Supporting poverty-stricken college students in smart campus

Wu, F., Zheng, Q., Tian, F., Suo, Z., Zhou, Y., Chao, K-M., Xu, M., Shah, N., Liu, J. & Xu, M.

Original citation & hyperlink:
Wu, F, Zheng, Q, Tian, F, Suo, Z, Zhou, Y, Chao, K-M, Xu, M, Shah, N, Liu, J & Xu, M 2020, 'Supporting poverty-stricken college students in smart campus', Future Generation Computer Systems, vol. 111, pp. 599-616.
https://dx.doi.org/10.1016/j.future.2019.09.017

DOI 10.1016/j.future.2019.09.017
ISSN 0167-739X

Publisher: Elsevier

Supporting poverty-stricken college students in smart campus

Fan Wu[a,b,c], Qinhua Zheng,[a,b,d] Feng Tian[a,b,c], Zhihai Suo [e], Yuan Zhou [f], Kuo-Ming Chao [g], Mo Xu [e], Nazaraf Shah [g], Jun Liu [e], Fei Li [e]

[a] *MOE Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, China*

[b] *National Engineering Lab of Big Data Analytics, Xi'an Jiaotong University, Xi'an, China*

[c] *Systems Engineering Institute, Xi'an Jiaotong University, Xi'an, China*

[d] *Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China*

[e] *Network information center of Xi'an Jiaotong university, Xi'an, China*

[f] *The SA Office of Xi'an Jiaotong university, Xi'an, China*

[g] *Department of Computer Science and Technology, Coventry University, CV1 2JH, UK*

keywords: smart campus, big data, poverty-stricken college students, identification, supporting system

Abstract:

Chinese colleges have formulated supporting polices to help poverty-stricken college students to deal with the barriers in their living and learning. The difficulty in fully collecting the required information related to student's financial status and the imbalanced-data classification problem caused by the small proportion of poverty-stricken students among total students makes it a challenging problem. This problem results in a heavy workload for the college staff to identify poverty-stricken students, determine the amount of corresponding subsidy, and execute the supporting polices in an efficient way. Therefore, this paper attempts to address the above-mentioned challenges by proposing a smart campus system, which makes use of campus big data to identify poverty-stricken students and support the decision-making on the subsidy for them. The proposed system can also alert the counselors to provide psychological support for students in trouble. The major contributions of this research are as follows. Firstly, in addition to the features of students' amount of consumption on campus and its statistical characteristics used in existing researches, this paper proposes new features that describe diversity of consumable commodities, preference of consumption location and price, and characteristics of students' campus activities. Secondly, in order to solve the problem of dataset imbalance, four imbalanced data processing methods (Subsampling, Resampling, Cost-sensitive learning and SMOTE) have been applied to produce four different experimental datasets, and five classification algorithms (Random Forest, J48, Naïve Bayes, SMO, Logistic regression) have been used to train the classification model on each dataset. The experimental results indicate that the model based on Resampling and Random Forest achieves the best performance in F1-measure of poverty-stricken students, among the combinations of four imbalanced processing methods and five classification algorithms. In addition, a method of quantization of subsidies, and strategies of early warning and counseling for students are also described in this paper. A system was developed based on the above-mentioned methods, which meets the needs of individualized and diversified support for poverty-stricken students. The methods and the proposed system have been put into practice, and it is serving more than 17,000 students. The system has significantly improved the efficiency and quality of student management, and reduced the workload of college staff.

Introduction

As a part of smart city, smart campus puts forward a vision which is composed of convenient campus network, innovative scientific research, transparent and efficient administrative management, as well as colorful campus culture [1]. In attempts to improving the quality and efficiency of school management and teaching, many researchers have conducted research on

different aspects of smart campus, including students' achievement predictions [2-8], students' social network discoveries [9-10] and student's behavior discoveries [11][28]. However, there is a special demand for universities in China to identify and support poverty-stricken students [25]. Poverty can lead to an inconvenient campus life for some students and may make them unable to fully involve themselves in campus life and learning, which has a certain degree of negative influence on their physical and mental development. Many universities have formulated and put in place supporting polices to help such students.

In the process of implementing students' support policies, it is necessary to identify poverty-stricken college students (shorted as PS students in this paper) and determine their amount of subsidies. However, traditional methods of identification of PS students and determination of their amount of subsidies are based on the materials submitted by students and the related policies [12], which are relatively inefficient.

As the development of campus network and the widespread application of RFID technology increases, many researchers used the big data from smart campus to identify PS students and provide support for them [14-15]. They analyzed the consumption data recorded by students' campus card, to extract features that describe each student's consumption level and consumption behaviors according to the characteristics of consumption frequency and amount, etc. Some machine learning algorithms such as k-means clustering, k-NN, C4.5 and association rule mining were applied to train classification models to identify PS students. Having analyzed the existing research, this paper summarizes several problems and difficulties that need to be addressed in identifying and supporting PS students based on campus big data.

