Project Title:

Predicting User Subscription Behavior in an EdTech Startup Using Logistic Regression and Business Intelligence Tools

---

1. Background & Business Objective

For over five years, your dance education startup offered online and offline learning options to aspiring dancers. The platform captured valuable user data including their intent for learning, type of dancer, genre preferences, location, and plan subscriptions. However, with the platform nearing its closure, the aim now is to leverage this historical data to extract actionable business insights and demonstrate full-stack data analytics capabilities for future employment.

This project aims to build an end-to-end data pipeline that begins with raw user data, processes and analyzes it, and ultimately produces a predictive classification model to forecast user subscription behavior. Additionally, the project will feature a Power BI dashboard for strategic recommendations.

2. Business Problem

The goal is to solve the following problems using historical user-level data:

- Can we predict what type of subscription plan a user is most likely to choose based on their learning intent, dancer type, location, and other factors?
- What user personas emerge from the data, and how do these personas align with the subscription behavior?

3. Goal

To build a logistic/multiclass classification model that predicts a user's subscription type based on user intent, profile attributes, and interests.

Additional goals:

- Clean and prepare real-world messy user data using SQL and Python.
- Perform exploratory analysis to understand user trends.
- Deploy insights using an interactive Power BI dashboard for storytelling and visualization.
- Package the project as a portfolio case study highlighting my role as a Data Consultant for my former startup.

4. Scope of Work

In Scope:

- Data ingestion, cleaning, and feature engineering using Snowflake SQL and Python.
- Exploratory Data Analysis (EDA), visualizations, and summary statistics.
- Modeling: 1. Logistic regression to predict plan type.

  2. Optional: churn prediction (active vs. inactive users).

- Dashboard development in Power BI.
- Documentation, reporting, and GitHub publication.

Out of Scope:

- Behavioral log analysis (no last login, session time available).
- Real-time prediction/deployment of model.

5. Tools & Technologies

- **SQL (Snowflake)**: Data modeling, cleaning, transformations
- **Python (Pandas, NumPy, Seaborn, Scikit-learn)**: EDA, feature engineering, logistic regression modeling
- **Power BI**: Interactive dashboards and storytelling
- **Jupyter Notebook**: Data analysis and modeling workflow
- **GitHub**: Code and documentation repository

6. Deliverables

- Cleaned dataset and Snowflake schema.
- Python-based EDA and logistic regression model.
- Power BI dashboard showing insights and KPIs
- GitHub repository with README, Project Proposal, and PACE document.
- Executive Summary for stakeholders.

Data Analyst & Consultant: Shijin Ramesh

Date: 31st July, 2025