Working Paper

Topic: Weather Forecasting

Students: Akritov Mark, Danielyan Anna

Instructor: Baghdasaryan Vardan

Introduction

When talking about the weather one can associate it with sun, clouds, wind, rain or even a thunderstorm. More commonly, we can say that weather is a mean of redistributing Earth's heat around its surface. Thus weather relates to almost all the aspects of our lives including agriculture, transportation and even distribution of humans around the globe and development of different cultures.

Accurate weather forecasts are crucial for planning our day-to-day activities. Almost every individual takes a look at the weather daily forecast, that helps us to make more informed decisions, and may even help keep us out of danger.

Modern weather forecasting incorporates different computer models, research, and measurement of common trends and patterns. With the application of these tools, accurate forecasting of the weather can be made up to about three to five days in advance. Forecasting further than that is not meaningful, because of the atmospheric conditions such as temperature and wind direction are very complex.

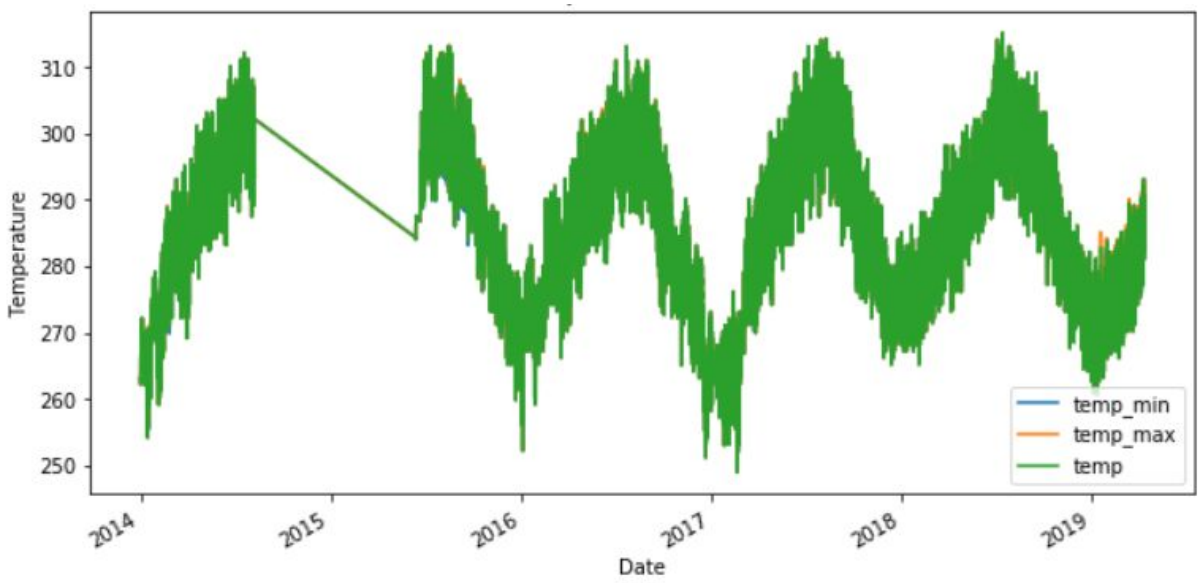
In this particular working paper, we are going to do a temperature forecasting using weather dataset of Yerevan. The motivation behind it is the prediction of daily temperature based on the features and conditions available on the observed data.

The work will be developed as follows: first, the data will be presented in detail discussing included features to be used in the modelling stage. Further, we will examine a few models that might be consistent with the collected data and theory behind it. As a culmination, the final models will be structured and presented. The endpoint of the work will be the sum-up of the main findings of the analysis.

Data

The paper analysis is conducted on the Yerevan weather data for the period from January 1, 2014, to April 11, 2019. Below the minimum, maximum and average temperatures are plotid. As it is expected there is a high correlation between these 3 variables so the simplicity we will take the average temperature as our dependent variable and drop the other two.

Graph 1. Temperature in time



In addition, the data is modified to be in days because of the missing observations for over a year (Graph 1) when observing the initial data in hours, while the temperature is converted to Celsius. Thus, the final data to be used consists of 9 variables (including dependent) and is shown in Table 1.

Table 1. Data Description

dt_iso	temp	temp_min	temp_max	pressure	humidity	wind_speed	wind_deg	clouds_all
2015-06-12	11.0235	11.0235	11.0235	869.0	87.0	1.0	189.5	0.0
2015-06-13	11.7770	11.7770	11.7770	871.0	85.5	1.0	175.5	0.0
2015-06-14	14.4400	14.4400	14.4400	871.0	83.0	1.0	15.0	0.0
2015-06-15	14.1330	14.1330	14.1330	867.0	68.0	1.0	3.0	0.0

Methods

Taking into consideration that we have time series data, before digging deep into models, first the stationarity of the dependent variable, which in our case is the daily temperature for Yerevan, will be checked. For this purpose, we will run a Dickey-Fuller test where our null hypothesis is that we have a non-stationary data. In case it is rejected, we can claim that our data is stationary, meaning that the process has no unit root, and doesn't have a time-dependent structure.

