Time Series Forecasting for Hourly Order Prediction

By Khawaja Hassan Abbas

Submitted to Central European University Department of Economics & Business

In partial fulfilment of the requirement for the degree of Master of Business Analytics

Budapest, Hungary

June 2022

Table of Contents

Problem Definition:

The focus of this research paper resolves around employing machine learning algorithms on the time series data for our client who want to predict the number of orders coming in the next hour based on the trend and other exogenous factors. The motivation to dig deep into this is to understand the dynamics of the industry and the major logistic problems faced due to fluctuation of the orders. The uncertainty and high variation in the order directly impact the functional efficiency of the business leading to higher delivery lead time and insufficient allocation of delivery shifts. This research will provide the base to further complex module that can be employed to serve the specific need of the company.

The pertaining issue that the company is facing is that due to high seasonality in the orders on hourly basis the logistic team is facing a dilemma of how many shifts are to be allocated in the next 24 hours. Therefore, our research will act as an additional analytical layer to their decision-making process that will leverage them to make more data supported decision. To summarize our problem statement is:

"Predicting the number of orders based on split of hour"

Moving average:

Let's start with the most basic methodology which is naïve hypothesis which states that future values will be derived from the present value. In other words, the average of x previous values will define the future value and that is why we will be using moving average. The first step here was just predicting the orders for next hour by simply taking the average of last 24 hours or the daily order from last day.

Exponential Smoothing

However, the drawback of this approach is that it failing to account for the weekly seasonality in the data and marking them as anomalies each week. Therefore, this leads us to our next complex model of weighted average where we will be assigning weightage to preceding values in order to avoid the capturing of false positive anomalies.

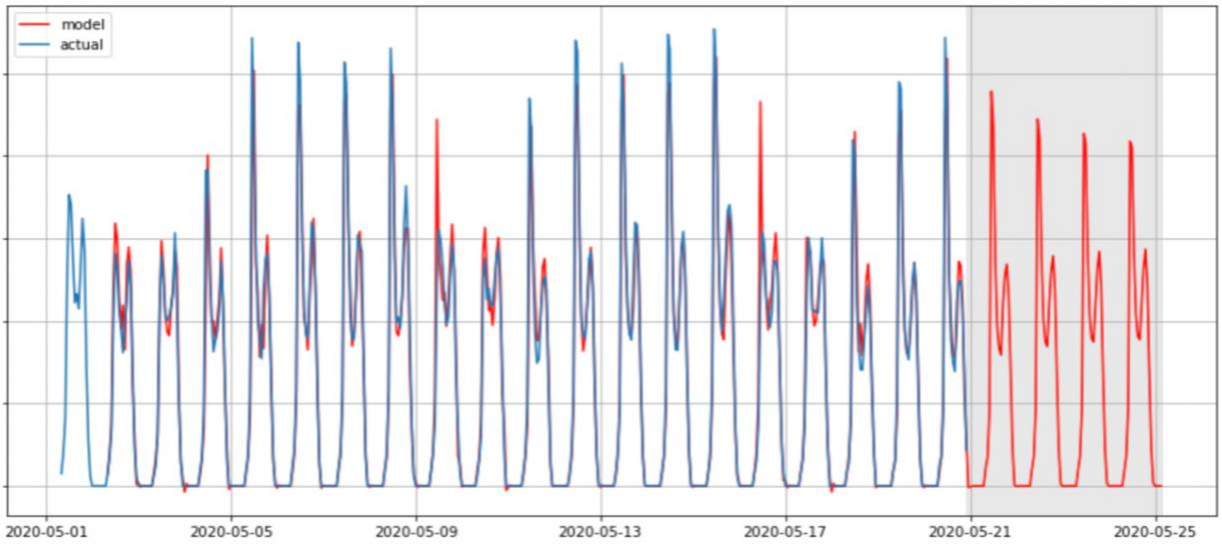
Holt Winter (Triple Exponential Smoothing)

Now since we have seen how exponential smoothing working out for us, we will be diving straight into a much more complex technique which is also known as triple exponential smoothing or Holts Winter smoothing. Our final Holt-Winter method module has slen value of 24 depicting the hourly seasonality for the day. Moreover, we provide our function with 100 n predictors so that we will be having the prediction for the next 4 days.