(1) Existing research only used students' campus consumption data to identify PS students. However, students' off-campus consumption data is not available, which causes deviation between the students' financial status reflected by campus consumption data and their real status. How to effectively use other kind of students' campus data and extract effective features to alleviate the impact of lacking off-campus consumption data is a difficult problem.

(2) The identification of PS students is a typical imbalanced data classification problem, the number of PS students vs. total students is less than 1:4, and the classification models trained on the original dataset are statistically bias towards the average student, which leads to the poor identification performance of PS students.

(3) There is a practical demand for determining the amount of subsidy according to the individual situation of each PS student as well as providing psychological support, which are not available in most of existing supporting systems based on big data of smart campus.

Our big data platform at Xi'an, Jiaotong University, provides well-collected and integrated data of all students' related departments which enables us to overcome the above-mentioned difficulties. This paper uses various kinds of student campus data in order to mine the implicit features of their living habits or behavioral rules related to their economic conditions, and to obtain the identification model of PS students by employing a machine learning approach. An identification and supporting system for PS students has been developed to address these challenges. The main contributions of this paper are as follows:

- On the basis of the existing research and the theories of pedagogy and psychology, this paper puts forward new implicit features, including personal and familial situation, consumption, location and price preference, campus activity, and consumption of diverse items. These features effectively alleviate the difficulty of identifying poverty-stricken students caused by the lack of consumption data outside of the school.
- A well performed classification model for PS students is obtained by using the newly extracted features and conducting imbalanced data processing and selecting it from results

of extensive experiments based on combinations of four imbalanced preprocessing methods and five classification methods.

- By employing above methods, an identification and supporting system for PS students based on campus big data is developed. In which a method of subsidy quantization incorporated, the method of which calculates the subsidy of each PS student according to the values of some specific features and the individual situation of each PS student is described. In addition, by analyzing the changes in students' behavior during a certain period of time, it tracks PS students' psychological state and provides early warning of the students in trouble. This early warning feature can help counselors provide timely psychological counseling support to such students. The system is integrated into our university's big data platform. The system has been put into practice and it helps to improve the efficiency and quality of student management including reduction in the workload of college staff.

The contents of this paper are structured as follows: Section 2 describes the related work. Section 3 proposes the framework and working mechanism of the identification and supporting system for PS college students. Section 4 describes the experiments and the result analysis of identification of PS students. Section 5 provides a brief introduction to big data platform and detailed information about support and analysis system for student development. Finally, section 6 concludes the paper.

Related work

The aspects of student's behavior living on campus mainly includes consumption, studying, social activities, etc. Student's behavior analysis is one of the foundations of intelligent management in smart campus. It uses campus big data to analyze student behavior and status, which can help administrators better understand students' situations and needs and provide timely and effective service according to students' demands.

The existing student behavior analysis mainly includes academic achievement prediction, student social network analysis and discovery and student abnormal behavior detection. The big data of smart campus used in student's behavior analysis contains different kinds of records collected by intelligent terminal devices of campus card systems and teaching and management data collected from departments' and units' systems. Liu T, Liu Y, Fan S et al. uses campus card records to analyze the student social network [9-11]. Conijn R, Hu Y et al. studies the early warning of students' online learning scores using data collected from systems such as Moodle and LMS [2][4]. Ashwin S and Gayathri R use students' learning behavior data for classification and clustering to predict students' academic achievements [3].

The existing research on identification of PS students can be presented from three aspects: data and features, identification of PS students and support for PS students.

The data used in existing research includes student's basic personal information, family situation information, consumption records, and based on them, two kinds of the features are extracted [16-20]. One kind is the personal and family features including gender, family annual income, etc. The other kind is consumption features that describe consumption amount/frequency and their statistical characteristics. In addition, there are some psychological studies which provide new perspective for feature extraction of PS students. Fuqiang G studies the relationship between consumption behavior and consumption psychology of PS students [21].

Various machine learning methods have been used to identify PS students. Huaifeng Q uses K-means on consumption data in canteens and finding the "optimal group" of PS students

[20]. Wenjuan W, Chaowen W et al. uses improved K-means, C4.5 and Apriori algorithms to find the relationship between PS students and their consumption behavior in canteen according to the consumption level of students [14] [22]. Although the above research can improve the efficiency of the identification of the PS students, however the performance of these models is poor.

Support for PS students mainly includes quantification of the subsidy and supporting method selection. At present, most universities give PS students first-class, second-class and third-class subsidy according to their poverty level [22]. There is little research on determining the amount of subsidy according to the individual situation of each PS student.

In the exploration of multi supporting methods, universities such as South China University of Technology and University of Electronic Science and Technology track the consumption behaviors of students on campus and conducted telephone and face to face interviews with students who have abnormal consumption behaviors to learn their recent family or individual condition to provide second time subsidies if necessary. Moreover, Zhejiang University has helped PS students to achieve diversified quality training by setting up ten comprehensive quality training centers and three practice projects for PS student.

