Online Education Data Analysis and Prediction

Abstract

Due to the arrival of the epidemic, online education is growing rapidly and gradually becoming one of the main teaching methods. **Kdd-cup 2015 competition dataset** comes from Tsinghua's XueTangX online course platform, which details the learning data of about 120,000 students and about 8 million learning log records. This paper uses **data visualization** to observe user characteristics and analyze the main characteristics of students who skip classes; uses **decision trees** and **Adaboost boosting methods** to predict whether students **skip classes** and observe the contribution of each attribute; and finally, uses qualitative and quantitative analysis to determine the characteristics of the skipping group.

Keywords: data visualization, decision trees, Adaboost boosting method, skipping

1 Review of the original kdd-cup competition information

Competition: Predict dropout rate on MOOC platforms

Description: Students' high dropout rate on MOOC platforms has been heavily criticized, and predicting their likelihood of dropout would be useful for maintaining and encouraging students' learning activities. Therefore, in KDD Cup 2015, we will predict dropout on XueTangX, one of the largest MOOC platforms in China.

The competition participants need to predict whether a user will drop a course within next 10 days based on his or her prior activities. If a user C leaves no records for course C in the log during the next 10 days, we define it as dropout from course C For more details about log, please refer to the Data Descriptions.

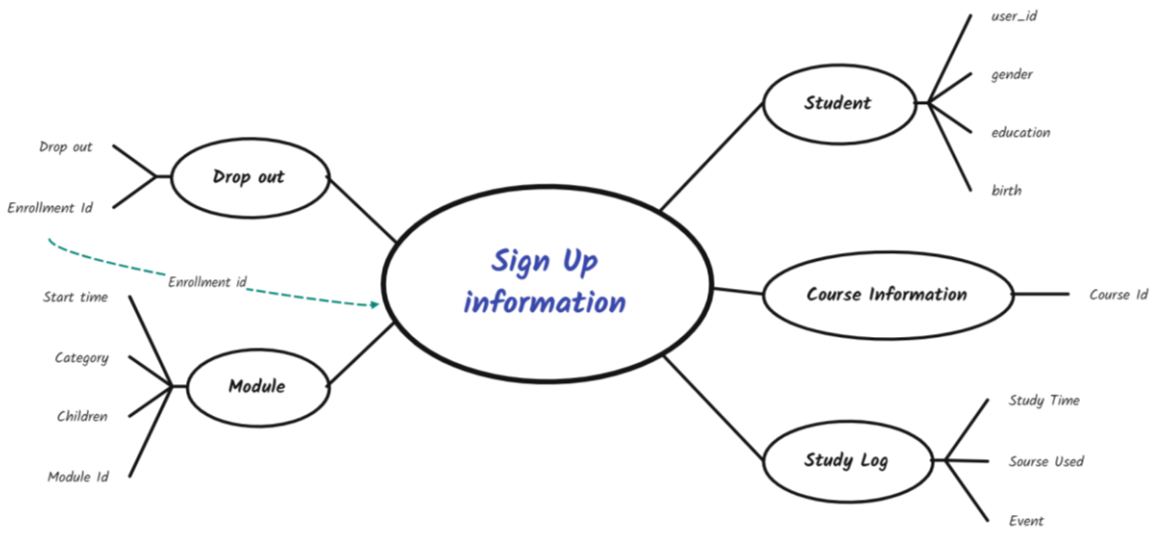
2 Introduction to the Dataset

2.1 Overview of the dataset

This dataset was obtained from the 2015 kdd-cup Data Mining competition. The

experimental dataset size is a total of 10 files, respectively user_info.csv, log_train.csv, log_test.csv, course_info.csv, data.csv, object.csv, enrollment_test.csv, enrollment_train.csv, truth_test.csv, truth_train.csv.

2.2 Relationship of the dataset



2.3 Data set meaning

2.3.1 user_info.csv

This dataset contains 9627149 users. Each row represents the corresponding information for a user and each column is an attribute for that user. The field descriptions are as follows:

(1) **user_id**: This field is the id of the user in the MOOC online education platform.

