Time Series Forecasting Challenge Report

Filippo Brozzi, Manuel Cecere Palazzo, Daniele Manfredonia

January 21, 2022

Abstract

In this document we'll shortly present how we approached the challenge, which was our workflow and the issues we faced, how we decided to overcome some of them, and finally describe our best models, explaining the rationale of our choice and the results they produced

1 The challenge and the sequence

We were given a multivariate time series consisting of 68000 samples of seven different variables, and asked to produce a forecast model to correctly predict 864 future samples. Since no information on the nature of the data was provided, we first tried to get some statistical clues over the dataset, by performing some analysis that could produce useful results for our task.

2 Data Inspection

2.1 Stationarity analysis

We first checked if our series were stationary. The most commonly used test to see this is the Augmented Dickey Fuller test, where the null hypothesis is the time series possesses a unit root and is non-stationary. So, if the P-Value in ADF test is less than the significance level (0.05), you reject the null hypothesis.
A stationary time series is easier to predict, hence the need of this analysis.
As shown below, all the 7 variables fulfill the bound of the P-value, and so we can successfully claim that they are indeed stationary.

Series	**p-value**
Sponginess	4.427e-29
Wonder Level	8.583e-22
Crunchiness	0.0
Loudness on impact	1.016e-20
Meme creativity	5.714e-10
Soap slipperiness	0.001
Hype root	0.0

2.2 Relationships between variables

Observing some similarities between the structure of the variables, we decided to perform an analysis to search for correlated time series.
We then made a test to discover the **correlation coefficient** bounding couple of variables at the same step.
As shown in Figure 1, we found out that Hype Root was deeply correlated to Hype root, Crunchiness was with Loudness on Impact, while other couples of variables showed a correlation factor to be considered insignificant. Obviously, a correlation at the same time-step provides limited information on the dependencies between our variables, since it does not take into account past samples.
To have a better picture about these relationships, we used the **Granger causality test** for the first 20 lags to determine if one time series would be useful to forecast another.
If that was the case, the results could have given us hints to understand if it was better to decompose our problem in smaller problems.

