Towards Long-Term Coverage and Video Users Satisfaction 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 satisfaction of their customers to minimise the churn rate: however, collecting user feedbacks through surveys is a cumbersome task. In this work we explore the possibility of predicting the long-term user satisfaction 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 user satisfaction prediction from network measurements is a very challenging task: we therefore point out possible action points to be implemented to improve the prediction results.**

***Index Terms*—User satisfaction prediction, Cellular Networks**

I. Introduction

According to recent Cisco estimates, by 2021 mobile cellular networks will connect more than 11 billion mobile devices and will be responsible for more than one fifth of the total IP traffic generated worldwide [1]. 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 general questions: as an example the Net Promoter Score (NPS) survey asks users 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 user satisfaction relative to 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. 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 satisfaction of their customers starting from network measurements only.

In this paper we explore this possibility and predict the customer satisfaction relative to two cellular network aspects: *network coverage* and *video streaming*. The former is a necessary service for every user: 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. The satisfactions relative to the two aspects play a big role in a customer decision to leave its current operator for a better one. Differently from related works in the area analyzing short-term Quality-of-Experience (QoE) of different network services, we focus here on the long-term satisfaction, 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 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 user satisfaction prediction in mobile cellular networks, while Section III describes the datasets under consideration. Section IV describes the task of

ISBN 978-3-903176-18-8 2019 IFIP ©2019 IEEE

coverage satisfaction prediction, commenting on the choice of the input features and the obtained results; Section V does the same for video streaming services. Finally, Section VI provides a discussion on the obtained results.

II. Related Work

The problem of estimating the user satisfaction 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 Quality of Experience (QoE) of a user accessing 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. Network data is gathered either from network-side traffic traces or directly by user terminals [6]. 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 [7], the downlink bandwidth or the access Round Trip Time [8], whose correlation to user satisfaction is well established [9]. 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 [10] and [3] the rebuffering ratio (duration of video stalls relative to duration of the video) is binarised using a threshold value of 0.1 [11]. In general, the reported accuracy of such short-term video QoE prediction approaches is satisfactory, higher than 80% in most cases. Other works focus on web browsing QoE: in [12] authors use as QoE metrics the web session length and the website abandonment rate. Results suggest that the web QoE is very sensitive to inter-radio-access-technology (IRAT) handovers and signal-to-noise-ratio. Differently from these works, here we focus on long-term satisfaction 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: such measure may capture more easily than short-term QoE users and in turn areas of the network which impact most on the churn rate. For what concerns coverage, several works in the past have studied the possibility of using crowdsourced measurements for predicting radio maps [13], [14]. However, to the best of our knowledge, such works only focus on objective coverage measurements and do not take into account the user satisfaction.

III. Datasets

This work uses two country-wide datasets coming from Vodafone, one of the major european mobile operators: a user-side network measurements dataset and a ground-truth user satisfaction 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 satisfactions dataset contains data relative to roughly 30k users, as just a subset of users reporting network-related measurements actually answer the proposed satisfaction surveys.

A. User-side Network Measurements

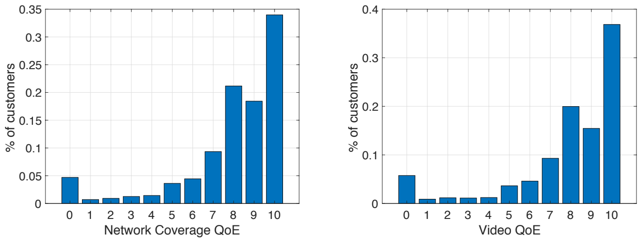
The first dataset is obtained through a Vodafone-branded mobile application installed on a subset of the operator customers' equipments and running in background under their consent. The application periodically logs several active and passive network measurements relative to low-level network indicators (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. Beside the measurement itself, the application provides also additional information, such as the measurement timestamp or the ID and location of the base station to which the user is connected. 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.
- Signal to Noise Ratio (SNR) Daily Minimum (s_d^{min}), Maximum (s_d^{max}) and Average (s_d^{avg}) in dB.
- Reference Signal Received Quality (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 $\mathcal{N}_C$ 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, we focus on network measurements strictly related to YouTube mobile sessions. Information are again sampled on a daily basis, this time considering both 3G and 4G access technology (i.e., RAT $\in$ $\{$3G, 4G$\}$), as follows:

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



(a) (b)

