TÉLÉCOM SUDPARIS

VAP DANI

DATA SCIENCE AND NETWORK INTELLIGENCE

Report

Analysis of Quality of Experience in Video Streaming

for the exam

Data Science - Theory to Practice

Professor	Student
Alessandro MADDALONI	Pratistha THAPA

DECEMBER 2024

Academic Year 2024-25

Table of Contents

TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

Introduction

Customer satisfaction and loyalty are integral to businesses, especially in a market where competitors offer similar services and customers have greater choices. Similarly, the quality of user experience is crucial for Internet Service Providers and Over The Top platforms, as poor experience can result in customer churn eventually reducing revenue [1]. To better understand the influence of Quality of Service (QoS) indicators on users' Quality of Experience (QoE), the report analyses a crowdsourcing campaign conducted in and around Paris for the PoqeMoN project. The study involved 181 participants aged between 19 and 38 years using 9 Android devices with different attributes (models and OS versions) for generating 1560 samples.

The investigation on YouTube video streaming experiences on Android devices across our major mobile network operators in France: Orange, SFR, Bouygues, and Free. The ratings are provided by the testers based on their experiences based on several QoE Influence Factors (QoE IFs) including video parameters from the VLC media player for measuring Quality of Service (QoS) indicators, network information, device characteristics, user profiles, and feedback. These ratings played a significant role in the calculation of the Mean Opinion Score (MoS) which is a primary measure for user satisfaction. To summarize, this dataset has been well utilized to correlate the subjective experiences and the objective metrics for predicting Quality of Experience (QoE) in the telecom environment.

The report showcases detailed processes and checkpoints undertaken in the data exploration journey which were critical to ensure an unbiased result from the developed model.

Goals of the Analysis

By leveraging data science techniques and algorithms with the user experience indicators and their opinion scores the analysis aims to achieve the following goals:

Anomaly Detection

Outliers and anomalies are data points that turn a final picture upside down if ignored, thus, it's very important to treat it any if it exists. By identifying anomalies or feedback that deviates from ordinary user behavior potential errors in the result could be reduced, ensuring a true outcome of the analysis. Employing statistical methods and clustering approaches will be used for anomaly detection in metrics like buffering time and MOS. With this analysis the technical outliers and feedbacks could be addressed for better service optimization.

User Retention Analysis

Analyzing user retention or their likelihood to continue with a mobile network service based on the Mean Opinion Score (MOS) and metrics like buffering time is a very interesting and critical subject. The outcome of this analysis could be further utilized to identify the dissatisfied users and addressing early to reduce their churn. Based on the user experience influence factors for the dissatisfied users, optimizations could be made to improve overall satisfaction and build loyalty.

For this users are categorized based on their Mean Opinion Score (MOS) into the following

Retained Users: MOS >= 4.

At-Risk Users: MOS < 4.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial process that is not just limited to summarizing data but it is also about going deeper into the data to understand its nitty gritty. It uncovers hidden patterns and provides clarity on the relationships between variables. These are essentials for discovering a new cluster/group based on their homogeneity that was unseen. EDA and data exploration techniques help in challenging the initial assumptions, creating avenues for further analysis to solve problems.

Assumptions

Based on initial eyeballing of the data and conventional wisdom on the dataset, the assumptions are as follows:

1. The presence of missing values or null values
2. Video Quality has the biggest influence on MOS

EDA Protocols

- Understanding data shapes
- Summary statistics
- Handling null and duplicate values

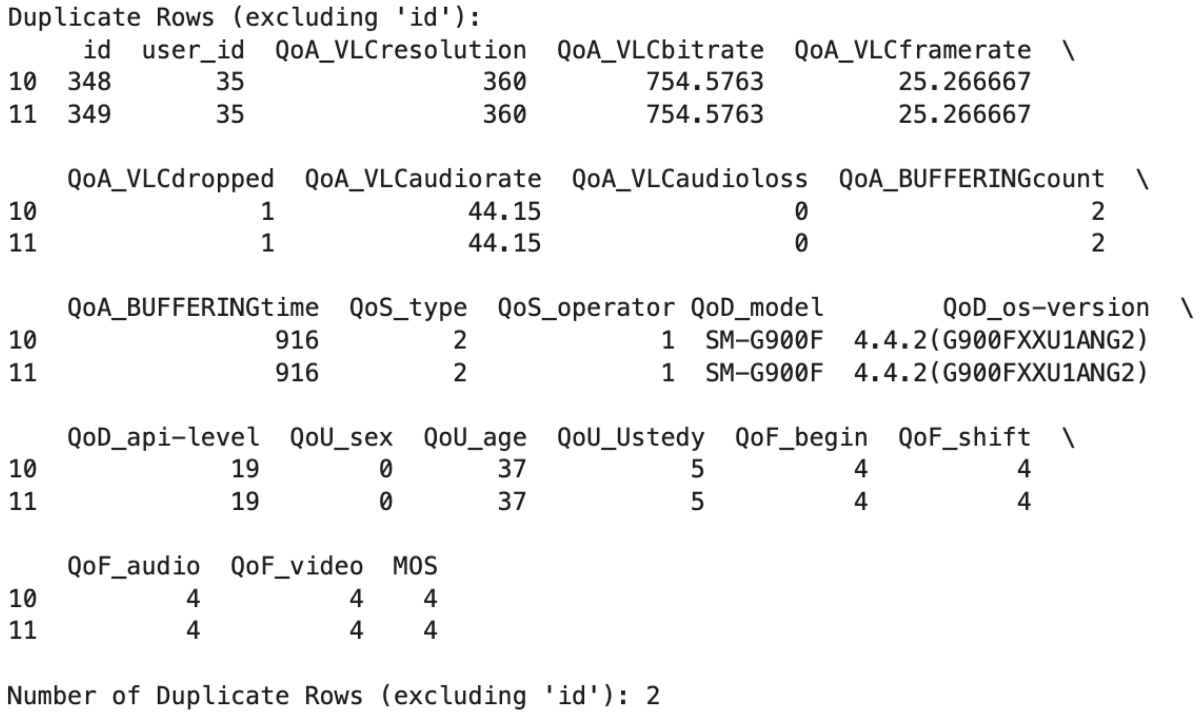
- Corelation Analysis
- Exploring relationships between variables
- Exploring categorical rariable distributions
- Anomaly Detection

EDA Results

The dataset consists of 1543 rows and 23 columns (features) where the target variable is the Mean Opinion Score. It consists of twenty-one continuous and two categorical variables. Identifiers include id and user_id to label records and users uniquely. Quality of Attributes (QoA) captures technical video metrics like resolution, bitrate, frame rate, dropped frames, buffering count, and duration. Quality of Service (QoS) includes service type and operator details, while Quality of Device (QoD) specifies device models, operating system versions, and API levels. Quality of User (QoU) covers demographic data like age, gender, and study group. Quality of Feedback (QoF) measures user interaction with audio and video quality scores. Finally, MOS (Mean Opinion Score) represents overall user satisfaction on a scale of 1 to 5. This comprehensive dataset provides insights into technical performance, user demographics, and satisfaction levels.

Null and Duplicated Values

While finding information about the nulls and duplicate values in the dataset it was unexpected to find that there were no such values. The first column ‘id’ is a unique identifier for each row, excluding which it was found that there were two duplicated rows. The duplicate row was removed and the updated dataset shape was 1542 X 23.



```
Duplicate Rows (excluding 'id'):
     id  user_id  QoA_VLCresolution  QoA_VLCbitrate  QoA_VLCframerate  \
10  348       35                360        754.5763         25.266667
11  349       35                360        754.5763         25.266667

    QoA_VLCdropped  QoA_VLCaudiorate  QoA_VLCaudioloss  QoA_BUFFERINGcount  \
10               1             44.15                 0                   2
11               1             44.15                 0                   2

    QoA_BUFFERINGtime  QoS_type  QoS_operator QoD_model        QoD_os-version  \
10                916         2             1  SM-G900F  4.4.2(G900FXXU1ANG2)
11                916         2             1  SM-G900F  4.4.2(G900FXXU1ANG2)

    QoD_api-level  QoU_sex  QoU_age  QoU_Ustedy  QoF_begin  QoF_shift  \
10             19        0       37           5          4          4
11             19        0       37           5          4          4

    QoF_audio  QoF_video  MOS
10          4          4    4
11          4          4    4

Number of Duplicate Rows (excluding 'id'): 2
```

Figure 1: Duplicated Values

Summary Statistics

Metric	Value
Average Bitrate	520.52 Kbps
Average Framerate	25.00 fps
Dominant Resolution	360p
Mean Buffering Time	6,164 ms
Mean Buffering Count	1.39 events
Average User Age	29 years
Gender Distribution	85% Male
Mean MOS	3.7
MOS Range	3–5
Average Dropped Frames	1.22

Table 1: Summary Statistics

The dataset primarily features quality metrics, buffering statistics, user demographics, and feedback scores. The average bitrate is 520.52 Kbps, while the average framerate is 25.00 fps. The resolution is predominantly 360p, with minor variations observed. Buffering metrics reveal a mean buffering time of 6,164 milliseconds and an average of 1.39 buffering events per session. Regarding user demographics, the average user age is 29 years, with a gender distribution skewed towards males, as indicated by a mean gender value of 0.85 (where 1 represents males). Feedback metrics, as measured by the Mean Opinion Score (MOS), show an average score of 3.7, with most ratings falling within the range of 3 to 5. Lastly, the analysis of dropped frames reveals an average of 1.22 frames per session, although a high standard deviation suggests occasional spikes in frame drops.

