DATA ANALYTICS

**Final Project exploratory analysis**

TELECOM DATASET FOR DYNAMIC NETWORK HOPPING

horizontal line

# 

# Introduction

In today’s day and age we see that artificial intelligence has become a commonly used term even amongst the general population. However the objective of years of research is to exchange and yet make it safely accessible to the general public while still making calculations and the decision making process more accurate. The objective of the problem statement is to perform an exploratory analysis on real time simulated data collected from an Irish hill station with Two cell operators specified as A and B in the dataset along with different mode of transport which includes bus, car , train , tram and walk and this data would help us distinguish which cell tower located, where it is located, how it is connecting to differently paced vehicles all come under this very dataset.

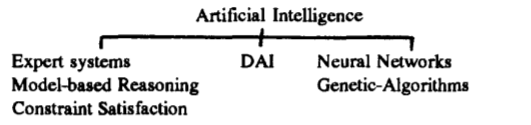
Scope of our research however can go beyond just an exploratory analysis but also trying to implement a recently popularised reinforcement learning method called Graphical Neural Networks (GNN) . With the growing population and massive demand for the ever expanding telecom networks dynamic hopping mechanism could prove to be quite useful and more accurate compared to traditionally installed system calculators.

### How do we approach the solution to this problem and what we aim to even find with the principal analysis of this very dataset? A brief about our dataset would start by our analysis on the features of the dataset. The dataset talks about a total of 20 attributes as listed below:

### *Timestamp Longitude Latitude Speed Operatorname CellID NetworkMode RSRP RSRQ SNR CQI RSSI DL\_bitrate UL\_bitrate State NRxRSRP NRxRSRQ ServingCell\_Lon ServingCell\_Lat ServingCell\_Distance*

Here the latitude and longitude were plotted for different modes of transport for operator A and B to get an overview of the cell tower location and the no range area for the dynamic allocation of network allocation to take place.

### However, doing so it is seen that standard clustering or classification algorithms may or may not be sufficient to identify which cell tower to hop to and traditional methods all use calculation of either euclidean or manhattan distances for allocation which makes signal clarity drastically different within just a couple of meters. Traditional methods had their pros where the call tower allocation could be found in O(n) time and so far that has proven to be the fastest method but it completely disregards call network quality, weather forecast, historic data analysis etc. which could play a vital role tower prediction for a smoother transition especially when there is extremely fast or extremely slow transition between regions.



**LITERATURE SURVEY:**

*Nivedhitaa R:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. no** | **Parameters:** | **Paper 1:** | **Paper 2:** | **Paper 3:** |
| 1. | **Name of paper:** | Artificial intelligence in telecommunications | Using Opportunistic Caching to Improve the Efficiency of Handover in LTE with a PON Access Network Backhaul | Analysis of smartphone user mobility traces for opportunistic data collection in wireless sensor networks |
| 2. | **Author of paper:** | Claude Muller, Evan H Magill, Patrick Prmw and D Geoffrey Smith | David Stynes, Kenneth N. Brown, Cormac J. Sreenan | Xiuchao Wu, Kenneth N. Brown, Cormac J. Sreenan |
| 3. | **Published year:** | Global telecom 1993 | May 2014 | 12 July 2013 |
| 4. | **Dataset used:** | Bell labs, asynchronous telecom network simulator dataset | LTE in POV metropolitan areas | Mobile Data Challenge 2012 (by Nokia) Workshop |
| 5. | **Summary:** | The paper summarizes about the different AI modeling techniques which can prove to be successful in analysis of telecom datasets: 1.Expert based modelling  2. Constraint based modelling.  3.neural network algorithms  The detailed study of their pros and cons along with some understanding about their application in industry standard datasets with their correlation and relevance to growing network technology. | The paper talks about core core network principles which can be understood and implemented with AI in order to perform efficient hopping and scheduling of signal packets across the cell towers.  This dataset is extremely similar to our problem dataset at hand and they talk about caching on packing (IN, WAIT AND OUT ) queuing , delays, and transmission rates.    It also provides us insights on the scheduling algorithms for packets which are waiting for their signal to be strong enough for transmission like FIFO, BEST FIT etc. | This paper summarizes the fact that the increasing ubiquity of smartphones coupled with the mobility of their users will allow the use of smartphones to enhance the operation of wireless sensor networks.  In addition to accessing data from a wireless sensor network for personal use, and the generation of data through participatory sensing, it proposes the use of smartphones to collect data from sensor nodes opportunistically. For this to be feasible, the mobility patterns of smartphone users must support opportunistic use.Analysis on the dataset from the Mobile Data Challenge by Nokia, helps identify the significant patterns, including strong spatial and temporal localities. These patterns should be exploited when designing protocols and algorithms, and their existence supports the proposal for opportunistic data collection through smartphones. |
| 6. | **Nature of model:** | Two models were described: 1. Supervised learning euclidean distance based  2. Neural network based model with various hidden layers (reinforcement learning) | Reinforcement learning with programmed queues and scheduler algorithms | Multivariate Polynomial regression, random forest regression.  Basics into intelligent networks(GNN) |
| 7. | **Accuracy and error:** | Here since this was a descriptive paper accuracy and error wasn’t estimated | Error rate of 0.06 wrt packet transfer | Alpha of 0.05 and accuracy 0.78 which is considered quite good |
| 8. | **Pros:** | Brought about the highlights and use case of various approaches to solve the dataset at hand.  Comprehensive documentation on analysis and error correction of the same | Their experimental evaluation is an intricate study of the dataset which shows the clear cut analysis that training set performs extremely well with the model as they have developed.  EDA is done to a point wherein they have also found out that they don't require a validation data set and it still fits great with the test set. | Here there is more emphasis on the telecom and network path detection mechanism in wireless medium which essentially helps understanding the domain in depth and even the dataset is made open source available for students to work on and provide insights on the data. |
| 9. | **Cons:** | Doesn’t cover the mathematical model and its requirement for each of the specified models.  Connectivity of time complexity and space complexity for a programmer's perspective was not tallied since power of computation wasn't sufficient in 1993. | shown that there is a large amount of redundant data transmission as part of the LTE handover process. Which isn’t clearly conveyed in the paper which gives us as anlysissers for more scope of research on the very same issue | This paper however hasn't done feature scaling or dimensionality reduction even though we may notice that some attributes have near to 0 significance wrt Y. |

