Weather Prediction for Ahmedabad using Machine Learning

Umang Kamdar
AU1841069
Ahmedabad University
umang.k@ahduni.edu.in

Jainesh Patel
AU1841101
Ahmedabad University
jainesh.p@ahduni.edu.in

Vatsal Patel
AU1841103
Ahmedabad University
vatsal.p1@ahduni.edu.in

Shubh Shah
AU1841122
Ahmedabad University
shubh.s1@ahduni.edu.in

I. Abstract

Weather prediction and forecasting are becoming important day by day, as it's applications makes human life more comfortable in dealing with uncertain events like: Heatwaves, Droughts, Blizzards, Hurricanes, etc. In this paper, we have predicted temperature by implementing different varieties of Machine Learning algorithms like: Multiple Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression. Furthermore, the data-set features involved for predicting temperature tended to change as per the location, consequently we had to restrict our model to data-set of a particular region i.e, Ahmedabad City. We also tried to predict humidity along with temperature as multivariate regression. We also tried for rainfall prediction. We further restricted the use of certain features for forecasting, since some of them were immaterial, according to our brainstorming sessions and some of them were found unassociated, according to mathematical analysis. After a series of actions, we were able to successfully execute various Machine Learning algorithms on the data-set and generate a good accuracy.

II. Key Words

Machine Learning, Correlation, Regression, Weather Prediction.

III. Introduction

In real life we need to deal with many uncertain weather problems. We use the weather prediction in Agriculture sector, Airport environments and some other important events. Thus, in order to forecast the chaotic nature of atmosphere we need to use some statistical approach. Only mathematical approach will not give us the insights of the weather prediction. We need past data of the all the features like temperature, humidity, wind index and others. We need to use algorithms which will include the past data and statistical calculations for predicting the weather.

Our project is to predict the weather specifically for Ahmedabad city. We collected the historical weather data of each day from 2015 to 2020 specifically of Ahmedabad City. The data-set included various features related to weather.

For weather prediction we chose Regression as the machine learning algorithm, as our weather data has real values, so the best algorithm for predicting temperature was Regression. There are many regression techniques and we tried some of them for testing and compare the accuracy of the prediction. First of all we tried Multiple Linear regression and in which we got good accuracy and after that we tried Polynomial, Lasso and Ridge regression. We got very interesting results in all of them.

First of all, we started with predicting the temperature and we looked for the model which give us best accuracy. After doing that, we move on to rainfall prediction. And at last, for weather prediction, many features are dependent on each other so we can take one or more features as dependent feature and predict the output. We thought of predicting humidity along with temperature, so we used Multivariate Regression to predict the same.

IV. Literature Survey

A simple weather prediction was performed using Multiple Linear Regression. In that model, a time-series data of weather was first collected. Temperature and relative humidity were predicted after doing feature selection based on correlation results [2]. Pearson correlation is used to find linear relationship between continuous variables [3]. Most of the time the data is noisy and it has null values and missing values so it needs to clean the data first [4]. To generalize the model well, Regularization is used to reduce the over fitting problem. [5].R square is statistical measure which gives information about how much amount of the variation for target variable determined by the regression. p value is a statistical measure which is used in regression analysis. When there are more than one dependant variables in the dataset then to predict those variables, multivariate regression is used. Three Machine Learning models - Multiple Linear Regression (MLR), Support Vector Machine (SVM) and Artificial Neural Network (ANN) were compared for finding better Weather Prediction model. It was concluded that MLR was better for temperature prediction than others [6]. Extra Tree classifier and other classifiers like decision trees and random forest is studied [7]. A two-level feature selection technique is proposed to get the significant features. The first level targets ranking the subset of features dependent on high information gain entropy in decreasing order. The second level broadens additional features with a better discriminative ability over the initially ranked features [8]. Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. To determine set of features which gives best prediction results an algorithm is designed based on generalized simulated annealing method (GSA) of optimization [9]. An R simulation was carried out whose results showed that the Wrapper Methods, Sequential Forward Selection and Sequential Backward Elimination were better than the Filter Methods, Correlation based Feature Selection and Information Gain in selecting the correct features [10]. Wrapper method is more accurate but it is computationally expensive so it is prefer to use when features are less than 20.[11]

