PREDICTIVE ANALYSIS ON RAINFALL STATISTICS IN MAHARASHTRA

Sarthak Chudgar
Department of Instrumentation and Control Engineering
Vishwakarma Institute of Technology
Pune, Maharashtra, India
sarthak.chudgar17@vit.edu

Johnathan Fernandes
Department of Instrumentation and Control Engineering
Vishwakarma Institute of Technology
Pune, Maharashtra, India
johnathan.fernandes17@vit.edu

Aneesh Poduval
Department of Instrumentation and Control Engineering
Vishwakarma Institute of Technology
Pune, Maharashtra, India
aneesh.poduval17@vit.edu

***Abstract* - Rainfall plays an integral role in the lives of millions of people worldwide. This is especially true for an agriculture heavy country such as India. Not only do we depend on rain as a source of fresh water, we also use it to irrigate farms and for rainwater harvesting.**

Given our dependence on rain, prior knowledge about the weather forecast is essential for optimal rainwater utilization. This is where predictive analytics comes in.

***Keywords: Rainfall, Prediction, Maharashtra, Weather, Rain*.**

I. INTRODUCTION

Predictive analytics is the 3rd step in big data analytics. It involves the previous two steps, i.e. descriptive and diagnostic analytics, and aims to build a model which can analyze data trends and predict future trends in order to help individuals better prepare themselves.
In this project we will carry out predictive analytics on weather data over the state of Maharashtra to construct a predictive model which will be able to predict the amount of rainfall in millimeters.
While the primary focus of this project is in the field of agriculture, it also plays a vital role in other fields such as disaster management

II. ANALYTICS

A. Descriptive Analytics

This step involves obtaining data of our variable (rainfall, in this case) and possible related factors (weather properties that may or may not affect rainfall).
We then visualize our dependent variable (rainfall) with each of the independent variables (other factors).

We obtain our data from the National Centers for Environmental Prediction (NCEP) website. They have constructed a Climate Forecast System Reanalysis (CFSR) which provides worldwide weather readings from as far back as 1980

Figure 1: NCEP Map GUI

The NCEP website provides a map GUI to choose an area, and then provides all data from that area. While the selection includes 657 stations, Maharashtra only contains about 370 of them. We use Tableau to extract those stations. From the system, we obtain data on the following:

Weather Factor	Units & Remarks
Precipitation	Millimeters
Humidity	Fraction
Location	Latitude and Longitude
Temperature	°C, (Minimum and Maximum)
Wind speed	Meters per second
Solar Coverage	Mega joules per square meter

.Table 1: Selected Features

B. *Diagnostic Analytics*

Diagnostic analytics is closely related to descriptive to the point where they are usually carried out simultaneously. It entails using the previously created visualizations to determine the relations between our independent and dependent variable.

