DA3_assignment3_swimming pool ticket sales

Tamas Stahl

14/02/2021

Aim of the analysis

This is the third and final assignment for data analysis three. The tasks for assignment three included building a daily predictive model that can forecast the sales of outdoor pools in Albequerque 12 months ahead. The assignment is based on the use case study of chapter 18 of the data analysis curriculum, with the difference to predict for all the outdoor pools not only a selected one.

Data

My task was to use the Albequerque (ABQ) swimming pool dataset. The swimming pool admissions dataset contains information about the number of people that were admitted into one of the city pool facilities of the city of Albequerque. The dataset contains information on type of admission, location, quantity, admission cost, date. Please find the csv file under raw data in the corresponding GitHub repo of mine.

Data preparation

The data is transaction level ticket sales for each swimming pool in ABQ, available due to the open data policy of the city. This is admin data with a lot of observations, around 1.5 million, and is collected automatically from point of sales or terminals. The data is dated from 1999 to November 2017. In order to be up-to-date I used the time period between January 1, 2010 and December 31, 2016, which is exactly 7 years.

Furthermore, I focused on ticket types that reflect normal business: adult, teen, senior, child and toddler. I aggregated the transaction-level data into daily frequency. The task was to include all outdoor pools, for which I used the website: https://www.cabq.gov/parksandrecreation/recreation/swimming/outdoor-pools. Here I could filter down to the outdoor pools and water sprays.

I decided to drop the water sprays so eventually I ended up with 7 pools that qualified as outdoor pools. The pools have the following codes: 'AQEI01', 'AQEJ01', 'AQMP01','AQRG01','AQSP01', 'AQSV01', 'AQWP01. The first two letters stand for the city and the next two letters are the abbreviations of the names of the outdoor pools, like 'EI' stands for 'Eisenhower Pool'.

ETA

From Figure 1 we could distinguish clearly that there is seasonality in our data. The left chart shows only the year 2015, meanwhile, the chart on the right shows us all the years in our work set (2010-2015). It is important to highlight that there is strong seasonal variation in daily volumes, but there is no visible trend.

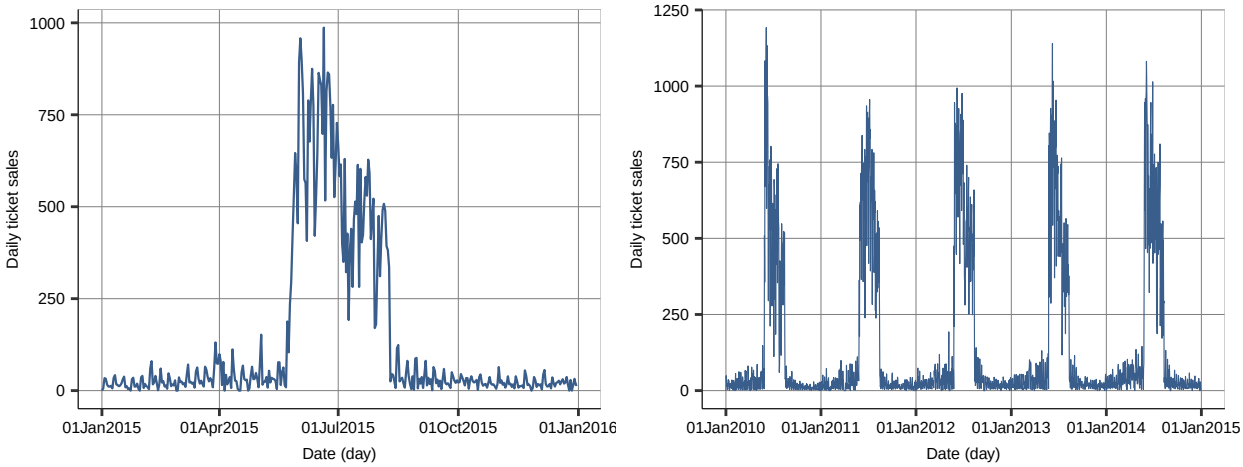

Figure 1: Seasonality in tickets sold

Just like in the in-class use case study the plotted time series show the importance of monthly seasonality but it is not the ideal method to capture the magnitude of those seasonal differences. For that I would recommend using boxplots plotted for days of the week and months of the year.

It is easy to spot the strong monthly seasonality, with way more visitors in the summer months than in any of the other months of the year. Furthermore, it is easily visible that weekends are way more frequent than weekdays.

