Forcasting Mobile Network Traffic

Prashant Kumar Ray

This report analyses 'Telecommunications Activity' dataset which has data recorded as measure of the level of interaction between the users and the mobile phone network. Specifically, it analyses 'Internet traffic activity' of the dataset which represents the number of CDRs (Call Detail Records) generated inside a given area (square Id) during a given time. After analysis of data, a machine learning model is proposed which tries to forecast the future internet traffic.

1 Hardware description

Google Colab platform is used for the data analysis and model building. Google Colab is a free Jupyter notebook environment running wholly in the cloud. The Google Colab usage Intel(R) Xeon(R) CPU @ 2.20GHz. The available RAM size is around 12 GB.

2 Environment setup

Following setup is done for this project:

- PySpark framework is used for data analysis. Therefore, PySpark library is installed in the Google Colab environment.
- Google Drive is used to load the 'Telecommunications activity' dataset for the analysis and model creation. Therefore, Google drive is mounted in the Google Colab environment to load the dataset.

3 Datset description

The 'Telecommunication Activity' dataset is used for this project as mentioned in the Homework.pdf file. The dataset can be downloaded from the link already given in the same file. The dataset has 62 text files which contain the information of the level of interaction between the users and the mobile phone over the period of two months (Nov 2013, Dec 2013). The total size of 'Telecommunication Activity' dataset is 19.4 GB. The dataset has 8 columns(1) which are listed below in order:

1. Square Id
2. Time Interval
3. Country code
4. SMS-in activity
5. SMS-out activity
6. Call-in activity
7. Call-out activity
8. Internet traffic activity

'Square id' is the identification string of a given square of Milan/Trentino GRID. 'Time Interval' represents the time (in milliseconds) at which the data was recorded. 'Internet traffic activity' represents number of CDRs generated inside a given 'Square id' during a given 'Time Interval'. Rest all fields are not relevant for this analysis as already mentioned in the Homework.pdf file.

4 Data preprocessing

The dataset is imported using PySpark framework for the cleaning, analysis and filtering the data of our interest. As already mentioned in the Homework.pdf file, only the three fields 'Square id', 'Time Interval' and 'Internet traffic activity' are relevant to the homework, other fields are removed. Default column names are renamed with appropriate names and 'Internet traffic activity' column is casted from string type to double type. We can observe null values present in the 'Internet traffic activity' column. These null values are replaced with 0.

5 Exploratory Data Analysis

- In the analysis, it is found that the area with square id 5161 has observed the most internet traffic during the given two months duration.
- Due to hardware constraint, I could not plot the PDF plot for the overall internet traffic in the city for the given two month period as asked in Task 1.2. Therefore, I have plotted the PDF for the region with square ids 5161 as shown in fig 1. The PDF plot indicates that the value for internet traffic activity in square id 5161 varies between 0 to 8000. However, probability of finding a data point in lower range is higher in comparison to higher ranges.
- The Box plot for the square ids 5161, 4556, 4159 are shown in fig 2. For square id 5161, it is observed that 75% of internet traffic has value less than 2500 approximately. Any value above 6000 might be considered as outliers. Similar analysis can be done for the rest two square ids using their box plots.
- Fig 3 shows the internet traffic patterns in the three square ids during first two weeks of Nov 2013. It can be observed that the internet traffic is lower than usual in square id 4159 on 02 Nov, 03 Nov, 09 Nov and 10 Nov in year 2013. These are the dates which are weekends. The possible reason for such internet traffic pattern can be that the square id 4159 can be an industrial or business area where many working professionals works during the weekdays. During the weekends, the traffic caused by these working professionals is absent. Hence, the less traffic is observed during the weekend as compared to weekdays. While in square id 5161, higher amount of internet traffic is observed during the weekends in comparison to weekdays.
- Fig 4 shows the internet traffic pattern for a given date. Here we can observe that the peak network hours are different in each square ids. In square id 5161, the peak traffic is found to be around 12:00 PM while in square id 4556, it is around 9:00 PM. In square id 4159, internet traffic seems to be almost uniform during 9:00 AM to 5:00 PM.
  Also the amount of traffic is different for every square id which may be considered as approximate population ratio of given two square ids. Square id 5161 is possibly resides in the city centre while square id 4159 may be situated in the outskirts of the city.

