PACE Strategy

PLAN

Business Question: Can we predict a user's subscription plan based on their profile, dancer type, learning goals, and genre interests?

Context: Understanding subscription behavior is critical to optimizing pricing, targeting, and retention strategies in ed-tech businesses. Your dataset, although lacking behavioral metrics, contains rich intent and user-type attributes. This makes it suitable for multiclass classification, user segmentation, and persona-based marketing strategy.

Goal: To develop a logistic regression-based classification model that predicts plan type (e.g., Basic, Standard, Full Access) based on user attributes and learning intent.

Target Variable: subscription_plan mapped to discrete categories:

- Basic (free)
- Full Access 7 (7 days)
- Full Access 30 (30 days)
- Full Access 180 (6 months)
- Full Access 365 (1 year)

Stakeholders:

- Founder & Director
- Hiring managers and recruiters (portfolio showcase)

ANALYZE

1. Data Exploration:

- Identify and treat missing values in columns like type_of_dancer, future_aspirations, location
- Analyze distributions of user types, plan types, countries
- Understand correlations between user intent (purpose, aspirations) and subscription plan
- Check class imbalance in plan.name and use resampling if necessary
- Visualize relationships using bar plots, heatmaps, violin plots

2. Assumption Checks & Ethics:

- Treat and encode multi-genre columns logically (e.g., binary flags or dominant category)
- Avoid bias against geographic or economic location
- Consider implications of over- or under-predicting high-tier plans (revenue projections, over-offers)
- Be transparent about interpretability using feature importance and coefficient weights

3. Criteria for Modeling Approach:

- Multiclass classification with both categorical and ordinal variables
- Logistic regression for interpretability; optionally compare with Decision Tree
- Evaluate model on accuracy, precision, recall, F1-score per class

CONSTRUCT

1. Data Preprocessing:

- Clean non-standard and missing values
- Encode:
  - type_of_dancer (ordinal)
  - genres_of_interest (multi-hot encoding or most frequent)
  - purpose_of_learning, future_aspirations (label/one-hot)
  - location, country (group sparse values as "Other")
- Derive:
  - signup_month, subscriptions, repeated_subscriber, etc.
  - Map plan.name to 3–4 simplified classes for modeling

2. Model Building:

- Split dataset (train/test: 70/30 or 80/20)
- Train logistic regression model (Scikit-learn)
- Evaluate performance using accuracy, precision, recall, F1-score, ROC-AUC, Confusion matrix and Feature importance interpretation.

3. Ethical Implications:

- Avoid promoting discriminatory patterns (e.g., only users from metro cities shown premium plans)
- Emphasize model interpretability over black-box performance

EXECUTE

1. Evaluate Model Performance:

- Select best performing model based on F1-score and class-level performance
- Document feature impacts (e.g., advanced dancers → more likely to choose premium plan)

2. Dashboard with Power BI:

- Plan distribution by: Country, purpose of learning, dancer type
- KPIs: Conversion likelihood by user type, Segment-wise churn risk (optional), Geographic visualization of user base
- Include recommendations panel

3. Final Deliverables:

- Exported insights and visualizations
- GitHub repo with Jupyter notebooks and documentation
- Executive summary PDF with findings and recommendations

4. Recommendation:

- Target specific user personas with custom pricing plans
- For future platforms, collect behavioral signals to improve prediction accuracy