an ARIMA model by order (1,2,4) with the lowest AIC value equals to 4378.997. The prediction sketch for the test dates of the ARIMA model is below:

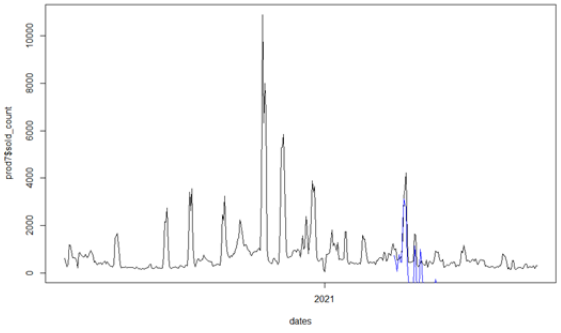

Chart 13: ARIMA predictions of the product 7

We have the linear regression model forecast for the test period:

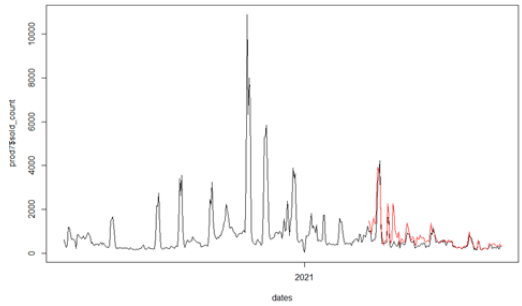

Chart 14: Linear regression model predictions of the product 7

In this model, we just used linear regression model's forecasts. Because the ARIMA model gave us 679% error with negative values of predictions we can see on Chart 13.

8. 73318567 - Plaj Giyim Bikini Ustü 1

We created our train times series object started from 23.01.2021 until 30.04.2021, and the test data starts from May 1st, 2021, to today. There was almost no data in 2020.

When we check the correlations between sales data and price, favored_count, basket_count, and category_sold data; we found high correlation with basket_count (0.98), and favored_count (0.772); so, we used them in our models.

We created an ARIMA model by order (2,1,0) with the lowest AIC value equals to 588.6244. The prediction sketch for the test dates of the ARIMA model is below:

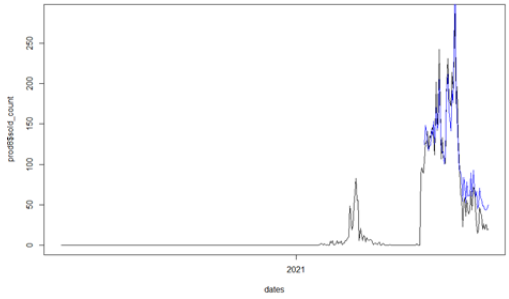

Chart 15: ARIMA predictions of the product 8

We have the linear regression model forecast for the test period:

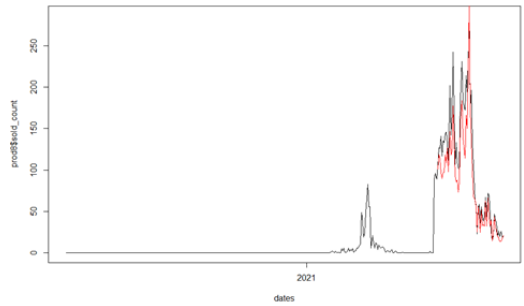

Chart 16: Linear regression model predictions of the product 8

ARIMA model had the error rate equals to 0.3889271, linear model had 0.1955627. Since the letter was more accurate, we just used it in our forecasts.

9. 32737302 - Plaj Giyim Bikini Ustü 2

We created our train times series object started from 20.02.2021 until 30.04.2021, and the test data starts from May 1st, 2021, to today. There was almost no data in 2020 and January and February 2021, there are exception in days of summer, 2020.

When we check the correlations between sales data and price, favored_count, basket_count, and category_sold data; we found high correlation with basket_count (0.949), and favored_count (0.734); so, we used them in our models.

We created an ARIMA model by order (1,0,0) with the lowest AIC value equals to 492.2612. The prediction sketch for the test dates of the ARIMA model is below:

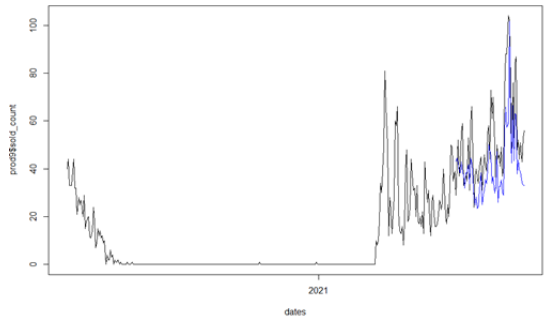

Chart 17: ARIMA predictions of the product 9

We have the linear regression model forecast for the test period:

