[Manipal University Jaipur]

Rainfall Prediction model using Random Forest

NICOLAS PONT

REG NO. 229310179

Index

1. Abstract---------- 1

2. Timeline---------- 2

3. Actual Model---------- 3

4. Applications---------- 8

5. Conclusion---------- 10

Abstract

This study presents a model that processes rainfall data across various subdivisions within the Indian peninsula using Random Forest Regression, which is an ensemble learning technique that leverages multiple decision trees. The model was developed as part of this Datathon project. This model makes forecasts regarding the amount of precipitation that will fall on a particular subdivision during a specific year or month. In addition, this project used a variety of data analysis tools, such as scikit-learn, Pandas, Matplotlib, and seaborn, in order to evaluate the effectiveness of the dataset. The findings were displayed graphically for easy comprehension using various representations. The website retrieves data from our model and displays it to users.

Timeline

The Actual Model

Data Collection and Cleaning

We collected the rainfall data for various subdivisions in India from kaggle which was sourced by the India Meteorological Department's website. The data was in the form of CSV files, which we loaded into a Pandas dataframe for cleaning and analysis.

The data contained missing values, which we replaced with the mean of the column. We also converted the categorical variables, such as month and subdivision, into numerical variables for machine learning model training.

Data Exploration

We explored the data using various data visualization libraries such as Matplotlib and Altair. We created interactive visualizations such as line charts and bar graphs to understand the trends and patterns in the data.

We found that the rainfall patterns were different across different subdivisions and years, with some subdivisions experiencing more rainfall than others. We also found that the rainfall was highest during the monsoon season, which occurs from June to September.

Machine Learning Models

1. Linear Regression
2. Random Forest

Linear Regression is a supervised learning algorithm used to predict a continuous target variable based on one or more predictor variables. The goal is to fit a linear model that best describes the relationship between the predictor variables and the target variable.
In a simple linear regression model, there is only one predictor variable, and the relationship between the predictor and target variable is a straight line. The equation of the line is given by:

$y = mx + b$

where y is the target variable, x is the predictor variable, m is the slope of the

line, and b is the y-intercept.

In a multiple linear regression model, there are multiple predictor variables, and the relationship between the predictor variables and the target variable is a hyperplane in higher dimensions. The equation of the hyperplane is given by:

y = b0 + b1x1 + b2x2 + ... + bn*xn

where y is the target variable, x1, x2, ..., xn are the predictor variables, b0 is the intercept, and b1, b2, ..., bn are the coefficients of the predictor variables.

The goal of linear regression is to estimate the values of the coefficients b0, b1, b2, ..., bn that best fit the data. This is done by minimizing the sum of the squared errors between the predicted values and the actual values. The most common method used for this is Ordinary Least Squares (OLS).

Once the coefficients are estimated, the model can be used to make predictions on new data.

Random Forest is a supervised machine learning algorithm used for classification and regression tasks. It is an ensemble method that combines multiple decision trees to create a more accurate and robust model.

Random Forest is a supervised machine learning algorithm used for classification, regression, and other tasks that involve predicting an output variable from a set of input variables. Random Forest is an ensemble method that combines the results of multiple decision trees to produce a more accurate and stable prediction.

The basic idea behind Random Forest is to create multiple decision trees using random subsets of the data and random subsets of the features. This introduces randomness into the model, which helps to reduce overfitting and improve the accuracy of the predictions. The output of each decision tree is then combined to produce a final prediction.

Here are the key steps involved in building a Random Forest model:

Data preparation: Random Forest can handle both categorical and continuous variables, but the data needs to be prepared in a way that is suitable for the model. This may involve cleaning the data, handling missing values, and transforming the data into a suitable format.

Feature selection: Random Forest can handle a large number of features, but it's important to select the most important ones to reduce the complexity of the model and improve its accuracy. This can be done using techniques such as correlation analysis, feature importance analysis, and dimensionality reduction.