Correlation Analysis

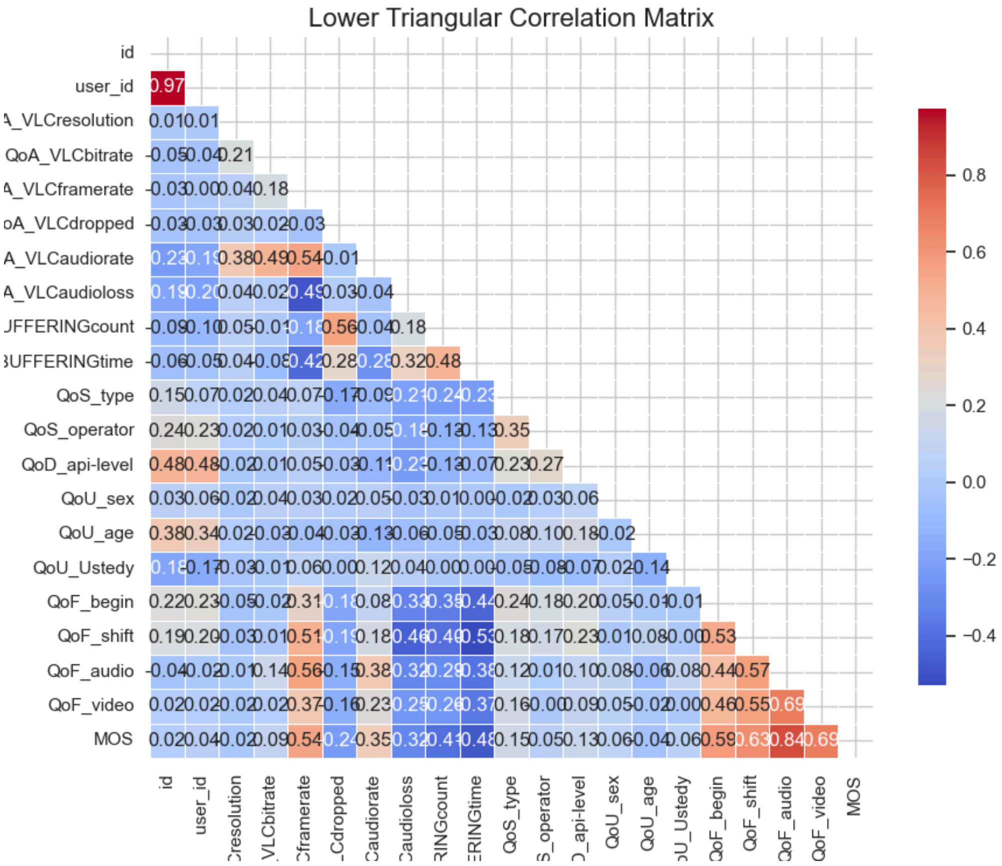


Figure 2: Correlation Matrix

The correlation analysis reveals several key insights into the relationships between the features. Strong positive correlations are observed between QoA_VLCaudiorate and QoA_VLCframerate, as well as between QoA_BUFFERINGtime and QoA_BUFFERINGcount, which is expected given their related nature. These correlations suggest that these pairs of features are closely connected and may measure similar aspects of the dataset. On the other hand, features such as QoA_VLCdropped and QoF_audio exhibit weak or no significant correlation with most other variables, indicating that they might provide independent and unique information about the data.

However, the presence of highly correlated features also raises the possibility of redundancy in the dataset, which could negatively impact model performance. To address this, dimensionality reduction techniques such as Principal Component Analysis (PCA) could be employed. These methods can help reduce the dataset's dimensionality while retaining the most critical information, ensuring that the models remain efficient and robust. These insights are crucial for guiding feature selection and optimizing preprocessing strategies for improved model performance.

Relationship Between Variables

QoA_BUFFERINGtime and MOS Relationship (Mean Opinion Score)

The relationship between buffering time and MOS was examined to find that there is a large variation in QoA_BUFFERINGtime, with a mean of 6164 ms and a maximum value of 329271. Normalization was applied to ensure all features were on the same scale, preventing any one variable from dominating the analysis due to differing units or magnitudes. This ensures optimal model performance and visualization clarity.

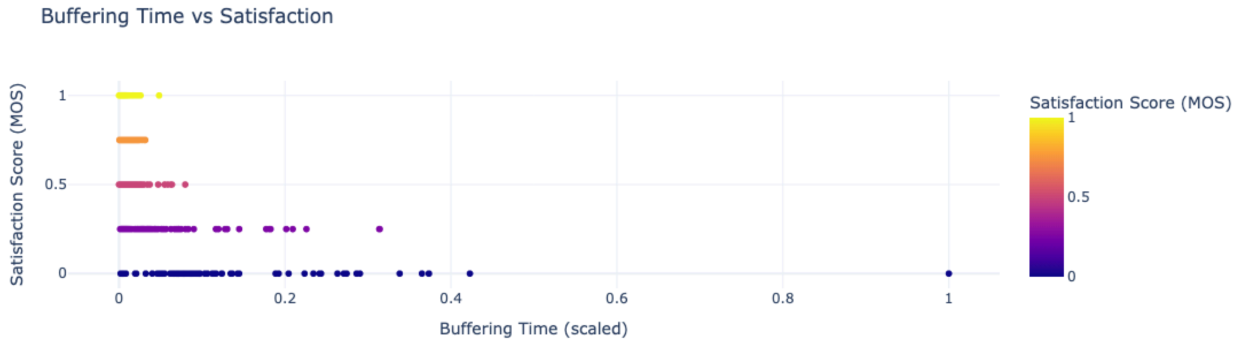


Figure 3: QoA_BUFFERINGtime and MOS Relationship

The scatter plot illustrates the relationship between QoA_BUFFERINGtime and MOS (Mean Opinion Score). A clear negative correlation is observed, indicating that as QoA_BUFFERINGtime increases, the MOS tends to decrease. This suggests that higher buffering durations negatively impact user satisfaction.

The MOS values are predominantly clustered at discrete levels (1, 2, 3, 4, and 5), indicating a categorical or discrete distribution of the score. Most data points are concentrated at lower buffering times, suggesting that the majority of users experience relatively low buffering durations. However, a few outliers are evident at exceptionally high buffering times (e.g., above 250,000), which may represent anomalous user experiences or extreme cases.

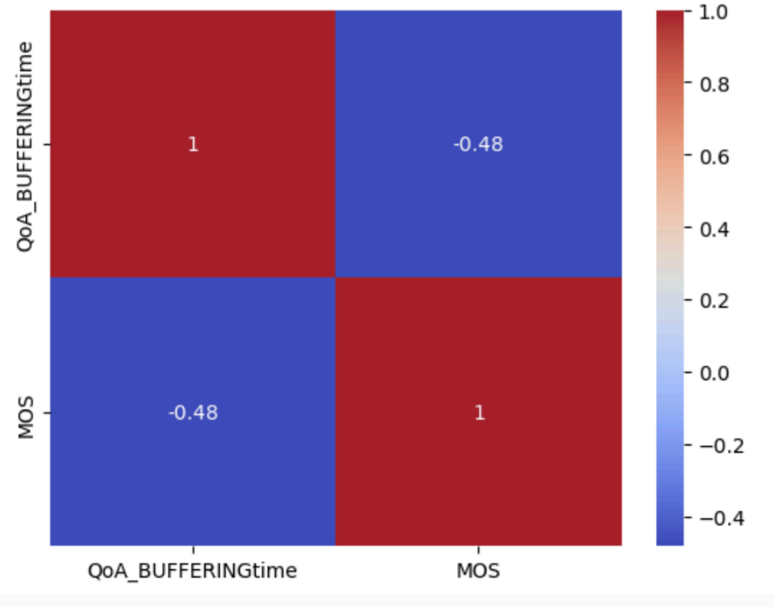


Figure 4: QoA_BUFFERINGtime and MOS (Mean Opinion Score)

This analysis highlights the significant influence of buffering time on user satisfaction. Reducing buffering time could lead to improved user retention and satisfaction levels, as suggested by the observed trend in the data. These insights underline the importance of QoA_BUFFERINGtime as a key feature for predicting user retention and satisfaction.

