Predictive Business and Finance [EM1415]*

Instructor: Dr. Dario Palumbo

October 17, 2023

Project: The Forecasting Tourism 2010 Competition (JIF)

Background

Tourism is one of the most rapidly growing global industries and tourism forecasting is becoming an increasingly important activity in planning and managing the industry (from kaggle.com).

Problem Description

The 2010 tourism forecasting competition had two parts. In part I, competitors were tasked with producing forecasts for the next four years, given 518 series of annual tourism data

Available Data

The data consists of 518 annual series, each related to some (undisclosed) tourism activity. Tourism activities include inbound tourism numbers to one country from another country, visitor nights in a particular country, tourism expenditure, etc. The series differ in length, ranging from 7-year series to 43-year series. They also differ in the order of magnitude of values. The data is available at `www.kaggle.com/c/tourism1/Data`. download the file "tourism_data.csv" (this will be also available on the Moodle page of the course).

Assignment Goals

This case will give you experience with forecasting a large number of series. While the competition goal was to achieve the highest predictive accuracy, the goal of this case is to highlight different aspects of the forecasting process that pertain to forecasting a large number of series. For example, easy-to-use visualization tools have a significant advantage for visualizing a large number of series.

Note that the multiple series in this case are not likely to be forecasted together. However, it is common to forecast the demand for a large number of series, such as products in a supermarket chain.

Another aspect of this case is the use of different predictive measures, and handling practical issues such as zero counts and summarizing forecast accuracy across series.

*This project is based on previous material by Dr. Anthony Osuntuyi. All the remaining errors are mine.

The assignment provides guidelines for walking you through the forecasting process. Remember that the purpose is not winning the (already completed) contest, but rather learning how to approach forecasting of a large number of series.

Assignment

1. Plot all the series (an advanced data visualization tool is recommended) - what type of components are visible? Are the series similar or different? Check for problems such as missing values and possible errors.

2. Partition the series into training and validation, so that the last 4 years are in the validation period for each series. What is the logic of such a partitioning? What is the disadvantage?

3. Generate naive forecasts for all series for the validation period. For each series, create forecasts with horizons of 1,2,3, and 4 years ahead (F_{t+1}, F_{t+2}, F_{t+3}, and F_{t+4}).

4. Which measures are suitable if we plan to combine the results for the 518 series? Consider MAE, Average error, MAPE and RMSE.

5. For each series, compute MAPE of the naive forecasts once for the training period and once for the validation period.

6. The performance measure used in the competition is Mean Absolute Scaled Error (MASE). Explain the advantage of MASE and compute the training and validation MASE for the naive forecasts.

7. Create a scatter plot of the MAPE pairs, with the training MAPE on the x-axis and the validation MAPE on the y-axis. Create a similar scatter plot for the MASE pairs. Now examine both plots. What do we learn? How does performance differ between the training and validation periods? How does performance range across series?

8. The competition winner, Lee Baker, used an ensemble of three methods:

- Naive forecasts multiplied by a constant trend (global/local trend: "globally tourism has grown "at a rate of 6% annually.") - Linear regression - Exponentially-weighted linear regression

(a) Write the exact formula used for generating the first method, in the form $F_{t+k} = ...$ (k = 1, 2, 3, 4)

(b) What is the rational behind multiplying the naive forecasts by a constant? (Hint: think empirical and domain knowledge)

(c) What should be the dependent variable and the predictors in a linear regression model for this data? Explain

(d) Fit the linear regression model to the first five series and compute forecast errors for the validation period.

(e) Before choosing a linear regression, the winner described the following process:

"I examined fitting a polynomial line to the data and using the line to predict future values. I tried using first through fifth order polynomials to find that the lowest MASE was obtained using a first order polynomial (simple regression line). This best fit line was used to predict future values. I also kept the R^2 value of the fit for use in blending the results of the prediction."

What are two flaws in this approach?

(f) If we were to consider exponential smoothing, what particular type(s) of exponential smoothing are reasonable candidates?

(g) The winner concludes with possible improvements one being "an investigation into how to come up with a blending *ensemble* method that doesn't use much manual twerking would also be

of benefit". Can you suggest methods or an approach that would lead to easier automation of the ensemble step?

(h) The competition focused on minimizing the average MAPE of the next four values across all 518 series. How does this goal differ from goals encountered in practice when considering tourism demand? Which steps in the forecasting process would likely be different in a real-life tourism forecasting scenario?

Tips and Resources

- The winner's description of his approach and experience: `blog.kaggle.com/2010/09/27/https://medium.com/kaggle-blog/phil-brierley-on-winning-tourism-forecasting`.

- Article "The tourism forecasting competition", by Athanasopoulos, Hyndman, Song and Wu, International Journal of Forecasting, April 2011. `www.robjhyndman.com/papers/forecompijf.pdforhttps://doi.org/10.1016/j.ijforecast.2010.04.009`.