WEATHER FORECASTING USING LINEAR AND LOGISTIC REGRESSION MODELS

ABSTRACT

Weather forecasting has become a main field of interest during the recent years. Machine learning methods have shown good results in this area. Algorithms like support vector machine, decision tree, random forest and artificial neural networks have been used for this purpose in various studies. However, in this particular study, the aim is to create models out of algorithms like linear regression and logistic regression to make weather forecasts. Also R Studio and R language will be used to create the related models.

INTRODUCTION

Weather forecasting is basically predicting the climate of a region by using different methods and relevant data. Accurate prediction of weather is not completely possible due to the dynamic nature of the earth's atmosphere. Weather forecasting has its uses in agriculture, energy industry, aviation, communication etc. The weather is also one of the most important factors in deciding the outcome of a war.

Statistically, a seven-day forecast can be right only about 80% of the time. In olden times people depended on patterns to forecast climatic changes. There are 4 types of forecasting. They are

1. Analog method - A day in the past should be found which would be similar to the present conditions.
2. Climatology method – Averages are calculated for a specific day with respect to the data collected on that same specific day in the past. For example, August 1st of 2021 should be compared with August 1st of any year in the past.
3. Persistence and Trends methods – It relies on past trends for prediction.
4. Numerical Weather Prediction – Algorithms are used by computer softwares for prediction.

We can use a variety of machine learning algorithms to predict data such as linear regression, logistic regression, random forest, decision tree, nearest neighbours, naive bayesian and support vector machine.

LITERATURE REVIEW

SI No.	Author	Findings
1	Gowtham Sethupathi. M, Yenugudhati Sai Ganesh, Mohammad Mansoor Ali (2021)	Logistic regression is more efficient than random forest when there are massive datasets to predict rainfall prediction.
2	Mark Holmstrom, Dylan Liu & Christopher Vo (2016)	Forecasting was done for seven days based on the data of two days. Though functional regression was able to capture trends in the data, linear regression outperformed it.
3	G.Sujatha, Chinta Someswara Rao & T Srinivasa Rao	Humidity was predicted using logistic regression, decision tree, k nearest neighbours, naïve Bayesian, support vector machine and random forest methods. Random forest had the highest accuracy.
4	Fahad Sheikh, S. Karthick, D. Malathi, J. S. Sudarsan & C. Arun	Decision tree algorithm was found to handle the weather dataset better.
5	Tanvi Patil & Kamal Shah	Linear regression was used to predict minimum and maximum temperatures as well as wind speed. Logistic regression is used in rainfall prediction.
6	Suvendra Kumar Jayasingh, Jibendu Kumar Mantri & P. Gahan	Decision tree, support vector machine and multilayer perceptrons was used and it was found that decision tree performed better.
7	Y.Radhika and M. Shashi	Support Vector Machine performs better than Multi Layer Perceptron trained with back propagation. When parameters are selected suitably, support vector machines can replace neural network based models.
8	Neeraj Kumar & Govind Kumar Jha	Artificial Neural Network and MATLAB was used and the model was trained on 60 years of past data and showed good results.

9	Nazim Osman Bushara & Ajith Abraham	Fourteen base algorithms were used and correlation coefficient of all base classifiers was found to be more than 0.8.
10	S. Santhosh Baboo & I.Kadar Shereef	Back propagation neural network was used and was successful in improving convergence.

OBJECTIVES

To forecast the temperature of a given day using linear regression algorithm.

METHODOLOGY

a) Algorithm

A supervised learning algorithm like multiple linear regression is going to be used in this study.

i) *Assumptions* –For forecasting the temperatures using linear regression, the relationship between dependent and independent variables is assumed to be linear, the residuals will have a constant variance, residual errors are assumed to be normally distributed and residual error terms are independent.

ii) *Criteria* - For temperature forecasts, independent variables should be continuous in nature. If they are categorical, dummy variables should be used.