QoA_VLCbitrate (video bitrate) and MOS (Mean Opinion Score)

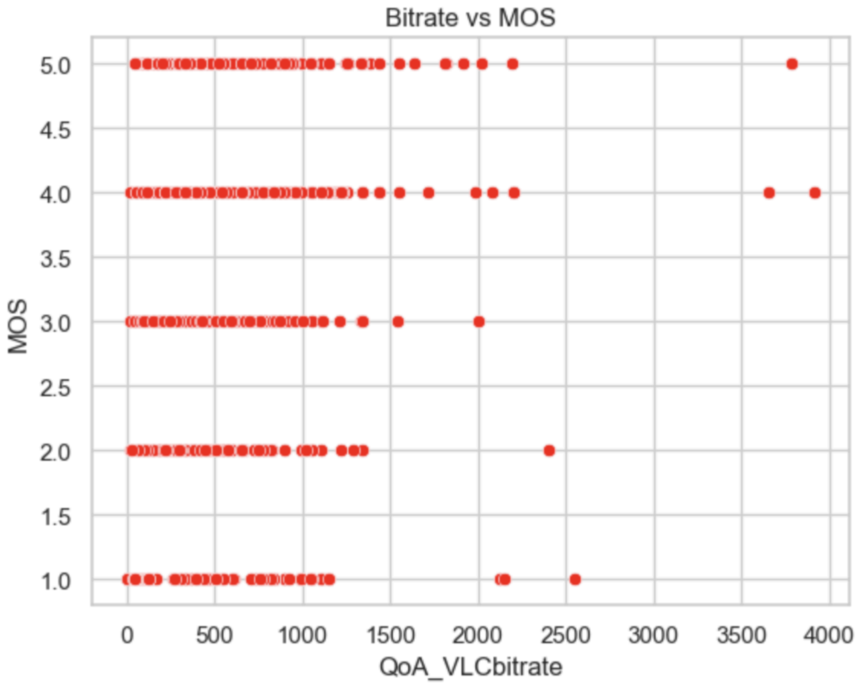


Figure 5: QoA_VLCbitrate and MOS

The scatter plot explores the relationship between QoA_VLCbitrate (video bitrate) and MOS (Mean Opinion Score). The analysis reveals that there is no strong linear correlation between bitrate and user satisfaction, as evidenced by the wide spread of MOS values across different bitrate levels. Users rate videos with high satisfaction scores (MOS of 4 or 5) over a wide range of bitrates, indicating that bitrate alone is not a definitive factor influencing the user experience.

Interestingly, some users provide low satisfaction scores (MOS of 1 or 2) even when the bitrate is high (e.g., above 2000). This suggests that other factors, such as buffering time or playback interruptions, might outweigh the positive impact of a high bitrate. These observations support the hypothesis that higher buffering time and frequency negatively affect MOS, even in cases where the bitrate is high. This analysis highlights the importance of focusing on factors like buffering and playback smoothness in addition to bitrate to enhance user satisfaction. While maintaining an adequate bitrate is essential, prioritizing consistent and interruption-free playback will likely lead to higher retention and satisfaction rates.

Buffering Count and MOS (Mean Opinion Score)

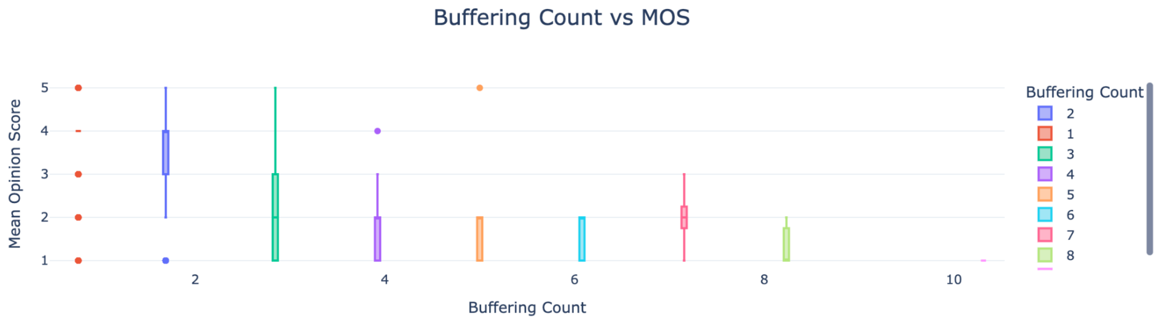


Figure 6: Buffering Count and MOS (Mean Opinion Score)

The plot highlights the impact of buffering count on user satisfaction as measured by the MOS. It is evident that as the buffering count increases, the MOS generally decreases, indicating a negative correlation. For lower buffering counts (e.g., 1 or 2), users tend to give higher satisfaction ratings, as reflected in the median MOS around 4 or 5. However, as buffering counts rise (e.g., 6, 7, or higher), the MOS drops significantly, with most values clustering around 1 or 2.

This trend suggests that frequent buffering significantly degrades the user experience, leading to lower satisfaction ratings. Additionally, there is variability within each buffering count level, particularly at lower counts, where the range of MOS values is wider. This variability indicates that while buffering count is a major factor, other elements such as video resolution or bitrate may also influence user satisfaction.

Framerate (fps) and Mean Opinion Score (MOS)

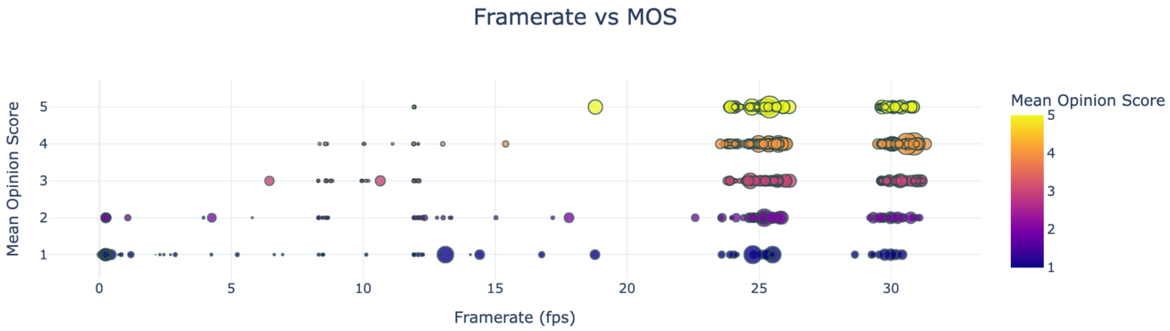


Figure 7: Framerate (fps) and Mean Opinion Score (MOS)

The plot suggests a positive association between higher framerates and user satisfaction as measured by the MOS. For framerates around 25–30 fps, the majority of MOS values cluster near the higher end of the scale (4–5), indicating that smoother video playback is correlated with better user experiences. Conversely, lower framerates (e.g., below 10 fps) are associated with lower satisfaction scores, with MOS values concentrated around 1–2.

However, there is variability in the satisfaction scores even within the same framerate range. For example, at higher framerates (25–30 fps), a small proportion of users still report low satisfaction scores. This suggests that while framerate is a critical factor, other variables, such as buffering time or bitrate, might contribute to the overall user experience.

The findings highlight the importance of maintaining a high framerate (preferably 25 fps or above) to enhance user satisfaction. Additionally, addressing other factors like buffering and dropped frames is essential to ensure consistently high MOS ratings across all users.