After modifying the data to be stationary, we will build two main models that are:

- ARDL (Autoregressive Distributed Lag) model which includes lags of both dependent and independent variables, as its name suggests. If we manage to include the right number of lags of y (temperature) and x (the rest of the features), we can eliminate serial correlation in the errors. ARDL model can also be transformed into a model with only lagged x's that go back to infinite past.
- SARIMAX (Seasonal Autoregressive Integrated Moving Average with the exogenous term) model which is used for time series forecasting with univariate data containing trends and seasonality. This is an extension to ARIMA that explicitly models the seasonal element in univariate data.

Modelling will be followed by an in-sample simulation. That is estimate the model up to April 11, 2019 (last observation of the data) then predict our dependent variable for the next 3 days.

Results

Summing up our main findings further, the key results are shown above including stationarity checking test and model processing according to it.

Initially, an Augmented Dickey-Fuller test (Table 2) for temperature variable showed that the absolute value of the ADF test statistics is less than the critical value at 5% of significance, |-2.0094|<|-2.864|, so we fail to reject the null hypothesis which is an indicator of non-stationary variable, whereas for the first difference ADF (Table 3) shows that temperature is stationary.

Table 2: ADF test for temperature

```
ADF Statistic: -2.009479
p-value: 0.282433
Critical Values:
        1%: -3.435
        5%: -2.864
        10%: -2.568
 Failed to reject HO: data is non-stationary
```

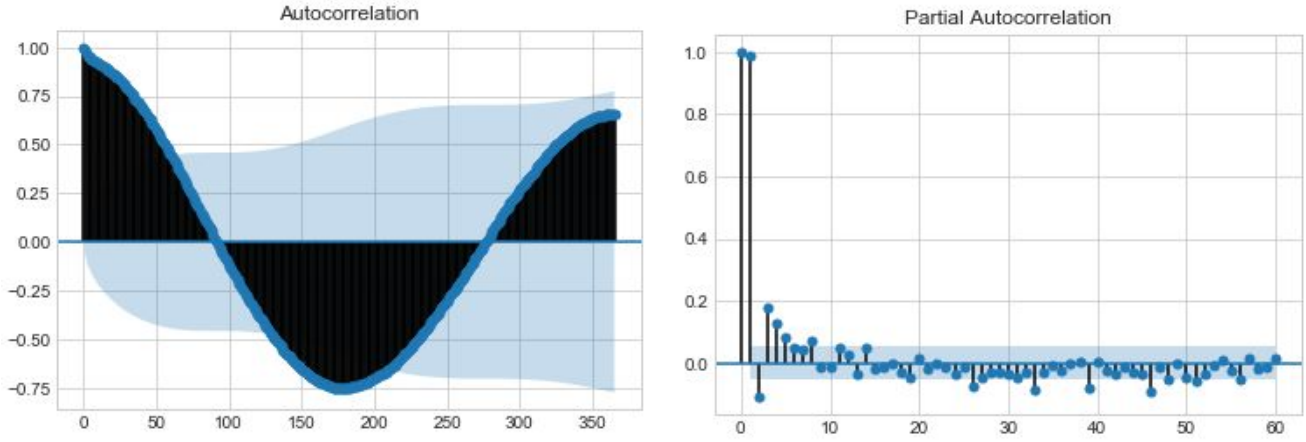
Table 3: ADF test for the first difference in temperature

```
ADF Statistic: -18.198600
p-value: 0.000000
Critical Values:
        1%: -3.435
        5%: -2.864
        10%: -2.568
 Reject HO: data is non-stationary
```

For introducing the ARDL model, the autocorrelations and partial autocorrelation in the dependent variable are reviewed in Graph 2**.** Here one can see that in the case of autocorrelation, a significant correlation within lags of the temperature goes up to 60 as each lag is being compared to the one before

it. Meanwhile, there is a totally different picture for the partial autocorrelation as is shows the correlation between each lag and the initial temperature observation (in our case it is June 13, 2015).

Graph 2. Autocorrelation and Partial Autocorrelation of temperature



Now, when we have enough information on our data, we can start our predictions including all above-mentioned details into the model.

The brief output of ARDL model can be found in Table X3. Lags of all independent variables are included in the model together with the first difference of temperature. Latter as one can remember came from fixing the non-stationarity of the data. Also, only the first lags of independent variables included in the model for making it more simple and precise. As one can see, all the included variables are statistically significant, while the speed of the wind at the current time was excluded for ending up with an insignificant coefficient in the model. From the table above it is obvious that the variables impacting on the temperature are actually its first lag and difference which actually makes sense when referring to the problem in general approach.

