Geoffrey Leigh

Capstone: Big Data Analytics

Prof. Jaume

May 7, 2017

*Results of the Project*

The results from my project were to predict the levels of pollution in near the near-term.

The output will be a mean value for that day. Any levels about the mean have a binary response. I would classify any value higher than the mean as unsafe and anything below as safe. As I was dissecting the data, there were strong correlation with Carbon and Sulfates on a daily level. I also had a high F-score with all the independent variables. (See png files on github). I then proceed to use Times series forecast to predict make predictions.

Six Time Series Models were used to best predict the four pollutants. The time series models used were Time Series Regression, and ARIMA Model. Taking the best model with less error rate for each pollutants.

Linear Model: Linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

ARIMA Model: ARIMA models are the most general class of models for forecasting a time series which can be made to be “stationary” by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). A random variable that is a time series is stationary if its statistical properties are all constant over time.  *A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion*, i.e., its short-term random time patterns always look the same in a statistical sense.  The latter condition means that its *autocorrelations* (correlations with its own prior deviations from the mean) remain constant over time, or equivalently, that its *power spectrum* remains constant over time.  A random variable of this form can be viewed (as usual) as a combination of signal and noise, and the signal (if one is apparent) could be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also have a seasonal component.  An ARIMA model can be viewed as a “filter” that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts.

 A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series *forecasting* is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time