Dropped frames (QoA_VLCdropped) and Mean Opinion Score (MOS)

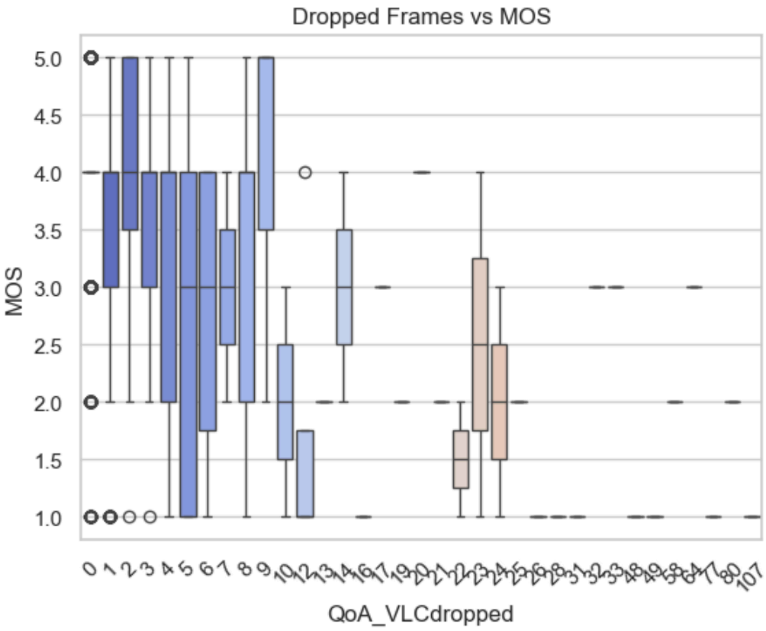


Figure 8: Dropped frames (QoA_VLCdropped) and Mean Opinion Score (MOS)

The plot indicates that an increase in dropped frames is associated with a decline in MOS, reflecting the negative impact of dropped frames on user satisfaction. For lower levels of dropped frames (0–5), MOS values are distributed more widely, with a significant proportion of higher ratings (4–5). This suggests that minimal frame drops result in a more satisfactory user experience. However, as the number of dropped frames increases beyond 10, the MOS values become concentrated at lower levels (1–2), indicating a consistent decline in satisfaction.

The presence of variability in MOS at lower dropped frame counts highlights the potential influence of additional factors, such as buffering time, bitrate, or framerate. Nonetheless, the overall trend confirms that dropped frames significantly degrade video playback quality, which directly affects user satisfaction.

Exploring Categorical Variable Distributions

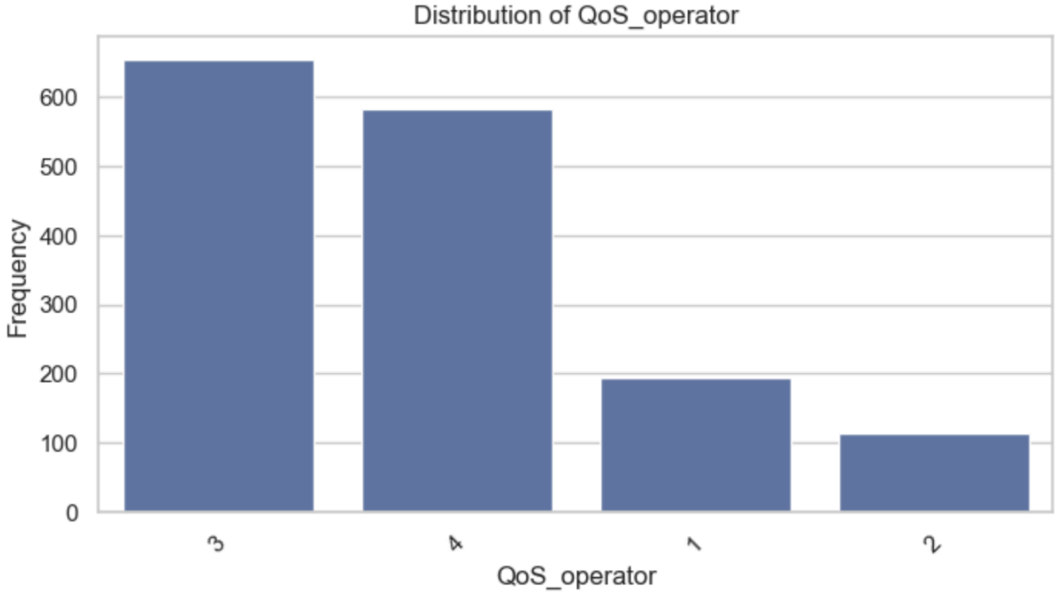


Figure 9: Distribution of QoS_operator

The distribution of QoS_operator reveals that the feature is unevenly distributed among its categories. The majority of entries belong to category 3, followed closely by category 4, which together dominate the dataset. Categories 1 and 2 have significantly lower frequencies, with category 2 being the least represented.

This imbalance indicates that the dataset is skewed toward certain operators, which might influence the model's ability to generalize across less frequent categories. Such imbalances in categorical features could lead to biased predictions, particularly if the underrepresented categories are critical to the target variable.

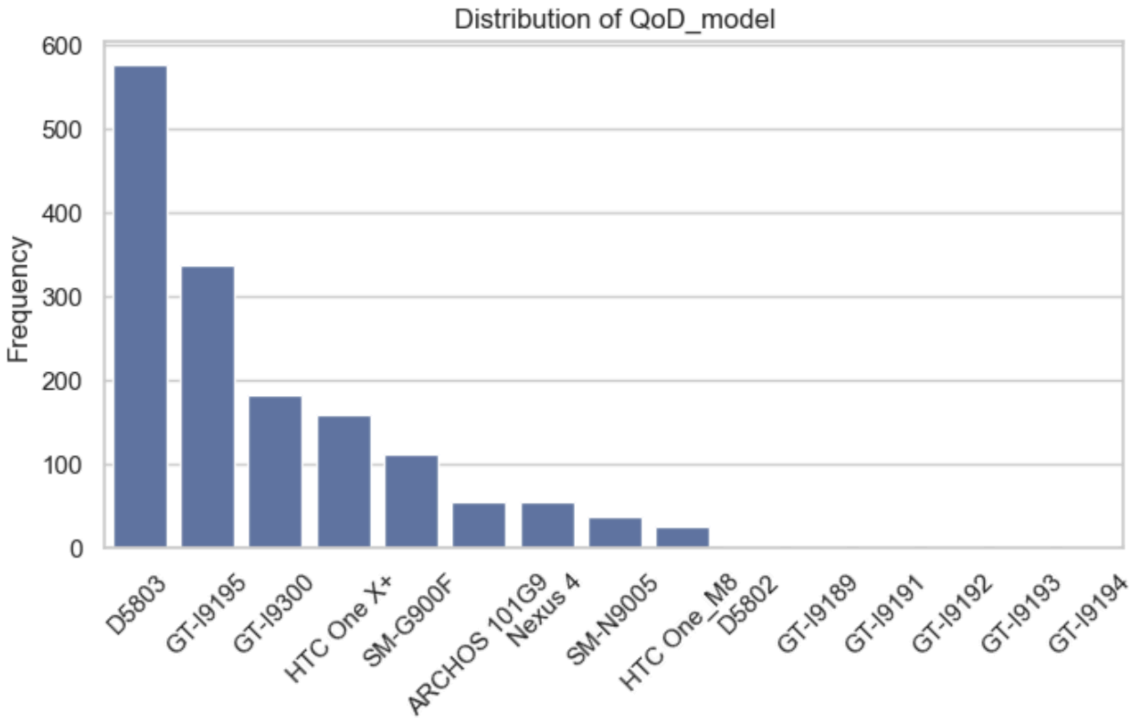


Figure 10: Distribution of QoD_model

The bar chart illustrates the distribution of the QoD_model feature, which represents different device models in the dataset. The distribution is highly imbalanced, with a few models dominating the dataset. The most frequent device model is D5803, which accounts for a significant portion of the entries, followed by GT-I9195 and GT-I9300. These top three models represent the majority of the data points, while the remaining models, such as HTC One X*, SM-G900F, and others, appear less frequently.

Notably, several models, including GT-I9191, GT-I9193, and GT-I9194, have extremely low frequencies, indicating that they are underrepresented in the dataset. This imbalance may pose challenges for machine learning models, as predictions for underrepresented device models might be less reliable.

