DSCI 632 Udemy Courses Analysis

Fernando Ramirez – Amit Nijsure – Evan Falkowski

March 2022

1 Abstract

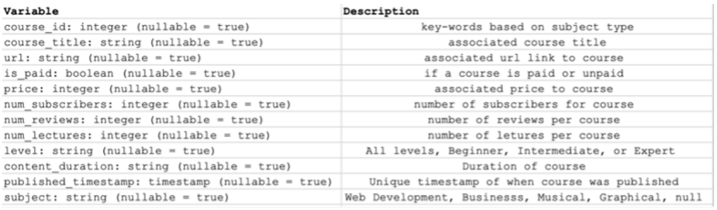
Udemy Courses Analysis implements Pyspark machine learning to predict if a course is paid or unpaid, as well as the predict the cost associated with paid coursing using data preprocessing and machine learning modeling. Udemy is one of the largest online learning platforms than span thousands of courses and hundreds of instructors. In this analysis we aim to use classification/regression modeling to draw insight to categories of classes offered and potentially the relationship of paid courses based on content duration, price, and number of subscriber reviews. Ultimately, we evaluate our models using Logistic Regression (LR), Decision Trees (DT), Linear SVM with summary characteristics Precision, Recall, and F1 scores, as well as RMSE MSE (regression). Paid courses were able to be accurately predicted – see discussion. However, regression model to predict prices of courses lacked accuracy, thus predictability – see discussion. We believe further engineering is needed or there is no obvious correlations between these features and target variable (courses).

2 Introduction

Udemy is one of the largest online learning MOOC (massive open online courses) platforms globally. Udemy's platform contains more than 15-thousand courses and 40-million students. The critical attributes in this dataset which we will use as predictions include the type of courses offered, number of reviews based on popularity of a course/courses and if a course is paid or unpaid. We aim to answer the following questions within the Udemy dataset using prediction models of regression and classification (1): predict the number of subscribers based on given courses (2): number of reviews given course features (3): revenue of each subject-type (4) price cost of courses (5) level of courses based on paid or unpaid filtering. Future work of this Project would involve predicting if a course a paid or unpaid, as well as, forecasting new course to maximize Udemy revenue.

3 Udemy Data Set

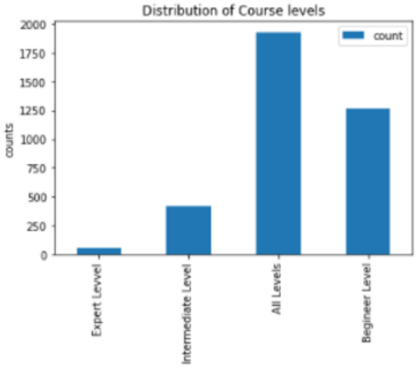
The dataset of interest is a free public csv file from Kaggle containing 3678 unique values and 12 columns over 4-unique subjects (Business Finance, Graphic Design, Musical Instrument, and Web Design). We provide the link to the dataset in our GitHub repository. Key features include is paid, price, of_review, content duration's, and level. As part of the EDA we explore what types of content are most successful to Udemy from the lens of the business/learning platform Udemy, as well as, the end user instructors (content creators) and students (consumers) alike.

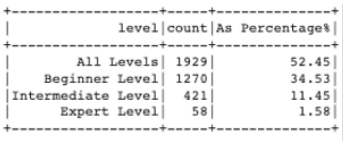


Variable	Description
course_id: integer (nullable = true)	key-words based on subject type
course_title: string (nullable = true)	associated course title
url: string (nullable = true)	associated url link to course
is_paid: boolean (nullable = true)	if a course is paid or unpaid
price: integer (nullable = true)	associated price to course
num_subscribers: integer (nullable = true)	number of subscribers for course
num_reviews: integer (nullable = true)	number of reviews per course
num_lectures: integer (nullable = true)	number of letures per course
level: string (nullable = true)	All levels, Beginner, Intermediate, or Expert
content_duration: string (nullable = true)	Duration of course
published_timestamp: timestamp (nullable = true)	Unique timestamp of when course was published
subject: string (nullable = true)	Web Development, Businesss, Musical, Graphical, null

