Rainfall Prediction

A ML LAB Project Report Submitted
in Partial Fulfillment of the Requirements
for the completion of

Machine Learning Lab course

by

BASU VERMA
(142002007)

under the guidance of

Dr Sahely Bhadra

IIT PALAKKAD

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD
PALAKKAD - 678557, KERALA

Acknowledgement

I would like to thank Dr. Sahely Bhadra for providing us the opportunity to do a project to get a practical exposure to what we learnt in the Machine Learning course, we thank Ms. Shikha Mallick, Ms. Abilasha for their valuable tutorials,feedback and support. We thank each other for our mutual understanding and support, it has been a great experience.

Jan 30, 2021
IIT Palakkad

Group g4

Abstract

Rainfall is one of the most important phenomena of the natural system. In Kerala, agriculture largely depends on the intensity and variability of rainfall. Therefore, an early indication of possible rainfall can help to solve several problems related to agriculture, climate change and natural hazards like flood and drought. Rainfall forecasting could play a significant role in the planning and management of water resource systems also. In this study, the univariate Autoregressive model was used to forecast monthly rainfall for twelve months lead-time for thirty-four rainfall stations of Kerala. The best AR model was chosen based on the root mean square error. A validation check for each station was performed on residual series. Residuals were found white noise at almost all stations. The predicted results from the selected models were compared with the observed data to determine prediction precision. We found that selected models predicted monthly rainfall with a reasonable accuracy. Therefore, year-long rainfall can also be forecasted using these models.

Contents

List of Figures

Part. 1

Introduction

1.1 Motivation

We need to build a model for local rainfall prediction in 1 to 7 days advance. Data has rainfall data captured in the various weather stations. you need to predict rainfall in all these station. For example, if one can feed rainfall data to 31/12/2020 of all station in your model, it will be able to predict rain fall for all station form 1/1/21 to7/1/21.

1.2 Data description

Given data was rainfall data for 34 stations from year 1970 to 2016 with date with frequency as Day and RF column showing rainfall on particular day.

1.3 Tools and Technologies

Programming languages	Python
Python libraries	Pandas,Numpy,Statsmodel,Joblib, Matplotlib
Other Tools	Jupyter Notebook,Colab Notebook, VS-Code

Table 1.1

Part. 2

Methodology

In this section we are going to discuss in detail about what testings we have done and information about which model we are using.

2.1 Data Preprocessing

Converted negative rainfall values to positive thinking it might be typing error.

Missing Value imputation:
For every dataset, first missing dates are filled with nan values. Nan values are then replaced for each date by the average rainfall data for other years on the same date (Orange colour graph represents the original values and blue colour graph represents the imputed values).