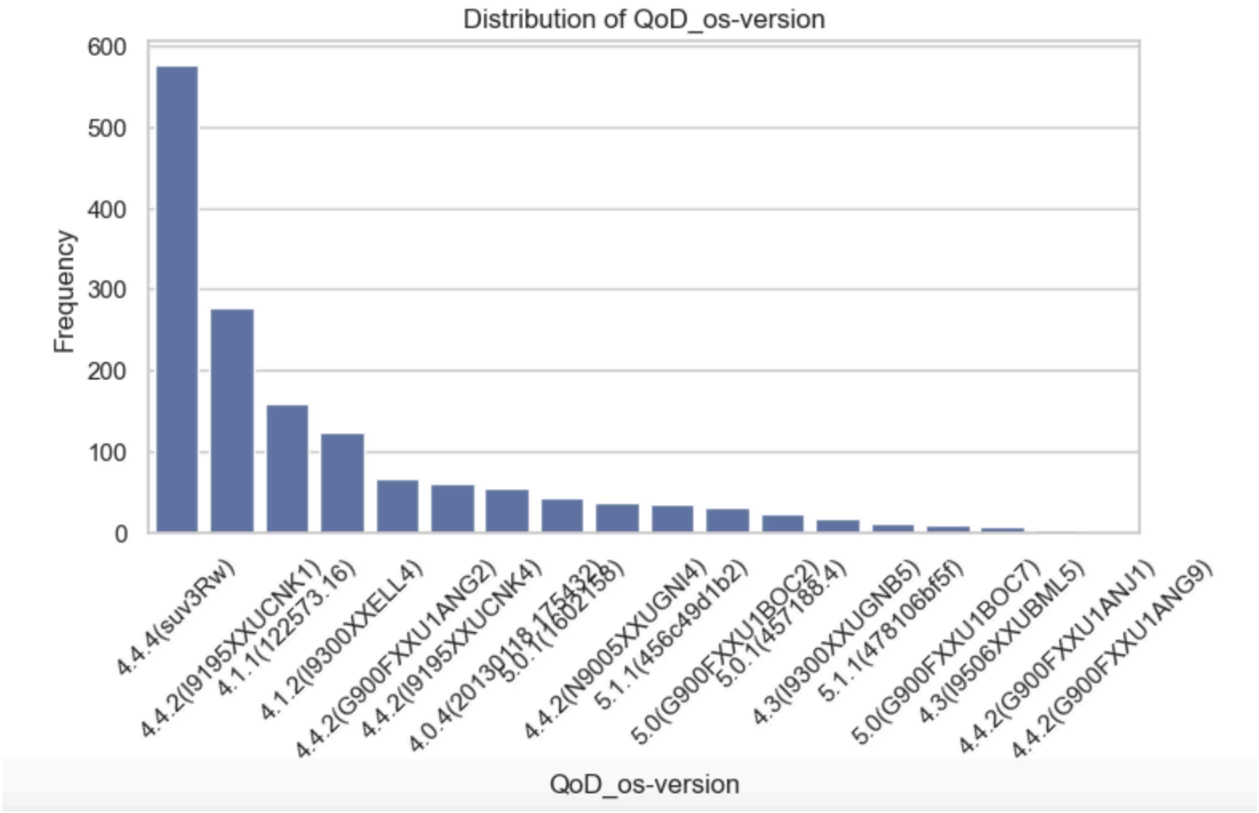


Figure 11: Distribution of QoD_os-version

Figure 11 shows a bar chart representing the distribution of the QoD_os-version feature, which shows the frequency of different operating system versions in the dataset. The distribution of QoD_os-version reveals a significant imbalance, with a few versions dominating the dataset. The 4.4.4 (suv3Rw) version is the most frequent, with a count exceeding 500, far surpassing other versions. It is followed by 4.4.2 (I9195XXUCNK1) and 4.1.1 (n22573.6), which are moderately represented. The remaining versions, such as 4.4.2 (G900FXXU1ANG2) and 5.1.1 (456c49d1b2), have much lower frequencies, with many appearing only a handful of times.

This skewed distribution indicates that a large portion of the dataset is concentrated on a limited number of operating system versions, while many versions are underrepresented. Such imbalances may influence the performance of machine learning models, as they may be biased toward the more frequent operating system versions.

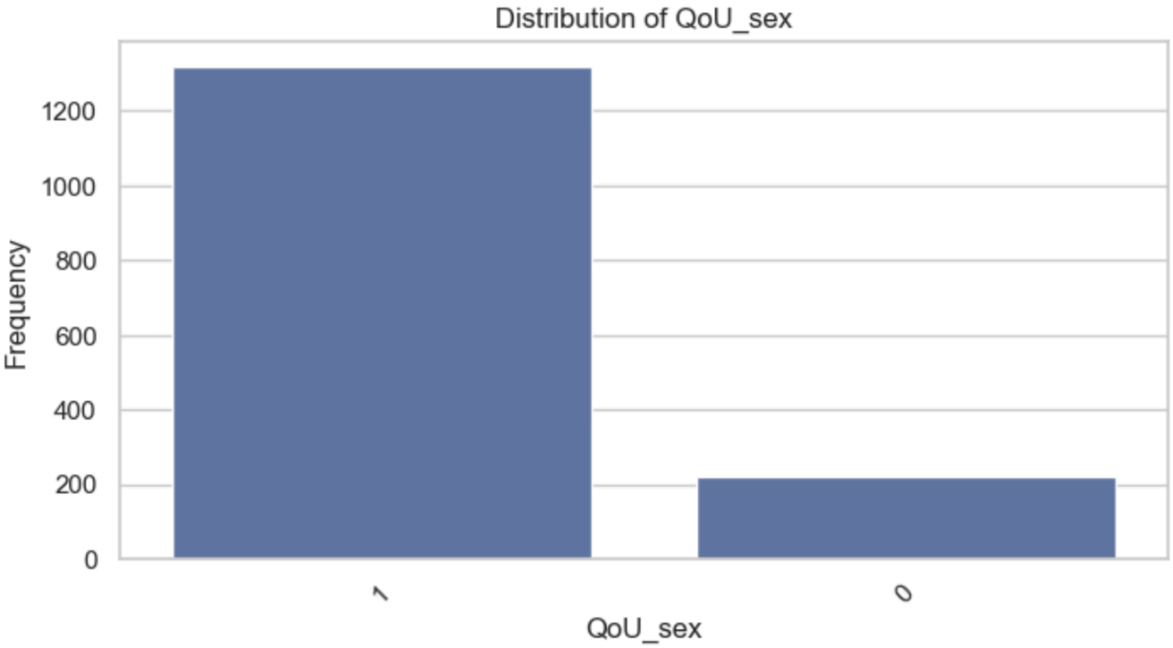


Figure 12: Distribution of QoU_sex

The distribution of QoU_sex is highly imbalanced, with the majority of entries belonging to category 1, which likely represents one gender (e.g., male). This category accounts for over 1200 entries, significantly outnumbering category 0, which represents the other gender (e.g., female) with fewer than 300 entries. This imbalance indicates that the dataset is heavily skewed towards one group, which could potentially lead to biased predictions in machine learning models.

Exploring Target Variable Distributions

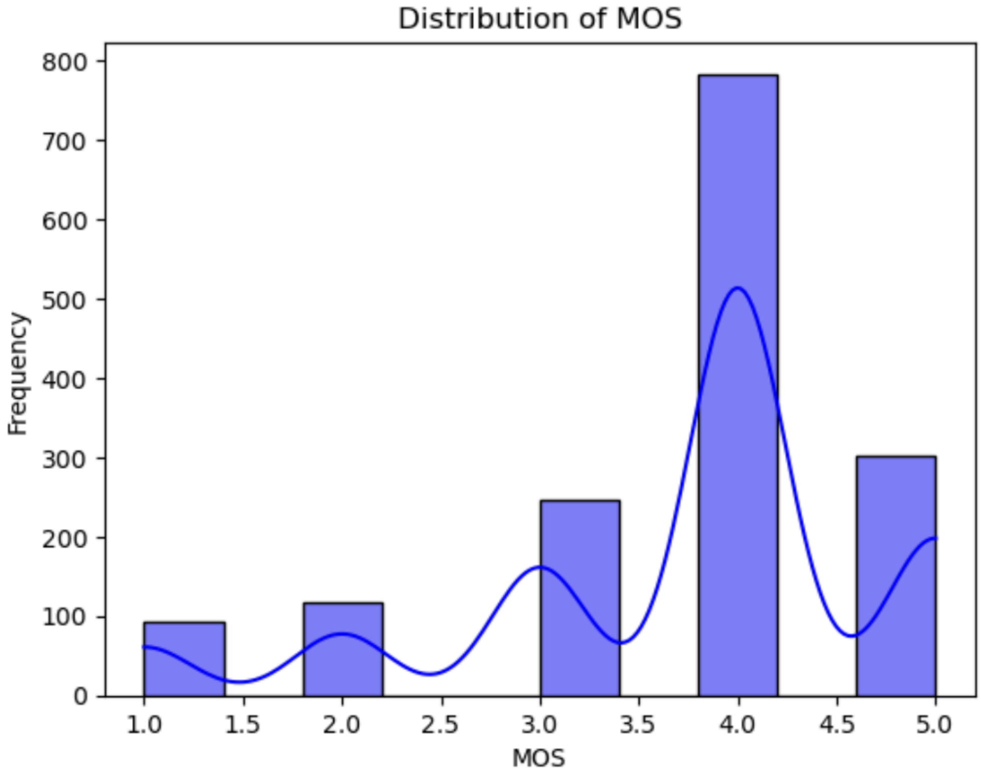


Figure 12: Distribution of MOS

The MOS distribution shows a clear right-skewed pattern, with the majority of ratings clustered around the higher end of the scale. The most frequent value is 4.0, which indicates that users tend to provide relatively high satisfaction scores. Lower scores, such as 1.0 and 2.0, are less frequent, suggesting that negative user experiences are less common in the dataset. Scores of 3.0 and 5.0 also have notable frequencies but are less prominent compared to 4.0.