In addition, there are pedagogical and psychological studies on the psychological problems of PS students. According Desheng K's study, the psychological problems of PS students in college are mainly manifested in the interweaving of inferiority, low self-esteem, anxiety, sensitivity and lack of self-confidence in interpersonal communication, and too frugal in the use of money [24]. At the same time, the authors also point out that PS students with poor academic performance are more likely to have psychological problems, which makes them difficult to adapt to the campus environment and have a negative impact on their future development [24]. The psychological problems of PS students may lead to a vicious circle of study and psychology, it is necessary to provide psychological support to the PS students who have abnormal psychological status.

Based on the existing research, this study makes use of big data of smart campus, and describes a smart campus system, which uses big data collected from campus to identify poverty-stricken students and support the decision-making on quantization of subsidy for them, even remind counselors to provide psychological support for the students being in troubles.

Mechanism of identification and supporting system for poverty-stricken college students

The development of the Internet of things technology, cloud computing technology, and the wide spread application of intelligent terminals on campus resulted in generation of a large volume of different types of data. This research effort exploits the availability of the campus data by developing an identification and supporting system for PS students based on the big data of smart campus. A high-level architecture of this system is shown in Fig 1.

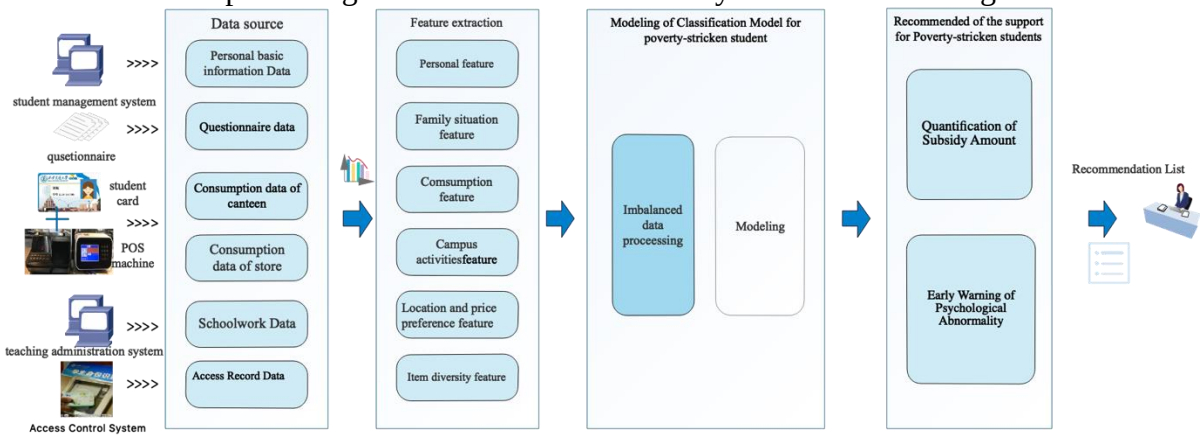

Fig 1 Working mechanism of identification and supporting system

Fig 1 shows the working mechanism of the proposed identification and supporting system. It contains four modules: Data source, Feature extraction, Modeling of classification model for PS student and Recommendation of the support for the Poverty-stricken students. Firstly, the Data source module collects relevant data of students' campus life through various terminals based on campus card system, student management system and teaching administration system. The terminals include computers of the various functional departments, POS machines for campus consumption and access control system based on RFID, etc. Secondly, the Feature extraction module extracts and chooses effective features from collected data for recognition of PS students. Thirdly, the Modeling module adopts an imbalanced data processing method (Resampling method is selected) on collected data, and then uses a machine learning method (Random Forest algorithm is used) to train the classification model for PS students. Finally, recommendation of the support for PS students module gets the recommendation list of PS students based on the classification result of the model, and the amount of subsidy can be determined according to the values of certain features of each PS student. The system can also give an early warning of psychological status of the abnormal students to counsellors according to the features related to their psychological status. Each module is described in detail as follows:

Data source: It includes student's personal information such as registration, questionnaire data about students' family background, consumption data in campus canteens and stores, enter-and-exit records of certain places (including dormitory, library, stadium, teaching building, etc.), and their grades. The six types of data collected in this paper come from three data sources. The first data source is in a student management system which contains personal and basic information data. This data is entered into the system when students are enrolled in the college. It contains the following items: student ID, name, province, admission category, examinee number, birth date, major, subject, class, political outlook, nationality, household registration type and graduation category, etc. The second data source is concerned with family background and it is obtained through questionnaire, which is filled out by every freshman during their first month in the college. The questionnaire includes the following items: the number of family member who are under 18 years, the number of family member who are over 65 years, whether their parents are still alive, whether they are divorced, their parents' educational level, their parents' occupation types, their relatives' occupation types, family income, and family debt, etc. The third data source is concerned with the use of campus card system. The student card system manages and records every student's campus data related to consumption, study, and daily life activities etc. When students consume products in campus canteens, stores and so on, they pay by swiping campus card (pre-paid) on POS machines. The system records the amount of consumption, time and POS machine's location. When students enter or leave a library, gymnasium, dormitory and other venues, they get access by swiping their campus card. The entrance guard subsystem records time, and place and other related information. In addition, teaching management subsystem records the information of course selection, achievement, evaluation of teaching, rewards and punishments, etc. Table I -II in Appendix are examples of the original records of campus card system. Refer to section 4.1 for the detailed description.

