Conclusion of Rainfall Project

Vellore daily Rainfall dataset is sparse in nature and has higher stochastic component than deterministic component(seasonality). Despite making the series stationary by differencing, ARIMA does not perform well due to sparse data, which in statistics is called **intermittent demand data**. In order to improve forecasting accuracy, a plethora of techniques have been implemented ranging from machine learning to deep learning techniques. Given our usecase is time series data and regression techniques cannot be modeled on such data, we resort to **feature engineering**. In doing so, we reconstruct our series to a regular dataset with one-hot encoded new features such as month,year,day,district and is_rain. Furthermore, we apply linear regression, SVR and XGBoost on our new dataset. **Xgboost** has seen to yield the best test data MAE(mean absolute error) of **0.899. It is even able to handle sparse data and can be applied on the dataset as a whole, unlike other techniques which required to create a separate dataset for each district.**