IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

RAIN PREDICTION USING MACHINE LEARNING

[1]VIJITHRA NAIR, [2]MEGHA MATHEW, [3]SWETA BHATTACHARJEE, [4]ARASHDIP SINGH, [5]PROF.PAYEL THAKUR

[1]UG Student, [2]UG Student, [3]UG Student, [4]UG Student, [5]Asst. Professor,
Computer Department,
Pillai College of Engineering, New Panvel, India

Abstract : As agriculture being the key point of survival, Rainfall is the important source for its cultivation. Rainfall prediction has always been a major problem as prediction of rainfall gives awareness to people and to know in advance about rain so as to take necessary precautions to protect their crops from rain. A particular dataset is taken from Kaggle community and this project predicts whether it will rain tomorrow or not by using the rainfall in dataset. CatBoost model is implemented in this project as it is an open sourced machine learning algorithm, and features great quality without the parameter tuning, categorical feature support, improved accuracy and fast prediction. CatBoost model is a gradient boosting toolkit and two critical algorithms classical and innovative are introduced to create a fight in prediction shift present in currently existing implementations of gradient boosting algorithms. CatBoost performed very well giving an AUC (Area under curve) score 0.8 and ROC (Receiver operating characteristic curve) score as 89. ROC is called as an evaluating curve whereas AUC presents a degree or measure of separability as the model is skilled enough to distinguish between classes. An Exploratory data analysis is done to examine data distribution, outliers and provides tools for visualizing and understanding the data through graphical representation. A dashboard is implemented to showcase the information that is represented in datasets i.e. any changes in the data will result in different types of graphs. A linear SVC (Support vector classifier) provides a best fit hyperplane that divides the data and feeds some features to the classifier to detect what the predicted class is and results in desired output.

IndexTerms **- ARIMA, CatBoost, Random Forest, Rainfall prediction, XgBoost**

1.INTRODUCTION

The world's welfare is agriculture. The achievement of agriculture is dependent on rainfall. It also helps with water resources. Rainfall information in the past helps farmers better manage their crops, leading to economic growth in the country. Prediction of precipitation is beneficial to prevent flooding that saves people's lives and property. Fluctuation in the timing of precipitation and its amount makes forecasting of rainfall a problem for meteorological scientists. To overcome these problems, a machine learning technology is used which is predictive analysis that is a branch of data mining which predicts the future probabilities and trends.

Prediction is the phenomenon of knowing what may happen to a system in the near future. Since rainfall is the major causes of calamities like floods and typhoons, predicting the occurrence of rainfall will help us to be prepared for these calamities. The basic procedures involved are first identifying an initial model, second repeatedly changing the model by removing a predictor variable based on a criteria and then terminating the process when a model which fits the data well. Here, various rainfall prediction projects were developed using multiple linear regression and other models.

This proposed method uses Australian meteorological dataset to predict the rain fall. Usually machine learning algorithms are classified into two major categories i.e. unsupervised learning and supervised learning. All of the clustering algorithms come under the supervised machine learning. Even though many models have developed, it is necessary for doing research using machine learning algorithms to get accurate predictions. The error free prediction provides better planning in the agriculture and other industries, Henceforth we have used the CatBoost model for faster and accurate prediction.

2. LITERATURE SURVEY

2.1 CatBoost model*:*

According to Shihab Ahmad Shahria this model follows a technique that is additionally divided into four parts: (a) data pre-processing that involves the gathering of pollutants and other alternative meteorological data alongside with correction of missing values; (b) analysis associated with relationship between meteorological parameters or variables which of the pollutants; (c) feature importance involves the screening of meteorological parameters and conjoints the pollutants in air right before the

operation on the models; (d) the application of models like ARIMA-ANN, ARIMA-SVM, PCR, DT and CatBoost.[1]

2.2 XgBoost model:
According to Nikhil Tiwari, Generally, XGBoost is fast when compared to other implementations of gradient boosting. Szilard Pafka performed few benchmarks comparing the performance of XGBoost to different gradient boosting techniques and bagged decision trees. XGBoost focuses on structured and tabular datasets on classification and regression based on its predictive modelling problems. The result is the algorithm for competition winners on the Kaggle competitive data science platform. Gradient boosting is an approach where new models are created that predict the residuals or errors of previous models and then summed up together to make the final prediction. It is called so because it uses a gradient descent algorithm to minimize the loss when new models are added. This approach supports both regression as well as classification predictive modelling problems.[2]

2.3 Random Forest:
Urmay shah's Random forest is a tree based model, it is a collection of many tree models. Different tuning parameters are used for tuning the model. In random forest one of the parameters shows exactly how many trees should be more used to get the accurate results. Random forest works really well with high variant and low biasing models. It is observed that after 250 number trees error rate doesn't change. Hence 250 number of trees are restricted to in the forest. Random forest method is excellent for light rain predictions as it gives the best accuracy. It also performs well for the no rain, moderate rain, and for light rain.[3]