Feature extraction and selection: By analyzing and using the big data of smart campus, the features used to identify PS students are extracted as shown in the left two columns of Fig 1. Then, feature selection methods (correlation analysis and information gain are used in this paper) are adopted to delete irrelevant features. Finally, the feature set for model training is obtained, it includes five feature categories, which are personal and family situation, consumption, campus activities, location and price preference and consumption of diverse items. This module provides the foundation for the identification of PS students, quantification of the amount of subsidy, and early warning of student's psychological abnormality.

Modeling of Classification Model for poverty-stricken student: This module includes two steps: imbalanced data processing and modeling. The imbalanced data processing is to solve the problem of dataset imbalance, which leads to the poor identification performance of poverty-stricken students. In this step, four frequently-used imbalanced data processing methods are adopted, including Subsampling, Resampling, Cost-sensitive learning and SMOTE[13]. Among them, Resampling, Subsampling and SMOTE solve imbalanced problems at the data level by resampling or constructing new data and generating balanced datasets. Cost-sensitive learning introduces a cost matrix in model training which adjusts the bias of the model and solves the imbalanced data problem at the algorithmic level. Comparing the performance of the above four methods, we choose Resampling as the imbalanced data processing method. In the Modeling step, we compare the performance of the models trained by multiple machine learning algorithms and select the model with the best performance (Random Forest algorithm is selected in this paper).

Recommendation of the support for Poverty-stricken students: It includes quantification of the amount of subsidy and early warning of student's abnormal psychological state. In this step, the amount of subsidy can be determined according to the individual situation of each PS student, and counselors can be reminded to provide psychological counseling to PS students with abnormal psychological status. Refer to 5. 2 and 5.4 for the more detailed descriptions.

Finally, a recommended list of PS students and the amount of their subsidies will be sent to counselors to help them in their decision making.

Experiments and their result analysis of identifying poverty-stricken college students

4.1 Datasets

The experiment uses two original datasets, ***data16*** and ***data17,*** collected from the students who registered in 2016 and 2017, over the first two months (usually September and October of the year) in their first year of university life, respectively. ***data16*** includes 3,244 students (including 835 PS college students) and 1,750,000 records from the data source in Fig 1. ***data17*** includes 3,151 students (including 659 PS college students) and there are total 1,880,000 records in this data set. Note that, the reason behind selecting the data of the first two months is that the application and subsidy of each student is checked before the end of that period and subsidy will be transferred into each PS student's corresponding account after the third month of their first year.

4.2 Features of identifying PS students

After exploring ***data16*** and ***data17***, five kinds of the features for identifying PS students are selected. These features are personal and family situation, consumption, campus activity, location and price preference and consumption of diverse items. The personal and family situation features are deemed as a kind of static feature, as well as consumption, campus activity, location and price preference and consumption of diverse items.

The data items of students' basic information and family situation questionnaires are shown in Table III-IV in Appendix, each of which is a personal profile and family status.

The entities and record samples of consumption data and access records are shown in Table1-2. The process of extracting features from them is as follows:

Table 1 Entities and record samples of consumption data

Student ID	Name	Record time	Operation	pay out (cent)	balance (cent)	Location	POS id	Description
2********3	song**	8/18/17 6:16 PM	consumption	-350	9650	Material supply center\Xinhua Store\	3	Xinghua Store
2********8	wang*	8/18/17 5:31 PM	consumption	-800	18740	Service Center\East Campus\Kangqiao Store	2	Kangqiao Store

2********8	wang*	8/18/17 5:14 PM	consumption	-60	19940	Service Center\East Campus\Kangqiao Store	2	Kangqiao Store

Table 2 Entities and record samples of access record data

Student ID	Name	Record time	Location	Description
2********1	Xie*	8/23/15 9:38 PM	No.10 Medical building	enter
2********7	Hu**	8/23/15 10:00 AM	Wen Zhi Academy	exit
2********7	Hu**	8/23/15 10:00 AM	Wen Zhi Academy	exit

Firstly, by analyzing the original records, we concluded four dimensions (time, place, price, and item) of these data, and their concept hierarchies.