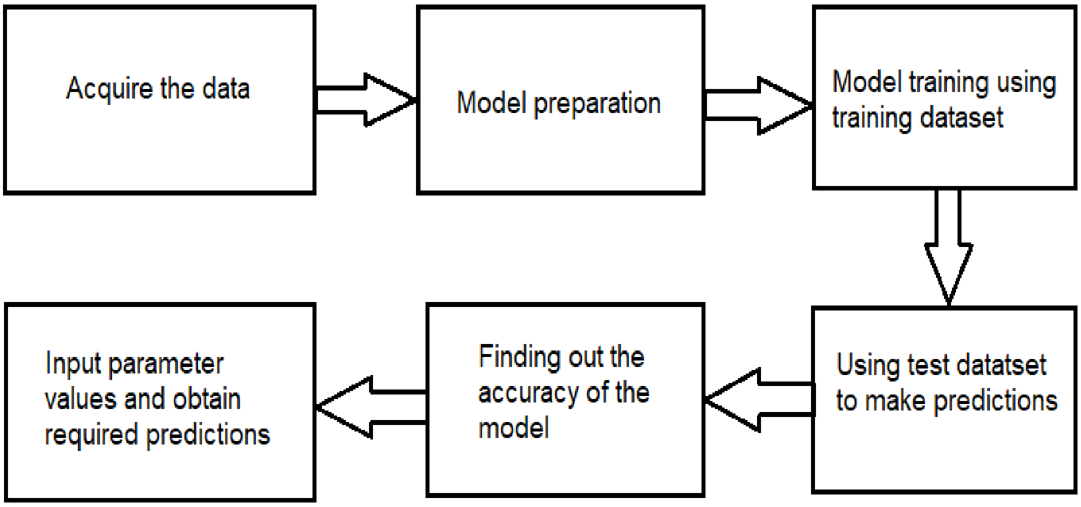
iii) *Available options* – Recurrent Neural Networks, lasso regression, linear regression can be used for temperature prediction.

iv) *Selection* – In this study linear regression will be used for temperature prediction.

v) *Justification* - There is a cause and effect relationship between the variables. We need to predict a continuous variable in the case of temperature prediction.

b) Architecture

Logical Design


Acquire the data
Model preparation
Model training using training dataset
Using test datatset to make predictions
Finding out the accuracy of the model
Input parameter values and obtain required predictions

Technical Design

Data has been downloaded from the link https://www.kaggle.com/rajatdey/jaipur-weather-forecasting and required variables like Mean_Temp, Mean_Humidity, Mean_Dewpoint, Mean_Pressure and Mean_Rainfall was found out in Microsoft Excel. This excel sheet was imported into R studio and coding was done.

Deployments

R Studio-1.4.1717 and R 4.1.1 have been used in this study.

Evaluation criteria

R squared, root mean squared error and mean absolute error can be used to evaluate linear regression model to predict temperature.

c)Code

Environmental setting and Dependency

In R studio, libraries like caTools, modelr and car have been used.

