Capstone Project

Medhat Fawzy

Machine Learning Nanodegree Programme

1. Domain Background:

A typical 4G network consists of several nodes connected to each other, where each node serves users in the surrounding area. While users are accessing network services, their mobile phones record Key Performance Indicators (KPIs) which can help network operators assess their service quality.

There are many KPIs which evaluate different aspects of the network. For example, RSRP (Reference Signals Received Power) is a KPI measuring network coverage in the user's location. Traffic Volume is another KPI which measures how much data has been consumed by the user.

2. Problem Statement:

The challenge is to visualise the network data of mobile phones of operators (A – B – C). Each operator has its coverage, RSRP, traffic volume (Download – Upload), etc. for its users then try to get insights to evaluate the competition. Also based on the data given,

1. Assuming that the coverage next week will improve in that polygon compared to the competition (i.e., RSRP will get better than other operators). What would be the impact on downlink and uplink traffic volumes?
2. Samsung devices are the main handsets in our network. Can you predict the traffic volume growth, uplink and downlink, over time for these devices and compare it to the competitors?

3. Datasets and Inputs:

We will be provided with two crowdsourced datasets: RSRP and Traffic Volume. Each dataset has the corresponding KPI measurements collected from mobile phones of different users over a week. It also includes the user's location, operator, phone model and other information. A detailed description of each field can be found in DataDescription.xlsx.

4. Solution Statement:

For the dashboard, I'll be using the Holoviz tools such as Holoviews, Datashader, Colorcet and Panel. The Dashboard consists of four maps: Users during a time period, a heat map of the users of each operator, a hex map showing the density of users in a specific region and a bar chart of RSRP per device type per operator.

For the ML part, we'll use the prophet model to predict the traffic volume growth over time for Samsung devices and compare it to the competitors. Also, a regression model will be used to predict the impact of RSRP improvements on traffic volume.

5. Evaluation Metrics:
The R^2 metric will be used to evaluate the regression model.
Part of the dataset will be used to validate the model to verify the time-series model accuracy.

6. Project Design:

6.1 RSRP data cleaning:
First, some EDA will be applied and some data cleaning as well. A column is dropped and the Timestamp column is converted to DateTime type.
After cleaning the data it will be saved to be used later in the dashboard.

6.2 TrafficVolume data cleaning:
Same as with the RSRP data. Some EDA first, then the country column dropped and the timestamp column converted to DateTime type. The data is also saved for later usage in the dashboard.

6.3 Merging data:
After cleaning the two raw datasets, I combined them into a single dataset that we are going to use with our machine-learning models. The data is grouped by the common columns except for the Timestamp column, this column is broken into its components and the rows are matched against the date, hour and if they happen in the same period of 15 minutes.
Data is also saved for later usage.

6.4 Machine Learning:
After the cleaning and merging process, I take the merged data to the Machine learning part. I test some regression models to see how they perform and make some conclusions.