Comparison of Student Interaction Between Paid and Unpaid Courses

Michael Bemus

Purdue University Fort Wayne

Abstract

The question of pricing plans for online service businesses is fundamental to the services a business produces. Often, we see the division between high quality, paid content versus mass marketable, unpaid content in online services. This paper attempts to analyze the efficacy of a mixed pricing approach as attempted by Udemy, an online education platform. This paper applies the t-test to measure the performances of paid and unpaid courses on their website. The analysis finds that paid courses do better on Udemy by all measures that were examined. However, this paper does not find the difference in performance to be predictable through logistic methods. Future research could create a more stratified examination of price compared to performance, as well as apply more complex machine learning methods to create pricing predictions.

keywords: t-test, logistic regression

Comparison of Student Interaction Between Paid and Unpaid Courses

Introduction

For online service businesses, the question of whether to use a paid or an unpaid model for their business activities is a fundamental decision that can shape almost every aspect of how the business presents itself. Under a paid model, companies present their services with an esteemed sense of quality. They justify their prices through the respect of their brand and the quality of the services they provide. Under an unpaid model, companies present their services as "for the consumer." Their services often use advertisements or premium content, which generate profit for the company. However, the baseline service is usually tailored for simplicity, relying more on what the consumers need than what else they might desire.

Both models have their pros and cons, and neither one is mutually exclusive. Udemy is an example of a hybrid model. The company provides virtual education over a variety of subjects like Development, IT & Software, Entrepreneurship, and more (Hossain, 2022). Through both paid and unpaid courses that are offered in 79 different languages, Udemy attracts a wide target market of learners that are interested in a variety of fields Hossain (2022).

Udemy provides both paid and unpaid courses for users to learn from (Hossain, 2022). The unpaid courses attract customers to the wide range of topics on their website (Hossain, 2022). Then, the paid courses offer more quality knowledge to their customers, with prices ranging from $0.10 to almost $1,000 (Hossain, 2022).

This can be a very useful strategy for businesses with plenty of material to provide as both paid and unpaid. However, it begs the question of which users prefer. Do customers interact more with unpaid course material? Or does the high quantity of paid material offer more interest to regular users? In the following sections, I will explore this question and see if there is a predictable trend between user interaction and a course's pricing model.

Methods

In the following sections, I will discuss the data set used for this analysis, as well as the hypotheses and tests used for my experiment.

Data Set

For my analysis, I used the Udemy Courses data provided by Hossian Hossain (2022), a contributor on the online data sharing service *Kaggle*. In it, Hossian Hossain (2022) provides data on more than 209,000 courses, collecting columns such as content length, published time, language, and topic. In this experiment, I utilized the columns indicating whether the course is paid or unpaid, the number of subscribers, the number of comments, the number of reviews, and the average rating. Provided in the results, I also examined the correlations between those variables, as well as the number of lectures and the content length.

Hypotheses

For my analysis, I started with two hypotheses. The first I examined was that paid courses see less user interaction than unpaid courses. To test this, I first assumed the null hypothesis that paid courses on average have more subscribers, comments, and reviews than unpaid courses. For my alternative hypothesis, I claimed that paid courses on average have fewer subscribers, comments, and reviews.

For my second hypothesis, I wanted to test whether unpaid courses would have lower average ratings than paid courses. For my null hypothesis, I assumed unpaid courses have higher average ratings than paid courses. Then, for my alternative hypothesis, I claimed that unpaid courses have lower average ratings than paid courses.

I used these two hypotheses to test the efficacy of the unpaid versus paid business models. If the data shows a proficient unpaid model, that would likely imply unpaid courses are being used more frequently than paid courses to attract new users. In addition, if Udemy has a proficient paid model, that would imply paid courses have higher ratings, and therefore are higher quality, than unpaid courses.

Experiment

To test these hypotheses, I ran t-tests of each individual comparison to evaluate their validity. While there are many forms of the t-test, they generally follow the same process. First, find the t-statistic. Second, compute the probability of that t-statistic. Last, evaluate that probability against some probability level, α.

