EAI Endorsed Transactions
on Smart Cities

Research Article EAI.EU

Big Data in Telecom Industry: Effective Predictive Techniques on CDRs

Sara ElElimy and Samir Moustafa*

Computational and Data Science and Engineering, Skolkovo Institute of Science and Technology

Abstract

Mobile network operators start to face many challenges in the digital era, especially with high demands from customers. Since the mobile network operators have considered a source of big data traditional techniques are not effective with new era big data, internet of things (IoT) and 5G, as a result handling effectively different big datasets becomes a vital task for operators with the continuous growth of data and moving from long term evolution(LTE) to 5G therefore, there is an urgent need for sufficient big data analytic to predict future demands, traffic, and network performance to fulfill the requirements of the fifth generation of mobile network technology. In this paper, we introduce data science techniques using machine learning and deep learning algorithms: the auto-regressive integrated moving average(ARIMA) Bayesian-based curve fitting, and recurrent neural network(RNN) is employed for a data-driven application to mobile network operators. The main framework included in models is an identification parameter of each model, estimation, prediction, and final data-driven application of this prediction from business and network performance applications. These models are applied to Telecom Italian Big Data challenge call detail records (CDRs) datasets. The performance of these models is found out using specific well-known evaluation criteria that show that ARIMA (machine learning-based model) is more accurate as a predictive model in such a dataset as the RNN (deep learning model).

Received on 20 April 2020; accepted on 29 May 2020; published on 04 June 2020

Keywords: Big Data Analytics, Machine Learning, CDRs, 5G.

doi:10.4108/eai.13-7-2018.164919

1. Introduction

Operators of mobile networks began to move to the fifth generation from the fourth generation, which is an upcoming and promising solution for meeting the requirements of wireless broadband. Additionally, they have started looking for some innovative solutions for facing challenges and providing a satiable customer experience with the management of the complex network by efficient backhaul resource handing [1]. Telecom organizations and researchers have been studying a diversity of techniques for big data management adequately for discovering unknown knowledge and patterns from the collected information obtained from operators and help organizations in providing smart services for achieving reduced expenditure and resources.

[*] Corresponding author. Email: samir.mohamed@skoltech.ru

With the fast uptake in mobile applications and services, requesting demands for infrastructures in wireless network. For 5G requirements and KPIs are to support exploding in mobile traffic, provide low latency so this raised need for real-time decision and network resources management and optimization to maximize and increase customer satisfaction and enhance user experience. Using traditions methods to achieve these requirements and overcome different problems become a challenge to telecoms.

Tradition techniques start to be useless in this area so industry and academia start to search and create more effective new techniques to deal with this tremendously increase of data and raise the question of how the telecoms deal with:

1. Enormous data sizes (Various systems generated a huge amount of log data and reached Giga-Tera byte).

2. Different sources (generated from different sources e.g., routers, switches, applications, operating systems, etc.).

3. Heterogeneity (Different format, structures, terms of terminology, etc.).

These questions and challenges are the main problems statement for this work, and how telecoms benefit from applying ML/DL on different datasets, and what kind of application can be achieved using these techniques that are exiting and traditional ones.

In this paper, we are investigating the analysis and application-driven by big data in the telecommunication industry concerning operators of mobile networks for the fifth generation and current networks in their operational and business aspects, implementing different ML/DL techniques driven by big data on data gathered from a telecommunication network and applying different models of prediction for predicting traffic. Moreover, in the end, how different results and applications are brought by big data analytic in comparison with traditional methods. Also, it will be discussed how they are beneficial for business and operational activities, companies, and how this can be utilized and in which types of applications.

2. Analytic Tools and Data Sources for Telecoms

2.1. Telecom Data Sources

Operators of mobile networks form a source and carrier of big data because of the penetration of mobile users have increased significantly [2], and organizations utilized traditional techniques before transactions from the analytic of big data. These techniques pay less attention to operational data, and they do not concentrate significantly on transnational data. The analytic of big data is essential in several ways in comparison with traditional methods. For instance, the compressor transmits data, and useful data are defined by the analytic of big data [3]. In large part of an application, decision-making in real-time is a benefit of using analytic of big data by monitoring the development and infrastructure of network performance. Several smart services will be supported and provided by MNOs with the analysis of sources and types of data [4].