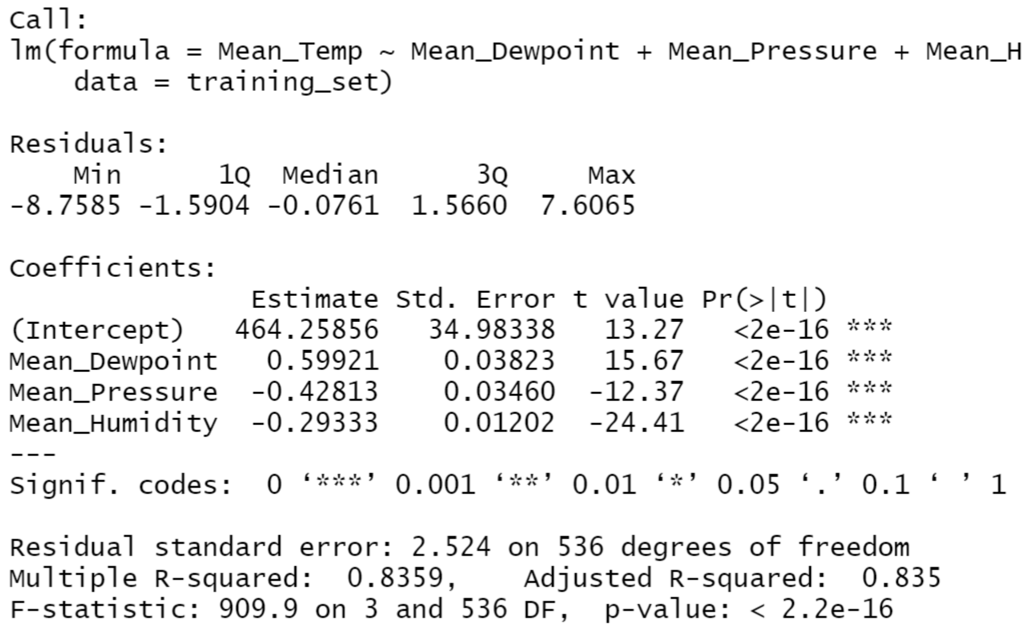
Data store link and Code Repository

The link containing data used in this project as well as the code used in R language are stored in github. The URL is https://github.com/susy23/mlr.git

Result

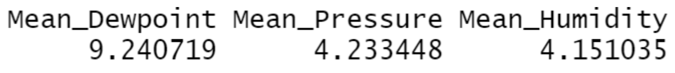
Test results

First, the data is split into training and test datasets. The training dataset contains 540 rows and test dataset contains 136 rows of data. After splitting, a multiple linear regression model is fitted o the training set whose summary is given below.



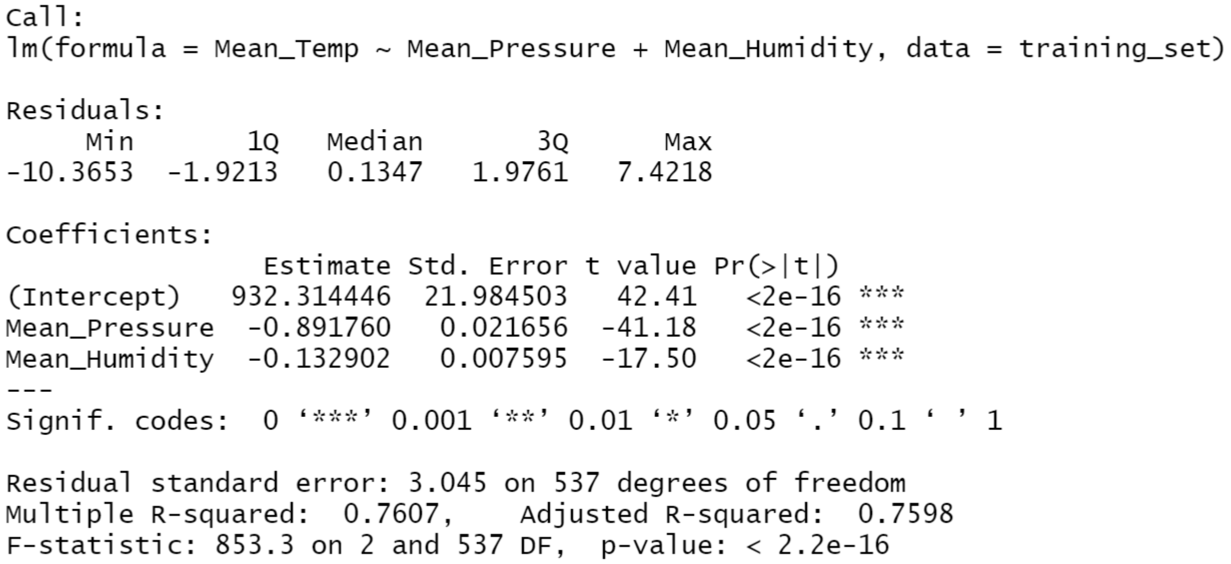
```
Call:
lm(formula = Mean_Temp ~ Mean_Dewpoint + Mean_Pressure + Mean_Humidity,
    data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7585 -1.5904 -0.0761  1.5660  7.6065

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   464.25856   34.98338   13.27   <2e-16 ***
Mean_Dewpoint   0.59921    0.03823   15.67   <2e-16 ***
Mean_Pressure  -0.42813    0.03460  -12.37   <2e-16 ***
Mean_Humidity  -0.29333    0.01202  -24.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.524 on 536 degrees of freedom
Multiple R-squared:  0.8359,    Adjusted R-squared:  0.835
F-statistic: 909.9 on 3 and 536 DF,  p-value: < 2.2e-16
```

