Why That Rating? Explainable Data-Driven Opinion Score Distribution Models for Video QoE

Edward Austin, Tomasz Lyko, Nicholas Race
{e.austin, t.lyko, n.race}@lancaster.ac.uk
School of Computing and Communications, Lancaster University, UK

***Abstract*—Driven by the need to better reflect and understand audience quality of experience content providers and deliverers are no longer solely using models for the Mean Opinion Score but also the opinion score distribution. In tandem, motivated by the desire to understand how these models predict the QoE and which features contribute positively to it, there has been increased emphasis on explainable QoE modelling. Recently the advantages of directly explainable models - where model outputs can be explained using the inputs and the model's inner workings- have been championed. These models provide clear insights into how different features contribute positively, or negatively, to QoE. To date, research into directly explainable methods has focussed on MOS, rather than opinion score distribution, modelling. To bridge this gap we discuss the feasibility of using multinomial logistic regression for directly explainable MOS modelling, demonstrating our approach on a short form streamed video dataset and showing that it compares favourably to other explainable methods.**

***Index Terms*—QoE, Machine Learning, Opinion Distributions**

I. Introduction

As [1] describe, the ways in which we stream and view video content has become increasingly complex. This has led to a demand for new QoE models that accurately reflect viewer experience, as traditional standards models for streamed video assume that the content is viewed in a standardised way, which is often not the case. This, in turn, gives media creators and providers an inaccurate understanding of their audiences perceptions and their Quality of Experience (QoE).To overcome this issue machine and deep learning methods have been employed for data-driven QoE modelling. These take results from a subjective test, and information about the content, network conditions, or viewer and train a QoE model.

Motivated by the desire to better understand audience experiences, there has been growing interest in modelling the distirbution of opinion scores rather than the traditional Mean Opinion Score (MOS) value for example the works of [2], [3] and [4]. Unlike the MOS this provides probabilities that a viewer will give a particular rating - typically on the 5-point ACR scale, outlined in ITU Rec. P.910 [5]. The advantage of this is that it models the spread of opinion scores, capturing the varied experiences of the audience and the subjectivity of their ratings. As such it allows content providers and deliverers to better model their audiences perceptions.

In parallel, and continuing this trend towards better understanding of QoE, the explainability of QoE models has received increasing attention [6], [7]. As outlined by [8], there are two types of explainable model: a directly explainable model, where it is clear from the fitted model itself how inputs map to outputs, and so which features best explain the model prediction; and a black-box model that is explainable through post-hoc procedures such as SHAP or LIME values. [8] argue that the directly explainable approach is preferable as it provides clear conclusions as to which features are most influential on the QoE, and how the model arrived at an output. In contrast post-hoc testing can be misleading as it is a global approximation over all inputs and so can either be wrong due to model misfit, or unrepresentative of particular scenarios.

To date, however, the focus of directly explainable modelling has been on explaining models for the MOS, rather than opinion score distributions. Given the aforementioned benefits of distribution modelling in this short paper we seek to bridge this gap. To do this we discuss the use of multinomial logistic regression for QoE modelling and answer two questions:

1) How can opinion score distributions be modelled in an directly explainable way, identifying which features explain why a user has given a certain rating?
2) How does this approach compare to other opinion score distribution modelling methods?

The benefit of directly explainable modelling for opinion score distributions is that the practitioner is able to identify which variables have most influence on a given rating - answering the question of "why that rating" - without the need to rely upon potentially inaccurate post-hoc approximations. This gives an understanding as to which features positively, or negatively, influence QoE.

II. Related Work

As [9] and [10] discuss, in the era of big data there has been significant interest in using both machine and deep learning methods to model QoE for video quality assessment. Whilst traditionally attention focussed on MOS modelling [11], [12], there has been a growing interest in opinion score distribution modelling. Example works for video quality assessment include [13], [14], [15] and [16], with the latter two approaches also using a regression model for probability prediction related to multinomial logistic regression but not focussing on explainability. Similarly, contributions for image quality assessment include those of [17], [18] and [19].

Recently, there has been an emphasis on explainable modelling for QoE estimation [20]. This is where the models themselves, or quantities derived from them, are used to

explain why an input gives a particular output, or identify which features are important for predicting model outputs. As [8] explain, the ideal approach to this would be where the model coefficients themselves allow the user to explain the important features in the model, or how an input leads to an output. [8] demonstrate this explainability using regression to model the MOS, and [21], [6] use decision trees.