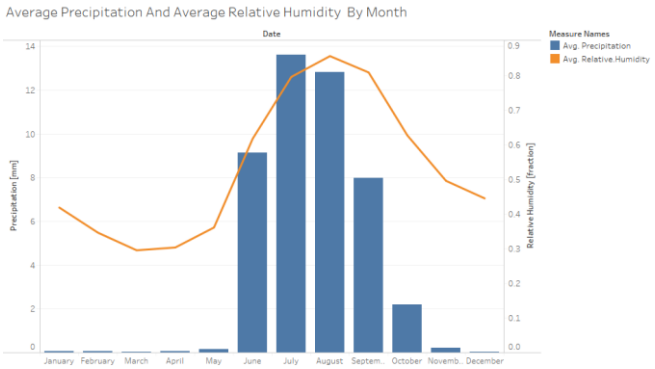

Figure 2: Average Precipitation and Average Humidity by Month

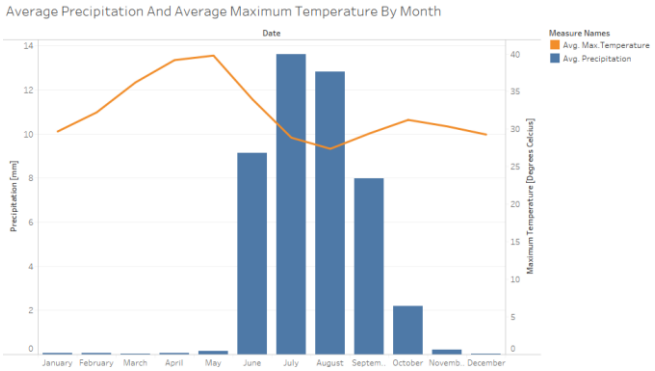

Figure 3: Average Precipitation and Average Maximum Temperature by Month

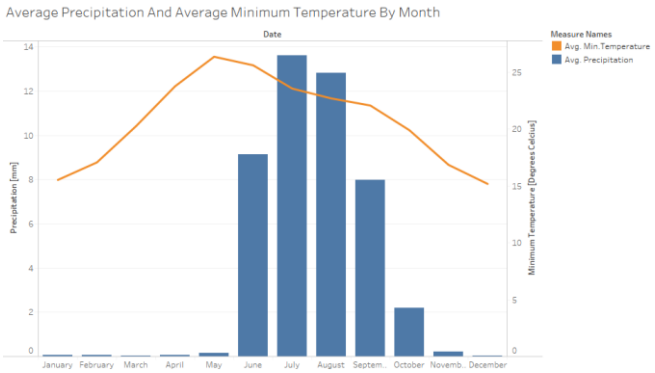

Figure 4: Average Precipitation and Average Minimum Temperature by Month

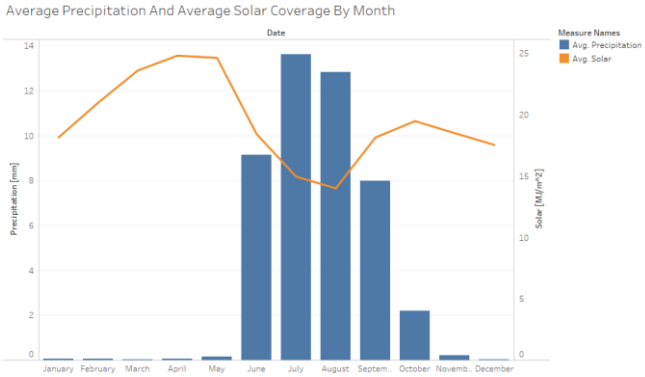

Figure 5: Average Precipitation and Average Solar Coverage by Month

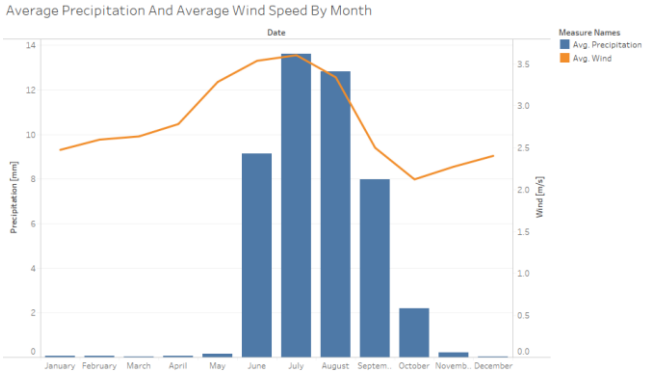

Figure 6: Average Precipitation and Average Wind Speed by Month

We use Tableau, a well-known business insights tool to visualize the data into various easy to understand charts.

C. *Pedictive Analytics*

Our third and final step involved importing data into a suitable program to construct a prediction model. Due to the sheer volume of the data considered, normal analysis applications such as MATLAB and Microsoft Excel are not suitable. Hence we use Apache Spark along with the R programming language to import the data. The initial step in building a predictive model is to clean up the data. This involves removing undefined null values and outliers.

```
> colSums(is.na(MH))
               X             Date         Latitude        Longitude
               0                0                0                0
 Max.Temperature Min.Temperature             Wind Relative.Humidity
               0                0                0                0
           Solar    Precipitation
               0                0
```

Figure 7: Analyzing null values

As observed, the dataset contains no null values.

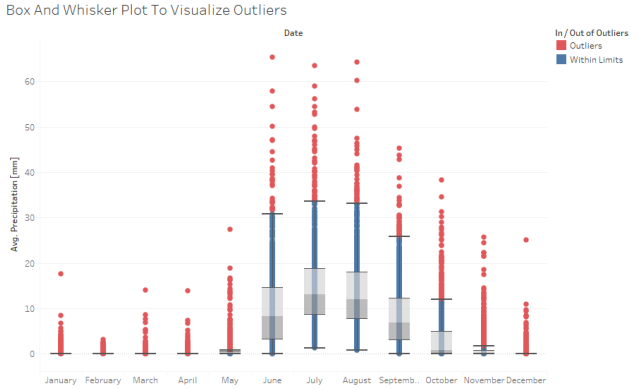

Figure 8: Analyzing outliers

Using tableau, we visualize the outliers in a box and whisker plot and remove them. In our plot, (Figure 8) each dot represents a single day.