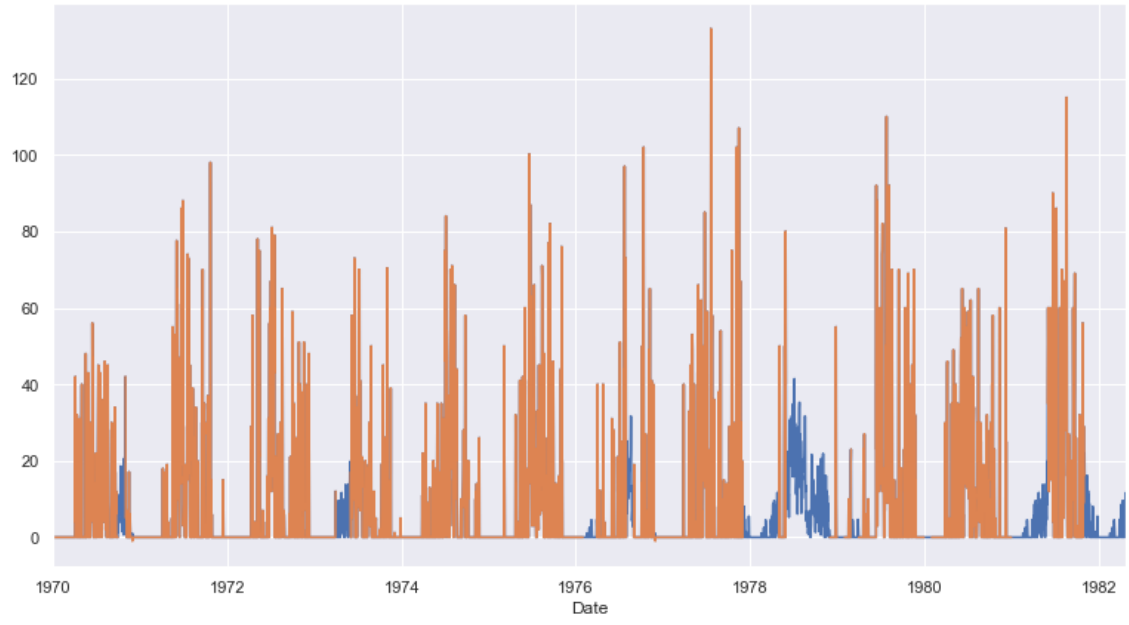

Fig. 2.1 Result after missing value Imputation

2.2 Testing

For testing data is splitted in 70-30 ratio. As our dataset is time series so splitted is done sequencially. Which means last values are saved as test data and first 70% as train data.

2.2.1 Autocorelation

An auto regression model makes an assumption that the observations at current and previous time steps are useful to predict the value at the next time step. This relationship between variables is called correlation. If both variables change in the same direction (e.g. go up together or down together), this is called a positive correlation. If the variables move in opposite directions as values change (e.g. one goes up and one goes down), then this is called negative correlation. We can use statistical measures to calculate the correlation between the output variable and values at previous time steps at various different lags. The stronger the correlation between the output variable and a specific lagged variable, the more weight that autoregression model can put on that variable when modeling. Again, because the correlation is calculated between the variable and itself at previous time steps, it is called an autocorrelation. It is also called serial correlation because of the sequenced structure of time series data. The correlation statistics can also help to choose which lag variables will be useful in a model and which will not. Interestingly, if all lag variables show low or no correlation with the output variable, then it suggests that the time series problem may not be predictable. This can be very useful when getting started on a new dataset. In this tutorial, we will investigate the autocorrelation of a univariate time series then develop an autoregression model and use it to make predictions.

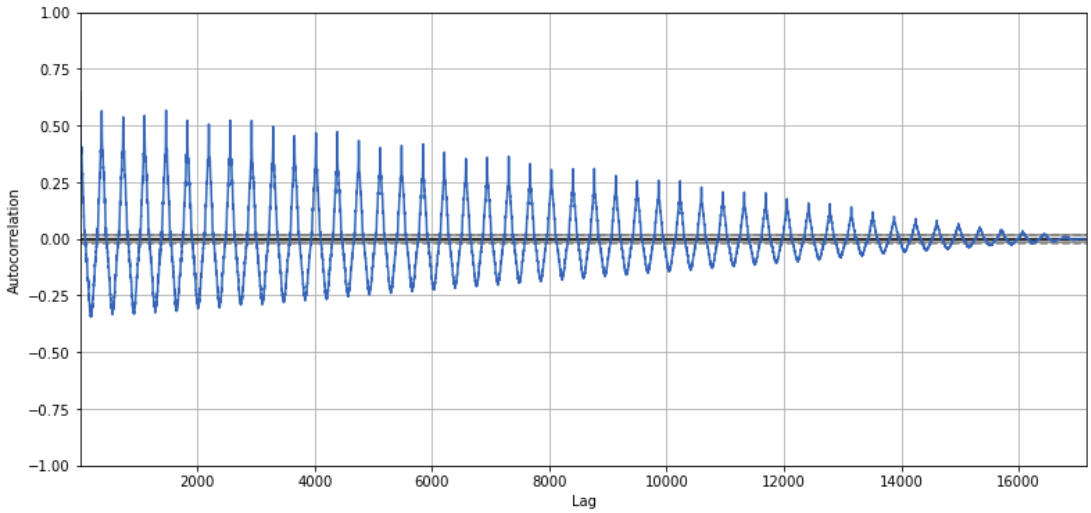

Fig. 2.2 Autocorrelation plot

Root mean squared error for finding optimum number of lags

We have calculated rms errors for different lags for all documents. Also we have checked time required for prediction using each lag value in which 365 is found to be optimum.

2.3 Forecasting

We have tried different models for forcasting like ARIMA, SARIMAX, AR with respective cross validation techniques. After trials we found Autoregression to be appropriate model for our dataset giving optimum results compared to others.

2.3.1 Autoregression

In an autoregression model, we forecast using a linear combination of past values of the variable. The term autoregression describes a regression of the variable against itself. An autoregression is run against a set of lagged values of order p.

Prediction can be calculation using equation,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where c is a constant, ϕ_n are lag coefficients up to order p, and ε_t is white noise.

