Modelling 5G Data Using Tree-Based Machine Learning Models

P. Mithillesh Kumar and M. Supriya

Abstract 5G or fifth generation is the latest in the communication technology which is being researched worldwide as a successor to the current 4G technology. 5G operates on higher bandwidth with higher data rates of the order of Gbit/s. 5G is estimated to play a major role in the development of smart cities and IoT use cases. Lumos 5G is one of the groups researching on the topic. In this paper, the throughput obtained under various conditions is analysed as a regression model in machine learning with the features as continuous variables. It is observed that the newer tree machine learning models are performing better on the dataset than the traditional tree models. This is verified by performing a tenfold cross-validation check on the best performing models.

Keywords 5G · Regression · Throughput · Trees · Cross-validation

1 Introduction

The need for higher bandwidths and higher data rates is increasing day to day with the rate of increase of Internet adoption and the increased need to be satisfied. Higher the bandwidth, higher is the rate of data transfer, analysis and the development of the society. The bandwidths are graded using data rates, and the developments are given by generations. The generations range from the first generation to the fourth generation. The current generation of data transfer being implemented in the cellular networks is called the fourth generation of wireless connection or Long-Term Evolution (LTE). This is turning insufficient with the rapid increase in the number of devices connected to the Internet and the network load generated these days.

P. Mithillesh Kumar (✉) · M. Supriya
Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: bl.en.p2dsc20023@bl.students.amrita.edu

M. Supriya
e-mail: m_supriya@blr.amrita.edu

D. Gupta et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 473,
https://doi.org/10.1007/978-981-19-2821-5_7

In order to cater to the needs of the society, the latest improvement to the cellular network generations is the fifth generation of wireless connection which has a bandwidth of the order of up to 10 GBit/s [1]. As the world is becoming more and more virtual and the devices become more and more smart these days, the network traffic has increased a lot. The increased data rates reduce the latency of the network, reduce the information loss, increase network availability, more reliability, higher network load handling capacity and more consistency of the network to the end-user. 5G is based on orthogonal frequency-division multiplexing (OFDM) modulation method that reduces interference between multiple channels [2]. The 5G technology uses the sub-6 GHz and the mmWave frequency spectrum. 5G is expected to be a driving factor for a 13.1 trillion dollars in the global economy. 5G technology is expected to have a major impact on the adoption and development of the latest technologies such as the virtual reality (VR), Internet of Things (IoT) and the artificial intelligence (AI). In IoT, it is expected to have a major impact on the edge computing technology. The current global air standard for the 5G technology is the 5G-NR (New Radio) developed by third Generation Partnership Project (3GPP), and the first version of specifications was released in 2017. It can handle the applications that are time consuming with the current telecommunication standard with ease and enhance the productivity.

The 5G technology is limited by the factors such as the inability to penetrate through the walls, as the frequency is high, the losses and the dissipation are also higher. This limits the range of 5G network within a short range, and its signal can also be affected even by the air. This range of 5G is much lesser than that of the 4G network. The initial costs of implementation are high, the current cellular devices are incompatible with the upgraded network, and this leads to an additional cost of upgrading to a new device to the consumers. Adaptive modulation and coding scheme (MCS) is used to ensure minimal data loss in 5G networks [3]. When the error rate falls below a threshold, the network operates on lower MCS to reduce the error rate. Here speed of the network is compromised in order to ensure minimal data loss.

This work focuses on modelling the 5G data based on the conventional and unconventional machine learning models to understand and predict the performance of the network under multiple conditions as a regression problem. The paper is structured as follows: Sect. 2 providing the literature survey; Sect. 3 describing the dataset and the features involved; Sect. 4 providing the exploratory data analysis; Sect. 5 providing the analysis and its results; Sect. 6 concludes the discussion.