The Variance Inflation Factor for this model is as shown below



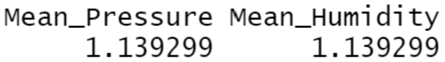
```
Mean_Dewpoint Mean_Pressure Mean_Humidity
     9.240719      4.233448      4.151035
```

Therefore, Mean_Dewpoint is omitted from the model in order to reduce the multicollinearity, and a new model is created whose summary is given below.



```
Call:
lm(formula = Mean_Temp ~ Mean_Pressure + Mean_Humidity, data = training_set)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3653  -1.9213   0.1347   1.9761   7.4218 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   932.314446  21.984503   42.41   <2e-16 ***
Mean_Pressure  -0.891760   0.021656  -41.18   <2e-16 ***
Mean_Humidity  -0.132902   0.007595  -17.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.045 on 537 degrees of freedom
Multiple R-squared:  0.7607,	Adjusted R-squared:  0.7598 
F-statistic: 853.3 on 2 and 537 DF,  p-value: < 2.2e-16
```

The VIF for this modified model is as shown below

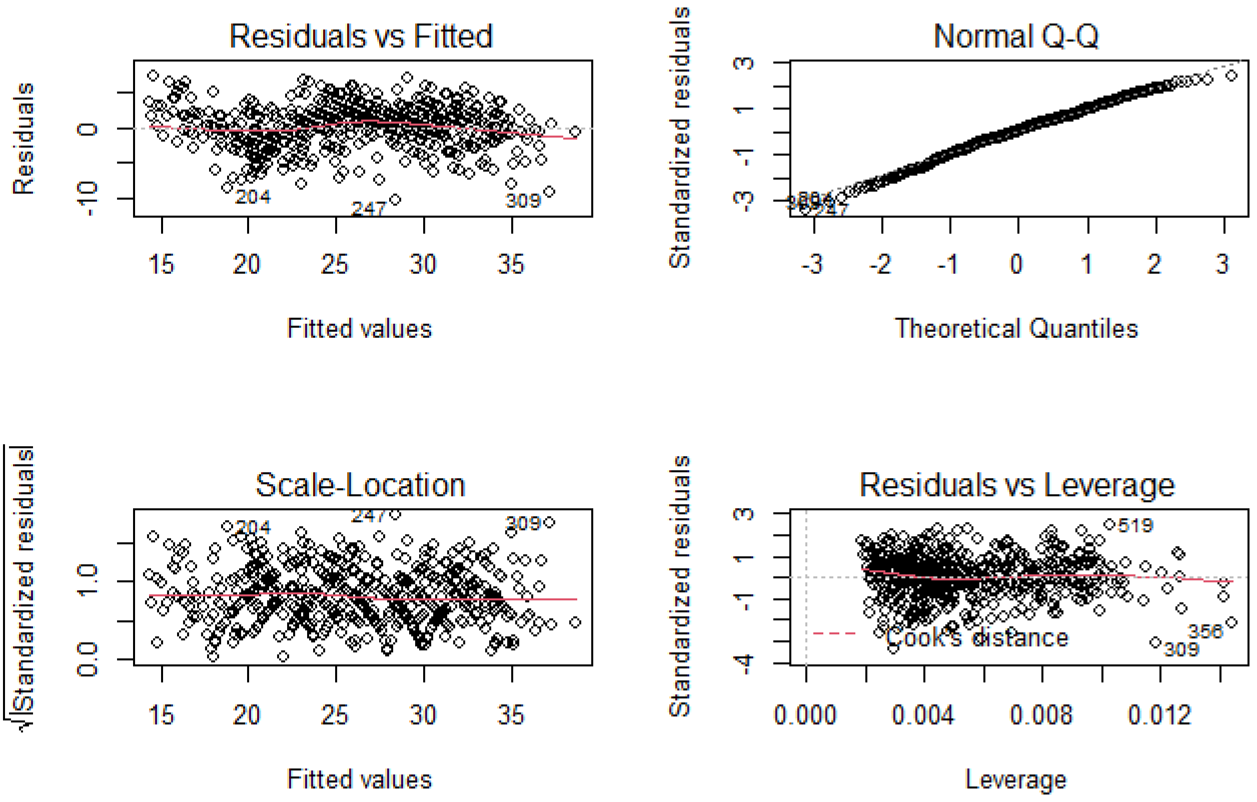


```
Mean_Pressure Mean_Humidity 
     1.139299      1.139299 
```

