Machine Learning Project

DengAI: Predicting Disease Spread

Group No. - 12

Competition link :- https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/#features_list

Problem statement

The task is to predict the number of dengue cases each week based on the environmental factors like amount of precipitation, minimum and maximum temperature, humidity etc which are responsible for its spread in San Juan and Iquitos.

Motivation: Early prediction can help the authorities strategically plan and control the Dengue epidemic.

Related Work

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), Ensemble learning technique.
- Mean Square Error (MSE) and Mean Absolute Error (MAE) is used as an evaluation metric.
- The preprocessing technique used is removal of Data instances with missing values.

2) **Sathler, Carlos. 2017.”Predictive Modeling of Dengue Fever Epidemics: A Neural Network Approach”**
- The algorithm used in this work is Random Tree Regressor, LSTM (long-short term memory recurrent neural network), GRU.
- Mean Absolute Error (MAE) is used as an evaluation metric.
- The preprocessing technique used is removal of features like week_start_date and precipitation_amount_mm and “reanalysis_sat_precip_amt_mm that are nearly 100% correlated.
- The mean absolute error for this model came out to be 22.8077.

Dataset

This problem statement is an ongoing competition on site drivendata.org. Dataset is provided in the competition.

SPECIFICATIONS

- No. of features - 20
- Training Set contains 1456 samples.
- Test set contains 416 samples.

week_start_date	1994-05-07
total_cases	22
station_max_temp_c	33.3
station_avg_temp_c	27.7571428571
station_precip_mm	10.5
station_min_temp_c	22.8
station_diur_temp_rng_c	7.7
precipitation_amt_mm	68.0
reanalysis_sat_precip_amt_mm	68.0
reanalysis_dew_point_temp_k	295.235714286
reanalysis_air_temp_k	298.927142857
reanalysis_relative_humidity_percent	80.3528571429
reanalysis_specific_humidity_g_per_kg	16.6214285714
reanalysis_precip_amt_kg_per_m2	14.1
reanalysis_max_air_temp_k	301.1
reanalysis_min_air_temp_k	297.0
reanalysis_avg_temp_k	299.092857143
reanalysis_tdtr_k	2.67142857143
ndvi_location_1	0.1644143
ndvi_location_2	0.0652
ndvi_location_3	0.1321429
ndvi_location_4	0.08175

Preprocessing and Feature Selection

1. **Data Imputation:** Handling missing values in the dataset .

   The following 2 techniques are used,

   - Replacing the missing values of a feature with **mean value** for that feature.
   - **Regression Imputation**: Taking the feature with missing values as a function of other features, and use that to predict the missing value.

2. **Normalization:** Standardization (Z- score)

$$x' = \frac{x - \bar{x}}{\sigma}$$

3. Feature "week_start_date" was dropped.

   This is because timescale is set by year and the week of year features, it is stationary with no dependence on the start of the interval.

Feature Selection

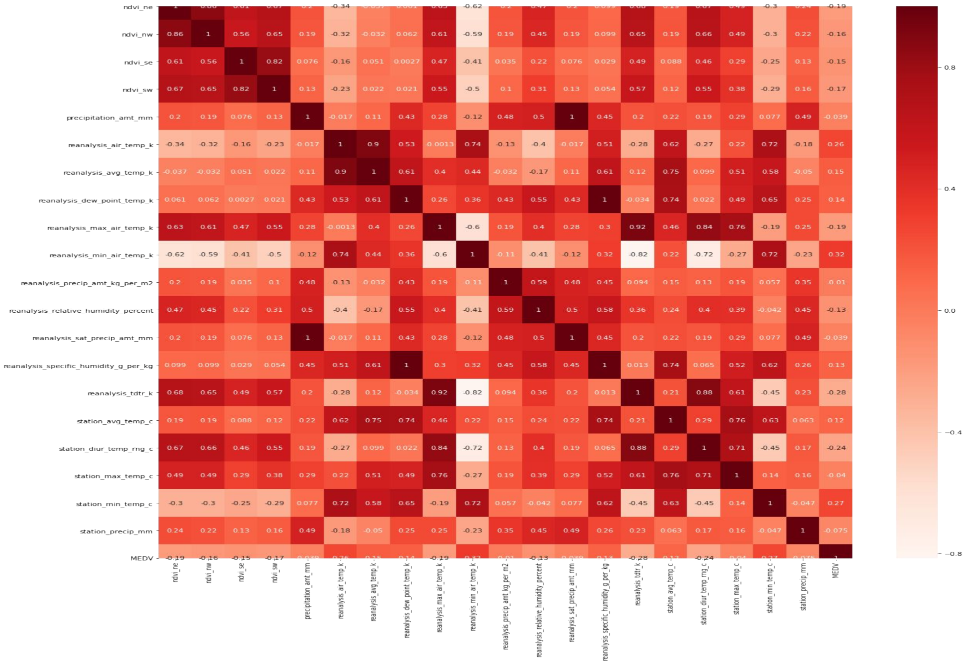
Correlation Values

The techniques used to find the correlation between the features include **Heatmap, Recursive Feature Elimination** and **Best k fit**.