Classifies sources of data for Telecoms as operator and subscriber data, external and internal data sources [3], core network levels, cell, subscriber, and KPI deep classification for different networks [5]. When it comes to analytic tools, some of the main tools, as defined by the previous studies, include methods of machine learning modeling, data mining, and statistical modeling [6]. Actually, with current development and improvement in data analytic, networks based on big data have formed an attractive area of research for numerous researchers around the globe [7], [8]. Additionally, in the industrial sector, researchers recently developed and studied frameworks for big data management in an efficient manner in mobile networks.

2.2. Contribution of CDRs or Call Details Records

In mobile operators, CDRs were considered essential in for finical aspects. However, in the period of big data, applications driven by it are obtaining attention by researchers in industrial and scientific aspects because datasets of CDRs are full of information associated with communication among numerous users along with how, when, and with whom they are communicating.

The analysis of CDRs datasets has become quite a significant and exciting research area [9] because numerous uses associated with these datasets provided by it for different purposes of research resulting in the improvement of dataset management techniques, development of analytic techniques, and analysis types from several perspectives with the use of big-data methods. When it comes to telecom operators, Orange is recognized as one of the biggest, and the first challenge, "D4D Challenge" was launched in 2013. They invited different candidates through this challenge from around the globe.addition to it, and access was provided to massive datasets of CDRs for developing objectives of their customer satisfaction and infrastructures as a source of gaining more revenues. Successful outcomes have resulted in scientific work, which encouraged the organization to launch a second challenge during the mobile conference of NET in April 2015 [9]. In Europe, Telecom Italian is also a recognized mobile operator that faces the same challenges of big data, and2014, Big Data Challenge's first edition was launched by its [10].

3. Techniques and Methodology

In the analysis of these datasets, different techniques and methods are utilized. Some of the techniques utilized in this work include data visualization, prediction, and clustering. We followed the framework for obtaining the optimum outcomes from datasets.

Pre-processing is the first step, and it is considered an essential step while using massive data, and in understanding the hidden patterns existing in the data. The next step is concerned with defining analysis type and necessary tools for it, the application type is driven by it, and which type of information might be needed for it. Finally, based on the results, the best applications are determined for this analysis.

3.1. Data Set

Millions of records are included in a dataset between December and November 2013. In 2014, these datasets were a component of the Big Data Challenge of Telecom Italian. It was quite ironic and included different types of telecommunications, including electricity data, weather forecasting, news, and social networking. Telecom Italian has formed an original dataset with the connotation of some specific labs. The institutes included in them are:

- Fondazione Bruno Kessler.
- EIT ICT Labs.
- Trento and Trento RISE Institute.
- Milan Polytechnic University.
- MIT Media Labs.

Before the first dataset is released, the attention of partakers is considered. The demand is nevertheless being increased at the competition's end for datasets, which has become an initiative or measure towards "Open Big Data." Datasets, following [10], were freely published for improving the dataset used in the society.

Telecom Italian generated a dataset that is a consequence of evaluation or calculation upon the call detail records for subscribers of Milano City. CDRs record user activities for billing and network management, but our research focuses on the use of dataset for different applications rather than utilizing it for traditional activity.

Information included in dataset described in [10], it consists of main eight variables:

- Square ID: the Square ID, which is the portion of Milan GRID.
- Time Interval: The start of the time interval can be stated as the number of milliseconds passed till 1st January 1970 from the Unix Epoch at UTC. In addition, of 10 minutes (600000 milliseconds) to this value, the time interval can be achieved.
- Country Code: It is the local code of a country for phones.
- SMS-in Activity: The SMS activity is receiving the inside square ID throughout the time interval
- SMS-out Activity: The SMS activity is sending the inside square ID throughout the time interval.
- Call-in Activity: The Calls activity is receiving the inside square ID throughout the time interval.
- Calls-out Activity: The SMS activity is issuing the inside square ID throughout the time interval.

- Internet Traffic Activity: The Internet Traffic activity is issuing the inside square ID throughout the time interval and by the state of the user all these activities are recognized from the country code.

