Model Governance: Optimizing Music for Video Success on YouTube

Business Analytics Capstone - BA723

Saksham Tiwari-301334725
Centennial College
David Parent
August 16, 2024

Table of Content

Validation, Monitoring and Governance

1.0.Variable Level Monitoring

Variable Level Monitoring is essential for ensuring that the key predictors used in the YouTube music video views prediction model remain consistent and relevant over time. This process involves continuously tracking the behavior of these variables to detect any shifts, anomalies, or drifts that could impact the model's performance. By establishing baseline statistics during the model development phase, such as the mean, standard deviation, and distribution ranges, we set a reference point against which future data is compared.

1.1. Model Build Variable Level Statistics

Providing a foundation to the behavior of each key variable in the dataset during model development phase the statistics are essential in establishing a baseline against which future data will be monitored. By capturing the detailed metrics such as mean, median, standard deviation and distribution ranged, we can ensure that the model's input remains consistent and relevant over time.

Key variables:

1. Danceability:
   - Mean- 0.645
   - Median- 0.640
   - Standard Deviation- 0.123
   - Range: 0.1- 0.97
2. Energy:
   - Mean- 0.710
   - Median- 0.705
   - Standard Deviation- 0.150
   - Range- 0.12 to 0.99
3. Loudness:
   - Mean: -5.6db
   - Median: -5.8db
   - Standard Deviation: 4.1 db
   - Range: -60db to 0 db
4. Tempo:
   - Mean: 115 BPM
   - Median: 118 BPM
   - Standard Deviation: 25 BPM
   - Range: 60 BPM to 200 BPM
5. Duration_minutes:
   - Mean: 3.5 minutes
   - Median: 3.4 minutes
   - Standard Deviation: 1.0 minute
   - Range: 1.5 minutes to 7.5 minutes
6. Licensed (True/False):
   - True: 80%

- False: 20%

7. Official Video (True/False):
   - True: 85%
   - False: 15%

These baseline statistics are critical for ongoing model monitoring. Any substantial changes in these values when new data is introduced could indicate data drift or other anomalies, which could degrade the model's performance. Regular comparison of future data against these statistics helps maintain the model's accuracy and effectiveness, ensuring it continues to deliver reliable predictions aligned with business goals.

1.2. Acceptable Ranges

The predefined thresholds for each key variable in the model, derived from the statistical analysis conducted during the model development phase. These ranges are to ensure that incoming data have a reference to what the model was developed upon not only looking to set a standard but as a scope for improvement. But to work on the model base we establish a framework to detect anomalies, outlier, and potential drift that could impact the model's performance.

Acceptable ranges for each variable are determined typically using the mean ± 3 standard deviation or other thresholds.

1. Danceability:
   Range: 0.3 to 1.0
   This range covers majority of the observed data.
2. Energy**:**
   Range**:** 0.4 to 1.0
   Values below 0.4 are considered atypical and may not represent the typical energy levels of successful tracks.
3. Loudness**:**
   Range**:** -20 dB to 0 dB
   This range includes the most common loudness levels for music tracks, filtering out extremely low values that are less likely to correspond to commercial tracks.
4. Tempo**:**
   Range**:** 60 BPM to 200 BPM
   This range accommodates a wide variety of music genres while excluding tempos that are either too slow or too fast for typical tracks.
5. Duration_minutes**:**
   Range**:** 2 to 6 minutes
   Tracks outside this range are likely to be either too short or too long for typical consumption, which could affect their viewership.
6. Categorical Variables (Licensed, Official Video)**:**
   Range: 0 or 1
   These binary variables should only take values of 0 or 1. Any deviation indicates a data entry or processing error.

1.3. Caps & Floors

Caps and Floors are limits imposed on continuous variables in the model to manage extreme values that could skew predictions or lead to model instability. By capping (setting an upper limit) and flooring (setting a lower limit) values, we ensure that the data fed into the model remains within a controlled range, preventing outliers from disproportionately influencing the model's outputs.

