Midterm Project

MA 678

Xiaojia Liu

2020.12.03

Abstract

In the year 2020, with the outbreak of COVID-19, many students who previously had no experience in distance learning started to explore online sessions. Online course providers including Udemy and Coursera have experienced significant boomings in their enrollment.

In this project, the topic I'm investigating is the potential cause for one course under financial category to become popular on Udemy. By looking into the correlation between number of subscribers and other factors, I seek to illustrate how these factors are correlated with the number of subscribers for certain course.

After serval attempts, I observed that the number of review and number of published lectures are most positive correlated with the number of subscribers.

It is notable that the number of subscribers we use here can only reflect how popular a course is at that moment, instead of how one course has become popular. What's more, we can only tell how these factors are correlated with the popularity, but we should stay cautious and not taking them as the reason for one course to become popular. It could be the case that number of subscribers affects the predictors, in contrast.

Introduction

I. Background

The main reason for me to explore on this topic is my interest in finance industry. By learning what are the possible reasons for an online finance course to become popular on Udemy, I wish to know what are the factors I should look at, and use those as the criteria to select one for the upcoming winter break.

In summary section, I will interpret the result, what it may tell us about popularity, and most importantly what are the limitations.

II. Dataset

The dataset contains 19 columns and 13608 rows as observations. Among the 19 columns, id and title does not provide enough information for modeling and thus are not included as predictors.

To explain the variables contained in this dataset:

- id : the course ID of selected course
- title : the unique names of the courses available under the development category on Udemy.
- url: the URL of the course
- is_paid : a boolean value displaying true if the course is paid or false if otherwise
- num_subscribers : the number of people who have subscribed that course
- avg_rating : the average rating of the course
- avg rating recent : the recent changes in the average rating
- num_reviews : the number of ratings that a course has received
- num_ published_lectures : the number of lectures the course offers

- num_ published_ practice_tests : the number of practice tests that a course offers
- created : the time of creation of the course
- published_time : the time of the course's publishing
- discounted_ price_amount : the discounted price which a certain course is being offered at
- discounted_ price_currency : the currency corresponding to the discounted price which a certain course is being offered at
- price_ detail_amount : the original price of a particular course
- price_ detail_currency : the currency corresponding to the price detail amount for a course

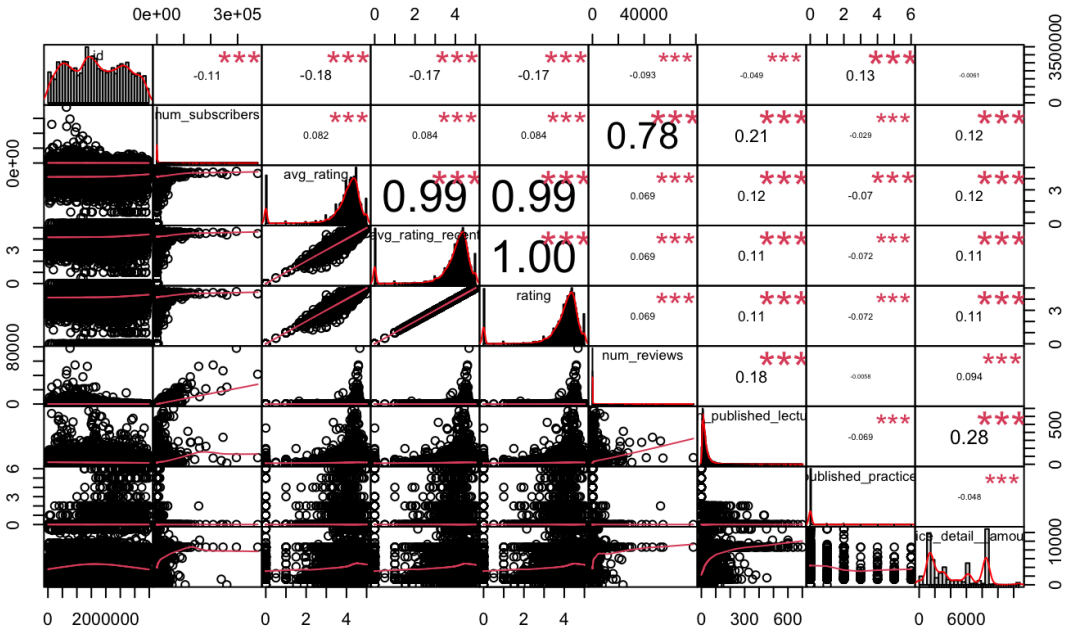
This dataset is obtained from Kaggle under: https://www.kaggle.com/jilkothari/finance-accounting-courses-udemy-13k-course. The data was last updated in September, 2020, so it lacks the information for the months after.