6 Model creation

6.1 Algorithm

In this project, Random Forest Regression algorithm is used for the forecasting of mobile network traffic hourly. Random Forest is a supervised learning algorithm that is based on the ensemble learning method. It can be used to solve both Classification and Regression tasks. The name "Random Forest" comes from the Bagging idea of data randomization (Random) and building multiple Decision Trees (Forest).
Like most ML methods, Random Forest have no awareness of time. On the contrary, they take observations to be independent and identically distributed. This assumption is violated in time series data which is characterized by serial dependence. However, Random Forest algorithm can be used in time series forecasting, both univariate and multivariate dataset by creating lag variables and seasonal component variables manually.
The dataset has the internet traffic information available for every 10 minutes. The following steps are performed to prepare the dataset that is used for the Random Forest Regression model training for hourly forecasting:

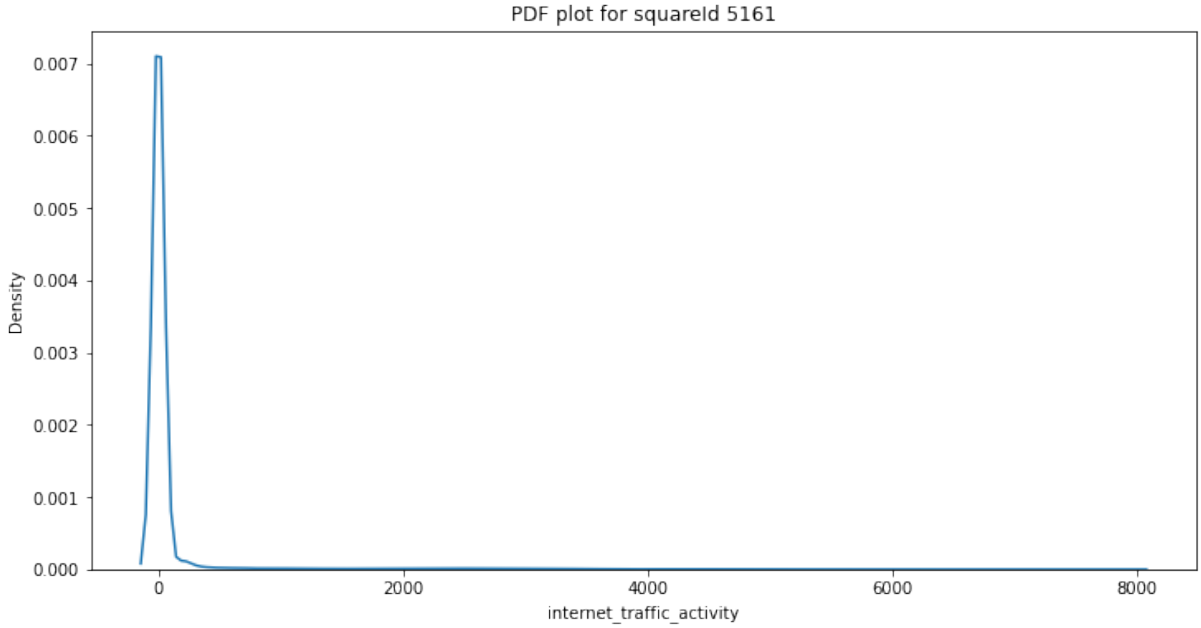

Figure 1: PDF plot of internet traffic over the two month (Nov2013-Dec2013) in square id 5161