2 Literature Review

A random forest (RF) regression model is implemented, and the performance of the model is estimated on the mean absolute error (MAE) and root mean square error (RMSE) in [4]. The best case obtained is a MAE of 163 with a RMSE value of 241, and the gradient boosted regressor (GBR) model implemented has given the best MAE of 100 with a RMSE of 154. Also, the history-based harmonic mean is modelled, and the implementation of an ordinary least squares (OLS) regression has generated a MAE of 231 and a RMSE of 340 on the throughput when analysed for a short-term prediction.

A logistic regression-based decision tree (DT) model for the accident injury severity which has shown a misclassification rate of 30%, and the cross-validation-based confusion matrix gives a misclassification rate of 32% in [5]. A DT model has achieved a 100% accuracy in condition monitoring of a milling tool in [6]. A RF model implementation for analysing pressure in suddenly expanded aerodynamic flows is applied in [7]. The analysis shows that the RF regression model has performed better than the K-means model for a nonlinear target variable prediction. Bayesian model proposed and implemented in [8] holds a better result with an $F1$-score of 0.785 while dealing with the data related to the consequences of construction accidents. The useful life prediction of lithium ion batteries by applying a adaptive extended kalman filter and genetic algorithm optimized support vector regression model has shown the best performance with the least MAE and RMSE in [9].

A support vector machine classifier (SVC) and linear discriminant analysis model has detected driver drowsiness in [10]. A SVC has been implemented in [11] for 2D indoor localization using RFID. A SVM-based heart disease detection system is modelled in [12]. A grinding wheel condition monitoring system using the acoustic characteristics is modelled in [13] applying CART and SVC. The extra trees (EXT) model has performed much better than the other models in the determination of bubble point pressure and the oil formation volume factor in [14]. The prediction of rock mass class on classification and regression tree-based AdaBoost model has achieved an $F1$-score of 0.77 and an accuracy of 0.865 in [15]. Gradient boosted regression (GBR) model is implemented for the degradation of prismatic cells, where the model has performed closer to the best model which is built upon the gradient boosted model in [16]. An extreme gradient boosted regressor (XGB) model on the interfacial tension between oil and injected gas has shown the least error in predicting the target variable with a R^2 value of 0.997 in [17]. A XGB model has been implemented in [18] to predict the sales of Big Mart, and the results show that it has outperformed the existing models. A construction cost prediction problem is modelled using a hybrid natural–light gradient boosted regression (LGB) model which has given an RMSE of 0.5 with a R^2 value of 0.99 in [19].

From the literature survey, it is noted that every model has performed the best on particular use case and the data is modelled using multiple models for a particular use case. In this work, multiple models are applied on the 5G dataset [20] as a regression problem, the analysis is performed, and the best model is identified.

3 Dataset Description

The dataset sourced from IEEE dataport site [20] contains about 68,118 records of data collected by means of 300 km of walking, 130 km of driving and 35 TB of data download. The data is collected over a loop area of 1300 m loop length. The dataset contains the following features such as the run number, sequence number, abstract signal strength, latitude, longitude, moving speed, compass direction, NR status/connection status, received signal strength indication (RSSI), reference signal received power (RSRP), reference signal received quality (RSRQ), reference signal signal-to-noise ratio (RSSNR), raw signal strength power (nrssRsrp), raw signal strength quality (nrssRsrq), raw signal strength signal-to-noise ratio (nrssSinr), throughput, mobility mode, trajectory direction and tower ID. All these data were collected using an Android API. Here the features such as the moving speed is recorded in m/s, the compass direction is measured in degrees, the connection status is a categorical variable with three options such as the connected, not restricted and none, mobility mode is also a categorical variable with two options such as walking and driving indicating the means of motion, and trajectory direction is also a categorical variable with two options such as clockwise or anticlockwise within the planned trajectory of motion and tower ID indicating to which tower the device is currently connected. The throughput is the target parameter in this work which is the rate of output obtained for a given condition at a given instance of time. All the other parameters indicate the signal qualities at a given run and sequence number. The run and sequence numbers are used for identification purposes. The data has been recorded over observing the throughput and the signal quality, the signal power with and without barriers between the carrier or the device and the tower to which it is connected. Here the accuracy of the data collected is subject to the accuracy and performance of the API.

