Towards Long-Term Coverage and Video QoE Prediction in Cellular Networks

Andrea Pimpinella*, Alessandro E. C. Redondi*, Iacopo Galimberti†, Francesco Foglia†, Luisa Venturini†
*Dip. Elettronica, Informazione e Bioingegneria, Politecnico di Milano
†Vodafone Group, Network Engineering and Delivery
Email: name.surname@polimi.it or name.surname@vodafone.com

***Abstract*—Network operators are interested in continuously monitoring the Quality of Experience of their customers to minimise the churn rate: however, collecting user feedback through surveys is a cumbersome task. In this work we explore the possibility of predicting the long-term QoE relative to network coverage and video streaming starting from user-side network measurements only. We leverage country-wide datasets to engineer features which are then used to train several machine learning models. The obtained results suggest that, although some correlation is visible and could be exploited by the classifiers, long-term QoE prediction from network measurements is a very challenging task and requires future investigations.**

***Index Terms*—QoE prediction, Mobile Cellular Network, Network Intelligence**

I. Introduction

According to recent Cisco estimates, by 2021 mobile cellular networks will connect more than 11 billion of mobile devices and will be responsible of more than one fifth of the total IP traffic generated worldwide. Aware of these facts, mobile network operators are constantly monitoring and improving their access networks in order to give the best possible service to users and reduce failures, with the final goal of generating profit. This goal can be reached on the one hand by attracting as many new customers as possible and on the other hand trying to minimise the churn rate, i.e. the percentage of customers that, due to an unsatisfactory service, stop their subscriptions and move to a different operator.

In order to monitor the level of satisfaction of their customers, network operators often rely on surveys and questionnaires. Standard tools exist to capture the level of satisfaction of users through generic questions: as an example the Net Promoter Score (NPS) survey asks user to indicate the likelihood of recommending the network operator to a friend or colleague on a scale from 0 to 10. In addition to such a generic survey, operators often ask customers to reply on very specific questions related to the Quality-of-Experience (QoE) of certain mobile network services (network coverage, voice and video quality, etc.), which can better highlight possible problems in the network. Based on the results of such surveys, operators have some clues on which services should be boosted up and possibly where: as an example, the operator can invest money to increase the bandwidth or the output power available at a certain base station.

Unfortunately, performing customer feedback surveys is costly and cumbersome for operators, mainly due to the generic poor cooperative attitude of customers: most of the times, the survey proposed to customers are in fact returned empty. At the same time, network operators has several ways of gathering objective measurements from their customers: radio statistics and channel quality indicators can be obtained at the Radio Access Network (RAN), while advanced measurements such as throughput or latency can be measured with deep packet inspection (DPI) devices and network probes nowadays commonly installed at the network gateways (GGSN in 3G networks or PGW for 4G networks). Additionally, operators may obtain network measurements directly from the user terminals with specific applications running in background and installed under the user consent. Leveraging the renovated interest in machine learning and artificial intelligence techniques, operators may therefore attempt to predict the QoE of customers starting from network measurements only, rather than relying on user surveys.

In this paper we explore this possibility and predict the user QoE relative to two cellular network aspects: network coverage and video streaming. The former is a necessary service for every user as trivially, no network activities can be performed without radio coverage. The latter has become more and more important for network operators in the last few years, as video already constitutes the majority of mobile traffic [1]. Combined, the QoE relative to the two aspects may play a big role in a customer decision to leave its current operator for a better one.

Differently from related works in the area, we focus here on the long-term QoE, i.e., the satisfaction reported by a user relative to a period of time spanning several weeks. We base our study on country-wide datasets obtained from the Vodafone cellular network, containing both user-side activity measurements and ground truth satisfaction feedbacks. We describe the features that can be extracted from those datasets and we report on the prediction results obtained when using such features to train different machine learning models.

The remainder of this paper is structured as it follows: Section II summarises the related works in the area of QoE prediction in mobile cellular networks, while Section III describes the datasets under consideration. Section IV describes the task of coverage QoE prediction, commenting on the choice of the input features and the obtained results; Section V does the same for video QoE. Finally, Section VI provides a discussion on the obtained results and on future work directions.