Beyond the directly explainable approaches, several authors have also considered post-hoc approximations for model explainability including SHAP and LIME values, or permutation testing. Examples of authors using SHAP include [22], [23], [24], LIME include [25], [26], and [27] use permutation testing as part of their random forest approach. As discussed in the Introduction, however, these post-hoc approximations can be inaccurate, and are often aggregated globally over all inputs and so can be misleading when used to interpret individual settings. This may give practitioners an incomplete understanding of what influences their viewer's QoE.

III. Methodology

A. Background and Mathematical Notation

Mathematically, we consider a set of n videos that have been rated by m viewers in a subjective test, with y_{ij} representing the jth viewer rating for the ith video in the test. Furthermore we let $\boldsymbol{x_i} = \{x_{ik}\}_{k=1}^{K}$ be the K model features for the ith video. Our model uses the random variable Y_{ij} for the absolute category rating of the jth user for the ith video. Under this model the opinion score probabilities given the data are $P(Y_{ij} = r \mid \{x_{ik}\}_{k=1}^{K}), 1 \leq r \leq 5$; this can be interpreted as the probability that the jth user gives a rating of r to the ith video. We assume that the $Y_{ij} \mid \{x_{ik}\}_{k=1}^{K}$ follow an underlying opinion score distribution. This distribution captures the probability of a user giving a particular rating to a video given the viewing conditions.

B. QoE Modelling With Multinomial Logistic Regression

Under the assumption that each test subject rates a video independently of another subject a viable approach for estimating the opinion score distribution is to use multinomial logistic regression. For each of the r ratings in the subjective test this model contains a set of weights for the K viewing condition factors. These weights are denoted by $\{\beta_{r,k}\}$, and can be estimated using the regression modelling. The opinion score probabilities for the ith video are then given by:

$$P(Y_{ij} = r \mid \boldsymbol{x_i}) = \begin{cases} \frac{\exp\left(\sum_{k=0}^{K} \beta_{r,k} x_{i,k}\right)}{1+\sum_{s=2}^{5} \exp\left(\sum_{k=0}^{K} \beta_{s,k} x_{i,k}\right)} & 2 \leq r \leq 5 \\ \frac{1}{1+\sum_{s=2}^{5} \exp\left(\sum_{k=0}^{K} \beta_{s,k} x_{i,k}\right)} & r = 1. \end{cases} \tag{1}$$

$$P(Y = r \mid \boldsymbol{x}) = \begin{cases} \frac{\exp\left(\sum_{k=0}^{K} \beta_{r,k} x_{k}\right)}{1+\sum_{s=2}^{5} \exp\left(\sum_{k=0}^{K} \beta_{s,k} x_{k}\right)} & 2 \leq r \leq 5 \\ \frac{1}{1+\sum_{s=2}^{5} \exp\left(\sum_{k=0}^{K} \beta_{s,k} x_{k}\right)} & r = 1. \end{cases} \tag{2}$$

Here $\beta_{r,0}$ is the intercept term, and so $x_{i,0} = 1$ for all i. The weights of this model can be estimated from a sample of data using maximum likelihood - for details see for example [28].

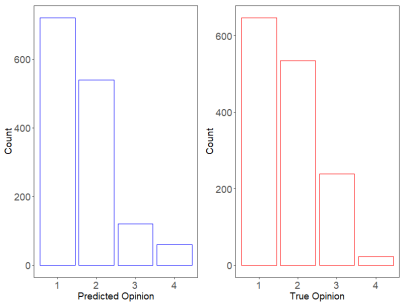

Fig. 1. Plot showing predicted and observed distribution of opinion scores for videos streamed at $10Mbit/s$ in $1080p$ subject to 0.2% packet loss.

Once the regression model has been estimated it can be used as a model for understanding and predicting QoE based on new input. To predict the opinion score distribution for a new video with features $\boldsymbol{x_{N+1}}$ we would compute $P(Y_i = r \mid \boldsymbol{x_{N+1}})$ using equation (2), for each $1 \leq r \leq 5$.

C. Explainability

To obtain the most important features one analyses the exponentiated values of the model coefficients, $R_{r,k} = \exp(\beta_{r,k})$ [29]. The value of $R_{r,k}$ measures how much the kth feature affects the likelihood of a user giving a video a rating of r compared to the baseline rating, which in this case is $r = 1$. If $R_{r,k} > 1$ then the kth feature makes a rating more likely to be r rather than 1, whilst if $R_{r,k} < 1$ then it is less likely.