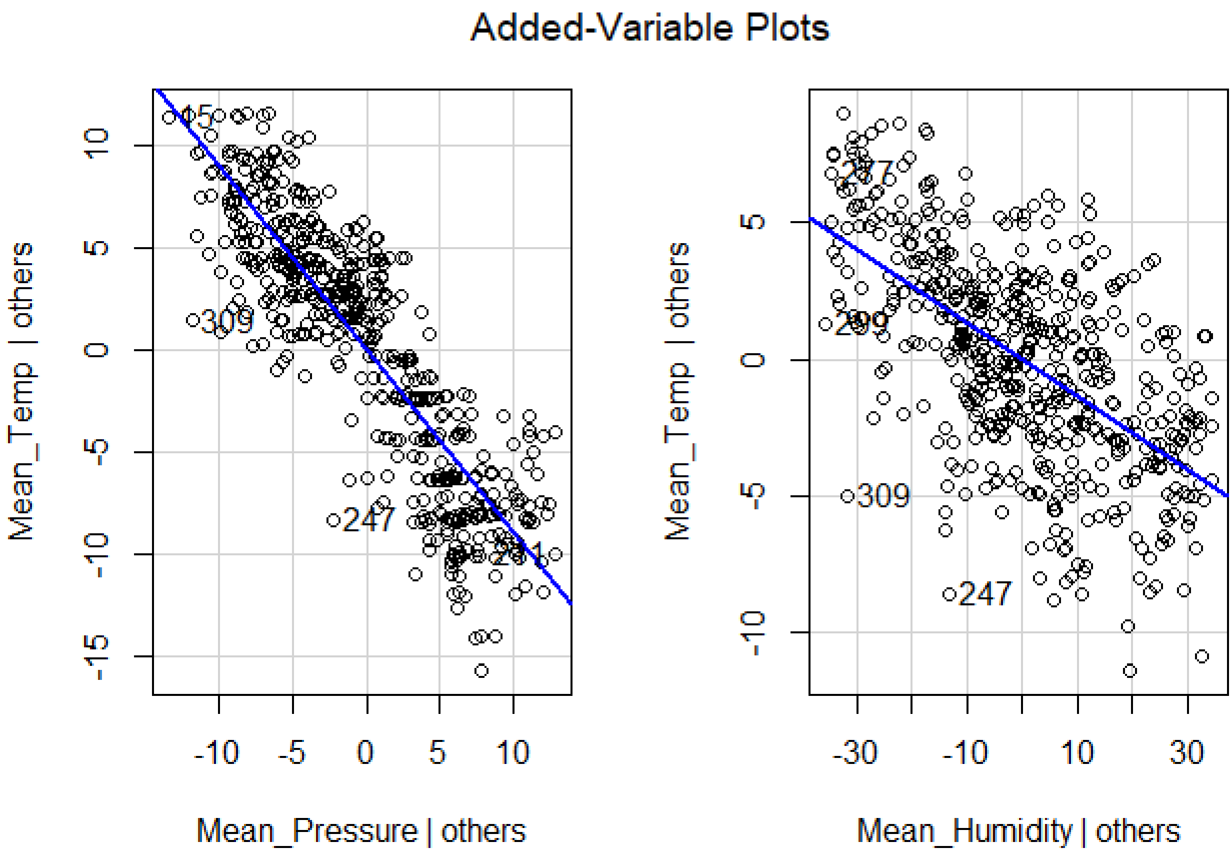
Hence, the multiple linear regression equation for temperature of a given day is