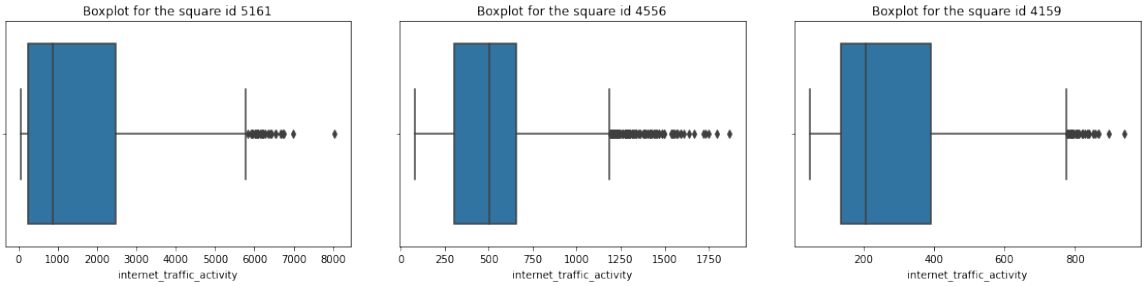

Figure 2: "Boxplots for square id 5161, 4556 and 4159

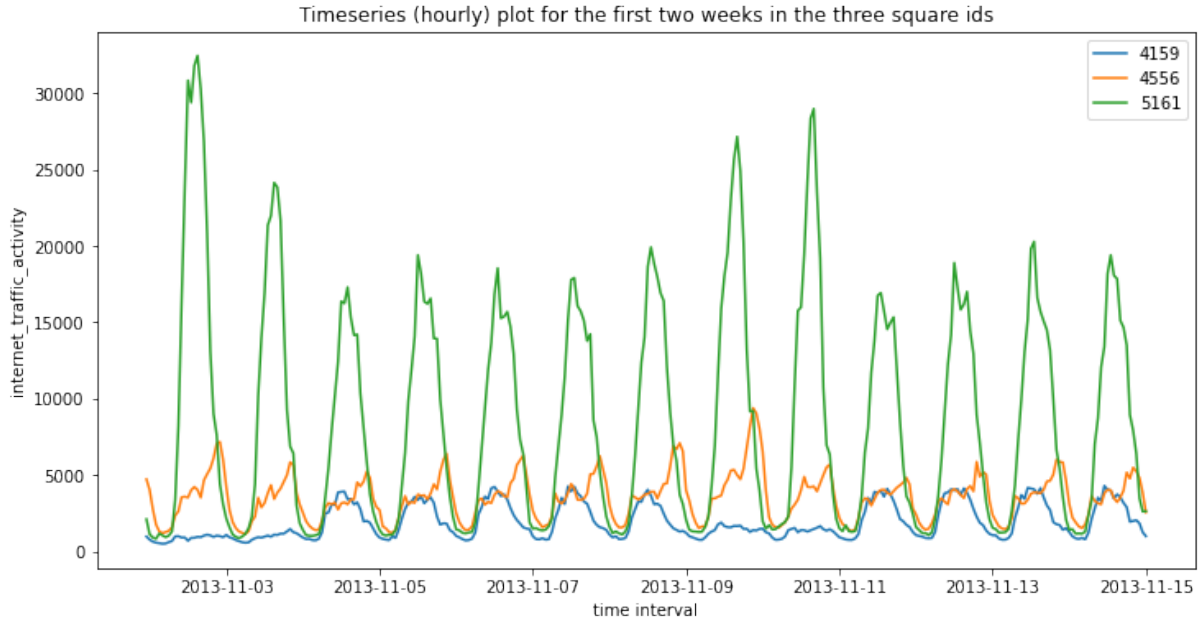

Figure 3: Timeseries plot of internet traffic over the first two weeks of Nov 2013

