Team Name: DS420 – Factoria

6/3/2018

Capstone Project

**KDD Cup of Fresh Air**

Breathing; It’s the most essential part of life. As populations increase and more vehicles enter the roadways (especially in large, dense cities), a focus for maintaining air quality is critical. Studies show the health repercussions for breathing in certain air pollutants is extremely high. Particulate Matters (PM) of 2.5µm or less are said to be one of the deadliest forms because, if inhaled, can penetrate deep into the lungs and even enter the blood stream. This can go on to cause DNA mutations, neural system damage, cancer, and unfortunately premature death. Obviously, these health risks are enough to create a lot of concern in the general population of big, condensed cities. The governments of these cities (we’ll be focusing on Beijing and London) are extremely interested in finding out if any data that is already captured can be analyzed in order to identify patterns that would lead to air quality improvements. Also, in the immediate future, these predictions could help identify certain locations in the city that will have high air pollutants, leading to governments sending out warnings that people should stay indoors.

The main pollutants that we focused on in our predictions were the ones that can cause the most damage to the human population. Those are PM2.5, PM10, and 03. The challenge was to be able to predict the levels of these pollutants at every hour for 48 hours. Beijing has 35 air quality station locations, and London has 13 locations that we were required to provide predictions for. Doing the math, we had predictions for 48 stations, and each station would have 48 hours of predictions for each of the 3 pollutants (48\*48 = 2,304 rows of data and 3 columns for each pollutant).

The major downside of this project was not being able to see the results right away. Yes, we could break up the data into a training set and testing set (which we did), and see how our predictions compared to actuals, but the results weren’t as understandable. This is because the KDD Cup created a complex calculation they used to evaluate our submissions (The Symmetric mean absolute percentage error – “SMAPE”). For example, even if our R-squared result was awful when testing against our regression training model, it might surprisingly perform relatively well when submitted for evaluation. That was what caused some confusion around this project; We had to wait 3 days after each submission, and then we were able to see how our predictions of these 3 values matched the actual air quality data using the SMAPE evaluation. We had to make sure we documented what we did for each submission, so that we could understand how our different models and model parameters behaved. There was also the fact that some things are unpredictable; one day the air quality could be due to something that we don’t collect data on, and our model performs poorly. The next day, the same model could perform extremely well because that unknown factor is no longer present. In this way, sometimes it was difficult to understand why we received a poor score (or stellar score!).

The data that we did have available to us was the hourly 2017 historical data for each of the air quality stations in which we had to predict. Along with this, we had up to the hour API data that we could bring in at any time which we were able to take advantage of. Before our submissions, we could bring in the latest which at least helped in our ARIMA time-series model. We also were able to extrapolate more features with the features that were given to us. For example, with the date/time column, we could obtain the hour, the month, whether it was a weekday or not, holiday, etc. (We’ll expand on this feature engineering topic later in this paper). Along with air quality station data, for each city location (Beijing and London), we had weather station data which was collected in a grid fashion (imagine a square grid placed over the city). These stations collected data on weather (ex. Rain, Partly cloudy, sun, etc.), temperature, pressure, humidity, wind direction, and wind speed. We ended up using historical weather data and future forecast data in our linear regression model during the last couple of days in the competition. For our other models and rest of the days, we ended up using only historical weather data. This could have been helpful in improving our model, but we didn’t find the time or have the necessary knowledge to add it into our models until late in the competition. We will talk more about how we utilized this weather data later in this paper. Other than what was provided on the KDD cup website, we did not use outside resources as it was against the rules of the competition.

As far as submitting each of our models, it would have been helpful to set up something that automatically collected the API data, ran our model, and submitted on a daily-basis. Since we are amateurs in this field and did not get help in setting something up like this, we had to make sure that we were available to submit before UTC midnight (5pm Pacific time) which sometimes proved to be difficult. That being said, let’s get into the details of our project and what we were able to accomplish. In this paper we will discuss how we analyzed our data and how we created features with the data that was given to us. Furthermore, we will describe how we selected the models we used, and how we changed different model parameters to improve and obtain the best results.