We have a few types of Call Detail Records for generating the datasets which are related to these activities:

Before the first data-set is released, the attention of partakers is considered. The demand is nevertheless being increased at the competition's end for data-sets, which has become an initiative or measure towards "Open Big Data." Datasets, following [10], were freely published for improving the dataset used in the society.

Information included in dataset described in [10], it consists of main eight variables:

- Received SMS: Every time when a user receives an SMS.
- Sent SMS: Every time when a user sent an SMS.
- Incoming Call: Every time when a user receives a call.
- Outgoing Call: Every time when a user issued a call.
- Internet: Every time when a user starts or end an internet connection.

Throughout the similar internet connection one of the below restrictions is reached :

- 15 Minutes after producing the final CDR
- 5 MB after producing the final CDR

This Data-set was formed by accumulating the above stated records, to deliver Internet Traffic, SMSs and Calls activities. The level of collaboration between users and mobile network is calculated through this. For instance, more SMS sending by a user results in more activity of the SMSs sent by the user. The SMSs and Calls activities are having the similar scale of sizes “Therefor they are analogous to each other”. According to (Data Telecom, 2014), Data-sets are combined in four-sided cells gird, as shown in Figure 1.

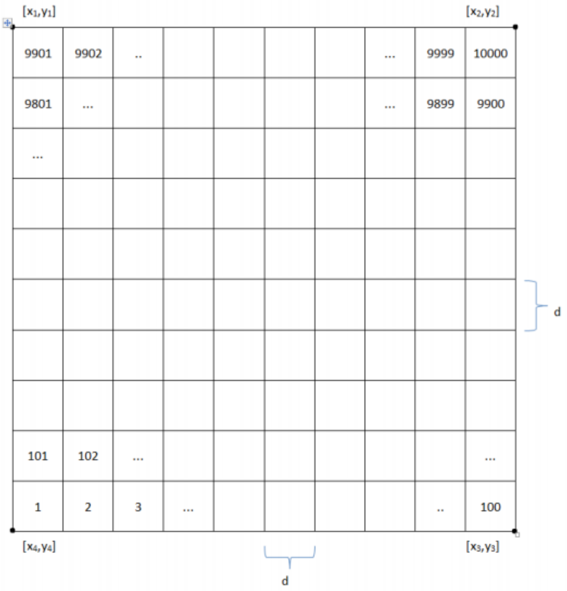

Figure 1. “The area of a Milan is composed of a grid overlay, which is 1,000 squares having the size of 235*235 meters.” The grid is probable with WGS84 (EPSG:4326) standard

3.2. Methods and Models

In these sections, the adopted methods are explained:

- Data visualization: using the right type of visualization brings insight into the data analysis process. Explanatory Data Analysis(EDA) executed in a proper order to study and expound the dataset. The aim of conducted data analysis, to discover the restriction of data, data patterns, and which unavailable or missing variables.
- Clustering: Clustering procedures, in the data mining field, constitute some important methods [11] due to their significant-high abilities for deducing connections among different data objects.

  Scientists have primarily utilized them for investigating datasets for the tracing of mobile. On different networks acquired from mobile networks, K-means is implemented the most, and in other works, including [12] and [13], it provides satisfactory results.

  The techniques of clustering are accepted, either a separated approach or hierarchical approaches. Hierarchical techniques arrange items into a

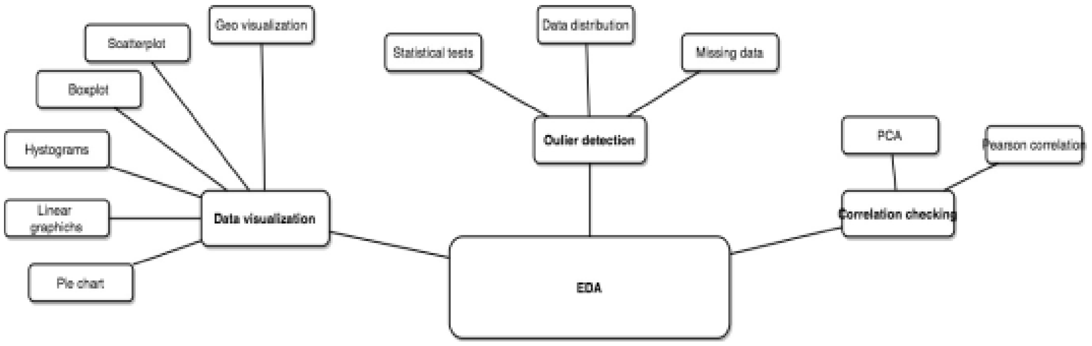

Figure 2. Diagrams to show the Explanatory Data Analysis(EDA) relations.

hierarchical structure, which can visually be represented diagrammatically.

