

ML TASK REPORT

The original data was messy and chaotic which was eventually fixed and cleaned to be tested against the singer's scribbles and to build a machine learning model to predict the future show success.

The data proved that every venue behaves differently, just as the singer suspected;

1] **The Holy Grounds/V Alpha** - The monk's theory is real. The venue has strict volume limits.

After correcting the negative volume readings and data errors; it was found that volume control is critical here.

2] **The Vampire's Den/V Beta** - The singer's note about 'goths' and timing was correct. We grouped shows by time (like morning, afternoon, evening, late night) and found that the late night shows behave completely different than afternoon ones.

3] **The Snob Pit/V Gamma** - For evaluating this, 'price_bucket' was especially created and over this we evaluated 'price_shock' feature to measure how expensive a ticket is compared to other venues.

The audience here is highly sensitive to pricing.

4] **The Mosh Pit/V Delta** - 'Pure chaos'; the chaos index (volume levels x crowd density(crowd size / total max capacity)) was generated which became a key predictor for this venue. more packed the venue, louder it is subsequently adding the higher energy levels.

Furthermore, the code was used to check whether the scribbles theories were true or just hallucinations.

- **Tuesdays are cursed** : This is partially true. The average crowd size on Tuesdays is approx 498 lower than the weekly average of 533. It's not a total disaster, but audience drop is noted.

- **Full moon = magic** : false. The singer is totally wrong. The full moon shows actually had lowest average crowd size (approx 491) compared to other moon phases.

- The drummer on other hand was right; the full moon theory is just confirmation bias
- Spandex vs Leather jacket : True; The singer felt that leather jackets were mid. The data agrees as the leather shows had lower audience (512 avg). On other hand, spandex and denim performed significantly better (541+ average)
- The rain sucks : True; rainy and stormy weather showed lower crowd averages compared to clear skies. The singer's mood might be affected, but the audience definitely takes a hit.

*Model choice - I chose **random forest regressor.***

The model selection started with simple linear regression assuming straight line relation. But the data was too complex; v_delta is chaotic, v_gamma being price sensitive and so on which proved themselves as curves and interactions.

A simple linear regression model was tested at first but it **failed** to capture the complex chaos at those mosh pits.

Furthermore, single decision trees were prone to overfitting ie memorising the training data rather than learning general rules.

The random forest regressor thus was chosen as it is incredibly stable & averages out hundreds of decision trees while being resistant to outliers; handling the complex, non linear patterns much better.

The best RMSE ie root mean squared error of 51.82 was achieved on validation set.

Hyperparameter tuning - here GridSearchCV was used to test different settings. The tuned model performed almost exactly the same as the default model with improvement *very small (0.0002)*.

This assures and meant how the feature engineering was strong enough in the model as it didn't need much tweaking to find perfect signal.