DengAI: Predicting disease spread

Akshyta Katyal(2017216) – Chinmay(2017274) – Anushika Gupta(2017135)

1. Abstract

Dengue fever, a fast-emerging mosquito-borne disease, affects millions across the globe every year. A predictive system that can minimize the damage and loss in advance is essential to tackle its outbreak efficiently. In this work, machine learning algorithms such as Linear Regression (Ridge and Lasso Regression), Elastic Net, Random Forest Regression, Support Vector Regression, Bayesian Ridge Regression, Multilayer Perceptron Regression and Gradient Boosting Regression are used to predict its outbreak and are evaluated using mean absolute error as an estimate to find the most suitable model for prediction. The Dataset consists of the values for factors like temperature, rainfall and relative humidity on which the occurrence of Dengue depends in a particular geographical region (San Juan and Iquitos). Using this data, the model learns to predict the number of cases of Dengue and hence its outbreak in that region. It is observed that Bayesian Ridge Regression forecasts the dengue outbreak with least error in prediction.

2. Introduction

In recent years fatal dengue fever has become prevalent in several parts of the world. We are interested in creating a system that can predict its widespread occurrence (number of Dengue cases per week) in the areas of San Juan and Iquitos using the information about the environmental factors like amount of precipitation, temperature, humidity etc. responsible for its spread. Dataset is provided in a competition on the drivendata.org platform.

3. Related Work

The following work has been done on this problem statement –

1. P.Muhilthini, B.S. Meenakshi, S.L. Lekha, S.T. Santhanalakshmi. 2018.” Dengue Possibility Forecasting Model using Machine Learning”

The algorithm used in this work is Gradient Boosting Regression (GBR). Mean Square Error (MSE) and Mean Absolute Error (MAE) has been used as an evaluation metric.

2. Sathler, Carlos. 2017.” Predictive Modeling of Dengue Fever Epidemics: A Neural Network Approach”

The algorithm used in this work is Random Forest Regression and LSTM (long-short term memory recurrent neural network). MAE is used as evaluation metric.

4. Methodology

Our task is a regression problem. The dataset for this problem contains meteorological data for the areas which are analyzed to find the features that give the best prediction.

4.1 Dataset and Evaluation

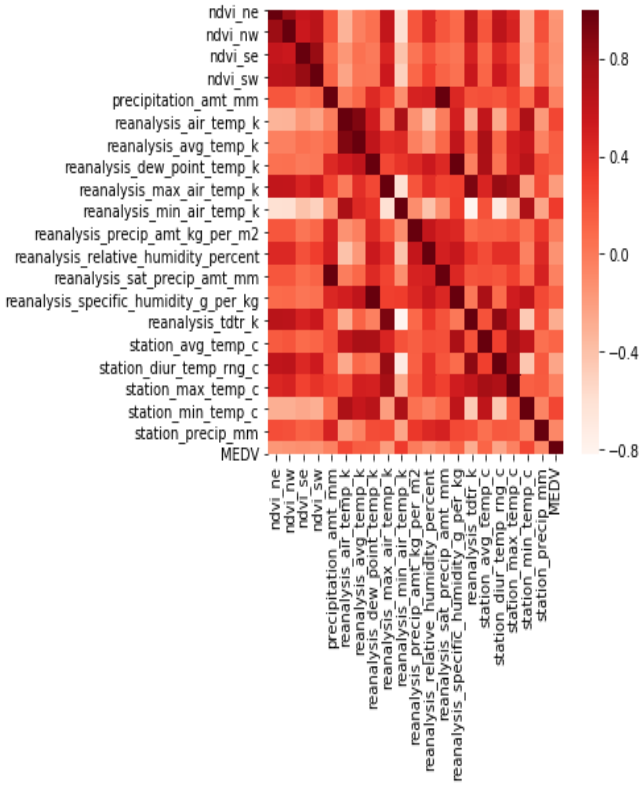
Dataset is provided in the competition on the platform drivendata.org. It consists of 20 features, 1465 training samples and 416 testing samples. Dataset is split into 80 -20 training set and validation set. Thus, validation set consists of 293 samples.

4.2 Preprocessing and Feature Selection

1) The dataset contains a lot of missing values for certain features. These missing values are handled using the technique of Imputation. Simple (Mean) Imputer is used to replace the NaN entries for a particular feature with the mean of the values for that feature in the dataset.

2) Normalization: All the feature values are normalized to 0 mean and unit variance Gaussian values.

3) Using heatmap, all the features significantly correlated with the target value are selected. Also, one among the features highly correlated with each other is dropped so as to ensure a set of independent features.



4) Other techniques used in selecting the most relevant features include Recursive Feature Extraction and Select K Best.

5) The feature “week_start_date” was dropped because timescale is set by year and the week of year features, so it’s stationary with no dependence on start date.

4.3 Outlier Removal Techniques

In order to remove the presence of outliers in the dataset, the techniques like Isolation Forest, Local Outlier Factor and One Class SVM are used. These functions output a negative value for the points that are outliers to the dataset.

