Developing a Ranking System for Udemy's "Data & Analytics" courses

Written by:

Danfeng Cao, Haonan Wang,

Rajasa Kakkera, and Ana Parra Vera

Table of Contents

Executive Summary

A leading online learning platform Udemy has more than 100,000 courses and keeps adding more courses every month. The objective of this report is to help Udemy in predicting the relevancy score and course labels to improve their search result for students based on the course content details, instructor details and the length of the course. This model will facilitate course designers to identify the areas of improvement to improve the visibility and popularity of the existing course content.

To build the proof of concept, the “Data & Analytics” category was considered. A total of 1,019 courses were available in the given domain. Majority of the ratings for the course work was between 3.5 and 4.5 and the price range for the courses was in between $25 to $150.

These insights were taken into consideration while building an ensemble of machine learning models to predict the relevancy score and the course label. The model predicts the relevancy score and the course label with an accuracy of 91%. Language has a very significant impact on the relevancy score. Also, course features like whether it is downloadable, and price have higher impact on the relevancy score. Courses with higher relevancy score directly impact the course label. Hence, Udemy should use this model to improve the search relevance of courses and to understand what needs to be done to improve the course popularity.

Introduction

Udemy is an online educational platform aimed at helping students and professionals to master new skills. The website provides more than 100,000 online video classes. As a marketplace service provider with close to 24 million students, Udemy must ensure that the students get relevant search results when they query for a course.[1] To show better search results, Udemy considers relevance score into account and ensures that students get to see the right courses on querying. Relevance score is calculated based on the ratings, reviews that are received from the students.

Besides the relevancy score, students might be curious to know a course's label (best seller, or highest rated). These labels can help the students know which courses are currently preferred by other students using the Udemy platform.

The aim of our project is to develop a machine learning model to predict the relevancy score and the label. This model will help Udemy in improving their search relevance and in understanding the courses to be prioritized. Hence, whenever a new course is added to the website, the model will predict the relevancy score and the label and basis the prediction Udemy can decide the placement of the course on the website.

Thus, this project can help Udemy in improving the students studying experience.

[1] Erbay, Hamza. "Learn about Udemy Culture, Mission, and Careers: About Us." Udemy About, Udemy About, about.udemy.com/.

Data Characteristics

To develop the model, a dataset was created by extracting data for 25 key variables for all the analytical courses from the Udemy website. The dataset contained 1019 rows of data. Exploratory Data Analysis (EDA) was performed on the dataset and the steps are explained below

a) **Data cleansing:** Cleansing the data is very important as the data quality improves and will improve efficiency of the model. The datatype of all the columns were checked and the necessary changes were made for the respective columns. For e.g.: the float datatype 'price' was converted to integer by converting the currency from dollars to cents. The columns not relevant in predicting the relevancy score were dropped from the dataset.The column 'caption' had null values and hence was imputed with 'No caption' .

b) **Univariate analysis:** For all the columns, boxplots and histograms were plotted to identify the outliers and the spread of the data. (see **Appendix 1**). The key finding was that 50 % of the average rating score for all the courses was in the range of 3.6 to 4.5. For 86 courses there were no ratings since they were relatively new courses.There were 6 courses for which more than 150000 students rated the course. There were 4 courses for which more than 90,000 students had enrolled. These are the courses for which the number of people who rated for the course is also very high. Hence, we can say that there is a positive correlation between the number of students who rated the course and the students who enrolled for the course. Also, 78% percent of the courses are taught is English. There are other language options like Portuguese and Spanish

c) **Bivariate Analysis:** A heatmap and scatterplot matrix was created for the dataset to analyze the correlation between the variables (see **Appendix 2**) . The number of lectures, the no of hours of videos and the number of articles were positively correlated. Price is negatively correlated to the average number of students enrolled for a particular instructor

Model Development, Estimation and Results

Upon cleaning the data, we created indicator variables for categorical variables such as label and language. Considering the wide range in value for various variables, in order to make it easier for us to explain the effect of different variables on dependent variables, we standardized all numerical data using StandardScaler from the Python package scikit-learn. Then, the data was divided into a training and testing set to develop two models. One is a regression model on relevancy score, and the other is a multinomial logistic regression model for label prediction.