V. Implementation

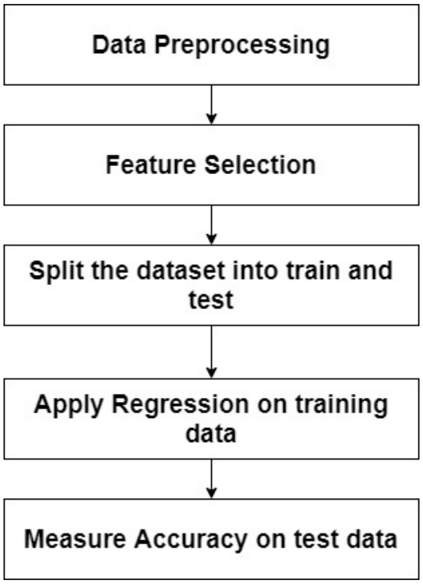
Data Preprocessing
Feature Selection
Split the dataset into train and test
Apply Regression on training data
Measure Accuracy on test data

Fig. 1. Algorithm

- Data Gathering : We took a data-set of Ahmedabad city from [1] using API key of worldweatheronline.com. We collected the historical weather data of each day from 2015 to 2020. The data-set included various features related to weather which were: humidity, totalSnow, sunHour, precipMM, uvIndex, moon illumination, moonrise, moonset, sunrise, sunset, DewPointC, visibility, HeatIndexC, WindChillC, WindGustKmph, cloudcover, pressure, tempC, winddirDegree, windspeedKmph.
- Data Pre-processing : We removed few of the columns which were irrelevant or redundant or the columns which had all null values. So, we removed columns like maxtempC, mintempC, moonrise, moonset, sunrise, sunset, location, totalSnow, FeelsLikeC. Further, we used feature selection process to remove less important columns.
- Feature Selection
  1) Correlation: It is used to see that how features are correlated with each other. If correlation value is positive than it is positively correlated and if correlation value is negative than it is negatively correlated.
     To make the feature independent of each other we performed correlation between the features and observe the correlation values. Based on the values, we see that some features are highly correlated. We set threshold value to obtain the features which are dependant on some other features of the dataset. We found that the some of the columns have correlation value greater then 0.9. We removed those features as feature selection process and reduce the features from our dataset.
  2) Extra Tree Classification: It is a type of group learning technique which combines the results of multiple uncorrelated decision trees collected in a ”forest” to output it’s classification result. Each Decision Tree is constructed from the original training sample. Then, at each test node, each tree is provided with a random k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria. This random sample of features leads to the creation of multiple uncorrelated decision trees.
     We calculate the entropy of the data using below equation.
     $$\text{Entropy }(S)=\sum_{i=1}^{c}-p_i\log_2\left(p_i\right)$$
     After that we have to create the decision trees to be constructed such that Gain will be:
     $$\operatorname{Gain}(S,A)=\text{ Entropy }(S)-\sum_{v\in\operatorname{Values}(A)}\frac{\left|S_v\right|}{|S|}\text{ Entropy }\left(S_v\right)$$
     At the end we computed total information gain for the each features. And we removed the features whose values were the lowest ones.
  3) Information Gain: It is a type of Filter method in which Reduction in Entropy is calculated after transformation of a dataset, which is further used to evaluate information gain of each feature against the target which is used for selecting feature.
  4) Backward Feature Elimination: It is a type of Wrapper method in which we feed all the features to the model. We check the performance of the model and then iteratively remove the worst performing feature one by one until we reach a stage where removing a feature results in no improvement in performance of the model.
     The performance metric used here to evaluate feature performance is pvalue. If the pvalue is above some threshold value, then we remove that feature, else we keep it.

  Next we performed different regression techniques which are described below.
- Multiple linear Regression: Initially, we used inbuilt function of sklearn library to predict the temperature. For that, we import all required libraries and split the data set into training and testing data set. Then by using

inbuilt function of linear regression, we trained the model. Based on predicted parameters, we test the model on testing dataset and find the accuracy. After understanding of mathematics behind the multiple linear regression, we build custom function for regression. We used cost function under gradient descent method to find the good parameters.

$$cost = \frac{1}{2n} * \sum_{i=1}^{n} (f(x(i)) - y(i))^2$$

Here $f(x)$ is hypothesis function which is $\theta^T X$. Here actual value is subtracted from the hypothesis function value and its value is squared. n is the total number of training data.

To get good parameters we should minimise the cost function.For that gradient descent method is used.

$$\theta_{new}(i) = \theta(i) - \alpha * \frac{d}{d\theta}(cost)$$

$$\theta_{new}(i) = \theta(i) - \alpha * \frac{1}{n}(f(x) - y)X^T$$

Here α is learning rate and we iterate over multiple times such that we get at the situation where $\theta_{new} = \theta$ for gradient descent algorithm. By using this method we can find best parameters to predict the temperature accurately.

