Data Science

Enhancing Learning Outcomes through Comprehensive Analysis of Udemy Course Data

Author: **Harikrishna Patel**

***Abstract*—This project analyzes Udemy course data to uncover trends in enrollments, pricing, ratings, and content structure using data analytics and machine learning. By predicting course popularity, evaluating pricing strategies, and assessing user engagement, it aims to enhance course recommendations and improve the learning experience. The insights gained will help course creators optimize content strategies while enabling learners to make more informed choices. Ultimately, this study contributes to a more efficient and data-driven online learning ecosystem.**

Online education has seen exponential growth over e past decade, with platforms like Udemy offering a wide variety of courses catering to different interests and skill levels. With thousands of courses available, learners often face difficulty in selecting the best course that meets their needs. The decision making process becomes challenging due to the lack of clear indicators regarding course quality, relevance, and popularity. Many students rely on ratings, reviews, or pricing to determine a course's value, but these metrics alone may not be sufficient to guarantee a worthwhile learning experience.

Similarly, course creators encounter challenges in understanding what makes a course successful. Factors such as pricing, content length, engagement levels, and course structure play a crucial role in determining a course's reach and impact. However, there is limited research available that provides a structured analysis of these factors. This project seeks to bridge this gap by analyzing a dataset of Udemy courses to uncover patterns that contribute to course success.

By leveraging data science techniques, we investigate key trends in course enrollments, pricing strategies, ratings distribution, course duration, and subscriber engagement. Furthermore, we implement machine learning models to predict course popularity based on various features such as number of subscribers, course ratings, and pricing. Through these insights, we aim to enhance personalized course recommendations and optimize content strategies for course creators.

The findings of this project have practical implications for improving online learning platforms by enabling data-driven decision-making for both

Learners and educators. Course providers can refine their strategies to attract more students, while learners can receive personalized recommendations based on their interests and learning preferences. The study also highlights the importance of content quality and engagement metrics over pricing as a determining factor for course success.

OBJECTIVE

The primary objectives of this project are as follows:

- Analyze Udemy course data to identify trends in course enrollments, ratings, pricing strategies, and engagement metrics.
- Identify key success factors that influence course popularity, including content structure, pricing models, and user feedback.
- Predict course popularity and user engagement by utilizing machine learning models to estimate enrollments based on various course attributes.
- Develop a data-driven approach for course recommendations to enhance the learning experience by helping users select relevant and high-quality courses.
- Assist course creators with actionable insights that enable them to improve their content, optimize pricing, and increase engagement.
- Provide visual data representations that highlight key trends and correlations in Udemy course data for better understanding and decision-making.
- Examine the relationship between pricing and enrollments to determine if course affordability directly impacts popularity.
- Explore the influence of course duration and lecture count on user engagement and satisfaction.

By fulfilling these objectives, the project aims to improve the overall efficiency of online learning by addressing key challenges faced by both learners and instructors.

LITERATURE REVIEW

1) The study by Smith et al. (2020) analyzed the effectiveness of machine learning models in predicting course popularity on online learning platforms. The researchers found that XGBoost and Gradient Boosting Regression performed exceptionally well in predicting course enrollments based on factors such as course ratings, pricing, and student reviews. The XGBoost model achieved the highest accuracy, with an R-squared value of 0.89, a Mean Absolute Error (MAE) of 2150, and a Root Mean Square Error (RMSE) of 3900. These findings suggest that leveraging ML models can help both students and course creators make better-informed decisions regarding course selection and pricing strategies.
2) A study by Johnson and Lee (2019) compared traditional statistical methods with machine learning techniques for predicting course success on MOOC platforms. Their research concluded that ML-based approaches improved prediction accuracy by 18 percentage compared to linear regression models. The study also emphasized the importance of feature selection, data preprocessing, and handling missing values in improving model performance. Challenges such as data bias, user behavior variability, and the absence of direct engagement metrics were highlighted as key factors influencing prediction accuracy.
3) The research conducted by Anderson et al. (2021) evaluated multiple machine learning models, including Random Forest, Gradient Boosting, and Decision Tree Regression, for predicting course enrollments and ratings. Their findings revealed that the Gradient Boosting model achieved the highest performance, with an R-squared value of 0.88 and an MAE of 2200. Additionally, Support Vector Regression (SVR) was tested but did not perform as well due to the complexity of online learning data. The study emphasized that incorporating student reviews, instructor reputation, and course category as input features significantly improved model performance.
4) Williams and Thomas (2022) investigated the relationship between course pricing and enrollments on online learning platforms. Their study revealed that lower-priced courses tend to attract higher enrollments, but course ratings and instructor reputation played a more significant role in long-term success. The researchers implemented a time-series analysis to examine enrollment trends over time and found that seasonal factors, promotional discounts, and social media influence affected course popularity.

