BIG DATA ANALYTICS

PROJECT REPORT

MOBILE PHONE ACTIVITY IN A CITY

Lidia Tesfahiwet Kidane

Masters Student

Machine Intelligence

African Institute for Mathematical Sciences

25.05.2020

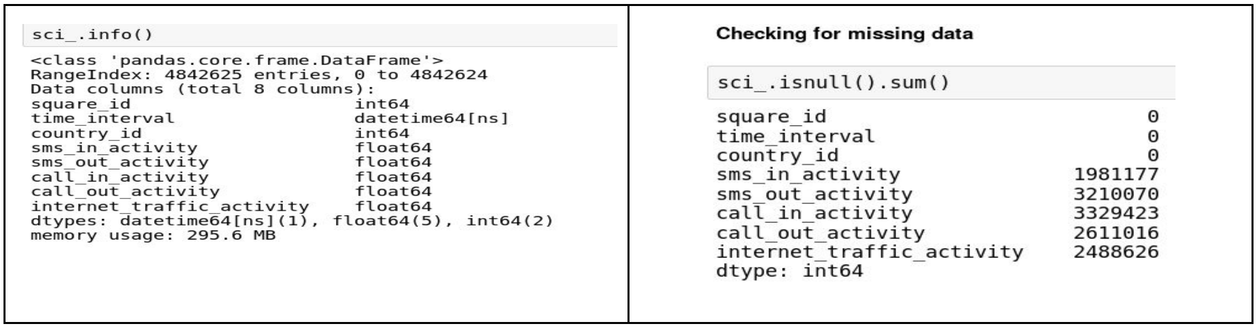
INTRODUCTION

A call detail record (CDR) is produced by a telecommunication equipment that provides information on telecommunication transactions and activities. This report contains an analysis on the telecommunication dataset from the Telecom Italia first edition of the Big Data Challenge contest which was designed to stimulate the creation and development of innovative technological ideas in the Big Data field. Due to our limitation on resources, we limited our analysis to the sms-call-internet-mi-2013-11-01.txt file. In this task, we are going to provide analysis by extracting some important information from the telecommunication data. Given that data, the purpose is not to infer any interaction for single users, but to analyse the variability of the volume of the activity with respect to time and space.

The information provides good insight on customer needs and overall provides a view of customers for different purposes. Here, we will discuss customers' activities with respect to time of the day and grids to generate meaningful insights on customer's behaviours as well as the network's functionality.

DATA Pre Processing

The dataset contains 5 million records of customer activities for 24 hours. The file contains CDR records from Telecom Italia cellular network. The CDR record provides information for call, internet and sms costumes' activities for 10,000 grids in the province of Milan and the country associated with those activities.



```
sci_.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4842625 entries, 0 to 4842624
Data columns (total 8 columns):
square_id                  int64
time_interval              datetime64[ns]
country_id                 int64
sms_in_activity            float64
sms_out_activity           float64
call_in_activity           float64
call_out_activity          float64
internet_traffic_activity  float64
dtypes: datetime64[ns](1), float64(5), int64(2)
memory usage: 295.6 MB
```

Checking for missing data

```
sci_.isnull().sum()

square_id                        0
time_interval                    0
country_id                       0
sms_in_activity            1981177
sms_out_activity           3210070
call_in_activity           3329423
call_out_activity          2611016
internet_traffic_activity  2488626
dtype: int64
```

Fig 1: Information about the data

As shown in Fig 1 above we have missing values from the columns in the dataset. Since this dataset shows all the telecommunications activity that took place within a particular time interval, the missing values indicate that the particular activity did not occur at that particular time instant. Thus, we replace all the missing values with zero for computation purpose

Analysis 1: Variability of volume of traffic with respect to time of day

Network Congestion occurs when a network is not able to adequately handle the traffic flowing through it. This is because there are more users on the network (Over-Subscription) during particular hours of the day (peak period). Moreover, it is also necessary to know which activity has the highest usage rate and contributes more to the network congestion.

To find the peak hours of the day we grouped the data based time_interval and calculated the Total activity. As we can see in Fig 2 below, the peak hours are 10, 11, 16, 17. However, as we are taking one day data this might change on analysis on different days. We can conclude that we have the network congestion decrease during night time relatively.