4 Exploratory Data Analysis

The data set is compiled with a description of 3678 unique values and 12 columns over four unique subjects including (Business Finance, Graphic Design, Musical Instruments, and Web Design). For our analysis our team sought out the following questions in preparation in EDA and throughout training our models. (1): Use relevant features in association with Udemy courses to predict if a course is available is paid or unpaid (classification). Some key features for our analysis include, is paid, price, of_subscribers, of_reviews, num_lectures, level (beginner, Intermediate, or Expert), and content duration's. From our data set we have no missing values, and no further data cleaning is needed. However, course ID, and course title should not be included in the analysis as they might contributed to skewing data analysis based on unique object values. Furthermore, some categorical values are unbalanced and may be latter pursued in the analysis. Figure 1 shows the distribution of the most course level distribution and as a percentage of the data set (no missing values).

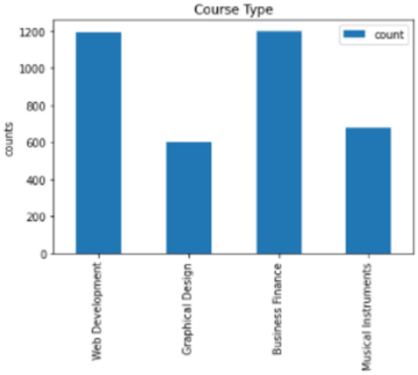


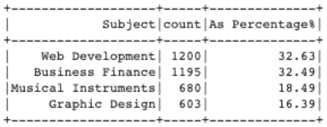


```
+------------------+-----+--------------+
|             level|count|As Percentage%|
+------------------+-----+--------------+
|        All Levels| 1929|         52.45|
|    Beginner Level| 1270|         34.53|
|Intermediate Level|  421|         11.45|
|      Expert Level|   58|          1.58|
+------------------+-----+--------------+
```

Figure 1

Figure 2 shows the distribution of Course Type distribution and as percentage of the data set. All levels lead the Udemy skill level which accounts for 52.45% of the data set. Both Web development and Business Finance lead the course type with 32.63% and 32.49% of the data set.





```
+-------------------+-----+--------------+
|            Subject|count|As Percentage%|
+-------------------+-----+--------------+
|    Web Development| 1200|         32.63|
|   Business Finance| 1195|         32.49|
|Musical Instruments|  680|         18.49|
|     Graphic Design|  603|         16.39|
+-------------------+-----+--------------+
```

Figure 2

Finally, an interesting EDA characteristic is the relationship between price and number of courses at different price points. There is no relationship of bucket of prices of courses that fit all paid courses. One may assume as price rises, fewer people will be willing to pay, but that is not the case with Udemy classes. The prices of courses fluctuate for paid and unpaid courses.

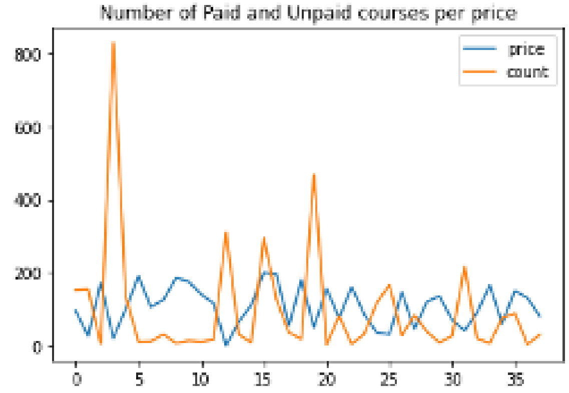


Figure 3

5 Methodology

5.1 Data Preprocessing

The machine learning tasks focused on are considered binary classification tasks, true or false. In preparation for ML tasks, all missing data (row) should be removed, string indexer, OHE (one hot encoding), vector assembler, data splitting should be applied prior to applying relevant models. Overall, the data set did not need additional preprocessing other than removing course_id, course_title and course_urls from analysis. This is due to unique values mapped to each course that does not advance the analysis and may skew / bias data predictions throughout the analysis caused by sparse data set splitting in OHE. String indexer was applied to both level and subject features. Vector assembler was applied to feature columns followed by general the pipeline builds and transformation of the Udemy data set (Figure 3). For feature vector column and target claim "Is paid" was used as the target variable. After applying all the following method, the train and test data set were split according to an 80% train, 20% test (validation) data sets. This analysis also explored how ML modeling can be used to predict the price of courses.

