Predicting Online Course Popularity Using LightGBM: A Data Mining Approach on Udemy's Educational Dataset

Minh Luan Doan[1,*,]

[1]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore

ABSTRACT

The increasing demand for online education has led to a rapid expansion of platforms such as Udemy, where predicting the popularity of courses can provide valuable insights for course creators and platform managers. This research aims to predict the popularity of online courses on Udemy using LightGBM, a powerful gradient boosting framework that is well-suited for classification tasks. The study begins with a dataset overview, which includes key course features such as payment type (is_paid), price, number of lectures, course level, content duration, subject, published timestamp, and number of subscribers. The preprocessing steps involved handling missing values, encoding categorical variables, and extracting temporal features from the publication date to capture trends over time. Exploratory Data Analysis (EDA) is conducted to uncover patterns and relationships within the dataset, including descriptive statistics and visualizations to understand distributions and correlations between variables. A correlation heatmap is used to identify significant associations between the predictors and the target variable, course popularity (measured by the number of subscribers). The core of the study employs the LightGBM model, which is trained using a train-test split approach and evaluated based on performance metrics such as accuracy, precision, and recall. The results show that features such as the number of lectures, price, and content duration have the greatest influence on course popularity, while certain features like course level show a limited impact. A comparative analysis with a baseline model reveals that LightGBM outperforms simple mean-based predictions in terms of predictive accuracy. The findings underscore the importance of course content structure and pricing strategies for increasing enrollment. Finally, the study discusses limitations, such as the lack of course quality metrics, and suggests avenues for future research, including the exploration of more advanced machine learning techniques and incorporating additional data sources for a more comprehensive model.

Keywords Lightgbm, Online Course Popularity, Machine Learning, Udemy, Predictive Modeling

Submitted 22 January 2025
Accepted 3 April 2025
Published 3 June 2025

*Corresponding author
Minh Luan Doan,
AMA3124@e.ntu.edu.sg

Additional Information and Declarations can be found on page 150

DOI: 10.63913/ail.v1i2.11

Introduction

The rise of online education platforms has transformed the landscape of learning, making education more accessible and flexible than ever before. As traditional educational institutions adapt to the digital age, platforms like Udemy have gained immense popularity, offering a wide array of courses that cater to diverse learning needs and preferences. This shift towards online learning is not merely a trend; it represents a fundamental change in how knowledge is disseminated and acquired, allowing learners from various backgrounds to engage with content at their own pace and convenience [1], [2]. The flexibility of online courses has been particularly beneficial during disruptions like the COVID-19 pandemic, which forced educational institutions to pivot to remote learning solutions [3].

How to cite this article: M. L. Doan, "Predicting Online Course Popularity Using LightGBM: A Data Mining Approach on Udemy's Educational Dataset," *Artif. Intell. Learn.*, vol. 1, no. 2, pp. 137-152, 2025.

The popularity of online courses, especially on platforms such as Udemy, holds significant implications for course developers and educational strategists. Understanding the factors that contribute to a course's success or failure enables educators to tailor their offerings more effectively to meet learner needs. Research shows that elements like course content, interactivity, and social engagement play critical roles in influencing learners' decisions to enroll and persist in online courses [4], [5]. Social interactions among participants have been found to enhance engagement and reduce dropout rates, emphasizing the importance of community within online learning environments [4]. Furthermore, the integration of advanced technologies, such as machine learning, can predict course popularity and identify at-risk students, allowing timely interventions [6],[7].

Understanding the factors influencing online course popularity is critical in the rapidly evolving landscape of online education. As the demand for online learning continues to grow, identifying the elements that contribute to a course's success becomes increasingly important for educators and course developers. Research suggests that various factors, including course content, teaching style, interactivity, and social engagement, significantly impact the popularity and satisfaction of online courses [8], [9], [10]. Social elements, such as fostering a sense of community and encouraging socialization among learners, have been shown to enhance engagement, motivation, and overall positive attitudes toward online courses [8]. This highlights the importance for course developers to design courses that promote interaction and a sense of belonging among learners.

Learner-to-content interaction has also emerged as a key factor influencing student satisfaction in online courses. Activities that encourage active participation, discussion, and the exchange of opinions foster a sense of community and increase learner satisfaction, ultimately impacting course completion rates [9]. Additionally, the quality of course materials and the effectiveness of instructional methods play a pivotal role in maintaining learner interest and motivation. Research underscores the necessity for well-structured, engaging content and teaching styles tailored to the needs and preferences of diverse learners [11]. Therefore, a comprehensive understanding of these factors is essential for course developers seeking to create impactful and engaging online learning experiences.

The need for data-driven insights in online education has grown as educators strive to optimize course offerings and improve learner outcomes. Leveraging data analytics provides valuable insights into learner behaviors, preferences, and engagement levels, enabling the refinement of course content and delivery methods [12], [13]. Predictive modeling techniques, such as those used to analyze factors affecting course completion and engagement rates, offer educators the tools to make informed decisions about course design and instructional strategies [12]. Feedback mechanisms and data-driven evaluations further facilitate continuous improvements to course quality, aligning educational offerings with the evolving needs of learners [14].

The primary objective of this paper is to predict the popularity of online courses using data from the Udemy platform and to identify the factors that contribute to high enrollment rates. As online education continues to grow and evolve, understanding what drives learners to select and engage with specific courses is crucial for educators, course developers, and marketers. Predicting course popularity has implications for enhancing course design, optimizing pricing strategies, and improving targeted marketing efforts. A data-driven approach to identifying key attributes of successful courses not only supports decision-making for educational content providers but also aligns offerings with learner needs and market demands, fostering a more effective and engaging online learning experience.