This distribution highlights that the dataset is slightly imbalanced, with a bias toward higher satisfaction ratings. While this skewness is reflective of a generally positive user experience, it may influence the predictive modeling process, as the model could become biased toward predicting higher scores.

Anomaly Detection

Anomaly detection was conducted using two complementary methods: the Z-score-based approach and the Isolation Forest algorithm, each offering unique advantages in identifying outliers in the dataset.

The Z-score method was used to flag data points that deviate significantly from the mean, based on a threshold of ±3 standard deviations. Z-scores were computed for two variables: QoA_BUFFERINGtime (buffering time) and MOS (Mean Opinion Score). This straightforward statistical approach identified 26 anomalies (1.7% of the data), representing records with extreme buffering times, satisfaction scores, or both. The remaining 1516 records (98.3%) were classified as normal.

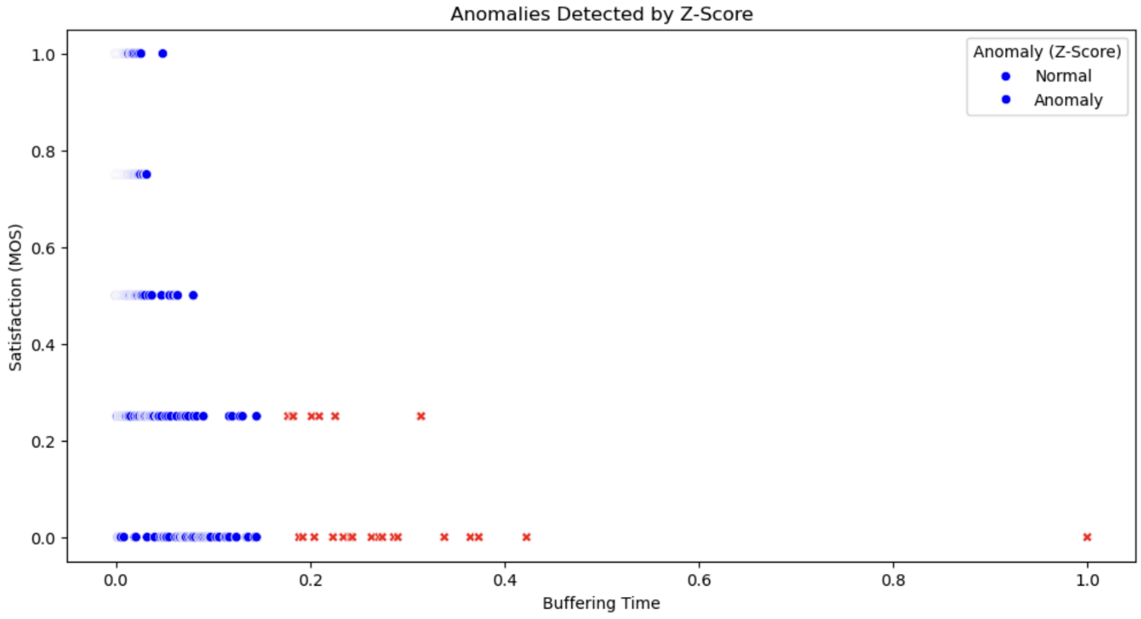


Figure 13: Anomaly Detected by Z-Score

The Isolation Forest algorithm was applied with a contamination factor of 5%, targeting approximately 5% of the dataset as potential anomalies. This method considers multidimensional feature relationships and identified 77 anomalies in total. Among these, 26 anomalies were also detected by the Z-score method, confirming their significance. Additionally, Isolation Forest flagged 51 anomalies that the Z-score method did not capture. These additional anomalies represent subtler deviations or multidimensional irregularities, which are beyond the scope of linear methods like Z-score analysis.

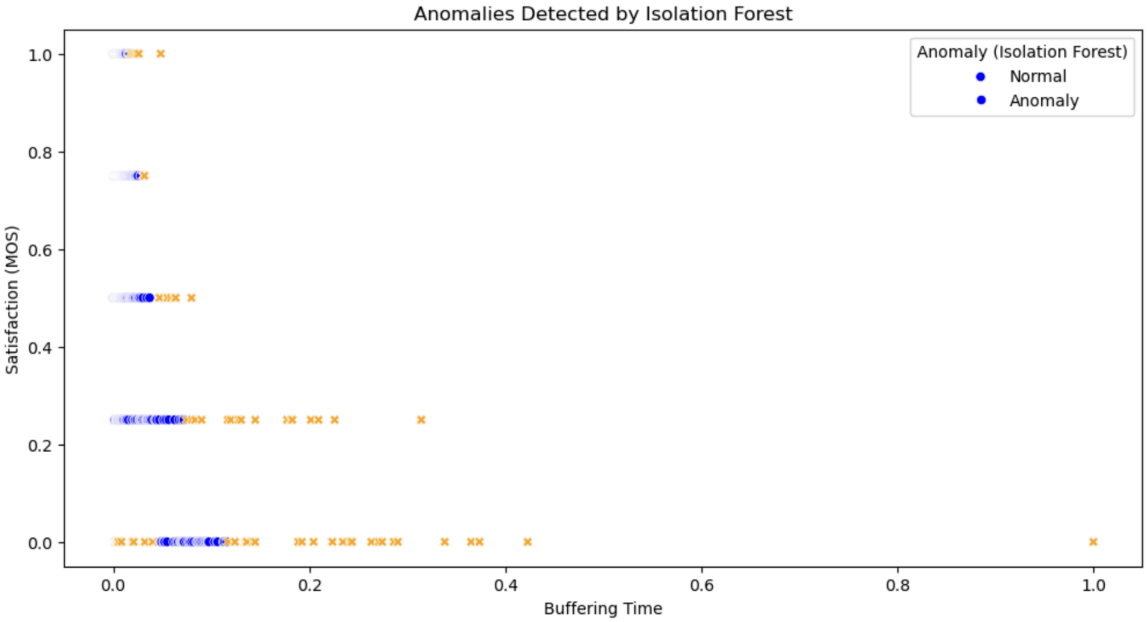


Figure 14: Anomaly Detected by Isolation Forest

Interestingly, no anomalies were flagged by the Z-score method alone, highlighting that Isolation Forest effectively incorporates linear anomaly detection while extending its capabilities to more complex, non-linear patterns. This makes Isolation Forest particularly valuable for datasets with intricate relationships between features.