(2) **gender**: this field is the user's gender.

(3) **education**: this field is the user's education level.

(4) **birth**: this field is the year of birth of the user.

2.3.2 log_train.csv & log_test.csv

This dataset contains 8,157,277 entries of 120,542 users taking the course. Each row represents information about a user's participation in a specific course of study. The fields are described as follows:

(1) **enrolment_id**: the enrolment id of the user who selected a particular course.

(2) **time**: the duration of the user's course.

(3) **source**: the resource (event) operated on, for example: browser and server.

(4) **event**: indicates the type of event the user is performing, there are 7 types of events: problem, video, access, wiki, discussion, navigate, page _close.

(5) **object**: the object to read or browse, course module number.

2.3.3 course_info.csv

This dataset contains 6411 courses. Each row represents specific information about a course and each column is an attribute of the course. The fields are described as follows:

(1) **id**: the user id.

(2) **course_id**: is the course id.

(3) **start**: is the opening time of the module to students.

(4) **end**: is the end time of the module for students.

(5) **course_type**: is the type number of the course.

(6) **category**: the category of the course, e.g. computer, engineering, etc.

2.3.4 data.csv

This dataset contains the start and end times of the course. Where each row is the time information of the course and each column is the relevant time attribute of the course:

(1) **course_id**: the id of the course.

(2) **from**: is the start time of the course.

(3) **to**: is the end time of the course.

2.3.5 object.csv

This dataset contains information on 27,250 modules in courses. Each row in this file describes a module in a course, including its category, its sub-modules, and the time of publication. These modules represent different parts of the course, such as chapters, sections, online video material, exercises, etc. Modules are organised into a tree structure, with each course containing several chapters; each chapter contains several sections, and each section contains several objects (videos, exercises, etc.):

(1) **course_id**: course.

(2) **module_id**: module number.

(3) **category**: type of module.

(4) **children**: sub-modules of the module.

(5) **start**: the time when the module is available to students.

2.3.6 enrollment_train.csv & enrollment_test.csv

This dataset contains 120,542 records of a user attending a course. Each row in this file describes the record corresponding to a user's participation in a course, and each column is an attribute of that record:

(1) **enrollment_id**: registration number.

(2) **username**: student number.

(3) **course_id**: course number.

2.3.7 true_train.csv & true_test.csv

This dataset contains 120,542 records of whether or not a user skipped a class. Each row in this file describes whether or not a user has missed a class. Each column is an attribute of that record:

(1) **id**: user id.

(2) **tag**: corresponds to whether or not the user skipped class, indicated by 0 or 1.

3 Object of experience

(1) **Statistics** of the overall user profile distribution, such as age, education distribution, to obtain the overall user profile of the platform.

(2) **Process the raw data** to obtain and **visualize the difference** between students who skip classes and those who do not. This is an intuitive way to show the relationship between whether or not a student skips a class and their usual learning behavior. This allows us to easily observe the characteristics of students who skip classes from the statistical graph of the data.

(3) **Train a classifier** that can make predictions about whether students skip school or not, and see if the **contribution** of each feature in the classifier is related to the statistical graph of the data.