Regression Model on Relevancy Score

Relevancy score is an attribute provided and calculated by udemy, which is used for determining the sequence of all courses while users search for courses by keywords or categories. Other service providers like Google, Facebook and Instagram also have the same attribute. Facebook, for example, uses the Facebook Relevance Score to estimate how well the ads resonate with the target audience. To create the score, Facebook looks at the ad draft, considers the audience's potential positive and negative feedback, and boils it down to a score of 1 to 10. And the score will update based on the actual feedback received.[2] The higher the score, the more likely Facebook is to put the ads in front of the audience. Udemy's relevancy score works similarly. The higher the relevancy score, the more likely the course will be shown at the top of the list. Since the calculation principle behind this score has not been shared by Udemy, we decided to explore how they calculate the relevancy score for courses. A linear regression is used to investigate the principle.

With regard to models, two models were run to predict relevancy score—basic linear regression model and lasso regression model, and information criteria was used to select the best model. The results are shown in the table below.

	Number of Used Variables	**OOS R-Squared**	**AIC**	**AICc**	**BIC**
Basic Linear Regression	40	0.9962	1608.2	1612.6	1801.0
Lasso Regression	22	0.9968	1414.5	1415.9	1522.7

For the basic linear regression model, we put all 40 variables into the model to predict score, and the Out-of-Sample (OOS) R-squared is 0.9962, which is very high. Besides, we also tried to use 5-fold cross validation to increase accuracy, and the OOS R-squared only increased by 0.001. For lasso regression, we first tested the best alpha and 0.012 was

[2] "Relevance Score for Facebook Ads" Facebook for Business, www.facebook.com/business/news/relevance-score.

selected to penalize variables. Lasso regression selected 22 variables to estimate relevancy score, and the OOS R-squared reached to 0.9968, which is slightly higher than basic regression model. Considering the reduction of number of variables, lasso regression performed better. To make sure that our result is reliable, we also tested multicollinearity by calculating VIF. According to the formula and definition of VIF, when VIF is larger than 10, serious multicollinearity problem exists. The result of VIF is shown in the **Appendix 3.1**. Fortunately, multicollinearity did not exist and most VIFs are close to 1, which proves that our result is accurate and reliable.

In addition to OOS R-squared, we also use information criteria to select the best model, including AIC, AICc and BIC. According to the table, it is clear that the lasso regression model is better than the basic linear model.

Multinomial Logistic Regression Model on Labels

As stated earlier, the label variable on the data reflects whether each course is categorized as a “Best Seller”, a “Highest Rated”, or a “Normal” course. A normal course is a course that does not have this special label, which is the case for the majority of courses (see **Appendix 3.2**) In order to predict such categories, we trained a multinomial logistic regression on the training dataset that contains 75% of all courses.

A multinomial logistic regression was chosen since the label values are not ordered. As an extension of logistic models, multinomial logistic regressions allow us to predict the probability of membership in each category.[3] Therefore, for each course on our test

[3] Williams, Richard. "Multinomial Logit Models - Overview" University of Notre Dame, https://www3.nd.edu/~rwilliam/stats3/Mlogit1.pdf

dataset, we calculated the probability of it being under each label, and used a threshold of 50% to categorize each course. By comparing the actual labels of the testing data against the predicted label (based on the probabilities mentioned above), we created an accuracy matrix to reflex the performance of our model:

	Predicted value			
Actual value		'BestSeller'	'HighestRated'	‘Normal’
	'BestSeller'	231	2	1
	'HighestRated'	13	2	0
	‘Normal’	6	0	0

The accuracy is the total number of courses that were correctly labeled by our multinomial logistic regression model. In this case, we have 233 correctly classified courses divided by the total of 255 courses available in the testing dataset. Therefore, the accuracy of our model is 91.37%.

The most significant variables that led to the “Normal” label are whether captions exist in a course, whether the course is taught in English, the rating score, price, number of lectures, average number of students who have rated the instructor(s), the average value of the instructor(s), and the course’s relevancy score (see **Appendix 3.4** for details).