4.4 Models

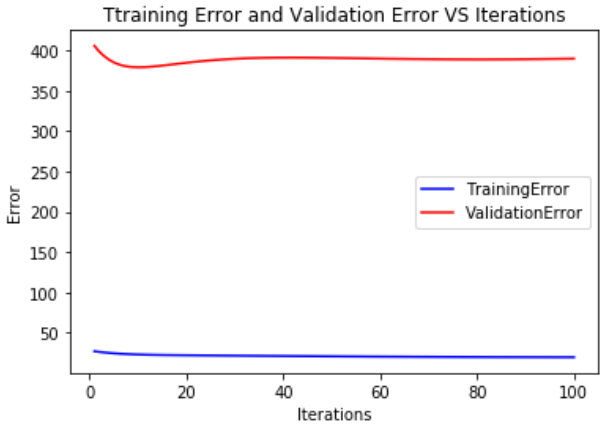
- Gradient Boosting Regressor
- Linear Regressor
- Random Forest Regressor
- Support Vector Regressor
- Multilayer Perceptron Regressor
- Elastic Net
- Bayesian Ridge Regression

4.5 Evaluation Metrics

Our task is a competition on the DrivenData.org platform. The evaluation metric used as a criterion for judging is Mean absolute error. Hence, the ultimate comparison of the performance of our various models is done using Mean absolute error as an evaluation metric. Although we have also used other evaluation metrics like mean squared error, R2 score and maximum error.

5. Result

1) This graph shows the plot of validation error and training error versus no. of iteration for linear regression.

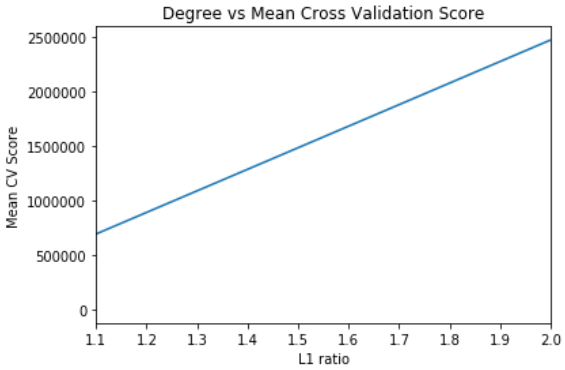


From this plot it can be seen that model is underfitted using linear regression model.

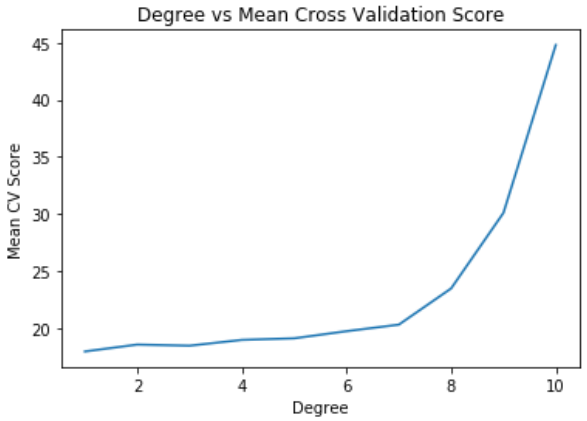
2) For Elastic Net the optimal hyper parameters obtained are **'alpha'**: 0.01, **'l1_ratio'**: 0.9 and **'C'**: 0.5, **'degree'**: 1, **'epsilon'**: 0.001, **'kernel'**: 'linear' for Support Vector Regressor.

.

- For ElasticNet, L1 ratio vs Mean Cross Validation Score is plotted as:



- We plotted the graph between degree of polynomial and mean absolute error to decide the degree of Support vector regression.



This graph shows that error is increasing as degree of polynomial is increasing that is why we used linear kernel in Support Vector Regression

The table below shows the results of baseline model on training set.

Regularization	R2 score	Maximum error	Mean squared Error	Mean Absolute Error
None	0.1487	388.39	1753.9	23.421
L1	0.1384	384.51	1723.5	22.736
L2	0.1473	389.66	1729.7	21.843

The table below shows the results of baseline model on validation set.

Regularization	R2 score	Maximum error	Mean squared Error	Mean Absolute Error
None	0.1287	381.78	1723.8	21.305
L1	0.1302	381.48	1720.8	21.251
L2	0.1267	384.52	1727.8	20.924

Table shows the results of Elastic – Net, Support Vector Regressor and Random Forest Regressor learning technique on validation set.

Model	R2 score	Max Error	MSE	MAE
Elastic Net	0.13	381.31	1878.39	21.21
Support Vector Regressor	0.05	404.29	1718.77	17.16
Random Forest Regressor	0.12	329.00	1732.44	20.67

The evaluation metric used for selecting the best model is Mean absolute error as evaluation of system is done on the competition which uses mean absolute error for evaluation. The mean absolute error of SVR model is minimum on the validation set among the three models that we proposed to work on.

We performed some more experiments to achieve better results. These experiments involved some more advanced models. Corresponding result on training set is as follows:-

Model	R2 Score	MSE	MAE	Maximum Error
Gradient Boosting Regression	0.14	1674.7	20.20	380
Neural Network Regression	0.66	637.8	14.67	236
Bayesian Ridge Regression	0.14	1623.9	14.27	443

And, corresponding result on the validation set is as follows:

Model	R2 Score	MSE	MAE	Maximum Error
Gradient Boosting Regression	0.5	1691.9	20.4	378
Neural Network Regression	0.27	1430.7	19.43	289
Bayesian Ridge Regression	0.05	1872.6	17.76	402

6. Conclusion

We performed extrinsic evaluation on our models by submitting the test results on driven-data competition. The following were the best results we obtained.

Model	Mean Absolute Error
Support Vector Regression	30.48
Bayesian Ridge Regression	26.67
Gradient Boosting Regression	27.85
Multilayer Perceptron Regression	29.56

We obtained best result on Bayesian Ridge Regression with a ranking of 1830 among 7710 participants.

7. References

[1] P.Muhilthini, B.S. Meenakshi, S.L. Lekha, S.T. Santhanalakshmi. 2018.”Dengue Possibility Forecasting Model using Machine Learning”

[2] Sathler, Carlos. 2017.”Predictive Modeling of Dengue Fever Epidemics: A Neural Network Approach”