Findings Report – Machine Learning Task 1

The World Tour of ’99: The Data Detective

Name: Prakash Kumar
Program: B. Tech In Data Science and Artificial Intelligence
Event: Coding Club IIT Guwahati – Coding Week
Task: Crowd Energy Prediction

---

1. Data Cleaning and Preparation

The provided tour log dataset contained several real-world data quality issues that required careful preprocessing before analysis and modeling. These included inconsistent date formats, mixed currencies in ticket pricing, missing and zero sensor readings, and extreme outliers caused by logging or sensor failures.

Dates were standardized into a uniform datetime format to enable temporal feature extraction. Ticket prices originally recorded in USD, GBP, and EUR were converted into USD using the exchange rates specified in the task description. Missing or zero values in sensor-related features were treated as systematic failures and handled through appropriate imputation or exclusion depending on context. Clearly impossible values were identified as outliers and removed to prevent distortion of model learning.

Special care was taken to avoid data leakage. Any features that were recorded after the show were excluded from the modeling process to ensure that predictions reflected only information that would realistically be available before a concert. After cleaning, the dataset was consistent, reliable, and suitable for exploratory analysis and predictive modeling.

---

2. Exploratory Data Analysis & Venue-wise Findings

Exploratory Data Analysis (EDA) revealed that Crowd Energy is highly venue-dependent, with different factors influencing audience response at each location. The lead singer’s scribbles were treated strictly as hypotheses and tested against observed data.

V_Alpha – The Holy Grounds

Crowd energy at V_Alpha showed sensitivity to sound-related and regulatory constraints. Shows that approached higher noise levels or exceeded implicit thresholds tended to experience a drop in crowd energy. This supports the singer’s hypothesis regarding noise limits at this converted monastery venue.

However, ticket price did not show a strong or consistent relationship with crowd energy at this venue, indicating that audience engagement here is less price-sensitive than assumed.

Singer's Theory Verdict: *Partially correct*

V_Beta – The Vampire's Den

At V_Beta, timing emerged as one of the strongest predictors of crowd energy. Late-evening and night shows consistently resulted in higher crowd energy compared to earlier time slots. This aligns closely with the gothic nightclub atmosphere of the venue and validates the singer's belief that timing matters significantly here.

Other factors such as pricing and sensor readings had comparatively weaker influence.

Singer's Theory Verdict: *Mostly correct*

V_Gamma – The Snob Pit

V_Gamma exhibited a clear relationship between ticket price and crowd energy. Extremely low prices appeared to reduce the sense of exclusivity, while very high prices negatively impacted attendance-driven energy. The highest crowd energy was observed within a moderate price range.

This confirms the singer's suspicion that pricing plays an important role, but also reveals that the relationship is nonlinear rather than straightforward.

Singer's Theory Verdict: *Correct but oversimplified*

V_Delta – The Mosh Pit

V_Delta showed the highest variance in crowd energy and the least predictable behavior. Traditional predictors such as price, timing, and basic sensor features did not consistently explain energy levels. This suggests that unobserved factors, such as crowd composition or spontaneous audience behavior, may dominate at this venue.

This chaotic pattern explains the singer's conflicting and unreliable memories about what drives energy at the Mosh Pit.

Singer's Theory Verdict: *Inconclusive*

---

3. Feature Engineering

Based on insights from EDA, several new features were engineered to improve model performance. Temporal features such as show hour and day were extracted from standardized dates. Ticket prices were normalized, and interaction features between venue type and pricing or timing were introduced to capture venue-specific behavior.

Features suspected of introducing leakage were explicitly excluded. The resulting feature set balanced expressiveness with robustness, allowing the model to generalize better under distribution shifts and unseen categories in the test data.

---

4. Model Selection and Hyperparameter Tuning

A tree-based regression model was selected for this task due to its ability to capture nonlinear relationships and complex feature interactions without requiring extensive manual transformation. Such models are also robust to outliers and mixed feature scales, which suited the nature of the dataset.

Hyperparameter tuning was conducted using k-fold cross-validation to ensure reliable performance estimation. Key hyperparameters—including tree depth, number of estimators, and learning-related parameters—were systematically explored over defined ranges. Model performance using tuned hyperparameters was compared against default settings.

The tuned model demonstrated a clear improvement over the baseline configuration, validating the importance of proper hyperparameter optimization. This process ensured that the final model was not overfitted and could generalize effectively to unseen test data.

---

5. Conclusion

This analysis highlights that crowd energy is influenced by different factors across venues, and no single rule applies universally. While some of the lead singer’s theories were supported by data, others were incomplete or unreliable, reinforcing the need for evidence-based decision-making.

Through rigorous data cleaning, thoughtful feature engineering, and validated model tuning, a robust regression model was developed to predict crowd energy effectively. The findings provide actionable insights for future tour planning and demonstrate the practical application of machine learning in a noisy, real-world setting.

End of Report