Project Proposal: Predicting User Engagement in AI-Powered Educational Platforms

Course: Statistical Methods II -STATS 547

1. Project Title: Predicting User Engagement in AI-Powered Educational Platforms.

2. Objective:

This project aims to analyze a dataset from an AI-based educational platform to understand the factors influencing user engagement. Through statistical modelling, we will identify and interpret the main predictors of engagement, such as the number of sessions, time spent on the platform, content type, and user demographics. These insights will help enhance the platform's content delivery and improve user retention.

3. Data Sources:

To analyze user engagement on an AI-powered educational platform, we used a dataset containing course information, such as course title, organization, rating, difficulty level, and enrollment numbers. However, since the dataset didn't include direct engagement metrics, we generated synthetic data based on course characteristics.

- **Converting Enrollment Numbers:** We first converted enrollment values into a numeric format, enabling us to use this variable in calculations.
- **Simulating Session Count:** Using the enrollment numbers as a basis, we generated an estimated session count for each course. Courses with more enrollments were assigned a higher session count, simulating the likely frequency of user interactions.
- **Session Duration Based on Difficulty:** Session duration was generated based on course difficulty. For example, beginner-level courses typically had shorter session times, while advanced courses had longer average session durations, reflecting the time users might spend on challenging content.
- **Completion Rate Estimation:** We simulated course completion rates based on course rating and difficulty. Courses with higher ratings and lower difficulty were more likely to have higher completion rates, indicating higher user satisfaction and engagement.

This synthetic data allows us to model user engagement effectively and explore patterns related to course features such as difficulty and rating. It provides a basis for applying statistical analyses to better understand the factors influencing user engagement on educational platforms.

4. Statistical Methods

Our analysis will focus on the following statistical methods to evaluate and predict user engagement:

- **Multiple Linear Regression:** We will quantify the relationship between user engagement (e.g., completion rate, session time) and various predictors (e.g., user demographics, frequency of use, content difficulty). This will help determine the key factors that influence engagement levels.
- **Logistic Regression:** This method will model the probability of a user completing a course based on predictors like course content, user background, and session frequency.
- **Factorial Designs:** We will analyse interactions between factors such as content type and user demographic (e.g., age or education level) to observe how these combined variables affect engagement.
- **Randomized Block Design:** This technique will help us control for categorical variables like education level, ensuring that differences in engagement are not biased by such variables.
- **Non-Linear Regression:** If applicable, we will explore non-linear trends in engagement over time, such as examining how engagement decays or grows with repeated interactions.

5. Steps and Methodology

- **Data Cleaning and Exploration:** We will start by examining and cleaning the dataset in R, identifying any missing values or outliers and visualizing key engagement metrics.
- **Exploratory Data Analysis (EDA):** Using visualization techniques in R (e.g., ggplot2), we will analyse the distribution of engagement metrics across demographic groups and course types, identifying patterns or trends.
- **Model Building:** We will implement each statistical model, testing hypotheses about user engagement, and evaluating the relationship between predictors and engagement levels.
- **Analysis and Interpretation:** The results from each model will be analysed to determine which factors significantly impact engagement. We will interpret these findings and discuss their implications for improving platform design and user retention.
- **Model Validation:** Cross-validation techniques, such as k-fold cross-validation, will be applied to assess model performance, using metrics like R-squared for linear models and AUC for logistic regression.

6. Expected Outcomes

By the end of this project, we aim to:

- Identify the primary factors influencing user engagement on educational platforms.
- Generate predictive models that can inform content delivery strategies and improve user retention.
- Provide practical recommendations for educational platform developers based on our findings.

7. Tools and software

R Language - We will use R for data processing, modeling, and visualization, utilizing libraries such as dplyr, ggplot2, car, lmtest, nls, and caret.