THE CURIOSITY CUP 2025
A Global SAS® Student Competition

Title: Gamification in Education: Enhancing Engagement and Learning Outcomes

The Gunners Team: Bhekamuzi Dhlamini, Nash Balraj and Phuti Mphaki

Abstract

This study explores the role of gamification in modern education, emphasizing its potential to enhance student engagement and learning outcomes. By analyzing a dataset of Coursera courses, the study identifies key trends and insights into the effectiveness of gamified elements in online education, using Coursera course ratings as a proxy for effectiveness. The methodology involves data-driven analysis to evaluate the impact of gamification on student enrollment and course ratings. Data analysis was conducted using SAS® OnDemand for Academics™, focusing on variables such as course ratings, student enrollments, and retention rates. Furthermore, the results suggest that gamified learning environments correlate with higher engagement and improved course ratings, particularly in beginner-level and short-duration courses. Key findings highlight the increasing adoption of gamification in education and its positive influence on student participation and satisfaction.

Introduction

Concept of Gamification in Education: Gamification refers to the application of game-design elements, such as point systems, leaderboards, and rewards, in non-gaming contexts like education. It aims to increase student motivation and engagement by incorporating interactive and competitive elements into learning activities. "Coursera's mission is rooted in serving the world through learning so everyone – regardless of location, socioeconomic status, or personal circumstance – has the power to unleash their full potential"(Coursera 2023).

Importance of Gamification in Modern Teaching: In contemporary education, gamification has become a crucial tool for educators striving to make learning more engaging and effective. It helps bridge the gap between traditional teaching methods and the digital preferences of modern learners, fostering an interactive and rewarding learning experience.

Scope and Relevance to Modern Teaching Methodologies: The rise of e-learning platforms has facilitated the integration of gamification into educational content. This study examines the relevance of gamification in online education, particularly on Coursera, and its impact on student engagement and course success rates. "Visualizing and incentivizing the mastering of specific abilities or competencies. Rewarding users for achieving different levels of proficiency can motivate them to continue engaging with an app or platform, and can also help users track their progress"(Adjust, 2024).

Key Research Question: How does gamification influence student engagement and learning outcomes in online courses?

Supporting Questions:

1. What are the key factors that correlate with higher course ratings?

2. Which types of certificates (e.g., Course vs. Specialization) tend to receive better ratings?
3. Which machine learning model (e.g., regression tree) best predicts course ratings based on course features?

Methodology

Data Preparation

- The dataset, sourced from Kaggle, contains 1,000 observations with 12 variables. It was imported into SAS® OnDemand for Academics™ for analysis.
- Data cleaning steps included handling missing values (applying the 5% rule), converting text-based numeric values, and managing duplicate records.
- Feature engineering introduced a new variable, retention rate, calculated as:
  - $Retention\ Rate = \frac{course\ reviews}{course\ student\ enrolled} \times 100$
- After cleanup, the dataset now contains 958 observations, 13 variables, No missing values, No duplicates and a newly engineered retention rate variable for further analysis.
- Refer to Output 1 to see an overview of the dataset from SAS®

Exploratory Data Analysis (EDA)

- Correlation analysis was performed to identify relationships between course ratings, enrollments, and engagement metrics. (refer to Output 2.)
- Visualization techniques such as histograms and box plots were used to assess data distribution.

Key Takeaways from the Correlation Matrix:

1. **Weak Positive Correlation (0.1733) Between Course Rating & Students Enrolled**
   - Higher enrollments slightly correspond to better ratings, but the relationship is not strong.
2. **Moderate to Strong Positive Correlation (0.6078) Between Students Enrolled & Course Reviews**
   - Courses with more students tend to receive more reviews, showing that enrollment size influences engagement.
3. **Minimal Correlation (0.0106) Between Course Rating & Retention Rate**
   - Retention rate is not a reliable proxy for engagement, as it does not significantly impact course ratings.
4. **Moderate Positive Correlation (0.3432) Between Course Reviews & Retention Rate**
   - More reviews generally indicate higher retention, but bias should be considered, as only engaged learners tend to leave reviews.