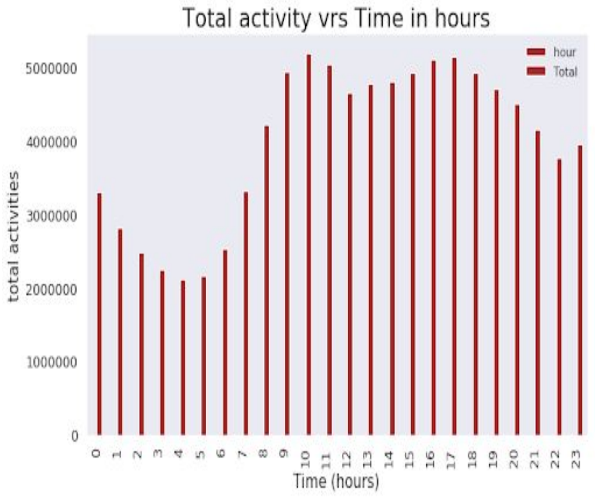


Fig 2: Total activities in 24 hours

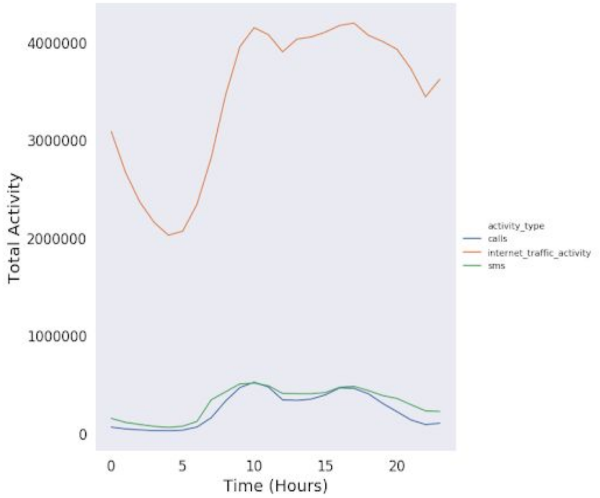


Fig 3: Distribution of activities per day

As shown in Fig 3, we have a distribution of internet traffic, calls and sms activities per hour of the day. We can observe that the most utilized service is internet service. In addition, we can observe that the three activity types have the same trend. Which indicates that clearly the customer’s usage depends on the hour of the day.

Analysis 2: Variability of the volume of the traffic with respect to space (grids)

Visualization in Fig 4, which is a bar chart depicts top 10 Countries by Total Activity. It shows France has the highest activity.

