Prediction challenge

Anas Zakroum

21/03/2021

Contents

Follow this link for the code

1 Introduction

The goal of this study is to produce a model that is able to predict the daily number of bicycles passing through a sensor from midnight to 09:00 AM on April 2nd. The sensor is placed in Albert 1st totem.
Towards that aim, we are given a data set consisting of the totem's daily snapshots for over a one year period. Multiple snapshots can be taken during the same day. ***We assume that the sensors are reset every day at midnight and from that we develop our framework.***

2 Methodology

As a first step, we extract a consistant time series from the data in which the values are the daily number of bicycles that passed through the sensor from 00:00 AM to 09:00 AM.
To get these values, we consider the last value taken before 09:00 AM and the first value taken right after 09:00 AM, then, we infer the value at 09:00 AM with a basic linear regression (two points).

Formally, let us denote by:
a : the number of bicycles that passed through the sensor right before *9:00 AM.*
b : the number of bicycles that passed through the sensor right after *9:00 AM.*
m_a : the number of minutes elapsed from midnight to the time a is taken.
m_b : the number of minutes elapsed from midnight to the time b is taken.

The estimated number of bicycles $\hat{V}$ that passed through the sensor from *midnight* to *9:00 AM* is given by:

$$\hat{V} = \left\lceil \frac{b-a}{m_b - m_a} \times 540 + \frac{a \times m_b - b \times m_a}{m_b - m_a} \right\rceil$$

When only one snapshot is taken during the day, we perform the regression using two points : the first one at midgnight, and the second one at the time the snapshot is taken.
These estimations are then computed for each day, which results in a time series consisting of the estimated number of bicyles that passed through the sensor from *midnight* to *9:00* AM, as shown in Figure 1.

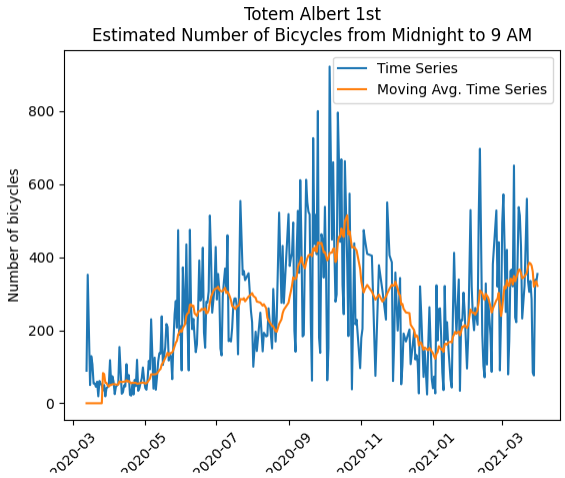

Figure 1: Estimated number of bicycles from midnight to 9:00 AM.

We observe that the time series is non-stationary in terms of first and second moments with significant fluctuations. These latter may be due to exogenous factors such as weather conditions, holidays or pandemic conditions.
Because of the fluctuations, we think it is more convinient to perform the prediction over the ***time series trend*** rather than on the ***time series itself***. In that regard, we smoothen it by computing the moving average over a 14-days window in an attempt to catch the trend, as shown in Figure 1 (orange curve).

Next, to model the time series, we use the following auto-regressive model.

$$X_t = w_0 + \sum_{i=1}^{p} w_i X_{t-i} + \varepsilon_t$$

Where p is the order of the autoregressive model that we consider as design parameter. We vary p in $\{1, 2, \ldots, 14\}$.

To train the AR(p) model, we turn our time series into a supervised learning data set in which the features are the sequences $\{X_{t-p+1}, X_{t-p+2}, \ldots, X_t\}$ and the target is X_{t+1}. We split this data set into two subsets: a training set (75% of the data) and a testing set (the remaining 25%).
The features are then scaled to zero mean and unit standard deviation. To estimate the model's parameters, we use the normal equation:

$$\hat{W} = (X'X)^{-1}X'y$$

Where X is the feature matrix, y is the target vector and $\hat{W}$ are the estimated parameters.
Finally, we use the R^2 score for each $p \in \{1, \ldots, 14\}$ to select the best auto-regressive order. The R^2 scores are reported on the testing set.

3 Results

We use the scikit-learn python package which contains implementations of the linear regression model and functions to perform data preprocessing tasks. Figure 2 shows the R^2 scores in function of the auto-regressive order. We found that the optimal order is $p = 9$.

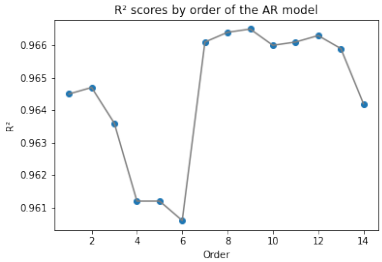

Figure 2: R2 scores for each value of the design parameter p

Finally we perform the prediction using $p = 9$ as the order of the autoregressive model.
The predicted number of bikes passing through the sensor between 00:00 AM and 9:00 AM is 307 bicycles.

4 Limitations

In order to estimate the number of bicycles between midnight and 9 am, we used a linear model. Even though this assumption is convenient due to its simplicity, it is certainly not optimal. A better approach could be for instance to collect for some days more data between midnight and 9 am in smaller time windows, then to perform a curve fitting without necessarily assuming the linearity hypothesis. However, we opted for the linear estimation because of the lack of such data; the majority of records contain very few data between midnight and 9 am (less than 3 records).