After cleaning up null values and outliers, we scale and center the data points so as to maintain zero mean and unit variance. This reduces model computing time and improves model prediction.

```
> summary(MH)
       X                Date          Latitude       Longitude      Max.Temperature
 Min.   :      1   1/1/1980:    309   Min.   :15.77   Min.   :72.50   Min.   :11.21
 1st Qu.: 897646   1/1/1981:    309   1st Qu.:18.27   1st Qu.:74.06   1st Qu.:28.30
 Median :1795290   1/1/1982:    309   Median :19.51   Median :75.31   Median :31.01
 Mean   :1795290   1/1/1983:    309   Mean   :19.26   Mean   :75.82   Mean   :32.16
 3rd Qu.:2692935   1/1/1984:    309   3rd Qu.:20.45   3rd Qu.:77.50   3rd Qu.:35.56
 Max.   :3590580   1/1/1985:    309   Max.   :21.70   Max.   :80.62   Max.   :55.34
                   (Other) :3588726
 Min.Temperature       Wind          Relative.Humidity      Solar          Precipitation
 Min.   : 0.80   Min.   : 0.09359   Min.   :0.0304   Min.   : 0.00   Min.   :  0.000
 1st Qu.:17.35   1st Qu.: 1.93048   1st Qu.:0.3229   1st Qu.:16.97   1st Qu.:  0.000
 Median :21.50   Median : 2.53952   Median :0.5298   Median :20.03   Median :  0.000
 Mean   :20.84   Mean   : 2.81174   Mean   :0.5425   Mean   :19.12   Mean   :  4.514
 3rd Qu.:24.19   3rd Qu.: 3.40621   3rd Qu.:0.7832   3rd Qu.:23.47   3rd Qu.:  2.438
 Max.   :37.63   Max.   :19.14967   Max.   :0.9868   Max.   :29.54   Max.   :421.259
```

Figure 9: Original data values

```
> summary(transformed)
    Latitude          Longitude       Max.Temperature   Min.Temperature       Wind
 Min.   :-2.3293   Min.   :-1.5139   Min.   :-3.8367   Min.   :-4.0779   Min.   :-2.1115
 1st Qu.:-0.6627   1st Qu.:-0.8019   1st Qu.:-0.7052   1st Qu.:-0.7093   1st Qu.:-0.6846
 Median : 0.1706   Median :-0.2323   Median :-0.2098   Median : 0.1346   Median :-0.2115
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.7956   3rd Qu.: 0.7645   3rd Qu.: 0.6241   3rd Qu.: 0.6817   3rd Qu.: 0.4618
 Max.   : 1.6289   Max.   : 2.1886   Max.   : 4.2468   Max.   : 3.4176   Max.   :12.6918
 Relative.Humidity      Solar          Precipitation
 Min.   :-2.01711   Min.   :-3.1038   Min.   :-0.3543
 1st Qu.:-0.86510   1st Qu.:-0.3483   1st Qu.:-0.3543
 Median :-0.04988   Median : 0.1477   Median :-0.3543
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.94800   3rd Qu.: 0.7065   3rd Qu.:-0.1630
 Max.   : 1.74995   Max.   : 1.6923   Max.   :32.7078
>
```

Figure 10: Scaled data values

After cleaning up and scaling our data, we proceed to construct a random forest regression model in Spark using data from 1980 to 2010.

III. Prediction Model And Working

After cleaning up and scaling our data, we proceed to construct multiple prediction models in Spark using data from 1980 to 2010. After thorough testing, we determine that a random forest regression model has the best performance. We further test multiple models using differing numbers of trees and compare their root mean square error (in fig. 11) and determine that a model with 10 trees has the least error.