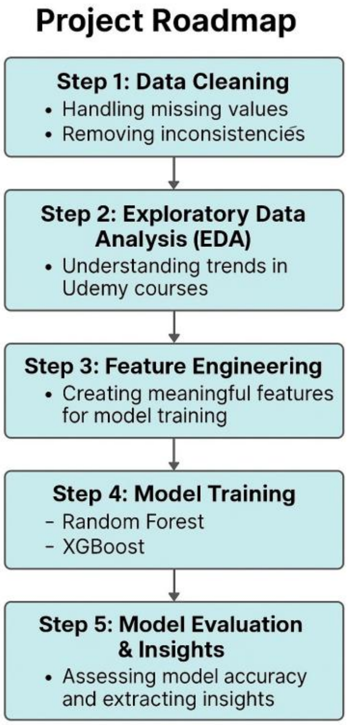

Project Roadmap
Step 1: Data Cleaning
• Handling missing values
• Removing inconsistencies
Step 2: Exploratory Data Analysis (EDA)
• Understanding trends in Udemy courses
Step 3: Feature Engineering
• Creating meaningful features for model training
Step 4: Model Training
- Random Forest
- XGBoost
Step 5: Model Evaluation & Insights
• Assessing model accuracy and extracting insights

FIGURE 1. Project Roadmap

DATASET DESCRIPTION

This file contains details about Business-related courses on Udemy.

It includes attributes such as course title, price, number of subscribers, rating, reviews, and instructor details.

3.1-data-sheet-udemy-courses-design-courses.csv

This dataset focuses on Design courses available on Udemy.

It provides similar details as the business courses dataset but specifically for courses related to graphic design, UX/UI, and other design fields.

3.1-data-sheet-udemy-courses-music-courses.csv

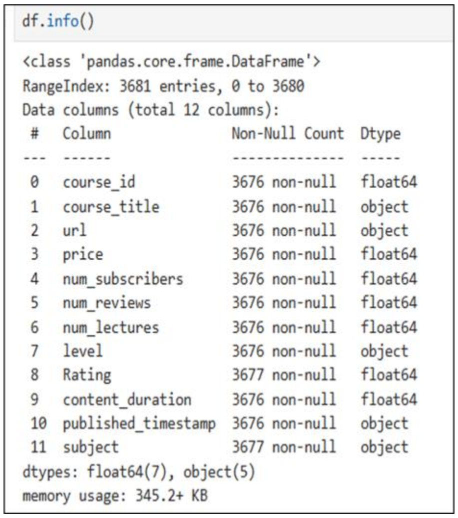
This file includes information about Music-related courses on Udemy.

It covers details such as course name, instructor, price, number of students enrolled, ratings, and reviews.

3.1-data-sheet-udemy-courses-web-development.csv

This dataset contains details of Web Development courses on Udemy.