Course Ratings as Primary Target Variable:

Course ratings provide a more balanced and consistent measure of learner satisfaction and engagement compared to retention rates, which exhibit a highly skewed distribution with extreme outliers. Ratings offer a more reliable reflection of overall course quality, whereas retention rates may be influenced by voluntary reviews, introducing potential bias. Courses with low enrollments but a high number of reviews can further distort retention metrics,

making them less representative of all learners. While retention rate remains a valuable indicator, it should be interpreted with caution, recognizing its limitations as a review-based metric. Flagging it with appropriate caveats ensures a more accurate assessment of course performance. Refer to Figure 2. Retention Rate and Figure 3. Course Rating to see the distribution.

Model Selection

- Given the non-normal distribution of data, non-parametric model (Regression Trees) were used to predict course ratings.
    - Refer to Figure 4. Check for Normality
    - Since, Kolmogorov-Smirnov **p(<0.0001) < 0.05**, reject H_0 → Data is NOT normal → Consider non-parametric models.
    - The relationship appears **curved** and **non-linear**, consider non-parametric models.
- Feature selection included course duration, difficulty level, organization category (refer to Figure 1.), and enrollment numbers.
- Applied 70-30 split for training and test dataset

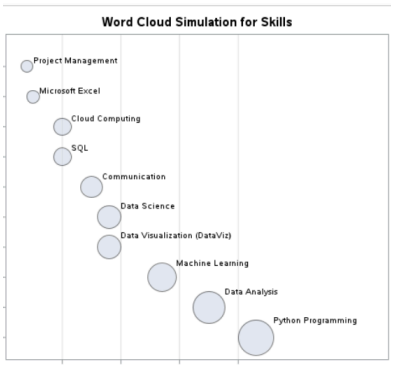
```
DATA proj.coursera_scaled_model;
    LENGTH organization_category $20; /* Increase character length to 20 */
   SET proj.coursera_scaled_model;
   IF INDEX(LOWCASE(course_organization), "university") > 0 or
      INDEX(LOWCASE(course_organization), "school") > 0 THEN
      organization_category = "University";
   ELSE
      organization_category = "Tech Company";
   DROP course_organization;
RUN;
```

Figure 1. Creating a New Variable organization_category

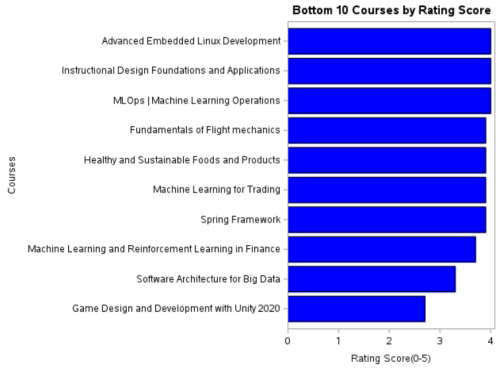
Results and Discussion

What are the key factors that correlate with higher course ratings?

The analysis of course ratings reveals distinct patterns in course organization, difficulty level, and skill focus. The top-rated courses cover a diverse range of subjects and are evenly split between educational institutions and technology companies, highlighting a balanced distribution. Most of these courses are at the beginner level, making them more accessible to a wider audience. Additionally, Python emerges as the most dominant skill among top-rated courses, indicating its relevance across various fields. In contrast, the lowest-rated courses are predominantly offered by universities and tend to focus on technical subjects, often requiring advanced knowledge. Many of these courses are classified as intermediate or advanced, reflecting the difficulty of acquiring skills such as big data and machine learning. This suggests that courses covering complex technical skills may face lower ratings, potentially due to challenges in content delivery or learner engagement. Therefore, we are applying predictive modelling to identify variables' importance and see if difficulty type, organization and skills correlate with course rating.



Display 1. Top Skills



Display 2. Bottom 10 Courses

Which types of certificates (e.g., Course vs. Specialization) tend to receive better ratings?

The analysis of course certificate types shows that Courses and Specializations dominate the offerings, with Courses being the most popular choice among learners. During analyses, we created a box plot of course ratings which indicated that Guided Projects and Courses tend to receive the highest ratings, followed closely by Professional Certificates, while Specializations exhibit more variability and lower-rated outliers. This suggests that learners value shorter, skill-focused learning formats over longer, more comprehensive programs.

Which machine learning model (e.g., regression, decision tree, random forest) best predicts course ratings based on course features?