Recommendations and Managerial Implications

Increasing Relevancy Score

As we have discussed before, relevancy score is related to the position of the course in the search result. A course with a higher score will be located at the top of the list, which makes it more likely for users to find and even take that course. Thus, it is essential for course providers to increase the relevancy score in order to maximize exposure of their courses and generate more profits. Our model explains how to increase such exposure.

As mentioned earlier, the lasso regression model was the best model to predict the relevancy score, and 22 variables are related to it. The coefficients of these variables are shown in the **Appendix 1.3**. Although it might sound counterintuitive, language is the factor which influences the relevancy score most. If the instructor teaches courses in English, the relevancy score will increase by 99.37. However, if the instructor uses traditional Chinese or Indonesian, the relevancy score will decrease by around 4 points. While 4 points may seem like a small value, it is actually pretty significant because the scores vary from course to course in decimals (i.e. from 110.93 to 110.98). In addition, courses with a higher price, more downloadable resources and more reviews will be more preferred, but courses which have more instructors and are updated more recently will have lower scores. Course designers could properly adjust their course designs on the basis of our model to achieve higher popularity.

Predicting Labels

Our model can predict with approximately 91% accuracy whether a course will stand out from the rest by being categorized as “Best Seller” or “Highest Rated”, as opposed to having the common label of “Normal”. By knowing which variables are most statistically significant in the relationship with the label “Normal” (see **Appendix 3.4** for details), course designers can take them into consideration in order to make a course stand out from the rest. In particular, it is recommended that courses aim to obtain a higher rating score, price, relevancy score, number of lectures available, and average number of students who have rated the instructor(s) in order to deviate from the “Normal” category. This result in a better outcome, whether it is a “Best Seller” or “Highest Rated”, the course will still stand out from the rest.

Conclusion

In order to differentiate “Data & Analytics” courses from Udemy, we have developed two models. The first model is a lasso regression model on courses’ relevancy score, and the other is a multinomial logistic regression model for predicting a course’s label. The lasso regression model has an R-squared of 0.9968, while the multinomial regression has an accuracy of 0.9137. While the R-squared of the first model seems suspiciously high, we have checked for multicollinearity and there is none.

We have seen how the course being talked in English is one of the most significant factors that leads to a higher relevancy score. Other significant factors are a higher price, number of downloadable resources and number of reviews for the course. On the other hand, having less instructors and having been updated more recently will yield a lower relevancy score.

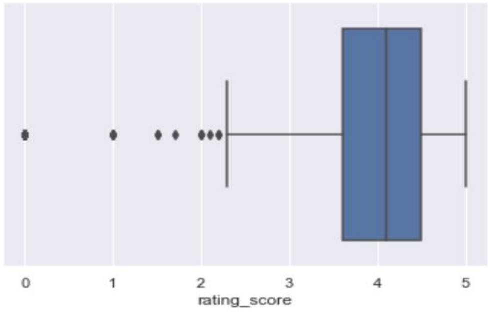
It is interesting to note that the price is a common significant variable across the models. It impacts both the relevancy score as well as the probability of a course standing out (being labeled as “Best Seller” or “Highest Rated”, as opposed to “Normal”). The relevancy score itself is another significant factor for the label prediction, as well as the rating score, number of lectures available, and average number of students who have rated the instructor(s).

These factors are important for improving the search relevance of courses, as well as understanding what makes a course stand out from the rest. Therefore, Udemy’s course designers should consider the recommendations presented on this report in order to maximize a course’s visibility, popularity, and therefore, profitability.

