Predicting Student Success in Online Courses

Abhi

October 13, 2024

Abstract

This report presents a comprehensive analysis for predicting student success in online courses using synthetic data generation with CTGAN, hypothesis testing for generated data, clustering analysis, and predictive modeling techniques. The aim is to identify at-risk students and provide early interventions to enhance their learning outcomes. We generate synthetic data using CTGAN and compare its correlation with the original data to ensure that key relationships between features are preserved. This report details the dataset, methodology, results, and insights gained from the study.

1 Introduction

Online education platforms provide valuable learning opportunities for students worldwide. However, predicting student success in such platforms is a significant challenge, as many students drop out before course completion. In this study, we leverage historical student data to predict course completion rates and identify at-risk students.

We employ machine learning models and clustering analysis to explore student behavior. The study also involves comparing the relationships in original and synthetic datasets to ensure that key features and dependencies are maintained.

2 Methodology

The methodology is divided into several phases: data preprocessing, synthetic data generation using CTGAN, hypothesis testing for generated data, feature engineering, clustering, and predictive modeling.

2.1 Data Preprocessing

The dataset was preprocessed by removing columns that could introduce bias, such as **region** and **major**. These columns can introduce bias and limit the model to only the provided options, as we are not given all possible majors and regions. This incomplete representation may result in a model that doesn't generalize well to the full student population. The final dataset included the following features:

- Age
- Gender

- Year
- Logins per Week
- Videos Watched
- Time Spent on Platform
- Average Quiz Score
- Average Score Across Courses
- Course Completion Ratio

2.2 Synthetic Data Generation Using CTGAN

To generate the dataset and simulate new student profiles, we used the Conditional Tabular Generative Adversarial Network (CTGAN). This technique helps in generating synthetic data that mimics the distribution of the original data, preserving the relationships between variables.

The CTGAN model was trained on the original dataset and used to generate synthetic student records. These records are intended to represent new students while maintaining the statistical characteristics of the original dataset.

2.3 Hypothesis Testing for Synthetic Data Validation

To ensure that the synthetic data closely resembles the original data, hypothesis testing was conducted as part of the data generation process. We used the following tests:

- **Two-sample t-test:** This was applied to check if the mean values of key features (e.g., logins per week, average quiz score) were statistically similar between the original and synthetic datasets.
- **Kolmogorov-Smirnov Test:** This test compared the distributions of continuous features in the original and synthetic datasets to ensure that they were similar.

The null hypothesis was that there is no significant difference between the means and distributions of the original and synthetic data. If the p-value was greater than 0.05, we failed to reject the null hypothesis, implying that the synthetic data closely matched the original.

2.4 Correlation Analysis

To further assess the quality of the synthetic data, we calculated the Pearson correlation coefficients for both the original and synthetic datasets. The correlation matrices were visualized using heatmaps to clearly illustrate the relationships between features, such as the correlation between average quiz score and course completion ratio.

Figures 1 and 2 display the heatmaps for the original and synthetic datasets, respectively. These visualizations allow for an intuitive comparison of key relationships preserved in the synthetic data.

The correlation results indicate that the synthetic data effectively preserves the relationships between key features, as evidenced by the similarity in heatmap patterns.