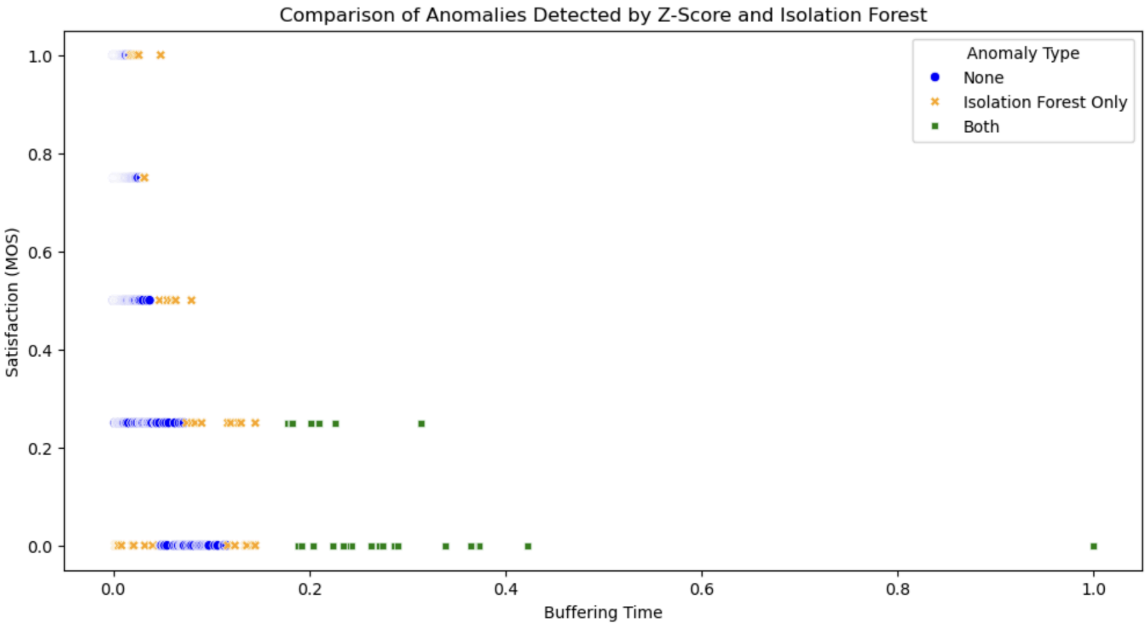


Figure 15: Anomaly Detected by Z-Score and Isolation Forest

In summary, the Z-score method is simple and effective for identifying extreme deviations in individual features, while the Isolation Forest excels at uncovering multidimensional and non-linear anomalies. Combining these methods provides a comprehensive understanding of unusual data points, ensuring critical anomalies are not overlooked. This dual approach enhances the robustness of data preprocessing and model development, offering deeper insights into potential outliers and their impact on the dataset.

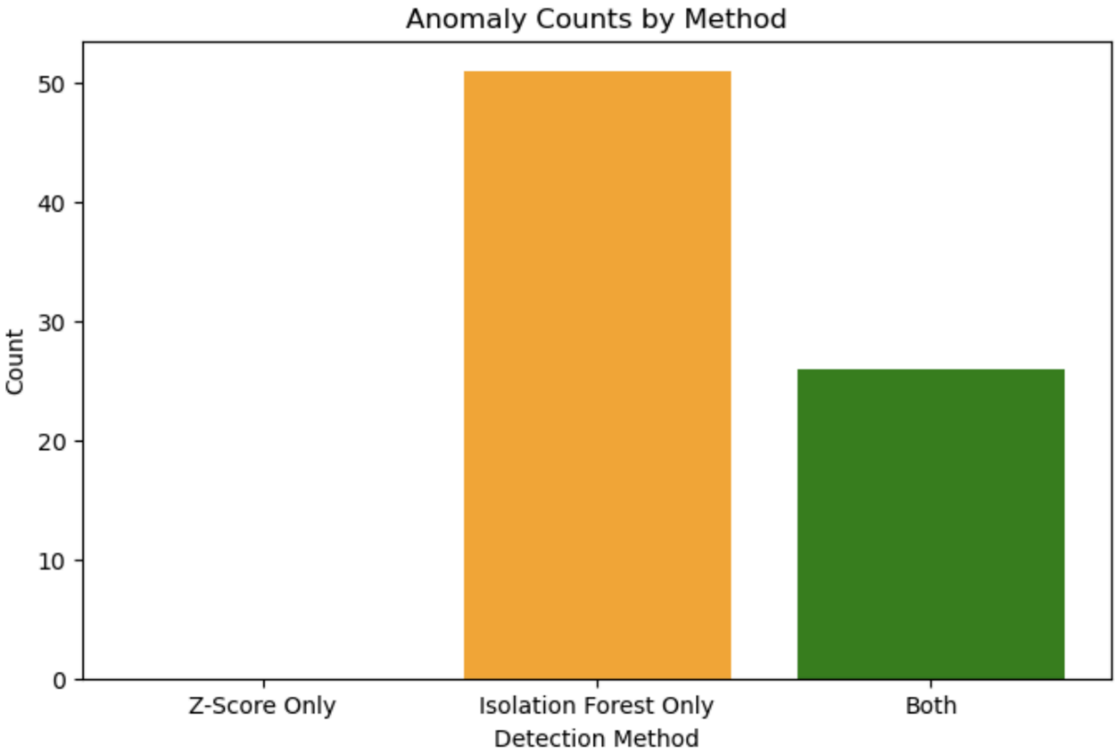


Figure 16: Anomaly Count by Different Methods: Z-Score and Isolation Forest

Model Development

Data Science Protocols Used

Feature Engineering

Derived features such as interaction terms between QoA metrics or aggregations for categorical variables might have been created to enhance predictive power. The target variable, retention, was derived from the MOS column in the dataset. A threshold-based approach was applied to categorize retention outcomes:

- Entries with MOS values greater than or equal to 4 were assigned a retention value of 1, indicating a positive outcome.
- Entries with MOS values below 4 were assigned a retention value of 0, indicating a negative outcome.

This transformation ensured the target variable was binary, and suitable for classification models. Once the target variable was created, the original MOS column was dropped from the dataset as it was no longer required for analysis.

Model Training and Evaluation

This section details the training, evaluation, and comparative performance of several machine learning classifiers, including Random Forest, XGBoost, LightGBM, and Logistic Regression.

Random Forest Classifier

The Random Forest classifier is an ensemble method that constructs multiple decision trees during training. It uses a majority-voting approach for classification and averages predictions for regression. By randomly selecting subsets of the data and features, Random Forest effectively reduces overfitting and improves generalization.

The Random Forest classifier achieved strong baseline performance on the test set, with the following metrics:

- Accuracy: 90.71%
- Classification Report
  - Precision: 0.84 (Class 0), 0.94 (Class 1)
  - Recall: 0.85 (Class 0), 0.93 (Class 1)
  - F1-Score: 0.84 (Class 0), 0.93 (Class 1)
- Confusion Matrix
  - True Positives: 303
  - True Negatives: 117

- False Positives: 20
- False Negatives: 23

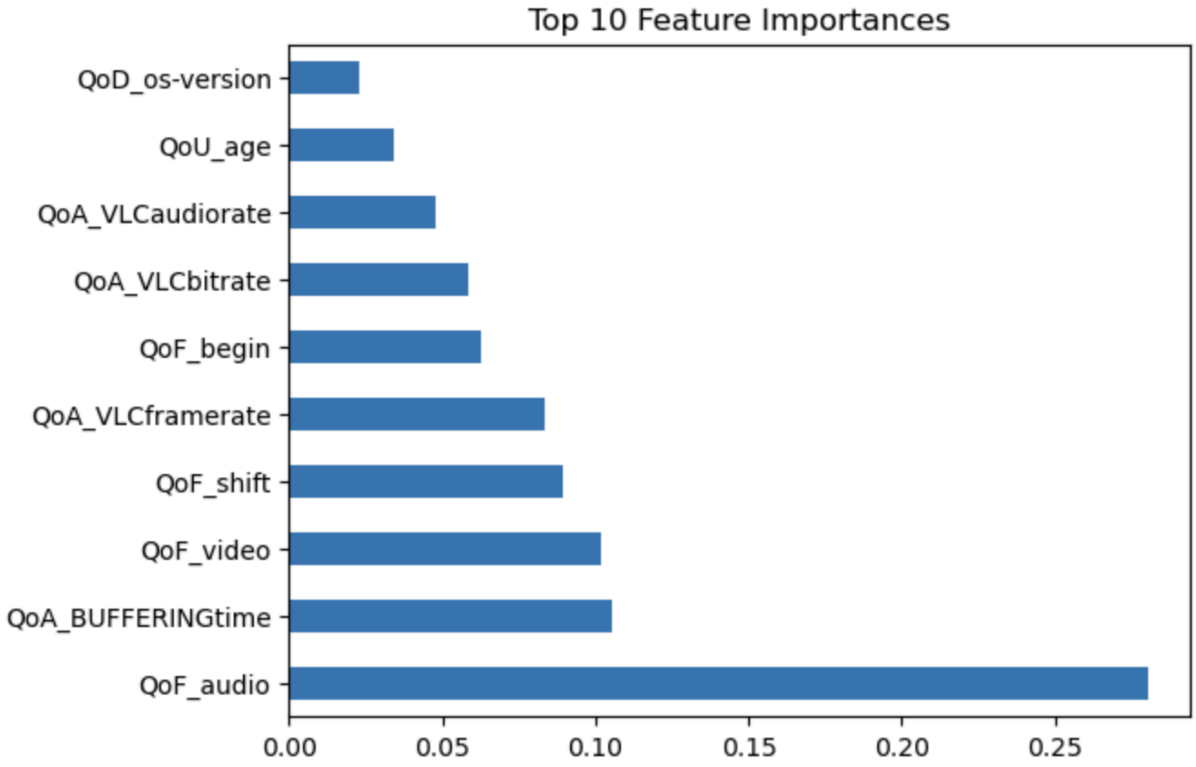


Figure 17: Top 10 Feature Importances

Feature importance analysis revealed the most significant predictors of the target variable:

1. QoF_audio: 28.04%
2. QoA_BUFFERINGtime: 10.51%
3. QoF_video: 10.18%

This analysis highlights the substantial contribution of audio quality (QoF_audio) to the classification outcome, followed by buffering time and video quality. These insights provide actionable knowledge about the relative importance of features in the prediction process.