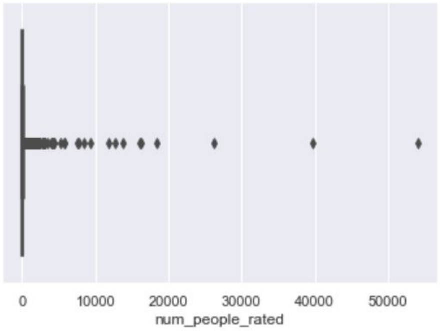
Appendix

1 Univariate analysis

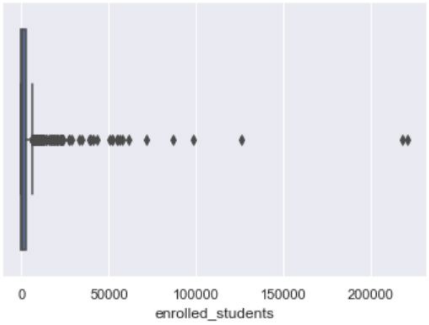
1.1 Rating Score boxplot



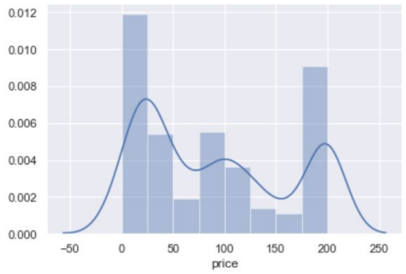
1.2 Number of people who rated the course