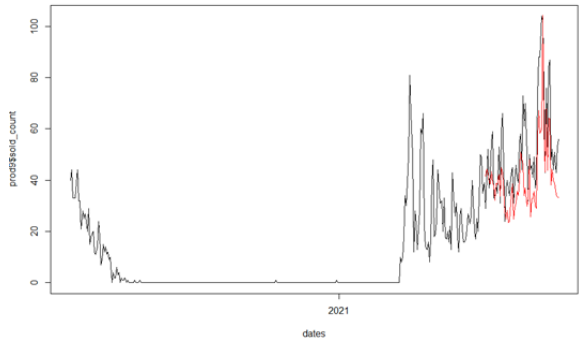

Chart 18: Linear regression model predictions of the product 9

Both models had MAPE value around 22%, we used their mean to forecast sales because of this.

Results

We used linear regressions and ARIMA models to forecast the future sales of the products. We decided to use which of them according to forecast graphs for test periods and error rates (MAPE/MAE). For some products, these models gave accurate predictions but for some they did not. As mentioned above, groups were graded according to their predictions. Leaderboard of the competition is below:

Competition	Group8	Group1	Group6	Group3	Group5	Group2	Group14	Group11	Group10	Group9	Group15	Group12	Group1
Day-12-June	18	16	10	14	15	19	17	7	9	13		8	
Day-13-June	17	19		11	15	12	14	16	7	9		10	
Day-14-June	14	17	16	15	18	12	13	19	10	8		9	
Day-15-June	13	19	16	17	11	14	7	18	8	10		6	
Day-16-June	18	14	19	17	13	15	11	16	7	9	12	10	
Day-17-June	13	17	18		10	16	12	15	14	11	19	7	
Day-18-June	11		18	19	9	16	13	16	12	7	17	14	
Day-19-June	16	9	7	14	12	15	17	6	11	8	10	18	
Day-20-June	18	12	13	7	10	17	19	11	9	14	16	8	
Day-21-June	11	19	18	8	16	9	12	13	10	14	17	15	
Day-22-June	19	15	18	12	10	5	8	13	17	11	14	9	
Day-23-June	18	19	17	11	10	9	7	6	16	14	15	13	
Day-24-June	19	13	11	15	18	8	7	5	12	9	16	14	
Day-25-June	19	15	12	18	14	11	8	5	16	17	10	9	
Day-26-June	17	16	18	19	12	13	11	7	14	9	15	10	
Grand Total	**241**	**220**	**211**	**197**	**193**	**191**	**176**	**173**	**172**	**163**	**161**	**160**	**1**

Table 3: Leaderboard of the competition

At the first week, results show that we could not make our predictions accurately. We changed our code a little for better forecasts. The changes we made:

- We changed regressors of some models. Even if we found a high correlation, some variables had affected the predictions adversely or vice versa. We could see the difference easily on the graphs and MAPE values.
- Some ARIMA models had gave very wrong results. For these products, we just used the linear regression models to forecast the future sales. Product 7 (31515569) and Product 8 (73318567) were these models.
- We adjust dates according to avoid empty data.

At the second week, we had better results than the first week. But our predictions were not stable. Products that have worse error rate such as Product 3 (4066298) gave accurate forecasts some days, but some days predictions were bad. This made our ranks for each day different from each other.

Conclusions and Future Work

All teams have submitted their predictions for 15 days. We can easily see the competition environment among groups. Some groups got better points at the first week, some of them made their rise at the second week. We improved our predictions at the second week by little changes in the program. But of course, these were not enough to predict the best sales. There were some problems.

Data could be decomposed. We could seek for seasonalities to make our predictions better. If there were more data in the past, we could find different seasonalities at different frequencies.

We could submit integers for predictions. Because we forecast the products' sales numbers, they should be integers. Even our predictions' integer part was correct, there were errors because of the decimal points; so, maybe it increased our WMAPE value every day. Especially for the Product 1 (48740784) had very low sales numbers, it affected a lot more.

We could search data for the same products from competitors of the Trendyol such as Hepsiburada, Amazon, N11 etc. Maybe these data affected our Trendyol sales data. Or we could find data of similar products such as looking other coats for our Product 1 (48740784).

Google Trends data could be used. Daily search numbers of products or their categories may have correlations between the products. These outer data may improve our models to forecast better.

The main inadequacy of our code is it contains just 2 types of models (ARIMA and Linear Regression). If we tried to build different types of models such as Naive Forecast or GAMs, we could make better forecasts. In the future, these changes may improve the predictions a lot.

Code Link

Here is the link of our group's source code for forecasting.