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

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. User Satisfaction Dataset

The second dataset contains ground truth feedbacks of a subset of the operator customers on their satisfaction relative to different aspects of the received service (e.g. network coverage, video streaming quality, voice quality, data speed, etc.). The feedbacks, collected by the operator through individual surveys, are reported in form of satisfaction grades on a scale from 0 (*fully dissatisfied user*) to 10 (*fully satisfied user*), and each user gives a different answer for each investigated network item. We extract from this dataset only the feedbacks relative tonetwork coverage and video streaming, creating 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 survey responses for coverage and $\mathcal{Q}_\mathcal{V}$ contains 6264 survey responses for video streaming, where an entry of any of the two datasets has the following form: {user_id, date, QoE}. Note that, out of the initial 30k survey responses available in the dataset, only about 44% of those are actually related to coverage and video services. Fig. 1 shows the distribution of satisfaction grades for the two considered services. As one can see, both distributions are highly skewed, with the majority of users reporting positive feedbacks. It is possible to discretise the grades into two classes, with respect to a predefined satisfaction 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 roughly 19% 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 of $\mathcal{T}$.

IV. Network Coverage

First, we focus on predicting the satisfaction of a user relatively to the experienced network coverage. We take a supervised learning approach, leveraging the network measurements in dataset $\mathcal{N}_\mathcal{C}$ and satisfaction grades 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 satisfaction grade on coverage in the five-months period of interest. These are limited to 4680 users (i.e., roughly 15% of the total available survey responses).

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 satisfaction feedback issued at day d (i.e. day d is the survey response date) 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 users' satisfaction depends only on their short-term activity (i.e., what happened in the days closest to the survey response date), while large values of n assume longer-terms correlations. Therefore, 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 satisfaction on coverage depends on the fraction of time that the user has passed in full, limited or no service as well as on 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.

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, Fig. 2 shows the distribution of F_{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 Fig. 2(b), which now looks more similar to a Gaussian bell. A similar transformation is also applied for Z_n. Distributions of other service time features are not shown for space limits. We observe that such a transformation gives benefits in terms of features correlation with users' satisfaction. Figure 3 shows the Correlation Relative Improvement (CRI) we get from F_n^{tr} with respect to F_n, for n in the range $1 \ldots 30$ days. For a given user memory length n, CRI is computed as follows:

$$CRI_n = \frac{corr(F_n^{\text{tr}}, y) - corr(F_n, y)}{corr(F_n, y)} \qquad (6)$$

where y is the vector of users' satisfaction votes and *corr(a,b)* refers to the Pearson's correlation coefficient between a and b. As one can observe, correlation increases due to logarithmic transformation for each considered user memory length, with an average CRI of 2 (i.e. on average correlations after transformation are three times larger). Generally speaking, gains are higher for longer-term look-up periods, where the improvement peak is reached by F_{26}^{tr} which is almost 10 times more correlated with users' satisfaction than F_{26}. Similar considerations can be done for Z_n^{tr} with respect to Z_n, this showing that users' satisfaction depends more likely on their long-term rather than short term activity.

Beside network measurements related features, we believe it is worth to add to the model some features related to user location, as in principle the satisfaction of a user regarding a network service may depend on the geographical area where the service is most frequently experienced by that user. With this aim, for a given user we define $\mathcal{U}_i$ as the set of base stations i visited within a period of 30 days preceding the user's survey response date. Also, we refer to $Lat(i)$ and $Long(i)$ as the latitude and longitude of base station i, respectively. Moreover, we compute the Active Time $AT(i)$ of a given user with respect to base station i as follows:

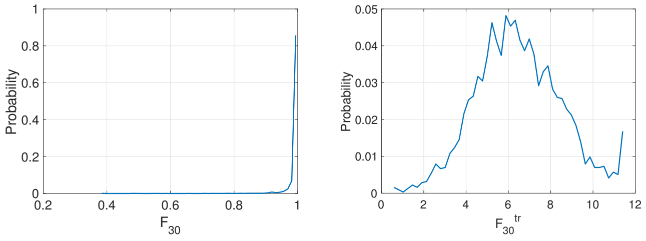
$$AT(i) = F_{30}^i + L_{30}^i + Z_{30}^i \qquad (7)$$

which is non-zero if and only if base station i belongs to $\mathcal{U}_i$ for that user. Note that superscript i on service time indicators refers to the fact that to compute $AT(i)$ we consider the total service time experienced by the user with respect to base station i only. Finally, we compute the *Average Latitude* (LAT_{avg}) and *Average Longitude* ($LONG_{avg}$) of a user as follows:

$$LAT_{avg} = \frac{\sum_{i \in \mathcal{U}_i} AT(i) \cdot Lat(i)}{\sum_{i \in \mathcal{U}_i} AT(i)}, \qquad (8)$$

$$LONG_{avg} = \frac{\sum_{i \in \mathcal{U}_i} AT(i) \cdot Long(i)}{\sum_{i \in \mathcal{U}_i} AT(i)}. \qquad (9)$$

which correspond to the weighted average of the latitude and longitude of the base stations visited by the user, where the weights correspond to the base stations visit times of the user itself. Adding these two features to the model we get a total of $60 + 180 + 2 = 242$ features per user.



(a) (b)

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

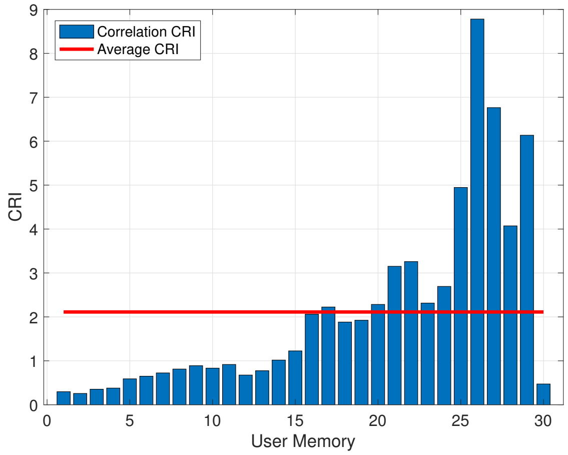
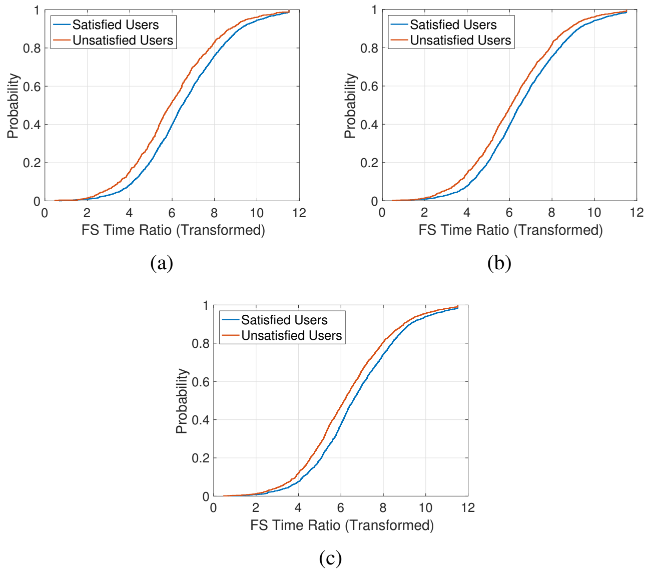


Fig. 3. Correlation Relative Improvement for Full Service time features.

B. Satisfaction Threshold Selection

In this section we comment about the tuning of the Satisfaction Threshold $\mathcal{T}$. On the one hand, satisfaction thresholds lower than 6 (e.g. 5 or 4) would include all those users giving a vote equal to 6 (or 5) in the class of *Satisfied* users, even though such a vote is far from the maximum possible (i.e. 10). On the other hand, considering a satisfaction threshold equal to 9 would include in the same class only those users giving a 10, which is a too severe approach. Therefore, we decide to look for the best $\mathcal{T}$ in the range $6, 7, 8$. The aim is to tune $\mathcal{T}$ such to let the classifier extract as much information as possible from the features, conditioned to the two classes of Unsatisfied and Satisfied users. With this in mind, it is worth observing some of the class-conditional Cumulative Distribution Functions (CDFs) of the above described features for the three considered values of $\mathcal{T}$. Figures 4 and 5 show the class-conditional CDFs of F_{30}^{tr} and Q_{30}^{avg} for $\mathcal{T} = \{6, 7, 8\}$. Generally speaking, the larger the gap between the red and blue curves, the more the information associated to the features is conditioned to the observation samples' class label. As one can see, Figures 4 shows that satisfied users (blue curves) have more likely experienced longer full service periods than unsatisfied users, disregarding the satisfaction threshold. For instance, considering $\mathcal{T} = 6$ we can see that almost 98% of satisfied users had a fraction of full service time greater than