*Kunal Mujoo:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. no** | **Parameters:** | **Paper 1:** | **Paper 2:** | **Paper 3:** |
| **1.** | **Name of paper:** | Beyond throughput: a 4G LTE dataset with channel and context metrics. | Estimating Cell Capacity from Network Measurements in a Multi-Service LTE System. | Performance Evaluation of Multivariate linear Regression Model with LDA and PCA. |
| **2.** | **Author of paper:** | Raca, Darijo; Quinlan, Jason J.; Zahran, Ahmed H.; Sreenan, Cormac J. | J.A. Fernandez-Segovia, S. Luna-Ramirez, M. Toril, J.J. Sanchez-Sanchez | Krishan Pal, Mayank Sharma – Amity University. |
| **3.** | **Published year:** | 2018-06 | 2015 | 2020 |
| **4.** | **Dataset used:** | https://www.kaggle.com/aeryss/lte-dataset | N/A | N/A. 200+ attributes. |
| **5.** | **Summary:** | This paper provides a high-level overview of the dataset which provides insight into different mobility patterns across both mobile operators, with respect to application throughput, variation in bandwidth, and channel metrics. It defines many possible use cases like Commercial VR technology with rapidly increasing download demands which could be implemented using HAS(HTTP adaptive streaming) mechanism. Another use-case is handover analysis and prediction to generate a large number of realistic traces which alleviates the need for manually collecting and testing vast amounts of data. This is the first publicly available dataset that contains throughput, channel and context information for real-time analysis of a production 4G network. | Network dimensioning is a very critical task. We need to estimate future traffic demand and upcoming network capabilities to detect network bottlenecks in advance. Several multivariate linear regression equations are used to estimate the value of service – specific QOS indicators collected on a cell basis. | Regression analysis is continuously performed on key attributes in the dataset to predict system health by anomaly detection and accordingly preventive and corrective measures are taken. LDA and PCA are key techniques performed for dimensionality reduction. Independent variables with high correlation predict dependent variables with better accuracy. Then use plain variate or multivariate linear regression. |
| **6.** | **Nature of model:** | - | Multiple linear regression models are used. It is a supervised machine learning algorithm. | Simple/Multivariate Linear Regression model. LDA is supervised and PCA is unsupervised. |
| **7.** | **Accuracy and Error:** | - | Full Model predicts TH8qci accurately, but fails to predict VoLTE SatisfUsRatio due to the low VoLTE traffic in the available dataset, which proved to be less than 0.24 VoLTE connections per cell and hour on average. | Average prediction error in Multivariate linear regression method is relatively high (21.89%), if not used in conjunction with any of the dimensional reduction techniques (PCA or LDA). With PCA error is reduced to 19.56% and with LDA error is reduced to 14.57%. |
| **8.** | **Pros:** | HAS is the most popular streaming format. As connection capacity changes over time, the size of data must vary to adapt to the varying network conditions. This technology is fundamental to help accommodate growth of delivery over wireless networks. | Measurement-based method to estimate LTE cell capacity. Unlike analytical approaches, it has low computational load. Reasonable accuracy for services with QCI8. | LDA does not require any pre-processing, which reduces computational costs. |
| **9.** | **Cons:** | Interoperability and economics. Also, the manufacturer’s ability to delegate the selection of quality of content to the device. Many devices grab the highest capacity available on the networks, leading to wastage of resources. | Poor accuracy for services with QCI1. Solved currently(2021) due to increased VoLTE in the past 5 years | Relatively high average prediction error. |