2.4 ARIMA Model:
CMAK Zeelan Basha states ARIMA MODEL(AutoRegressive Integrated Moving Average) is used for time series predictionand analysis and forecasting. It contains four methods and is proposed by Box and Jenkins. The following are the four steps used in the ARIMA model. Stage-1: Identification of a series of responses is done in the first stage which is used in calculating the time series and autocorrelations using statement IDENTIFY Stage-2: In this stage Estimation of the previously identified variables is done and also the parameters are estimated using the statement ESTIMATE. Stage-3: Diagnostics correction of the above gathered variables and parameters are all implemented in this stage. Stage4: In this stage the predicting values of time series are forecasted which are future values, using the ARIMA model using the statement FORECAST. The parameters used in this model are p,d,q which describes 'p' as the number of lag observations, 'q' as the degree of differencing and as the moving average order.[4]

2.5 Summary of Related Work

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

Literature	ARIMA	CatBoost	Hybrid
Shihab Ahmad Shahria et al. 2021[1]	Yes	Yes	Yes
Nikhil Tiwari et al. 2020 [2]	No	Yes	No
Urmay Shah et al. 2018 [3]	Yes	No	No
CMAK Zeelan Basha et al. 2020 [4]	Yes	No	Yes

The overview of comparison of different parameters are given in Table 2.

Table 2 Summary of literature survey

Literature	**Performance Parameters**
Shihab Ahmad Shahria et al. 2021[1]	CatBoost,ARIMA-ANN, ARIMA-SVM, DT
Nikhil Tiwari et al. 2020 [2]	Neural Networks, Support Vector Regressor (SVR), Elastic Net , Ridge Regression , Lasso Regression , Linear Regression , XGBoost , Random Forest , Bagging Regressor, Gradient Boosting Regressor
Urmay Shah et al. 2018 [3]	ARIMA(Auto-Regressive Integrate d Moving Average), SVM(Simple Moving Average Model), Decision Tree, Holt Winter, , Random Forest, Neural Network method ,Seasonal Naive method
CMAK Zeelan Basha et al. 2020 [4]	ARIMA Model(Auto-Regressive Integrate d Moving Average), Artificial Neural Network, Logistic Regression, Support Vector Machine and Self Organizing Map

3. PROPOSED WORK

The Proposed system consists of using the CatBoost Algorithm. It performs very well giving an AUC (Area under curve) score 0.8 and ROC(Receiver operating characteristic curve) score as 89.ROC is evaluating curve and AUC presents degree or measure of separability as this model is capable of distinguishing between classes. An Exploratory data analysis is done to examine data distribution, outliers data through graphical representation.

3.1 SYSTEM ARCHITECTURE

The system architecture is given in Figure 1. Each block is described in this Section.

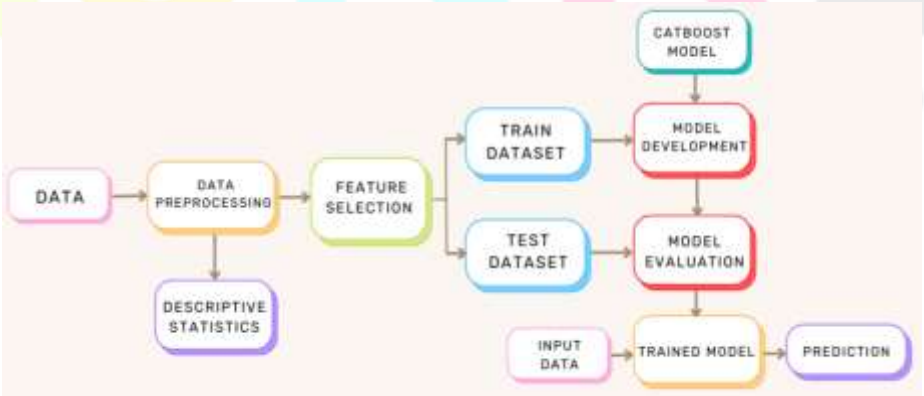

Fig. 1 Proposed system architecture

3.1.1 Cardinality check for Categorical features:

The accuracy and the performance of a classifier depends not only on the model that we use, but also on how we preprocess data, and what kind of data we are feeding the classifier to learn and adapt

Many Machine learning algorithms like Linear Regression, Logistic Regression, k-nearest neighbors, etc. can handle only numerically allowed data, so encoding categorical data to numeric becomes a mandatory step. But before jumping into encoding part, check the cardinality of each categorical feature.

Cardinality: The number of unique values in every categorical feature is known as cardinality.