II. Related Work

The problem of estimating the user QoE relative to different services (video streaming, web browsing, etc.) in cellular networks has been subject to increasing attention in the past few years. Most works focus on the QoE of video streams, mostly in form of unencrypted [2] or encrypted [3], [4], [5] YouTube contents. Generally, these works focus on short-term QoE, i.e., they estimate the QoE of individual video sessions starting from flow-level features extracted from each video network traffic traces, such as flow size and duration, average throughput and statistics on RTT and packet losses. QoE is obtained either directly through subjective user feedbacks in form of Mean Opinion Scores (MOS), or more often it is substituted by objective QoE metrics such as number of video stalls or buffering ratio [6], whose correlation to user satisfaction is well established [7]. QoE estimation is generally performed as a supervised classification task: when objective QoE metrics such as video stalls are used, they are quantised into discrete classes. As an example, in [8] and [3] the re-buffering ratio (duration of video stalls relative to duration of the video) is binarised using a threshold value of 0.1 [9]. In general, the reported accuracy of such short-term QoE prediction approaches is satisfactory, higher than 80% in most cases.

Differently from these works, here we focus on long-term QoE prediction, i.e., the satisfaction reported by a user is relative to a period of time spanning several weeks and not to an individual session. For what concerns coverage, several works in the past have studied the possibility of using crowdsourced measurements for predicting radio maps [10], [11]. However, to the best of our knowledge, such works only focus on objective coverage measurements and do not take into account the Quality of Experience of users.

III. Datasets

This work uses two country-wide datasets coming from one of the major european mobile operators: a user-side network measurements dataset and a ground-truth Quality of Experience dataset. Both datasets, in which users details are anonymised, are relative to a period of five months from May 2018 to November 2018 (excluding July and August to avoid summer seasonal biasing). The network measurements dataset contains data relative to roughly 500k users, while the QoE dataset contains data relative to roughly 30k users, as just a subset of users reporting network-related measurements actually answer the proposed QoE surveys.

A. User-side network measurements

The first dataset is obtained through a mobile application installed on a subset of the operator customers' equipments and running in background under their consent. The application periodically logs several active and passive network measurements relative to low-level network indicators (e.g., average cell signal strength and channel quality indicators, daily time spent by the user in full or limited service conditions, etc.) as well as application level indicators (e.g., session downlink/uplink data volume, duration and throughput) of different applications run in foreground by the user. We are interested in measurements relative to (i) network coverage and (ii) video streaming. Regarding the former, the following measurements are available for each day d and only for 4G Radio Access Technology (RAT):

- Daily Full Service Time, (f_d): the total time in seconds a user has reported full service in day d.
- Daily Limited Service Time, (l_d): the total time in seconds a user has reported limited service (emergency service only) during the day d.
- Daily No Service Time, (z_d): the total time in seconds a user has reported no service in day d.
- SNR Daily Minimum (s_d^{min}), Maximum (s_d^{max}) and Average (s_d^{avg}) in dB.
- RSRQ Daily Minimum (q_d^{min}), Maximum (q_d^{max}) and Average (q_d^{avg}).

The extraction of such counters allows to obtain a new dataset $\mathcal{N}_\mathcal{C}$ for coverage network measurements, containing entries of this form: $\{$user_id, date d, f_d, l_d, n_d, s_d^{min}, s_d^{max}, s_d^{avg}, q_d^{min}, q_d^{max}, $q_d^{\text{avg}}\}$.