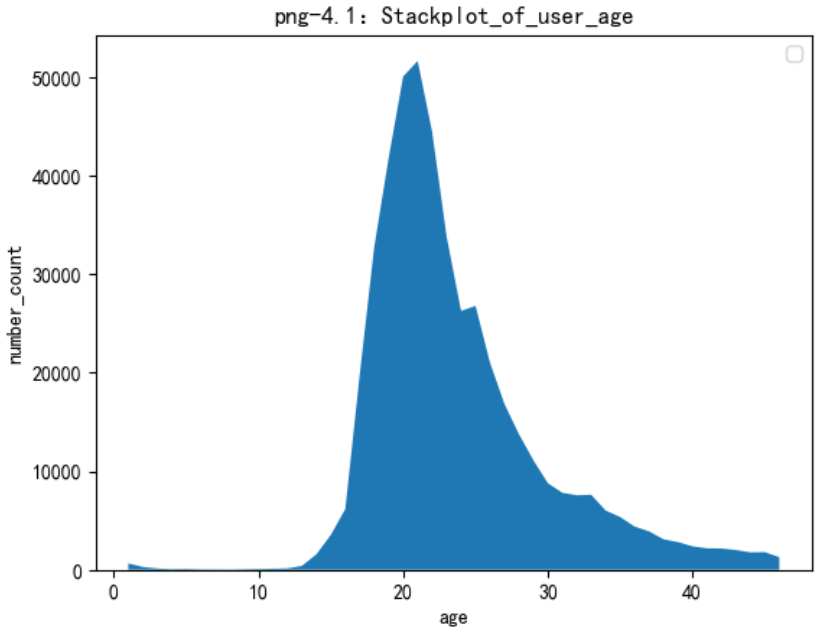
4 Distribution of users' features

This task is mainly done using **MySQL database**. After importing user_info.csv into the database, write **query statements** to obtain statistics and visualize them in python.

4.1 age

Query sql code: Because there are outliers and unreasonable values in the data (for example, some people were born after 2015, but the data was collected in 2015), records with age less than 0 and data larger than 3 times the standard deviation of the mean will be filtered out in the statistics.

```
select 2015-BIRTH age,count(BIRTH) count
from USER_INFO
where 2015-BIRTH>0 and 2015-BIRTH<(select avg(2015-BIRTH)+3*stddev(2015-BIRTH) from USER_INFO where 2015-BIRTH>0)
group by (BIRTH)
order by age
```

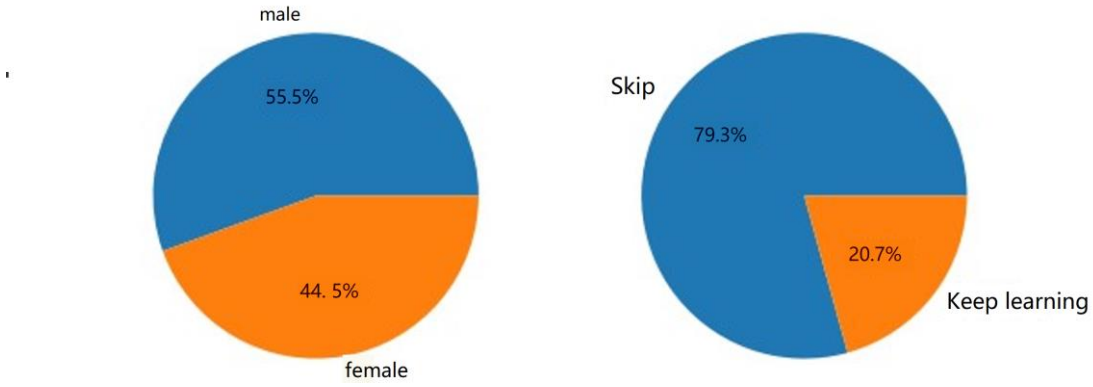


Analysis: According to Figure 4.1, it can be found that the user age is concentrated between **18 and 25 years old**. It can be guessed that the main users of online education are **undergraduate students and postgraduate students**, mainly to complete the teaching tasks assigned by teachers.