Determining Caps and Floors**:**

Caps and floors are typically set based on the distribution of each variable during the model development phase, often using the mean ± 3 standard deviations or based on domain knowledge.

- Danceability:
  - Floor: 0.3 (Minimum allowed value)
  - Cap: 1.0 (Maximum allowed value)
  - Values below 0.3 or above 1.0 are rare and may not represent realistic track characteristics.
- Energy:
  - Floor: 0.4 (Minimum allowed value)
  - Cap: 1.0 (Maximum allowed value)
  - Tracks with energy levels outside this range are uncommon and could lead to inaccurate predictions.
- Loudness:
  - Floor: -20 dB (Minimum allowed value)
  - Cap: 0 dB (Maximum allowed value)
  - Extreme loudness values outside this range could represent noise or erroneous data.
- Tempo:
  - Floor: 60 BPM (Minimum allowed value)
  - Cap: 200 BPM (Maximum allowed value)
  - Tempos below 60 BPM or above 200 BPM are unlikely for most commercial music genres.
- Duration_minutes:
  - Floor: 2 minutes (Minimum allowed value)
  - Cap: 6 minutes (Maximum allowed value)
  - Tracks shorter than 2 minutes or longer than 6 minutes are less common and may skew the model's predictions.

1. Monitoring and Adjustment: The capped and floored values are monitored over time. If trends shift and certain values begin to regularly approach or exceed these limits, the caps and floors may be revisited and adjusted as necessary.
2. Impact on Predictions: By controlling for extreme values, caps and floors help maintain the consistency and accuracy of the model’s predictions, ensuring that the outputs are not disproportionately influenced by anomalous inputs.

1.4. Missing Values

Handling missing values effectively is crucial for maintaining the accuracy and reliability of the predictive model. The approach taken depends on the nature of the data, the proportion of missing values, and the importance of the affected features within the model.
Importance of Addressing Missing Values:

1. Preserving Data Integrity: Missing values can introduce bias and reduce the representativeness of the dataset. Proper handling ensures that the model is trained and evaluated on complete and accurate data.
2. Ensuring Model Accuracy: Unaddressed missing values can lead to incorrect model predictions. By imputing or excluding these values, we maintain the model's predictive power and avoid distortions in the output.
3. Consistency in Data Inputs: Consistent handling of missing data ensures that the model receives inputs that align with the expectations set during the training phase, reducing the risk of encountering unexpected errors during prediction.

Approaches to Handling Missing Values:

1. Numerical Variables:
   - Imputation Using the Median: For numerical features like Danceability, Energy, Loudness, Tempo, and Duration_minutes, missing values are imputed using the median value of the feature. The median is less sensitive to outliers than the mean, making it a robust choice for imputation.
   - Example: If Tempo has a missing value, it would be replaced with the median Tempo value from the dataset.
2. Categorical Variables:
   - Imputation Using the Mode: For categorical features like Licensed and Official Video, missing values are imputed using the mode, which is the most frequently occurring category. This approach ensures that the most representative category is used to fill in the gaps.
   - Example: If the Licensed feature has a missing value, it would be replaced with either True or False, depending on which is more common in the dataset.
3. Flagging Missing Values:
   - In cases where missing values are significant or indicative of an underlying issue, a flag is created to track and monitor the presence of missing data. This flag can then be used as an additional feature in the model, helping to account for the impact of missing data on predictions.
4. Exclusion of Missing Values:
   - If the proportion of missing values is very low (e.g., less than 1%), and the impact on the dataset is minimal, the affected rows may be excluded entirely. This approach is typically reserved for cases where imputation is not feasible or could introduce bias.

Monitoring Missing Values:

1. Regular Checks: The presence of missing values is regularly monitored to ensure that the data pipeline remains robust. Any sudden increase in missing values may indicate a problem in data collection or processing that needs immediate attention.

2. Dynamic Imputation: As new data is introduced, the imputation strategy may be adjusted based on the evolving distribution of the data. This ensures that the model continues to receive accurate and representative inputs.
3. Impact Assessment: The impact of imputed values on the model's predictions is periodically assessed. If imputation is found to be affecting model accuracy, alternative strategies, such as more sophisticated imputation techniques or retraining the model, may be considered.