Methods

I. Exploratory Data Analysis

The first thing I did in EDA was to explain the NA values. Empty values are replace with NA. Also for courses with not discount, instead of having NA as discount amount, an amount of 0 makes more sense.

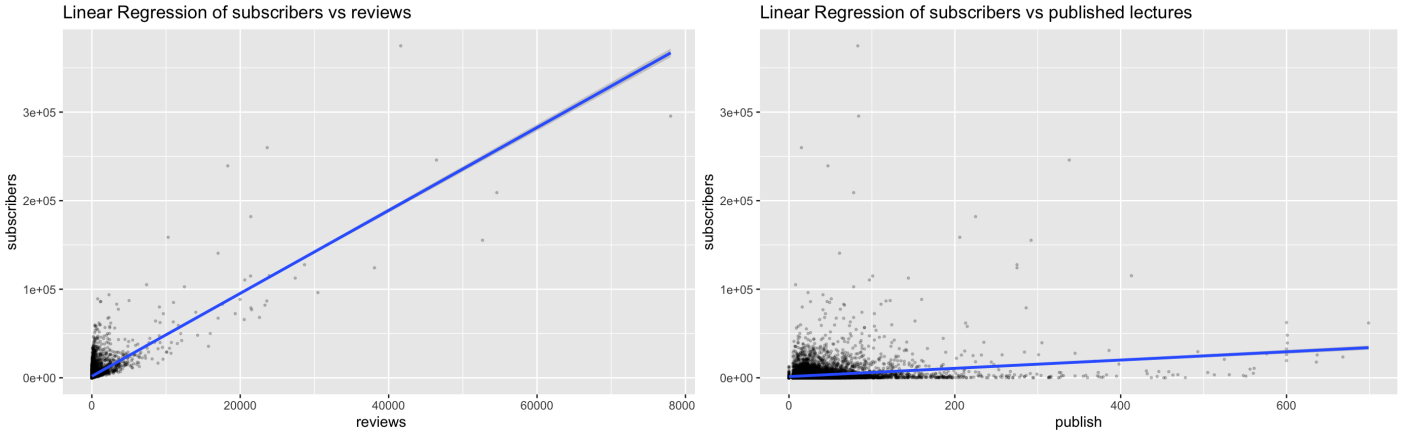
A correlation matrix was drawn in order to visualize the relationship between subscribers and other variables. As shown in the figure 1.1, the num_subscribers is most correlated with num_reviews with a correlation value as 0.78, with the second most correlated variable being num_published_lectures. The "***" symbols next to the correlation value are the significant values, which in this case, corresponding to extremely small p-value near 0. It is notable that this matrix also indicates a potential correlation between num_reviews and num_published_lectures. It makes sense because the more published lectures available for a certain online course, the subscribers have more suggestions to make and more potential for leaving some reviews. What's more, this correlation matrix is based on pearson correlation, and these significant symbol only represent the p-values and the implication of why p-values being significant should be evaluated separately.



II. Models

The complexity of the dataset makes it difficult to decide on which model to apply. Under this case, I picked the 2 most correlated variables, drew scatterplot for each, in order to explore the distributions and check if there's potential available model fit.