The value of this is that a practitioner can easily identify which factors contribute positively, or negatively, to their QoE. As a practical example, content deliverers and providers will be most interested in large values of $R_{4,k}$ and $R_{5,k}$ as these are the features that are most likely to increase the probability that the viewer gives a rating of 4 or 5. Alternatively if every coefficient is much less than 1 then the feature most influences a rating of 1, and so negatively affects QoE.

D. Explaining Model Fit

To further support practitioners in explaining the suitability of their QoE model, goodness-of-fit testing can be performed to establish theoretical guarantees on the accuracy of the estimated QoE model. This is achieved through a likelihood ratio test that compares the fitted model with a null model containing only an intercept term and no other features [28]. The statistic for this test, $\mathcal{D}$, is given by

$$\mathcal{D} = 2\left(\mathcal{L}_{fitted} - \mathcal{L}_{null.}\right) \tag{3}$$

Here $\mathcal{L}^{fitted}$ is the log-likelihood of the fitted model using the procedure outlined in this section, and $\mathcal{L}^{null}$ is a multinomial logistic regression fitted using only an intercept term. For a

sample of observations with estimated probabilities $P(Y_i \mid \boldsymbol{x_i})$ this is equal to

$$\mathcal{L}^{null} = \sum_{r=1}^{5} \sum_{y_{i,j}=r} \log\left(P(Y_i^{null} = r \mid \boldsymbol{x_i})\right)$$

$$\mathcal{L}^{fitted} = \sum_{r=1}^{5} \sum_{y_{i,j}=r} \log\Big(P(Y_i^{fitted} = r \mid \boldsymbol{x_i})\Big).$$

If $\mathcal{D} > \chi^2_{(K)(R-1)}(0.95)$, where $\chi^2_{(K)(R-1)}(0.95)$ represents the 95th quantile of the χ^2 distribution with $(K) \times (R-1)$ degrees of freedom, then there is evidence that the fitted model is a significantly better fit for the data. Here $R = 5$.

IV. Case Study

To demonstrate the feasibility of this explainable modelling approach we analyse data from a short-form streamed video subjective test produced by [30], which has not yet been analysed in the QoE modelling literature. This dataset contains 424 10 second reference videos that haven been presented to 60 test subjects using a video streaming testbed, giving a total of 25440 ratings. The videos cover a mix of static and dynamic shots, are delivered in 1080p, 1440p, and 4K, and contain a range of impairments caused by packet loss in the streaming session. The dataset contains the following features: bitrate, resolution, % of lost packets, ranging from 0% to 1% across the session, and two full reference metrics, Structural Similarity Index and the VMAF.

A. Model Fit

We split the streaming video dataset into a 50:50 train-test split along the videos - so 212 videos are in the train set, and 212 are in the test set. In terms of sample size, various rules have been proposed for the minimum required for multinomial logistic regression. As described by [31] 10 samples per predictor is sufficient, and so in this case we would need 200 videos. We then fit the multinomial logistic regression to the training data using all five features after they have been standardised to $[0, 1]$. Using the likelihood ratio test we obtain $D = 14931.78$, and this exceeds the critical value of the χ^2_{20} distribution, which is 31.41, and so there is strong evidence that the estimated model is a good fit to the data.

To establish the accuracy of the fitted model we use the test dataset features to predict the opinion score distributions, recording a MSE of 0.092. We will compare this against other distribution prediction methods in the next section, however to identify the comparative accuracy of each fitted distribution with the observed distribution we use a Chi-Squared Test, finding no evidence at the 5% level that any predicted distribution is different to the observed. We show a predicted and observed distribution comparison in Figure 1.

In terms of explainability we present the coefficients $R_{r,k}$ in Table 1. Given the very small values of $R_{r,k}$ for packet loss, notably for higher rating scores, packet loss can be identified as having the most negative impact on QoE. For a rating of 2 and 3 the most influential factors are the SSIM and resolution,

$R_{r,k}$	Bitrate	Resolution	Packet Loss	SSIM	VMAF
2	0.69	1.17	0.49	3.85	0.25
3	0.74	1.07	0.48	3.04	0.25
4	0.74	0.95	0.12	0.03	1.51
5	0.19	0.50	0.01	0.01	6.86

TABLE I
COEFFICIENTS $R_{r,k}$ EXPLAINING THE IMPORTANCE OF THE FEATURES IN THE MODEL FITTED TO THE STREAMED VIDEO DATA.

however the coefficients are smaller for 4 and 5 are so these features can be explained as improving QoE but to a limited extent. For the higher QoE ratings the coefficients show that the VMAF is the most influential factor, suggesting that this metric provides the best indicator of high QoE.