SARIMA (Seasonal Auto-Regressive Integrated Moving Average)

Let's focus now on constructing ARIMA model and gradually add the seasonality factor in it as well. Based on the results of the of our statistical test we can conclude that our data does not have a unit root or in other words they are reflecting a constant mean and variance. Therefore, we can reject our Null hypothesis and state that our series is stationary. Now we need to account for the seasonality which we will be doing by subtracting the series value with value multiplied by the defined season which is 24 in our case.

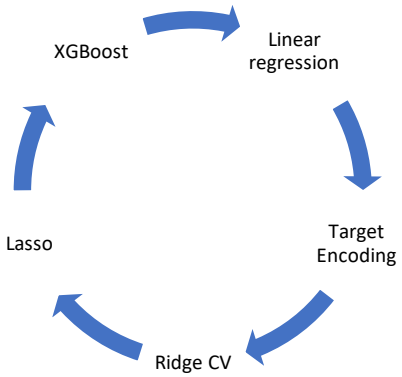
This model consists of following components and its usually written as SARIMA (p, d, q) (P, D, Q, s). There are 7 components, and we need to come up with the combination that result in the most optimal prediction. Therefore, rather than testing different combination for each of them we will make range for each of them and let the function run to find the most optimal combination that result in lowest AIC value.



Looking at the prediction above our model is showing a gradual decrease as we approach closer to weekend. This validate that our model is not only capturing our daily seasonality but also the weekly seasonality of our series

Univariate Vs Multi-Variate Time

In reality, there are many other variables that along with the time factor that explain the movement of our series. To begin with we will be calculating the lag of our time series and date/time features. The following provides the illustration of our building blocks and series of modelling techniques which was adopted to make our model more sophisticated and fetch for better model performance. Each of the mentioned model below gave us the mean absolute error percentage and the variable importance. The best model here turns out to be XGBoost with the lowest error term but once again this is a black box method and tends to overfit the model. Having said that Lasso model with the help of lasso we can further narrow down our research by focusing on the nonzero coefficients and using them for further modelling.



Limitation of the model

- The data that we are currently working with is a very simplified data which lacks some classification features which would have been useful in making our model more refined and realistic.
- The number of orders coming in any district is based on the number of vendors located in those area.
- Our model tends to highlight the total number of orders but not the district where orders are generated from

Way Forward

The current model and data pipeline has provided us with framework for further modification and building more sophisticated and complex module. orders timing varies when compared with the other districts. One important thing that we need to still dig deeper is having an optimal loss function and the cost associated with it. Moreover, using this information our next model should be Prophet model of Facebook which is even more sophisticated model and might help to capture better seasonality. Nevertheless, we need to keep the modelling process on going in search of finding an optimal model. With the help of Deep learning and Machine Learning technique we can further create model that might has more explainable outcomes and cater for specifically to our needs.

Conclusion

The purpose of this thesis paper was to construct and implement statistical models incorporating time series & exogenous variables to evaluate impact our order fluctuations. For the purpose of modelling, we made sub-groups of data for monthly, daily, and hourly. In search of finding optimal forecasting model, we applied various dynamic methodology ranging from complex SARIMAX model to simple linear regression to account for our high-dimension data. Lastly, based on our forecasting modules we found out that SARIMA outperformed other basic models in determining not only the daily seasonality but also the weekly variation in the dataset. To conclude, one need to understand the trade-off between quality and cost-associated with each approach. In some cases, the results from SARIMA model might be very much collated with simple linear regression but it requires hours of manipulation in search of finding the right parameters.

Lesson Learned

The most important take away from this research is learn how to break down the problem in smaller abstracts and then target them individually, rather than looking for the ultimate solution. Moreover, before you start with coding it is very important for you to begin with building a firm knowledge base around the modelling technique and understanding the dynamics of your industry. Most of the irregularity and outliers in the data get explained if you have good grip on the industry knowledge. Lastly, having good communication and guidance from your supervisors is the key for a successful project.