For what concerns video, only the measurements relative to the YouTube mobile application are retained. Information are again sampled on a daily basis, this time considering both 3G and 4G access technology (i.e., RAT $\in \{$3G, 4G$\}$):

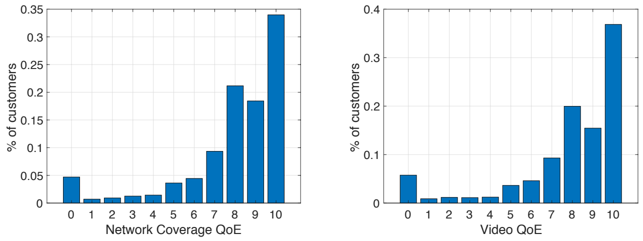
- Daily Download Time: (t_d^{RAT}): the total time in seconds a user has downloaded video data using either 3G or 4G radio access technology.
- Daily Download Volume (v_d^{RAT}): the total video data volume a user has downloaded using either 3G or 4G.
- Daily Maximum Data Session Volume (w_d^{RAT}): the maximum video data volume downloaded in a single session using either 4G or 3G.
- Daily Maximum Data Session Throughput Peak (p_d^{RAT}): the maximum throughput experienced in a single video session using either 3G or 4G.

We construct a new dataset $\mathcal{N}_\mathcal{V}$ containing only video network measurements, where entries have this form: $\{$user_id, date d, t_d^{3G}, t_d^{4G}, v_d^{3G}, v_d^{4G}, d_d^{3G}, d_d^{4G}, w_d^{3G}, w_d^{4G}, $p_d^{3G}, p_d^{4G}\}$.

B. QoE dataset

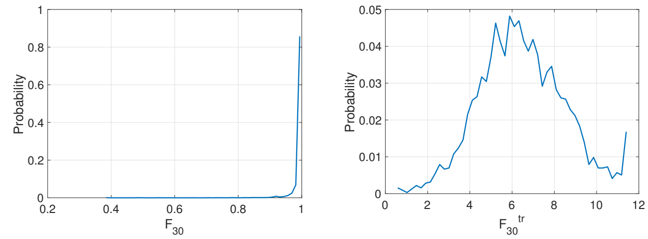
The second dataset contains ground truth user feedbacks on QoE relative to different network services, collected by the operator through individual surveys. The feedbacks are reported in form of satisfaction grades on a scale from 0 to 10. We extract from this dataset only the QoE feedbacks relative to network coverage and video streaming and create two distinct datasets $\mathcal{Q}_\mathcal{C}$ and $\mathcal{Q}_\mathcal{V}$: considering the five months period of analysis, $\mathcal{Q}_\mathcal{C}$ contains 7045 QoE feedbacks for coverage and $\mathcal{Q}_\mathcal{V}$ contains 6264 feedbacks for video streaming. Each entry has the following form: $\{$user_id, date, QoE$\}$.

Figure 1 shows the distribution of the QoE for the two services. As one can see, both distributions are highly skewed, with the



(a) (b)

Fig. 1. Distribution of user QoE feedbacks for (a) network coverage and (b) video streaming.



(a) (b)

Fig. 2. Probability density function of F_{30} and F_{30}^{tr}

majority of users reporting positive feedbacks. It is possible to discretise the QoE votes into two classes, with respect to a predefined threshold $\mathcal{T}$: users whose vote is less or equal than $\mathcal{T}$ are grouped together as *Unsatisfied* users, while the opposite happens for *Satisfied* users. As an example, the percentage of users unsatisfied with network coverage is 22% when $\mathcal{T} = 6$. It is not trivial to decide which threshold value should be used: in the following we show results for different values $\mathcal{T}$.

IV. Network Coverage QoE

First, we focus on predicting the network coverage QoE of a user. We take a supervised learning approach, leveraging the network measurements in dataset $\mathcal{N}_\mathcal{C}$ and QoE feedbacks in dataset $\mathcal{Q}_\mathcal{C}$. We are interested only in those users that appear in both datasets, i.e. we consider network measurements of those users having issued a QoE feedback on coverage in the five-months period of interest: these are limited to 4680 users (i.e., roughly 15% of the total available feedbacks).