As suggested by the 2 scatter plots, there might be some linear correlation underlay. However, from the simple linear model we can see that the y, num_subscribers, spans out greatly towards larger values, which makes the models not representative enough for courses with subscribers lower than 10000.



After careful research, I acknowledge that I didn't find a valid way to solve this problem. Some article argued that if the outlier does not change the results but does affect assumptions, one may drop the outlier, and remain those outliers otherwise (Karen)

Under this argument, the following models are evaluated: 1) Multi-regression of (subscribers ~ reviews * publish) ; 2) Multilevel linear model of subscribers vs review with publish being the random factor; and 2 models repeat previous ones but limit the data evaluate only to those courses with subscribers less than 100000.

Result

To measure and compare the performances of the four models, I use Cross Validation method.

It turns out that there is no significant difference between using multilinear regression and multilevel regression.

The RMSE for Multilinear is 12595.27 and for Multilevel regression is 12660.86.

Discussion

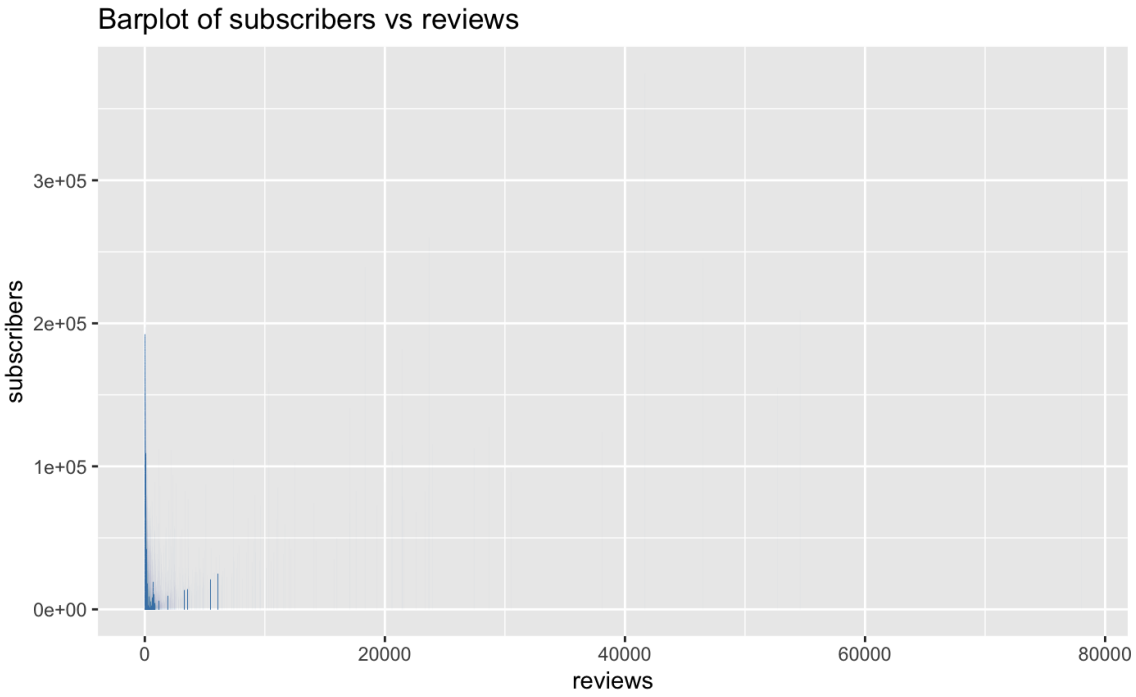
Results Implications

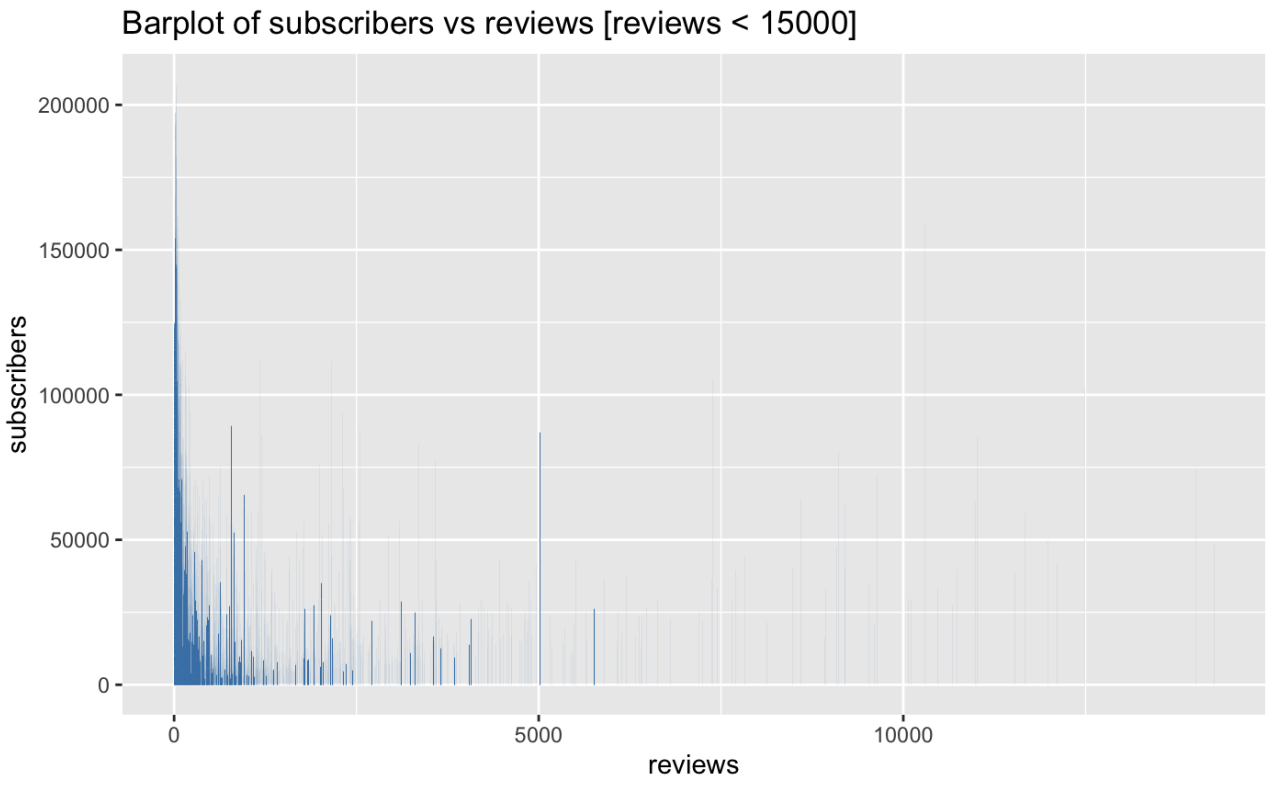
The most influential factor is number of review, and second one being number of published lectures.

As suggested by the comments for my short presentation, the description of the course definitely worth to be taken into account. In fact, I have spend much time looked up the methods to analysis keywords and the way to numerically represent their effects. However, I wasn't able to complete the string operation in r, so this will definitely be improved if time allows.

Another limitation of this project is the way data collected. This brings me back to a problem I've talked about earlier in this report. The dataset contains only the information at the moment it was

collect. As a result, it will be impossible and inaccurate to analysis their effects, which should be measure within a time period, of certain factor on the subscriber.


Barplot of subscribers vs reviews
subscribers
3e+05
2e+05
1e+05
0e+00
0
20000
40000
60000
80000
reviews



Reference

Karen Grace-Martin, Outliers: To Drop or Not to Drop, https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/ , 2015

Appendix