A feature with a enormous number of distinct/ unique values is a highest cardinality feature. A categorical feature with hundreds of zip codes is the best example for a high cardinality feature.

This highest cardinality feature poses many threats like it will increase the number of dimensions of data when the feature is encoded. This is bad for the model.

There are many ways to handle the high cardinality, one can also feature engineering and the other is simply dropping that feature if it doesn't add any value or meaning to the model.

3.1.2Handling Missing Values:

Machine learning algorithms can't handle missing values and cause problems. So they need to be addressed in the first place. There are many techniques to identify and impute missing values.

If a dataset contains missing values and loaded using pandas, then missing values get replaced with NaN(Not a Number) values. These NaN values can be identified using methods like isna() or isnull() and they can be imputed using fillna(). This process is known as Missing Data Imputation.

Missing values in Numerical Features can be imputed using Mean and Median. Mean is sensitive to outliers and median is immune to outliers. If you want to impute the missing values with mean values, then outliers in numerical features need to be addressed properly.

3.1.3 Outliers detection and treatment:

An Outlier is an elaborated observation that lies associate abnormal distance from the opposite values in a given sample. They'll be additionally detected using visualisation (like boxplots, scatter plots), Z-score, applied math and probabilistic algorithms, etc.

3.1.4Feature Importance:

Machine Learning Model performance relies on features that are used to train the model. Feature importance mentions which features are way relevant to build a model.

Feature Importance also refers to the techniques that assign a score to input/label features based on how useful and good they are at predicting a target variable. Feature importance helps in Feature Selection.

3.1.5 Splitting Data into training and testing set:

train_test_split() is a method of model_selection class that is used to split data into training and testing sets.

Feature Scaling: Feature Scaling is a technique used to scale, normalize and standardize data in range(0,1). Each column of a dataset has different values, that's how it helps to scale data of each column to a common level. StandardScaler is a class used to implement the feature scaling.

4. REQUIREMENT ANALYSIS

The implementation detail is given in this section.

4.1 Software

Operating System	Windows 10
Programming Language	Python

4.2 Hardware

Processor	2 GHz Intel
HDD	180 GB
RAM	2 GB

4.3Dataset and Parameters

Dataset from Kaggle contains of about 10 years of daily weather observations from various different locations across Australia. Prediction of next-day rain by training classification models on the target variable Rain Tomorrow and various parameters including Date, Location, Mintemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, etc. The missing

values are usually handled by random sample imputation to observe the variance categorical Vvlues like location, wind direction are handled by using Target guided Encoding Outliers are handled using IQR and boxplot Imbalanced Dataset was handled using SMOTE (Synthetic Minority Oversampling Technique).

5. IMPLEMENTATION

Catboost model is implemented in this project as it is an open source machine learning algorithm, and features great quality with parameter tuning, categorical feature support, improved accuracy and fast prediction. Cat boost model is a gradient boosting toolkit and two critical algorithms classical and innovative are introduced to create a fight in prediction shift present in currently existing implementations of gradient boosting algorithms.

Catboost performed very well giving an AUC (Area under curve) score 0.8 and ROC (Reciever operating characteristic curve) score as 89. ROC is an evaluating curve and AUC represents degree or measure of separability as the model is capable of distinguishing between classes. An Exploratory data analysis is examined to data distribution, outliers and also provides tools for visualizing and understanding the data through the graphical representation.

5.1 Encoding of Categorical Features:

Most Machine Learning Algorithms like the Logistic Regression, Support Vector Machines, K Nearest Neighbours, etc. cannot handle categorical data. Hence, these categorical data need to converted to numerical data for modeling, which is called Feature Encoding.
There are many feature encoding techniques like One code encoding, label encoding.