A. Feature computation

As a first step, we engineer features from the daily measurements in $\mathcal{N}_\mathcal{C}$. We start with the assumption that a user's QoE feedback at day d depends on its experience in the previous n days $d-1, d-2, \ldots, d-n$. A first question is how to dimension n, which controls the memory of a user. Small values of n assume that the QoE feedback depends only on the short-term activity (i.e., what happened in the days closest to the answer), while large values of n assume longer-terms correlations. Instead of making a strong choice on this parameter, we compute features for all values of n in the range $1 \ldots 30$ (i.e., we assume that the maximum user memory is one month) and we let the learning model select the best inputs. We assume that the QoE on coverage depends on the fraction of time that the user has passed in full, limited or no service as well as the signal quality observed by the user during those days. Therefore, we first compute for each user the *Cumulative Full Service Time Ratio*, F_n as:

$$F_n = \frac{\sum_{i=d-n}^{d} f_i}{\sum_{i=d-n}^{d} f_i + l_i + n_i} \tag{1}$$

Similarly, we compute the *Cumulative Limited Service Time Ratio* (L_n) and the *Cumulative No Service Time Ratio* (Z_n) changing the numerator in (1) with l_i or z_i, respectively. Note that $F_n + L_n + Z_n = 1$, which means that one out of the three features can be excluded from the model as linearly dependent from the other two, for each selected user memory length n. In the following, we will just consider F_n and Z_n. This process creates $30 \times 2 = 60$ feature per user.

For what concerns channel measurements, we compute the *Minimum of Daily Minima* (S_n^{min}), *Maximum of Daily Maxima* (S_n^{max}) and *Average of Daily Averages* (S_n^{avg}) of SNR as:

$$S_n^{\text{min}} = \min_{i=d-n}^{d} s_i^{\text{min}}, \tag{2}$$

$$S_n^{\text{max}} = \max_{i=d-n}^{d} s_i^{\text{max}}, \tag{3}$$

$$S_n^{\text{avg}} = \frac{\sum_{i=d-n}^{d} s_i^{\text{avg}}}{n}. \tag{4}$$

Similarly, we compute the same triplets for RSRQ measurements (Q_n^{min},Q_n^{avg} and Q_n^{max}). This process creates additional $30 \times 3 \times 2 = 180$ features per user, for a total of 240 features per user.

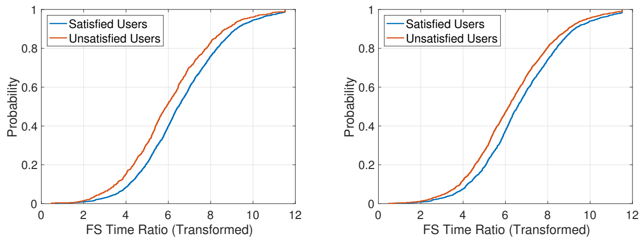
It is important to check the statistical distribution of each computed feature, as a large portion of machine learning methods assume that input features are characterised by a Gaussian distribution. We observe that the channel measurements features are already Gaussian distributed, while this is not true for the service time ratios. As an example, Figure 2 shows the distribution of FR_{30}; as one can see, the distribution is not Gaussian since the majority of users have reported full time service ratio close to 1. To make the data distribution more similar to Gaussian we apply a log-like transformation as follows:

$$F_n^{\text{tr}} = -\log(1 - F_n). \tag{5}$$

The corresponding transformed distribution is shown in Figure 2(b), which now looks more similar to a Gaussian bell. A similar transformation is also applied for Z_n.

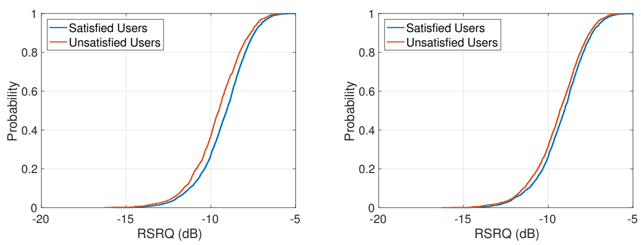
B. Threshold selection

It is worth observing some of the class-conditional cumulative density functions (CDF) of such features for different values of the threshold $\mathcal{T}$. Figures 3 and 4 show the class-conditional CDFs of F_{30}^{tr} and Q_{30}^{avg} for $\mathcal{T} = \{6, 8\}$: as one can see, satisfied users (blue curves) have more likely experienced longer full service periods than unsatisfied users. Considering $\mathcal{T} = 6$ we can see that almost 98% of satisfied users had a



(a) (b)