Hierarchical algorithms can follow an organized method or separated one. However, partitioned clustering algorithms e.g., ISODATA and K-means, directly group objects into numbers of categories K.A relevant comment is that hierarchical algorithms can also be used in categorizing objects into a definite number of categories, which can be finished by ending the algorithm at the required point/level. In all instances, there is no stipulated rule to determine the definite number of categories, the decision still remains either ascertained definitely relying on the accordance to certain clustering quality measures or knowledge about the data. inner-cluster distances.

- Standardization: Standardizing a vector most often means subtracting a measure of location and dividing by a measure of scale. For example, if the vector contains random values with a Gaussian distribution, you might subtract the mean and divide by the standard deviation, thereby obtaining a "standard normal" random variable with mean 0 and standard deviation 1, So standardizing the internet traffic before modeling will help in prediction.

Table 1. Show the ARIMA model parameters.

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with constant
Auto-regression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

- Prediction: For mobile operators, it is considered necessary in making decisions associated with network optimization, and as a part of ML. ARIMA model is one of the most renowned algorithms of prediction, as explained in [14]. It is significant for time series data in both static and practical manner.

$$y_t = C + \sum_{i=1}^{p} \varphi_i y_{t-i} + \epsilon_i \tag{1}$$

The following are special models from ARIMA:

y_{t-i} and ϵ_i are respectively the actual value and the random error at the time t, $\varphi_i (i = 1, 2, 3, \ldots, p)$ are the model parameter and is a constant, the integer is known as the order of the model [15].

RNN model is another adopted model, model with many layers on the basis of short and long-term memory is referred to as LSTM. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell [16]. It consists

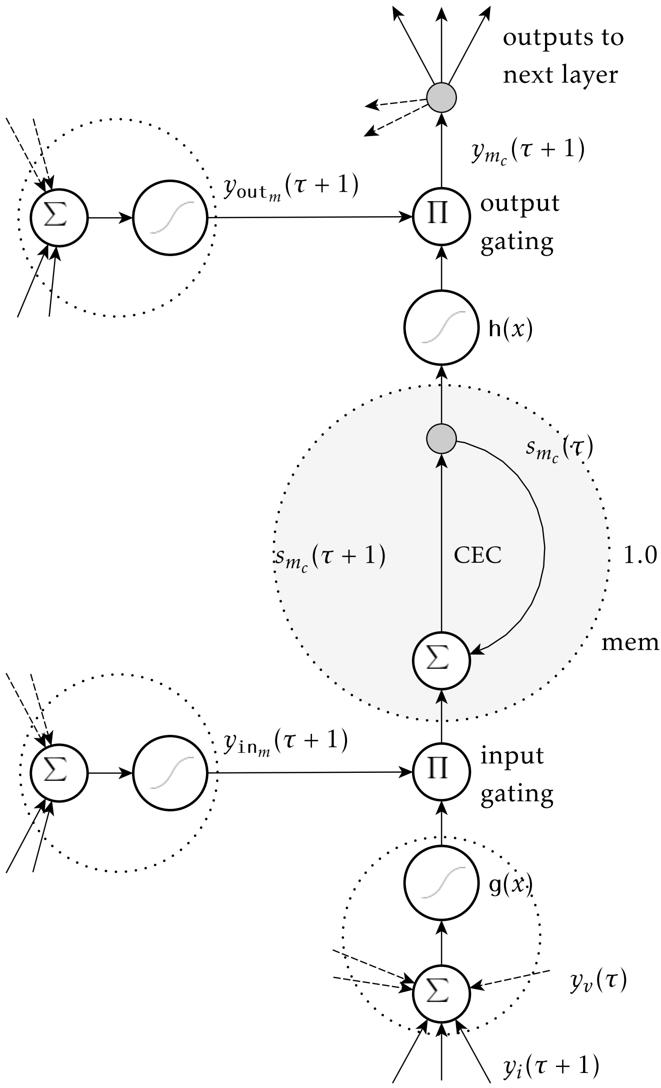

Figure 3. A standard Term Short Memory (LSTM) memory block, and the cell output is calculated by multiplying the cell state by the activation of the output gate.