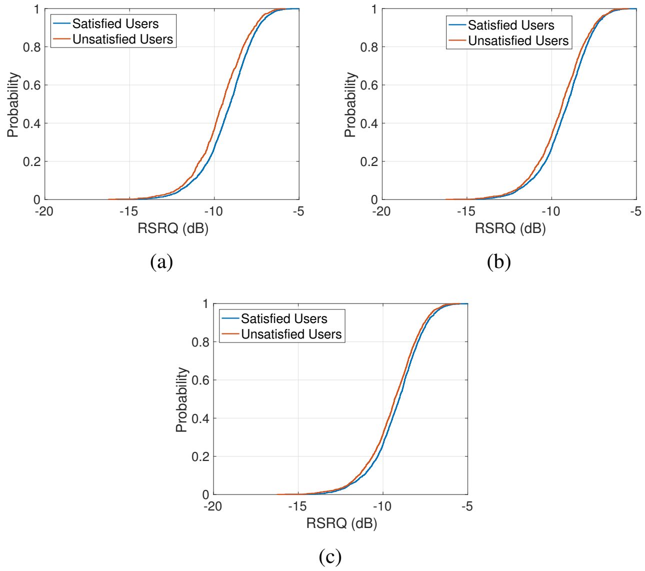


(a)

(b)

(c)

Fig. 4. Class-conditional CDF of F_{30}^{tr} at threshold $\mathcal{T} =$ (a) 6, (b) 7, (c) 8.



(a)

(b)

(c)

Fig. 5. Class-conditional CDF of Q_{30}^{avg} at threshold $\mathcal{T} =$ (a) 6, (b) 7, (c) 8.

95% (corresponding to $F_{30}^{\text{tr}} = 3$), while this is true only for 93% of unsatisfied users. Similarly, for what concerns signal quality measurements, the CDFs in Fig. 5 show that satisfied users have higher median values of RSRQ of about 0.5 dB. In general we observe that $\mathcal{T} = 6$ maximises the difference in the class-conditional CDFs compared to other thresholds, both for F_{30}^{tr} and Q_{30}^{avg}. The same result is observed from the CDFs of the other features, which are not shown here.

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 times (i.e. F_n and Z_n) and user locations. 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 satisfaction ground truth votes 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), v) Linear Discriminant Analysis (LDA), vi) AdaBoost (a more complex classifier following the paradigm of *ensemble* learning with boosting which uses decision trees as first level learners, AB) and vii) Multi Layer Perceptron (a neural network, in this study with a single hidden layer, NN). All classifiers but GNB and LDA 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. Similarly, tree-based classifiers (DT, RF and AB) require to set parameters such as the maximum depth of the trees and the tree splitting criterion. Also, NN classifier requires to optimize the hidden layer structure, by tuning the number of neurons. To tune such hyper-parameters, we proceed with a *grid search* on a set of candidate values as follows. First, according to a k-fold cross-validation strategy with $k = 5$, the original dataset of 4680 observations and ground truth pairs is divided into five folds with splitting ratios 80% (*Training set*) and a 20% (*Test set*). Secondly, we focus on a given pair of Training Set and Test Set. We apply to the Training Set a further 5-fold cross-validation, such that it is divided into five folds with splitting ratios 80% (*Sub-Training Set*) and 20% (*Validation Set*). Each Sub-Training Set is then trained with each classifier's hyper-parameters candidate values. Prediction performances are then evaluated on the corresponding Validation Set. At end of this (*inner*) cross-validation process, we can select the classifiers' best hyper-parameters (i.e. those maximising, per each classifier, the prediction results on the Validation Set) that are used to train each model in the outer cross-validation loop (i.e. the original Training Set). Finally, the trained models prediction performances are tested on the Test Set. Note that this procedure is repeated 5 times, one per each Training Set selected by the outer cross-validation loop. The results we will show correspond to the average results across the different Test Sets. In particular, for each observation of a given Test Set, the tested classifiers output the probability that the observation belongs to the *Unsatisfied* class. By thresholding such probability with different Prediction Thresholds (PT), 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