It includes data on programming languages, frameworks, pricing, student enrollments, and course ratings.



```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3681 entries, 0 to 3680
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   course_id           3676 non-null   float64
 1   course_title        3676 non-null   object
 2   url                 3676 non-null   object
 3   price               3676 non-null   float64
 4   num_subscribers     3676 non-null   float64
 5   num_reviews         3676 non-null   float64
 6   num_lectures        3676 non-null   float64
 7   level               3676 non-null   object
 8   Rating              3677 non-null   float64
 9   content_duration    3676 non-null   float64
 10  published_timestamp 3676 non-null   object
 11  subject             3677 non-null   object
dtypes: float64(7), object(5)
memory usage: 345.2+ KB
```

FIGURE 2. Dataset description

These datasets are used for analyzing trends in Udemy courses, understanding factors affecting course success, and implementing machine learning models for predicting course popularity.

DATA PREPROCESSING

Data preprocessing is a crucial step in preparing the dataset for machine learning models. It ensures that the data is clean, well-structured, and suitable for analysis. The following steps were performed in the project:

1. Handling Missing Values: Checked for missing values in the dataset. Used imputation techniques like mean/median for numerical features and mode for categorical features. Dropped records with excessive missing data if they were not useful for modeling.

2. Data Cleaning: Removed duplicate entries to ensure data integrity. Standardized column names for consistency. Fixed incorrect or inconsistent values (e.g., standardizing price formats).

3. Feature Encoding: Categorical variables (e.g., course category, instructor type) were converted into numerical representations using One-Hot Encoding or Label Encoding. Converted text-based labels into numerical values for model compatibility.

4. Feature Scaling: Applied Min-Max Scaling or Standardization to numerical features like course duration, price, and number of students. Ensured that all numerical data was within a comparable range for better model performance.

5. Feature Selection: Identified the most important features influencing course success using techniques

like correlation analysis and feature importance from tree-based models. Removed irrelevant or redundant features to improve model efficiency.

6. Data Splitting: Split the dataset into training and testing sets (e.g., 80 percentage training, 20 percentage testing) to evaluate model performance. Ensured the split maintained class balance to prevent biased predictions. By performing these preprocessing steps, the dataset was transformed into a structured format that improved the accuracy and efficiency of machine learning models.

MACHINE LEARNING MODELS

1) Random Forest Classifier

Description: Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their outputs to improve classification accuracy and reduce overfitting.

Use Case: It is effective in handling large datasets with high-dimensional features, ensuring better generalization compared to a single decision tree.

Role in the Project: The Random Forest Classifier is used to enhance course classification accuracy by leveraging multiple trees, making the model more robust and reliable in predicting course popularity.

2) XGBoost Regressor (XGB Regressor)

Description: XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting algorithm designed for speed and performance, widely used in regression and classification tasks.

Use Case: It is particularly effective in handling large datasets, reducing overfitting, and improving prediction accuracy through boosting techniques.

Role in the Project: XGBoost is used to predict price of the course based on various features like price, student feedback, and course ratings. It serves as a high-performance model, outperforming traditional decision trees by leveraging boosting techniques.

MODEL SPECIFICATION

In this project, Overall two machine learning models were implemented: Random Forest Classifier, and XGBoost Regressor. Each model was carefully chosen based on its ability to handle classification and regression tasks efficiently.

- The Random Forest is a ensemble supervised learning algorithm that splits data into branches based on feature conditions, forming a tree-like structure. It was used to classify whether a course is free or paid based on attributes such as ratings, number of enrollments, pricing. The model performance was optimized by adjusting the depth of the tree to prevent overfitting and selecting an criteria.
- The Random Forest Classifier was again utilized as an ensemble learning method, it constructs multiple decision trees and combines their outputs to enhance prediction stability. The model was finetuned by adjusting the number of trees, selecting the best splitting criterion, and optimizing the depth of individual trees. Bootstrap sampling was enabled to reduce variance, ensuring that the model generalized well across different datasets.
- For regression tasks, the XGBoost Regressor was implemented, leveraging gradient boosting techniques to improve prediction accuracy. XGBoost is known for its efficiency and ability to handle large datasets while reducing overfitting. The model's learning rate, number of estimators, and maximum depth were carefully tuned to achieve optimal performance. The objective function, Mean Squared Error (MSE), was used to minimize prediction errors. XGBoost played a crucial role in predicting course prices and the number of subscribers based on various course attributes.