*Stuti Patel:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. no:** | **Parameters:** | **Paper 1:** | **Paper 2:** | **Paper 3:** |
| **1.** | **Name of the paper:** | LTE QoS Parameters Prediction Using Multivariate Linear Regression Algorithm | A Quantitative and Comparative Evaluation of Key Points Selection Algorithms for Mobile Network Data Sets Analysis | Discovering Computer Networks Intrusion using Data Analytics and Machine Intelligence |
| **2.** | **Author of the paper:** | Mourad Nasri, Mohamed Hamdi | David Cortes-Polo, Luis Jimenez, J.L. Gonalez-Sanchez, Javier Carmona-Murillo | Oluwafemi A. Sarumi, Adebayo O. Adetunmbi |
| **3.** | **Published year:** | 2019 | June 2021 | July 2020 |
| **4.** | **Dataset used:** | N\A | Telecom Italia, 2014 edition of big data challenge (real data set) | Network Security Laboratory Knowledge Discovery and Data Mining (NSL-KDD) dataset and the University of New South Wales–NB 2015 (UNSW-NB15) dataset |
| **5.** | **Summary:** | The paper talks about the different QoS requirements for the wide range of mobile applications available. We use a data set that contains the evolution of the 4 KPIs for a wireless operator for six months. Then it evaluates the model by calculating the residuals. Finally, it discusses several case studies to show the potential of the optimum linear function. | Different key points selection algorithms are investigated to make a comparative evaluation and analyze the performance of algorithms which use different approaches to select these points. The effect of key points is statistically evaluated using selection algorithms in the clustering, performance is measured. | In this paper, a comparison between two intrusion detection systems is shown–one that uses the association rule data mining approach–Apriori and the other that adapts the use of a machine learning technique–Support Vector Machine (SVM). Evaluation results show that SVM performs better than Apriori in terms of accuracy, while Apriori gives a better performance in terms of testing speed. |
| **6.** | **Nature of the model:** | Machine learning multivariate linear regression technique. |  | SVM and Apriori analytics |
| **7.** | **Accuracy and error:** |  | Average accuracy expressed in percentage of the methodology applied to the 24 interval synthetic data set using, IEA, OSP and SGA for β random is 70.06, 87.83 and 70.11 respectively. | The accuracy using the UNSW-NB15 dataset was 64.09(filter+apriori) and using the NSL-KDD dataset was 67(filter+apriori). 15 features were selected. |
| **8.** | **Pros:** | This method allows optimization engineers to have an accurate idea about the current quality of service provided to the customers and take appropriate actions to manage the network resources. | These key points selection algorithms are used in a real scenario to evaluate the impact of the different approaches in the analysis, which fit in very well with the data model used for the test set | An efficient intrusion model for preventing network intrusion attacks in real-life application scenarios. The proposed model using filter selection method and SVM classifier on UNSW-NB15 datasets in this paper gives better performance than other previously proposed models. |
| **9.** | **Cons:** |  | In the random data set, the overall accuracy  decreases, reducing the similarity of the classes, which implies a reduction of the algorithm  accuracy, similar to the IEA algorithm. | Number of selected features is only 15 which might lead to an inaccuracy in predictions when a heavy load of features are involved. |

