# Sales Forecasting

## Exploratory Analysis

Exploratory Analysis of the Dataset and key insights realized

The individual and total course sales of the company are seasonal in nature and have stayed constant in the given period.

Of the 600 courses, over 80% are from the Software marketing and the Development domain, while the rest of the courses belong to the – finance and Business. And hence, sales of the courses follow a similar ratio.

Of the five factors that may have potential impact on the sales – User traffic and short promotion could explain 75% of the change in value of sales, while long promotion, public holidays, and competition metric had a minimal impact.

## Approach

A brief on the approach, which you have used to solve the problem

The univariate series to be predicted here is sales for each course. The primary step was to analyze if the series was dependent on any of the given explanatory variables. The given five features were regressed against dependent variable, sales. The linear regression model revealed that short promotion and user traffic were significant features and could explain 75% of the variation in sales.

This linear regression model built to find relationships between variables was used to build the ARIMA model with 2 regressors initially. Eventually, user traffic was dropped as it was not available in the test dataset.

The auto ARIMA model was first built for course #1, the approach was to apply the model so built for the first course 1 to rest of the 599 courses. On the basis of the observation that overall sales and the individual course sales series was stationary, the same model was applied to all the 600 courses.

Here, another approach could have been to forecast for the overall sales for each day and proportionate the course sales from the forecasted total sales on the basis of the past proportions of each course sale in the total course.

## Data-preprocessing / Feature engineering

What data-preprocessing / feature engineering ideas really worked? How did you discover them?

Time series can be run only when the following three factors are understood well. 1. How well we understand the factors that contribute to it; 2. how much data is available; 3. whether the forecasts can affect the thing we are trying to forecast. (Hyndman, R. and Athanasopoulos, G., 2018)

In the sales series, we had data available for more than two years for each course, through regression modeling and correlation mapping, it was figured that two explanatory variables could significantly explain the variation and direction of sales – primarily user traffic and short promotion.

Correlation mapping noticeable positive correlation between sales and short promotion and user traffic, which meant that any increase in either of the two variables could reflect in increase in sales. So, the direction of the value of sales was understood.

Here, a lagged difference of sales was also considered as an explanatory variable initially to increase the accuracy of prediction, however, it didn’t significantly explained the variation to the sales, also the presence of this variable also did not increase the accuracy of the model.

## Final Model

What does your final model look like? How did you reach it?

Primarily three models were tested to predict the sales for each course – ARIMA, Auto ARIMA and Neural nets.

The predictions were made based on AUTO Arima as it showed the least Root Mean Squared Logarithmic Error (0.12) in comparison of the other two models. The model included short promotion as an exogenous variable, with seasonal factor turned to TRUE. The final ARIMA model included AR(2), I(0) and MA(1) terms, as was also observed with the ACF and PACF plots of the series. The residuals of the model so generated resembled white noise which is what is needed to make accurate predictions.

Cross-validation was not used here to test the efficiency of the model because cross-validation “can be applied to any model where the predictors are lagged values of the response variable” (Hyndman, R., 2016). In the current model, the explanatory variables are changing with sales each day, hence the model is trained on the entire training dataset.

The accuracy of the predictions is checked using RMSLE value.

## Interpretation of the final model

Interpretation of your final model. Which features are important, and which are not?

As mentioned earlier, the final model was forecasted with short promotion as the explanatory variable, although user traffic could explain a fair share of the variation in sales. This factor wasn’t available in the test dataset, which threw off the model while forecasting on the test set. Here, I’d like to mention that this is my first attempt at forecasting, and hence my limited understanding, could have resulted in dropping this very significant variable.

Features like competition metric, long promotion and public holidays did not add much to the model and didn’t explain the variation in sales much, hence they were not considered in the final model.

References

1. Hyndman, R. and Athanasopoulos, G., 2018. Forecasting. [Melbourne]: O Texts, online open-access textbooks.