TITLE PAGE

Project Title: Analysis of Online Course Completion Rates

Intern Name: [INARA SAMINA.M]

Internship Organization: [UNIFIED MENTOR]

College/University: [FRANCIS XAVIER ENGINEERING COLLEGE]

ABSTRACT

This project examines the variables that affect the completion rates of online courses by utilizing simulated student engagement data. Online learning platforms encounter substantial obstacles due to high dropout rates, which diminish the effectiveness of digital education. Through exploratory data analysis (EDA), feature engineering, and predictive modeling, this research uncovers trends and patterns that determine a learner's likelihood of completing a course. The objective is to furnish actionable insights to course designers and educational platforms to enhance learner engagement and retention.

INTRODUCTION

The emergence of online learning platforms such as Coursera, Udemy, and edX has broadened access to education. Nonetheless, these platforms are confronted with a considerable challenge: a significant proportion of learners do not finish the courses in which they enroll. This project seeks to analyze student behavior alongside course attributes to discern the factors that drive course completion. Recognizing these elements can assist in refining platform design and improving learner outcomes.

PROBLEM STATEMENT

Despite robust enrollment figures in online courses, the completion rates remain markedly low. Comprehending the influences that lead learners to complete or abandon a course is essential for enhancing course engagement and educational results.

OBJECTIVE

1. To analyze student engagement metrics and demographic characteristics.

2. To identify critical factors influencing course completion.

3. To develop a model that predicts the likelihood of course completion.

4. To propose enhancements for course design and learner support.

DATASET OVERVIEW

The dataset utilized in this study is synthetically generated and simulates 200 learners enrolled in online courses. It encompasses features such as:

Age, Gender, Country

Course Length, Course Difficulty

Video Completion Percentage, Assignments Completed

Login Frequency, Time Spent

Completion Status (Target Variable)

Tools and Technologies Used

Python * pandas * numpy * matplotlib, seaborn * scikit-learn

Jupyter Notebook / Google Colab

Excel (for data review)

Data Cleaning and Preprocessing

Eliminated missing or invalid values (not applicable to synthetic data).

Encoded categorical variables (e.g., Gender, Country, Course Difficulty).

Scaled numerical features (e.g., Time Spent, Video Completion Percentage).

Exploratory Data Analysis (EDA)

Analyzed the distribution of course completion across demographics.

Identified correlations between engagement metrics and completion rates.

Visualized trends using bar charts, heatmaps, and histograms.

Feature Engineering

Created an Engagement Score that combines logins, time spent, and video completion.

Generated interaction features such as login rate per week.

MODELING

```
import pandas as pd
import numpy as np
from random import randint, uniform, choice

np.random.seed(42)
n = 200

data = {
    "Student_ID": range(1, n + 1),
    "Age": np.random.randint(18, 50, n),
    "Gender": np.random.choice(["Male", "Female", "Other"], n),
    "Country": np.random.choice(["India", "USA", "UK", "Canada",
"Australia"], n),
    "Course_Length_Weeks": np.random.randint(2, 16, n),
    "Course_Difficulty": np.random.choice(["Beginner",
"Intermediate", "Advanced"], n),
```

```python
    "Video_Completion_Percentage": np.round(np.random.uniform(0, 100,
n), 2),
    "Assignments_Completed": np.random.randint(0, 10, n),
    "Total_Logins": np.random.randint(1, 50, n),
    "Time_Spent_Hours": np.round(np.random.uniform(1, 100, n), 2),
}

df = pd.DataFrame(data)
df["Engagement_Score"] = (
    df["Video_Completion_Percentage"] * 0.4 +
    df["Assignments_Completed"] * 0.2 +
    df["Total_Logins"] * 0.2 +
    df["Time_Spent_Hours"] * 0.2
)

df["Completed_Course"] = df["Engagement_Score"].apply(lambda x: 1 if
x > 60 else 0)

print(df.head())

completion_rate = df["Completed_Course"].mean()
print(f"\nSimulated Course Completion Rate: {completion_rate *
100:.2f}%")
```

KEY FINDINGS

Learners who completed a greater number of assignments and spent more time on the platform exhibited a higher likelihood of completing the course.
Course difficulty had a moderate impact; beginner-level courses demonstrated elevated completion rates.
Increased login frequency and video completion percentage were strongly correlated with success.

RECOMMANDATIONS

Send reminders to users exhibiting decreased engagement.
Break extended courses into smaller modules to mitigate dropout rates.
Integrate gamification strategies to encourage consistent participation.

LIMITATIONS

The synthetic dataset may not accurately represent real-world learner behavior.
Absence of time-series data limits the analysis of engagement over periods.
There is a lack of psychological or motivational data.

CONCLUSION

This project illustrated how data analytics can reveal patterns in learner behavior that affect online course completion. By identifying critical factors such as engagement, time spent, and course structure, platforms can make informed decisions to support learners and enhance success rates.