4.2 Gender and skipping labels

```
select GENDER,count(GENDER) from USER_INFO group by GENDER;
select LABEL,count(LABEL) count from TRUTH_TRAIN group by LABEL;
```

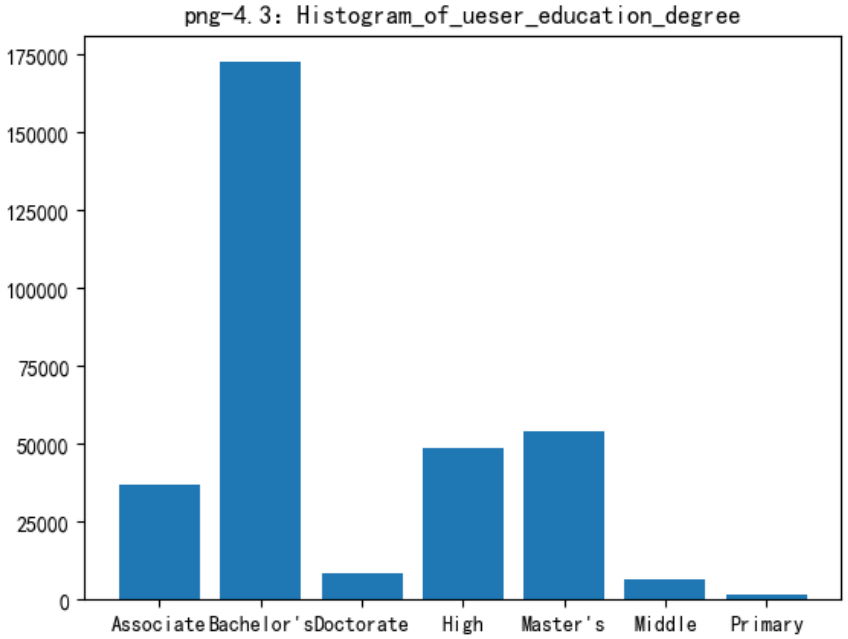
Figure 4.2: user gender and skip class label distribution



Analysis: The male student is slightly more than the female student, guess the reason that the male student is more than the female student overall. However, **the proportion of skipping classes is far greater than that of sticking to study**, which indicates that college students are more playful, and also represents that the data set is a seriously unbalanced data set.

4.3 Education background

```
select EDUCATION,count(EDUCATION) count from USER_INFO where EDUCATION is not null group by EDUCATION
```

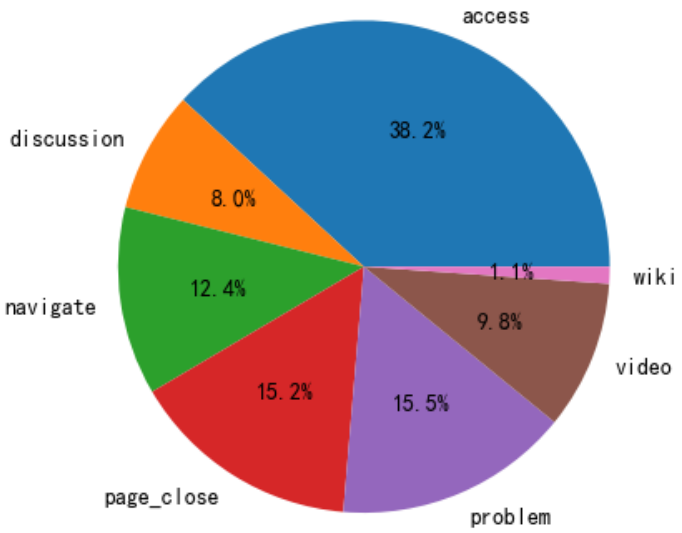


Analysis: **Undergraduates** account for the vast majority, which is also in line with the characteristics of more courses and heavy teaching tasks for undergraduates.

4.4 study events

```
select EVENT,count(EVENT) count from LOG_TRAIN group by EVENT
```