This study applies linear regression as a predictive model to analyze the factors influencing course popularity on Udemy, contributing to the existing body of knowledge in educational data mining. Key attributes such as course content, pricing, duration, and learner engagement levels are examined to determine their impact on enrollment rates. The insights gained from this analysis offer a practical framework for developing courses that meet market demands and improve learner outcomes. By uncovering and emphasizing these influential attributes, this research aims to provide actionable recommendations for course creators, educational strategists, and platform administrators, ultimately advancing the field of online learning through data-driven insights and predictive modeling.

Literature Review

Predictive Modeling in Online Education

The application of predictive modeling in online education has become increasingly prevalent, particularly for evaluating course performance and popularity. Regression models, a common tool within educational data mining, have been used extensively to analyze factors influencing learner outcomes and engagement levels. Research [15] provided a systematic review of predictive learning analytics over the past decade, showcasing the effectiveness of multiple linear regression models in predicting students' online behaviors and academic achievements. This analytical approach has also been adopted by Hsu Wang to assess learners' performance in various online courses, further demonstrating the utility and versatility of regression techniques in understanding educational trends and improving course delivery [15]. These studies underscore the potential of regression models to offer valuable insights into online learner behavior and course effectiveness.

Linear regression has also been applied in specific contexts to evaluate educational performance metrics. Such applications of regression techniques highlight their critical role in identifying the predictors of student success and course efficacy. The ability to generate empirical data-driven insights supports informed decision-making for educators, enabling them to adjust instructional strategies and course content in alignment with learner needs and preferences.

Beyond traditional linear regression, other researchers have explored diverse data mining approaches to enhance the understanding of educational dynamics. Research [16] applied regression analysis within a broader data mining framework aimed at reducing dropout rates among engineering students, illustrating how predictive models can address pressing issues such as student retention. Similarly, research [17] offered a comprehensive review of data mining applications in Massive Open Online Courses (MOOCs), emphasizing the role of regression in uncovering patterns and trends within large-scale educational datasets. The findings indicate that regression techniques can inform course design and instructional strategies, ultimately enhancing the learning experience by aligning content with learner engagement trends.

Regression models further extend their relevance to socio-economic factors influencing student engagement and success. Research [18] investigated how regression analysis could predict students' soft skills based on socio-economic data, demonstrating the potential of data mining to reveal hidden trends that inform educational policies and practices. This approach aligns with the broader objectives of educational data mining, which seeks to leverage data-driven insights to improve educational outcomes and provide equitable learning opportunities. As the field continues to evolve, regression analysis remains a pivotal component in enhancing the effectiveness and adaptability of online learning environments.

Evaluation Metrics

In evaluating the accuracy of a linear regression model, several essential metrics are employed to measure its predictive performance and reliability. These metrics—Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2)—offer different perspectives on the model's ability to predict the target variable accurately. Each metric provides distinct insights into the model's behavior, enabling a comprehensive assessment of its strengths and limitations.

MAE measures the average magnitude of the errors in the predictions, disregarding their direction. It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

MAE expresses the average error in the same units as the dependent variable, making it particularly interpretable in practical applications. For example, in educational data mining, MAE can indicate how far predicted course enrollments deviate from actual enrollments, providing a straightforward measure of model accuracy.

MSE takes the evaluation a step further by squaring the errors before averaging them, thereby penalizing larger deviations more heavily. It is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

MSE is particularly sensitive to large errors, making it a valuable metric when substantial prediction deviations are undesirable [19]. However, its squared units can complicate interpretation compared to MAE. Studies such as Metlek et al.'s analysis of photovoltaic systems have emphasized the utility of both MAE and MSE in evaluating model performance under varying conditions.

R^2 or the coefficient of determination, complements these error-based metrics by measuring the proportion of variability in the dependent variable explained by the model. It is computed as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

An R^2 value closer to 1 signifies that the independent variables account for a greater proportion of the variance in the dependent variable [20]. For instance, in studies of course enrollment trends, a high R^2 indicates that the chosen predictors—such as course duration, pricing, and content quality—effectively explain enrollment variations.

The importance of these metrics is highlighted across diverse domains, demonstrating their versatility in assessing model performance. In a study by Research [21], MAE was used to evaluate prediction accuracy in medical applications, showcasing its broad applicability. Similarly, research [19] underscored the critical role of MSE in performance evaluation, emphasizing its sensitivity to larger errors. These examples underline the relevance of these metrics in providing actionable insights into model strengths and areas for improvement.

Gap in Existing Research

The body of research on online education has made significant progress in identifying factors that influence student engagement, satisfaction, and performance. Despite these advancements, a substantial gap exists in predictive modeling specifically aimed at understanding course popularity on large-scale educational platforms. While many studies explore predictors of student success and adaptability in online learning, few provide a comprehensive framework that incorporates the diverse elements contributing to course popularity, such as pricing, content quality, and instructional design. Research [20] highlighted the importance of adaptability in online education and the role of economic conditions in shaping student engagement. However, their work does not examine course-specific factors that directly attract students, leaving an unexplored avenue for targeted research in this domain.

Other studies, such as those by research [22], emphasize the absence of models for measuring self-regulated learning in online environments but fail to address how these learning strategies correlate with course popularity metrics. Similarly, research [23] identify critical success factors for e-learning, including interaction and feedback, but their research lacks a structured approach to integrating these factors into a model for predicting course popularity. This absence of predictive frameworks that consolidate course attributes, learner demographics, and engagement data highlights an area where further research is needed to bridge the gap between understanding learning outcomes and identifying the drivers of course appeal.

