Udemy Dataset ML Model

1st Vishv Jetani
Computer Science and Engineering
Institute of Technology,Nirma University
Ahmedabad,India
jetanivishv@gmail.com

2nd Jatin Undhad
Computer Science and Engineering
Institute of Technology,Nirma University
Ahmedabad,India
jatinundhad33@gmail.com

3rd Urvik Jada
Computer Science and Engineering
Institute of Technology,Nirma University
Ahmedabad,India
jadaurvik@gmail.com

4th Jwal Shah
Computer Science and Engineering
Institute of Technology,Nirma University
Ahmedabad,India
Shahjwal123@gmail.com

***Abstract*—This document is based on the Udemy dataset Predictor application which is built in python language using various machine learning Algorithms and deploy the model in a streamlit app.**

Application major functionality consists on recommender system and classification.Whenever user choose one particular course for study from the udemy course and he want to find what are the other recommended course which is similar to that course which is chosen.along with that there are several courses under udemy if user want to find out subject of that particular course then it also can be find by this application.Particular Course is free or not that also can be classify this model.

***Index Terms*—component, formatting, style, styling, insert**

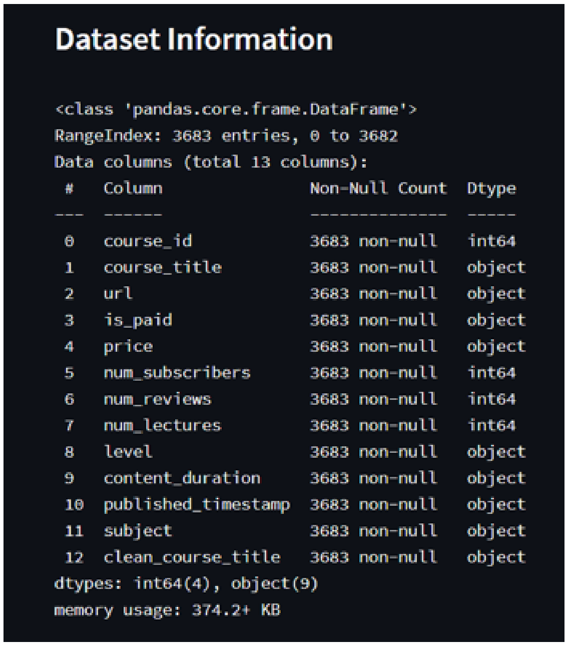
I. Introduction

The concepts which are used for build this application is cosine-similarity matrix,k-nearest neighbour,logistic regression.each of these is discussed later.The cosine similarity concept is used in finding the recommended course For given particular course.Knn is used for finding that whether this particular course is free or not and logistic regression is for classify the course in particular subject.Recommended system is very much used in several real-life application.

for Example:Whenever we buy a particular product from amazon/flipkart,it recommends that this product also should be bought.the inner mechanism that is used in such type of application is called Recommendation system.

II. Dataset Information

The dataset consists of details of 3683 courses.Courses are from the subject of Business finance,web development,musical instruments and graphic design.The course than udemy is providing is either free or paid.There are several other attributes like number of lectures,content duration,number of subscribers which take that course,number of people who review the course,the year in which the course was published..The level is one of the attribute which states that course is beginner level or intermediate level or expert level or applicable to all the levels. Total 13 columns present in the dataset,some of this is not used in the application like course-url and course-id.data is very much cleaned so there are no such data cleaning method is used but some of the pre-processing is done according the requirement,



Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3683 entries, 0 to 3682
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   course_id           3683 non-null   int64 
 1   course_title        3683 non-null   object
 2   url                 3683 non-null   object
 3   is_paid             3683 non-null   object
 4   price               3683 non-null   object
 5   num_subscribers     3683 non-null   int64 
 6   num_reviews         3683 non-null   int64 
 7   num_lectures        3683 non-null   int64 
 8   level               3683 non-null   object
 9   content_duration    3683 non-null   object
 10  published_timestamp 3683 non-null   object
 11  subject             3683 non-null   object
 12  clean_course_title  3683 non-null   object
dtypes: int64(4), object(9)
memory usage: 374.2+ KB
```

Fig. 1. Dataset Information

III. Recommendation System

This model gives the recommendation of courses based on user choose any particular course.

A. libraries that are used

- pandas
- neattext
- sklearn in which countVectorizer and cosine-similarity is used.

B. Build Model

Recommended system is useful while suggesting particular course from the given course.The given course is the only attribute in which machine learning algorithm is applied.The course title entry consists of various words which also includes special characters and some of the commonly used words like how to,In etc.This words can be exist in multiple course title and it increase the similarity between two courses but there are no actual contribution of this words.so in order to achieve greater efficiency,it is very necessary to remove this words.most commonly used words are called stopwords.This stopwords and special characters can be removed by simply applying methods of neattext library which removes the stopwords and special characters from the all course titles.

1) countVectorizer: whenever we want to build any machine learning model,it is necessary that all the data that want to use in numeric form only.if it is not then using such methods,this non numeric data must be converted into numeric data.In this case,it is necessary to convert course-title field to numeric data.this can be done using countvectorizer method of sklearn then course title is fit into countvector.In this method,all the value of course title is taken and all the distinct words that exist in the course-title field in dataset is work as a features and after that any particular course-title whatever word it has,all word in feature set is marked as 1 and rest other words are marked as 0.by default it becomes sparse matrix but sparse matrix ignores the zero values so it is converted into dense matrix.

2) cosine Similarity: cosine similarity is a term that is used to measure the similarit between the two vectors.it is a angle between two vectors.The formula for the same is described below;

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Fig. 2. cosine similarity formula

After the Countvector step is completed.this count vector is passed into the cosine similarity function.Cosine similarity matrix shows the similarity between all course-titles.After analyzing this cosine-similarity is very easy to identify which course is similar to the given course and after this similar courses will recommend to the user.

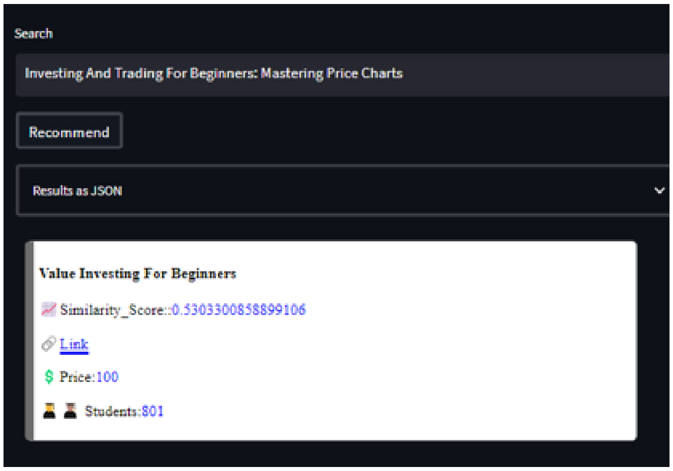
After completiton of this two step model is built.

C. Process

For testing recommendation system,the flow is user enter one particular course from the database and model will find similarity between entered course and all the other course from the dataset from the similarity matrix.This given results are sorted in descending order.According to User needs or requirements Number of Courses will be shown to user.