4 The Exploratory Data Analysis

This section presents the exploratory data analysis of the 5G dataset described in the previous section with a focus on the throughput. Figure 1a, b indicates the throughput obtained along the contour traversed during the data collection process. It is observed that the throughput is minimal at some locations and high close to 5G speeds at other locations.

Figure 2 indicates the throughput obtained from each of the tower, and it could be noted that few towers are low in throughput, whereas others are close to the 5G spectrum range.

Figure 3a, b indicates the regression plot when the throughput is measured while driving a car and walking along the contour, respectively. The line represents the ordinary least square (OLS) regression line of the plot.

Figure 4 indicates the throughput obtained with the connection status a categorical variable which has three states which are the not restricted, connected and none, indicating the connection status with the 5G network.

Figure 5a, b indicates the throughput obtained when the traversal trajectory is clockwise and anticlockwise, respectively, in nature with the angle of motion as measured by the API, and the lines originating from the centre indicate the throughput

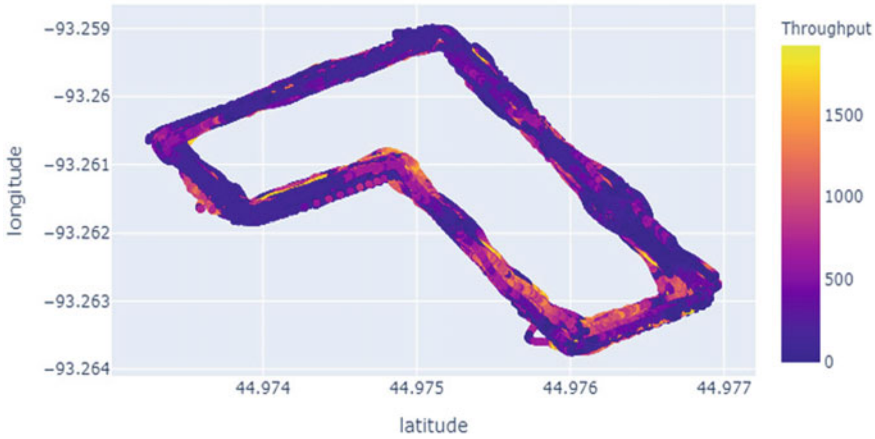

Fig. 1 Throughput obtained along the contour traversed

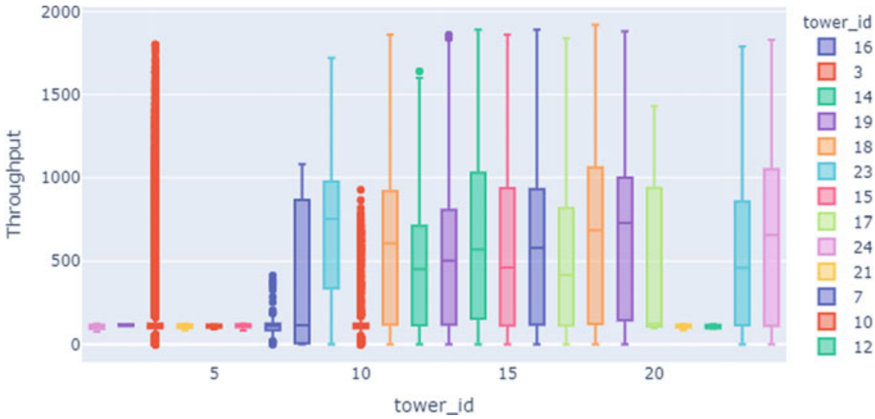

Fig. 2 Throughput obtained from the individual tower

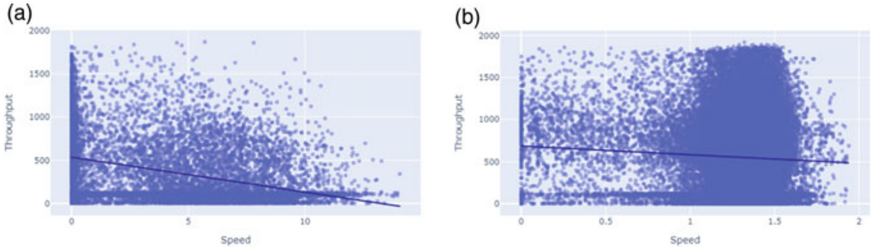

Fig. 3 Regression plot of throughput obtained **a** while driving a car and **b** while walking