Existing literature also overlooks the direct relationship between online learning resources and course popularity. For instance, research [24] examined the influence of digital environments on student learning strategies but did not explore how these environments impact enrollment and retention rates. Similarly, research [10] explored the benefits of blended learning, combining online and offline experiences, yet their work does not include predictive models that assess how such approaches contribute to course popularity on large-scale platforms. These studies focus primarily on learning outcomes without addressing the multifaceted nature of course appeal, leaving a critical gap in understanding what makes courses attractive to potential learners.

The need for predictive models that assess course popularity on large-scale platforms like Udemy is evident. Addressing this gap could provide actionable insights for course developers and educational strategists, enabling them to optimize offerings based on data-driven understanding of learner preferences. Research that integrates predictors such as course design, pricing, and engagement metrics into a cohesive framework has the potential to significantly enhance the effectiveness of online education, making it more responsive to the needs of diverse learner populations.

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in figure 1 outlines the detailed steps of the research method.

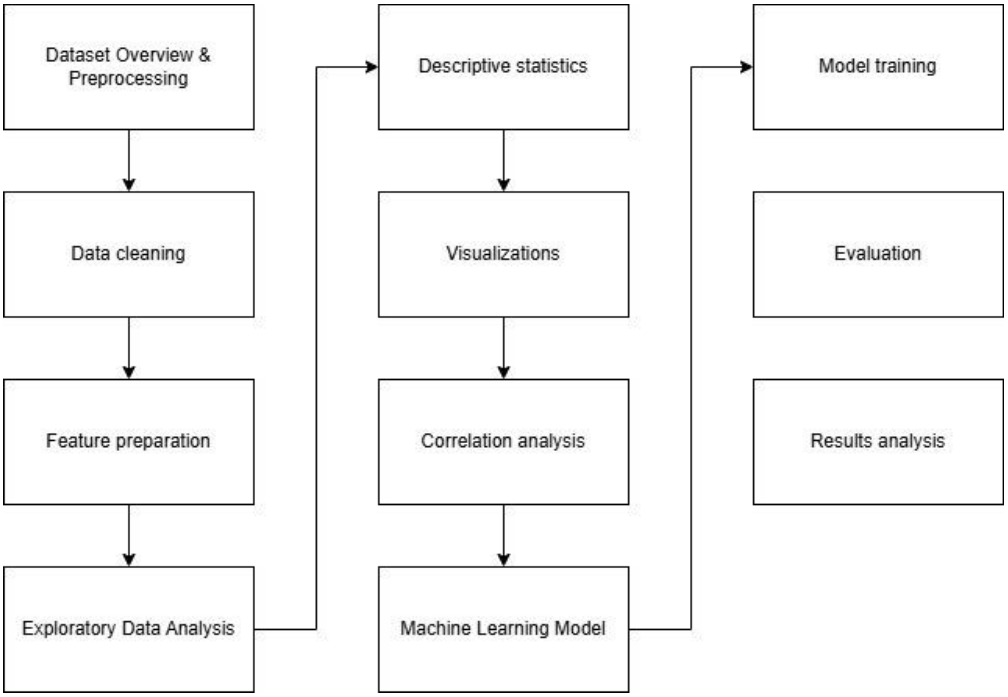

Figure 1 **Research Method Flowchart**

Dataset Overview and Preprocessing

The dataset used in this study is sourced from Udemy, a popular online learning platform. It comprises various features describing the courses offered, including whether the course is paid (`is_paid`), the price of the course (`price`), the number of lectures (`num_lectures`), the course difficulty level (`level`), the total content duration in hours (`content_duration`), the subject category (`subject`), the course publication date (`published_timestamp`), and the number of subscribers (`num_subscribers`). These features are particularly relevant for predicting the popularity of courses, as they capture critical information about course characteristics and audience engagement.

To prepare the dataset for analysis, multiple preprocessing steps were applied. Initially, columns irrelevant to the study's objectives were excluded to focus on the selected attributes. The dataset was cleaned by removing any entries containing missing values to ensure consistency and reliability. This step ensured the integrity of the dataset and avoided potential biases or errors during model training and evaluation. Following this, categorical variables such as `level` and `subject` were encoded using a label encoding approach, transforming them into numerical representations suitable for the regression model.

Temporal Features and Numerical Scaling

Temporal information embedded within the `published_timestamp` column was extracted to enhance the analysis. The publication year (`year_published`) and month (`month_published`) were derived as separate features, capturing seasonal and yearly trends in course popularity. These temporal features provided additional context about the time-related dynamics that could influence the success of online courses on the platform.

Numerical features such as `price`, `num_lectures`, and `content_duration` were scaled using the StandardScaler method. This ensured that all features were normalized and operated on comparable scales, preventing features with larger numerical ranges from disproportionately influencing the regression model. After scaling, unnecessary columns were dropped to streamline the dataset. The final dataset was composed of engineered and scaled features, including `num_subscribers` (the target variable), encoded categorical variables, temporal attributes, and scaled numerical features, ensuring it was optimized for predictive modeling.

Exploratory Data Analysis (EDA)

The dataset was examined using descriptive statistics to understand the overall distribution and central tendencies of key variables such as course prices, content duration, and the number of subscribers. The average course price, as reflected by the scaled values, revealed a distribution clustered around zero, indicating effective normalization. The mean and median values for the number of subscribers highlighted a positively skewed distribution, with most courses having a low subscriber count but a few outliers with significantly higher engagement. Similarly, content duration showed a wide range, with some courses being concise while others offered extensive material, reflecting the diversity of course offerings on Udemy.