- We also tried polynomial regression to predict the temperature. We used inbuilt functions and the libraries. We choose the degree 2 and degree 3 for this model.
- We also tried to introduce regularization by performing regression using Ridge regression and the Lasso regression.The cost function for the Ridge regression is described as

$$cost = \frac{1}{2n} * \sum_{i=1}^{n} (f(x(i)) - y(i))^2 + \sum_{j=1}^{m} \frac{\lambda}{2n} * (\theta(j)^2)$$

here λ is regularisation parameter. By minimising the cost function using gradient descent algorithm, we find the parameters which fit the model well and get good accuracy.

Cost function for Lasso regression is described as

$$cost = \frac{1}{2n} * \sum_{i=1}^{n} (f(x(i)) - y(i))^2 + \sum_{j=1}^{m} \frac{\lambda}{2n} * (\theta(j))$$

We chose among many various learning rate and regularization parameters to find out which value gave us the best accuracy.

- We tried the same regression techniques for rainfall prediction too. We get negative values from regression output which is not significant, so we converted the negative values to zero and then find the accuracy.
- When there are more than one dependant features in the dataset then we can use Multivariate regression. Here we consider that humidity is another feature that we can consider to predict along with temperature.To predict both the feature values, we consider temperature value and humidity as output values and use other features as input features.Then apply all the regression techniques described above and predict the output and measure the accuracy.

VI. Results

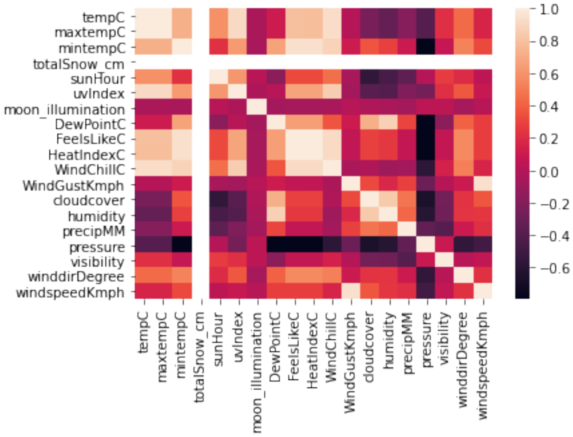

Fig. 2. Correlation of the features

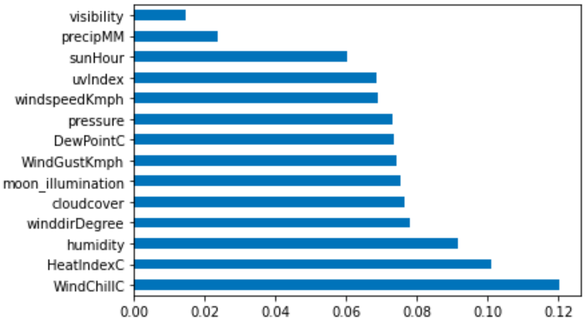

Fig. 3. Extra Trees Classifier

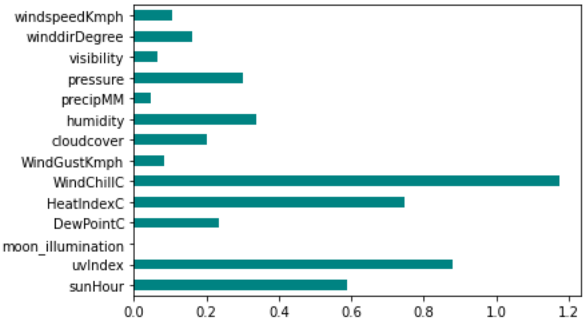

Fig. 4. Information Gain

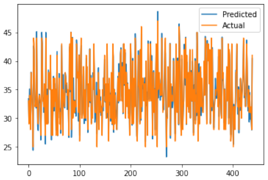

Fig. 5. Temperature prediction using polynomial regression

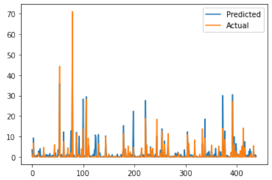

Fig. 6. Rainfall prediction using polynomial regression

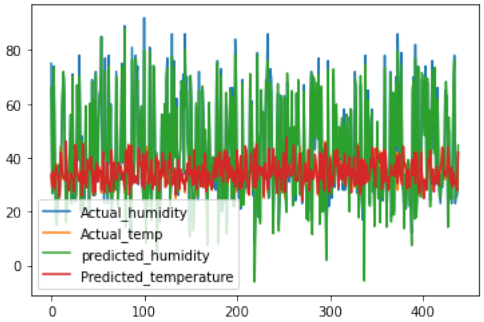

Fig. 7. Multivariate regression

The table shows accuracy of regression model to predict the temperature.