1) Concept Hierarchy of Time

The concept hierarchy of time is an ordered sequence, {month, day, hour}. Firstly, for month-level data operation, the data of one year are segmented monthly, and feature values are extracted at monthly intervals and formed a dataset. Therefore, for data of the year, we obtained twelve datasets, which are recorded as D_1, D_2, …, D_m and defined as $D = \{D_1, D_2, \cdots, D_m\}$, where m is between 1 and 12. Then, for day-level data operations, we defined four type labels (weekday, weekend, statutory-holidays, and winter-or-summer-vacation) for each day, which is denoted as $m_{weekday}$, $m_{weekend}$, $m_{holiday}$, $m_{sw-vocation}$ correspondingly, and defined as $M_{day} = \{ m_{weekday}, m_{weekend}, m_{holiday}, m_{sw-vocation}\}$. Note that, if one day is one of national statutory holiday of the years, the day is marked as statutory-holidays. Similarly, if one day is in winter or summer vacation of that school year, the day is marked as winter-or-summer-vacation. For hour-level operations, considering the schedule of students' studying and resting in college, this paper divides 24 hours of a day into four periods: morning, noon, afternoon and night, which are recorded as $T = \{t_m, t_{no}, t_{aft}, t_{nt}\}$. The division is shown in Table 3.

Table 3 Time division of a day

Division	Time interval
morning	6:00-11:00
noon	11:00-14:00
afternoon	14:00-18:00
night	18:00-24:00

2) Concept Hierarchy of items of food

Concept hierarchy of items of food sold in university canteen has two layers, the top layer and bottom layer. The top layer has two types (main food and other) and bottom layer has nine categories, shown in Table 4.

Table 4 Concept Hierarchy of Item and Price

Item type	Category
main food	rice
	wheaten food
	Chinese food
	Western-style food
other	Snacks
	Dessert
	cold dish
	Fruit
	Drink

3) Concept Hierarchy of Price of food

Concept hierarchy of Price of food has two layers, the top layer is price level and bottom layer is price. The price of consumption items (noted as p) in campus is grouped into two levels: level1 and level2, where level1 represents a high-price item ($p \geq 8$), and level2 represents a low-price item ($0 \leq p < 8$), as shown in Table5.

Table 5 Concept Hierarchy of Item and Price

Price level	Price (p)
level1	$p \geq 8$
level2	$0 \leq p < 8$

4) Concept Hierarchy of Location

The main places of the college are divided according to their functions and types, as shown in Table 6.

Table 6 Concept Hierarchy of Position

	Function	**Type**	**Location name**
Location	Study	library	library1, library2,…
		teaching building	teaching buding1,…
	Life	dormitory	dorm1,dorm2,…
		canteen	canteen1, canteen2,…
		store	store1, store2,…
	Campus activities	gymnasium	gym1,gym2,…

Then, based on the above conceptual hierarchies of four dimensions, a starnet query model[27] is constructed and shown in Figure 2. Based on the combination of different dimensions and different hierarchies, 268 features and their measures, such as frequency, count, sum, mean, variance, max/min, are formed and calculated.

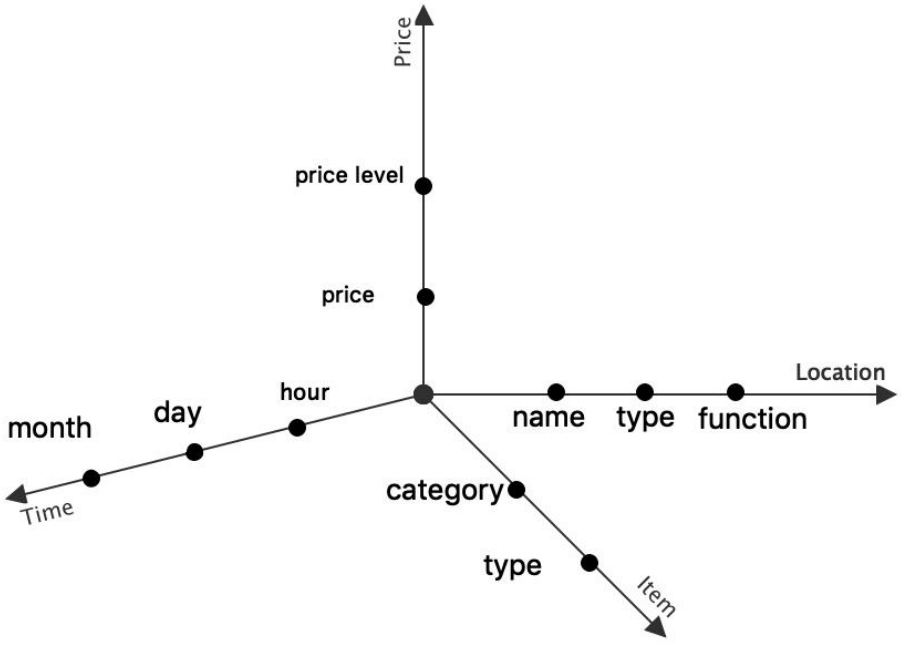

Fig 2 The starnet query model of campus card system data

Then, after the 268 features are filtered (in this paper, filtering out the features with 90% missing values), 135 features are obtained, as shown in Table V in Appendix. Furthermore, box-dividing, normalization and Boolean quantization operations carry on these feature values.

Finally, through many different feature selection methods (including method based on single type feature, method of manually combining different feature types, feature selection method

based on information gain, feature selection method based on correlation analysis etc. [13], [27]), 81 features are selected (shown in Table 7) due to their achieving outstanding classification performance.

