**Introduction**

As Internet evolves throughout years, online shopping became one of the most important sectors around world. Everyday, millions of people purchase many different items via online shopping websites. Trendyol is Turkey’s most popular online shopping website. In this project, we were expected to forecast the sales of eight different products in Trendyol.

Historical sales data from the 30April 2019 were given in dataset. Average price of the items sold, number of items sold number of visits to the particular product, number of times the product is added to favorites, number of times the product is added to basket, number of items sold from the same category, number of items sold from the same category brand, visits to the category of the item, total Trendyol visits in the given day were available in dataset.

Before using data for forecasting, the sales data of days having a price value -1 had to be removed. On these days, there were not any sale and these rows in data could cause problems to determine a good model for forecasting. In data of 2019, there were many rows in this format. So, starting date for training data for each product was different.

Groups competed according to forecasts between 15th May and 12th June. Groups had to submit their sales forecasts before 1:00 pm everyday. These forecasts are the forecasts for the day after submission. Last data which could be used for predictions were sales which occurred one day before the submission. Each day, groups were ranked according to their weighted mean absolute percentage error (WMAPE). Groups having the least WMAPE value become winner of the day and took 26 points. Overall ranking is determined with sorting groups by cumulative sum of daily ranking points.

Groups were free to choose their own approach to predict the sales. As Group 11, we used linear regression, naïve forecast approach with lags of two and seven days, autoregressive integrated moving average (ARIMA) and simple moving average. All approaches are explained in Approach part. Among these methods, method having least mean squared error or mean absolute error, depending on the cases of zero-demand, was used to predict next day’s sales.

**Related Literature**

Before starting submissions, DataCamp courses assigned to students in the IE360 course were studied:

*Forecasting in R, by Rob J.Hyndman:* Forecasting methods which can be used in R were included in this course. Plotting the data, training and test data concepts, exponential smoothing, ARIMA models and other advanced methods can be found in this course.

*Forecasting Product Demand in R, by Aric La Barr:* This course is very useful in this project because the main purpose of this project is also a demand forecast. Plotting the time series data, evaluating methods, components of demand, regression in time series and hierarchical forecasting are included in this course.

**Approach**

For this project, in order to forecast the next day’s demand for the given products on Trendyol, we have facilitated from four different forecasting methods.

* Naive Forecast:

Naive Forecasting is a technique that uses that last period’s actual values as the forecast for the next period. We built two different models using naive forecast approach with lags of two and seven days.

* Moving Average:

Moving average is a forecasting technique that displays the overall trend in a data set by taking the average of observations on a selected set of time period. In this project, we applied moving average method with different time periods. We built six models that take the average of the last observations that range from 2 days to 7 days.

* Linear Regression:

Linear regression is an approach that helps to predict the future values of a variable based on one or more explanatory variables. The case in which only one explanatory variable is used is named as simple linear regression. However, if more than one explanatory variable is used than it is called multiple linear regression. We used both of them to build models in this project. In order to forecast the number of sales, we considered the data that might be correlated with it and built two different regression models, one of which takes only price as an explanatory variable and the other takes both price and weekday/weekend difference into account. We did not include the rest of the data since they are not highly correlated to sales count.

* ARIMA:

ARIMA (Auto Regressive Integrated Moving Average) is a class of models that is used to make predictions from a time series data based on the past values, which means it describes the autocorrelations that exist in the data. ARIMA can be used to build a model for a time series that is ‘non-seasonal’, exhibits a certain pattern and is not a random white noise. An ARIMA model can be described by 3 terms: p, d, q where p stands for the order of the AR term, q represents the order of the MA term and d is the number of differencing necessary to make the time series stationary.

Before applying the methods, we have mentioned above, we manipulated the data since the methods were not applicable to raw data. From the data that was provided to us, we eliminated the ones that we considered were unrelated and took only the required information. Also, we have ignored the data that belonged to a date that was before the sale of the specific item began for each item. Finally, we added a column called ‘is\_weekend’ to build the second regression model.

We have applied a simple cross validation by splitting the data into two sets: training set and test set. Cross validation indicates how the model will perform on the unseen data. We decided to take the last 31 days as test data and previous data as the training data.

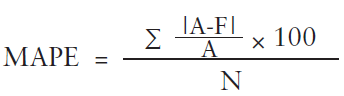
  We mostly preferred to use MAPE to select the best model since it is independent from scale and easy to interpret. However, for some of the products, due to lack of data, we were unable to calculate MAPE. For such products, we have made our decisions for the best model by using MSE.

In the first days of the project we have mostly used moving average models to make predictions since they gave better results when their MAPE or MSE values are considered. However, as project continued, ARIMA and regression models started to make more accurate forecasts, so we began to use them instead of moving average.

**Results**

As mentioned in the previous section, we used different forecast approach techniques which are naive forecast, moving average, linear regression and ARIMA model. It is possible to see that trend of the data changes over time. Hence, throughout the submission period we observed the performance of the forecast models we used and made changes accordingly. Initially, we mostly constructed our predictions with the moving average models since they have us the smallest MAPE or MSE values.

MAPE refers to Mean Absolute Percentage Error and is a measure of prediction accuracy of forecasting model. It expresses the accuracy as a percentage ratio defined by the formula ;



where A is the actual value, F is the forecast value and N is the time horizon.

The other measure of prediction accuracy which we used in our project is MSE. MSE refers to Mean Squared Error and measures the average of the squares of the errors which expresses the average squared difference between the predicted values and actual values. It is computed as;



where{\displaystyle Y}YYyYSAFASF Y is the actual value and Y- head is the predicted value.

In our predictions, we commonly preferred to use MAPE performance indicator since it is independent from scale and easy to interpret. However, for some of the products, due to lack of data, we were unable to calculate MAPE. For such products, we have made our decisions for the best model by using MSE. Based on the approach implemented, you can see the results of our prediction models.

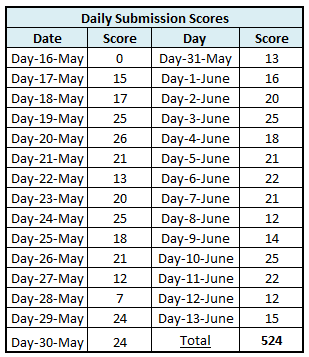


Table : The daily submission scores

Consequently, the forecast method we used in the model is dependent on the scores. Initially, we started with moving average models but in the later days we started to use regression and ARIMA models since they gave us more accurate predictions.

**Conclusion and Future Works**

As mentioned above, there are several approaches to create a forecast and we used four methods to understand which method should be used for each product in this project. We tested our approaches using test data set which created from existed time series for each product and choose least MAPE value. First days of submission, moving average model was worked well, but in forward days our Arima and regression models’ results were better. Because moving average model is responsive to current event and has been affected from daily changes more.

Regression models are used when there is relation between two (or more) elements in a time series. In project, ‘price’ and ‘is\_weekend’ variables are used to understand behavior of ‘sold\_count’. Arima models and regressions were resulted well in later days because models had more historical data to interpret and reduce autocorrelation.

In next implementations, basket\_count (number of specified product in basket), favorite\_count, (number of specified product in customer’s favorite list), or number of subsidiary products that are sold with primary product can be used to develop different regression models.