6. OUTPUT

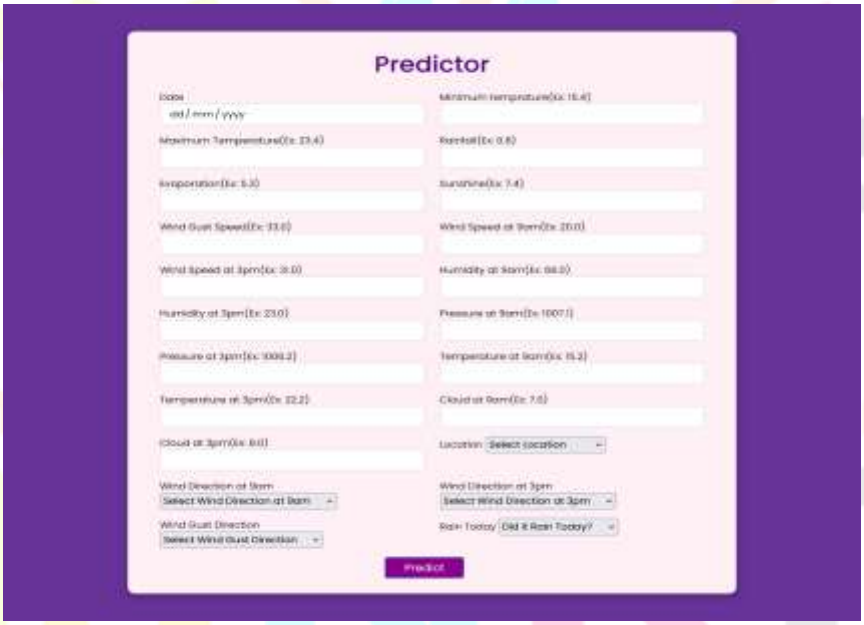

Fig. 2 Predictor Page

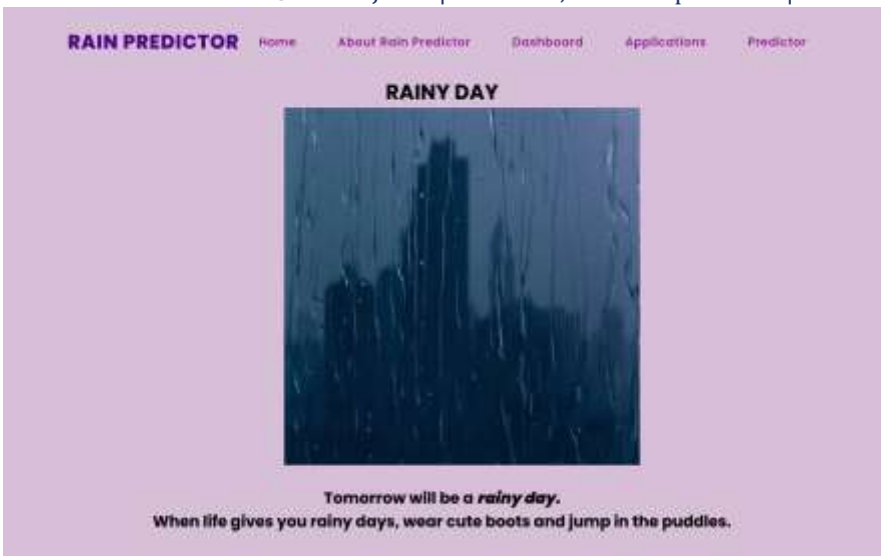

Fig. 3 Output(rainy day)

Fig. 4 Output(sunny day)

7. CONCLUSION

Rain fall prediction plays a big role in agriculture production. The growth of this agricultural products is based on the rainfall amount. So it is necessary to predict the rainfall of the season to assist the farmers in agriculture. In this paper, a gradient boost approach has also been proposed to forecast rainfall for a selected location in Australia. The proposed method predicts the rainfall for the Australian dataset for using decision tree algorithm (Catboost model) and provides the improved results in terms of accuracy and prediction. . The study also showed that the Catboost model was performing so much better than in months with the high annual averages compared to alternative/existing approaches. In future research, ensemble techniques will be used to combine the varieties of the models. Also, additional locations and datasets needed to be incorporated.

8.FUTURE SCOPE

In this project, as rainfall is dependent on the various parameters it is also required to study how other meteorological parameters affect the Rainfall prediction. We can also perform the same exercise on hourly data using various parameters to forecast next hour rainfall. This study can also be done using more observations for particular region or area, and design this kind of model on big data framework so that computation can be faster with higher accuracy.

9.ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof Payel Thakur for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

10. REFERENCES

[1] Yang Liu, Qingzhi Zhao, Wanqiang Yao, Xiongwei Ma, Yibin Yao and Lilong Liu, "Short-term rainfall forecast model based on the improved BP–NN algorithm", 2019. Available: https://www.nature.com/articles/s41598-019-56452-5[Submitted on 24 December 2019]

[2] Vishal Morde. "XGBoost Algorithm: Long May SheReign!",2019.Available:https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-longshe-may-rein-edd9f99be63d [Submitted on 8 April 2019]

[3] Ramya Bhaskar Sundaram, "An End-to-End Guide to Understand the Math behind XGBoost", 2018.Available:https://www.analyticsvidhya.com/blog/2018/09/an-endto-end-guide-to-understand-the-math-behind-xgboost/[Submitted on 6 September 2018]

[4] Available:https://www.kaggle.com/prashant111/ca tboost-classifier-in-python

[5] Aman Kharwal, "Rainfall Prediction with Machine Learning",2020.Available:https://thecleverprogrammer.com/2020/09/11/rainfallprediction-with-machine-learning/ [Submitted on 11 September 2020]

[6] Anamika Jha "CatBoost – A new game of Machine Learning:",2020. Available:https://affine.ai/catboost-a-new-game-of-m Ac hinelearning/ [Submitted on 21 October 2020