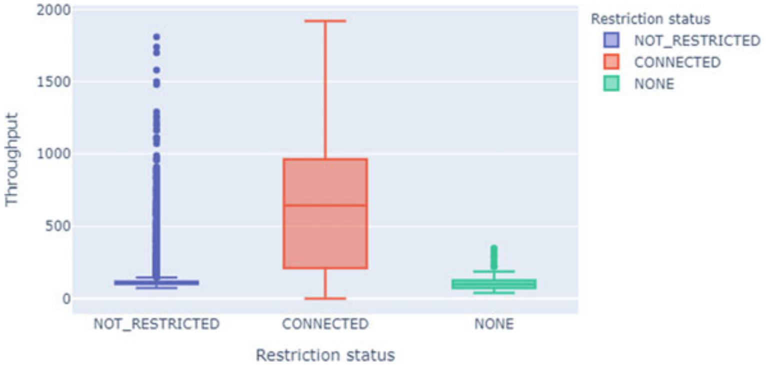

Fig. 4 Throughput obtained with the connection status

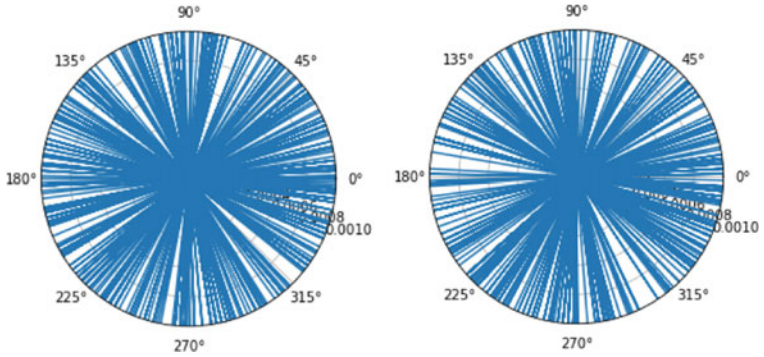

Fig. 5 Throughput obtained while direction of motion along the contour is clockwise and anti-clockwise with compass direction **a** as a polar plot when moving clockwise, **b** as a polar plot when moving anticlockwise

obtained at that particular angle. It is observed that the throughput is limited at some angles and in the 5G range at few other. Also, it has to be noted that there is significant difference in the throughputs obtained in case of the clockwise and anticlockwise directions of traversal. The accuracy of the plots is subject to the accuracy of the API. So, these categorical variables have been eliminated to reduce the modelling error, and the data is modelled as a regression problem with the other continuous variables.

5 Modelling and Analysis

The dataset is preprocessed to remove the null values detected, which are later replaced with 0. This preprocessing will enable the machine learning model to learn the dataset completely and present a model with higher accuracy. This in turn may reduce the errors and misclassification in the further process. Here the approach is to model the throughput selected as the target variable, and the data is modelled as a regression model not considering the categorical variables in the dataset. The categorical variables that are removed from the analysis are the abstractSignalStr, nrStatus, mobility mode, trajectory direction and tower ID. The less important features such as the run-num and seq-num are dropped from the analysis as they are just used for identification purposes. For modelling the data, the dataset is split into the standard train–test split ratio of 70:30. The models are applied onto the dataset as a regression problem.

$$\mathrm{MAE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|y_i - x_i| \tag{1}$$

$$\mathrm{MSE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - x_i)^2 \tag{2}$$

$$R^2 = 1 - (\mathrm{SSR}/\mathrm{SST}) \tag{3}$$

where

n	number of terms,
i	the ith term,
y_i	the actual value,
x_i	the calculated value,
MAE	mean absolute error,
MSE	mean squared error,
SSR	sum of squares of residuals,
SST	total sum of squares.