In such cases where user just enter keyword of any particular course so in that case cosine-similarity is not be count in this application.For that exceptional case the keyword that exist in any of the course is displayed to the user.

D. output



Investing And Trading For Beginners: Mastering Price Charts

Value Investing For Beginners

Similarity_Score::0.5303300858899106

Link

$ Price:100

Students:801

Fig. 3. Course Recommendation

In the given image user enter one course related to investment and trading and the application display courses which are very similar to this courses.

IV. Subject Prediction

This model predicts the subject of the given course from the course name. This model used logistic linear regression approach. For this purpose, model is completely relying on Udemy dataset for training purpose. Udemy dataset contains four different classes to train the model. These classes are 1) Business Finance 2) Graphic Design 3) Musical Instruments 4) Web Development. So that it is multiclass classification problem.

A. Libraries that are used

- pandas
- neattext
- sklearn: linear model for logistic regression will be imported for the model training and prediction

B. Data Pre-processing

For this classification problem we need only course names and their corresponding class name for prediction purposes. So that form given data we need to drop out the other features which are not required.

From the course names we found so many noisy words which are not related to corresponding class and not useful to predict any class so that it should be removed from the class names i.e., is, of, the, etc. for this purpose neattext library to clean those words from the course names.

Strings are not useful to train the classification problem, it should require to convert into numeric data so that, better classification can be done. Classes (output) converted into corresponding numeric values. For feature transformation fit-transform method has been used which convert the data in numeric format for each words found into training dataset.

Then this model will train through this data and classification will be happened using this fitted model.

C. Build Model

For building the model, logisticRegression() method of sklearn.linearmodel is used. Model will fit with sigmoid activation function which output through regression output.

Then this model to predict class in newly unknown data using the predict method.

for classifying the course to it's corresponding class, maths used (equation) is given below.

$$\phi(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(w\times x+b)}}$$

Fig. 4. logistic regression classification: activation function

D. Model Evaluation

1) Confusion Matrix: For testing the fitted model confusion matrix will be used in which it represents how much data is well classified and how much data is not classified well.

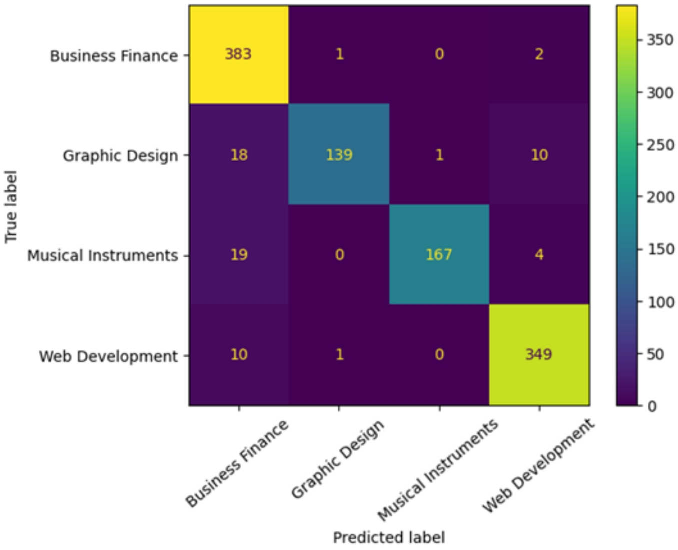
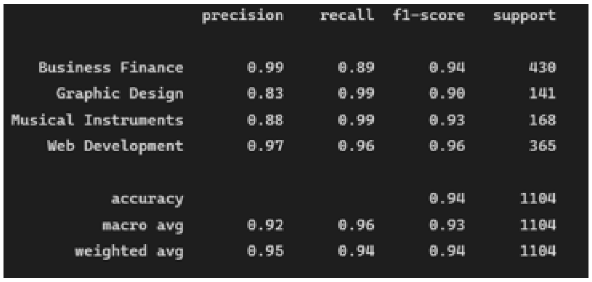


Fig. 5. logistic regression classification: confusion matrix

2) Classification Report: Using classification report one can able to understand how much the model is capable to predict the right class from the fitted model for the unknown data. classification report for the fitted model is given below. in which accuracy, precision, recall, f1-score measures will be given for measuring the accuracy.

E. output

Finally the output will be predicted by the model is given below.

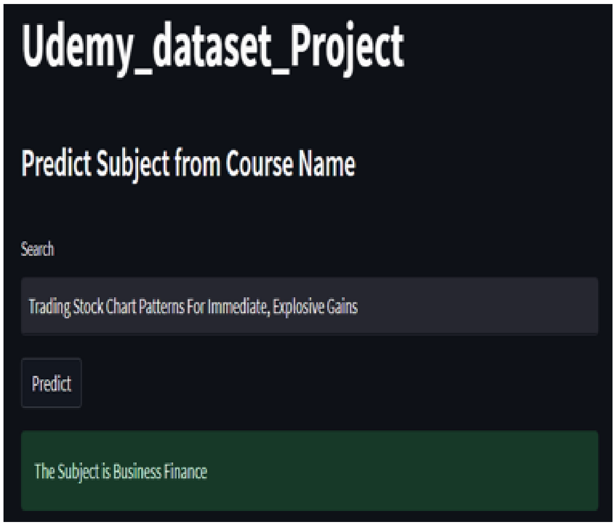


