DS 670 – Assignment 11 - Draft Manuscript

Aakash Parwani

April 07, 2017

Discussion:

Arhus is Denmark's second-greatest city and the cash related concentration of the Central Denmark Region. The city has a catchment zone of 1.2 million people inside a one-hour travel go and is especially connected with Copenhagen and Hamburg. The tenants of Aarhus live inside walking division of parks and recreational reaches, and inside a 15-minute bike ride of an immaculate coastline, and they advantage from close-by provisions of clean drinking water. In a nearby organized exertion with the business assemble and the city's various data foundations, the City of Aarhus will reduce the city's CO2 transmissions and make keen game plans and green advancement. Creative demonstrating wanders ensure exchange offers of home-created environment courses of action abroad, attract overall theory and fulfill the goal of being CO2-fair by 2030.

This document proposes a discussion of the analysis results obtained from the study of Aarhus city weather data set. And there is an attempt to compare the obtained results with the competitor articles analysis. Basically, document will be divided in sections and we will evaluate result of each section in terms of competitor attempt and proposed attempt.

1. **DATA:** CityPulse gives a game plan of open-source fragments and demonstrators for sharp city application engineers, offering access to and organization of the re-usable parts like: - Online-Data-Repository, CityPulse-3D-Map, City-Pulse-City-Dashboard, Data-Quality-Explorer, Social-Media-Analyzer.
   My Result: The weather data for the city of Aarhus in Denmark is open for examinations reason, available at Weather Data. The dataset is a social occasion of weather discernments from the city of Aarhus. Estimations are recorded from February 2014 – June 2014 and August 2014 – September 2014. So it saves most of the data cleaning time because all the datasets are of same format.

   Competitor: The datasets acquisition is not trivial task. Next sources were chosen subjectively. 66 places (www7.ncdc.noaa.gov; daily data): Nwso Agana (Guam; 1), Aarhus Lufthavn (Denmark; 2), Abbeville (France; 3), Aeropuerto Pettiros (Paraguay; 4), Amman Airport (Jordan; 5), Amsterdam AP Schiph (Netherlands; 6), Annaba (Algeria; 7), Ashgabat Keshi (Turkmenistan; 8), Auckland Airport (New Zealand; 9), Bangkok Metropolis (Thailand; 10), Beijing (China; 11), Ben-Guron Int. Airport (Israel; 12), Beograd-Surcin (Serbia; 13), Bogota-Eldorado (Colombia; 14), Brasilia-Aeroporto (Brazil; 15), Bratislava-Letisko (Slovakia; 16), Bruxelles National (Belgium; 17) etc. Because data is obtained from 66 places that causes linguistic problems and data format problem so it affects the total execution time of algorithm.

2. **PARAMETERS:** For weather analysis we are considering six important variables.
   - **Independent Variables:** Dew Point, Humidity, Pressure, Temperature, Wind Direction.
   - **Dependent Variable:** Wind Speed.

   My Result: For weather data analysis, exploration and prediction six variables are considered. My algorithm first calculates the relation between all independent variables. Then it investigates the relation between each independent and dependent variable. So this attempt is reliable because weather change depends on interaction of many parameters. And after consideration of relation (positive or negative) final algorithm is designed.

Competitor: For weather analysis only one variable **Air Temperature** is considered. Prediction of air temperature is performed using time series analysis. That means competitor is only considering air temperature and trying to understand the pattern followed by variable in the past years. This is not the reliable approach for prediction because as we discussed earlier weather components depend on other environment variables also. So it is important to first consider other relations and then come to a decision of modeling prediction.

3. **DATA CLEANING:** Data cleaning incorporates the disclosure and ejection (or cure) of missteps and abnormalities in a data set or database in view of the degradation or wrong segment of the data. Inadequate, off kilter or unnecessary data is perceived and after that either supplanted, adjusted or eradicated.
   My Result: Performed null values detection on training and test dataset. And found all the variables with 10 percent of null values. Then, designed a dynamic procedure to update those null values with mean of particular variable if null values existence is less than 10 percent, else update null values with centroid of the parameter data. This approach provides assurance that we are taking care of data cleaning.

   Competitor: In competitor article there is no discussion about data cleaning procedure. As data is already fetched from more than 60 different locations, data requires cleaning procedure to handle null values and wrong segments.

4. **DATA EXPLORATION:** Specialists routinely use data visualization programming for data exploration since it grants customers to quickly and essentially observe most of the vital segments of their dataset. From this movement, customers can perceive elements that are likely going to have entrancing discernments. By indicating data graphically - for example, through scatter plots or bar charts - customers can check whether no less than two components interface and choose whether they are awesome plausibility for further start to finish examination.
   My Result: Used intelligent **IDE Zeppelin** to first understand the distribution of all the parameters used in the study. Performed data aggregation to know the hidden facts of the dataset. Used **pearson correlation statistical method** to understand the relation between parameters used in the study. Designed bar charts and scatter plots to get visual understanding of relation between variables.
   There was **positive** relation between dewpoint & windspeed, **negative** relation between humidity & windspeed, **negative** relation between air pressure & windspeed, **positive** relation between temperature, windspeed & **positive** relation between winddirection & windspeed.

   Competitor: Competitor article has discussed close relation of precipitation and air temperature. But it fails to describe whether the relationship is positive or negative.

5. **FORECAST MODELS:** There are three types of weather forecast models **a).short range b).medium range c).long range.** Short range generally forecast to go out to **72 hours or less**. Medium range includes forecast from **3 to 7 days.** Long range includes forecast beyond **7 days.** As range increases model forecasting error also increases.

   My Result: Used short range forecasting model for prediction of wind speed. Short-range forecasts tend to focus on the exact details, such as temperature gradients, precipitation, and mesoscale phenomena. Multi linear regression analysis is performed. That considers relations between variables like: Temperature, Wind speed, Wind direction, precipitation etc. That improves the **accuracy** of model to 85 percent. And because of short-range forecasting **mean absolute error (MAE)** of my algorithm is 5.7^0 F. And there are chances of improvement.

   Competitor: Competitor has proposed long-range forecasting of average daily air temperature using inductive method. The forecasting method used has **significant errors** at long-range period (more than 20% in the mean). The model designed is combination of air temperature time series analysis for different places. So the **accuracy** depends on the number of places into consideration. Recent research based on the inductive methods showed possibility of the long-range (half-year lead-time) forecast with **mean absolute error** (MAE) up to 8^0 F.

Conclusion:

After the use of short-range forecast method there is significant improvement in the error rate of prediction. And consideration of relationship between different independent variables gave a good platform for prediction of **wind speed**. Statistical model used in this analysis is still under process of improvement. Soon, there will be inclusion of new parameters in the study that will be combination of existing parameters.

Future Enhancement:

Future enhancement of existing model would be to include time series analysis and prediction of air temperature.