TABLE I

PERFORMANCE OBTAINED FOR COVERAGE QOE PREDICTION

Case	Classifier	AUC	PT	TPR	FPR
ST-only	RLR	0.58	0.195	0.59	0.46
	GNB	0.57	0.136	0.54	0.42
	DT	0.55	0.185	0.53	0.43
	RF	0.57	0.178	0.54	0.42
	AB	0.59	0.347	0.55	0.42
	LDA	0.58	0.192	0.53	0.41
	NN	0.57	0.192	0.54	0.42
ST+SQ	**RLR**	**0.60**	**0.215**	**0.59**	**0.44**
	GNB	0.59	0.131	0.61	0.46
	DT	0.57	0.181	0.52	0.40
	RF	0.60	0.181	0.56	0.40
	AB	0.60	0.414	0.58	0.42
	LDA	0.59	0.175	0.63	0.49
	NN	0.59	0.192	0.56	0.44

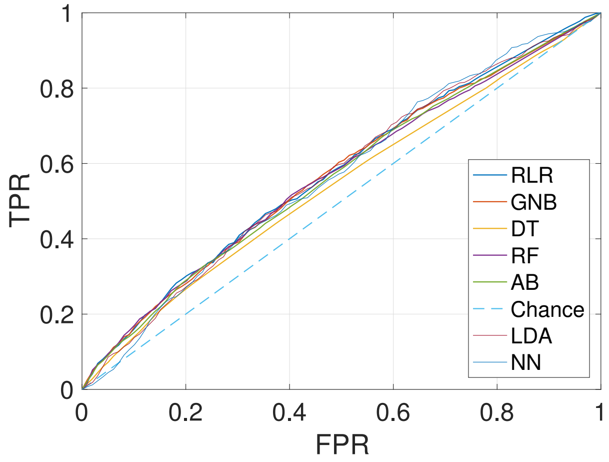


Fig. 6. ROC curve for coverage QoE prediction, case *ST-Only*, $\mathcal{T} = 6$.

value the performances of each classifier, the Area Under the Curve (AUC) is computed. Note that, for a random classifier, the AUC equals 0.5.

Figures 6 and 7 show the ROC curves of the different classifiers for the cases ST-Only and ST+SQ, 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], corresponding to an ideal condition of perfect prediction), by giving the corresponding values of TPR and FPR at the corresponding PT. 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 RLR correctly detects 59% of the *Unsatisfied* users, with a corresponding false alarm rate of 44%. Also, optimal PTs are usually lower than 0.5.

V. VIDEO STREAMING

Beside network coverage, we focus also on predicting users satisfaction on the quality of the experienced YouTube streaming sessions, joining the two datasets $\mathcal{N}_{\mathcal{V}}$ and $\mathcal{Q}_{\mathcal{V}}$. In this case, the number of users appearing in both datasets is limited to 1140 (i.e., less than 5% of the total feedbacks).

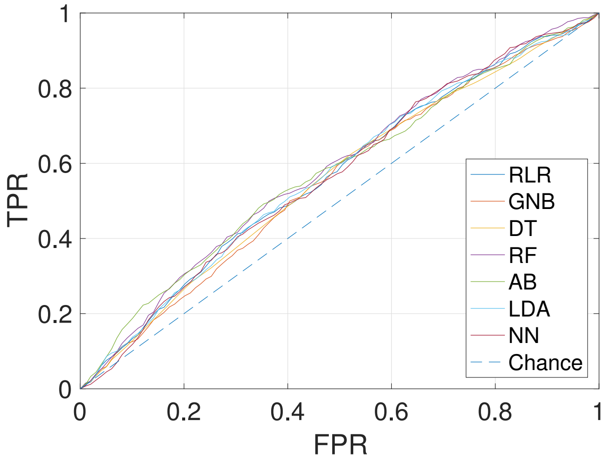


Fig. 7. ROC curve for coverage QoE prediction, case *ST+CI*, $\mathcal{T} = 6$.

