DATA SCIENCE PROJECT

1.0 TASK ONE

1.1 Goals of the assignment

The goal of this assignment is to check your understanding on machine learning and aspects of data analysis.

1.2 Expectations

From these assignment we will be looking at:

- Importation of the various libraries
- Merging data
- EDA: Handling anomalies in the data and visualizing the data.
- Feature engineering and feature selection.(You can go the extra mile and see how you can implement LassoCV and PCA in feature selection)
- Model development: Splitting data (opt. Check out cross validation and see its advantages and how it is implemented), training various models(Choose the best model), hyperparameter tune the best model, test tuned model on test data set, evaluate model.

1.3 ASSIGNMENT

The data is on the spread Dengue Fever in two cities. Your goal is to predict the total cases label. There are two cities, San Juan and Iquitos, with data for each city spanning 5 and 3 years respective. You will have to merge the data in the labels and features.

The features include:

- city – City abbreviations: sj for San Juan and iq for Iquitos
- week_start_date – Date given in yyyy-mm-dd format

NOAA's GHCN daily climate data weather station measurements:

- station_max_temp_c – Maximum temperature
- station_min_temp_c – Minimum temperature
- station_avg_temp_c – Average temperature
- station_precip_mm – Total precipitation
- station_diur_temp_rng_c – Diurnal temperature range

PERSIANN satellite precipitation measurements (0.25x0.25 degree scale)

- precipitation_amt_mm – Total precipitation

NOAA's NCEP Climate Forecast System Reanalysis measurements (0.5x0.5 degree scale)

- reanalysis_sat_precip_amt_mm – Total precipitation
- reanalysis_dew_point_temp_k – Mean dew point temperature
- reanalysis_air_temp_k – Mean air temperature
- reanalysis_relative_humidity_percent – Mean relative humidity
- reanalysis_specific_humidity_g_per_kg – Mean specific humidity
- reanalysis_precip_amt_kg_per_m2 – Total precipitation
- reanalysis_max_air_temp_k – Maximum air temperature
- reanalysis_min_air_temp_k – Minimum air temperature
- reanalysis_avg_temp_k – Average air temperature
- reanalysis_tdtr_k – Diurnal temperature range

Satellite vegetation - Normalized difference vegetation index (NDVI) - NOAA's CDR Normalized Difference Vegetation Index (0.5x0.5 degree scale) measurements

- ndvi_se – Pixel southeast of city centroid
- ndvi_sw – Pixel southwest of city centroid
- ndvi_ne – Pixel northeast of city centroid
- ndvi_nw – Pixel northwest of city centroid

1.4 References

Data, D. (n.d.). *Driven Data*. Retrieved from Dengue Fever: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/80/

2.0 TASK TWO

2.1 Goal

To check your understanding on unsupervised machine learning

2.2 Assignment

You own a mall and want to understand the target customers so that the sense can be given to marketing team and plan the strategy accordingly. With data about your customers like Customer ID, age, gender, annual income and spending score. Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.