Online education data analysis and forecast

Content

1. Abstract

> Due to the advent of the epidemic, online education has developed rapidly and gradually become one of the main teaching methods. kddcup2015 competition data set is from Tsinghua Xuetang online course platform, which records the learning data of about 120,000 students and about 8 million learning logs in detail. This paper uses data visualization to observe user characteristics and analyze the main characteristics of students skipping classes; The decision tree and Adaboost promotion method were used to predict whether students skipped class or not, and the contribution of each attribute was observed. Finally, according to the qualitative and quantitative analysis to determine the characteristics of skipping class group.

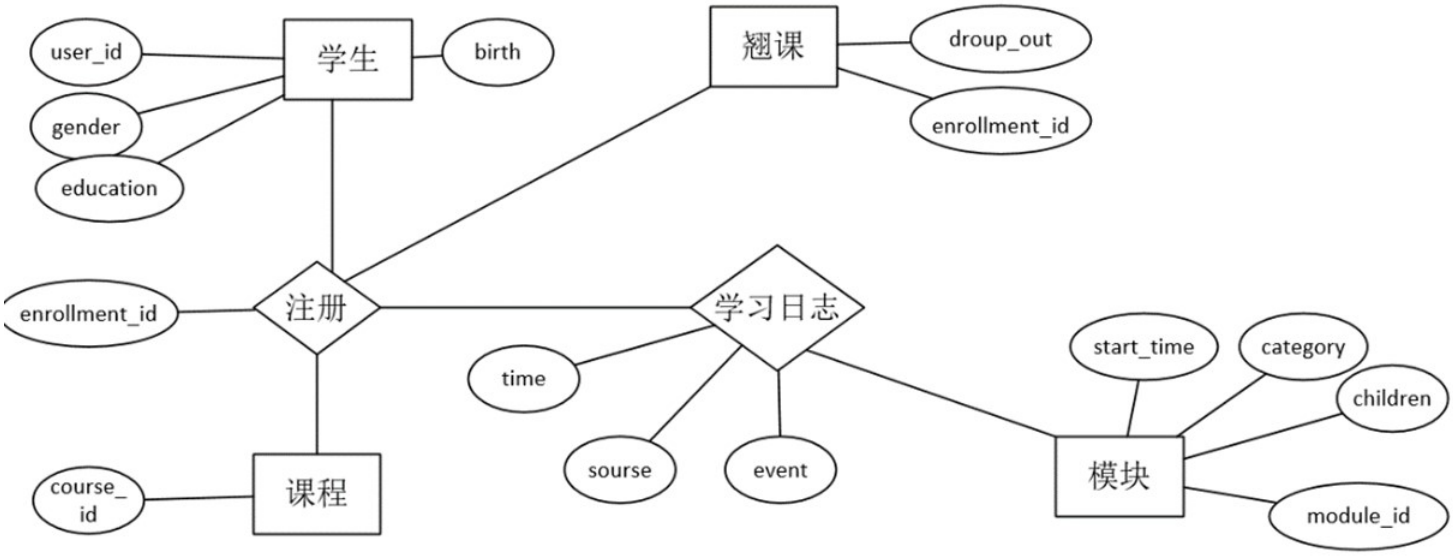
Keywords: data visualization, decision tree, Adaboost lifting method, skipping class

2. Introduction to the Dataset

2.1 Overview of the dataset

This dataset was obtained from the 2015 kddcup 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



3. Object of experience

3.1

Make statistics on the distribution of overall user characteristics, such as age and education background distribution, to obtain the overall user characteristics of the platform.

3.2

By processing the original data, the difference between skipping class and not skipping class is obtained, and it is visualized. To show the relationship between skipping class and learning behavior in an intuitive way. So that we can conveniently observe the characteristics of students skipping class from the data statistics chart.

3.3

A classifier can be trained to predict whether students skip class or not, and observe whether the contribution of each feature in the classifier is related to the statistical graph.