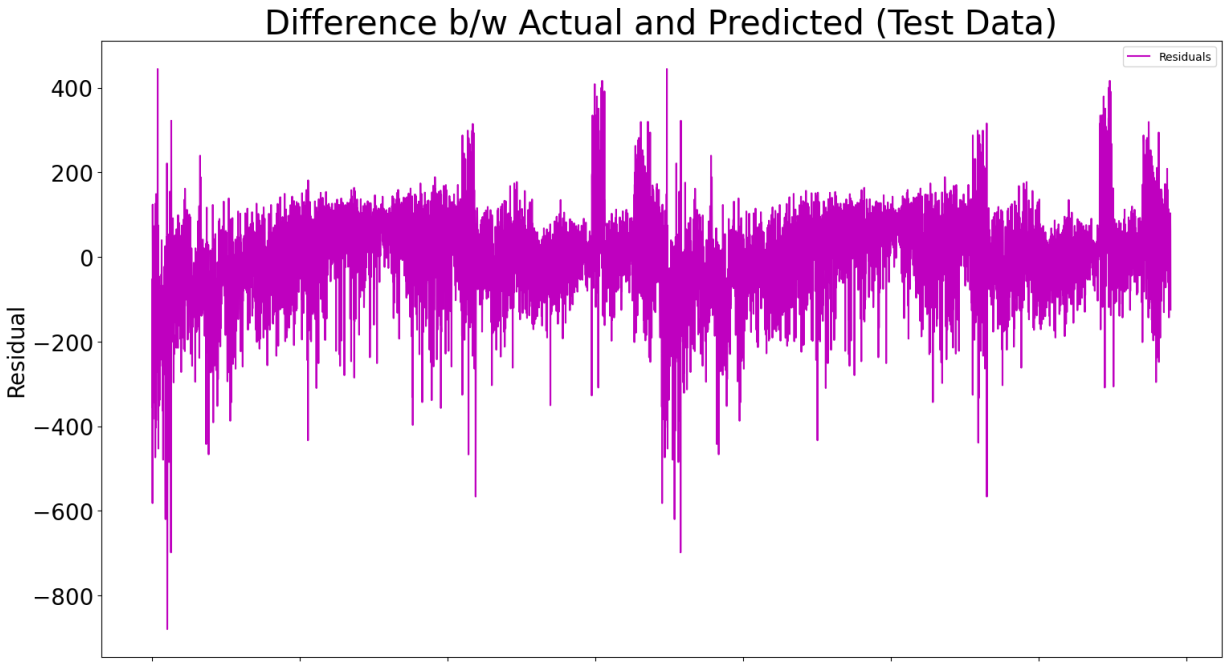
Building the decision trees: Random Forest builds multiple decision trees using different subsets of the data and different subsets of the features. Each decision tree is trained using a subset of the data and a subset of the features, and the output of each tree is combined to produce a final prediction.

Combining the decision trees: The output of each decision tree is combined to produce a final prediction. This can be done using techniques such as averaging, weighted averaging, or voting.