Fig. 3. Class-conditional CDF of F_{30}^{tr} at threshold $\mathcal{T} = 6, 8$



(a) (b)

Fig. 4. Class-conditional CDF of Q_{30}^{avg} at threshold $\mathcal{T} = 6, 8$

fraction of full service time greater than 95% (corresponding to $F_{30}^{\mathrm{tr}} = 3$), while this is true only for 93% of unsatisfied users. For what concerns signal quality measurements, the CDFs in Figure 4 show that satisfied users have higher median values of RSRQ of about 0.5 dB. In general, it can be seen that $\mathcal{T} = 6$ maximises the difference in the class-conditional CDF compared to other thresholds. The same result is observed on the CDFs of other features, which are not shown here for space limits.

C. *Prediction results*

The features computed in Section IV-A are used as input to several different supervised machine learning models. We distinguish here two separate cases: in the first one we use as input only those features relative to the service time (i.e. F_n and Z_n). We refer to this case as (*ST-Only*). In the second case we also add as input the features obtained starting from SNR and RSRQ signal quality measurements: we refer to this case as (*ST+SQ*). In both cases, the number of useful observations is 4680. At the selected threshold $\mathcal{T} = 6$, 1030 observations belong to the *Unsatisfied* class and the rest to *Satisfied* class. The features and the corresponding QoE ground truth values are input to the following supervised classifiers: i) Regularised Logistic Regressor (RLR), ii) Gaussian Naive Bayes (GNB), iii) Decision Trees (DT), iv) Random Forest (RF) and v) AdaBoost (AB), a more complex classifier following the paradigm of *ensemble* learning (with boosting) which uses decision trees as first level learners. All classifiers but (GNB) accept in input different hyper-parameters whose setup is not trivial and needs to be optimised. As an example, regularised logistic regression requires to determine the regularisation coefficient used to penalise features and reduce overfitting.

TABLE I
PERFORMANCE OBTAINED FOR COVERAGE QOE PREDICTION

Case	Classifier	AUC	Thr	TPR	FPR
ST-only	RLR	0.57	0.200	0.48	0.35
	GNB	0.57	0.085	0.58	0.46
	DT	0.55	0.214	0.51	0.44
	RF	0.57	0.177	0.52	0.41
	AB	0.57	0.314	0.56	0.46
ST+SQ	**RLR**	**0.60**	**0.364**	**0.54**	**0.36**
	GNB	0.60	0.098	0.61	0.46
	DT	0.56	0.237	0.53	0.42
	Rf	0.60	0.229	0.56	0.41
	AB	0.60	0.398	0.58	0.42

Similarly, tree-based classifier (DT, RF and AB) require to set parameters such as the maximum depth of the trees and the tree splitting criterion. To tune such hyper-parameters, we proceed in the following way. First, the original dataset of 4680 observations and ground truth pairs is divided into five folds with splitting ratio 80% (Training set) and a 20% (Validation set) according to k-fold cross-validation. Each Training Set is sub-divided again in 5-folds with the same splitting ratios for hyper-parameters tuning. The best hyper-parameters are chosen through a grid search on a set of candidate values. This inner-loop prevents the model from overfitting the data with respect to hyper-parameters values. Once such parameters are set, they are used to train each model in the outer cross-validation loop. Prediction results are evaluated on the Validation Sets, preventing overfitting and providing more generalisable the results. For each observation in the Validation Set, the tested classifiers output the probability that the observation belongs to the *Satisfied* or *Unsatisfied* class. By thresholding such probabilities with different values, one can compute the so called Receiver Operating Characteristic (ROC) curve, which shows the values of the True Positive Rate (TPR) and False Positive Rate (FPR) obtained by the particular classifier. The TPR is defined as the fraction of correctly detected *Unsatisfied* users, while the FPR is the fraction of *Satisfied* users which are incorrectly labeled as *Unsatisfied*. Additionally, to summarise in a single value the performance of each classifier, the Area Under the Curve (AUC) is computed. Figure 5 and 6 show the ROC curves of the different classifiers for the *ST-Only* and *ST+SQ* cases, respectively, while Table I reports the corresponding AUC values. Additionally, we also report in the Table the point on each classifier's ROC curve closest to the upper left corner (FPR=0, TPR=1), which corresponds to an ideal condition. The results obtained show that all classifiers perform at par and, unfortunately, quite poorly, with a maximum achievable AUC of 0.6. In general, adding signal quality features as input improves the classification task by 3%. Looking at the best working points in Table I, we can see that regularised logistic regression correctly detects 54% of the *Unsatisfied* users, with a corresponding false alarm rate of 36%.