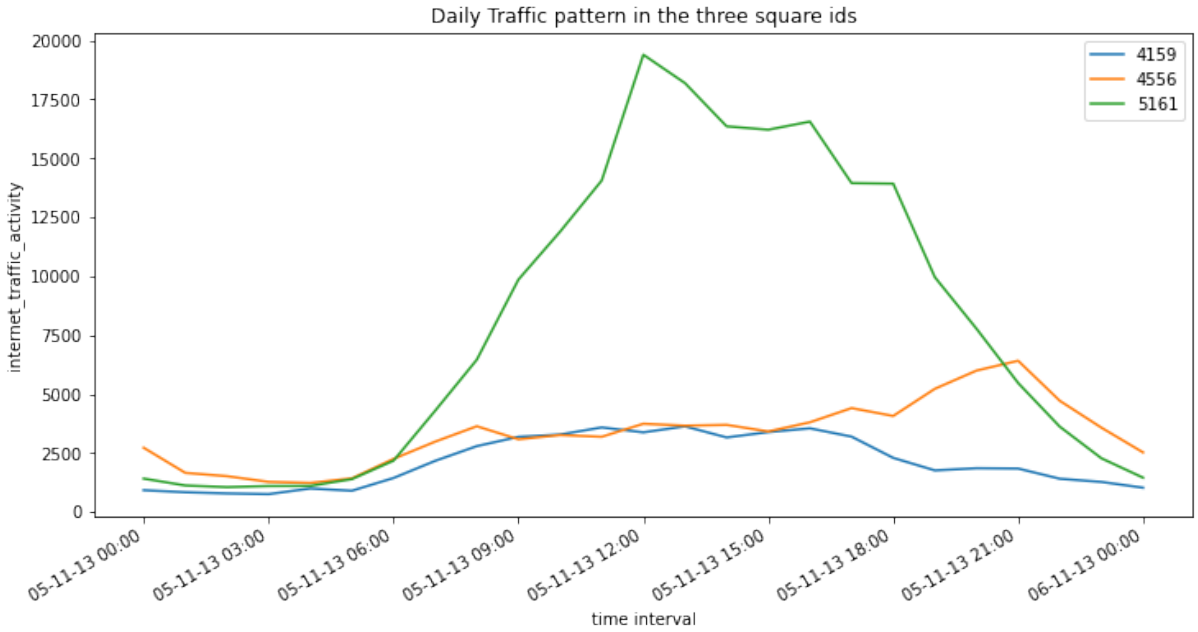

Figure 4: Timeseries plot of internet traffic pattern in a day in the square id 5161, 4556 and 4159

1. The data is resampled hourly. Originally it is available for every 10 minutes interval.
2. 24 lag variables are created as data pattern repeats at 24 hours.
3. A column for day of week is added as it has an effect on network traffic for example: square id 4159.

So, if we denote $x_a(t)$ as the traffic observed at area 'a' during time interval 't' and $\tilde{x_a}(t+1)$ is the estimate of future traffic at time (t + 1) in area 'a' then,

$$\tilde{x_a}(t+1) = F(x_a(t), x_a(t-1), x_a(t-2), x_a(t-3), ..., x_a(t-23))$$

6.2 Model evaluation

Fig [5,6,7] shows the actual and predicted network traffic timeseries plot for square ids 5161, 4556 and 4159 respectively. For the performance evaluation of the model, mean absolute error (MAE) and mean absolute percentage error (MAPE) are used which are listed in the Table 1.

Square id	MAE	MAPE
5161	1086.98	21.46
4556	405.67	18.06
4159	195.98	15.84

Table 1: Network traffic forecasting model performances in different square ids

6.3 Training and execution time statistics

Training time is the time taken by a model to train on a dataset, and the execution time represents the total time taken for computations, including data splitting, data preprocessing, and model evaluation. The training and execution times for the network traffic forecasting model is listed in the table 2.