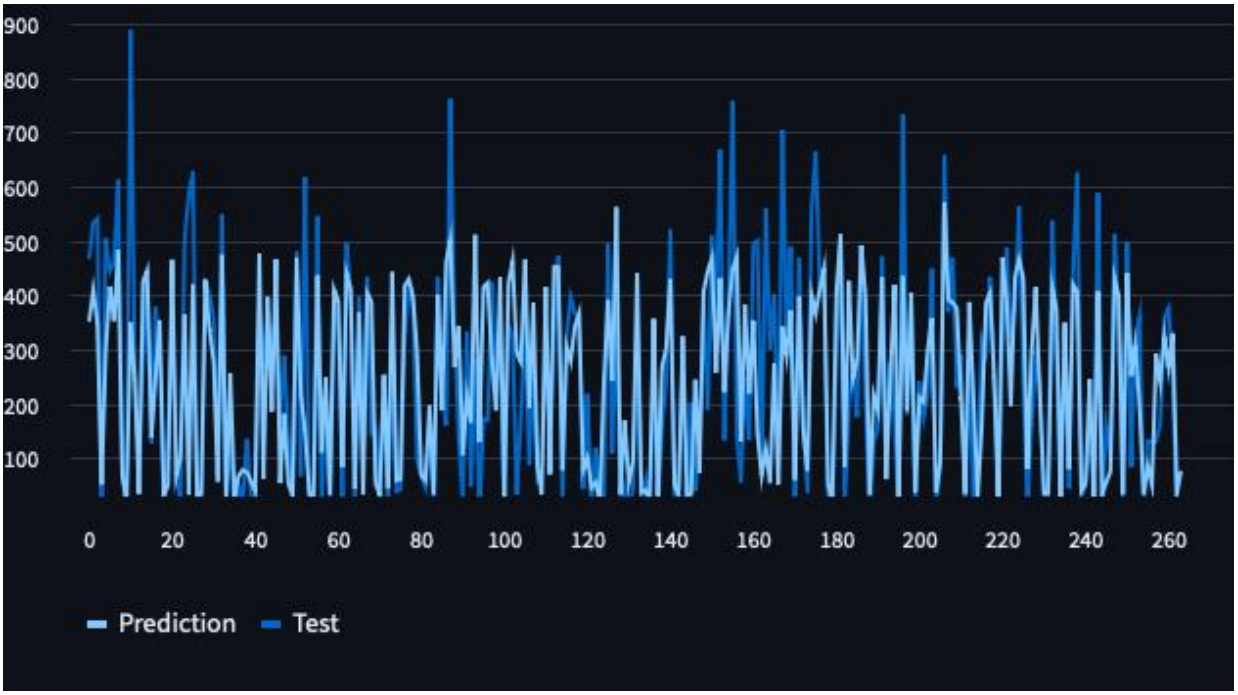
Tuning the parameters: Random Forest has several parameters that can be tuned to improve its performance, such as the number of trees, the depth of the trees, and the size of the subsets used for training. It's important to tune these parameters carefully to achieve the best possible performance.

Some advantages of using Random Forest include:
-It can handle a large number of features and is generally resistant to overfitting.
-It can handle both categorical and continuous variables.
-It can be used for both classification and regression tasks.
-It provides a measure of feature importance, which can be useful for feature selection.

However, there are also some limitations to Random Forest, such as:
-It can be slow to train on very large datasets.
-It may not perform as well as other algorithms on some types of data.
-It can be difficult to interpret the results of a Random Forest model.



Linear Regression



Model Fit for Random Forest Model

Clearly, the Random Forest Model is a better fit as the errors in Linear Regression are quite high. That is why we chose to implement RandomForest.