The analysis reveals that **log_course_reviews** is the most influential predictor of course ratings across all models. The unpruned regression tree, with 9 leaves, captures all possible splits and fits the training data well (*MAE=0.1063*), but the risk of overfitting remains due to excessive complexity. The test set results (*MAE=0.1002*) suggest minimal overfitting, reinforcing the tree's predictive consistency. However, pruning the tree to 6 leaves improved generalization by retaining meaningful splits while reducing complexity. The pruned training model (ASE = 0.0338, RSS = 9.6986) demonstrated a balanced trade-off between accuracy and interpretability. The pruned test model further confirmed its robustness, achieving **lower ASE (0.0242)** and maintaining strong generalization. Importantly, **university-offered courses and course difficulty levels (beginner vs. advanced) significantly impact ratings**, suggesting students prefer easier courses. The pruned tree model MAE and RMSE are close to the unpruned model, showing good generalization. Given the results, the **6-leaf pruned tree is the most appropriate model**, as it balances predictive accuracy, interpretability, and generalization while avoiding excessive complexity and overfitting. For further insights, you can refer to Figure 5. regression Trees

Conclusion

The analysis highlights key factors that influence course ratings, student engagement, and learning outcomes in online courses, particularly gamification. Course organization, difficulty level, and skill focus play crucial roles in learner satisfaction. Courses from technology companies, those targeting beginner-level learners, and those covering widely used skills

such as Python tend to receive higher ratings. In contrast, university-offered courses and those focused on complex technical subjects like big data and machine learning often receive lower ratings, likely due to their challenging nature and difficulties in maintaining engagement.

Gamification strategies, such as progress tracking, achievement rewards, and interactive learning experiences, can significantly enhance student engagement by making learning more interactive and rewarding. The preference for shorter, skill-focused formats like Guided Projects and Courses suggests that learners value structured yet engaging content that provides immediate feedback and a sense of accomplishment. This aligns with the increasing demand for micro-credentials, which offer targeted learning pathways and motivate learners through milestone-based progression (Coursera, 2024).

From a predictive modelling perspective, log_course_reviews emerged as the most influential factor in determining course ratings, reinforcing the role of engagement and social validation in online learning. The unpruned regression tree initially fit the training data well but risked overfitting. By pruning the model to six leaves, a balance between accuracy and generalization was achieved, with a lower ASE on the test set (0.0242), indicating better predictive reliability. The structured approach of the pruned model mirrors gamification principles by focusing on key decision points rather than overwhelming learners with excessive complexity.

Ultimately, this analysis underscores the importance of engagement-driven design in online learning. Gamification techniques that integrate structured progression, interactive elements, and skill-based incentives can enhance student motivation, leading to improved learning outcomes. Online course providers can optimize engagement by incorporating elements such as badges, leaderboards, and milestone rewards, which not only sustain interest but also improve overall course ratings and retention.

References

Adjust. "App Gamification: The Ultimate Guide." *Adjust*. Accessed [February 18, 2025]. Available at https://www.adjust.com/resources/guides/app-gamification/.

Baker Stein, Marni. "Coursera's 2023 Learner Outcomes Report Highlights the Real-World Impact of Online Learning." *Coursera Blog*. May 9, 2023. Available at https://blog.coursera.org/coursera-2023-learner-outcomes-report/.

Coursera. "Fastest Growing Job Skills of 2024." *Coursera Blog*. Accessed [February 18, 2025]. Available at https://blog.coursera.org/fastest-growing-job-skills-of-2024/.

SAS Institute Inc. "Example 2: Regression Tree Using the HPSPLIT Procedure." *SAS Documentation*. Accessed [February 14, 2025]. Available at https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/stathpug/stathpug_hpsplit_examples02.htm.

Tianyi Ma. "Coursera Course Dataset." *Kaggle*. Accessed [February 1, 2025]. Available at https://www.kaggle.com/datasets/tianyimasf/coursera-course-dataset/data.

Acknowledgements

Special thanks to SAS OnDemand for Academics for providing the analytical tools necessary for this research.