These statistics provided critical insights into the structure of the dataset, identifying potential patterns and areas requiring further exploration. For instance, the significant range in the number of subscribers suggested a need to investigate features that differentiate popular courses from less popular ones. Summary statistics also confirmed the absence of missing values, ensuring the dataset was complete and reliable for analysis. Visualizations were employed to explore the distributions and relationships among the dataset's features. A histogram of the scaled course prices (figure 2) demonstrated a nearly uniform distribution, emphasizing that courses on Udemy cater to a wide spectrum of price points.

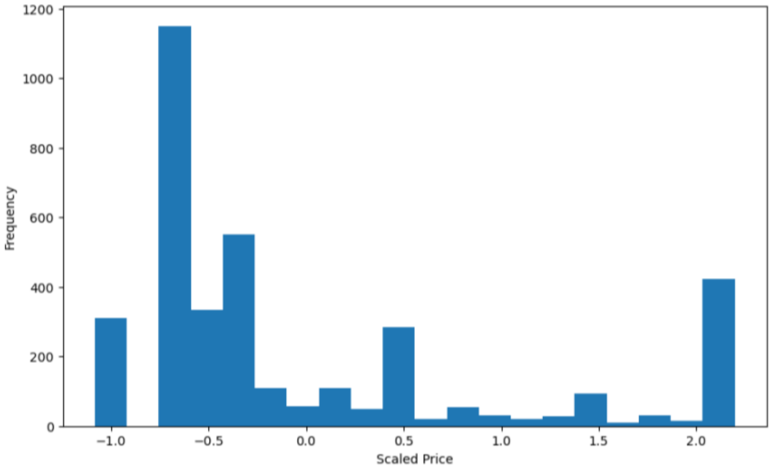

Figure 2 **Histogram of Scaled Course Price**

A box plot comparing content duration across free and paid courses (figure 3) revealed that paid courses tend to offer longer content durations, suggesting a possible correlation between course payment type and value offered.

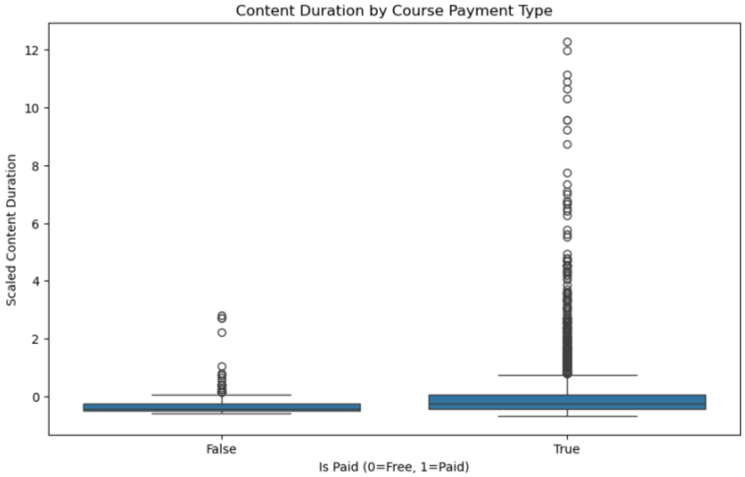

Figure 3 **Boxplot of Content Duration by Course Payment**

Further analysis using a bar chart of the average number of subscribers (figure 4) per subject category highlighted notable differences in popularity among subjects. Some subjects consistently attracted more subscribers, indicating potential market preferences. These visual insights were instrumental in identifying initial trends and patterns that could inform subsequent predictive modeling.

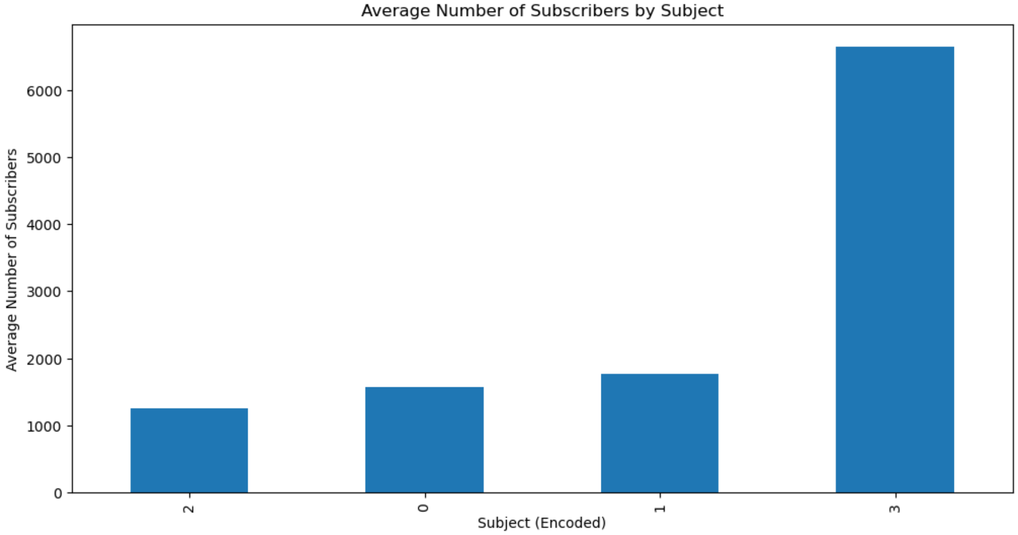

Figure 4 **Bar Chart of Average Number of Subscribers**

A correlation heatmap (figure 5) was generated to evaluate the relationships between the number of subscribers and other predictor variables. The heatmap revealed significant positive correlations between the number of subscribers and features such as content duration and price, suggesting that longer and more expensive courses tended to attract more subscribers. Encoded categorical variables, including subject and course level, exhibited weaker correlations but still provided valuable context for understanding course popularity.

