QoE prediction on LTE Networks

Alessandro Zito, Zhengchen Xu

August 2022

ID number: 890219 - 990061
Personal code: 10617579 - 10815279
Supervisor: Andrea Pimpinella

Contents

1 Introduction

The goal of this project is to develop a Supervised Machine Learning Classifier method to predict users' satisfaction related to the usage of YouTube Service. The Classifier will take as input (i.e., data measured directly at users terminals) and will output the corresponding users satisfaction class.

2 Data Analysis

Our dataset is composed 23.7 K unique survey responses (YouTube service) and it has been collected with the crowd-sourcing method. The dataset was given us in random distribution, that is not the best type of distribution that it is possible to use to train the model, cause it might lead to bad training. In figure 3, we have transformed our dataset from Random distribution to Gaussian distribution, in order to improve our model.

We've analyzed our dataset to see if there were some correlations between our features, in order to start to see if some feature may be helpful or not.

2.1 CDF

The meaning of the CDFs is that, if there is a gap between the distributions of the data conditioned to the satisfaction class of the corresponding users, it means that the information in the data is correlated to users satisfaction and thus can be learnt by a supervised classifier. In our dataset, only few features have no gap between Low QoE and High QoE; most of them have gap, which means they are correlated to the user satisfaction.

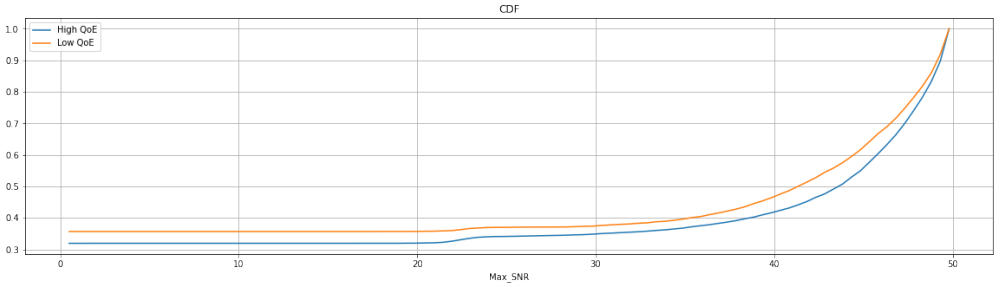

Figure 1: CDF of Signal-to-Noise Radio

2.2 Pearson's correlation coefficient

Pearson's correlation coefficient helps us to understand the correlation between the one feature and the user satisfaction. This coefficient says that the more absolute value is near to 1, there will be bigger correlation between two inputs; on the other hand, if the coefficient is near to 0, there will be a very low correlation. Trying to calculate this coefficient, we have seen that the correlation between our features and the user satisfaction is quite low (there are some feature that have more, and some less).

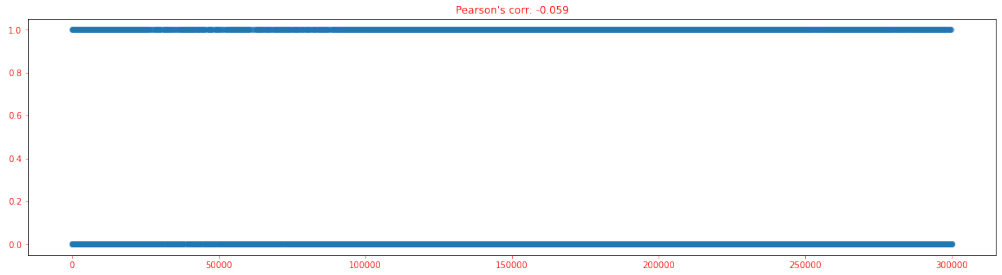

Figure 2: Pearson's correlation coefficient under Full LTE network

This is saying that the relationship between our 1st grade features and the user satisfaction is quite low, but we will see that we can combine more features in order to train better our model.

3 Features Selections

3.1 Data Transformation

ML Classifiers prefer in input Gaussian distribution features. Gaussian Processes, are a generalization of the Gaussian probability distribution. Gaussian probability distribution functions summarize the distribution of random variables, whereas Gaussian processes summarize the properties of the functions, e.g. the parameters of the functions. As such, you can think of Gaussian processes as one level of abstraction or indirection above Gaussian functions. Gaussian processes can be used as a machine learning algorithm for classification predictive modeling, and the processes are a type of kernel method, like SVMs, although they are able to predict highly calibrated probabilities, unlike SVMs.

