w207 Final Project - Helios

Temperature prediction using multivariate time series analysis

SIMON STOREY, DIVYA BABU & MAJID ALZEIN

GOALS

- Predict temperatures for a range of locations along the coast of the mainland USA
- Identify any trends / patterns between long and short term signals i.e. sea surface temperature, atmospheric pressure
- Identify correlations across multiple time series and use these in our forecasts
- Implement and benchmark performance of a number of supervised methods and compare these vs. traditional techniques using ARIMA

DATA FEEDS

- Data will be sourced from the following providers/aggregators
  - Berkeley Earth (http://berkeleyearth.org/source-files/)
  - NOAA (https://www.ncdc.noaa.gov/data-access)

- Data once transformed is simply a series of data points for each geo-station (with lat/long identifier)
  - % Station ID, Series Number, Date, Temperature (C), Uncertainty (C), Observations, Time of Observation

INVESTIGATION

- As part of the project we will look to
  - Investigate univariate and multivariate time series
  - Run classic timeseries analysis in R to identify key features of the datasets including trends, seasonality, drift
  - Use feature engineering to transform the time series into a format suitable for supervised learning extracting additional features for trends, seasonality
  - Run the following algorithms on the data
    - Forecast (Rob Hyndman) in R
    - Linear Regression
    - Gradient Boosted & Random Forest
    - Long / short term memory RNN
    - Prophet (Facebook Research) which uses an additive regression model

Useful links

- https://github.com/robjhyndman
- https://research.fb.com/prophet-forecasting-at-scale/
- https://machinelearningmastery.com/multi-step-time-series-forecasting/
- https://machinelearningmastery.com/feature-selection-time-series-forecasting-python/
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-276