1.3 Students enrolled for the course

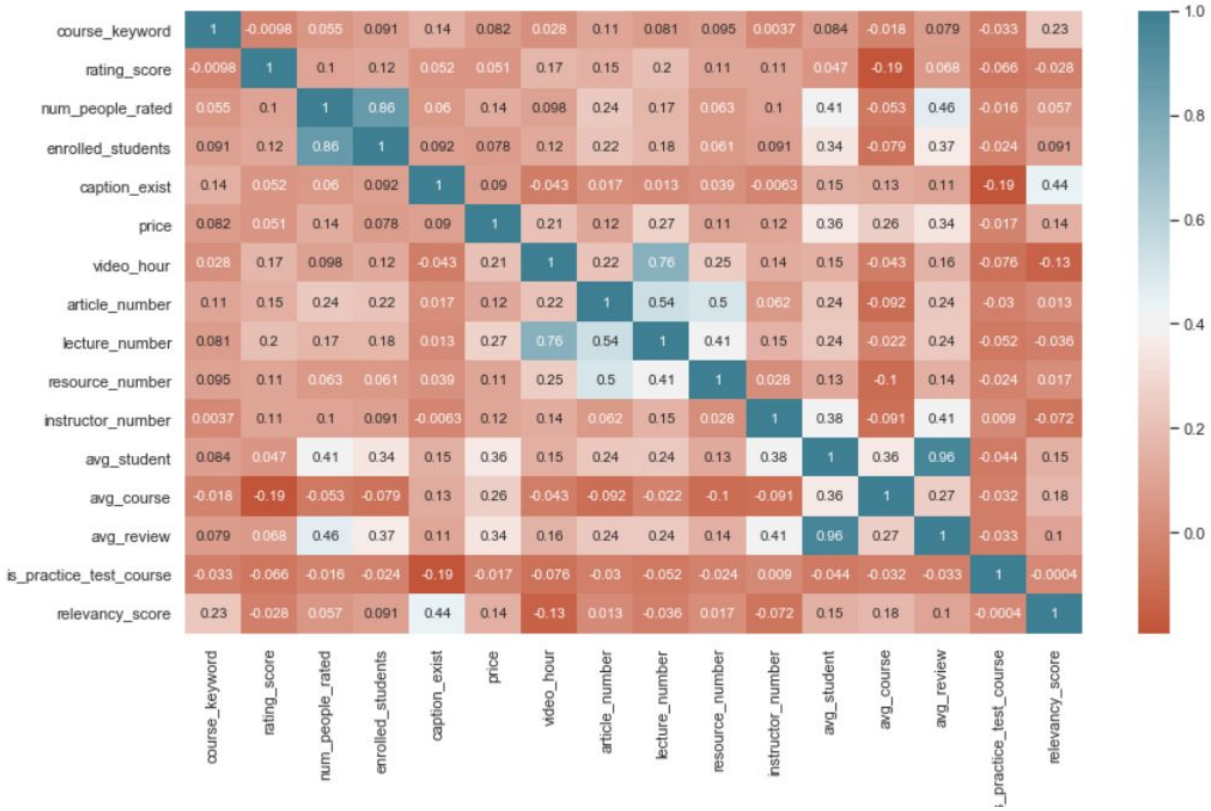


1.4 Price



2 Bivariate Analysis

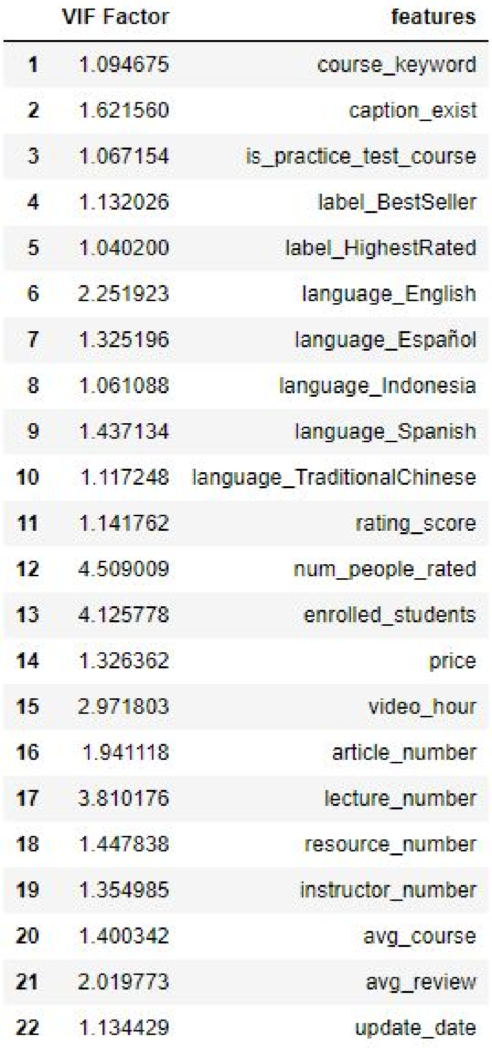
2.1 Heat Map



	course_keyword	rating_score	num_people_rated	enrolled_students	caption_exist	price	video_hour	article_number	lecture_number	resource_number	instructor_number	avg_student	avg_course	avg_review	is_practice_test_course	relevancy_score
course_keyword	1	-0.0098	0.055	0.091	0.14	0.082	0.028	0.11	0.081	0.095	0.0037	0.084	-0.018	0.079	-0.033	0.23
rating_score	-0.0098	1	0.1	0.12	0.052	0.051	0.17	0.15	0.2	0.11	0.11	0.047	-0.19	0.068	-0.066	-0.028
num_people_rated	0.055	0.1	1	0.86	0.06	0.14	0.098	0.24	0.17	0.063	0.1	0.41	-0.053	0.46	-0.016	0.057
enrolled_students	0.091	0.12	0.86	1	0.092	0.078	0.12	0.22	0.18	0.061	0.091	0.34	-0.079	0.37	-0.024	0.091
caption_exist	0.14	0.052	0.06	0.092	1	0.09	-0.043	0.017	0.013	0.039	-0.0063	0.15	0.13	0.11	-0.19	0.44
price	0.082	0.051	0.14	0.078	0.09	1	0.21	0.12	0.27	0.11	0.12	0.36	0.26	0.34	-0.017	0.14
video_hour	0.028	0.17	0.098	0.12	-0.043	0.21	1	0.22	0.76	0.25	0.14	0.15	-0.043	0.16	-0.076	-0.13
article_number	0.11	0.15	0.24	0.22	0.017	0.12	0.22	1	0.54	0.5	0.062	0.24	-0.092	0.24	-0.03	0.013
lecture_number	0.081	0.2	0.17	0.18	0.013	0.27	0.76	0.54	1	0.41	0.15	0.24	-0.022	0.24	-0.052	-0.036
resource_number	0.095	0.11	0.063	0.061	0.039	0.11	0.25	0.5	0.41	1	0.028	0.13	-0.1	0.14	-0.024	0.017
instructor_number	0.0037	0.11	0.1	0.091	-0.0063	0.12	0.14	0.062	0.15	0.028	1	0.38	-0.091	0.41	0.009	-0.072
avg_student	0.084	0.047	0.41	0.34	0.15	0.36	0.15	0.24	0.24	0.13	0.38	1	0.36	0.96	-0.044	0.15
avg_course	-0.018	-0.19	-0.053	-0.079	0.13	0.26	-0.043	-0.092	-0.022	-0.1	-0.091	0.36	1	0.27	-0.032	0.18
avg_review	0.079	0.068	0.46	0.37	0.11	0.34	0.16	0.24	0.24	0.14	0.41	0.96	0.27	1	-0.033	0.1
is_practice_test_course	-0.033	-0.066	-0.016	-0.024	-0.19	-0.017	-0.076	-0.03	-0.052	-0.024	0.009	-0.044	-0.032	-0.033	1	-0.0004
relevancy_score	0.23	-0.028	0.057	0.091	0.44	0.14	-0.13	0.013	-0.036	0.017	-0.072	0.15	0.18	0.1	-0.0004	1