3.2 Throughput

Throughput is a measurement that refer to the rate of successful message delivery over a communication channel in a communication network. We selected it as our main feature cause having this data. It can be used to pinpoint network impairments, reflect the viewer's watching experience. Consider the following applications: Calculating the throughput as

$$throughput = volume/time \tag{1}$$

and using it to train our model, we noticed that it has increased our ROC curve.

4 Classifier selection and optimization

In our case, we used Deep Learning method as the ML Classifiers, Keras as the most used deep learning framework, could be suitable for binary classification problem and easy to be implemented. Before training the model, the input dataset is the key factor. As we mentioned before, the given features is in random distribution, which might leads to bad result after model training. On the contrary, the parametric methods are powerful and well understood towards Gaussian distribution data. Therefore, Yeo Johnson Transformation method is used for data transformation, it can modify the distributional shape of a set of data to be more normally distributed. The result is shown in figure3, it can be clearly seen that after data transformation, the data is near Gaussian distribution.

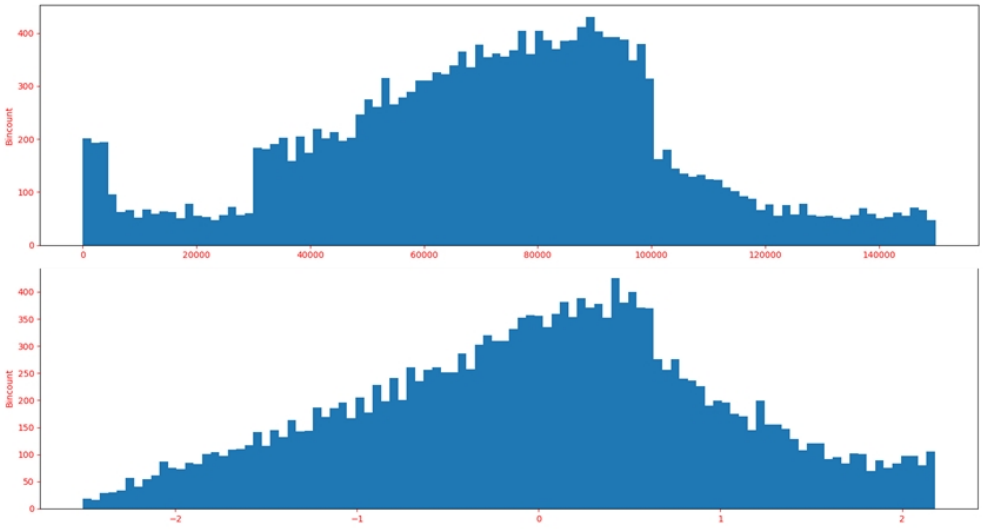

Figure 3: data transformation

After that, we divided our training dataset into SubTraining set and Validation set with the percentage of 80% to 20%. In order to feed the demand of binary classification purpose, we built up our Keras model after several experiments, considering the problem of over-fitting and under-fitting, to decide the number of units is important, we selected the units number based on the binary numbers, and evaluate the performance based on the accuracy. In the end, our model contains three Fully connected layer with first two activation function "relu", and the last activation function "sigmoid". After compiling the model, hyper-parameters like batch_sizes, epochs and verbose need to be tuning. Firstly, we maintained epochs=30, verbose=1 as primary parameters, and changed batch_size, after tuning in each time, we can get the model prediction accuracy based on the test set, and following figure4 is how batch_size related the the model accuracy. Secondly, we set the

batch_size=10, changed the epochs, as the result shown in figure5.