A. Features Computation

As done in section IV, we start by engineering features from the daily measurements in $\mathcal{N}_{\mathcal{V}}$. We assume that the satisfaction of a user regarding video quality reported at day d is somehow correlated with RAT-dependent features (3G/4G) analysed within 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*:

$$T_n^{\text{RAT}} = \sum_{i=d-n}^{d} t_i^{\text{RAT}}, \quad V_n^{\text{RAT}} = \sum_{i=d-n}^{d} v_i^{\text{RAT}} \qquad (10)$$

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

$$G_n^{\text{RAT}} = \frac{V_n^{\text{RAT}}}{T_n^{\text{RAT}}} \qquad (11)$$

- *Overall Average Throughput*:

$$A_n = \frac{V_n^{4G} + V_n^{3G}}{T_n^{4G} + T_n^{3G}} \qquad (12)$$

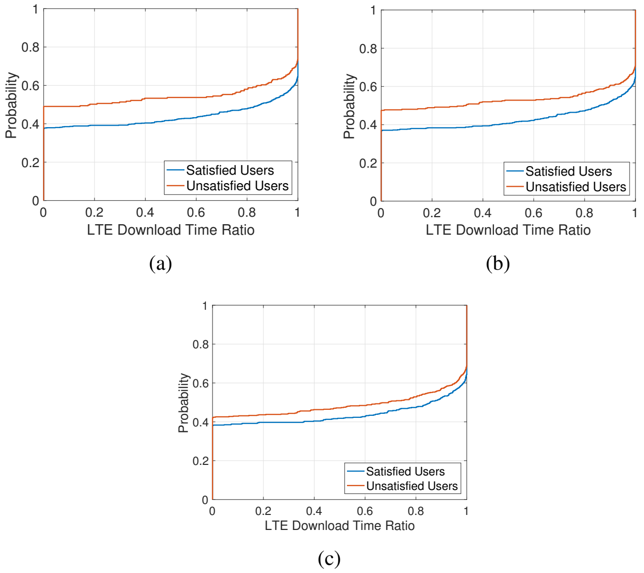
- *Overall maximum of Data Session Volumes and Throughput Peak*:

$$W_n^{\text{RAT}} = \max_{j=i-n}^{i} w_j^{\text{RAT}}, \quad P_n^{\text{RAT}} = \max_{j=i-n}^{i} p_j^{\text{RAT}}, \qquad (13)$$

- *Cumulative Download Time Ratio in 4G*:

$$R_n^{4\text{G}} = \frac{T_n^{4\text{G}}}{T_n^{3\text{G}} + T_n^{4\text{G}}}. \qquad (14)$$

We don't consider the cumulative download time ratio in 3G RAT since it is collinear with $R_n^{4\text{G}}$ (i.e., $R_n^{3\text{G}} + R_n^{4\text{G}} = 1$) and thus does not add any additional information. Computing

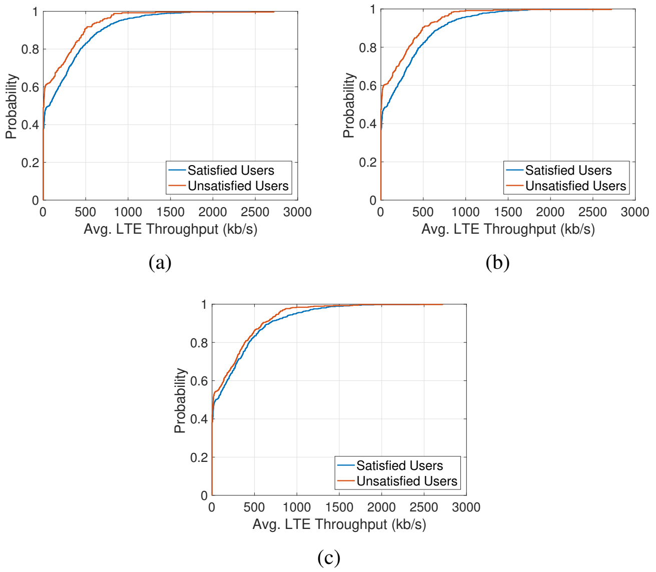


(a)

(b)

(c)

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



(a)

(b)

(c)

Fig. 9. Class-conditional CDFs of G_{n}^{4G} for $\mathcal{T}$ = (a) 6, (b) 7, (c) 8.

TABLE II
PERFORMANCE OBTAINED FOR VIDEO QOE PREDICTION

Classifier	AUC	PT	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
LDA	0.57	0.1529	0.56	0.43
NN	0.52	0.3752	0.51	0.49

such features for $n = 1 \ldots 30$, we create 360 features per user for what concerns video measurements. Adding to the model the Average Latitude (LAT_{avg}) and Average Longitude ($LONG_{avg}$) of each user in the dataset, we end up with 362 features per user for what concerns video streaming.