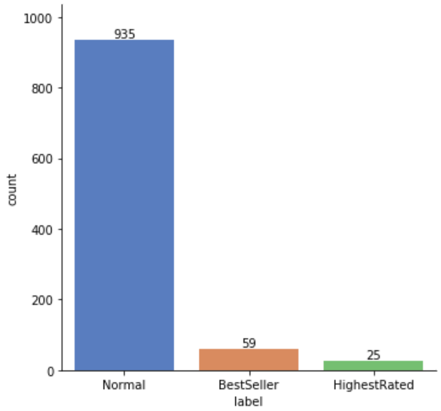
3. Modeling Results

3.1 VIF result for lasso regression

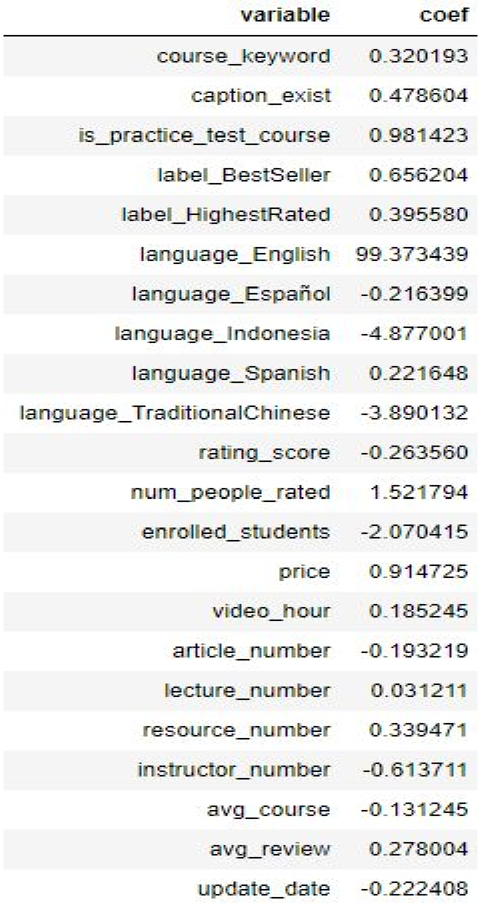


	VIF Factor	features
1	1.094675	course_keyword
2	1.621560	caption_exist
3	1.067154	is_practice_test_course
4	1.132026	label_BestSeller
5	1.040200	label_HighestRated
6	2.251923	language_English
7	1.325196	language_Español
8	1.061088	language_Indonesia
9	1.437134	language_Spanish
10	1.117248	language_TraditionalChinese
11	1.141762	rating_score
12	4.509009	num_people_rated
13	4.125778	enrolled_students
14	1.326362	price
15	2.971803	video_hour
16	1.941118	article_number
17	3.810176	lecture_number
18	1.447838	resource_number
19	1.354985	instructor_number
20	1.400342	avg_course
21	2.019773	avg_review
22	1.134429	update_date