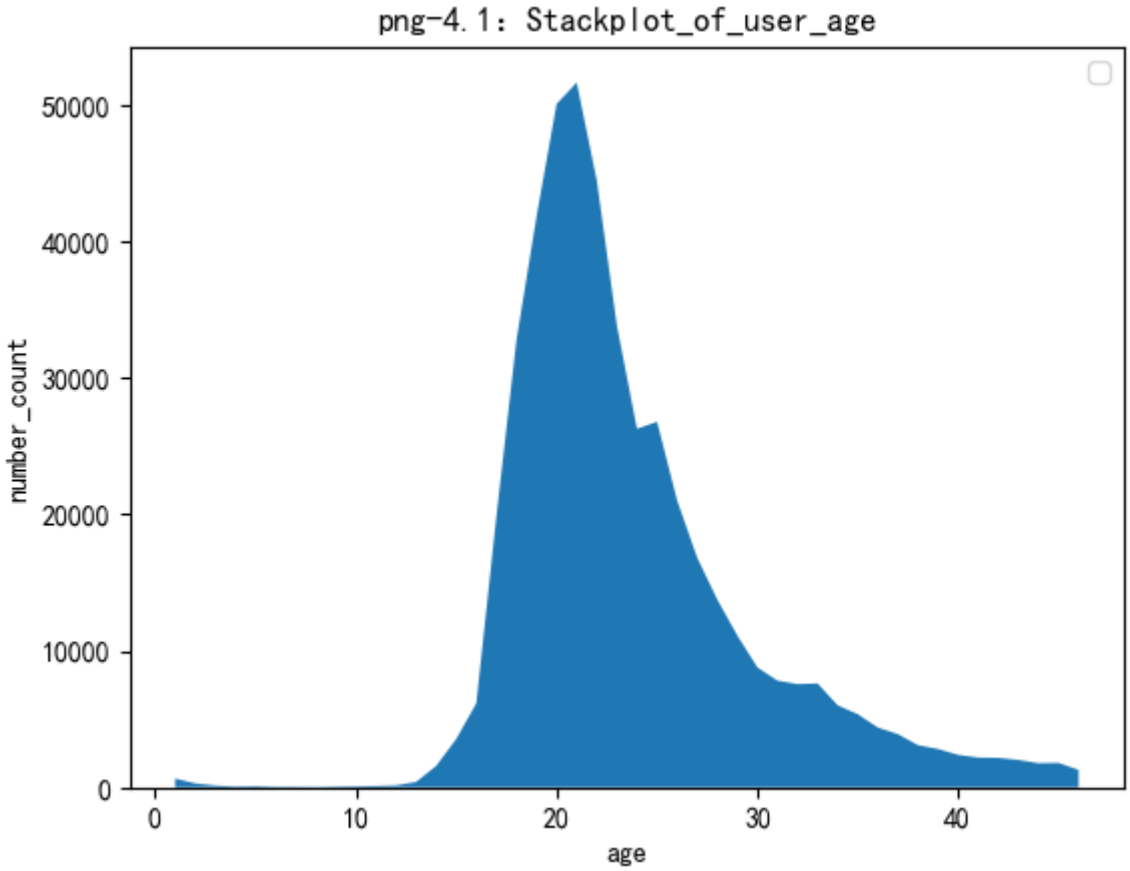
4. Distribution of user features

This task is mainly done using oracle 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 wh
group by (BIRTH)
order by age
```

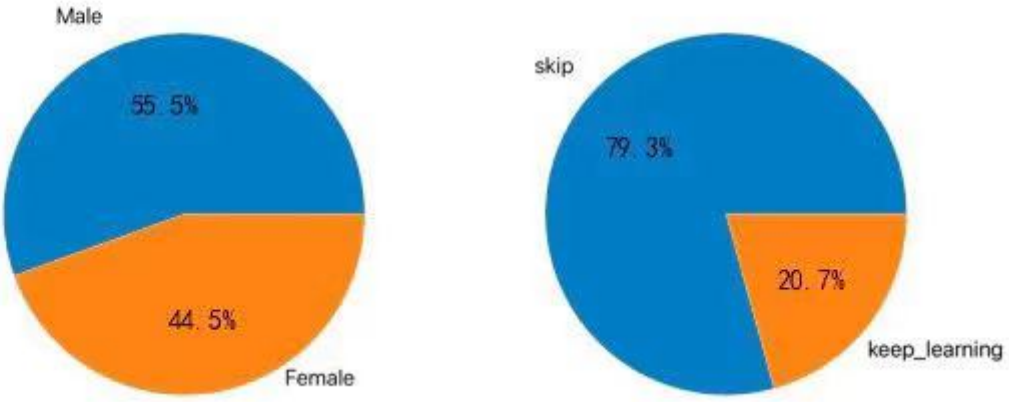


Analysis: According to Figure 4.1, it can be found that the user age is concentrated between 18 and 25 years old. It can be guessed that the main users of online education are undergraduate students and postgraduate students, mainly to complete the teaching tasks assigned by teachers.

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;
```