of memory blocks, and it can be trained with the use of black propagation. In this model, the issue of the gradient is gradually decreased [17].

$$f_t = \sigma(X_t * U_f + H_{t-1} * W_f) \tag{2}$$

$$\overline{C_t} = \tanh(X_t * U_c + H_{t-1} * W_c) \tag{3}$$

$$I_t = \sigma(X_t * U_i + H_{t-1} * W_i) \tag{4}$$

$$O_t = \sigma(X_t * U_o + H_{t-1} * W_o) \tag{5}$$

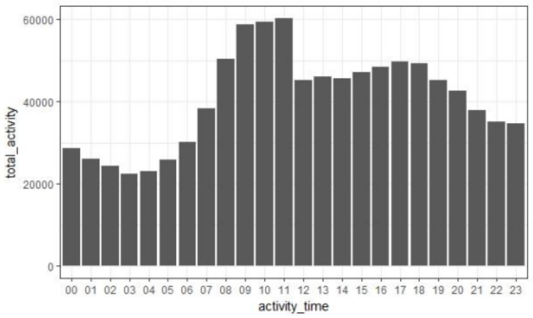

Figure 4. Daily activity.

$$C_t = f_t * C_{t-1} + I_t * \overline{C_t} \tag{6}$$

$$H_t = O_t * \tanh(C_t) \tag{7}$$

X_t = Input Vector , H_{t-1} = Previous Cell output C_{t-1} = Previous Cell Memory, H_t= Current Cell output , C_t = Current Cell Memory. W, U = weight vectors for forget gate (f_t), candidate (C),i/p gate (I) and $o/p(O)$ [18].

Both ARIMA and RNN are performed in a better manner in comparison with others for time series prediction [19].

3.3. Analysis of Data and Prediction Process

Generally, the base of our analysis is the data-intensive approach, and different techniques of machine learning are applied on datasets of CDRs because it contributes to the value of both business and scientific aspects. Three analyses have been performed in our work:

First analysis : The highest daily activity is identified in this analysis during a specific day. In addition to it, peak hours within a day are also identified. The first analysis's results were derived concerning total and time activity, while peak hours are 11, 10, and 9 AM, while 3 AM is not a peak activity hour.

In business aspects and network development, this result is quite beneficial because it will aid in the identification of which areas needs to be developed or requires more resources. It will also help in determined which country code or square grid develops more traffic due to which companies gain more revenues by targeting customers based on their geo-location. Additionally, with resource management, it decreases its costs and expenses.

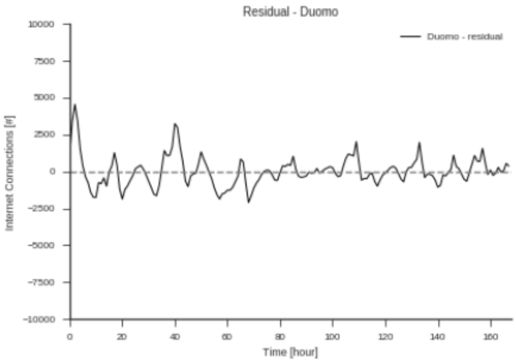

Figure 5. Analysis of Residuals

Second analysis: This analysis compares and illustrates the weekly usage of the internet in November for three ID cells portraying different areas for categories in the city of Milan. It also included nightlife area, university area, and downtown area. It was indicated by the results that the downtown area's peak is earlier than that of nightlife, phone calls are less in universities area on the weekends, and a decrease was experienced in the volume of calls.

In optimization and resource allocation, these observations will help by defining which area is fully loaded and at what time, and it can help in defining temporary solutions for different peak hours, such as the deployment of Pico cell.

Certain tests were carried out on the dataset to identify and select the proper and effective models for time series data. It is essential to discover trends, seasonality, and stationary of data.