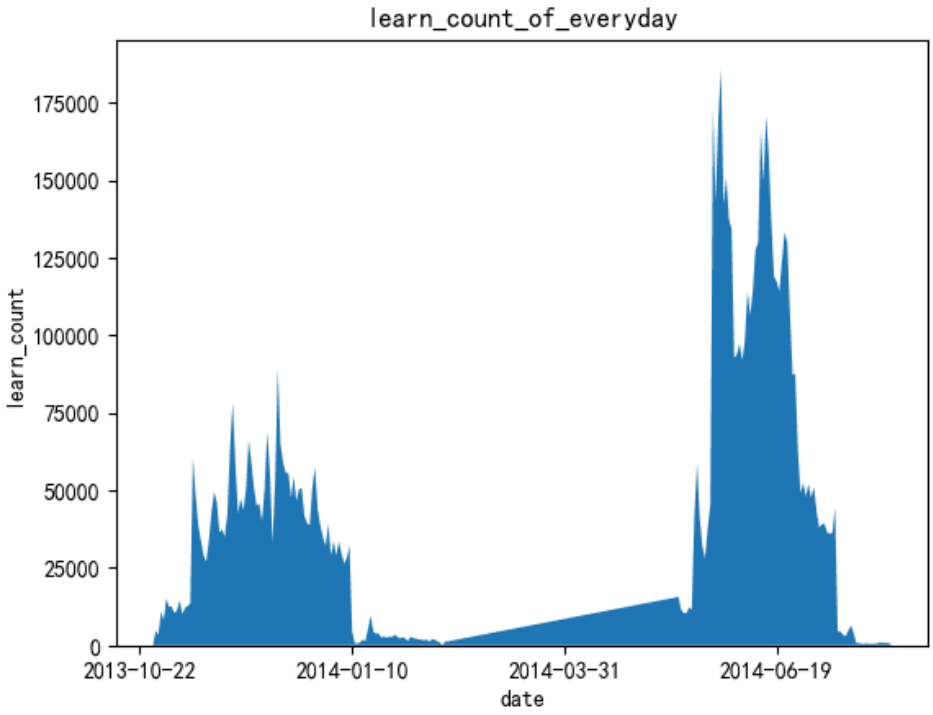
png-4.4: Distribution_Diagram_of_Events



Analysis: **access events account for the largest proportion**, while students mainly watch videos and do homework on online platforms, which reflects that students' learning attitude is not too serious, and they do not fully and effectively use online classes. At the same time, **the number of people who browse chapters and do homework is almost the same**. This article speculates that homework can only be entered by browsing chapters.

4.5 learning time

```
select COUNT(*) learn_count,substr(TIME,1,10) time from LOG_TRAIN group by
substr(TIME,1,10) order by substr(TIME,1,10)
```



Analysis: It can be seen from the chart that students' study time is mainly concentrated in **June to July** and **December to January**, while a relatively small number of students study from **August to October**, and it also shows that students' study time is usually **the final examination period**. It can be seen that most students study less in ordinary times, but study more seriously in the final.

5 Statistical modeling of skipping classes

In the data set, there are only 120,542 students with labeled individuals. Therefore, **true_test.csv** will be used for feature extraction in the following modeling. The only file that can extract features for students is **log_train.csv**. This file records every learning event of all students. It is a log file. This file will be modeled in subsequent modeling efforts. The main tool used is the MySQL database.

At present, the action data of each student can be extracted from log_train.csv as follows: **problem**: doing homework, **video**: watching video, **access**: reading other objects except video and homework of the course, **wiki**: reading Wikipedia of the course, **discussion**: Forum discussions, **navigate**: navigate the rest of the course, **page_close**: close the page.

Therefore, it is necessary to sort out the learning situation of each student from the log file.

5.1 Data Preprocessing

5.1.1 Grouping statistics

Separate statistics were made for each action to sort out the times of each student's corresponding actions and **separate tables** were built. For the 7 events contained in the log file, there were 7 tables for **one-to-one correspondence.** The specific statistical method takes statistical of event “problem” as an example, where the sql code is as follows:

```
create table problems(enrollment_id,problem) as select enrollment_id,count(event)
problem from log_train
where event='problem' group by(enrollment_id)
```

The table has two fields: **enrollment_id** corresponds to the student registration number, and **problem** the number of enrollments that the event triggered.

<Filter criteria>

	ENROLLMENT_ID	PROBLEM
1	1	87
2	3	138
3	4	6
4	5	170
5	6	2
6	7	94
7	9	6
8	12	0
9	13	150
10	14	12
11	16	64
12	18	17

Other events are the same.