7 Conclusion and possible improvements

In this report, 'Telecommunication activity' dataset is analysed in three different areas of the city. The amount of traffic and peak traffic hours is different for these areas. Effect of Weekends on network traffic

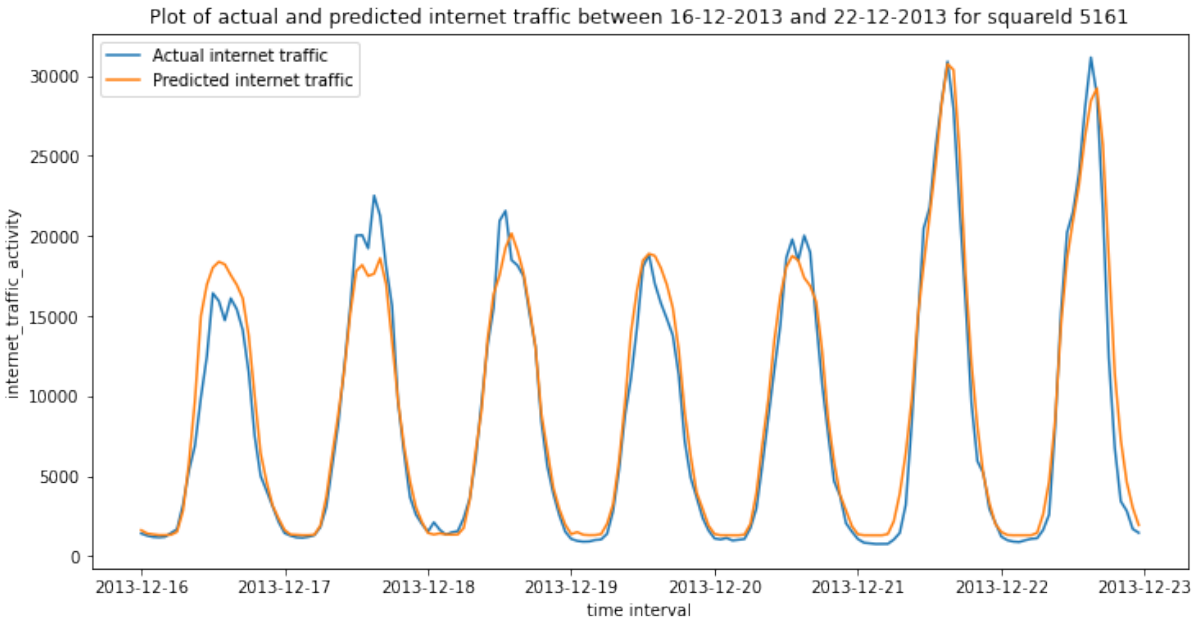

Figure 5: Actual and Predicted timeseries plot of internet traffic in square id 5161 from 16 Dec-22 Dec 2013