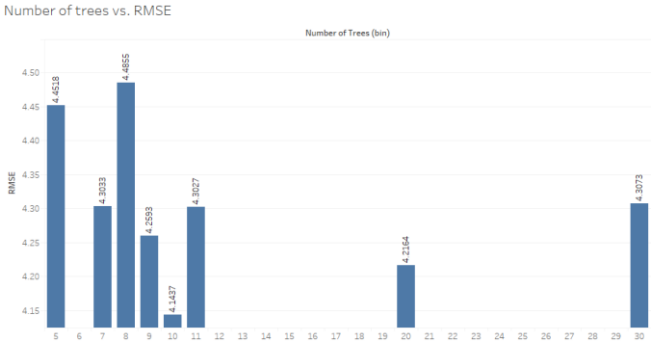

Figure 11: Number of trees vs. RMSE

IV. Model Results

In order to judge the performance of our model, we use testing data from the year 2011, and use this formula to predict the value of precipitation for each day. We first compare the predicted values to the actual values by using a scatter plot in Figure 12.

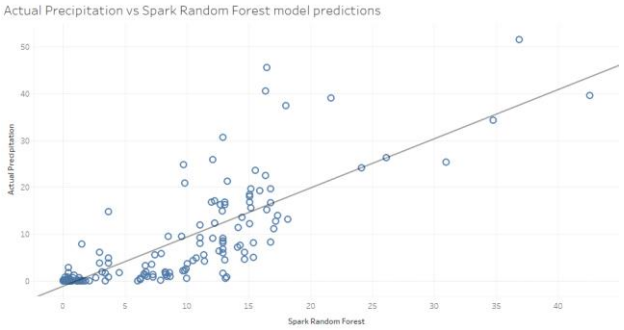

Figure 12: Actual vs Predicted values

The closer the trend line is to 45˚, the better the prediction. We also compare the monthly average predictions in Figure 13.

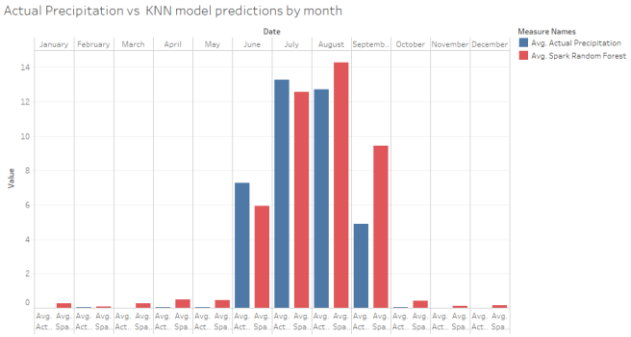

Figure 13: Monthly average predictions

Our model is able to predict rainfall with an average error of ±1.90 mm.

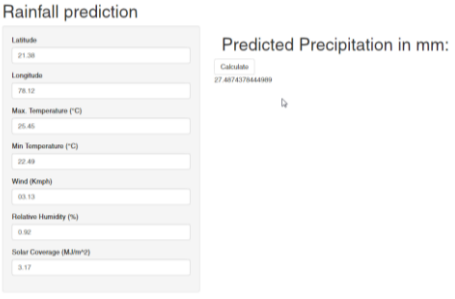

Figure 14: GUI

We utilized R and its "Shiny" package to build a basic GUI which allows a user to input the independent values and predicts the output with the click of a button.

V. Future Scope

We aim to improve our prediction by utilizing better modelling techniques and market this program in various countries.
We also plan on expanding this project to make it more versatile in terms of automatic data acquisition and user notification.

Conclusion

Through this research project we studied the importance of rainfall, its measurement techniques, and their shortcomings and devised a solution to overcome these by integrating predictive analytics techniques from the new and upcoming field of big data analytics.

Acknowledgment

We would like to thank Prof. Vijaykumar Bhanuse for giving us his constant support and this opportunity to work on this project under him. We would also like to thank our honorable director Mr. Rajesh Jalnekar and head of department prof. Dr. Shilpa Sondkar for their inspiration.

Bibliography

Dile, Y. T., R. Srinivasan, 2014. Evaluation of CFSR climate data for hydrologic prediction in data-scarce watersheds: an application in the Blue Nile River Basin. *Journal of the American Water Resources Association (JAWRA)* 1-16. DOI: 10.1111/jawr.12182

Fuka, D.R., C.A. MacAllister, A.T. Degaetano, and Z.M. Easton. 2013. Using the Climate Forecast System Reanalysis dataset to improve weather input data for watershed models. *Hydrol. Proc.* DOI: 10.1002/hyp.10073.

Lily Ingsrisawang ET. Al. Machine Learning Techniques for Short-Term Rain Forecasting System in the Northeastern Part of Thailand

National Centers for Environmental Prediction (NCEP) Climate Forecast System Reanalysis (CFSR) globalweather.tamu.edu