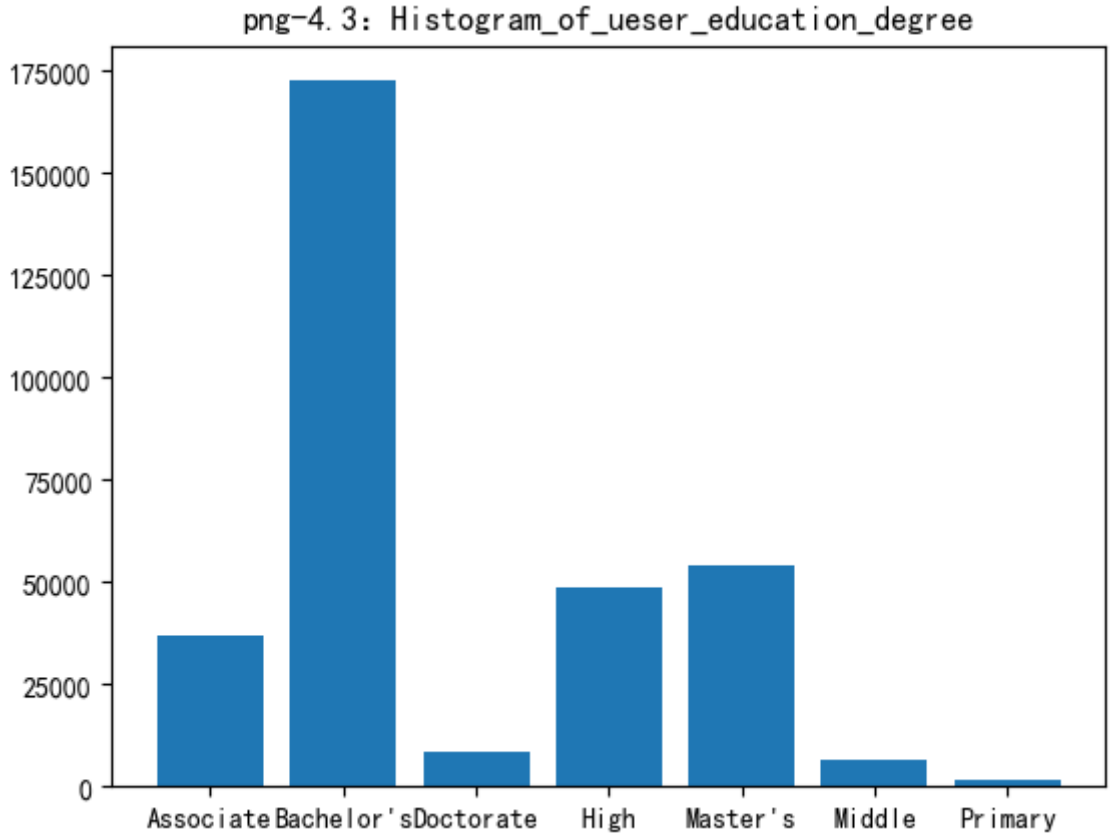
png-4.2: Pie-Char_of_urser_gender



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 EDUC
```

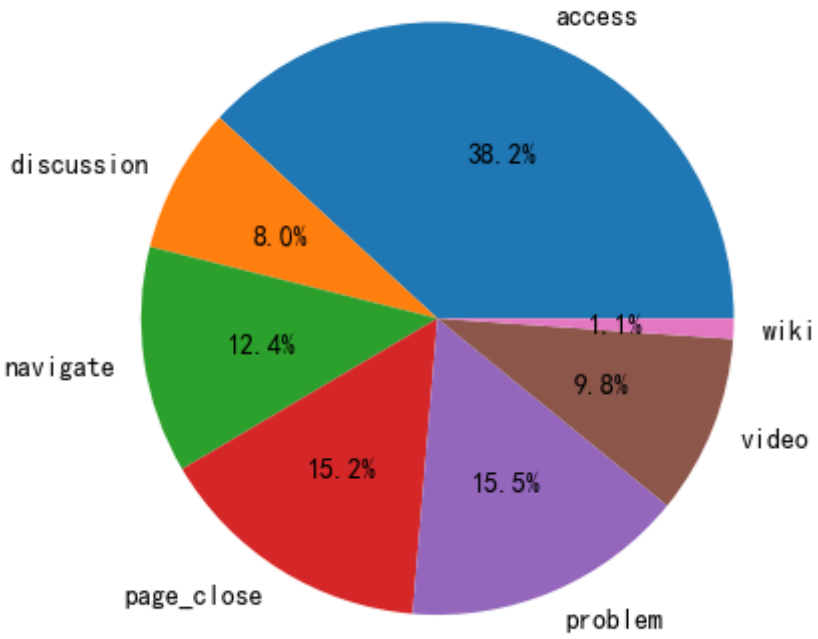


Analysis: Undergraduates account for the vast majority, which is also in line with the characteristics of more courses and heavy teaching tasks for undergraduates

4.4 Learning style

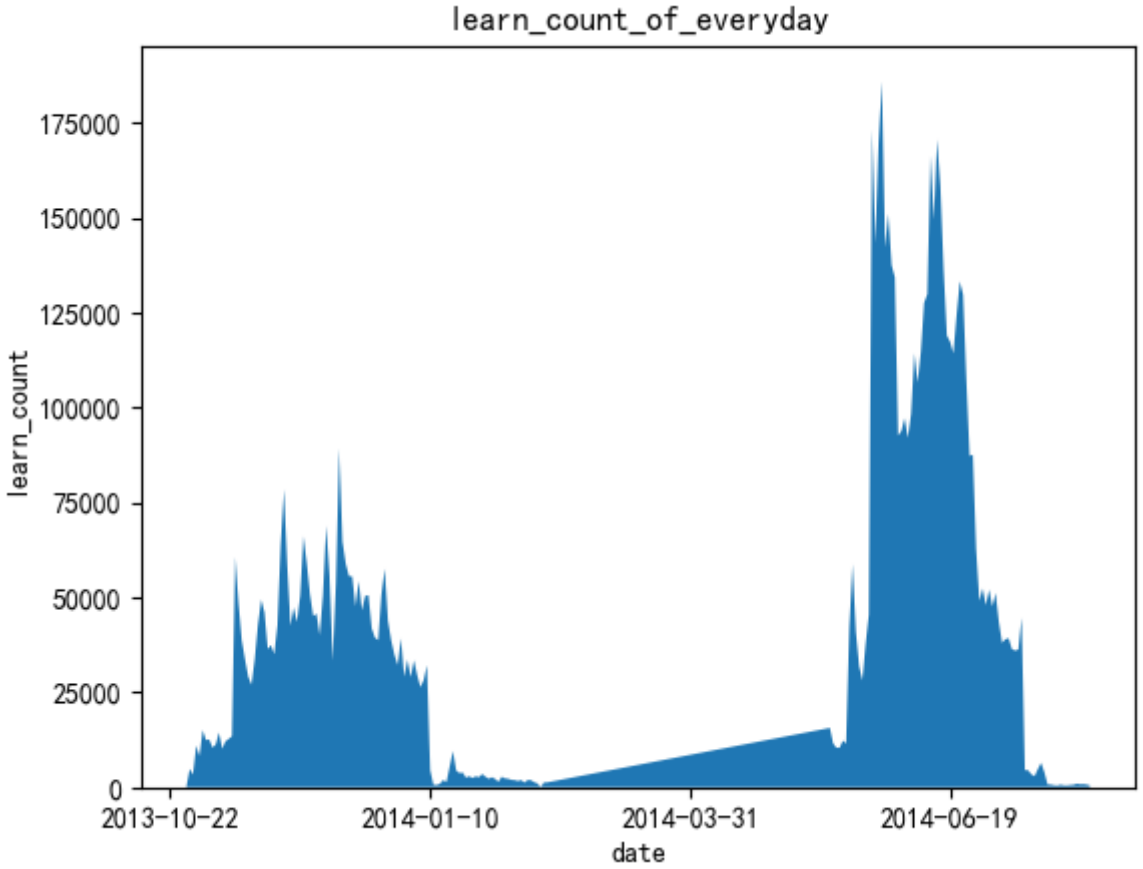
```
select EVENT,count(EVENT) count from LOG_TRAIN group by EVENT
```

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



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 oracle 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 