To compute the t-statistic, I used the following equation:

$$t = \frac{\bar{x}_{obs.} - \bar{x}_{hyp.}}{s}$$

where $\bar{x}_{obs.}$ is the observed mean of the variable, X, which represents the variable (paid or unpaid) being tested against $\bar{x}_{hyp.}$, which is the hypothesized mean, the mean of the variable that x_{obs} is compared to. s is the standard deviation of the two distributions.

After computing the t-statistic, I evaluated it on the CDF of the t-distribution with degrees of freedom equal to the number of observations in the data frame minus 2. Then, using an α level of 0.05, I compared the probability of that t-statistic to α, and if the t-statistic was smaller, I rejected the null hypothesis. Otherwise, I failed to reject the null hypothesis.

In addition to the t-test, I also attempted to fit the data to a logistic regression model using the variables compared in the above analysis: number of subscribers, number of comments, number of reviews, and average rating. The model uses likelihood estimation to find the best fit of the logistic equation:

$$y = \frac{1}{1+e^{-\beta X}}$$

which determines whether a value is 0 or 1 based on a pre-selected probability threshold. For this model, I used a threshold of 0.5 to evenly split the values.

To evaluate the model, I divided the data in an 80-20 split, with 80% of the data used for training and 20% of the data used for testing. I then found the f1-scores and confusion matrices for both training and testing.

Results

As shown by table 1, I found low t-statistics for all variables of hypothesis 1. Each one had a probability nearing 1. Because it is greater than my pre-established α, I fail to reject the null hypothesis.

For hypothesis 2, I found almost the opposite. I calculated a high t-statistic for the variable, which provided a p-value near zero. Because this is lower than my α, I reject the null hypothesis.

Addressing the logistic regression model, I found that the maximum likelihood estimate of the model produces an f1-score of 0.94 for paid courses and 0.01 for unpaid courses. This provides a general accuracy of 0.90 for the distribution on the training set.

On the testing set, the model had an f1-score of 0.94 of paid courses and 0.01 on unpaid courses. This gave a general accuracy of 0.89 on the distribution.

Discussion

As shown in the results, my first hypothesis was incorrect and the second was correct. This shows that, on these measures, paid courses perform better than unpaid courses. The generate more subscribers, comments, and reviews, as well as higher average ratings.

This suggests that Udemy has a greater focus on their paid content rather than their unpaid. This fits with the distribution of their course price plans. Of their total catalog, almost 22,000 of their courses are unpaid, whereas almost 188,000 of their courses are paid. For Udemy, the product is the paid courses. The unpaid courses are not necessarily their marketing focus.

An avenue for future research could be a price threshold for maximum user interaction on the platform. The data set provides a price column for every entry. By breaking up these prices into groups, one could draw comparisons between them to estimate the group that maximizes ratings and user traction.

Concerning the logistic model, Figure 2 shows that the confusion matricies generally identify values as paid, which provides relatively high accuracy because most

values fall under this category. However, it is not only predicting paid. Some values do receive unpaid predictions. However, the model predicts an unpaid value incorrectly more often than it predicts unpaid correctly.

Figure 1 provides some evidence as to why this occurs. None of the variables are too highly correlated with the variable "is_paid", and there is often not much difference between the x values of unpaid courses compared to paid courses.

The visualizations of Figure 2 provide further evidence for this. In subfigures 1 and 2, the green values represent unpaid courses, and the gray courses represent paid courses. Even in the 2-D space, it is hard to find clusterings of unpaid courses outside of the clustering of paid courses.

These low correlations returns high numbers of false negatives. This could potentially suggest that Udemy should change their pricing plan for some of their courses.

In future analysis, the use of different machine learning techniques could design a better fit for the model. Other methods such as Random Forests or Neural Networks might be able to create better predictions about the data.

Conclusion

Though Udemy has a mixed strategy for their paid and unpaid services, this analysis shows that they focus most on their paid content. Paid courses generally performs better than their unpaid courses. However, under the logistic regression model, this is not a predictable trend.

Future research could better segment the data for analysis. This would provide more information about the optimal pricing plan on the platform. In addition, other machine learning models could provide better predictions for which courses are free compared to the courses that are paid for.

