Omdena Silicon Valley Chapter- Xtreme Weather Forecast

Task 1 Exploratory Data Analysis

Task-Leader: Vishu Kalier

Task-Coleader: Pooja

Things We Need To Work On Now

- Evaluation of Null Values, we evaluated them by looking at adjacent columns. The Null values area case of NMAR (Not Missing At Random).
- The null values are specific to weather stations (Ccsm3 and Ccsm30). Also, we can try to evaluate the missing values using KNNImputer. The KNN Imputer works similarly to the KNN Algorithm. Here we have to take care of the number of neighbors (n) we are choosing.
- We can also evaluate the null values on the basis of Ensembled Mean column. Since, the ensembled mean column gives the best computation in general cases, so we can directly paste the ensembled mean value in the missing rows whether it is of Temperature or Pressure.

Columns Which We Can Remove

- We can try eradicating these columns:-
- 1. The entire El-Nino Columns since we have only two years data and El-Nino trend lies between 2010-2030. So we cannot determine the El-Nino Trends within just two years (time range is pretty small).
- 2. Since MEI ratio determines the El-Nino factors, so we can take the MEI ratio.
- 3. MJO basically deals with sub-seasonal climate, so it better to not eradicate these columns.

Distribution of Parameters

Distribution of Parameters in the Dataset:-

- 1. Static – Latitude and Longitude.
- 2. Dynamic Factors (environmental ones)- Pressure, Precipitation, Humidity, Atmospheric Precipitate, Elevation, Geo-potential Height. They are independent variables.
- 3. Dynamic Factors (computed ones)- Weather stations data (for Temperature, Temperature-0 and Precipitation). They are dependent variables.
- 4. Climatic Regions – They are also dependent of dynamic environmental factors (the column can be considered as a dependent one).
- 5. El-Nino Parameters – To be eradicated.
- 6. Time and Date- they are in chronological or sequential order. So they must be taken care of while training the network and used to check the trends.

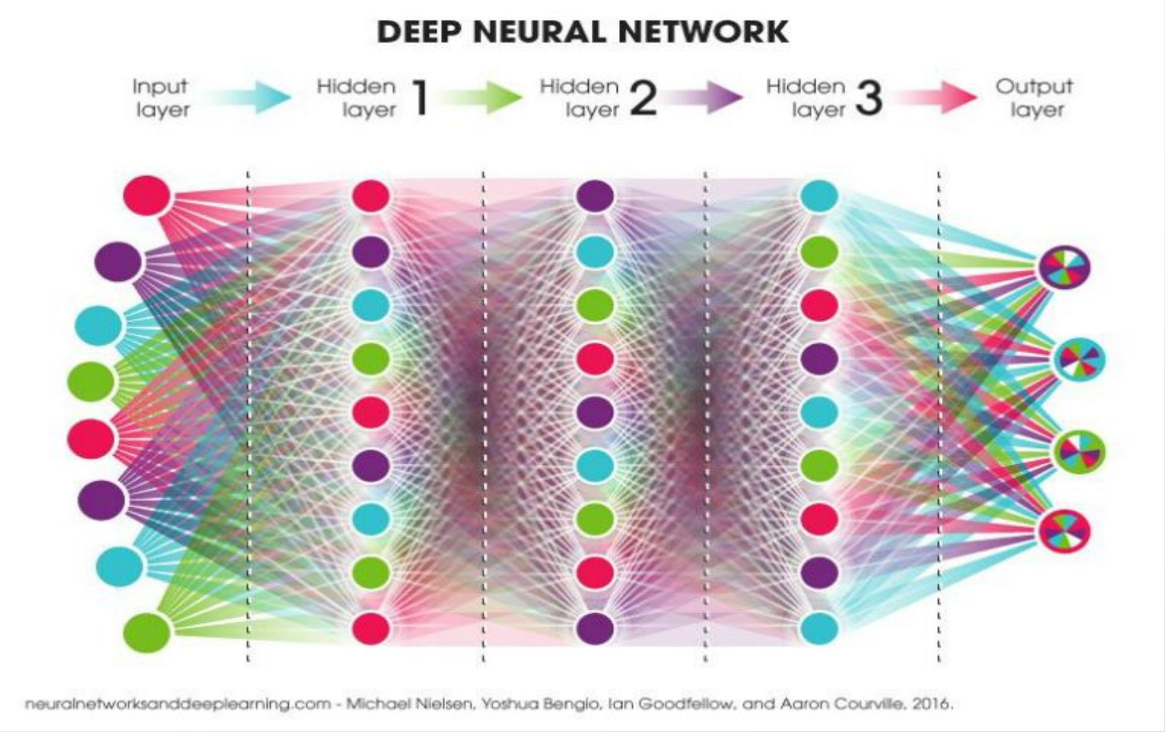
What Will Be The Use Of EDA?