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 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 l
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	50	48
2	60	40
3	55	20
4	82	18
5	75	72
6	90	30
7	141	101
8	167	18
9	214	15
10	190	31
11	195	8
12	244	30
13	257	39
14	284	6
15	261	10
16	264	2
17	265	8

Others 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,
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 an 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.

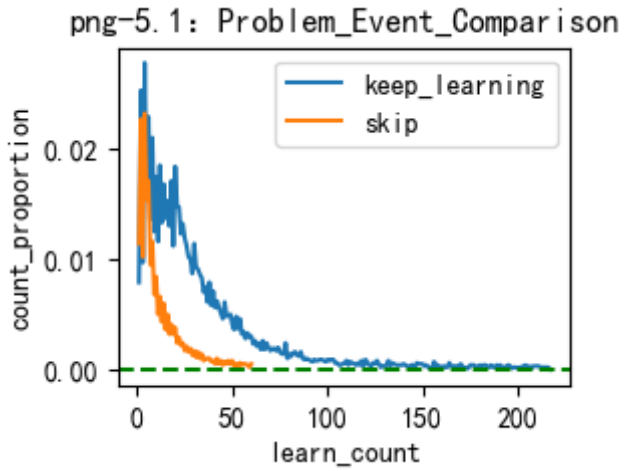
```
con = create_engine('oracle+cx_oracle://hadoop:root@127.0.0.1:1520/database');
sql="select * from STATISTICAL_INFORMATION";