B. Satisfaction Threshold Selection

As done for coverage, it is worth observing the class-conditional CDFs of the features in order to assess the optimal value of the Satisfaction Threshold $\mathcal{T}$. Figures 8 and 9 show the class-conditional CDFs for $\mathcal{T} = 6$, 7 and 8 of R_{30}^{4G} and G_{30}^{4G}. We observe again that $\mathcal{T} = 6$ maximises the difference in the class-conditional CDFs. For that threshold, we observe in Fig. 8 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%. Similarly, observing Fig. 9, 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. Note that the result regarding the best satisfaction threshold is confimed also by the CDFs of the other features, which are not shown for space limits.

C. Prediction Results

We adopt the same workflow described in Section V-C to compute the prediction performances of different learning models. Figure 10 shows the ROC curves of the tested classifiers 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, which scores an AUC value of 0.58. The best working point identified is at 57% of correctly labeled *Satisfied* user (and a corresponding FPR of 43%) for a PT equal to 0.244. The only underperforming classifier turns to be the neural network, scoring an AUC of 0.52, meaning that the chosen structure with a single hidden layer does not suite this model. We believe that a neural network with a more-than-one layer structure would yield better performances: this will be subject to further investigation.

D. Impact of Location-related Features

It is worth to analyse the impact of the features LAT_{avg} and $LONG_{avg}$ in terms of prediction performances, both for coverage and video streaming. Figures 11 (a) and (b) show the gain relative to the AUC score that we get adding to the models the two location-related features. For what concerns coverage, considering the case ST+SQ, the best performing classifier is RF, which improves the AUC score by 2%. Note that similar results are observed for the case ST-Only, which are not shown here. On the other hand, we get for Video Streaming a maximum improvement of 3%, which is reached by AB classifier. In general, we observe that adding the average users location to the considered models we get on average 1% higher AUC scores compared to a model based on objective network measurements only.

VI. DISCUSSION AND CONCLUDING REMARKS

In this work we have commented on the possibility of predicting the long-term coverage and video satisfaction starting from user-side network measurements. The results obtained

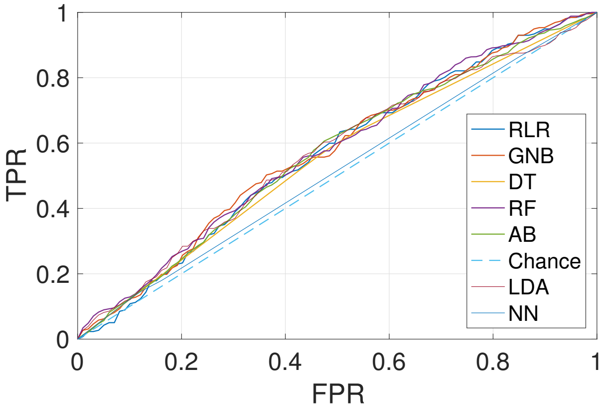
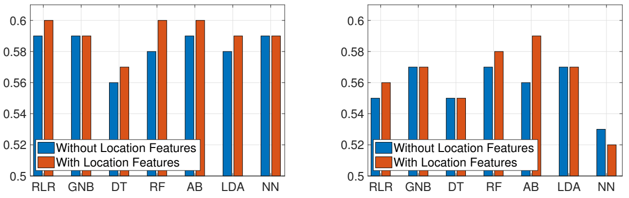


Fig. 10. ROC curve for $Video\ Streaming$ QoE prediction, $\mathcal{T} = 6$.



(a)

(b)

Fig. 11. AUC scores with and without location-related features for cases a) $ST + SQ$ and b) $Video\ Streaming$.

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 user satisfaction feedbacks could be exploited and can be used from the operator as a starting point to identify possible problems in the network. Some interesting points can be raised:

- Compared to short-term QoE estimation, long-term satisfaction 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. Future work will focus on adding commercial-related features (e.g., data plan type and fee) as well as user profile related features (i.e. age, sex, customer type, etc.) to the models tested in order to capture also these type of factors.
- 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). 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 impact of individual prediction errors on the overall task of network problems detection is under investigation.

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] P. Casas, A. D'Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz, "Predicting qoe in cellular networks using machine learning and in-smartphone measurements," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[7] 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.

[8] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, and R. Schatz, "Next to you: Monitoring quality of experience in cellular networks from the end-devices," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 181–196, 2016.

[9] 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.

[10] 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.

[11] 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.

[12] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, "Modeling web quality-of-experience on cellular networks," in *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014, pp. 213–224.

[13] 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.

[14] 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.