Since we are not allowed to reduce the columns, other than El-Nino because that would lead to underfitting due to loss of data points or dimensions:-

- 1. In these types of cases, where we are performing regression but we cannot reduce the dimensionality so, we start identifying clusters among variables. The problem category changes from Regression to Regression + Classification (performing Regression on the basis of Clusters or Classification).
- 2. Try developing or extracting relationships among columns which you all can do.
- 3. The Clusters formed will work as a bias function or bias parameters for the Neural Network. Using Clusters and SVR we can manipulate the dependent variables and extract more features.

How Will EDA help in Model Evaluation

Now we have to categorize the model, as an example, we can predict the dependent variable Y from the independent variable X.
So the main intuition is to draw a boundary line separating the dependent and independent variables.
Since, we cannot overlook any variable since, we have reduced the columns to a good extent, so it is advisable to use a Dense Neural Network and not a Sparse Neural Network probably Dense DLSTM.



neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

Visualization Tools for EDA

- We can use a variety of visualization tools:-
- 1. Heatmaps – correlation for columns.
- 2. Bar Plots – Mean and Standard deviation and other mathematical intuitions.
- 3. Line Plots – Observing the Time Series frame and the trend in general.
- 4. Contour Plots – Visualizing the slope, descent, intensity and depth.
- 5. Histograms – For General Distribution functions and probability distributions.
- 6. Scatter Plots – Distribution of variables (data points in general).
- 7. Violin Plots – To get the mean-line, median-line, quartiles.
- 8. Box Plots – To get the outliers in the various columns.
- 9. Pie Charts – To get the distribution density of unique column values.
- 10. Quiver Plots – To check the vectors (especially for wind directions).

Feature Scaling and Column Minimization

- Feature Scaling and Column Minimization:-
- 1. We should look for the range of the column data. The range will provide us more insights about the scaling factor.
- 2. Then using the appropriate scaling techniques and then cross-validating them is of the main priority.
- 3. PCA (Principal Component Analysis) can be used for finding the most relevant columns for the target temperature column.

Mathematical Concepts To Ponder

- Mathematical Concepts which will come in handy:-
- 1. Matrices, rank of a matrix, column and row transform of a matrix.
- 2. Slope of a surface and first and second order partial derivatives.
- 3. Distributions like Normal, Gaussian, Standard, Exponential, etc.
- 4. Clear and concise understanding of Mean, Median, Variance, Standard deviation and Probability deviation.

With a little bit of touch of these mathematical essential, one can draw very powerful insights even from a simple plot as well.

Approach Regarding EDA and Model Development

- These are my few opinions regarding EDA and Model Development:-
- 1. The main priority is to identify the independent and dependent variables after excluding the El-Nino Parameters.
- 2. Since DL goes with EDA, we should start working on DL as well. The first we can do is to get the accuracy of the raw data to training data with a simple regression line (the base case).
- 3. After the independent variables are enumerated, we can work by splitting the dataset on basis of seasons and other factors. Using proper visualizations will behave as a booster.
- 4. The clustered formed can be used as an optimizer or loss function or bias parameter or weights to train the Neural Network. Just to remind, we are working to Improve the performance and not degrade it, so the best way to work is by hit and trial method, because even if we have small time range data the number of parameters (dimensionality is pretty vast, 246 columns).

Opinions of the Team Members