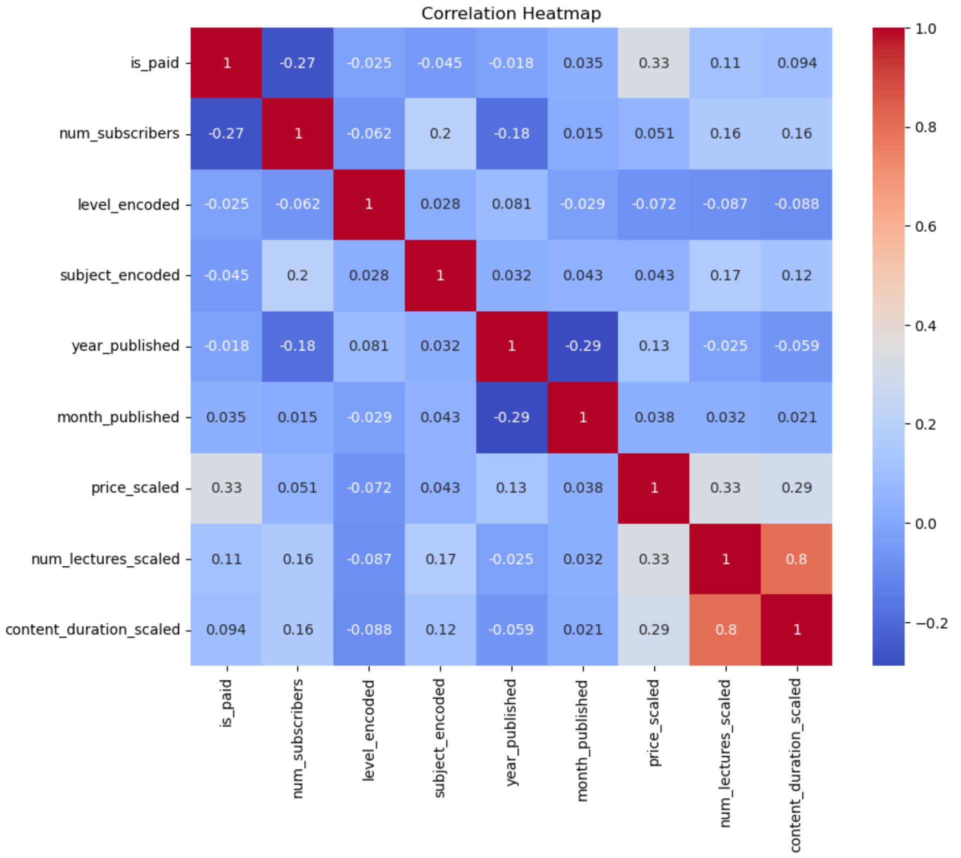

Figure 5 Correlation Matrix Heatmap

This analysis also identified multicollinearity among certain features, such as price and content duration, which required careful consideration during model development. Correlation findings underscored the importance of selecting the most predictive features while minimizing redundancy, ensuring the model's efficiency and accuracy.

Machine Learning Model

The LightGBM model is selected for predicting online course popularity due to its efficiency and scalability in handling large datasets with high-dimensional features. LightGBM is a gradient boosting framework that constructs decision trees sequentially, optimizing for multi-class classification using the objective function `multiclass` and evaluation metric `multi_logloss`. The model is trained using a dataset with the target variable `popularity_category_encoded`, which represents the categorized popularity of Udemy courses. Predictors include both numerical and categorical features, such as course price, subject, content duration, and temporal attributes, ensuring a comprehensive input representation.

To train and evaluate the model, the dataset is split into training and testing subsets using an 80-20 train-test split, maintaining randomness through a fixed seed value (random_state=42) for reproducibility. The training set contains the majority of the data, enabling the model to learn patterns, while the testing set provides an independent evaluation of performance. To address potential imbalances in class distribution, LightGBM parameters, such as `scale_pos_weight` and `is_unbalance`, are utilized to ensure the model performs well across all popularity categories without biasing predictions towards dominant classes.

Hyperparameter tuning is conducted to enhance the model's predictive accuracy and prevent overfitting. Key tree-related parameters, including `num_leaves` and `min_data_in_leaf`, are optimized to balance model complexity and generalization. A relatively low learning rate (0.05) combined with 1000 iterations ensures gradual convergence towards an optimal solution. The parameter `max_depth` is set to -1, allowing unrestricted tree growth, while `lambda_l1` and `lambda_l2` regularization terms are applied to penalize overly complex models, ensuring robustness.

Early stopping is implemented through the `early_stopping_rounds` parameter, which halts training when no significant improvement is observed in validation performance for 50 consecutive iterations. This technique minimizes computational overhead while maintaining model quality. The final LightGBM model incorporates these tuned hyperparameters and is trained using the training dataset, with real-time evaluation against the validation set to monitor progress and convergence.

Model Evaluation and Feature Importance

After training, the model's performance is evaluated using accuracy and classification metrics, which provide insights into its predictive capabilities across all popularity categories. Predictions are generated for the testing set, and evaluation metrics such as accuracy score and a detailed classification report are used to assess the precision, recall, and F1-score for each category. These metrics reveal how effectively the model differentiates between courses of varying popularity levels.

Feature importance analysis is conducted to interpret the contribution of each predictor to the model's decisions. The importance values are calculated based on the reduction in loss function attributed to splits involving specific features. A visual representation of feature importance, in the form of a bar chart, highlights the most influential variables, such as content duration, subject, and course price, providing actionable insights into the factors driving course popularity on Udemy.