APPENDIX

Output 1. Output from a PROC CONTENTS Statement

The CONTENTS Procedure

Data Set Name	PROJ.COURSERA_DATA_CLEANED	Observations	958
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	02/20/2025 11:23:10	Observation Length	7376
Last Modified	02/20/2025 11:23:10	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

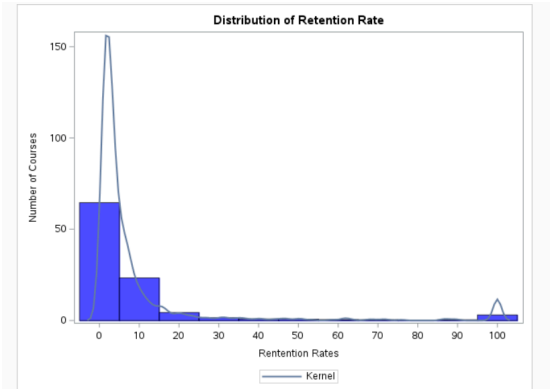
#	Variable	Type	Len	Format	Informat	Label
3	course_certificate_type	Char	24	$24.	$24.	course_certificate_type
11	course_description	Char	4473	$4473.	$4473.	course_description
6	course_difficulty	Char	12	$12.	$12.	course_difficulty
2	course_organization	Char	50	$50.	$50.	course_organization
5	course_rating	Num	8	BEST.		course_rating
12	course_reviews	Num	8			
9	course_skills	Char	960	$960.	$960.	course_skills
8	course_students_enrolled	Num	8	COMMA15.		course_students_enrolled
10	course_summary	Char	1579	$1579.	$1579.	course_summary
4	course_time	Char	17	$17.	$17.	course_time
1	course_title	Char	101	$101.	$101.	course_title
7	course_url	Char	128	$128.	$128.	course_url
13	retention_rate	Num	8			

Output 2. Output from a PROC CORR Statement

Pearson Correlation Coefficients, N = 958

	course_rating	course_students_enrolled	course_reviews	retention_rate
course_rating course_rating	1.00000	0.17332	0.10830	0.01059
course_students_enrolled course_students_enrolled	0.17332	1.00000	0.60776	-0.09864
course_reviews	0.10830	0.60776	1.00000	0.34315
retention_rate	0.01059	-0.09864	0.34315	1.00000

Figure 2. Retention Rate



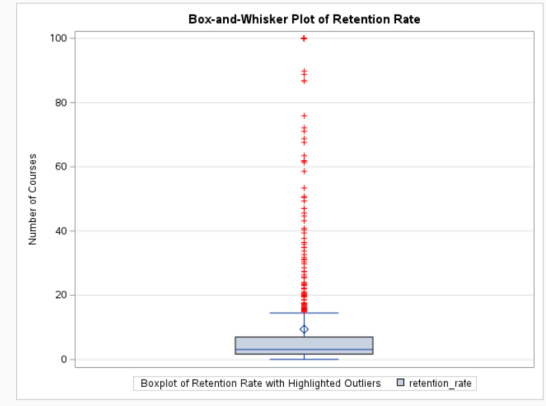


Figure 3. Course Rating

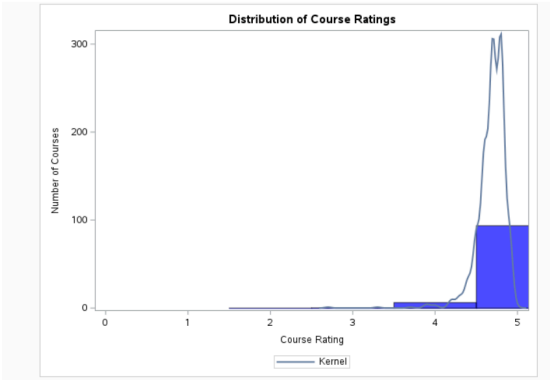
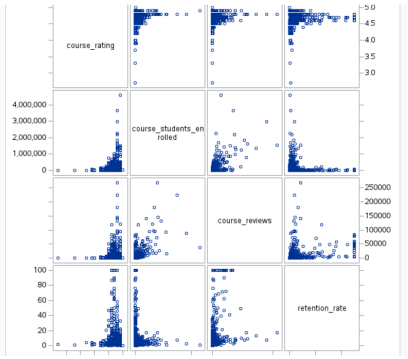


Figure 4. Normality Check



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.77172	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.215274	Pr > D	<0.0100
Cramer-von Mises	W-Sq	7.562893	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	41.95732	Pr > A-Sq	<0.0050

Figure 5. Regression Trees