Time Series Project

Winter 2017

By: Ritika Lohadiya

Executive Summary:

The goal of this time series project was to use the Kaggle data set on Pokemon sighting data to forecast the number of pokemons that will appear at a particular time in New York, USA. Our data consisted of approximately 300,000 sightings observed in August, 2016 from around the world.

Data Transformation:

As the data obtained had irregular time intervals, our first task was to transform the data into a standard time interval. By aggregating the initial data set I converted it into per minute intervals i.e. per minute total number of Pokemon sightings. Other variables in the dataset included: Wind Speed, Wind Bearing, Pressure, Temperature and Population density. I had data of pokemon sightings from all the major cities in the world. Hence, I culled down the dataset to include sightings only from New York city.Thus, our final dataset included of 6246 observations.

Linear Regression analysis:

Initially I used a very basic linear regression model to forecast the number of pokemons in New York. However, diagnosis of residuals suggested serial correlation and thus linear regression did not seem to be a good model for such a time series data.

Dynamic Regression Model:

Since, I had data on some X variables that could have an impact on the number of Pokemon occurrences in New York, I decided to use dynamic regression model for our analysis. Before running the model, I did some exploratory analysis to see whether there were any trends or seasonality in the dependent and independent variable or any transformation was required.Since, not much variability was observed, no transformation was applied.The ACF and PACF function suggested non-stationarity in both dependant and independant variables and thus I took the first difference to make the series stationary.The stationarity of the differenced series was also confirmed through ADF and KPSS test. After making the series stationary, I used the training data to run an AR(2) model. Through the analysis of arima errors of the AR(2) model, I came up with different sets of p and q values to be used by the MA,AR,ARMA models.After running these different models, ARIMA(4,0,2) came up as the best model with the lowest AIC,even though auto.arima suggested an ARIMA(4,0,0) model.

Forecasting:

After diagnosing the residuals of our finalised model, I used the model to forecast using the observation from the test data.The mean absolute error of our forecasts came to be around 11 pokémons.

Conclusion:

I used ARMA(4,2) model to forecast total number of Pokemon occurrences with an MAE of 11 pokemons. A similar approach can be used to forecast events occurring with irregular frequency.