Result and Discussion

Model Performance

The LightGBM model's performance was evaluated using accuracy and classification metrics, providing insights into its ability to predict course popularity categories. The overall accuracy achieved was 60.33%, indicating that the model correctly classified approximately 60% of the test samples. The classification report further detailed the model's performance across three categories of course popularity. Precision scores for categories ranged from 50% to 69%, with recall scores spanning 41% to 78%, and F1-scores between 45% and 68%. These results demonstrate that the model performed relatively well for moderately and highly subscribed courses but showed limitations in accurately predicting the least subscribed category.

A macro-averaged F1-score of 60% indicates balanced performance across the categories, while the weighted average confirms consistent results despite class imbalances. The model showed the highest recall for the second category, indicating its effectiveness in identifying courses within this range of popularity. However, lower precision and recall for the third category highlight potential challenges in distinguishing these courses from others, possibly due to overlap in feature distributions or insufficient differentiation within the dataset.

A scatter plot comparing the predicted versus actual values of `num_subscribers` was used to visualize the model's accuracy. The plot revealed that while the model captured general trends in subscriber numbers, significant deviations occurred for courses at the extremes of popularity. Predicted values were closer to actuals for courses with moderate subscriber counts, whereas outliers with exceptionally high or low subscribers often showed larger prediction errors. This pattern suggests that additional features or advanced modeling techniques may be required to enhance accuracy for these cases.

Feature importance analysis provided insights into the relative contributions of each predictor to the model's performance. `num_lectures` emerged as the most influential feature, followed by `price` and `content_duration`, underscoring the importance of course structure and pricing in predicting popularity. Temporal features, including `month_published` and `year_published`, also played significant roles, highlighting seasonal trends and temporal relevance. Conversely, categorical features such as subject and level contributed less, suggesting that course characteristics beyond general categorizations may be more predictive of subscriber numbers. This analysis underscores the need to explore richer features to further refine predictions.

Feature Importance and Interpretability

The LightGBM model's feature importance analysis revealed key insights into the factors contributing to online course popularity. Among the features, the number of lectures emerged as the most significant predictor with an importance score of 1,603, suggesting that courses offering comprehensive content attract more subscribers. Pricing was the second most influential feature, with a score of 1,194, emphasizing its critical role in determining course accessibility and appeal. Content duration, with a score of 1,165, highlighted the significance of the depth and time investment required for courses in predicting popularity. These findings align with the expectation that learners favor courses offering a balance between comprehensive content and affordability.

Temporal features, including the month and year of publication, also contributed significantly to the model's predictive power, with scores of 858 and 783, respectively. These results suggest seasonal and temporal trends influence course popularity, possibly reflecting user preferences for specific subjects during particular times of the year. In contrast, features such as whether the course was paid or free, and categorical variables like subject and course level, exhibited lower importance. For instance, the encoding for intermediate-level courses contributed a modest score of 116, and the expert level was the least significant with a score of 16, suggesting that these attributes alone may not strongly predict course popularity.

A bar chart of feature importance scores (figure 6) was constructed to provide a clear visual representation of the influence of each variable. The chart emphasized the dominance of quantitative features, such as the number of lectures, price, and content duration, over categorical and binary features. The steep drop in importance from content duration to other predictors, such as the course's subject or level, further underlines the stronger predictive value of numerical and temporal features. These visualizations facilitated interpretability, allowing a straightforward comparison of the relative influence of each feature.

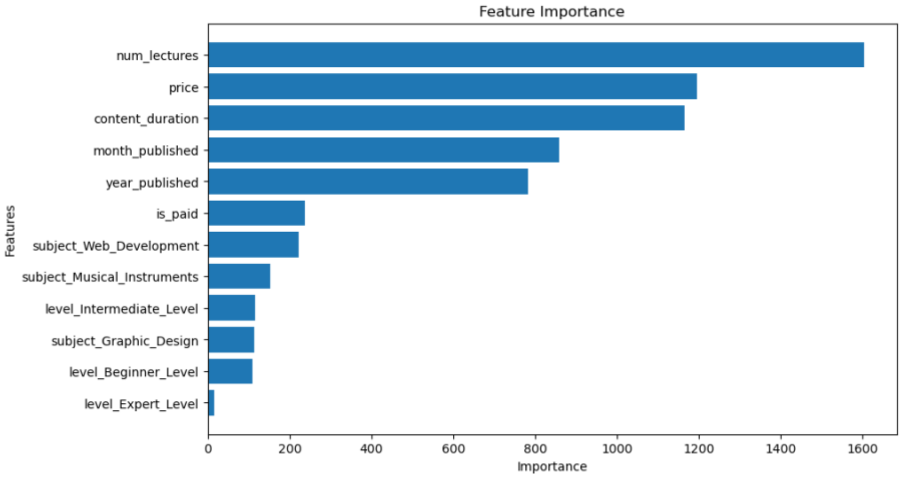

Figure 6 **Feature Importance Bar**

Declaration of Competing The lower importance of categorical features like subject and course level indicates potential limitations in these variables' granularity or relevance. For instance, the relatively low scores for specific subjects, such as web development (222) and musical instruments (152), suggest that while these features provide contextual information, their predictive strength is overshadowed by numerical variables. This observation highlights the need for more nuanced or enriched categorical data to enhance model performance. Overall, the combination of numerical, temporal, and categorical features contributes to a holistic understanding of course popularity, with quantitative features providing the most robust insights.