Table 7 List of selected features

Category	List of features selected
Personal/Family features	***Existing features:*** gender, number of family members (4), father's education level, mother's education level, father's job type, mother's job type, family deposit, family debt, family economic level. ***Newly extracted features:*** **parent alive, parent divorced, physical condition of parents (5), sibling condition (2), family position level (10), household car, province, subject, major, ethnic groups, graduation type**
Consumption features	***Existing features:*** monthly expense in canteen, monthly expense in store, monthly consume times in canteen, monthly consume times in store, mean-every purchase expense in canteen, mean-daily expense in canteen, var-every purchase expense in canteen, var-daily expense in canteen, max-daily expense in canteen, mean-daily breakfast expense, mean-daily lunch expense, mean-daily afternoon expense in canteen, mean-daily dinner expense. ***Newly extracted features:*** **max-daily dinner expense, var-daily dinner expense, var-weekday daily expense in canteen, max-weekday daily expense in canteen, max-weekday every purchase expense in canteen, var-weekday every purchase expense in canteen**
Item diversity features	***Newly extracted features:*** **monthly consumption item type num, median-daily lunch location num, median-daily lunch type num, median-daily breakfast location num**
Location and price preference features	***Newly extracted features:*** **tendency of food type (1), noon top1 item type, price preference of lunch, afternoon top1 item type, night top1 item type, price preference of dinner, monthly canteen price tendency, noon canteen price tendency, weekday canteen price tendency**
Campus activity features	***Newly extracted features:*** **monthly canteen consume day num, monthly store consume day num, monthly morning consume day num, monthly noon consume day num, monthly afternoon consume day num, monthly night consume day num, weekday_weekend canteen single cost mean compare, weekday_weekend canteen daily cost mean compare, weekday_weekend store single cost mean compare, weekday_weekend store daily cost mean compare, number of weekdays consumption in the campus, number of weekends consumption in the campus**

* the newly proposed features in this paper are in bold, and the ***Existing features*** has been taken from Refs. [16-20].

Based on features presented in Table 6, we generated two datasets, ***original1*** and ***original2***, which are composed of all features in Table 1 extracted from ***data16*** and ***data17***, respectively. In addition, we also generated a dataset named ***compare1***, which is composed of the plain-texted features in Table 1 extracted from ***data16***. And the dataset obtained by merging by ***original1*** and ***original2*** is named as ***original3***.

The length limitation of the paper does not allow us to discuss each feature in greater detail. Therefore, we choose one feature for each kind of feature type (except Personal/Family feature) to briefly describe and analyze as follows.

- The average daily lunch expenses (is one feature of Consumption): The boxplot of average daily lunch expenses of PS students and ordinary students in ***original1*** are shown in Fig 3. It can be seen from Fig 3 that PS students generally spend less on lunch than ordinary students, because the median, mean and maximum of daily average lunch expenses of PS students are lower than those of ordinary students.

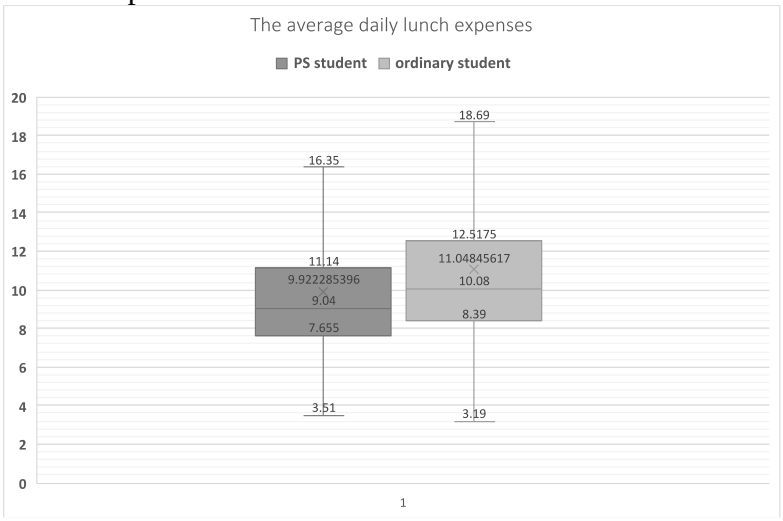

Fig 3 Boxplot of average daily lunch expenses of PS students and ordinary students in ***original1***

- Diversity of monthly consumption item type (is one feature of Item diversity): The boxplot of the number of monthly consumption item type (the item type list seen in Table IV of Appendix) for PS students and ordinary students in ***original1*** are shown in Fig 4. Fig 4 shows that PS students have less diversity of consumption items as compared to ordinary students, as median, mean and minimum of the number of monthly consumption item type of PS students are lower than those of ordinary students.