Hyperparameter Tuning

Hyperparameter optimization was conducted using GridSearchCV to enhance the model's performance. The best combination of hyperparameters was as follows:

- **max_depth**: None
- **max_features**: 'sqrt'
- **min_samples_leaf**: 4
- **min_samples_split**: 10
- **n_estimators**: 100

Despite hyperparameter tuning, the model's accuracy slightly decreased to **90.28%**, suggesting that the baseline configuration was already near optimal.

XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a powerful gradient-boosting algorithm that minimizes a regularized objective function. It optimizes both the loss function (e.g., logistic loss for classification) and a regularization term to improve model generalization. XGBoost builds trees sequentially, with each tree learning residual errors from the previous one.

After hyperparameter tuning, XGBoost demonstrated the highest accuracy among all model which was 90.05%.

The best hyperparameters identified during tuning were:

- **learning_rate**: 0.1
- **max_depth**: 3
- **min_child_weight**: 5
- **n_estimators**: 50
- **subsample**: 0.8

The model's superior performance is attributed to its ability to learn complex patterns through gradient boosting while controlling overfitting using regularization techniques.

LightGBM Classifier

LightGBM (Light Gradient Boosting Machine) is another gradient-boosting framework that employs histogram-based learning for faster computation. Unlike XGBoost, LightGBM uses a leaf-wise tree growth strategy, splitting leaves to maximize information gain. This approach makes it computationally efficient and highly scalable.

LightGBM performed competitively, achieving an accuracy of **90.28%** after hyperparameter tuning. The optimal parameters were:

- **learning_rate**: 0.05
- **max_depth**: 3
- **n_estimators**: 50
- **num_leaves**: 31

Although slightly less accurate than XGBoost, LightGBM's comparable performance and computational efficiency make it a strong contender in applications where training time is critical.

Logistic Regression

Logistic Regression is a linear model for binary classification that estimates probabilities using the sigmoid function. The decision boundary is determined by thresholding at 0.5, and the model optimizes parameters by minimizing the negative log-likelihood. Logistic Regression demonstrated respectable baseline performance with accuracy of 90.06%. While it lacks the flexibility of ensemble methods like Random Forest or gradient boosting models, its simplicity and interpretability make it a reliable choice for straightforward classification tasks.

Comparative Analysis of Models

Model	Accuracy	Key Insights

Random Forest	90.71%	Strong baseline performance; feature importance analysis highlighted QoF_audio as the top feature.
XGBoost	90.50%	Best-performing model after tuning; excels in learning complex patterns while managing overfitting.
LightGBM	90.28%	Comparable performance to Random Forest; faster computation and efficient handling of large data.
Logistic Regression	90.06%	Simplest model with respectable performance; limited by linear decision boundaries.

Table 2: Model Comparison

Interpretation

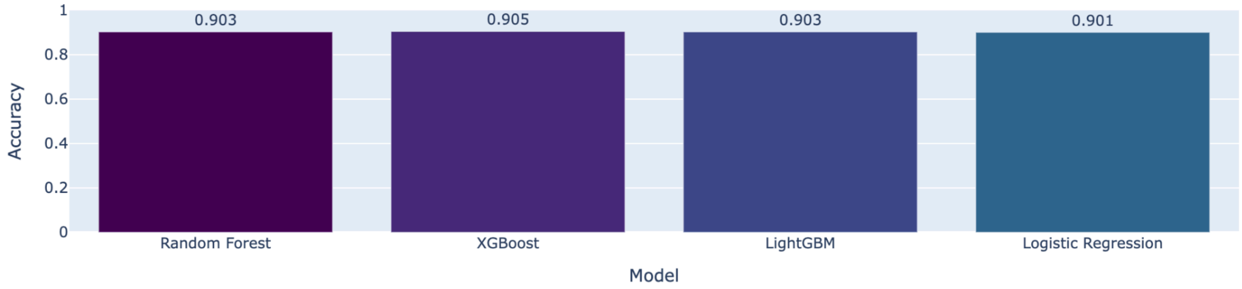


Figure 18: Model Comparison

XGBoost emerged as the best-performing model, achieving the highest accuracy of 90.50%. Its ability to handle complex interactions and regularization makes it well-suited for this task.

Random Forest provided valuable feature importance insights, emphasizing the impact of QoF_audio, QoA_BUFFERINGtime, and QoF_video on prediction accuracy.

LightGBM offered competitive accuracy while being computationally efficient, making it a viable alternative to XGBoost in time-sensitive applications.

Logistic Regression, while less sophisticated, maintained reliable performance, underscoring its value as a baseline classifier.

Overall, the choice of model depends on the application context, with XGBoost being the optimal choice for accuracy, Random Forest for interpretability, and LightGBM for efficiency.

Results

Exploratory Data Analysis (EDA)

The dataset consists of 1,543 rows and 23 features, reduced to 1,542 rows after removing duplicates. The target variable, MOS, exhibits a right-skewed distribution, with the majority of ratings clustered around higher satisfaction levels. Significant correlations were observed between QoA_BUFFERINGtime and MOS (negative correlation) and between QoA_BUFFERINGcount and MOS (negative correlation). Framerate showed a positive association with MOS.

Anomaly Detection

Z-score and Isolation Forest identified a combined total of 77 anomalies, highlighting extreme buffering times and satisfaction scores.

Model Performance

- XGBoost emerged as the best-performing model with an accuracy of 90.50%, leveraging gradient boosting for high predictive power.
- Random Forest offered strong baseline performance (90.71%) and valuable feature importance insights, emphasizing QoF_audio and QoA_BUFFERINGtime.
- LightGBM achieved comparable performance (90.28%), excelling in computational efficiency.
- Logistic Regression provided a reliable baseline accuracy of 90.06%.

Key Insights

- High buffering time and frequent buffering events significantly decrease user satisfaction.
- Video framerate positively impacts MOS, but its effect diminishes with severe buffering.

- Audio quality (QoF_audio) emerged as the most influential predictor of retention outcomes.

Limitations

By addressing the following limitations and enhancing the methodology, the project can achieve more robust results.

- The dataset is skewed, with certain device models, operating system versions, and network operators being overrepresented. This imbalance may affect the model’s ability to generalize across underrepresented categories.
- The Mean Opinion Score (MOS) relies on subjective user feedback, which can be influenced by personal biases or external conditions (e.g., mood, expectations).
- The dataset does not account for temporal changes (e.g., network congestion at different times of day), which could influence QoE metrics.
- The dataset's imbalance in categorical variables such as device models and network operators may bias predictions.

Conclusion

The analysis highlights the critical role of Quality of Service (QoS) metrics, particularly buffering time, buffering count, and audio quality, in shaping users’ Quality of Experience (QoE). By leveraging advanced machine learning models such as XGBoost, the study accurately identified at-risk users and provided actionable insights to optimize service quality. While maintaining a high bitrate and framerate is essential, prioritizing smooth playback and minimizing interruptions is key to improving user satisfaction and retention. The findings emphasize the need for proactive anomaly detection and feature optimization to address technical outliers and improve predictive accuracy. Future work should incorporate temporal analysis and investigate other potential QoE factors, such as network latency and environmental influences, to further enhance model robustness and generalizability.

References

[1] S. Moteau, F. Guillemin, and T. Houdoin, *Correlation between QOS and QOE for HTTP YouTube content in Orange cellular networks*, vol. 24. 2017, pp. 1–6. doi: 10.1109/latincom.2017.8240146.

Link to Code: https://github.com/Pratistha-99/Pokemon/tree/main