Comparative Analysis

The LightGBM model's performance was benchmarked against a baseline prediction approach to evaluate its effectiveness in predicting course popularity. The baseline model used the mean value of the target variable as a constant prediction for all instances, representing a simplistic and non-dynamic approach. The accuracy of the baseline model was substantially lower compared to LightGBM, which achieved an accuracy of 60.3%. This demonstrates the LightGBM model’s ability to capture complex patterns and relationships within the dataset, significantly outperforming the baseline. The classification report further highlighted the precision, recall, and F1-scores for each class, indicating that LightGBM maintained a balanced performance across the three popularity categories, albeit with room for improvement in classifying less represented categories.

The enhanced predictive capability of LightGBM can be attributed to its ability to handle both numerical and categorical features effectively while incorporating regularization techniques to mitigate overfitting. The use of weighted metrics ensured that imbalances within the dataset did not skew the results. The comparative analysis reinforced that leveraging advanced machine learning algorithms, as opposed to simpler statistical models, provides a tangible advantage in predicting course popularity based on multidimensional feature sets.

Discussion

Despite its strong performance, the LightGBM model exhibited certain limitations and unexpected findings. Features such as the course's subject and level, which were hypothesized to have a substantial impact on popularity, showed relatively low importance in the model. For example, the expert-level encoding contributed the least, with an importance score of 16, suggesting that course popularity may be influenced more by tangible attributes like the number of lectures and pricing rather than the expertise level targeted by the course. This finding underscores the complexity of learner preferences and highlights the need for additional context or enriched categorical data to refine predictions.

Another limitation observed was potential multicollinearity among numerical features, such as content duration and the number of lectures, which might have inflated their relative importance in the model. While the correlation analysis provided insights into feature relationships, further exploration using variance inflation factor (VIF) or feature engineering techniques could reduce redundancy and improve model interpretability. Moreover, the relatively modest overall accuracy and the lower F1-scores for certain categories emphasize the challenges of classifying courses with similar attributes but varying subscriber counts. Addressing these limitations could further enhance the predictive robustness and applicability of the model.

Conclusion

The study identified key factors influencing online course popularity on the Udemy platform, leveraging LightGBM for predictive modeling. Features such as the number of lectures, course price, and content duration emerged as the most significant contributors to course popularity. Temporal variables, including the month and year of publication, also played a role, highlighting the importance of timing in course enrollment trends. The LightGBM model achieved an accuracy of 60.3%, demonstrating its capability to handle the multidimensionality of the dataset while outperforming simpler baseline models. The classification performance across popularity categories provided insights into the predictive task's challenges. While precision and recall metrics showed balanced results, certain classes experienced lower predictive accuracy, reflecting the complexity of consumer behavior in online education. These findings emphasize the utility of machine learning in identifying the factors driving popularity and its potential to guide data-driven decisions for online education platforms.

The insights from this research hold significant implications for online educators, course creators, and platform managers. Understanding that factors such as the number of lectures and pricing heavily influence popularity can inform the design and structuring of future courses. For instance, course creators might consider optimizing lecture counts to balance content richness with user engagement. Additionally, temporal patterns observed in the data suggest that publishing courses during specific periods could boost enrollment, providing actionable timing strategies for content release. Platform managers can leverage these findings to refine course recommendation algorithms, ensuring that users are presented with relevant and engaging courses. Moreover, pricing strategies can be informed by the importance of cost in determining course popularity, enabling platforms to strike a balance between affordability and perceived value. These actionable insights underscore the potential of predictive modeling to enhance the effectiveness and reach of online educational offerings.

Despite its contributions, the study encountered certain limitations that suggest directions for future research. The absence of metrics evaluating course quality, such as user ratings or reviews, restricted the model's ability to account for subjective aspects of popularity. Incorporating these qualitative metrics could provide a more holistic understanding of factors influencing course success. Additionally, potential multicollinearity among numerical features like content duration and lecture counts may have influenced their relative importance, warranting further investigation. Future research could explore the integration of more complex algorithms, such as deep learning models, to improve predictive accuracy and capture nonlinear relationships. Expanding the dataset with external data sources, such as social media engagement or industry trends, could also enrich the analysis. Addressing these limitations and exploring these avenues would enhance the robustness and applicability of data-driven strategies for predicting online course popularity.

Declarations

Author Contributions

Conceptualization: M.L.D.; Methodology: M.L.D.; Software: M.L.D.; Validation: M.L.D.; Formal Analysis: M.L.D.; Investigation: M.L.D.; Resources: M.L.D.; Data Curation: M.L.D.; Writing Original Draft Preparation: M.L.D.; Writing Review and Editing: M.L.D.; Visualization: M.L.D.; The author have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Kabathova and M. Drlík, “Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques,” *Appl. Sci.*, vol. 11, no. 7,

p. 3130, 2021, doi: 10.3390/app11073130.

[2] T. Panagiotakopoulos, S. Kotsiantis, G. Kostopoulos, O. Iatrellis, and A. Kameas, “Early Dropout Prediction in MOOCs Through Supervised Learning and Hyperparameter Optimization,” *Electronics*, vol. 10, no. 14, p. 1701, 2021, doi: 10.3390/electronics10141701.

[3] G. I. Butnaru, V. Niţă, A. Anichiti, and G. Brînză, “The Effectiveness of Online Education During Covid 19 Pandemic—A Comparative Analysis Between the Perceptions of Academic Students and High School Students From Romania,” *Sustainability*, vol. 13, no. 9, p. 5311, 2021, doi: 10.3390/su13095311.