The parameters used for measuring the model performance are the mean absolute error (MAE), mean squared error (MSE) and the R-squared (R^2) values represented by Eqs. (1)–(3), respectively. MAE is the mean of the absolute difference between the actual and the predicted values. MSE is the mean of the square of the difference between the actual and the predicted value. R^2 is the measure of the correlation between the actual and the predicted values. SSR is the sum of the square of residues, and SST is the total sum of squares. Multiple models such as the decision tree (DT) model, random forest (RF) model, Naive Bayes (NB) model, support vector regressor (SVR) model, extra trees regressor (EXT) model, AdaBoost regressor (ABR) model, gradient boosted regressor (GBR) model, Xtreme gradient boosted regressor (XGB) model and light gradient boosted (LGB) model are applied on to the dataset. The performance of the DT model with different number of leaf nodes along with the tenfold cross-validated score of each of the case being analysed by varying the number of leaf nodes as 5, 50, 500 and 5000 is given in Table 1.

From Table 1, it could be noted that the model performance is increasing by increasing the number of leaf nodes of the decision tree model indicated by the reduction in the values of the parameters MAE and MSE. The performance of the models applied onto the 5G data which are the RF, NB, SVR, EXT, ABR, GBR, XGB and LGB is presented in Table 2. The performance of the models is measured using the parameters MAE, MSE and the R^2 values.

From the results in Table 2, EXT model has performed the best followed by the RF, DT model with 5000 leaf nodes, XGB and the LGB. The EXT model performance metrics are the MAE value of 140, MSE value of 46350 and the R^2 score of 0.783. As a validation of the performance of the models, cross-validation (CV) check is performed on the top five models, i.e. DT of 5000 leaf nodes, RF, EXT, LGB and XGB by dividing the dataset into ten partitions, i.e. a tenfold cross-validation is performed. The cross-validation scores of the top five models are given in Table 3.

From Table 3 of cross-validation scores, it is indicative that the EXT model has performed better than the other models applied onto the dataset.

Table 1 Performance of DT model

Number of leaf nodes in DT	MAE	MSE	R^2
5	256	125,144	0
50	232	106,202	0
500	201	85,083	0
5000	173	81,207	0

Table 2 Performance of other models

Algorithm	MAE	MSE	R^2
RF	145	47,746	0.777
NB	474	305,176	−0.425
SVR	391	219,580	−0.025
EXT	140	46,350	0.783
ABR	261	118,769	0.445
GBR	229	100,120	0.532
XGB	179	63,981	0.701
LGB	194	74,407	0.652

Table 3 Cross-validation scores of top five models

CV score (tenfold)	DT	RF	EXT	LGB	XGB
1	−0.1797	0.3815	0.3977	0.4256	0.3386
2	0.0467	0.489	0.5026	0.499	0.4775
3	0.1057	0.5312	0.5299	0.531	0.4932
4	−0.1097	0.4481	0.4638	0.4475	0.3604
5	0.1051	0.4217	0.4239	0.4359	0.3904
6	0.0604	0.4703	0.4794	0.5181	0.5189
7	−0.1729	0.4312	0.4593	0.4259	0.3684
8	−0.0044	0.4288	0.4385	0.4664	0.4365
9	0.0016	0.5046	0.5288	0.4514	0.4492
10	−0.0231	0.3854	0.3868	0.4331	0.4188

6 Conclusion

5G is one of the emerging technologies globally which is finding use case in almost every application and the successor to the current 4G technology. In this paper, the throughput obtained under various conditions is modelled as a regression problem eliminating the categorical variables. From the analysis, it is observed that, among the models being analysed, the EXT model has performed the best with a MAE of 140, MSE of 46350 and a R^2 score of 0.783, and the results are in accordance with the results of the tenfold cross-validated score. The future scope of work is to apply other machine learning models upon the data and identify the best model that can be applied onto 5G dataset as a regression problem.