3.2 Distribution of “Label” values

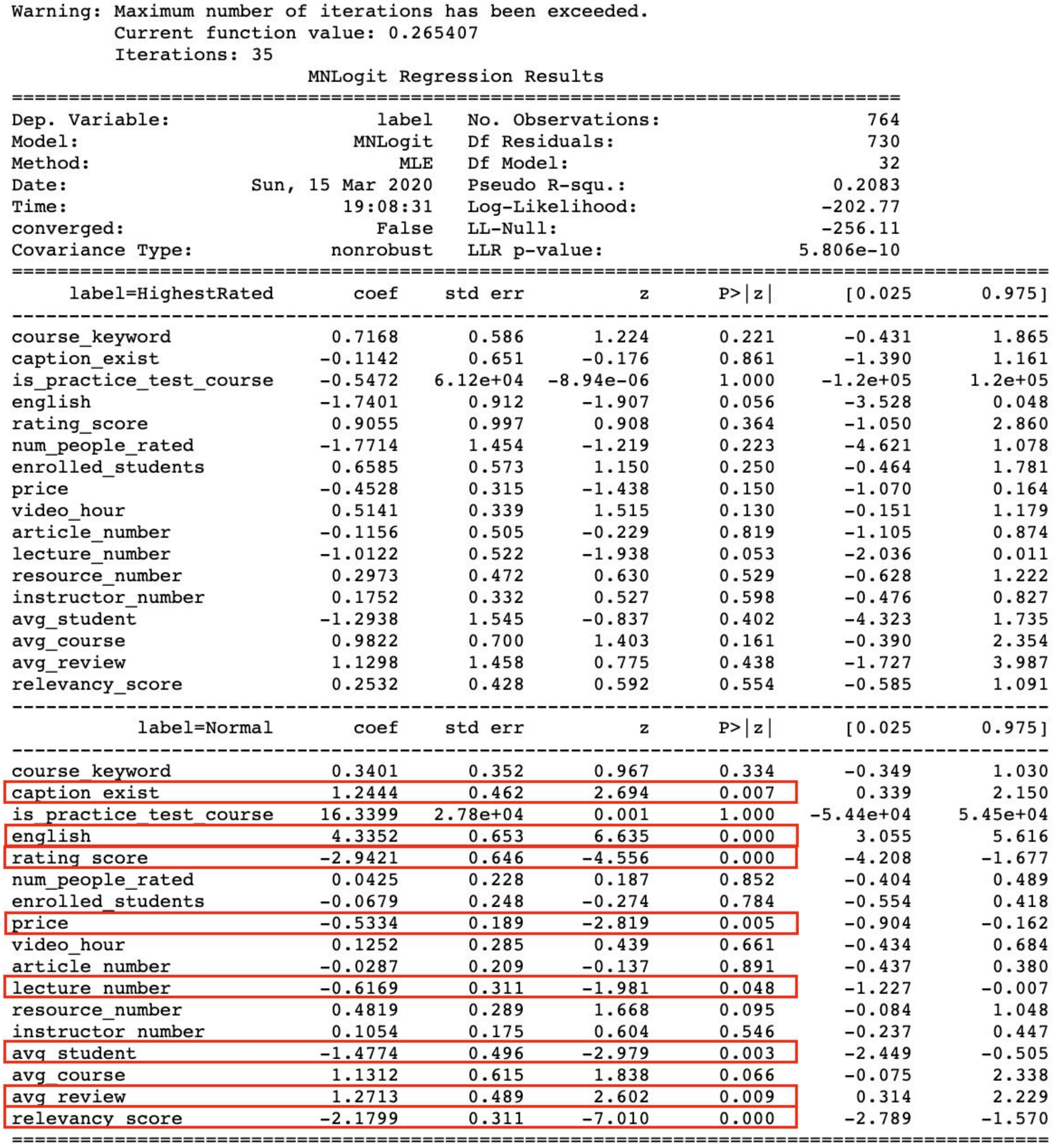


3.3 Variable coefficients from lasso regression model



variable	coef
course_keyword	0.320193
caption_exist	0.478604
is_practice_test_course	0.981423
label_BestSeller	0.656204
label_HighestRated	0.395580
language_English	99.373439
language_Español	-0.216399
language_Indonesia	-4.877001
language_Spanish	0.221648
language_TraditionalChinese	-3.890132
rating_score	-0.263560
num_people_rated	1.521794
enrolled_students	-2.070415
price	0.914725
video_hour	0.185245
article_number	-0.193219
lecture_number	0.031211
resource_number	0.339471
instructor_number	-0.613711
avg_course	-0.131245
avg_review	0.278004
update_date	-0.222408

