Findings Report – Crowd Energy Prediction

This analysis was conducted to predict Crowd Energy for future concerts and to understand the key drivers influencing audience engagement. The focus was on building a realistic, leakage-free, and interpretable machine learning pipeline suitable for real-world deployment.

1. Data Cleaning & Validation

Data cleaning prioritized realism and causality. Post-event variables such as merch sales were removed to prevent data leakage. Show dates were standardized and missing dates were imputed using venue-specific weekday probability distributions. Ticket prices were converted to a single currency, invalid sensor readings were corrected, and unrealistic values were filtered.

2. Exploratory Data Analysis (EDA)

EDA revealed strong venue-specific effects on crowd energy. The singer's hypothesis of lower Tuesday energy was partially supported, while pricing alone showed weak global correlation with energy. Crowd size and volume exhibited non-linear relationships, motivating feature engineering and non-linear models.

3. Feature Engineering

Non-linear transformations were introduced to better represent real-world behavior. Log-transformed crowd size captured diminishing returns, price-per-person reflected perceived affordability, and squared volume emphasized extreme loudness effects. These features improved model stability and generalization.

4. Model Selection & Validation

A Linear Regression model served as a baseline. Ensemble models including Random Forest and Gradient Boosting were evaluated using 5-fold cross-validation and RMSE as the metric. Although Gradient Boosting achieved marginally lower RMSE, Random Forest was selected due to its robustness, stability, and lower sensitivity to hyperparameters.

5. Key Takeaways

- • Crowd energy is highly venue-dependent.
- • Non-linear effects dominate audience engagement.
- • Model performance is limited by inherent human and environmental noise.