df=pd.read_sql(sql,con);
df["lable"]=df["lable"].astype("int")
cols=df.columns.tolist()
del cols[0];
del cols[-1];
for col in cols:
    df.loc[df[col]>df[col].mean()+3*df[col].std(),col]=df[col].mean();
```

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

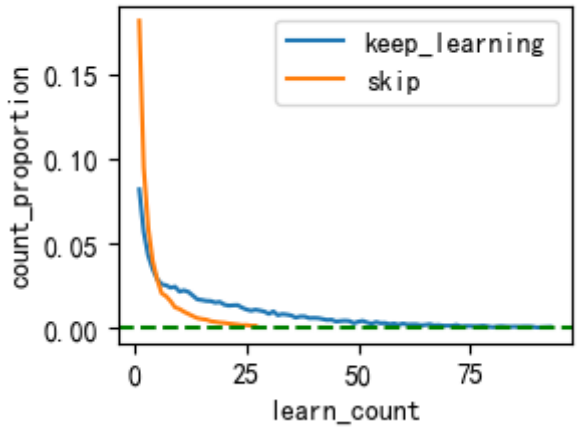
5.3 Visual analysis of skipping class or not

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 problem as an example:

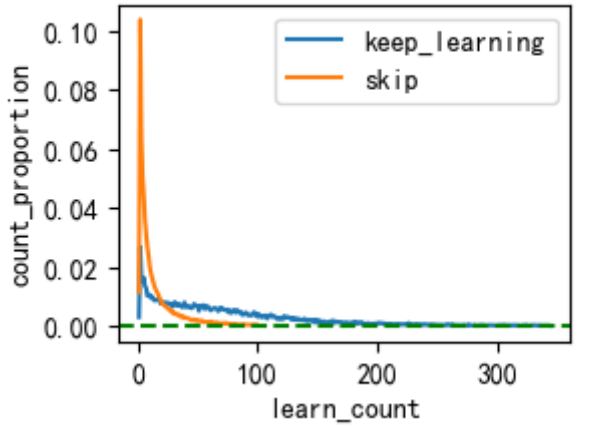
```
select PROBLEM,count(PROBLEM) count from statistical_information group by PROBLEM
```



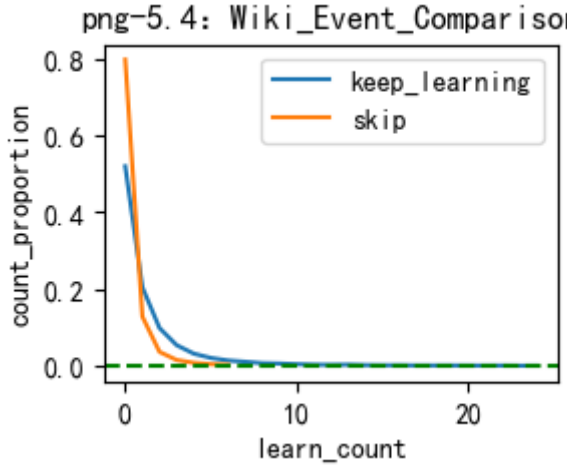
png-5.2：Video_Event_Comparison