1.5. Variable Drift Monitoring Tolerance

Monitoring for drift ensures that the model remains reliable and that its predictions are based on data that is consistent with what it was trained on.
Importance of Monitoring for Variable Drift:

1. Maintaining Model Accuracy: Changes in the distribution of variables can lead to a decline in model performance. By monitoring drift, we can identify when the model may need recalibration or retraining.
2. Early Detection of Data Shifts: Drift monitoring helps detect shifts in the underlying data early, allowing for proactive measures to prevent degradation in model performance.
3. Ensuring Data Consistency: Continuous monitoring ensures that the data fed into the model remains consistent with the data on which the model was trained, preserving the accuracy of predictions.

Setting Drift Tolerance:

- High-Importance Variables:
  - Loudness, Duration_minutes: Tolerance: ±5%
  - Explanation: These variables are critical to the model's predictions, and even small shifts in their distributions can significantly impact model performance. A tight tolerance ensures that any drift is detected early.
- Moderate-Importance Variables:
  - Danceability, Energy, Tempo: Tolerance: ±7%
  - Explanation: These variables are important but have a bit more flexibility in their acceptable range. Moderate tolerance allows for natural variations without compromising model performance.
- Low-Importance Variables:
  - Licensed, Official Video, Key_8.0: Tolerance: ±10%
  - Explanation: These variables, while still relevant, can tolerate more drift before affecting the model. A wider tolerance is acceptable, as minor shifts in these features are less likely to degrade the model's accuracy.

Monitoring and Responding to Drift:

1. Regular Monitoring: Continuous monitoring of the key variables is conducted to ensure they remain within their established drift tolerance. This is typically done through automated systems that compare the current data distribution against the baseline statistics.

2. Control Charts: Control charts are often used to visualize the drift over time. These charts display the variable's performance against the upper and lower control limits (set by the drift tolerance). If the variable crosses these control limits, an alert is triggered.
3. Impact Assessment: If drift is detected, the impact on model performance is assessed. This involves checking whether the drifted variable still correlates with the target variable in the same way it did during training.
4. Actions on Drift Detection:
    - Minor Drift: If the drift is minor and within tolerance, it may be monitored for further changes without immediate action.
    - Significant Drift: If the drift exceeds the tolerance level, the model may need recalibration, retraining, or even redesign to accommodate the new data distribution.
    - Model Retraining: If significant and persistent drift is detected, the model may require retraining with updated data to ensure that it continues to deliver accurate predictions.

2.0. Model Monitoring, Health & Stability

Model Monitoring, Health & Stability is essential to ensure that the YouTube music video views prediction model continues to deliver accurate and reliable predictions as data evolves. This involves tracking the model's performance over time, detecting any signs of degradation, and taking corrective actions to maintain its health and stability. Given the dynamic nature of YouTube content and viewer behavior, continuous monitoring is crucial to adapt to changing trends.

Importance of Model Monitoring:

1. Sustaining Predictive Accuracy: The YouTube music video views model relies on various features such as Loudness, Duration, Danceability, and others. As new music trends emerge or platform algorithms change, the data characteristics may shift. Continuous monitoring ensures that the model remains aligned with current trends and continues to provide accurate predictions.
2. Detecting Shifts in Data: Viewer preferences, content types, and engagement patterns on YouTube can change over time, potentially leading to shifts in the data distributions that the model relies on. Monitoring for these shifts helps in early detection, allowing for timely recalibration or retraining of the model.
3. Ensuring Business Continuity: Accurate predictions are essential for guiding decisions related to content creation, marketing, and promotions. Monitoring the model's health ensures that stakeholders can trust the insights provided by the model, supporting better decision-making.

Key Components of Model Monitoring:

1. Performance Metrics Monitoring:
    - The model's key performance metrics, such as RMSE, MAE, and MAPE, are tracked regularly. For example, an increase in RMSE might indicate that the model is becoming less accurate at predicting views, signaling a need for investigation.

