Capstone 1 Project Proposal

Background

Call Detail Record [CDR] describes a specific instance of a telecommunication transaction that passes through a network element. Every time a user performs a telecom activity such as send/receive SMS and calls, a CDR is generated. It contains information about the caller/sender ID, location, time, data used, etc. Millions and millions of such records are generated and is mainly used for billing purposes by the telecom company.

Client

Analysis and modeling of this time series data helps to identify usage patterns over a period of time and across geographical grids. This helps in decision making of resource allocation by the telecommunication company.

Datasets

Telecom Italia organized the 'Telecom Italia Big Data Challenge' in 2014, they provided data of two Italian areas: the city of Milan and the Province of Trentino. It is a rich, open multi-source aggregation of telecommunications, weather, news, social networks and electricity data. The data pertaining to the challenge have been released to the research teams under the Open Database License (ODbL) and is maintained by Harvard Dataverse.
For this project we will be using 2 months (November & December) of telecommunication activity data from the city of Milan, Italy and its census data.

Telecommunications activity:

This dataset represents the telecommunication events which took place within Milan from Nov-01-2013 to Jan-01-2014.
File type: TSV
File size: ~ 200MB – 300MB with ~4 million rows each file
Files: sms-call-internet-mi-2013-11-01.txt to sms-call-internet-mi-2014-01-01.txt
Columns:

- *Square id*: identification string of a given square of Milan GRID;
- *Time Interval*: start interval time expressed in milliseconds. The end interval time can be obtained by adding 600,000 milliseconds (10 min) to this value;
- *SMS-in activity*: activity proportional to the amount of received SMSs inside a given Square id and during a given Time interval. The SMSs are sent from the nation identified by the Country code;
- *SMS-out activity*: activity proportional to the amount of sent SMSs inside a given Square id during a given Time interval. The SMSs are received in the nation identified by the Country code;
- *Call-in activity*: activity proportional to the amount of received calls inside the Square id during a given Time interval. The calls are issued from the nation identified by the Country code;
- *Call-out activity*: activity proportional to the amount of issued calls inside a given Square id during a given Time interval. The calls are received in the nation identified by the Country code;
- *Internet traffic activity*: number of CDRs generated inside a given Square id during a given Time interval. The Internet traffic is initiated from the nation identified by the Country code;

- *Country code*: the phone country code of the nation.

Grid:

The original bundle of datasets come from various companies (telecommunications, weather, news, social networks and electricity) with different standards. To ease the comparisons of different geographical areas, the city of Milan's spatial distribution is aggregated in a grid with square cells. The area is composed of a grid overlay of 10,000 squares with size of about 235×235 meters. This dataset provides the geographical reference of each square which composes the grid in the reference system: WGS 84—EPSG:4326

File type: geojson

File size: 3MB

File: milano-grid.geojson

Columns:

- *square id*: identification string of a given square of the Milan or Trentino GRID;
- *Time Interval*: The cell geometry expressed as geoJSON and projected in WGS84 (EPSG:4326).

Census:

Istat provides census datasets that can be aggregated at Grid level for more insights.

Approach

- Identify high and low performing grids
- Month on month trend analysis
- Weekday level analysis
- Forecast the volumes for the first week of January using the top grids
- Clustering - Combine Grid and census data to identify grids of similar attributes that can be used as market segments

Deliverables

- Code
- Report
- Slide Deck

Citations

- Barlacchi, G. *et al.* A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci. Data*2:150055 doi: 10.1038/sdata.2015.55 (2015).
- Telecom Italia, 2015, "Telecommunications - SMS, Call, Internet - MI", https://doi.org/10.7910/DVN/EGZHFV, Harvard Dataverse, V1
- Telecom Italia, 2015, "Milano Grid", https://doi.org/10.7910/DVN/QJWLFU, Harvard Dataverse, V1