The formula for the model we have got can be written as follows:

$$temp = \beta_0 * pressure_t + \beta_1 * humidity_t + \beta_2 * wind\_deg_t + \beta_3 * clouds\_all_t + \beta_4 * temp_{t-1} + \beta_5 * clouds\_all_{t-1} + \beta_6 * pressure_{t-1} + \beta_7 * humidity_{t-1} + \beta_8 * wind\_speed_{t-1} + \beta_9 * wind\_deg_{t-1} + \delta temp$$

Table X3. ARDL output

pressure	5.898e-16	7.07e-17	8.344	0.000	4.51e-16	7.28e-16
humidity	-1.527e-16	8.07e-17	-1.891	0.059	-3.11e-16	5.7e-18
wind_deg	2.845e-16	3.21e-17	8.863	0.000	2.22e-16	3.47e-16
clouds_all	3.79e-16	4.34e-17	8.730	0.000	2.94e-16	4.64e-16
lag_temp1	1.0000	5.34e-17	1.87e+16	0.000	1.000	1.000
lag_clouds1	3.955e-16	4.44e-17	8.912	0.000	3.08e-16	4.83e-16
lag_press1	-2.012e-16	7.05e-17	-2.854	0.004	-3.4e-16	-6.29e-17
lag_hum1	-3.643e-16	8.01e-17	-4.550	0.000	-5.21e-16	-2.07e-16
lag_speed1	-4.458e-16	4.19e-17	-10.641	0.000	-5.28e-16	-3.64e-16
lag_deg1	3.469e-16	3.21e-17	10.792	0.000	2.84e-16	4.1e-16
temp_diff1	1.0000	2.03e-16	4.94e+15	0.000	1.000	1.000

Next, the errors of the model are predicting and tested for stationarity using the same ADF test (Table X4) as for the temperature case. Here, with 95% of confidence, we can claim that the obtained errors are stationary, which means that the model works as it should.

Table X4. ADF test for errors

```
ADF Statistic: -17.988852
p-value: 0.000000
Critical Values:
        1%: -3.435
        5%: -2.864
        10%: -2.568
 Reject HO: data is non-stationary
```

Unlike ARDL the SARIMAX model is built using only the initial variables and 2 lags of the dependent variable that is temperature. In addition, it also includes the seasonal element which makes the results of the model more accurate by taking into account the seasonal feature of the weather data.

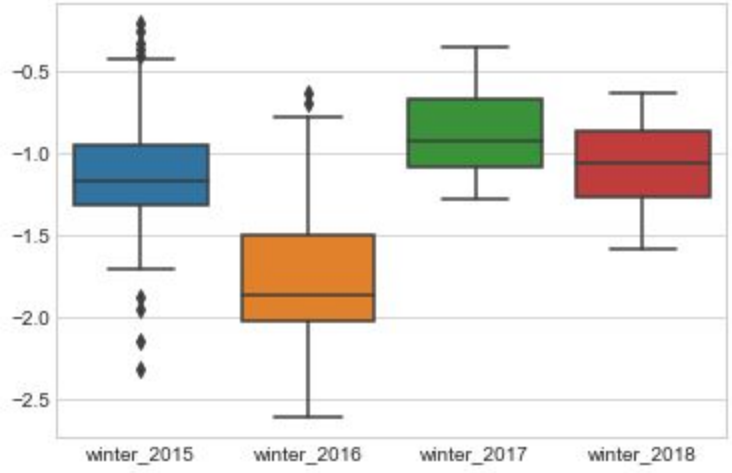
The key finding for this model is the temperature prediction for the upcoming three days results of which are sown in Table X5.

Table X5. Temperature forecasting for upcoming 3 days

Day	Temperature
April 12, 2019	9.778953
April 13, 2019	9.746843
April 14, 2019	9.777626

Another analysis that was conducted checks for the differences in means of the temperature during the winter period. To get the hypothesis of having the last two winters warmer compared to previous ones, the T-test was used which indicates that the mean differences are normally distributed only for the years 2018 and 2016. That is also obvious from the graph above.

Graph X2. Mean distribution of the temperature during the winter (2015-2018)



Last model that was used to forecast weather is Recurrent Neural Network with LSTM(Long short term memory) layers.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_4 (LSTM)                (None, 1, 64)             24064
_________________________________________________________________
dropout_3 (Dropout)          (None, 1, 64)             0
_________________________________________________________________
lstm_5 (LSTM)                (None, 1, 64)             33024
_________________________________________________________________
dropout_4 (Dropout)          (None, 1, 64)             0
_________________________________________________________________
lstm_6 (LSTM)                (None, 16)                5184
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 17
=================================================================
Total params: 62,289
Trainable params: 62,289
Non-trainable params: 0
_________________________________________________________________
```

From the RNN model's summary above we can see that model was fit on data with 5 hidden layers(from which 3 LSTM and 2 Dropout Sequentially stacked layers) and 1 Dense output layer. Each hidden layer besides last has 64 hidden units, the last one has 16 and the output Dense layer has 1(regression). There are 62289 trainable parameters in total.

Then model was compiled with RMSprop optimizer, and Mean squared error loss function. Mean absolute error was used as metric to evaluate model's performance. All mentioned hyperparameters are result of training model with different set of them(learning rate, optimizers, layer and unit count, epochs count, regularization etc.). Other approaches used to train model tend to overfit it.

The initial data was splitted to train and test sets and fitted to model.

The final evaluation on test set resulted mean absolute error equal to ~ 0.04 for standardized output.

Below we can see model predictive power visualized for actual and predicted values of final model for test set.