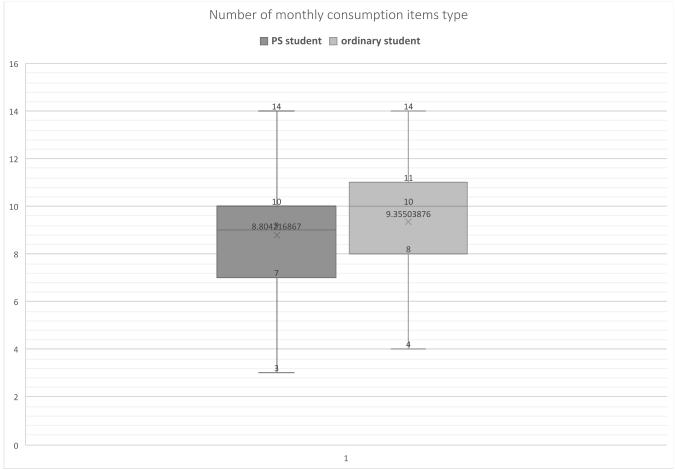

Fig 4 Boxplot of the number of monthly consumption item type for PS students and ordinary students in ***original1***

- Price preference of lunch (is one feature of Location and price preference): The price of consumption items on campus is grouped into two levels: **level1** and **level2**, where level1 represents a high-price item, and level2 represents a low-price item. The price preference for lunch is a ratio of the frequency **level1** items purchase which divides the frequency of **level2** items purchase in lunch per month. The boxplot of price preference of lunch for PS students and ordinary students in original1 are shown in Fig 5. We can conclude from Fig 5 based comparison of median, mean and

maximum/minimum of the price preference that PS students tend to choose low-price food at lunchtime.

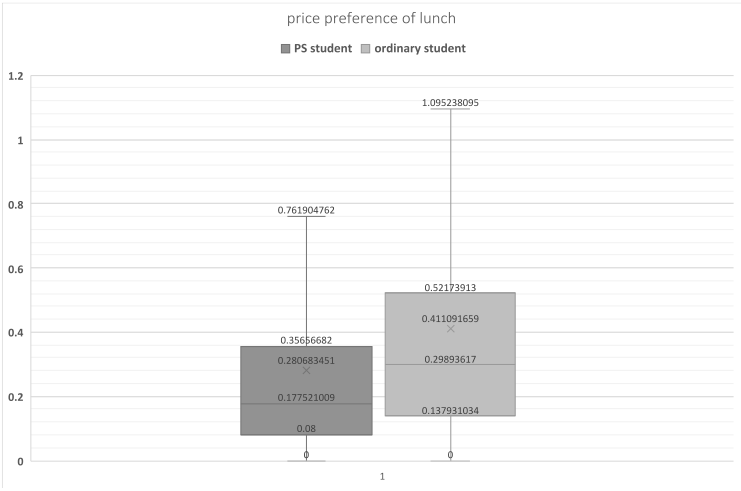

Fig 5 Boxplot of price preference of lunch for PS students and ordinary students in ***data16***

- Number of weekends consumption in the campus (is one feature of Campus activity): The boxplot of the number of days during weekends that relate to consumption on campus for PS students and ordinary students in ***original1*** are shown in Fig 6. Figure 6 indicates that median, mean and minimum value for this feature of PS students are higher than those of ordinary students, which means that PS students tend to consume on campus more often at weekends than ordinary students.

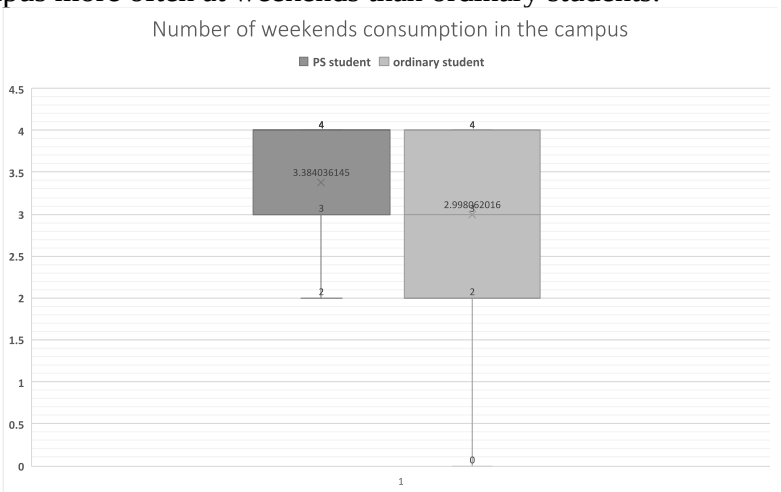

Fig 6 Boxplot of the number of weekends of consuming on campus for PS students and ordinary students in ***original1***

In addition, in order to understand the effect of the proposed features, the performance comparison experiments of the classification models trained on different feature sets are carried out. See 4.4.2 for details.

4.3. Classification model selection to identify poverty-stricken college students

In order to solve the problem of dataset imbalance, the process of constructing the classification model for poverty-stricken students in this paper consist of two steps.