V. VIDEO STREAMING QOE

Beside network coverage, we focus also on predicting users video streaming QoE joining the two datasets $\mathcal{N}_{\mathcal{V}}$ and $\mathcal{Q}_{\mathcal{V}}$. In

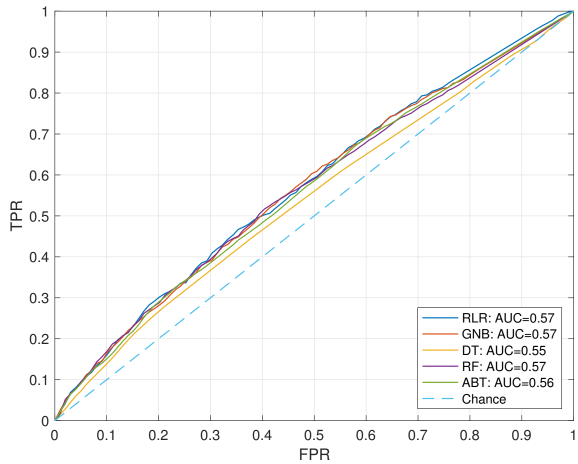


Fig. 5. ROC curve for coverage QoE prediction, case *ST-Only*, $\mathcal{T} = 6$

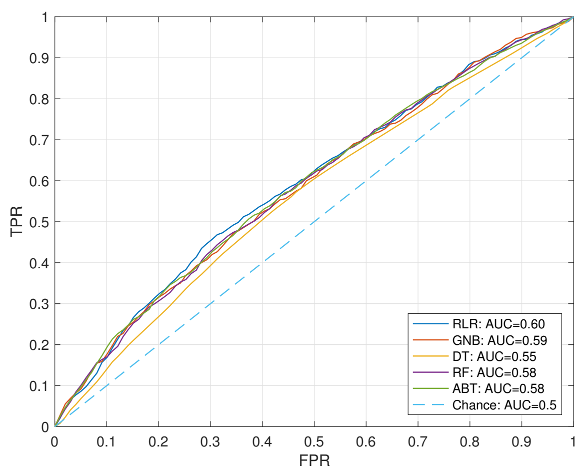
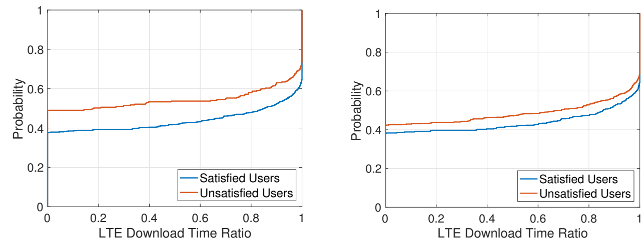
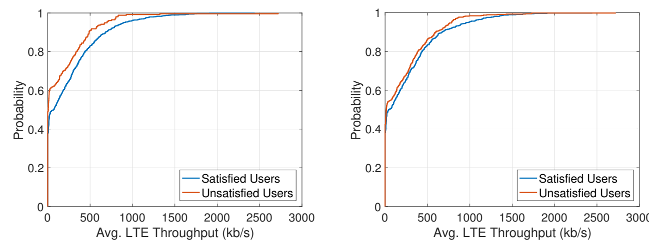


Fig. 6. ROC curve for coverage QoE prediction, case *ST+CI*, $\mathcal{T} = 6$



(a) (b)

Fig. 7. Class-conditional CDFs of R_{30}^{4G} for $\mathcal{T} = 6, 8$



(a) (b)

Fig. 8. Class-conditional CDFs of G_n^{4G} for $\mathcal{T} = 6, 8$

this case, the number of users appearing in both datasets is limited to 1140 (i.e., less than 5% of the total feedbacks).