5.1.2 Data integration

Using true_train.csv as the standard, left-join with other 7 tables and **create a new table**, sql code:

```
create table statistical_information as
select TRUTH_TRAIN.ENROLLMENT_ID
ENROLLMENT_ID,PROBLEM,VIDEO,ACCESS1,WIKI,DISCUSSION,PAGE_CLOSE,NAVIGATE,LABLE
```

```
from ((((((TRUTH_TRAIN left join PROBLEMS on TRUTH_TRAIN.ENROLLMENT_ID=PROBLEMS.ENROLLMENT_ID)
left join VIDEOS on TRUTH_TRAIN.ENROLLMENT_ID=VIDEOS.ENROLLMENT_ID)
left join ACCESSS on TRUTH_TRAIN.ENROLLMENT_ID=ACCESSS.ENROLLMENT_ID)
left join WIKIS on TRUTH_TRAIN.ENROLLMENT_ID=WIKIS.ENROLLMENT_ID)
left join DISCUSSIONS on TRUTH_TRAIN.ENROLLMENT_ID=DISCUSSIONS.ENROLLMENT_ID)
left join NAVIGATES on TRUTH_TRAIN.ENROLLMENT_ID=NAVIGATES.ENROLLMENT_ID)
left join PAGE_CLOSES on TRUTH_TRAIN.ENROLLMENT_ID=PAGE_CLOSES.ENROLLMENT_ID;
```

5.2 Data cleaning

After the integration was complete, a lot of **missing data** was found. This paper speculated that some students may have **never triggered the event** from the beginning to the end, resulting in **no learning record** of this item in the log. So **fill it with 0**.

Some abnormal data were also found in the **preliminary statistics** (for example, the number of times some accounts completed homework was as high as 800 times, while the course lasted only one year). Therefore, in this paper, according to **the principle of 3σ**, data less than (mean -3* standard deviation) or more than (mean +3* standard deviation) were regarded as abnormal data and modified to mean value.

```
con = create_engine('mysql + pymysql://hadoop:root@127.0.0.1:3366/LiuNian');
sql="select * from STATISTICAL_INFORMATION";
df=pd.read_sql(sql,con);
df["lable"]=df["lable"].astype("int")
cols=df.columns.tolist()
del cols[0];
del cols[-1];
for col in cols:
    df.loc[df[col]>df[col].mean()+3*df[col].std(),col]=df[col].mean();
```

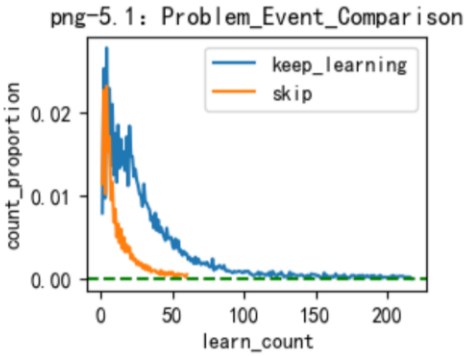
The final chart is shown below:

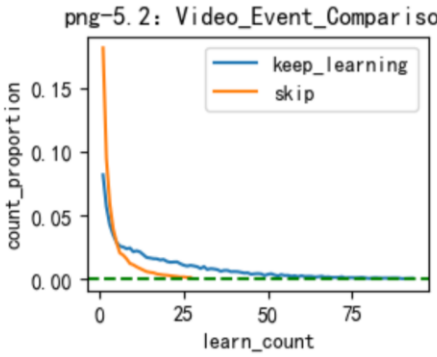
enrollment_id	problem	video	acces...	wiki	discus...	page_...	navig...	lable
2098	7.00000	8.00000	19.00000	1.00000	0.00000	13.00000	7.00000	1
2099	10.46249	6.61145	25.81831	2.00000	41.00000	10.26931	8.37309	0
2100	41.00000	19.00000	155.00000	6.00000	31.00000	10.26931	45.00000	0
2102	0.00000	0.00000	0.00000	1.00000	4.00000	0.00000	6.00000	1
2103	0.00000	0.00000	0.00000	3.00000	3.00000	0.00000	14.00000	1
2104	104.00000	6.61145	25.81831	5.00000	68.00000	10.26931	44.00000	0
2105	0.00000	3.00000	29.00000	2.00000	1.00000	3.00000	13.00000	1
2107	1.00000	14.00000	53.00000	1.00000	1.00000	21.00000	13.00000	0
2108	4.00000	14.00000	58.00000	4.00000	19.00000	30.00000	29.00000	0
2109	91.00000	6.61145	25.81831	2.00000	81.00000	10.26931	8.37309	0
2110	0.00000	1.00000	4.00000	0.00000	0.00000	2.00000	4.00000	1
2113	0.00000	0.00000	5.00000	2.00000	2.00000	3.00000	9.00000	1
2114	50.00000	6.61145	25.81831	2.00000	22.00000	10.26931	8.37309	0
2116	0.00000	13.00000	24.00000	6.00000	2.00000	14.00000	8.00000	1
2117	6.00000	5.00000	24.00000	3.00000	13.00000	2.00000	21.00000	0

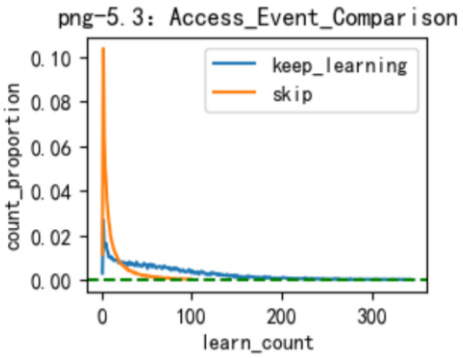
5.3 Visual analysis of skipping class or not

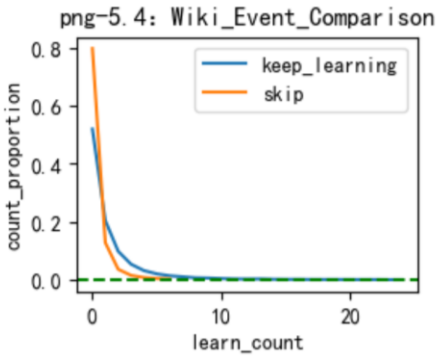
In order to better compare the difference of student groups between skipping class and not skipping class, this paper makes a **comparative analysis of the data visualization**. Because of the huge difference between skipping class and not skipping class, the **dimensional effect** should be eliminated. In this paper, the number of people corresponding to the times of each time is divided by the total number of people corresponding to the missing class label, which is converted into the frequency of events, and the **probability distribution diagram of seven events** in the case of skipping class and not skipping class is drawn. The sql code takes event “problem” as an example:

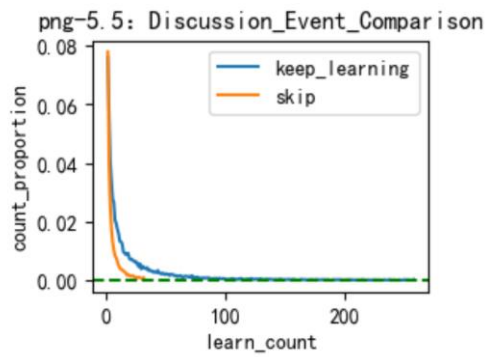
```
select PROBLEM,count(PROBLEM) count from statistical_information group by PROBLEM
```