Residuals analysis provides an indication if data is statistically stationary if the data is truly random noise, it can be classified as statistically stationary from Figure 5.

Another testing method is the Dickey-Fuller stationary test, which is a quantitative test for residuals analysis; its Null hypotheses represent that residual is not statically stationary.

Findings and results showed that the test statics is about -7, confirmed that residuals are statistically stationary.

Third analysis: In this analysis, three methods are implemented for prediction and modeling based on internet usage. ARIMA model is the first one, LSTM is the second model, and the last model is developed on the model which was utilized in the Kaggle Competition. This model was validated on different types of data weekly for determining if modeling for a week is efficient enough for having similar results and whether it can be implemented on datasets that are collected at different time intervals.

Table 2. Statistical Tests to show Dickey-Fuller stationary test.

Result of Dickey-Fuller Test:	
Test Statistic	-7.405407e+00
p-value	7.367220e-11
# Lags Used	1.000000e+00
Number of Observation Used	1.660000e+02
Critical Values(1%)	-3.470370e+00
Critical Values(5%)	-2.879114e+00
Critical Values(10%)	-2.576139e+00

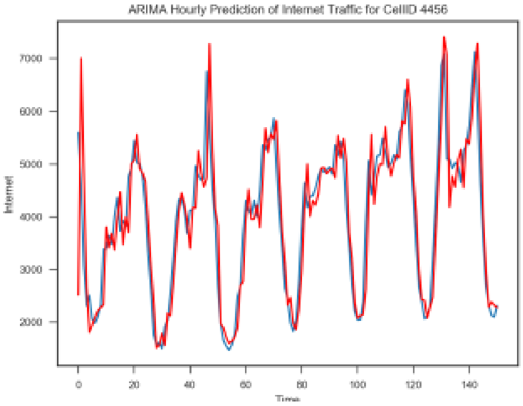

Figure 6. ARIMA Hourly Prediction of Internet Traffic for Cell ID 4456

- ARIMA
  For the datasets of one week, the applied model is ARIMA (2, 1, 0). Three ID cells will be focused upon first for the central regions, and the obtained results are portrayed in the Figures 6 and 7.

  Moving on, 9998 cells were the target, as illustrated in Figure 8.

- LSTM
  One input is included in this model for four blocks and a visible layer in the hidden layer. Meanwhile, in the output layer, there is a single input. Internet traffic prediction is shown in Figure 9 for 4456 cell ID every week.

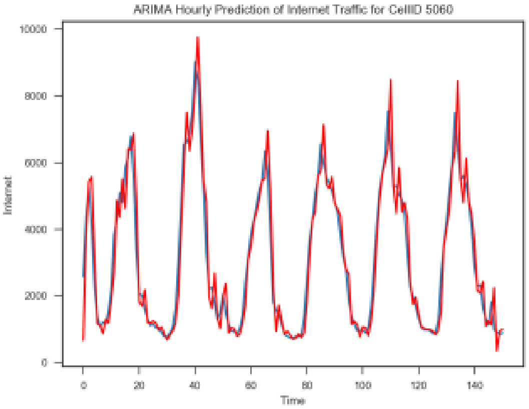

Figure 7. ARIMA Hourly Prediction of Internet Traffic for Cell ID 5060.