References

Hossain. (2022). *Udemy courses.* (https://www.kaggle.com/datasets/hossaingh/udemy-courses?select=Course_info.csv)

Table 1

Hypothesis Results

Hypothesis	**t-statistic**	**p-value**
1a. Subscribers	−27.36	1.00
1b. Reviews	−5.65	1.00
1c. Comments	−7.03	1.00
2. Ratings	64.09	0.00

Figure 1. Relation of Variables to Payment Variable

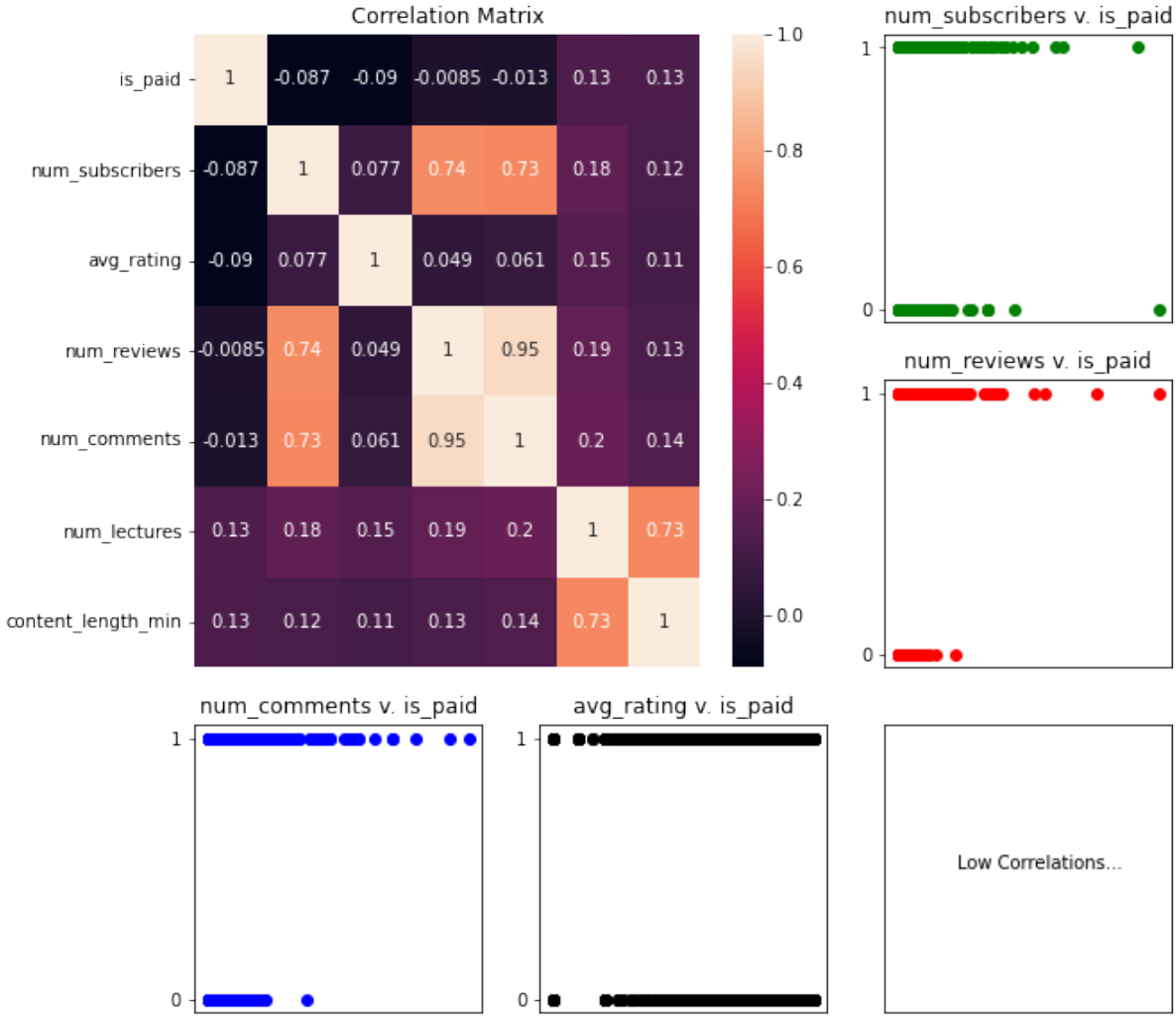

Figure 2. Payment Varaible Visualization and Logistic Regression Results