Mean_Temp=932.314446-0.891760*Mean_Presssure-0.132902*Mean_Humidity

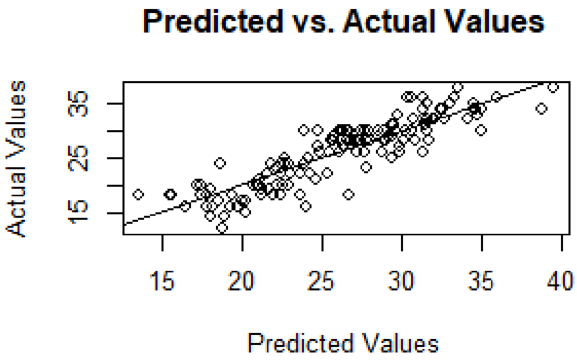
Plots for assumptions



1) Linearity assumption holds good because we can see that the red line is close to the dashed line.
2) Normality assumption holds good because the data falls along the diagonal line in the Q-Q plot.
3) In the scale - location plot, the red line is nearly a horizontal line. Homogeneity of variance(homoscedasticity) can be assumed.
4) In the residuals vs leverage plot, all cases are within Cook’s distance lines. Therefore, there are no influential cases (observations).

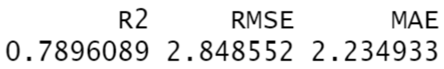


The relationships between the dependent variable and independent variable is got by controlling the other independent variables.



The plot of the predicted temperature values and actual data values in the test data set is given above.

Evaluation metrics



```
       R2     RMSE      MAE 
0.7896089 2.848552 2.234933 
```

The R-squared value, the RMSE value and the MAE value of the testing data with respect to the model are as shown above.

Findings

The R squared value of the test set is 0.7896 which is close to the adjusted R-squared value of the training data set which is 0.7598.

Limitations

1) Linear regression can find out only the linear relationships between the variables. It cannot analyse other underlying other patterns.
2) In this analysis, only mean values of dependent and independent variables have been used. Means cannot completely define a distribution.

Future works

Other significant variables can also be included to give better results.

Conclusion

Multiple linear regression has been used for this project. Even though low values for RMSE and MAE were not obtained, R squared value is optimum since it is above 0.7.

References

1. Gowtham Sethupathi, M., Ganesh, Y.S. and Ali, M.M., 2021. Efficient Rainfall Prediction and Analysis using Machine Learning Techniques. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(6), pp.3467-3474.
2. Holmstrom, M., Liu, D. and Vo, C., 2016. Machine learning applied to weather forecasting. *Meteorol. Appl*, pp.1-5.
3. Sujatha, G., Rao, S. and Srinivasa Rao, T., 2019. PREDICTION OF HUMIDITY IN WEATHER USING LOGISTIC REGRESSION, DECISION TREE, NEAREST NEIGHBOURS, NAIVE BAYESIAN, SUPPORT VECTOR MACHINE AND RANDOM FOREST CLASSIFIERS.
4. Sheikh, F., Karthick, S., Malathi, D., Sudarsan, J.S. and Arun, C., 2016. Analysis of data mining techniques for weather prediction. *Indian Journal of Science and Technology*, *9*(38), pp.1-9.
5. Patil, T and Shah, K. 2021. Weather Forecasting Analysis using Linear and Logistic Regression Algorithm. International Research Journal of Engineering and Technology (IRJET) . 08(06), pp. .
6. Jayasingh, S.K., Mantri, J.K. and Gahan, P., 2016. Comparison between J48 Decision Tree, SVM and MLP in Weather Forecasting. *International Journal of Computer Science and Engineering*, *3*(11), pp.42-47.
7. Radhika, Y. and Shashi, M., 2009. Atmospheric temperature prediction using support vector machines. *International journal of computer theory and engineering*, *1*(1), p.55.
8. Kumar, N. and Jha, G.K., 2013. A time series ann approach for weather forecasting. *Int J Control Theory Comput Model (IJCTCM)*, *3*(1), pp.19-25.
9. Bushara, N.O. and Abraham, A., 2014. Weather forecasting in Sudan using machine learning schemes. *Journal of Network and Innovative Computing*, *2*(1), pp.309-317.
10. Baboo, S.S. and Shereef, I.K., 2010. An efficient weather forecasting system using artificial neural network. *International journal of environmental science and development*, *1*(4), p.321.