3.4 Significant variables in Multinomial Logistic Regression



```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.265407
         Iterations: 35
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                  label   No. Observations:                  764
Model:                        MNLogit   Df Residuals:                      730
Method:                           MLE   Df Model:                           32
Date:                Sun, 15 Mar 2020   Pseudo R-squ.:                  0.2083
Time:                        19:08:31   Log-Likelihood:                -202.77
converged:                      False   LL-Null:                       -256.11
Covariance Type:            nonrobust   LLR p-value:                 5.806e-10
===========================================================================================
     label=HighestRated       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
course_keyword              0.7168      0.586      1.224      0.221      -0.431       1.865
caption_exist              -0.1142      0.651     -0.176      0.861      -1.390       1.161
is_practice_test_course    -0.5472   6.12e+04  -8.94e-06      1.000    -1.2e+05     1.2e+05
english                    -1.7401      0.912     -1.907      0.056      -3.528       0.048
rating_score                0.9055      0.997      0.908      0.364      -1.050       2.860
num_people_rated           -1.7714      1.454     -1.219      0.223      -4.621       1.078
enrolled_students           0.6585      0.573      1.150      0.250      -0.464       1.781
price                      -0.4528      0.315     -1.438      0.150      -1.070       0.164
video_hour                  0.5141      0.339      1.515      0.130      -0.151       1.179
article_number             -0.1156      0.505     -0.229      0.819      -1.105       0.874
lecture_number             -1.0122      0.522     -1.938      0.053      -2.036       0.011
resource_number             0.2973      0.472      0.630      0.529      -0.628       1.222
instructor_number           0.1752      0.332      0.527      0.598      -0.476       0.827
avg_student                -1.2938      1.545     -0.837      0.402      -4.323       1.735
avg_course                  0.9822      0.700      1.403      0.161      -0.390       2.354
avg_review                  1.1298      1.458      0.775      0.438      -1.727       3.987
relevancy_score             0.2532      0.428      0.592      0.554      -0.585       1.091
-------------------------------------------------------------------------------------------
           label=Normal       coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------
course_keyword              0.3401      0.352      0.967      0.334      -0.349       1.030
caption_exist               1.2444      0.462      2.694      0.007       0.339       2.150
is_practice_test_course    16.3399   2.78e+04      0.001      1.000   -5.44e+04    5.45e+04
english                     4.3352      0.653      6.635      0.000       3.055       5.616
rating_score               -2.9421      0.646     -4.556      0.000      -4.208      -1.677
num_people_rated            0.0425      0.228      0.187      0.852      -0.404       0.489
enrolled_students          -0.0679      0.248     -0.274      0.784      -0.554       0.418
price                      -0.5334      0.189     -2.819      0.005      -0.904      -0.162
video_hour                  0.1252      0.285      0.439      0.661      -0.434       0.684
article_number             -0.0287      0.209     -0.137      0.891      -0.437       0.380
lecture_number             -0.6169      0.311     -1.981      0.048      -1.227      -0.007
resource_number             0.4819      0.289      1.668      0.095      -0.084       1.048
instructor_number           0.1054      0.175      0.604      0.546      -0.237       0.447
avg_student                -1.4774      0.496     -2.979      0.003      -2.449      -0.505
avg_course                  1.1312      0.615      1.838      0.066      -0.075       2.338
avg_review                  1.2713      0.489      2.602      0.009       0.314       2.229
relevancy_score            -2.1799      0.311     -7.010      0.000      -2.789      -1.570
===========================================================================================
```

References

“Relevance Score for Facebook Ads” Facebook for Business, www.facebook.com/business/news/relevance-score

Erbay, Hamza. “Learn about Udemy Culture, Mission, and Careers: About Us.” Udemy About, Udemy About, about.udemy.com/

Williams, Richard. "Multinomial Logit Models - Overview" University of Notre Dame, https://www3.nd.edu/~rwilliam/stats3/Mlogit1.pdf