B. Comparing To Other Data-Driven OS Distribution Models

Although our analysis indicates that the multinomial logistic regression provides a good fit to our dataset, we must also assess how it compares to other opinion score distribution modelling tools. We test against several state of the art methods: a Support Vector Machine with a cubic kernel, a Random Forest, an MLP neural network, a mixture of binary regressions as proposed by [15] for distribution modelling, and a decision tree. Note the latter two methods are directly explainable. We fit each of these models to the streamed video dataset and present a comparison of the Mean Square Error (MSE) and Mean Kullback Leibler Divergence (Mean KLD) in Table 2. The former is the mean sum of squared difference, and the latter the mean KL divergence, between the predicted and observed probabilities for each video.

Method	MLR	SVM	RF	NN	BRs	DT
MSE	0.092	0.755	0.092	0.087	0.389	0.120
Mean KLD	0.70	2.62	0.72	0.71	5.53	0.77

TABLE II
TABLE COMPARING THE MSE AND MEAN KULLBACK-LEIBLER DIVERGENCE FOR MULTINOMIAL LOGISTIC REGRESSION (MLR), SUPPORT VECTOR MACHINE (SVM), RANDOM FOREST (RF), NEURAL NETWORK (NN), BINARY REGRESSIONS (MBRS), AND DECISION TREE (DT) MODELS FITTED TO THE STREAMING DATASET.

From the results in Table 2 we see that our approach has outperformed the other directly explainable methods, and the SVM, and is competitive with more complex tools such as the random forest and neural network. This gives further confidence that our proposed approach is valuable for QoE modelling as it not only estimates the opinion score distributions in an explainable manner but it performs similarly to other state-of-the-art methods.

V. Conclusion

We have demonstrated the feasibility of data-driven and explainable modelling for opinion score distributions using a multinomial logistic regression. This method allows practitioners to understand which input features explain why a viewer has given a certain rating. We have also shown how this model performs favourably compared to other directly explainable models, and competetive with alternative tools that require post-hoc explanations on a video dataset that has not yet been

analysed in the QoE literature. That said, it is important to note that these results are for only one dataset, and that this dataset contains only 424 videos (although 25440 datapoints), and so further work is to verify these conclusions on a broader variety of/larger datasets. Two additional avenues of research are: to build on [7] and explore how LLMS could be used to automate the distribution model explanations, or to investigate how ensembles of directly explainable tools could be used for more accurate and explainable QoE modelling.

ACKNOWLEDGMENT

This research was supported by UKRI EPSRC and BBC Prosperity Partnership AI4ME: EP/V038087. Code to reproduce the results is on Github at https://github.com/austine94/QOMEX25.

REFERENCES

[1] W. Moina-Rivera, M. Garcia-Pineda, J. Gutiérrez-Aguado, and J. M. Alcaraz-Calero, "Cloud media video encoding: review and challenges," *Multimedia Tools and Applications*, vol. 83, no. 34, pp. 81 231–81 278, 2024.

[2] T. Hoßfeld, P. E. Heegaard, and M. Varela, "Qoe beyond the mos: Added value using quantiles and distributions," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[3] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "Qoe beyond the mos: an in-depth look at qoe via better metrics and their relation to mos," *Quality and User Experience*, vol. 1, pp. 1–23, 2016.

[4] M. Seufert, "Fundamental advantages of considering quality of experience distributions over mean opinion scores," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.

[5] Recommendation ITU-T P. 910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union*, 2023.

[6] J. L. C. Bárcena, P. Ducange, F. Marcelloni, G. Nardini, A. Noferi, A. Renda, G. Stea, A. Virdis *et al.*, "Towards trustworthy ai for qoe prediction in b5g/6g networks," in *CEUR WORKSHOP PROCEEDINGS*, vol. 3189. CEUR WS, 2022.