A. Features computation

As done for the coverage QoE, we start by engineering features from the daily measurements in $\mathcal{N}_\mathcal{V}$ assuming that the user video experience at day d is somehow correlated with RAT-dependent features (3G/4G) in the previous n days. To give an example, it is reasonable to conjecture that a user that watched videos under 4G reports higher levels of satisfaction compared to a 3G-only user, since higher throughputs can be achieved with the former technology. Therefore, for a given user memory of length n, we compute:

- *Cumulative Download Time and Volume in 3G or 4G*:

$$T_n^{\text{RAT}} = \sum_{i=d-n}^{d} t_i^{\text{RAT}}, \quad V_n^{\text{RAT}} = \sum_{i=d-n}^{d} v_i^{\text{RAT}} \tag{6}$$

- *Average Throughput in 3G or 4G*:

$$G_n^{\text{RAT}} = \frac{V_n^{\text{RAT}}}{T_n^{\text{RAT}}} \tag{7}$$

- *Overall Average Throughput*:

$$A_n = \frac{V_n^{4G} + V_n^{3G}}{T_n^{4G} + T_n^{3G}} \tag{8}$$

- *Overall maximum of Data Session Volumes and Throughput Peak*:

$$W_n^{\text{RAT}} = \max_{j=i-n}^{i} w_j^{\text{RAT}}, \quad P_n^{\text{RAT}} = \max_{j=i-n}^{i} p_j^{\text{RAT}}, \tag{9}$$

- *Cumulative Download Time Ratio in 4G*:

$$R_n^{4G} = \frac{T_n^{4G}}{T_n^{3G} + T_n^{4G}}. \tag{10}$$

We don't consider the cumulative download time ratio in 3G RAT since it is collinear with R_n^{4G} (i.e., $R_n^{3G} + R_n^{4G} = 1$) and thus does not add any additional information. Computing such features for $n = 1 \dots 30$, we end up with 360 features per user for what concerns video measurements.

B. Threshold Selection

As done for coverage, it is worth observing the class-conditional CDFs of the computed features in order to assess the optimal value of the threshold $\mathcal{T}$.

Figures 7 and 8 show the class-conditional CDFs for $\mathcal{T} = 6$ and 8 of R_n^{4G} and G_n^{4G}. We observe again that $\mathcal{T} = 6$ maximises the difference in the class-conditional CDFs. For that threshold, we observe in Figure 7 that the median 4G download time fraction for video is just above 20% of the total download time, while for satisfied users it is almost 85%.

TABLE II
PERFORMANCE OBTAINED FOR VIDEO QOE PREDICTION

Classifier	AUC	Threshold	TPR	FPR
RLR	0.57	0.435	0.64	0.52
GNB	0.57	0.944	0.51	0.37
DT	0.55	0.2985	0.55	0.45
RF	**0.58**	**0.244**	**0.57**	**0.43**
AB	0.57	0.366	0.61	0.48

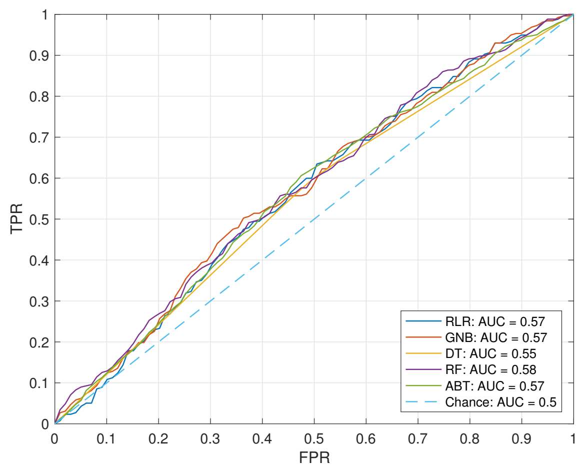


Fig. 9. ROC curve for video streaming QoE prediction, $\mathcal{T} = 6$

Similarly, observing Figure 8, 95% of the unsatisfied users experienced an average throughput less than 750 kbps, while it is above 1 Mbps for the same percentage of satisfied users.