2. Data Quality Checks:
   - Regular checks on incoming data ensure that it remains clean, consistent, and relevant. This includes monitoring for missing values, outliers, and ensuring that all variables remain within their acceptable ranges. For instance, a sudden change in the distribution of Loudness or Duration might indicate a shift in content trends that requires attention.
3. Variable Drift Detection:
   - As viewer behavior and content preferences evolve, key variables like Energy or Tempo may drift from their original distributions. Monitoring this drift helps identify when the model may no longer be accurately reflecting the current environment. Drift beyond set tolerances will prompt a review and possible model recalibration.
4. Model Retraining and Calibration:
   - Depending on the results of performance and drift monitoring, the model may require periodic retraining with the latest data. This ensures that the model adapts to new trends in YouTube content and viewer engagement.
5. Stability Assessment:
   - Stability assessments involve checking the consistency of the model's predictions over time. If the model starts showing greater variability in its predictions or becomes overly sensitive to specific features, it may indicate a loss of stability, necessitating recalibration.

Monitoring Framework:

1. Automated Monitoring System:
   - An automated system continuously tracks performance metrics, data quality, and variable drift. Alerts are triggered when any metric exceeds predefined thresholds, allowing for immediate investigation.
2. Scheduled Reviews:
   - Regular model performance reviews, conducted quarterly or semi-annually, ensure that the model's outputs remain aligned with current data trends and business needs. These reviews provide an opportunity to assess the model's overall health and make decisions about recalibration or retraining.
3. Control Limits and Action Thresholds:
   - Control limits are set for each performance metric and key variable. For example, if the RMSE exceeds a certain threshold, this would trigger an analysis to determine whether the model requires retraining.
4. Health Reports:
   - Periodic health reports document the model's performance, any detected drifts, and actions taken. These reports ensure transparency and provide a record of model adjustments over time.

2.1. Initial Model Fit Statistics

The Initial Model Fit Statistics provide a baseline evaluation of the model's performance during the training and validation phases. These statistics are crucial for understanding how well the model fits the data it was trained on and how accurately it can predict outcomes on unseen data.

In the context of the YouTube views prediction model, these fit statistics help establish a benchmark against which future performance will be compared.
Key Fit Statistics:

- Mean Error (ME): Measures the average difference between predicted and actual values. A near-zero ME indicates that the model's predictions are not systematically biased in one direction.
  - ME = 566,334.7585 for the Full Regression model.
- Root Mean Squared Error (RMSE): Quantifies the model's prediction error in the units of the target variable (YouTube views). RMSE is sensitive to large errors, making it a useful metric for identifying significant prediction inaccuracies.
  - RMSE = 24,625,953.6892 for the Backward Regression model.
- Mean Absolute Error (MAE): Represents the average magnitude of errors in a set of predictions, without considering their direction. MAE provides a straightforward measure of model accuracy.
  - MAE = 19,528,362.4278 for the Backward Regression model.
- Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE): These metrics express the prediction error as a percentage of the actual values, offering a scale-independent measure of accuracy.
  - MAPE = 306.9864% for the Backward Regression model.

These fit statistics establish a reference point for monitoring the model's health and stability over time. Any significant deviation from these initial values during production may indicate a decline in model performance, triggering further investigation.

2.2. Parameter #1

RMSE is selected as a key performance parameter because it reflects the model's ability to predict YouTube views with precision. RMSE is particularly valuable because it emphasizes larger errors more than smaller ones, making it sensitive to predictions that are far off the mark.
Why RMSE is Critical:

- Sensitivity to Large Errors: RMSE's sensitivity to large deviations is crucial for this model, as significant errors in predicting YouTube views can lead to poor decision-making.
- Interpretability: RMSE is expressed in the same units as the target variable (YouTube views), making it easy to interpret the magnitude of errors.
- Performance Monitoring: A rising RMSE over time would indicate that the model's predictions are becoming less accurate, signaling potential issues like data drift or changes in viewer behavior.