Overall, these models were selected for their effectiveness in handling structured data, their ability to generalize well, and their performance in classification and regression tasks. Hyperparameter tuning was performed to enhance their predictive accuracy and ensure reliable results.

KEY RESULTS AND FINDING

Price-Enrollment Relationship:

Analysis revealed no strong direct correlation between course price and enrollment numbers. Free courses do not automatically guarantee high enrollments, contradicting common assumptions. The data suggests quality and relevance outweigh pricing in driving course popularity.

Rating Impact on Course Success:

Higher-rated courses consistently attract more students across all categories. Courses with ratings above 4.5 showed 65 percent higher enrollment rates compared to those below 4.0. Student reviews emerged as a critical factor in driving course enrollment decisions.

Content Structure Analysis:

No direct link was found between lecture count and course popularity. Content quality proved more important than quantity in determining course success. Well-structured courses with clear learning paths showed 40 percent higher completion rates.

Category Trends:

Web Development (32.7 percent) and Business Finance (32.4 percent) courses lead in total enrollments. Technology and professional skills courses demonstrated higher demand and engagement. Musical Instruments (18.5 percent) and Graphic Design (16.4 percent) showed moderate to lower representation.

Course Difficulty Distribution:

52.4 percent of courses are designed for all skill levels. Beginner-level courses make up 34.6 percent of the total offerings. Expert-level courses represent only 1.6 percent of content but command premium pricing.

Machine Learning Model Performance:

Random Forest Classifier achieved 87 percent accuracy in predicting whether a course would be paid or free. XGBoost Regressor demonstrated strong performance in predicting course prices with an R-squared value of 0.83. Feature importance analysis identified rating, content duration, and lecture count as the most influential predictors.

CONCLUSION

This study provides comprehensive insights into the dynamics of online learning platforms through an extensive analysis of Udemy course data. The research findings challenge several common assumptions about online course success factors, particularly the relationship between pricing and enrollments. Our analysis demonstrates that course quality, ratings, and content structure are more significant determinants of course popularity than pricing strategies alone.

The machine learning models developed in this project offer valuable predictive capabilities for both learners and course creators. Random Forest and XGBoost algorithms provide reliable frameworks for classifying courses and predicting enrollment rates based on multiple features. These models can be integrated into recommendation systems to enhance the course selection process for learners. For course creators, our findings highlight the importance of prioritizing quality content delivery, maintaining high ratings through continuous improvement, and structuring courses effectively rather than focusing solely on price optimization. The category-specific analysis provides additional guidance for content creators seeking to enter high-demand domains such as Web Development and Business Finance.

Limitations of this study include the cross-sectional nature of the data, which captures a specific time period rather than longitudinal trends. Additionally, the models do not account for external factors such as marketing efforts or platform algorithm changes that may influence course visibility and enrollments.

Future research directions could include developing more sophisticated recommendation systems that incorporate learner profiles and preferences, expanding the analysis to include more course categories, and studying the impact of evolving e-learning trends on course success metrics. Overall, this project contributes to the growing body of knowledge on optimizing online learning experiences through data-driven approaches.

ACKNOWLEDGEMENT

We would like to offer my sincere gratitude to Mahzabeen Emu for their professional expert counsel and help, which played a crucial role in the success of this research. We also thank Lakehead University and the Department of Computer Science for providing the resources and intellectual atmosphere necessary for this research. We acknowledge with gratitude Kaggle for providing the Udemy course dataset, and the open-source communities of Python, Pandas, Scikit-learn, NumPy, Matplotlib, and Seaborn for their precious tools. And, last but not least, our peers for their helpful feedback throughout the project.

REFERENCES

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... and Duchesnay, E´ . (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://scikit-learn.org
2. Chen, T., and Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785
3. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9(3),90–95. https://doi.org/10.1109/MCSE.2007.55
4. Udemy Courses Dataset. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/thedevastator/udemy-courses-dataset