# **Project 2**

**Name: Caleb Thian Jia Le 田家樂**

**Student ID: NN6114035**

## About Data

The data created is about online courses. 10000 online courses are generated with 11 features, which are subject, subscribers, free, fee, reviews, avg. reviews, level, lectures, duration, published date and subtitles. The features’ definition and the way to generate them is listed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Definition | Way to generate | Graph |
| subject | The course’s involved area, expected only 1 value | Random choose among {Management, Photography, Social Sciences, Science, Personal Development, Design, Information Technology, Music} |  |
| subscribers | The number of subscribers of the course. | Sampling from a normal distribution N(7500,2500), each sample must no less than 0 |  |
| free | If the course is a free course, set as 1. | Bernoulli Trial with p=0.25 |  |
| fee | The fee of the course. | 0 if free, the others sample from exponential distribution (Exp(0.01)+10)\*10 |  |
| reviews | The number of reviews. | Sampling from uniform distribution with max = subscribers |  |
| avg. reviews | The average score from reviews. | Sampling from uniform distribution 0.5n, 0⩽ n ⩽10 and n is integer. (Note that the plot merges 4.5 and 5.0 into a bar, it is still a uniform distribution) |  |
| level | The level of the course. | Random choose among {Beginner, Intermediate, Expert} |  |
| lectures | The number of lectures provided. | Sampling integer > 0 from normal distribution N(10,5). |  |
| duration | The total duration of the lectures in minutes | Sampling average duration > 0 from normal distribution N(30,10) and multiply with number of lectures |  |
| published date | The published date of the course | Generate a random date between 2002.01.01 and 2021.12.31 |  |
| subtitles | If the course has subtitles, set as 1. | Bernoulli Trial with p=0.5 |  |

This data will like to recommend several courses according to these features. If a course will be recommended, at least 1 of the following rule should be satisfied:

1. subscriber > 12000
2. review >= 0.8\*subscriber and avg. reviews >= 4.5
3. level = beginner and fee <= 100
4. level = intermediate and fee <= 250
5. level = expert and fee <= 500

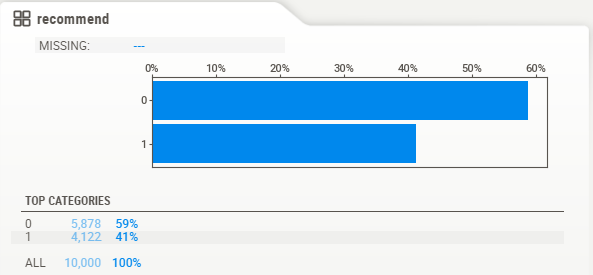
By these rules, the ratio of course that be recommended and not recommended is about 3:2. The plot below shows its distribution, 1 indicates the course is recommended.

Figure 1: Distribution of the course

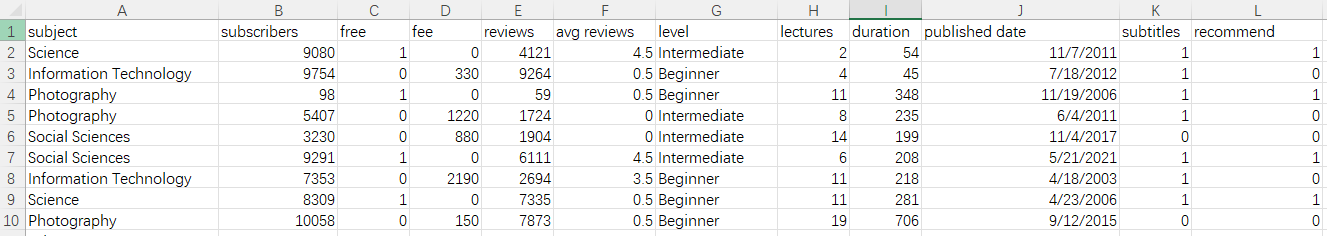


Figure 2: Example data generated

The details of the data can be viewed in EDA.html. The data is split into 4:1, i.e., 8000 for training and 2000 for testing.

## Data Pre-processing

Several data pre-processes are done on the dataset.

### Label Encoding

Label encoding on the categorical features, which are subject and level.

### Conversion

Convert the published date into timestamp so that classification can be run on it. Reviews also be converted into ratio but not integer. The reason to do this conversion will be discussed in the session discussing decision tree.

### Normalization

Normalization is done on numerical features, even though it doesn’t affect decision tree, it will improve the result of KNN and Naïve Bayes classifier.

## Result

Decision tree, KNN, and Naïve Bayes classifiers are tried to classify these courses. Discussion for each classifier will be made at each session and comparison will be made at the end of each session.

Decision