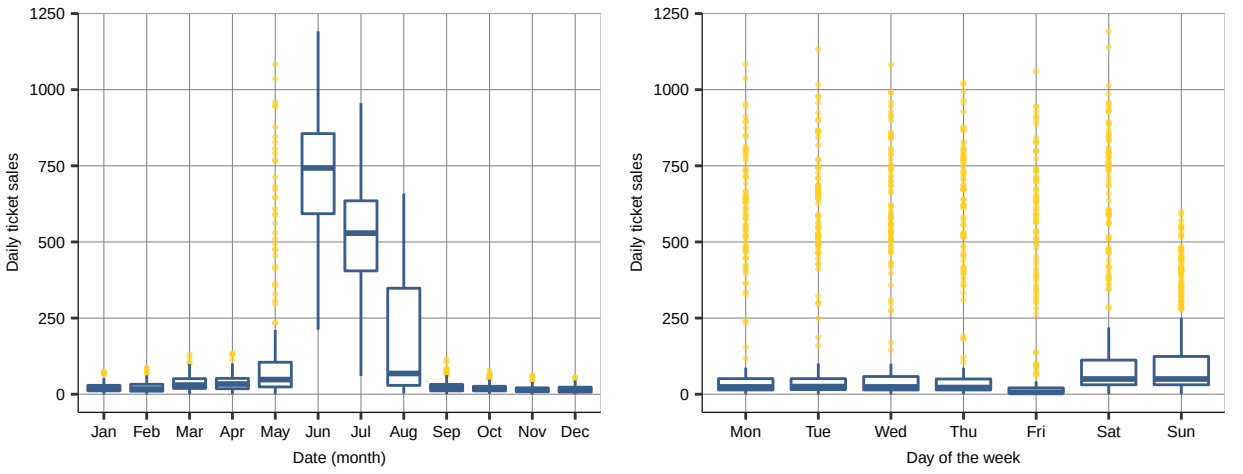

Figure 2: Monthly and daily seasonality in tickets sold

The last visualization for our seasonality in the data is a heatmap for days of months and tickets sold. The horizontal axis is days of the week, the vertical axis is months of the year and the color is the average daily sold tickets over the years.

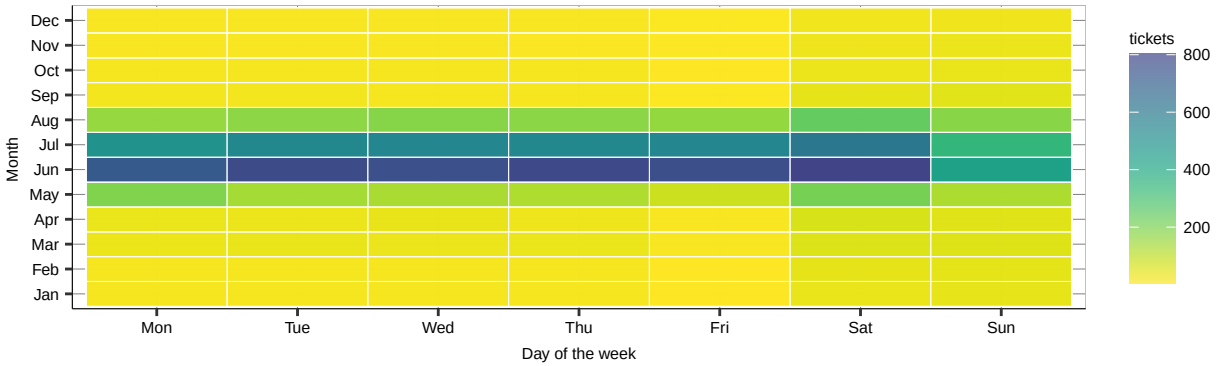

Figure 3: Heatmap

Model building

My task is to build at least three models to predict the tickets sale for the 7 outdoor pools in ABQ. Eventually, I ended up with the 7 models we used in class.

1. Model 1 has trend and monthly binary variables
2. Model 2 adds to Model 1 the day-of-the-week binary variables
3. Model 3 adds to Model 2 the holidays binary variables
4. Model 4 adds to Model 3 the school*days interaction
5. Model 5 adds to Model 4 the days*months interaction
6. Model 6 is Model 4 with log y variables
7. Model Prophet which is facebook's model building algorithm

As mentioned earlier our work set is the data between January 1, 2010 and December 31, 2015. Meanwhile, the holdout set is the data for the whole year of 2016.

Table 1 summarizes the results of our 6-fold cross-validation to select the best model. According to the results, our best-performing model is Model Prophet, which has an CV RMSE of 99.052 which is significantly better than the second best Model 5 with 106.297.

Table 1: RMSE for the 7 models

	CV RMSE
reg1	119.430
reg2	118.132
reg3	117.756
reg4	107.024
reg5	106.297
reg6	159.025
prophet	99.052

I decided to evaluate the actual fit of our two best models, Model 5 and Model Prophet, on th entire work set and applied its prediction on the holdout set. **The RMSE of Model 5 on the holdout set is 96.184. Meanwhile, it is 83.275 for the RMSE of Model Prophet on the holdout set.** We can see that both of the values are smaller than its cross-validated counterparts. This means that our patterns of association behind our prediction remained stable for the test set as well.