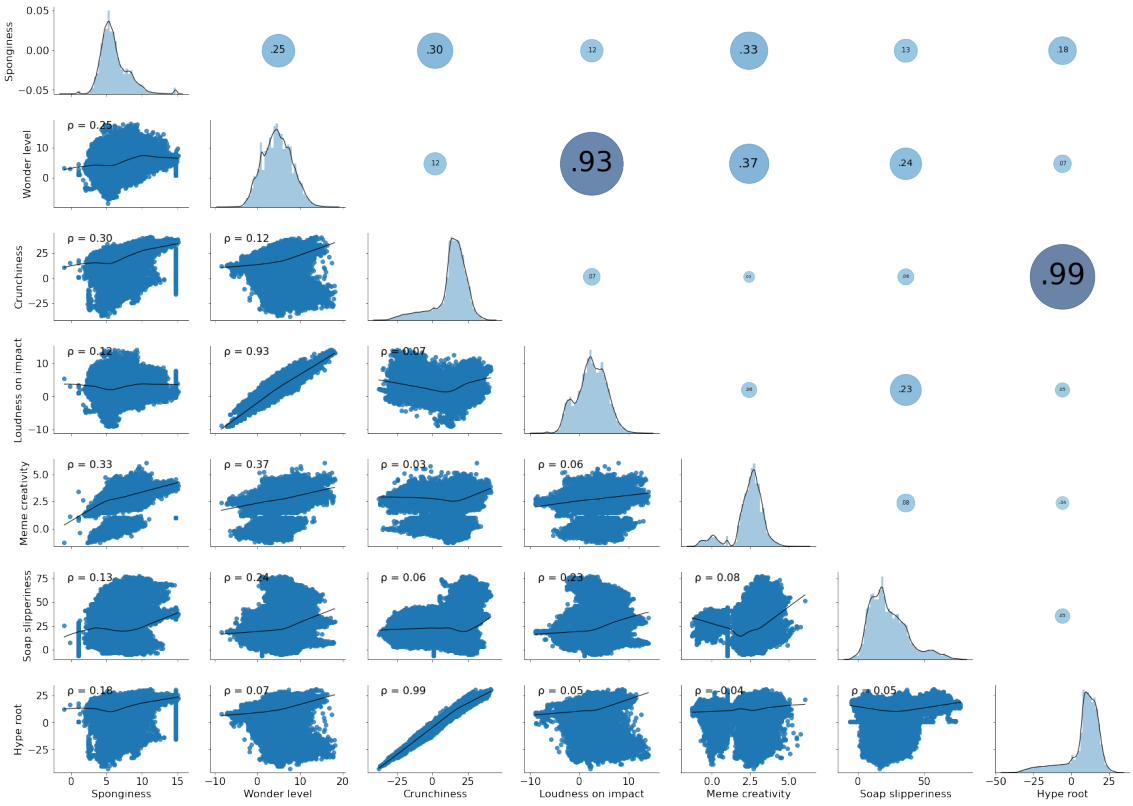

Figure 1: Correlation test results

The analysis concluded that Sponginess wasn't useful to predict Soap Slipperiness for some lags and Soap Slipperiness wasn't useful to predict Loudness on Impact for some lags. For the others, all the 7 variables were considered useful.

2.3 Autocorrelation and partial autocorrelation

These two function detect the correlation of a sequence with its past samples. More specifically the second one describe the direct influence of a sample at time t-k on the sample at time t, ruling out the influence of sample before time t-k on the sample at time t-k, while the first one does not perform this ruling out, hence carrying along the all the influence of past values.
In conclusion, as shown below as an example in Figure 2 the analysis gave us the hint that our series seemed to behave as autoregressive models with very short lags, as significant influence was detected only from recent past samples.

3 Models

3.1 Cross Validation and Normalization Choices

Since it is usually good practice to normalize the data in most models, we tried this approach to preprocess our time series. Nevertheless, the empiric results showed us worse results with respect to the non-normalized version, so we discarded this approach.
In addition, we decided to perform 5-fold cross validation on all of our models, to retrieve a better prediction of the final test set error. This method was particularly useful to select hyperparameters such as window and stride size.
However, as sometimes we trained and cross validate models to predict smaller windows of future samples than what was finally required to provide (the final 864 predictions were built in a regressive way), we noticed how our cross validation procedure was too optimistic as the window got smaller. To correct this behavior, we produced a new cross validation method, that would evaluate the model on predictions of length 864.

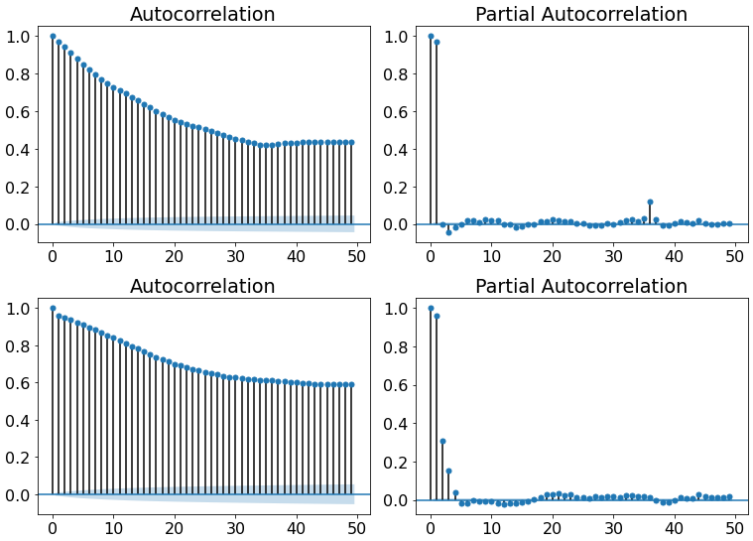

Figure 2: Autocorrelation and Partial Autocorrelation of two of the series: Sponginess (top) and Wonder Level (bottom)

3.2 Final Model

Our model consists of two bidirectional LSTM layers, the first one followed by a Conv1D and a MaxPooling1D layer, the second one followed by a Conv1D layer and a GAP layer. At the end a dense layer processes the final predictions. Our model takes as input the last 300 samples and predict the following 50. It then constructs the final 864 samples by concatenating the outputted 50 values and feeding back to the model, in a regressive way. To obtain better results, we introduced a callback to decrease learning rate if training reaches a plateau.
With this handcrafted model, we reached an RMSE of 4,2. [Figure 3]

3.3 Different architectures tried

In alternative to the bidirectional LSTM architecture, we tried and explored different approaches, such as the **GRU** architecture, a **sequence-to-sequence** model and **model-ensemble**, however, they produced no improvement, and so we discarded them.
We also tried to use the **percentage change** of all of our variables as an additional ingredient to produce our prediction, but again no improvement was reported.