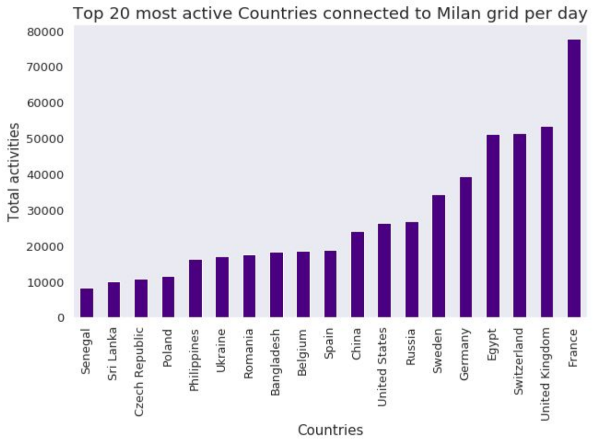


Fig 4: Top 20 Countries Connected to Milan

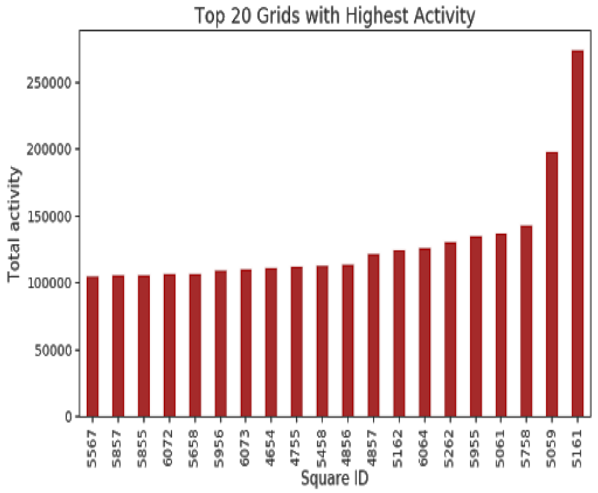


Fig 5: Top 20 Grids in Milan in terms of Activity

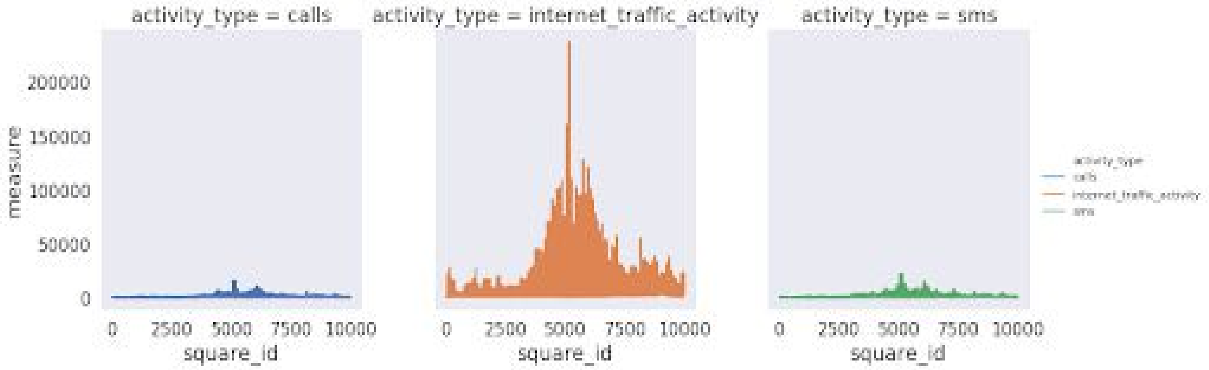


Fig 6: Distribution of Activities Per Grids

Here, we are going to infer variability of the volume of the traffic with respect to space. Our first bar plot in fig 5 investigates the top 20 grids in Milan with high activity rates throughout the day. We can observe some grids show high congestion throughout the day. This could be due to the location and population density. Moreover, we can observe the distribution of activities on space (grid). As Fig 6 indicates, internet service is a more utilized service across all the grids. However, it should be noted that the distribution of activities are the same across the grids.

Analysis 3: Clustering

We used the K-means algorithm to identify the k number of centroids, and then allocate every data point to the nearest cluster based on similarity. We utilized grid id (square id) and internet traffic activity to cluster the data points. It clustered the data points to 5 clustore. We can observe from Fig 8 our data points are clustered based on the Internet traffic activity. We can find clusters that are generating more internet traffic. This can help us target customers based on their geo location and usage pattern.

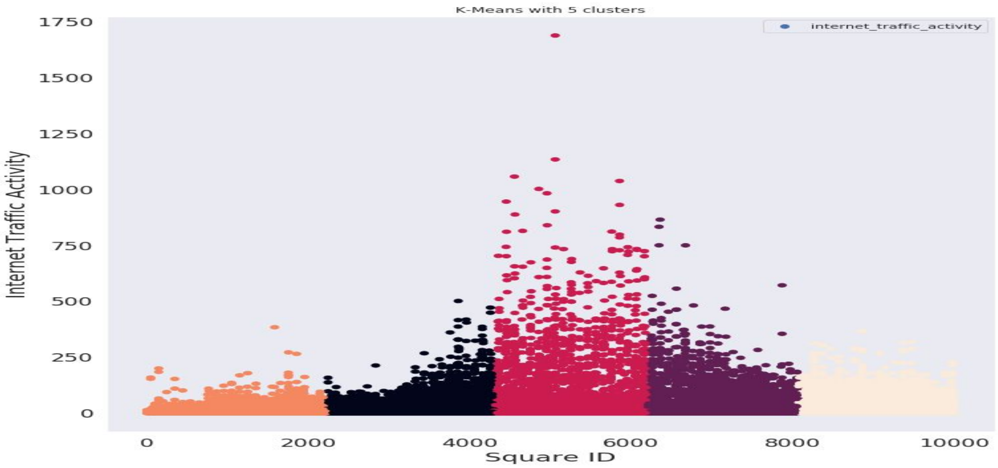


Fig 8: Clusters based on internet traffic and grid id

CONCLUSION

The dataset is very rich. However, due to resource constraint on day telecommunication data is utilized for this report. Using the visualizations, we can infer on the variability of the volume of the activities with respect to time and space. This report can be extended by adding weather and population dataset to add more dimensions and better analysis. Moreover, utilizing more datasets can give us better visualization of the trend.

REFERENCES

https://dataverse.harvard.edu/dataverse/bigdatachallenge?q=&types=datasets&sort=dateSort&order=desc&page=1

https://www.nature.com/articles/sdata201555?hash=152d20bc-ea8b-4a2e-af80-c4777b3c12b1

https://www.kaggle.com/marcodena/mobile-phone-activity/kernels