2.3. Parameter #2

MAE is selected as a second key performance parameter because it provides a straightforward measure of prediction accuracy by averaging the absolute errors, giving equal weight to all errors, regardless of their size.
Why MAE is Important:

- Simplicity: MAE is easy to understand and interpret, making it a useful metric for communicating model performance to stakeholders.

- Robustness: Unlike RMSE, MAE does not exaggerate the impact of large errors, offering a more balanced view of overall model performance.
- Monitoring and Comparison: Tracking MAE alongside RMSE allows for a more comprehensive understanding of the model's prediction quality. A consistent MAE paired with an increasing RMSE might indicate that while small errors are stable, large errors are becoming more frequent.

3.0. Risk Tiering (e.g., no action, report, refit, rebuild)

The purpose of risk tiering is to ensure that appropriate actions are taken based on the severity of the issue detected during model monitoring. By defining specific thresholds and corresponding actions, the model governance framework ensures that the model remains accurate, reliable, and aligned with business objectives.
Purpose of Risk Tiering:

1. Prioritized Response: Risk tiering helps prioritize responses based on the severity of the risk. This ensures that critical issues are addressed promptly, while less severe issues are monitored and managed over time.
2. Consistency in Decision-Making: Having a clear risk tiering system ensures that responses to model performance issues are consistent and aligned with predefined governance policies.
3. Minimizing Business Impact: By categorizing risks and defining actions, the model governance framework minimizes potential negative impacts on business decisions and outcomes.

Risk Tiers and Associated Actions:

1. Low Risk (No Action):
    - Description: Minor deviations in performance metrics or slight variable drift that are within acceptable tolerance levels.
    - Example: A slight increase in RMSE that remains within the predefined control limits.
    - Action: No immediate action required. Continue regular monitoring to ensure that the deviation does not escalate.
2. Moderate Risk (Report):
    - Description: Noticeable changes in model performance or variable drift that are close to exceeding predefined thresholds but do not yet pose a significant threat to model accuracy.
    - Example: RMSE or MAE approaching the upper control limit or variable drift nearing the tolerance level.
    - Action: Document the issue and report it to relevant stakeholders. Increase the frequency of monitoring and prepare for potential intervention if the trend continues.
3. High Risk (Refit):
    - Description: Significant performance degradation, variable drift that exceeds tolerance levels, or other issues that could materially impact the model’s predictions.

  - Example: A sustained increase in RMSE or MAE beyond the acceptable range, or a key variable's distribution shifting significantly from the baseline.
  - Action: Initiate a model refit or recalibration using recent data. This may involve adjusting model parameters, retraining the model with updated data, or revising the feature set to account for new trends.
4. Unacceptable Risk (Rebuild):
  - Description: Critical failures in model performance, severe data drift, or ethical concerns that render the current model unsuitable for use.
  - Example: A complete breakdown in the model's predictive accuracy, or a major shift in the underlying data that the current model cannot accommodate.
  - Action: Immediately halt the use of the current model. Begin the process of building a new model from scratch, incorporating updated data, revised features, and potentially new modeling techniques. Implement interim measures, such as using alternative models or manual decision-making processes, until the new model is ready for deployment.

Implementation of Risk Tiering:

1. Threshold Definitions: Clearly defined thresholds for each risk tier ensure that actions are taken based on objective criteria. These thresholds are set during the model development phase and are refined as more data becomes available.
2. Automated Alerts: Automated monitoring systems can be configured to trigger alerts when performance metrics or variable drifts exceed the thresholds associated with each risk tier. This enables timely and consistent responses.
3. Documentation and Reporting: For Moderate to Unacceptable Risks, detailed documentation is required. Reports should include the nature of the issue, the analysis conducted, the actions taken, and the outcomes. This documentation supports transparency and accountability in the model governance process.
4. Continuous Review: The risk tiering framework should be regularly reviewed and updated based on the evolving business environment, data trends, and model performance. This ensures that the risk management strategy remains effective and relevant.