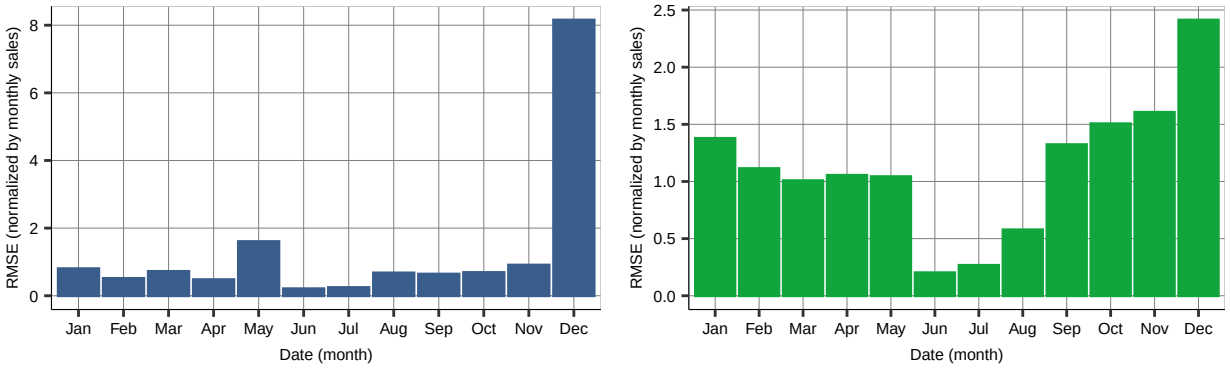

Figure 4: Montly RMSE on the holdout set (2016) - model 5 (left), model prophet (right)

From Figure 4 we can see that model 5 has unusually large errors in December. This could be the result of the difference in the holiday periods for Christmas, etc. Model Prophet (right chart) eventually performs much consistently. It has also the largest errors for December, however, it is by far not that outstanding as for model 5.

In order to fix the large errors of December, we should consider adding precise days for the Christmas period.

Pediction visualization

Figure 5 shows the predicted values of tickets to be sold vs the actual tickets sold for the 2016 (left chart) and for August, 2016 (right chart). The second chart is easier to interpret as we could see the difference between our prediction and the actual values with yellow.

It is easy to see that our prediction avoids the spikes of the actual values for August 2016. We can cl=onclude that our prediction does not take into account the highs and lows of the actual volumes.

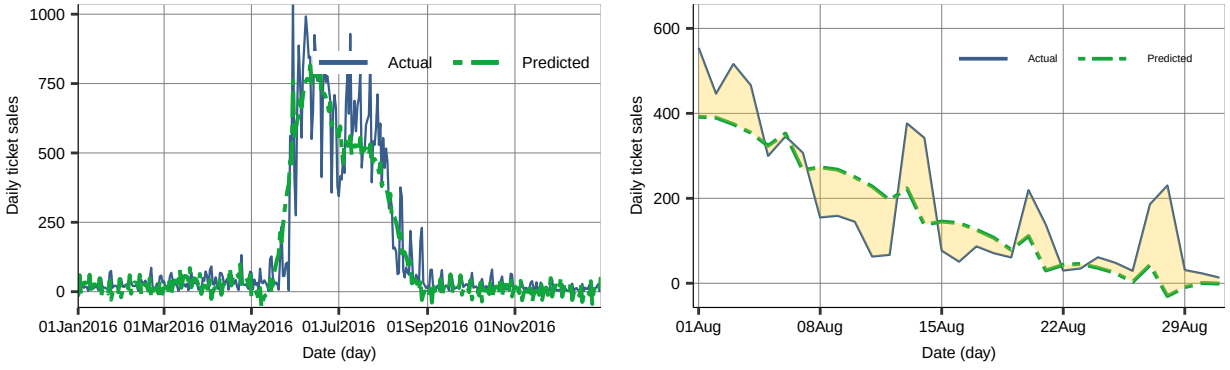

Figure 5: Predicted values - Daily ticket volume forecast (left), predicted vs actual in August 2016 (right)

Conclusion

I would like to highlight that my prediction had more data than in class due to the fact that instead of one outdoor pool I included all the outdoor pools in ABQ. We can conclude that the RMSE of my models are higher than of the use case study, which is due to the more data.

Also my prediction is more smoother than the one we did in class. Figure 5, right chart shows a smooth line. My educated guess is that the more outdoor pools resulted in better understanding of our data. The spikes were not that outstanding than in case of one pool.

I can conclude that seasonality can be established in case of the 7 outdoor pools and is important in case of long-horizon forecasts. Eventually the best model is Model Prophet, which outperformed all the other models.