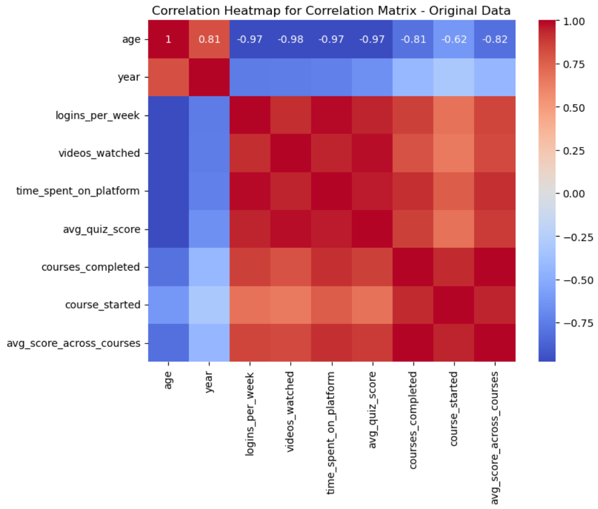

Figure 1: Heatmap of Correlation Matrix for Original Data

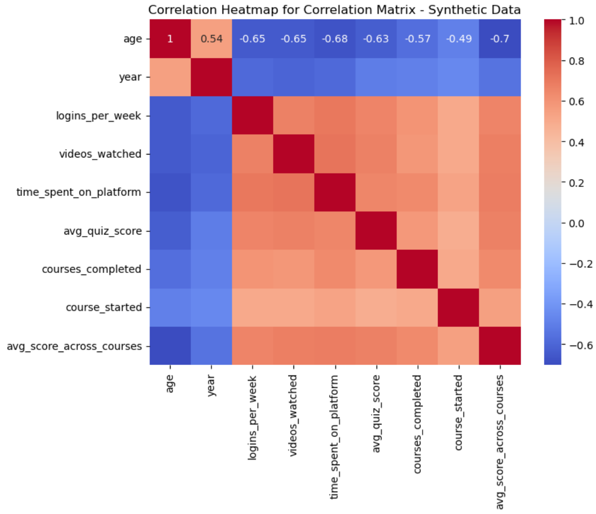

Figure 2: Heatmap of Correlation Matrix for Synthetic Data

3 Clustering Analysis

K-means clustering was applied exclusively to the synthetic dataset to identify distinct groups of students based on their engagement and performance. The optimal number of clusters was determined using the elbow method, allowing us to select the most informative clustering structure.

The clustering analysis provided valuable insights into the characteristics of students and highlighted key features that contribute to student success and risk. Features such as ‘Logins per Week‘, ‘Time Spent on Platform‘, and ‘Average Quiz Score‘ were found to be significant indicators of student engagement. By examining the clusters, we identified students with similar behaviors and performance levels.

3.1 Clustering Results

The clustering results revealed several distinct student groups, with a notable focus on a high-risk cluster characterized by low engagement and low course completion. This cluster predominantly consisted of students who logged in infrequently, spent minimal time on the platform, and achieved low average quiz scores.

The insights from the clustering analysis inform targeted interventions to minimize dropout risk among at-risk students. For instance, students identified in the high-risk cluster could benefit from:

- **Personalized Engagement Strategies:** Tailored communication and reminders to encourage regular logins and active participation in course activities.
- **Enhanced Learning Resources:** Additional support materials, such as tutorials or mentorship programs, to assist students in improving their quiz scores and overall understanding of course content.
- **Monitoring and Feedback:** Continuous tracking of student engagement metrics to provide timely feedback and assistance, helping to address potential issues before they lead to dropout.

By utilizing clustering analysis on the synthetic data, we can better understand the student population and implement data-driven strategies to enhance retention and success in online courses.

4 Predictive Modeling

Several machine learning models were applied to predict whether a student would complete a course using the synthetic dataset. These models included:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier
- XGBoost Classifier

The models were trained and evaluated on synthetic datasets. The evaluation metrics included accuracy, precision, recall, and F1-score.

5 Results and Discussion

5.1 Predictive Model Results

The performance of the predictive models across the synthetic dataset is summarized in Table 1. The Gradient Boosting Classifier performed best, achieving an F1-score of 0.8895.

