

The World Tour of '99: Crowd Energy Analysis

-- Akansha Das

Mechanical Engineering

OBJECTIVE:

The objective is to analyse factors influencing crowd energy for concerts of a particular band and build a machine learning model capable of predicting crowd energy using previous concert data. We aim to evaluate the lead singer's theories regarding the factors which may affect crowd energy

DATA OVERVIEW AND CLEANING:

Dataset contains information about concert venues, ticket prices, volume levels, weather conditions, moon phases, band outfits and resulting crowd energy on a scale of 0-100

Data cleaning steps included:

- Converting date and time fields into proper formats from strings
- Removing unique concert identifier (Gig_ID) and possible leakage variables (Merch_Sales_Post_Show)
- Handling missing values in numerical features by using median values
- Removing extremely unlikely outliers in the Crowd_Energy column which may be a result of incorrect data logs

EXPLORATORY DATA ANALYSIS (EDA):

EDA revealed several important patterns:

- Venue significantly affected the crowd energy with particular venues consistently producing higher crowd energy
- Volume levels had weak – to – moderate relationship with crowd energy, though variability remained high
- Ticket prices showed weak positive association indicating that higher ticket prices do not guarantee higher energy
- Weather, moon phases and band outfits showed minimal impact but are retained due to potential interaction effects

FEATURE ENGINEERING:

Feature engineering included:

- Extracting time based features like show hour, day and month
- Creating a weekend indicator
- Retaining categorical variables for encoding
- Dropping raw date- time fields after feature extraction

MODEL SELECTION AND TRAINING:

Random Forest Regressor was selected due to its ability to:

- Capture non-linear relationships
- Handle categorical and numerical features effectively
- Remain robust to outliers and noisy data

Categorical variables were encoded using one- hot encoding within a preprocessing pipeline ensure no data leakage.

Hyperparameter tuning indicates that the number of estimators must be 500 and max depth must be 10 for minimum Mean Average Error.

MODEL EVALUATION:

Model was evaluated on a validating set using Mean Absolute Error and R^2 score.

MAE was approximately 12.75 indicating average prediction errors of around 12.75 energy units.

R^2 was approximately 0.34 meaning the model explains around 34% of variance in crowd energy

CONFIRMATION OF LEAD SINGER'S THEORIES:

- Venue impact- Strongly supported by data
- Volume impact- Partially supported
- Ticket price impact- Weakly supported
- Weather, Moon Phase, Band Outfit impact- Largely unsupported as standalone predictors