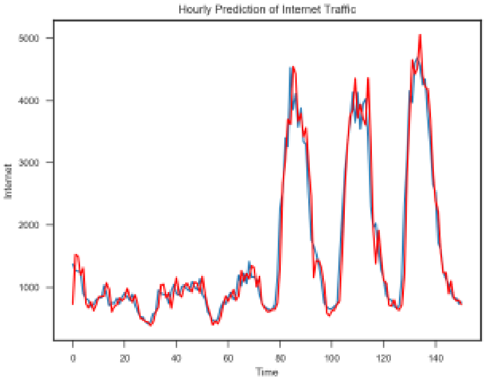

Figure 8. For all cells, Internet Traffic Hourly Prediction using ARIMA

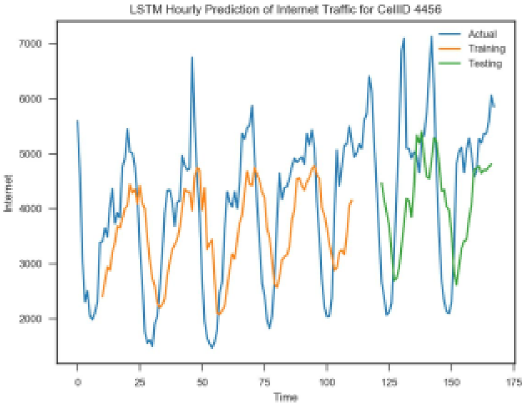

Figure 9. For 4456 cell ID, Internet Traffic Hourly Prediction using ARIMA

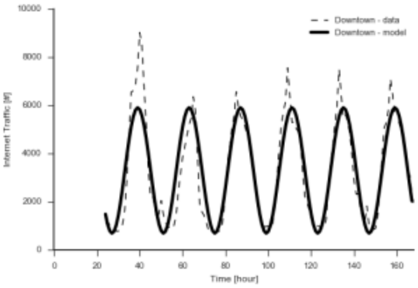

Figure 10. Downtown Area Results of Internet Traffic

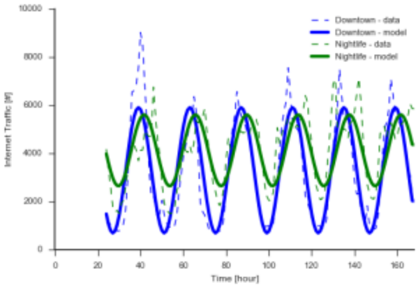

Figure 11. Nightlife and Downtown Areas and Internet Traffic Data

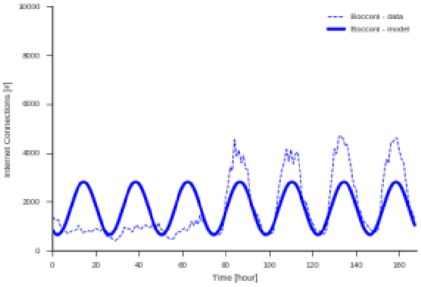

Figure 12. Universities Area and Internet Traffic Data

- Third Prediction Model
  In the Kaggle competition, this model was utilized where it was implemented on several periods in contrast without information. Generally, it is based on many datasets which are periodically set every twenty-four hours. Meanwhile, SIN behavior is exhibited by internet traffic, as portrayed in Figure 10.

  Moving on, this model is implemented in three areas, which are categorized from our analysis. Prediction results for nightlife and downtown are represented in Figure 11 for the area of universities in Figure 12.

Three models were applied for the prediction internet traffic based on hourly and weekly data Results explained that the prediction model of ARIMA is precise for the selected cells and with a 3 percent test set and 70 percent data set. It recognized that 21 percent of test sets and 69 percent training sets were not sufficient enough in cell/data ID. The obtained results, for the third model, it was indicated by the obtained results that this model is accurate and suitable for all the selected datasets with the university area being an exception. This area still has some issues, and it might be associated with the mobility of community patterns. The same conclusion as previous works was obtained for different dataset periods. Thus, it was determined that this model was suitable for all datasets.

Results have indicated that the application of predictive models and intelligent data analysis for the prediction of traffic are considered significant, and they play a vital role for mobile operators, which will be quite useful in the routing of traffic. It can indicate yearly prediction as well for supporting network optimization, resource allocations, self-organizing networks, and investment planning.

4. DISCUSSION

For MNOs, this research is dedicated to big data management and applying ML/DL techniques in an efficient manner in the sector of data-driven apps and the telecommunication sector. Comprehending the available data, which analytic tools are eligible and must be implemented, and which type of information or data should be collected are significant for any provider of service for harvesting the best results from the data. Big data is selected and applied in this work, and t is vital to recognize that techniques of machine earning and deep learning contribute significantly to both the industrial and academic sector and playing a significant role in wireless network application like network traffic prediction using different clustering techniques, it is possible to cluster mobile users based on CDR records and generate location-based recommendation system.