3.4 Different approaches following the results of the data inspection

In consideration of the results of our data inspection, we tried also to follow the indication obtained through our experiments. More specifically:

- The **low correlation** between variable lead us to try to build five different models. Three of them would predict a single variable using only its past values. Two of them would predict a couple of highly correlated variables using only past values of that specific couple of variable (i.e. Hype Root and Crunchiness to predict Hype root and Crunchiness)
- The **Granger Causality Test** results lead us to create two separate models: one to predict Loudness on Impact ruling out Soap Slipperiness and one to predict Soap Slipperiness ruling out Sponginess
- The **autocorrelation and partial autocorrelation** results lead us to try to train our model to predict a single future sample using a very limited window of past values. All of these experiments showed no improvement, and so those model were discarded.

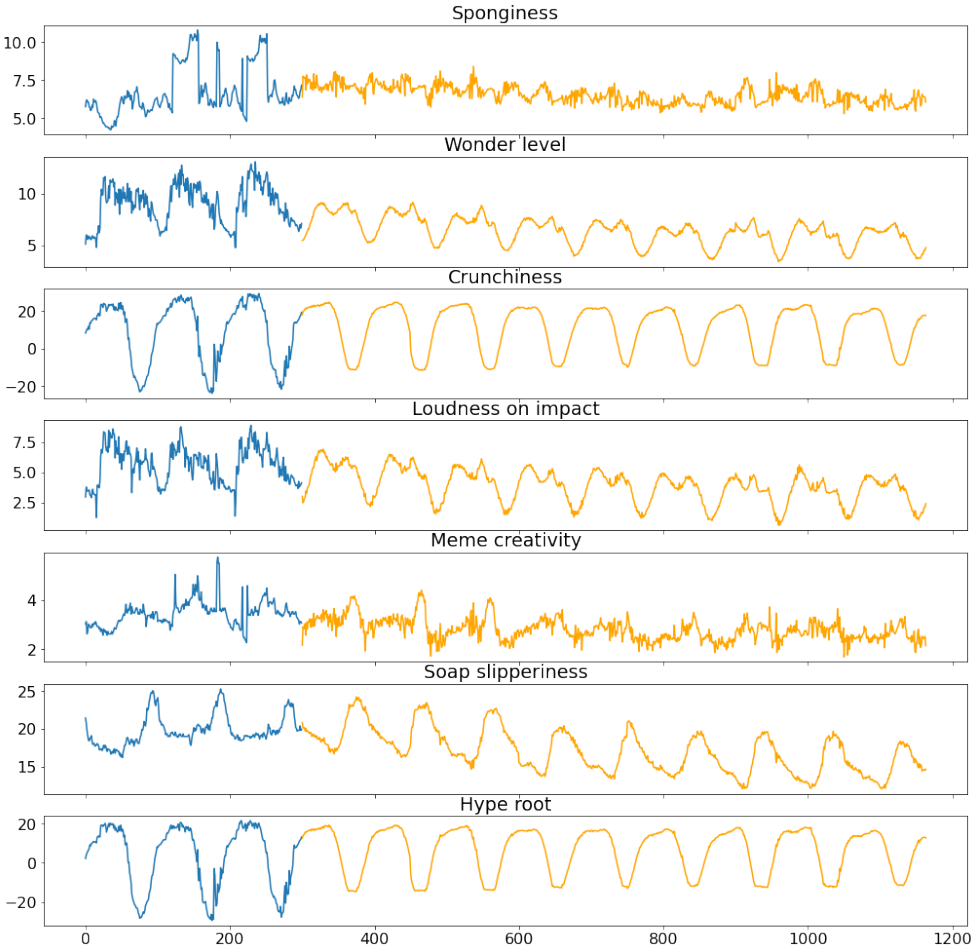

Figure 3: Future predictions per variables of our final models