```
                     precision    recall  f1-score   support

   Business Finance       0.99      0.89      0.94       430
     Graphic Design       0.83      0.99      0.90       141
Musical Instruments       0.88      0.99      0.93       168
    Web Development       0.97      0.96      0.96       365

           accuracy                           0.94      1104
          macro avg       0.92      0.96      0.93      1104
       weighted avg       0.95      0.94      0.94      1104
```

Fig. 6. logistic regression classification: classification report



Trading Stock Chart Patterns For Immediate, Explosive Gains

The Subject is Business Finance

Fig. 7. logistic regression classification: Subject Prediction

V. Free-paid Prediction

This model tells from the entered course details whether this course is paid or free.Whether Course is paid or free is depend on subject,difficulty of the course,content-duration of course,number of lectures,number of student that take that course,number of reviews for the course and year in which course was published.

From these features of the data,the model will be trained and it is used for prediction of free or paid for unseen data.

A. Libraries that are used

- pandas
- sklearn: For build model and evaluate model

B. Build Model

Like before in this model also,model can not working with categorical data so it is must to convert categorical features to numeric features.content-duration in string format so for this conversion for every content-duration,string part will be discarded and based on that string part is in hour form or minute form,the entire content-duration is converted into hour form. Year will be extracted from published-timestamp feature.for the level,subject and year feature there are few categories so based on particular category,such number is assigned so by this way this features are converted into numeric form.Target variable is in Boolean form so it also be converted into 0/1 numeric form.

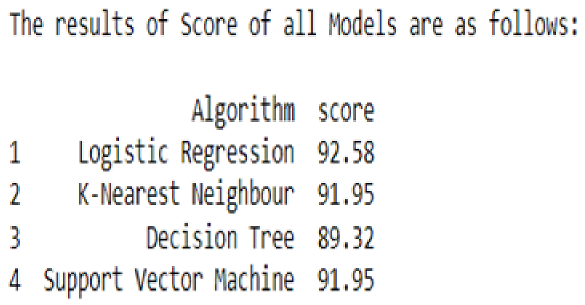
Now the train data is ready,y is a target variable and x is a features.

Now for which model is deployed for this application.several algorithms are used first and comparing the accuracy of the generated model on test data then analyses which is suitable for this application.

Based on these data,models will be trained.

Approaches that are used 1.logistic regression 2.K-Nearest Neighbour 3.Decision tree 4.Support vector machines

The Score of each of the model is described below:



```
The results of Score of all Models are as follows:

              Algorithm  score
1   Logistic Regression  92.58
2   K-Nearest Neighbour  91.95
3         Decision Tree  89.32
4 Support Vector Machine  91.95
```

Fig. 8. Score of Each Model

Although there is no significant difference in accuracy between this model.In this application KNN is used.

C. Testing and Evaluation

The model here is used is k-nearest-neighbour.The Accuracy of the Model is 91.95 Percentage.

1) Confusion Matrix: Confusion matrix is helpful for the evaluating the Model;

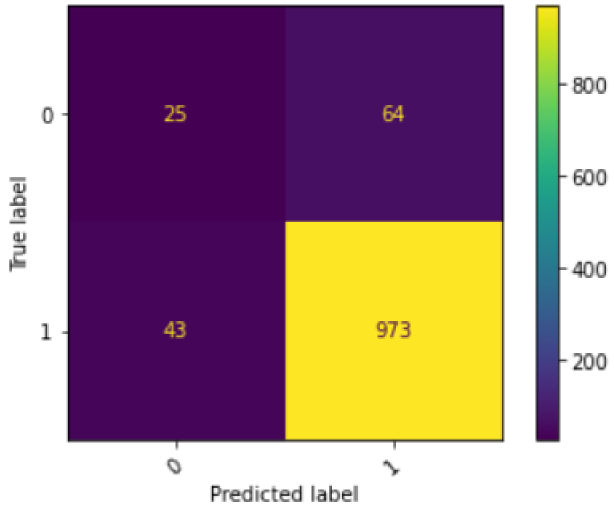
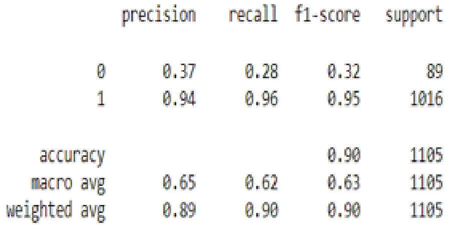


Fig. 9. Confusion Matrix for KNN

2) Classification Report:



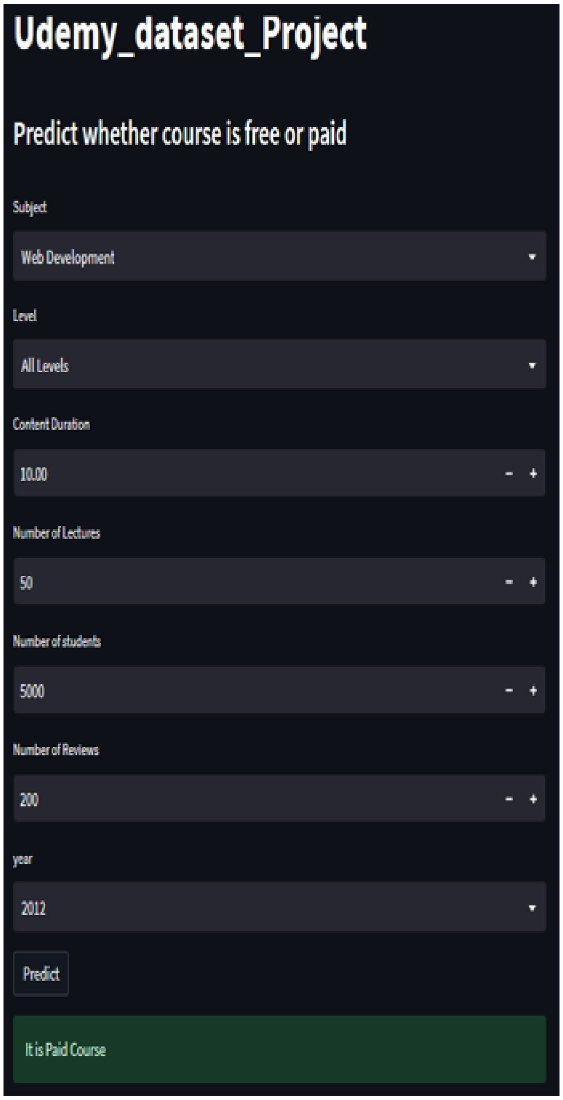
```
              precision    recall  f1-score   support

           0       0.37      0.28      0.32        89
           1       0.94      0.96      0.95      1016

    accuracy                           0.90      1105
   macro avg       0.65      0.62      0.63      1105
weighted avg       0.89      0.90      0.90      1105
```

Fig. 10. Confusion Matrix for KNN

VI. Output



Udemy_dataset_Project

Predict whether course is free or paid

Subject: Web Development

Level: All Levels

Content Duration: 10.00

Number of Lectures: 50

Number of students: 5000

Number of Reviews: 200

year: 2012

It is Paid Course

Fig. 11. Classification Report for KNN

VII. Integration and Deployment Of Model

The Integration of all model done and deployed it using streamlit app. The URL for the Deployed link is mentioned here : https://jetanivishv-machinelearningproject-app-7aniep.streamlit.app/

References

[1] Basics of CountVectorizer by Pratyakash jain

[2] Understanding Cosine Similarity And Its Application by Richmond Alake

[3] TF-IDF Simplified by Luthfi Ramadhan

[4] Bag-of-words vs TFIDF vectorization –A Hands-on Tutorial on Analytics vidhyalay

[5] streamlit Documentation

[6] Kaggle dataset sample for paid-free course

[7] sklearn documentation