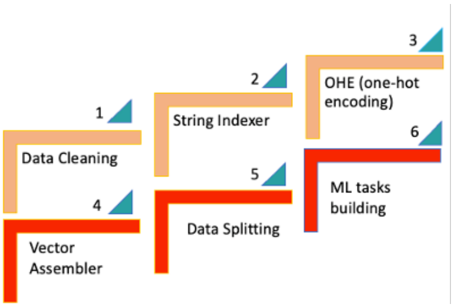


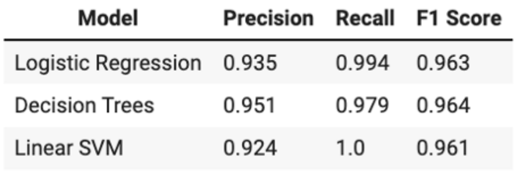
Figure 4

5.2 Machine Learning Modeling

For predicting if a course is paid or unpaid the machine algorithms used on both training and test data sets were LR (logistic regression) DT (decision tree), and SVM (support vector machine). After training models, models were evaluated based on performance metrics, including Precision, Recall, and F1 scores. All performance metrics were compared to the end-to-end machine learning pipeline build. For predicting the price of a courses Linear Regression, Decision Tree Regressor, and Random Forest Regressor was applied, however target variable switched to price.

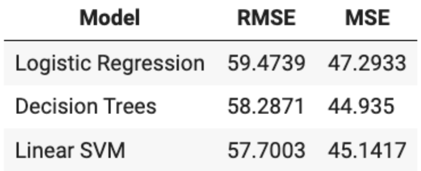
6 Results and Discussion

All models showed similar Precision, Recall, F1 score and Accuracy. Logistic Regression = 94%, Decision Tress = 95%, and Linear SVM = 93%. While Decision Trees scored the highest all models are comparable and the fitting of the model is above passing, given the typical threshold of greater than or equal to 80%. While all models have satisfactory performance in predicting whether a Udemy course is free based on given features, Decision Tree Classifier showed the highest Precision, Recall and F1 score.Logistic Regression, Decision Trees, and Linear SVM produced similar low RMSE at 59. 5, 58.3, and 57.7, respectively. Thus, indicating a good fit in reference to predicting course scores. On the other hand, the MSE was highest for logistic regression, but still comparable to all other models (Decision Trees and Linear SVM) Table 3.



Model	Precision	Recall	F1 Score
Logistic Regression	0.935	0.994	0.963
Decision Trees	0.951	0.979	0.964
Linear SVM	0.924	1.0	0.961

Table 2



Model	RMSE	MSE
Logistic Regression	59.4739	47.2933
Decision Trees	58.2871	44.935
Linear SVM	57.7003	45.1417

Table 3

7 Conclusion & Future Work

For predicting whether a course is paid or free given certain features it is possible to build classification tasks to forecast if a Udemy courses if free or paid based on given features and models described above. The Decision Tree should the highest accuracy in free or unpaid, while logistic Regression should the highest result in specific price prediction.

Other future work would include determination of optimal pricing for revenue maximization. This would include investigating the number of student willing to pay the highest amount per course or subject-matter. This approach would involve additional prediction on price range based on intrinsic parameters, however based on previous EDA and RMSE predictions Udemy and instructors should increase prices of courses to maximize revenue. The data set is either too small or there is no relationship between set price, content duration and maximum revenue.

Again, the model performance can be improved by collecting more data since the data set we used only contained 12 attributes for 3,000 + courses spanning four subjects. A more advanced analysis can be applied if more features were included.