References

1. https://www.qualcomm.com/5g/what-is-5g
2. https://www.qualcomm.com/media/documents/files/5g-research-on-waveform-and-multiple-access-techniques.pdf
3. https://www.3gpp.org//ftp//tsg_ran//WG1_RL1//TSGR1_17//docs//PDFs//R1-00-1395.pdf
4. Narayanan A, Ramadan E, Mehta R, Hu X, Liu Q, Fezeu RAK, Dayalan UK, Verma S, Ji P, Li T, Qian F, Zhang Z-L (2020) LUMOS5G: mapping and predicting commercial mmWave 5G throughput. In: Proceedings of the ACM internet measurement conference, IMC’20. Association for Computing Machinery, New York, NY, USA, pp 176–193
5. Rezapour M, Molan AM, Ksaibati K (2020) Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. Int J Transp Sci Technol 9(2):89–99
6. Mohanraj T, Yerchuru J, Krishnan H, Nithin Aravind RS, Yameni R (2021) Development of tool condition monitoring system in end milling process using wavelet features and hoelder’s exponent with machine learning algorithms. Measur J Int Measur Confed 173

7. Afzal A, Aabid A, Khan A, Khan SA, Rajak U, Verma TN, Kumar R (2020) Response surface analysis, clustering, and random forest regression of pressure in suddenly expanded high-speed aerodynamic flows. Aerosp Sci Technol 107:106318
8. Zhu R, Hu X, Hou J, Li X (2021) Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Safety Environ Protection 145:293–302
9. Xue Z, Zhang Y, Cheng C, Ma G (2020) Remaining useful life prediction of lithium-ion batteries with adaptive unscented kalman filter and optimized support vector regression. Neurocomputing 376:95–102
10. Madireddy R, Anudeep DSK, Poorna SS, Anuraj K, Gokul Krishna M, Balaji A, Venkat DJ (2021) Driver drowsiness detection system using conventional machine learning. Lecture notes on data engineering and communications technologies 58:407–415
11. Aravind Raamasamy S, Shanmuga Pradeep P, Mani Madhav Goud CH, Viswanathan Babu CA, Jayakumar M (2021) Analysis of machine learning algorithms for RFID based 2D indoor localization. Lecture notes on data engineering and communications technologies 58:229–242
12. Anusha M, Suresh K, Chandana M (2021) Earlier prediction on the heart disease based on supervised machine learning techniques. In: Proceedings—5th international conference on intelligent computing and control systems, ICICCS 2021, pp 1696–1703
13. Rameshkumar K, Mouli DSB, Shivith K (2021) Machine learning models for predicting grinding wheel conditions using acoustic emission features. SAE Int J Mater Manuf 14(4)
14. Seyyedattar M, Ghiasi MM, Zendehboudi S, Butt S (2020) Determination of bubble point pressure and oil formation volume factor: extra trees compared with LSSVM-CSA hybrid and ANFIS models. Fuel 269:116834
15. Liu Q, Wang X, Huang X, Yin X (2020) Prediction model of rock mass class using classification and regression tree integrated adaboost algorithm based on tbm driving data. Tunnell Undergr Space Technol 106:103595
16. Wang F-K, Mamo T (2020) Gradient boosted regression model for the degradation analysis of prismatic cells. Comput Indust Eng 144:106494
17. Zhang J, Sun Y, Shang L, Feng Q, Gong L, Kuankuan W (2020) A unified intelligent model for estimating the (gas + n-alkane) interfacial tension based on the extreme gradient boosting (XGBoost) trees. Fuel 282:118783
18. Ranjitha P, Spandana M (2021) Predictive analysis for big mart sales using machine learning algorithms. In: Proceedings—5th international conference on intelligent computing and control systems, ICICCS 2021, pp 1416–1421
19. Chakraborty D, Elhegazy H, Elzarka H, Gutierrez L (2020) A novel construction cost prediction model using hybrid natural and light gradient boosting. Adv Eng Inform 46:101201
20. https://ieee-dataport.org/open-access/lumos5g-dataset