MACHINE LEARNING

Task Released Deadline: 27th Dec

Exclusively for 1st-year students.

The World Tour of '99: The Data Detective

The Scenario

You are the newly hired data scientist for Electric Omen, a legendary rock band planning their massive comeback tour. The band has played hundreds of shows across four legendary venues, and now they need your expertise to ensure every future show is a riotous success.

Your mission: predict the Crowd Energy (a score from 0–100) for future venues. The band's manager believes that understanding what drives crowd energy will help them optimize setlists, pricing, and scheduling for maximum impact.

You have access to the band's historical tour logs (tour_logs_train.csv). However, fair warning, the band's roadies weren't exactly meticulous record-keepers. The data is chaotic, messy, and inconsistent.

You also have access to the Lead Singer's personal notes (lead_singer_scribbles.txt). These cryptic scribbles contain the singer's theories and observations about what affects crowd energy. IMPORTANT: The singer admits these notes are unreliable, contradictory, and possibly wrong. Treat them as hypotheses to test, not facts to assume.

Dataset

The below link contains the training and the test datasets, ensure that the test set is in no way used for the training of your model, doing so will lead to disqualification. (link)

The Four Venues

Each venue has its own personality, quirks, and hidden rules. Part of your job is figuring out what makes each one tick—and whether the singer's theories hold up.

- The Holy Grounds (V_Alpha) – A converted monastery. The singer has theories about noise limits here.
- The Vampire's Den (V_Beta) – A gothic nightclub. The singer believes timing matters, but isn't sure how.
- The Snob Pit (V_Gamma) – An exclusive venue. The singer suspects pricing plays a role, but can't remember the details.
- The Mosh Pit (V_Delta) – Chaotic crowds. The singer has conflicting memories about what drives energy here.

Your Mission

Primary Objectives

1. Data Cleaning & Wrangling: The data is messy. Dates are inconsistent. Prices have mixed currencies. Some readings are clearly errors. Your first job is to make sense of the chaos.
2. Exploratory Data Analysis (EDA): Use visualizations to discover patterns. Test the singer's hypotheses against the actual data—some may be right, some may be wrong, some may be partially true.
3. Feature Engineering: Create new features that capture the patterns you've discovered. Think about interactions, thresholds, and transformations.
4. Model Training & Hyperparameter Tuning: Train a regression model to predict Crowd_Energy. You must demonstrate hyperparameter tuning with cross-validation.
5. Prediction Submission: Generate predictions for the test set in the required format.

Hyperparameter Tuning Requirement
This is a mandatory requirement, not optional. Your submission must include:

- Justification for your model choice
- Documentation of hyperparameters explored (ranges, values tested)
- Validation strategy used during tuning (e.g., k-fold cross-validation)
- Final hyperparameter values with reasoning for selection
- Comparison of tuned model vs. default parameters

Bonus Objective: Revenue Optimization
The band's manager, Rick, wants to maximize profit at V_Gamma (The Snob Pit). He's given you the following business context:

- Venue capacity: approximately 800 seats
- Fixed costs per show: ~$5,000 (venue rental, crew, equipment)
- Variable cost per attendee: ~$8 (security, cleaning, insurance)
- Rick has noticed that crowd energy affects whether people stay, buy drinks, and buy merch
- Attendance may be affected by both ticket price and crowd energy—but Rick doesn't know exactly how

Rick's request: "Find me the ticket price that makes the most money. I don't care how you figure it out—just show me the math and justify it. If you tell me 'it depends on the moon' I'm docking your pay."
Your task:

- Derive your own revenue/profit optimization formula — we are NOT giving you the formula
- State your assumptions explicitly
- Use your trained model to simulate different price points
- Create a visualization showing your optimization curve
- Identify and justify the optimal price point

Warnings

Read these carefully. Ignoring them will result in failure.

The Scribbles Are UNRELIABLE
The Lead Singer's Scribbles contain theories, guesses, and half-remembered observations. Some are correct. Some are wrong. Some are partially true. Do NOT treat them as ground truth. Use them to generate hypotheses, then test those hypotheses against the actual data.

Data Leakage
Some columns in the dataset are collected AFTER the show. Using them to predict Crowd_Energy is cheating and will result in models that fail catastrophically in production. Think carefully about what information would actually be available BEFORE a show.

Red Herrings
Not every column matters. The singer has theories about many features, some are noise. A good data scientist knows what to keep and what to discard based on evidence, not intuition.

Data Quality Issues

- Dates are logged in multiple inconsistent formats (including text descriptions)
- Prices include different currencies: $, £, €. Exchange rates: £1 = $1.27, €1 = $1.09
- Some sensor readings have systematic failures (missing or zero values)
- Outliers exist from data entry errors (impossible values)

Distribution Shift
The test set may not have the same distribution as the training set. Models that overfit to training patterns will fail. Build robust models that generalize.

Unseen Categories
The test set may contain categorical values that don't exist in training. Make sure your pipeline handles unknown categories gracefully without crashing

Deliverables

- predictions.csv – Your predictions in the required format (link)
- analysis_notebook.ipynb – Complete analysis including EDA, cleaning, feature engineering, hyperparameter tuning, and modeling
- findings_report.pdf – 1 to 2 summary highlighting the key findings for each venue, which of the singer's theories were correct/incorrect, model choice justification
- (Bonus) revenue_optimization.pdf – Your formula derivation, assumptions, optimization analysis, and recommended optimal price with visualization.

Evaluation Criteria

You will be evaluated on:

Criterion	Weightage
Prediction Accuracy (RMSE on test set)	20
Data Cleaning Quality	15
EDA & Pattern Discovery	20
Feature Engineering	25
Hyperparameter Tuning	10
Code Quality & Documentation	10
Bonus: Revenue Optimization (with formula derivation)	+15

Note

- Using data leakage features (features that would not be available before the show)
- Submitting predictions in incorrect format
- Hardcoding test set answers
- No evidence of hyperparameter tuning (using only default parameters)
- Plagiarism or collaboration (this is an individual assessment)

This challenge is designed to test real data science skills, not just model fitting. We want :

- Critical thinking: Can you evaluate the singer's claims against evidence?
- Data intuition: Can you recognize leakage, noise, and real signals?
- Technical rigor: Can you properly tune and validate a model?
- Communication: Can you explain your findings clearly?
- Business sense: (Bonus) Can you translate a model into business recommendations?

Good luck. Rock on.

Deadline – 28th Dec
Submission Link

Join the WhatsApp Group

Having any doubts ? Contact us

Arkadeb : 9476256774
Manthan : 9512529935
Nilay : 9340251202
Pushpendra : 7351816883
Arnav : 7982390939
Naveen : 8074523143
Sarvesh : 7262006199