[7] N. Wehner, N. Feldhus, M. Seufert, S. Möller, and T. Hoßfeld, "Qoexplainer: Mediating explainable quality of experience models with large language models," in *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, 2024, pp. 72–75.

[8] M. Seufert and N. Wehner, "Explainable artificial intelligence for quality of experience modelling," *ACM SIGMultimedia Records*, vol. 15, no. 3, pp. 1–1, 2024.

[9] J. Klink, M. Łuczyński, and S. Brachmański, "Video quality modelling—comparison of the classical and machine learning techniques," *Applied Sciences*, vol. 14, no. 16, p. 7029, 2024.

[10] Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik *et al.*, "Video quality assessment: A comprehensive survey," *arXiv preprint arXiv:2412.04508*, 2024.

[11] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2351–2359.

[12] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: Overview, benchmark, and beyond," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–36, 2021.

[13] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying subjective quality evaluations for mobile video watching in a semi-living lab context," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 580–589, 2012.

[14] M. Seufert, "Statistical methods and models based on quality of experience distributions," *Quality and User Experience*, vol. 6, no. 1, p. 3, 2021.

[15] L. Janowski and Z. Papir, "Modeling subjective tests of quality of experience with a generalized linear model," in *2009 International Workshop on Quality of Multimedia Experience*, 2009, pp. 35–40.

[16] S. Göring, R. R. R. Rao, B. Feiten, and A. Raake, "Modular framework and instances of pixel-based video quality models for uhd-1/4k," *IEEE Access*, vol. 9, pp. 31 842–31 864, 2021.

[17] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[18] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 997–1005.

[19] Y. Gao, X. Min, Y. Zhu, X.-P. Zhang, and G. Zhai, "Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1641–1655, 2023.

[20] N. Wehner, A. Seufert, T. Hoßfeld, and M. Seufert, "Explainable data-driven qoe modelling with xai," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 7–12.

[21] A. Renda, P. Ducange, G. Gallo, and F. Marcelloni, "Xai models for quality of experience prediction in wireless networks," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2021, pp. 1–6.

[22] H. Gokcesu, O. Ercetin, G. Kalem, and S. Ergut, "Qoe evaluation in adaptive streaming: Enhanced mdt with deep learning," *Journal of Network and Systems Management*, vol. 31, no. 2, p. 41, 2023.

[23] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, "Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment," *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 396–405, 2023.

[24] S. Viriyavisuthisakul, S. Yoshida, K. Shiohara, L. Xiao, and T. Yamasaki, "Explainable ai for image aesthetic evaluation using vision-language models," in *2025 Conference on Artificial Intelligence x Multimedia (AIxMM)*. IEEE, 2025, pp. 62–65.

[25] X. Zhang, B. Joukovsky, and N. Deligiannis, "Quantitative evaluation of video explainability methods via anomaly localization," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 1235–1239.

[26] G. Buddhawar, D. Dave, K. Jariwala, and C. Chattopadhyay, "A spatio-temporal explainable deep learning approach for frame classification from book flipping videos," in *International Conference on Soft Computing and its Engineering Applications*. Springer, 2025, pp. 73–84.

[27] J. Akhtar, V. Verma, and A. Kumar, "Enhancing iot wi-fi networks through qos-qoe correlation: A random forest approach for predictive mean opinion score modeling," in *2024 IEEE Students Conference on Engineering and Systems (SCES)*. IEEE, 2024, pp. 1–6.

[28] J. S. Long and J. Freese, *Regression models for categorical dependent variables using Stata*. Stata press, 2006, vol. 7.

[29] J. Wooldridge, *Econometric Analysis of Cross Section and Panel Data, second edition*, ser. Econometric Analysis of Cross Section and Panel Data. MIT Press, 2010. [Online]. Available: https://books.google.co.uk/books?id=yov6AQAAQBAJ

[30] J. Frnda, M. Durica, J. C.-W. Lin, and P. Fournier-Viger, "Video dataset containing video quality assessment scores obtained from standardized objective and subjective testing," *Data in Brief*, vol. 54, p. 110458, 2024.

[31] V. M. de Jong, M. J. Eijkemans, B. Van Calster, D. Timmerman, K. G. Moons, E. W. Steyerberg, and M. van Smeden, "Sample size considerations and predictive performance of multinomial logistic prediction models," *Statistics in medicine*, vol. 38, no. 9, pp. 1601–1619, 2019.