Grab AI for SEA Submission

Harshi Mehrotra

Problem Statement: Traffic Management

**Executing the File:**

Please make sure following packages are loaded before executing the python jupyter notebook.

* Sklearn
* Pandas, Numpy, Datetime
* Networkx
* Keras/Tensorflow

The file uses the training data set as provided, directory of the training dataset should be same as python notebook.

**Background:**

The data looks to predict normalized traffic demand of a city. In the data provided normalized demand for a geohash6 location is given for frequent time intervals.

**Approach:**

We look to build features to help us better understand the area, the time feature and pattern of traffic nearby a geohash location. We leveraged time-series modelling - ARIMA, Linear Regression and Random Forest models to predict the traffic values.

**Solution:**

Features developed

* Time wise - Weekend/Weekday
  + We looked at daily peak traffic during office travel hours - 7 am to 10 am and 5 pm to 7 pm.
  + There was a clear identifiable pattern repeating every 6-7 days, it helped us distinguish the weekends from weekdays.
* Geographical proximity
  + We used graph network to identify the geohash points next to one another
  + The graph provided clear clusters of 4
  + We used lat-long to build a grid of all the geo-hash - identifying 4 closes location to every point. These 4 points were above, below, left and right of point, considerations were taken for edges.

**Method 1:**

* Grid modeling - assuming traffic flow, a point near to the geohash6 location would help predict demand in the location
* We identified traffic at points near the geohash location in directions - above, below, left and right. Traffic 15 min before are used as prediction covariates for traffic
* We used linear regression and Random Forest models - evaluating them on Mean Squared Error

**Method 2:**

* A simple time-series model for individual Geohash location, using the given time features
* We built and increasing time series with increment of 15 min, we used the past pattern of the demand to predict the demand for final days.

**Error Measurement:**

Upon checking the error values of Mean Squared Error.

**Result:**

Random Forest model produced the best outcome based on MSE error.