5.2 Insights from Predictive Modeling

The feature importance analysis from the predictive models revealed several key factors influencing student success in online courses. Among the most significant features were:

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.8155	0.8491	0.9327	0.8889
Random Forest Classifier	0.7967	0.8438	0.9118	0.8765
Gradient Boosting Classifier	**0.8160**	0.8474	0.9360	**0.8895**
XGBoost Classifier	0.8121	0.8494	0.9269	0.8864

Table 1: Model Performance on Synthetic Data

- **Average Quiz Score:** This feature demonstrated a strong positive correlation with course completion, indicating that higher quiz performance is associated with a greater likelihood of completing the course.
- **Logins per Week:** Increased frequency of logins was also a crucial factor, suggesting that regular engagement with course materials is vital for student retention.
- **Time Spent on Platform:** The amount of time spent engaging with course content positively impacted the likelihood of completion, highlighting the importance of active participation.
- **Course Completion Ratio:** This metric was crucial in assessing student progress and was closely linked to overall success in completing courses.

These insights emphasize the need for targeted interventions that can improve student engagement, such as personalized reminders, tutoring sessions, and additional resources aimed at enhancing quiz performance.

6 Recommendations to Minimize Risk of Dropout

6.1 Focus on Improving Platform Engagement

- **Time Spent on Platform:** Since time spent on the platform is a critical factor for course completion, consider implementing:
  - Learning Streaks: Offer rewards for students who log consistent hours on the platform.
  - Engagement Reminders: Automated reminders nudging students to revisit the platform.
  - Live Study Sessions: Organize live sessions to further enhance student interaction.

6.2 Enhance Overall Academic Performance

- **Avg Score Across Courses:** Overall performance across all courses plays a major role. It’s important to:
  - Offer Personalized Learning Plans: Students can benefit from tailored learning plans based on their performance.

  - Peer Study Groups: Organize virtual study groups to foster collaborative learning.
  - Academic Support: Provide academic support services for students who are struggling across multiple courses.

6.3 Improve Quiz Performance

- **Avg Quiz Score:** Quiz performance moderately impacts course completion. To address this:
  - Feedback and Reinforcement: Provide immediate and constructive feedback after quizzes to help students understand their mistakes.
  - Adaptive Quiz Design: Offer adaptive quizzes that adjust difficulty levels based on a student's performance.
  - Extra Practice Resources: Provide additional quizzes, flashcards, and mock tests to boost their performance.

6.4 Boost Logins and Engagement

- **Logins per Week:** Regular logins are a sign of active engagement. Some strategies include:
  - Gamification: Implementing badges, leaderboards, and achievements to encourage frequent logins.
  - Personalized Notifications: Send personalized messages or alerts to remind students to log in based on their activity level.
  - Weekly Challenges: Introduce weekly learning challenges that require consistent logins to complete.

6.5 Encourage Active Video Learning

- **Videos Watched:** While watching more videos correlates with higher completion rates, it is important to ensure students actively engage with the material. Recommendations:
  - Interactive Videos: Include quizzes and discussion prompts within videos to engage students more deeply.
  - Content Recommendations: Suggest videos based on their learning progress and areas of weakness.
  - Completion Tracking: Track video completion and suggest related videos or learning materials based on their watching habits.

6.6 Address Additional Factors

- **Monitor Low-Risk Indicators (Gender, Year, Age):** Although these features show minimal importance, it is still beneficial to:

- Track Progress: Even if gender, year, and age have minimal influence, monitor students' overall progress across various demographics to ensure no group is overlooked.

7 Conclusion

This study demonstrates that synthetic data generated by CTGAN can be effectively used to mimic real data for predictive modeling in online education platforms. The correlation analysis and hypothesis tests confirmed that the synthetic data closely mimicked the original data, preserving key relationships between variables.

The clustering analysis provided actionable insights into the characteristics of at-risk students, revealing groups with low engagement and performance. Strategies such as personalized engagement and enhanced learning resources can significantly improve retention rates.

Additionally, the predictive modeling results highlighted the importance of specific features, such as average quiz scores and login frequency, in determining student success. The Gradient Boosting Classifier emerged as a particularly effective model, achieving highest performance.

By leveraging both clustering insights and predictive modeling, educational platforms can implement data-driven strategies to enhance learning outcomes and support student success.