CDRs mining using these techniques then existing one expands its role and applications not only for finical usage, but also by extracting huge and important knowledge from this dataset introduces different application for telecoms:

1. Analyzing CDRs data can be provided demographic about genders and age where we can use RNN or CNN to predict these features of mobile users.
2. RNNs are employed to determine the metro density from massive CDRs data, they propose to identify the trajectory of the customer as a sequence of locations as input to RNN- model to handle this sequential data.
3. From code number information in CDRs, it is possible to predict tourist's locations and make business packages.

It has been proven by this practical work how benefits in the business and operational aspect of the telecommunication industry can be obtained with the effective application of techniques of Big Data instead of traditional techniques. Models like LSTM and ARIMA was applied for the prediction of traffic, and it was explained that results were quite beneficial in strategic and short plans for the operator. For the performance of our practical part, CDR database selection was based on the significance of the dataset for the MNO since it is indicated by our results that CDRs analysis has much significance beyond and currently in different areas like investment plans on the basis of optimization network, fault detection traffic prediction, network optimization, and resource allocation.

For future work, we will apply ML/DL techniques on different unlabeled datasets since mos-generated data in wireless network systems have these challenge able features, which required specific techniques.

Acknowledgement

This research is developed on the basis of a master thesis “Methods to Efficiently Handle Big Data in 5G Networks” 2018[1], double degree Erasmus + program between Higher School of Economics and UAS Technikum Wien. I would like to express my sincere gratitude to my academic supervisors and the professors and lecturers at the Big Data Systems program. Improvement is made during study at Skoltech.

[1] https://www.hse.ru/en/edu/vkr/219430036

References

[1] Zeng, D., Gu, L. and Guo, S. (2015) Cost minimization for big data processing in geo-distributed data centers. In *Cloud networking for big data* (Springer), 59–78.

[2] Bi, S., Zhang, R., Ding, Z. and Cui, S. (2015) Wireless communications in the era of big data. *IEEE communications magazine* **53**(10): 190–199.

[3] He, Y., Yu, F.R., Zhao, N., Yin, H., Yao, H. and Qiu, R.C. (2016) Big data analytics in mobile cellular networks. *IEEE access* **4**: 1985–1996.

[4] Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K. and Xiang, W. (2016) Big data-driven optimization for mobile networks toward 5g. *IEEE network* **30**(1): 44–51.

[5] Imran, A., Zoha, A. and Abu-Dayya, A. (2014) Challenges in 5g: how to empower son with big data for enabling 5g. *IEEE network* **28**(6): 27–33.

[6] Boccardi, F., Heath, R.W., Lozano, A., Marzetta, T.L. and Popovski, P. (2014) Five disruptive technology directions for 5g. *IEEE Communications Magazine* **52**(2): 74–80.

[7] Ramaprasath, A., Srinivasan, A. and Lung, C.H. (2015) Performance optimization of big data in mobile networks. In *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)* (IEEE): 1364–1368.

[8] Samulevicius, S., Pedersen, T.B. and Sorensen, T.B. (2015) Most: Mobile broadband network optimization using planned spatio-temporal events. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)* (IEEE): 1–5.

[9] Blondel, V.D., Decuyper, A. and Krings, G. (2015) A survey of results on mobile phone datasets analysis. *EPJ data science* **4**(1): 10.

[10] Italia, T. (2015), Telecom italia big data challenge. URL `https://dandelion.eu/datamine/open-big-data/`.

[11] Xu, R. and Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on neural networks* **16**(3): 645–678.

[12] Soto, V. and Frías-Martínez, E. (2011) Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*: 17–22.

[13] Liu, J., Chang, N., Zhang, S. and Lei, Z. (2015) Recognizing and characterizing dynamics of cellular devices in cellular data network through massive data analysis. *International Journal of Communication Systems* **28**(12): 1884–1897.

[14] Zhang, G.P. (2003) Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **50**: 159–175.

[15] Adhikari, R. and Agrawal, R.K. (2013) An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613* .

[16] Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural computation* **9**(8): 1735–1780.

[17] Sundermeyer, M., Schlüter, R. and Ney, H. (2012) Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

[18] Staudemeyer, R.C. and Morris, E.R. (2019) Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586* .

[19] Ho, S.L., Xie, M. and Goh, T.N. (2002) A comparative study of neural network and box-jenkins arima modeling in time series prediction. *Computers & Industrial Engineering* **42**(2-4): 371–375.