ANALYZING FACTORS INFLUENCING ONLINE COURSE COMPLETION

STAT 31631- Statistical Modeling

Department of Statistics and Computer Science

University of Kelaniya.

Academic Year 2022/2023

By

Group 04

CONTENT

1. INTRODUCTION

1.1 Background

Online education is growing rapidly, making it essential to understand factors that impact user engagement and course completion to optimize learning experiences.

1.2 Purpose and Objectives

The study aims to conduct a comprehensive regression analysis on user engagement metrics (e.g., course duration, interaction frequency, demographics) to identify key factors influencing course completion rates.

The goal is to provide actionable insights that can improve instructional design, learner retention, and overall course effectiveness.

1.3 Significance (Novelty and Advantages)

This study provides valuable guidance for educators and platform developers to refine strategies, improve course delivery, and create more engaging learning environments.

The analysis contributes to enhancing user engagement and educational outcomes, benefiting both learners and course providers.

2. PROBLEM STATEMENT

2.1 Specific Problem

Despite the rise of online education, low course completion rates remain a significant challenge due to varying factors affecting user engagement.

2.2 Knowledge Gap

There is limited understanding of how engagement metrics like interaction frequency, course duration, and demographics correlate with successful course completion.

2.3 Objective

The project aims to address this issue by conducting a regression analysis to identify and quantify the key factors influencing course completion, offering evidence-based recommendations to improve course design and student success rates.

3. METHODOLOGY

By following the complete methodology, we ensured that our data was precisely produced and thoroughly evaluated, establishing a foundation for subsequent inferential statistical analysis and model development to better understand the factors influencing online course completion This method will help us create a more accurate predictive model for future use.

3.1 Data Source

The dataset was obtained from Kaggle, focusing on user interaction with online courses.

3.2 Variables Selected

- **UserID**: Unique identifier for each user.
- **Course Category**: Type of course (e.g., Programming, Business, Arts).
- **Time Spent on Course**: Total time spent on the course (in hours).
- **Number of Videos Watched**: Insight into user engagement.
- **Number of Quizzes Taken**: Indicator of active participation.
- **Quiz Scores**: Average quiz performance (%).
- **Completion Rate**: Percentage of course content completed (dependent variable).
- **Device Type**: Device used (Desktop or Mobile).
- **Course Completion**: Whether the course was completed (1) or not (0).

3.3 Data Preprocessing

Check for Missing Values: No missing values were found in the dataset.

Removal of Columns: Variables with less than 60% data were omitted, but none were removed since no missing values existed.

3.4 Data Analysis

Descriptive Analysis: Summary statistics and visualizations (e.g., histograms, density plots) were used to explore data distribution.

Univariate Analysis: Distribution, central tendency, and dispersion of each variable will be examined using histograms, density plots, and box plots for outlier detection.

Multiple Linear Regression: To estimate the combined impact of variables on course completion. This includes model fitting, residual analysis, and variable selection.

Predictive Model Enhancement: Future validation and refinement to improve model accuracy.

4. RESULTS

4.1 Preliminary Analysis

A linear regression model was fitted using variables such as time spent on course, number of videos watched, number of quizzes taken, quiz scores, and device type.

Initial results showed that none of the predictors were statistically significant, with a low adjusted R-squared value (0.0002).

Best subset selection was applied to improve model accuracy, and forward subset selection identified a model with 5 variables as more effective.

- The final model indicated that:
  - Time spent on the course had a positive but small effect on completion rates.
  - The number of videos watched, and quizzes taken also contributed positively to course completion.
  - Quiz scores had a negative impact on course completion rates, and students in certain groups (e.g., science students) had lower completion rates compared to others.

4.2 Findings

1. Time spent, videos watched, and quizzes taken positively impacted completion rates.
2. Quiz scores negatively influenced course completion.
3. The final model explained only a small variance, suggesting room for improvement.

5. DISCUSSION

5.1 General Results

The results indicated limited predictive power in the model, as none of the variables showed strong significance.

The course completion rate was only slightly influenced by the factors considered, such as time spent and engagement metrics.

5.2 Relation to Objectives

The results align with the objective of understanding factors influencing course completion but suggest the need for additional variables or refined methods to enhance predictive power.

The weak correlation between the selected predictors and course completion rates suggests the model may benefit from incorporating other factors, such as user motivation or content difficulty.

6. CONCLUSSION

6.1 Significance of Time Spent on Course

- Time spent on the course has a positive but weak effect on completion rates, with a slight increase of 0.021 units per additional unit of time.
- While the p-value (0.0515) is slightly above the standard significance level of 0.05, it suggests a meaningful relationship worth further exploration.

6.2 Non-Significant Variables

- Variables such as **Number of Videos Watched**, **Number of Quizzes Taken**, **Quiz Scores**, and **M3** did not show statistically significant relationships with course completion.
- This suggests that increasing video consumption or quiz-taking does not necessarily lead to higher completion rates.

6.3 Model Fit and Limitations

- The model has low predictive power, with an R-squared value of 0.001 and an adjusted R-squared of 0.0005, indicating that only about 0.1% of the variation in course completion is explained by the included variables.
- Residual Standard Error (28.94) indicates wide variability, emphasizing the model's limited accuracy.

6.4 Key Insight

- The model's most valuable finding is the potential impact of time spent on course completion. This insight suggests focusing on strategies that increase engagement time.

6.5 Recommendations for Further Research

- The study highlights the need for further exploration of **time spent** as a factor and suggests incorporating other potential influences such as student motivation or course material quality.

- Future research should involve more comprehensive data and alternative modeling techniques to better understand the factors affecting course completion

7. INDIVIDUAL CONTRIBUTION

Student Number	Contribution
PS/2020/150	Multiple linear regression analysis (Coding and its interpretation) and report development, Final presentation development
PS/2020/174	Final conclusions and final presentation development
PS/2020/192	Univariate descriptive analysis (Coding and its interpretation) and report development
PS/2020/199	Multiple linear regression analysis (Coding and its interpretation) and report development, Development of project proposal
PS/2020/207	Development of project proposal and final conclusions
PS/2020/215	Univariate descriptive analysis (Coding and its interpretation) and report development
PS/2020/218	Development of project proposal and final conclusions
PS/2020/219	Development of project proposal and final conclusions, Multiple linear regression analysis (Coding and its interpretation) and report development