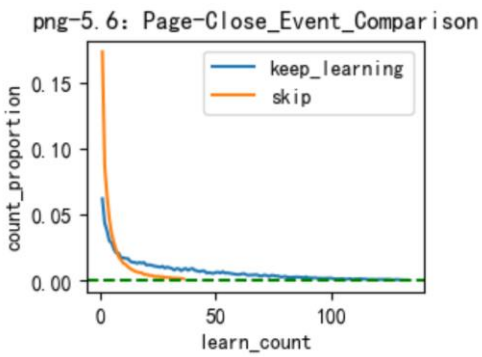


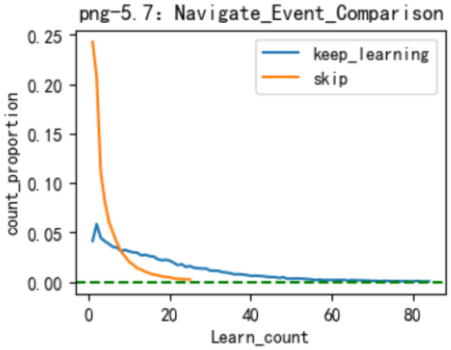












Analysis: From the above figure, it can be seen that there is **a big difference between Figures 5.1, 5.2, 5.3 and 5.6**. The two curves **in Figure 5.4 and 5.5 are similar**. But all of the graphs had one common feature: the **distribution** of frequency of skipping classes was more clustered in places with lower study times than the distribution of persistence, meaning that students who skipped classes learned less about events. This provides a strong support for our subsequent classifier modeling.

5.4 Constructing a classifier

The data mining part of this paper is completed based on **sklearn**.

5.4.1 The base classifier

Taking these 7 fields as the characteristics of students, this paper chooses decision tree as the basic classifier. Due to the **sufficient data volume**, there are about 120,000 samples. However, it can be seen from Figure 4.2 that there is a **serious class imbalance problem** in the data set, that is, the sample that skips class is about three times that of the sample that insists on learning. Therefore, the training set should be **resampled** to ensure class balance. This experiment adopts the **set-aside method**, and the test set is **30%** of the overall data set. Due to the serious imbalance of the class, the evaluation indexes of the experimental results include **accuracy, F1 score and G-mean** for comprehensive evaluation.

The experimental results are as follows:

```
DecisionTree:
accuracy: 0.8044410032353511
F1 score: 0.8782788296041308
G-Mean score: 0.8270386321882596
```

Contribution of each attribute:

0.09881	0.09897	0.44896	0.04068	0.08784	0.11117	0.11356

These 7 attributes correspond to the following events:

```
['problem', 'video', 'access1', 'wiki', 'discussion', 'page_close', 'navigate']
```

Therefore, **the contribution of access, page_close and navigate is high**, while **the contribution of wiki and discussion is low**. From the visual images in 5.3, attributes with high contribution degree of access, page_close and navigate have greater curve differences, while those with low contribution degree of wiki and discussion have higher curve coincidence degree. **The more differentiated the curve, the more information it provides to the decision tree**.

5.4.2 Promotion

In order to further improve the performance of the classifier, this paper will use the lifting method to obtain better training results. Historically Kearns and Valiant were the first to propose the concepts of "strongly learnable" and "weakly learnable". And in Schapire's later proof it turns out that these two things are equivalent.

Therefore, we can **use a linear combination of multiple basic classifiers to obtain a strong classifier.** Finding a basic classifier is not difficult. In this paper, the basic classifier is **decision tree**, and the lifting algorithm is the relatively common **Adaboost algorithm**. The evaluation indexes are still accuracy, F1 score and G-mean.

The experimental results are as follows:

```
AdaboostDecisionTree:
accuracy: 0.8399469070597019
F1 score: 0.902486690477795
G-Mean score: 0.8710919275311485
```

The improvement is obvious, with good improvements in all metrics compared to the basic classifier.

6 Summary

According to the visual analysis of the data and the contribution of each attribute of the

machine learning model, the students who skipped class completed the events less than those who insisted on learning, and many students who skipped class only triggered one or two events. From this point of view, there is still much room for improvement in online teaching.

7 Reference

[1]宋国琴，何春，章三妹.KDDcup2015 数据集研究[J].电脑知识与技术,2016,12(35):5-7.

[2]李航 统计学习方法