On studying the heatmap to find correlation between the features, 12 out of 20 features are found highly correlated. For example,

cor('ndvi_ne','ndvi_nw) = 0.89

cor('ndvi_se','ndvi_sw) = 0.78



Baseline:- Linear Regression

Score 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

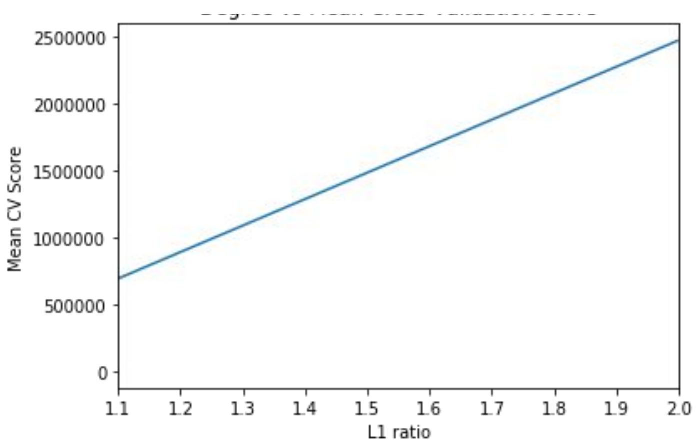
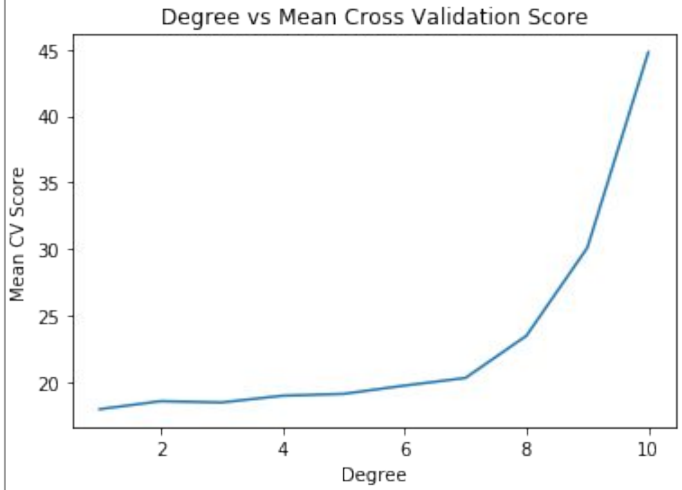
Score on Validation Set:-

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

Advanced Models

Average Cross Validation Error for number of folds = 10

Model	**R2 score**	**Maximum error**	**Mean squared Error**	**Mean Absolute Error**	**Best Parameters**
Elastic Net	0.13133197517	381.3127767086	1878.398473752	21.21590004469	**'alpha'**: 0.01, **'l1_ratio'**: 0.9
Support Vector Regressor	0.05065601882	404.2986039784	1718.770776872	17.16715236285	**'C'**: 0.5, **'degree'**: 1, **'epsilon'**: 0.001, **'kernel'**:'linear'
Random Forest Regressor	0.1244209088	329	1732.44520547	20.67808219178	**'N_estimator':** 250 **'Max_depth':** 4

<table>
<tr>
<th><u>Elastic Net (l1_ratio vs Mean CV score)</u></th>
<th><u>Support Vector Regressor (Degree of polynomial vs Mean CV score)</u></th>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
</table>

Some More Experiments

Result on Training Set

Model	r2 score	MSE	MAE	Maximum Error
Gradient Boosting Regression	0.14	1674.7	20.2	380
Neural Network Regression	0.66	637.04	14.67	236
Bayesian Ridge Regression	0.14	1623.9	14.27	443

Result on Validation Set

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

Extrinsic Evaluation of our model

Results Obtained on the final dataset:-

Model	Mean Absolute Error
Support Vector Regressor	30.4832
Bayesian Ridge Regression	26.6779
Gradient Boosting Regressor	27.8582
Multilayer Perceptron Regressor	29.5697

Best Result obtained :-
Score = 26.6779 Rank = 1830 (out of 7710)

Thank You

Submitted By :

Group 12

Anushika Gupta	2017135
Akshyta Katyal	2017216
Chinmay	2017274