C. Results - Video QoE

We adopt the same workflow described in Section V-C to compute the prediction performance of different learning models. Figure 9 shows the ROC curves of the tested classifier and Table II reports the corresponding AUC and best identified working points. Again, it can be seen that the different classifiers perform almost the same: the best performing classifier is the Random Forest one, which scores an AUC value of 0.58. The best working point identified is at 57% of correctly identified *Satisfied* user, with a corresponding False Positive Ratio of 43%.

VI. DISCUSSION AND CONCLUDING REMARKS

In this work we have commented on the possibility of predicting the long-term coverage and video QoE starting from user-side network measurements. The results obtained demonstrate that the task is complex and challenging, as all the different supervised machine learning classifiers used show quite poor performances. Nonetheless, a weak correlation between the engineered input features and the QoE user feedbacks could be exploited. Some interesting points can be made:

- Compared to short-term QoE estimation, long-term QoE prediction looks like a much more challenging task. The most direct explanation for this could lie in the way users reply to survey, which could be affected by many factors (e.g., value for money or other user-dependent standards) that network measurements alone cannot capture.
- Despite the availability of a country-wide dataset spanning several months, the actual number of ground truth observations we could use in this work was quite limited (15% of the total feedbacks for coverage and less than 5% for video), also compared to the number of features extracted. It is well known that data availability can greatly improve the performance of supervised machine learning methods: incentive strategies could be put in place by operators to retrieve as much data possible from their customers.
- Finally, we recall that one of the primary use of QoE prediction is to identify areas of the network or network elements with possible problems. Since each item is visited by many users, each one reporting a ground truth or predicted QoE value, it may be possible that misclassification errors are somehow alleviated when grouped on a single network element/area. The precise impact of individual prediction errors on the overall task of network problems detection considering realistic spatial distributions for the users is subject to ongoing work.

REFERENCES

[1] V. Cisco Mobile, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper," 2017.
[2] T. Mangla, E. Halepovic, M. Ammar, and E. Zegura, "Mimic: Using passive network measurements to estimate http-based adaptive video qoe metrics," in *Network Traffic Measurement and Analysis Conference (TMA), 2017*. IEEE, 2017, pp. 1–6.
[3] G. Dimopoulos, I. Leontiadis, P. Barlet-Ros, and K. Papagiannaki, "Measuring video qoe from encrypted traffic," in *Proceedings of the 2016 Internet Measurement Conference*. ACM, 2016, pp. 513–526.
[4] I. Orsolic, D. Pevec, M. Suznjevic, and L. Skorin-Kapov, "A machine learning approach to classifying youtube qoe based on encrypted network traffic," *Multimedia tools and applications*, vol. 76, no. 21, pp. 22 267–22 301, 2017.
[5] T. Mangla, E. Halepovic, M. Ammar, and E. Zegura, "emimic: Estimating http-based video qoe metrics from encrypted network traffic," in *2018 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2018, pp. 1–8.
[6] S. Moteau, F. Guillemin, and T. Houdoin, "Correlation between qos and qoe for http youtube content in orange cellular networks," in *Communications (LATINCOM), 2017 IEEE 9th Latin-American Conference on*. IEEE, 2017, pp. 1–6.
[7] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "A quest for an internet video quality-of-experience metric," in *Proceedings of the 11th ACM workshop on hot topics in networks*. ACM, 2012, pp. 97–102.
[8] V. Aggarwal, E. Halepovic, J. Pang, S. Venkataraman, and H. Yan, "Prometheus: Toward quality-of-experience estimation for mobile apps from passive network measurements," in *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*. ACM, 2014, p. 18.
[9] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 6, pp. 2001–2014, 2013.
[10] M. Molinari, M.-R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," in *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*. ACM, 2015, pp. 33–38.
[11] A. Lutu, Y. R. Siwakoti, Ö. Alay, D. Baltrūnas, and A. Elmokashfi, "The good, the bad and the implications of profiling mobile broadband coverage," *Computer Networks*, vol. 107, pp. 76–93, 2016.