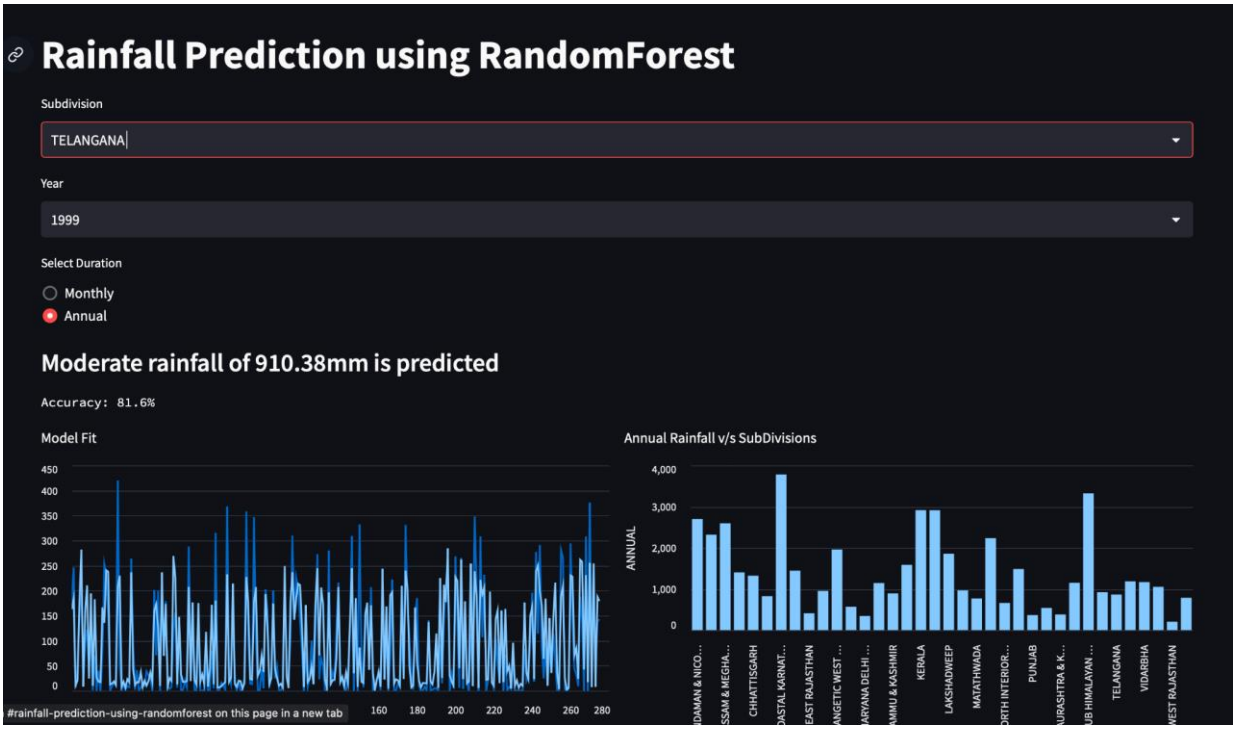
Feature importance

In addition to evaluating model performance, it's also important to understand which features are driving the predictions.
In this particular model, The Parameters used for prediction are -

1. Subdivision
2. Year
3. Month

User Interface

We created a user interface for the Rainfall Prediction using Streamlit, a Python library for building web applications. The user interface allows the user to select a subdivision, year, and duration (monthly or annual) and displays the predicted rainfall using the RandomForest model.



Accuracy Score: 75 - 85%

We also implemented other models such as Lasso and SVC in pursuit of higher accuracy scores but Random Forest proved to be more promising in each case.

Lasso - Training Accuracy: 28.2 , Testing Accuracy: 27.6
SVM - Training Accuracy: 10 , Testing Accuracy: 11.6

Applications

Agriculture

The ability to accurately forecast rainfall is extremely helpful to farmers. The farmers are able to plan their planting and harvesting schedules with the help of rainfall predictions. For instance, if it is forecast that there will be a dry spell, farmers may decide to plant crops that have a lower demand for water or they may choose to delay planting until the rains return.

The timing of when farmers irrigate their crops can also be determined with the help of rainfall forecasts. In order to save water and avoid overwatering the crops, they might decide to cut back on the amount of irrigation they do or even forego it entirely if heavy rains are forecast.

In addition, precise rainfall forecasts can assist farmers in making educated choices regarding crop insurance and risk management. For instance, if it is anticipated that there will be a drought, farmers may decide to purchase crop insurance in order to safeguard themselves against the possibility of financial loss.

Flood Control

The forecasting of rainfall can be of assistance to the authorities in a number of different ways when it comes to flood control. This would allow them to reduce the risk of flooding while still ensuring that households had access to water.

The authorities can use rainfall predictions to control flooding by storing excess rainwater during periods of heavy rainfall at suitable locations along the course of rivers in order to reduce the flow of water further downstream. A flood control measure of this kind can be carried out by building an obstruction similar to a dam across the river and collecting water in the portion of the river that is upstream in order to create a reservoir.

In addition, flood forecasting systems involve the predetermination of flood events, which is necessary to assist in the implementation of structural measures for mitigating flood damages, regulating and operating multipurpose reservoirs with the focus on controlling incoming floods, and evacuating affected people to safer places. The application of hydrological models in addition to the tools provided by GIS is of utmost significance in this regard.

Water Management

For efficient management of water resources, it is crucial to have a reliable method of predicting future rainfall patterns in regions with scarce water resources. When water managers have access to reliable precipitation projections, they can make more calculated choices about where and when to store and release water.

Conclusion

This machine learning model uses the RandomForest regression technique to predict the rainfall of any given year or month with an accuracy of about 75-90%(depending on the dataset). This technique overwhelms all the other machine learning algorithms we have tested so far such as LASO, SVM, LREG, RIDGE REG on the basis of accuracy. Integrating this model into a Streamlit based website also enabled us to make our model user-friendly and accessible for everyone.