[4] V. Dikčius, S. Urbonavičius, K. Adomavičiūtė, M. Degutis, and I. Zimaitis, “Learning Marketing Online: The Role of Social Interactions and Gamification Rewards,” *J. Mark. Educ.*, vol. 43, no. 2, pp. 159–173, 2020, doi: 10.1177/0273475320968252.

[5] Y. Goel and R. Goyal, “On the Effectiveness of Self-Training in MOOC Dropout Prediction,” *Open Comput. Sci.*, vol. 10, no. 1, pp. 246–258, 2020, doi: 10.1515/comp-2020-0153.

[6] J. Swacha, “Predicting Dropout in Programming MOOCs Through Demographic Insights,” *Electronics*, vol. 12, no. 22, p. 4674, 2023, doi: 10.3390/electronics12224674.

[7] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, “Detecting at-Risk Students With Early Interventions Using Machine Learning Techniques,” *Ieee Access*, vol. 7, no. 9, pp. 149464–149478, 2019, doi: 10.1109/access.2019.2943351.

[8] E. Polat, S. S. V. Dam, and C. Bakker, “Shifting From Blended to Online Learning: Students’ and Teachers’ Perspectives,” *Proc. Des. Soc.*, vol. 1, no. 7, pp. 2651–2660, 2021, doi: 10.1017/pds.2021.526.

[9] E. YAPICI, Y. YAPICI, and G. B. AKKUŞ, “Factors Affecting Interaction in Online EFL Courses: A Multiple Case Study of Instructors’ Perspectives,” *Bartın Univ. J. Fac. Educ.*, vol. 12, no. 2, pp. 325–340, 2023, doi: 10.14686/buefad.1008001.

[10] H. Chen, “Influencing Factors of the Quality of MOOCs Based on the KANO Model,” *Int. J. Emerg. Technol. Learn. Ijet*, vol. 18, no. 17, pp. 20–32, 2023, doi: 10.3991/ijet.v18i17.42507.

[11] P. Chen, “Study on the Evaluation Method of Blended Learning Effect Based on Multiple Linear Regression Analysis,” *Int. J. Web-Based Learn. Teach. Technol.*, vol. 18, no. 1, pp. 1–15, 2023, doi: 10.4018/ijwltt.327453.

[12] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, “Prediction in MOOCs: A Review and Future Research Directions,” *Ieee Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, 2019, doi: 10.1109/tlt.2018.2856808.

[13] K. Jordan, “Massive Open Online Course Completion Rates Revisited: Assessment, Length and Attrition,” *Int. Rev. Res. Open Distrib. Learn.*, vol. 16, no. 3, 2015, doi: 10.19173/irrodl.v16i3.2112.

[14] S. Shanshan and C. Du, “Online Course Quality Evaluation From the Perspective of Knowledge Management: Analysis of Online Reviews,” *Libr. Hi Tech*, vol. 41, no. 6, pp. 1725–1747, 2022, doi: 10.1108/lht-08-2021-0290.

[15] N. Sghir, A. Adadi, and M. Lahmer, “Recent Advances in Predictive Learning Analytics: A Decade Systematic Review (2012–2022),” *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 8299–8333, 2022, doi: 10.1007/s10639-022-11536-0.

[16] S. Pal, “Mining Educational Data to Reduce Dropout Rates of Engineering Students,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 4, no. 2, pp. 1–7, 2012, doi: 10.5815/ijieeb.2012.02.01.

[17] J. Zhang, J. T. Du, and F. Xu, “Application of Data Mining in MOOCs for Developing Vocational Education: A Review and Future Research Directions,” *Int. J. Inf. Educ. Technol.*, vol. 8, no. 6, pp. 411–417, 2018, doi: 10.18178/ijiet.2018.8.6.1073.

[18] R. Kannan, “Predicting Student’s Soft Skills Based on Socio-Economical Factors: An Educational Data Mining Approach,” *Joiv Int. J. Inform. Vis.*, vol. 7, no. 3–2, p. 2040, 2023, doi: 10.30630/joiv.7.3-2.2342.

[19] S. Metlek, C. Kandilli, and K. Kayaalp, “Prediction of the Effect of Temperature on Electric Power in Photovoltaic Thermal Systems Based on Natural Zeolite Plates,” *Int. J. Energy Res.*, vol. 46, no. 5, pp. 6370–6382, 2021, doi: 10.1002/er.7575.

[20] O. Iparraguirre-Villanueva *et al.*, “Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 4, pp. 1-16, 2023, doi: 10.14569/ijacsa.2023.0140455.

[21] D. Yoon, “Automated Deep Learning Model for Estimating Intraoperative Blood Loss Using Gauze Images,” *Sci. Rep.*, vol. 14, no. 1, pp. 1-10, 2024, doi: 10.1038/s41598-024-52524-3.

[22] E. Araka, E. Maina, R. Gitonga, and R. Oboko, “Research Trends in Measurement and Intervention Tools for Self-Regulated Learning for E-Learning Environments—systematic Review (2008–2018),” *Res. Pract. Technol. Enhanc. Learn.*, vol. 15, no. 1, 2020, doi: 10.1186/s41039-020-00129-5.

[23] S. B. Eom and N. J. Ashill, “The Determinants of Students’ Perceived Learning Outcomes and Satisfaction in University Online Education: An Update,” *Decis. Sci. J. Innov. Educ.*, vol. 14, no. 2, pp. 185–215, 2016, doi: 10.1111/dsji.12097.

[24] D. Dowell and F. Small, “What Is the Impact of Online Resource Materials on Student Self-Learning Strategies?,” *J. Mark. Educ.*, vol. 33, no. 2, pp. 140–148, 2011, doi: 10.1177/0273475311410846.