For example, an AR(1) model would follow the formula

$$y_t = c + \phi_1 y_{t-1} + \varepsilon_t$$

Where y_t is the prediction, ϕ_0 and ϕ_1 are coefficients found by optimizing the model on training data. This technique can be used on time series where input variables are taken as observations at previous time steps, called lag variables. For example, we can predict the value for the next time step (t+1) given the observations at the last two time steps (t-1 and t-2). As a regression model, this AR(2) model would look as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

and so on.

For prediction using autoregressive model, we have used open source python library called statsmodels. In which we have used statsmodels.tsa.ar_model.AR module. statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

2.4 Results

Rms errors for all stations for 7 days of prediction are given below:

Stations	**Rmse**
Malappuram-Angadipuram	1.449
Malappuram-Manjeri	2.099
Malappuram-Nilambur	2.112
Malappuram-Perinthalamanna	1.76
Malappuram-Ponnani	1.712
Malappuram-Tiruvangadi	1.157
Palakkad-Alathur	0.411
Palakkad-Alattur	1.887
Palakkad-Cherapalaseri	1.651
Palakkad-Mannarkad	0.9
Palakkad-Ottapalam	1.694
Malappuram-Palakkad	0.943
Palakkad-OBSY	0.309
Palakkad-Parli	1.011
Palakkad-Pattembi	0.586
Thrissur-Chalakudi	2.957
Thrissur-Enamakkal	2.834
Thrissur-Kodungallur	1.554
Thrissur-Mukundarpuram	2.149
Thrissur-Ollukara	1.772
Thrissur-Peechi	0.724
Thrissur-Thalipilly	1.574
Thrissur-Thrissur	0.985
Coimbatore-Anaimalai	0.558
Coimbatore-Attakatti	1.192
Coimbatore-Nirardam1	0.77
Coimbatore-Parambikulam	0.995
Coimbatore-Pollachi	0.909
Coimbatore-Sholiyarnagar	0.616
Coimbatore-Solayar	1.61
Coimbatore-Topslip	1.804
Coimbatore-Nirar	1.338
Coimbatore-Valparai	1.457
Palakkad-Chittur	1.675

Prediction result for next 7 days:

Following figure shows prediction results for next seven days for station Palakkad-Cherapalaseri

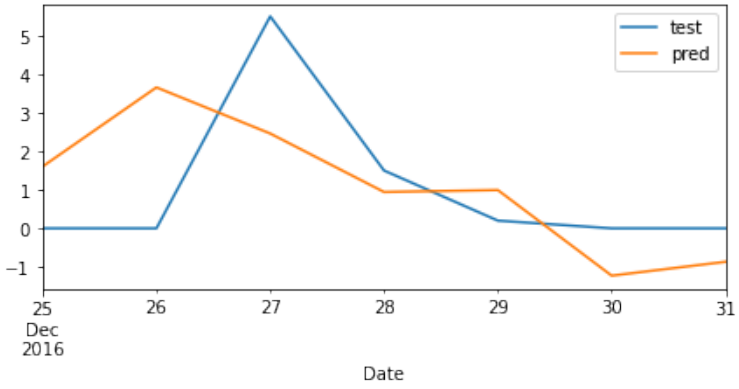

Fig. 2.3 Results

Here, based on data upto 24, Dec, 2016, next seven day prediction is made. Blue line showing actual values and orange line showing predicted values.

.

Part. 3

Conclusion

3.1 Conclusion

The techniques which we have learnt in our machine learning class like cross validation, missing data processing, time series analysis using Auto regression were used to build forecasting model. We have build this model with requirements given with scope for future updations.

3.2 Future-work

1. We want to increase the scope of this project with some user interface
2. We are willing to implement the prediction model using probabilistic model like Hidden Markov Model.

References

[1] Maulana Sidiq. Forecasting rainfall with time series model. *IOP Conference Series: Materials Science and Engineering*, 407:012154, 09 2018.

[2] Jason Brownlee. Autoregression Models for Time Series Forecasting With Python. `https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/#:~:text=Autoregression%20is%20a%20time%20series,range%20of%20time%20series%20problems.`, 2017.

[3] edureka. Time Series Analysis in Python. `https://www.youtube.com/watch?v=e8Yw4alG16Q&t=1114s`, 2018.

[4] Jason Brownlee. *Introduction to Time Series Forecasting With Python: How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery, 2017, 2017.