*Swayam S.A:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl No:** | **Parameters:** | **Paper 1:** | **Paper 2:** | **Paper 3:** |
| 1. | **Name of the Paper:** | Log Distance Path Loss or Log Normal Shadowing Model | Opportunities for AI/ML in Telecommunications Networks | Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework. |
| 2. | **Author of the Paper:** | Mathuranathan Viswanathan | Alberto Castro, Eduardo Grampín, Javier Baliosian, Matías Richart | Mohammed Al-Mashraie, Sung Hoon Chung, Hyun Woo Jeon |
| 3. | **Publishing Year:** | 30th of September, 2013. | 17th December, 2018. | 28th October, 2019. |
| 4. | **Datasets Used:** | N/A | N/A |  |
| 5. | **Summary:** | This paper is an implementation of the path-loss exponent (PLE). Log normal shadowing model is used to predict the propagation loss of a wide range of environments. This model also encompasses random shadowing effects that block signals by trees, buildings or any form of blockade. | While it is true that we are in the middle of one of the Artificial Intelligence hypes, it is also true that the combination of unprecedented computation-power and data availability with new variations of well seasoned Machine Learning algorithms is dramatically changing the optimization strategies for large ICT industries. Especially, the telecommunications industry has always had to deal with complex systems, stochastic contexts, combinatorial problems, and hard to predict users; Machine Learning-aided optimization was just waiting there to be used by telecoms. In this paper, we introduce some basic Machine Learning concepts, and discuss how it can be used in the context of telecommunications networks, particularly in optical and wireless networks. | Customer retention is one of the key challenges in the telecommunication industry. Companies may find customer churn prediction to be vital to the success of their operations because a careful analysis of churning may provide a crucial means to retain customers. Among potentially a vast amount of factors that impact the churn, it is critical to identify the most influential ones towards which customer retention efforts can be directed. In this paper, we compare the performance of different churn prediction models based on the real data obtained from a partner company. The prediction models include logistic regression, support vector machines, random forest, and decision tree. Furthermore, the push-pull-mooring (PPM) framework is utilized to study the effect of features on customers churn behavior from push, pull, and mooring perspectives. A partial least squares (PLS) regression is used to perform the PPM analysis. Furthermore, the behavior or churners and non-churners are analyzed.. |
| 6. | **Nature of Model:** | Free space propagation model. | The GN model of machine learning. | Churn prediction. |
| 7. | **Accuracy & Errors:** |  | Evaluations over the testing dataset  demonstrate that the QoT estimator achieves an overall prediction  accuracy of 95.79%, with the false positive rate being only  1.58%. | This proposed framework is shown to be very effective with high accuracy. |
| 8. | **Pros:** | This model can be used in the telecommunication industry to accurately calculate the signal strength. | This model helps to find the most suitable connection to a cell tower, with very high efficiency. | The prediction models include logistic regression, support vector machines, random forest, and decision tree. |
| 9. | **Cons:** | This model requires the geographic knowledge and lay of the land around the signal towers. |  |  |

**References:**

1. <https://cora.ucc.ie/handle/10468/6400>
2. <https://conferences.sigcomm.org/sigcomm/2013/papers/sigcomm/p363.pdf>
3. <http://www.jocm.us/uploadfile/2019/0610/20190610052324750.pdf>
4. <https://technews.tmcnet.com/live-streaming-coverage/topics/live-streaming-coverage/articles/256184-pros-cons-adaptive-bit-rate-streaming.htm>
5. <http://webpersonal.uma.es/~TORIL/files/2015%20CL%20multi-service%20dimensioning%20LTE.pdf>
6. <https://poseidon01.ssrn.com/delivery.php?ID=230074113006025031089070074092125026039073086004051025077099122072123085122071102093043031001119031112002026085074028006066122021091012079014090099004110092097002011093018003012071113006098070017093081096023025101024069028078018105090065111092112084116&EXT=pdf&INDEX=TRUE>
7. <https://www.ucc.ie/en/media/research/misl/2013publications/pmc2013.pdf>
8. <https://www.ucc.ie/en/media/research/misl/2014publications/lanman2014.pdf>
9. <https://sci-hub.se/http://ieeexplore.ieee.org/document/318205/?arnumber=318205&tag=1>
10. <https://www.researchgate.net/publication/343288869_Discovering_Computer_Networks_Intrusion_using_Data_Analytics_and_Machine_Intelligence>
11. <https://www.gaussianwaves.com/2013/09/log-distance-path-loss-or-log-normal-shadowing-model/>
12. <https://www.gaussianwaves.com/author/mathuranathan/>
13. <https://www.researchgate.net/publication/328439751_Opportunities_for_AIML_in_Telecommunications_Networks>
14. <https://www.sciencedirect.com/science/article/abs/pii/S0360835220302102?via%3Dihub>

**Github Repository Link:** <https://github.com/KunalMujoo/Data-aggregation-telecom>