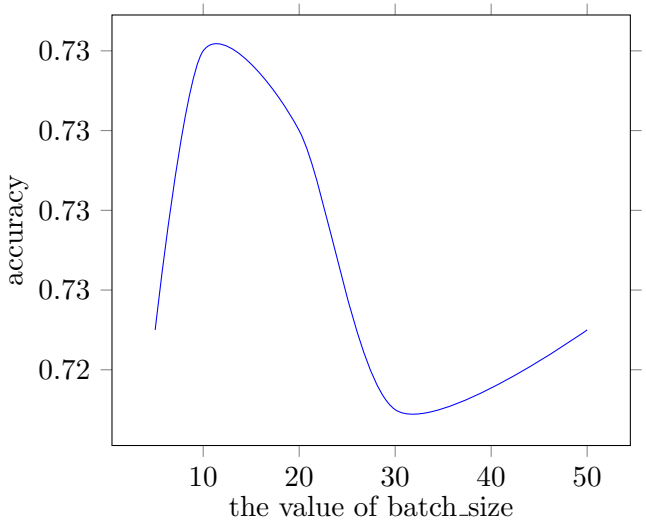

Figure 4: parameter batch_size

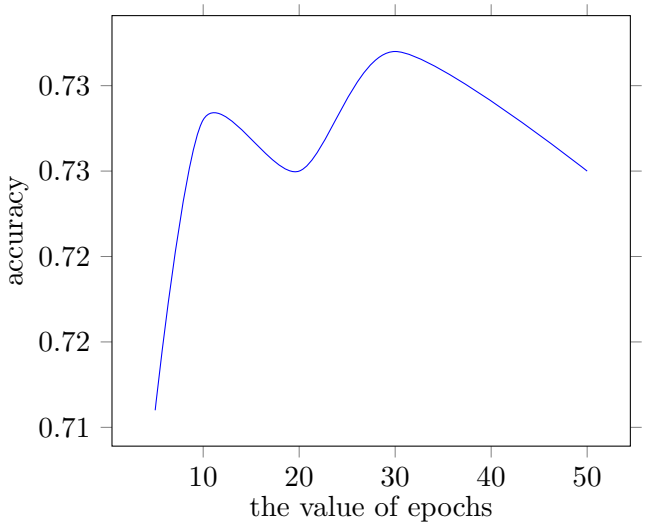

Figure 5: parameter epochs

Afterwards, our hyper-parameters can be settle, with batch_size=10 and epochs=30.

5 Prediction performance in ROC curve

In order to show the prediction performance, we use the Area under the ROC Curve (AUC), also by comparison, we build up a Random Forest model to show the prediction difference between Deep Learning method and RF method. As mentioned before, we used Control variables method to determine the hyper-parameters in the RF model, with the result of (max_depth=10,n_estimator=40).

Therefore, after plotting the figure6, we can see that our results can basically feed the demand of this project.

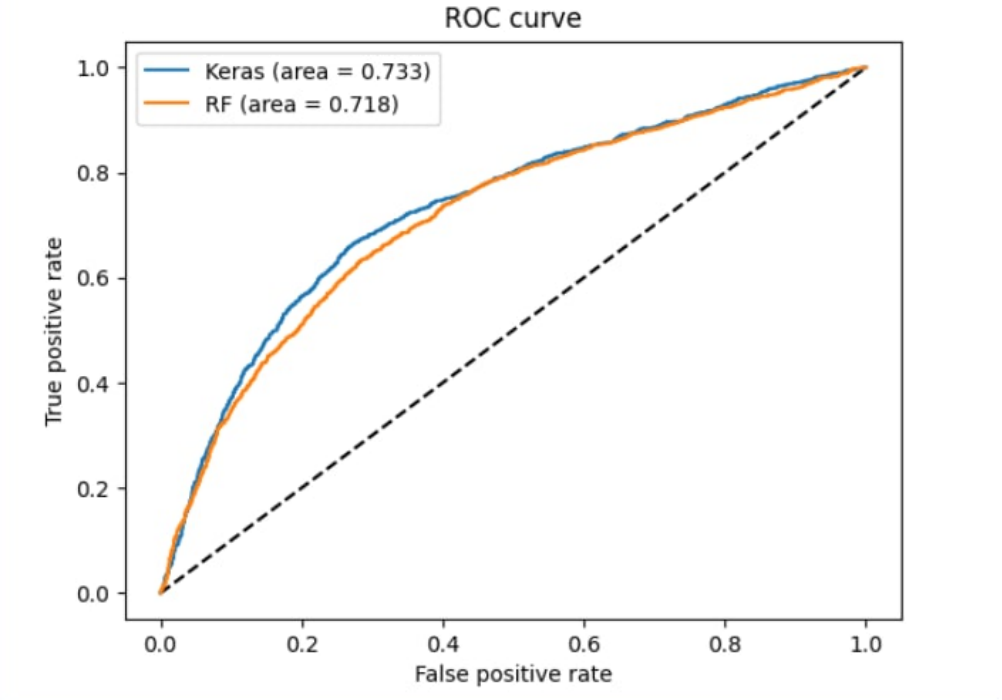

Figure 6: ROC curve