Machine Learning Algorithm	Accuracy
Multiple Linear Regression(Inbuilt)	0.9578
Multiple Linear Regression(Custom)	0.9535
Polynomial Regression(Degree=2)	0.9649
Polynomial Regression(Degree=3)	0.9596
Ridge Regression(Inbuilt)	0.9577
Ridge Regression(Custom)	0.9533
Lasso Regression	0.9577

In the Figure 2 - Correlation of the features, we set the threshold value equal to 0.9 and we removed those features which were greater than the threshold value. In the Figure 3 - Extra Tree Classifier, we selected the threshold value (0.06) for feature importance and remove the features which has less importance value than our threshold. Figure 4 shows the information gain of each feature to predict the temperature. We selected the threshold value (0.2) for feature importance and remove the features which has less importance value than our threshold. Next, we have used Backward Feature Elimination under wrapper methods. Here we have done in such a way that if p-value is greater than 0.05, then those features are removed. In Figure - 5, Temperature prediction using polynomial regression (deg=2) is shown. In Figure - 6, Rainfall prediction using polynomial regression (deg=2) is shown. We apply the same regression techniques for rainfall prediction too. But since the dataset contains more zero values for rainfall, our result is not much accurate. The best accuracy which we got for rainfall prediction is 0.7689. Figure - 7 represents the Multivariate Regression where 2 predicted features - temperature and humidity are compared against the actual features. For this, we got the accuracy around 0.9705.

VII. Conclusions

We found that among all the feature selection algorithms, the best accuracy was acquired by the Backward Feature Elimination. We performed four regression techniques for temperature prediction and rainfall prediction. We found that among all the regression algorithms, the best accuracy was acquired by the inbuilt polynomial regression with degree of 2. We found that degree 2 and degree 3 doesn't affect the accuracy much, but degree 2 is best. Next, we found that the accuracy of Multiple Linear Regression, Lasso Regression and Ridge Regression were almost similar. The accuracy result of custom made regression is also somewhat similar to inbuilt library functions.

Link to Github Repository:
https://github.com/Jainesh1009/CSE523-Machine-Learning-Amigos

References

[1] Ahmedabad Historical Weather. Worldweatheronline.Com, 2021, https://www.worldweatheronline.com/ahmedabad-weather-history/gujarat/in.aspx. https://dzone.com/articles/hashmap-

[2] Paras, Sanjay Mathur. ”A simple weather forecasting model using mathematical regression.” Indian Research Journal of Extension Education 12, no. 2 (2016): 161-168.

[3] ”What Is The Pearson Coefficient?”. Investopedia, 2021, https://www.investopedia.com/terms/p/pearsoncoefficient.asp.

[4] Folorunsho, Olaiya Adeyemo, Adesesan. (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies. International Journal of Information Engineering and Electronic Business. 4. 10.5815/ijieeb.2012.01.07.

[5] ”Mathematics for Machine Learning”. Copyright 2020 by Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. Published by Cambridge University Press.

[6] T. Anjali, K. Chandini, K. Anoop and V. L. Lajish, ”Temperature Prediction using Machine Learning Approaches,” 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 2019, pp. 1264-1268, doi: 10.1109/ICICICT46008.2019.8993316.

[7] Sharaff, Aakanksha Gupta, Harshil. (2019). Extra-Tree Classifier with Metaheuristics Approach for Email Classification. 10.1007/978-981-13-6861-5-17.

[8] Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F (2016) Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. PLoS ONE 11(11): e0166017. https://doi.org/10.1371/journal.pone.0166017.

[9] Sutter, J. and Kalivas, J., 1993. Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. Microchemical Journal, 47(1-2), pp.60-66.

[10] Ibrahim, Nuhu Hamid, H.A. Rahman, Shuzlina Fong, Simon. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. Pertanika Journal of Science and Technology. 26. 329-340.

[11] A. Shetye, Medium. [Online]. Available: https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b. [Accessed: 11-Apr-2021].