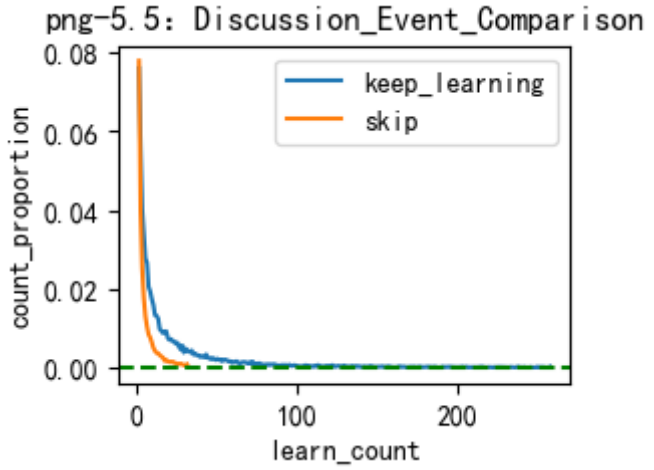


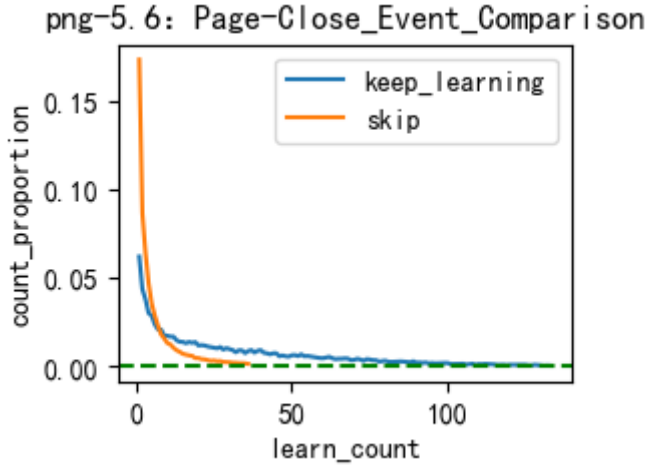
png-5.3：Access_Event_Comparison

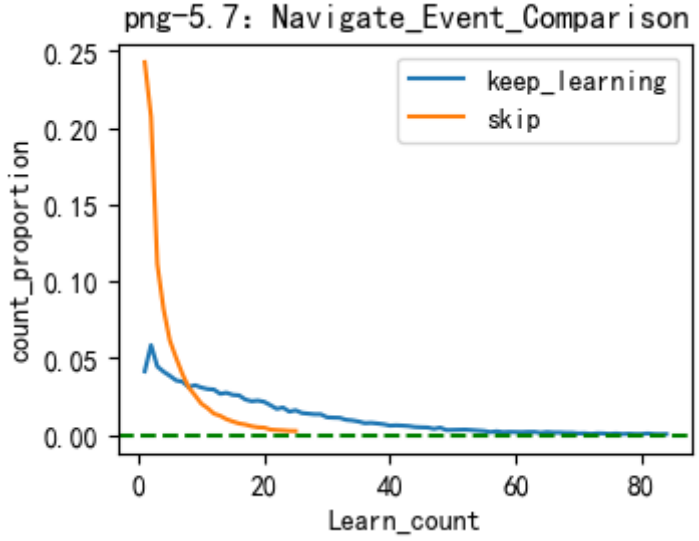


png-5.4：Wiki_Event_Comparison









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: 0.8069020822387523
F1 score: 0.8797216528583978
G-Mean score: 0.8289116417787024
```

Contribution:

0.09881	0.09897	0.44896	0.04068	0.08784	0.11117	0.11356

Corresponding fields:

```
['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.

```
AdaboostDecisionTree:
accuracy: 0.8370157343140779
F1 score: 0.9003718728870858
G-Mean score: 0.8687544805265787
```

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.