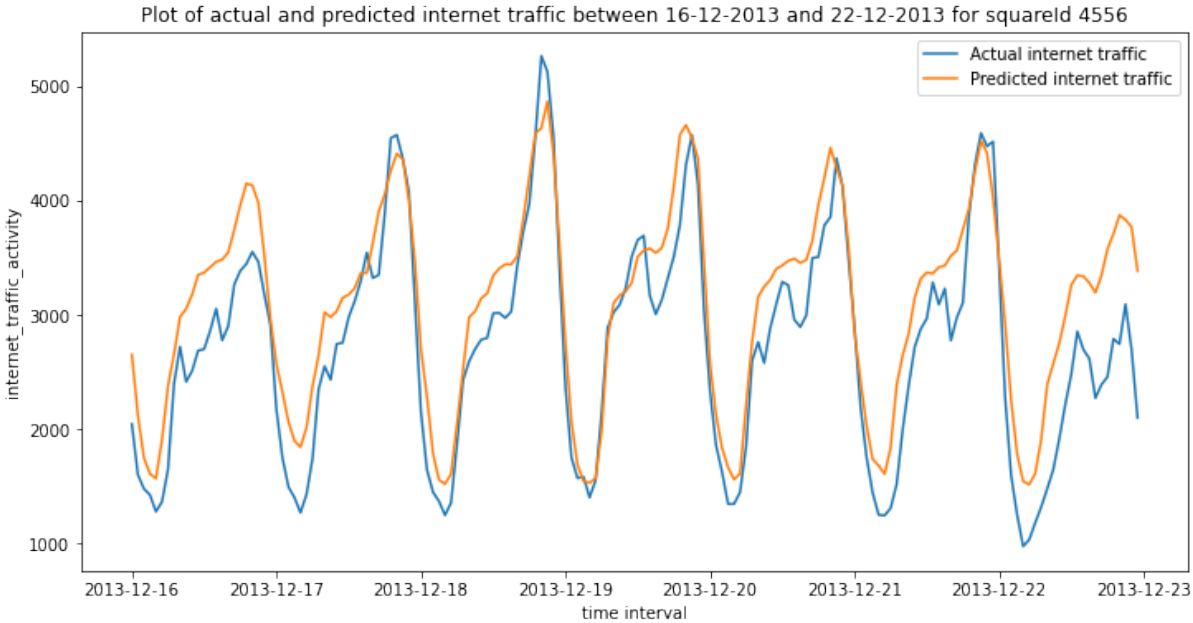

Figure 6: Actual and Predicted timeseries plot of internet traffic in square id 4556 from 16 Dec-22 Dec 2013

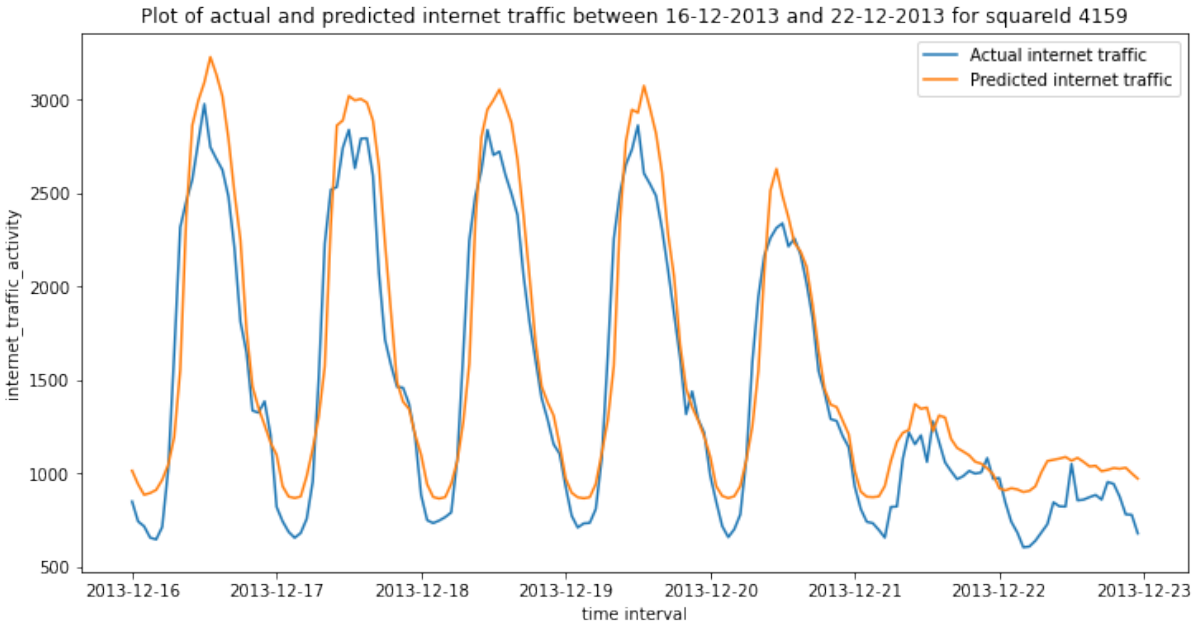

Figure 7: Actual and Predicted timeseries plot of internet traffic in square id 4159 from 16 Dec-22 Dec 2013

Square id	Training time (seconds)	Execution time (seconds)
5161	71.72	71.82
4556	75.51	75.58
4159	69.63	69.69

Table 2: Network traffic forecasting model training and execution statistics in different square ids

is also observed in some areas. Random Forest model seems to perform well in forecasting the future internet traffic in the three square ids in the city.
Outliers(unusual increase in network traffic) are not taken care of in this project. The outlier value may be replaced with the average value of the nearby data samples. This might improve the model performance.

References

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2(1):1–15, 2015.