- In Step1, imbalanced datasets are processed to generate the balanced datasets. Four imbalanced data processing methods are adopted, including Subsampling, Resampling, Cost-sensitive learning and SMOTE [13].

- Step 2 aims to train the classification models for poverty-stricken students by using different classification methods and select the best performing model. The classification algorithms used in the experiment are Random Forest, J48, Naïve Bayes, SMO, Logistic regression [13].

4.4 Experiments and results analysis

The experiments carried out in the paper are listed in Table 8. The results of each experiment will be introduced in following subsections.

Table 8 list of the purposes, datasets, classification methods adopted in each experiment

No.	Datasets adopted	Methods adopted	Describing Purposes of the experiments
#1	***original1, compare1***	Random Forest/Naïve Bayes/Logistic Regression/SMO/J48	Effectiveness validation of new features (4.4.1)
#2	***original1***	SMOTE/Cost-sensitive learning/ Subsampling/Resampling; Random Forest/Naïve Bayes/Logistic Regression/SMO/J48	Effectiveness validation of imbalanced data processing and selection of the best performed classification model (4.4.2)
#3	***original1, original2, original3***	Resample; Random Forest	Model stability verification (4.4.3)

Note that, the performance of each model is evaluated by 10-fold cross-validation on the datasets.

4.4.1 Experimental results and analysis of the validation of new features

Experiment #1 applied five algorithms (Random Forest, Naïve Bayes, Logistic Regression, SMO and J48) on ***original1*** and ***compare1*** to verify the improvement in performance of the proposed new features. The values of F1-measure are shown in Fig 7 and Fig 8.

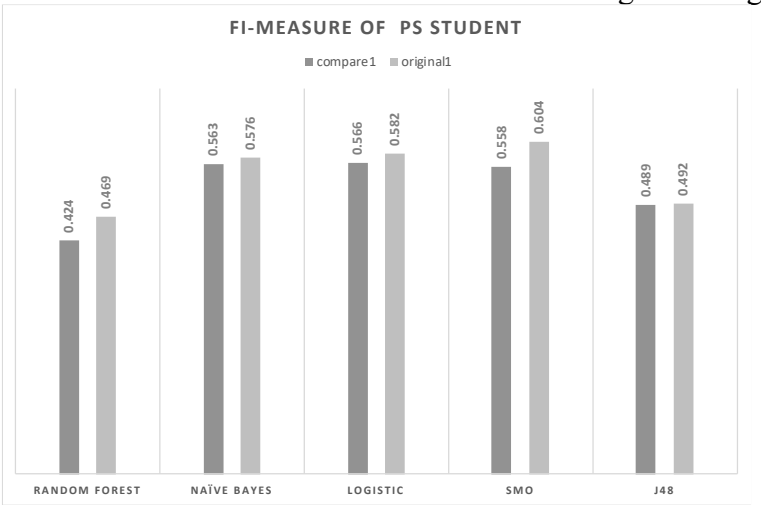

Fig 7 Values of PS student's F1-measure by five classification methods carried on ***original1*** and ***compare1***

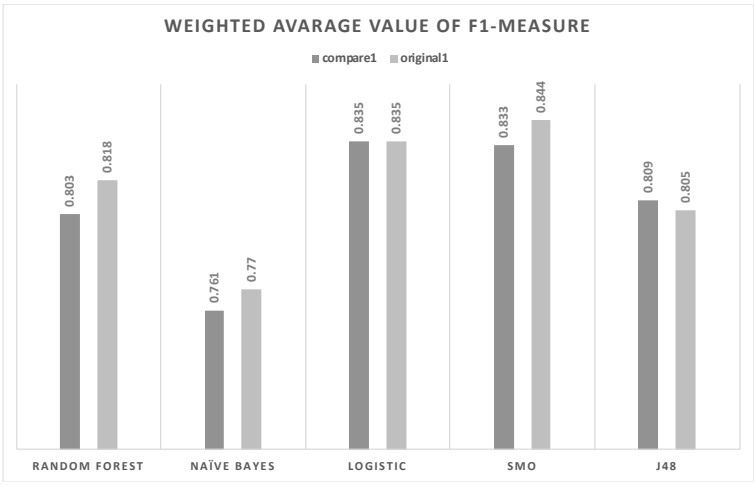

Fig 8 weighted average values of F1-measure by five classification methods carried on ***original1***and ***compare1***

The results of Experiment#1 show that the model trained on ***original1*** by SMO performed best among others and the value of F1-measure of PS students is 0.604, which is 8.24% higher than that of ***compare1***. The value of weighted average of F1-measure reached 0.844, as shown in Figures 7 and 8, respectively. It can be seen that new features are effective and can improve the performance of classification model of to a certain degree.

4.4.2 Experimental results and analysis of imbalanced data processing

Experiment#2 applied four imbalanced processing methods (SMOTE, Cost-sensitive learning, Subsampling and Resampling) and five classification algorithms (Random Forest, Naïve Bayes, Logistic Regression, SMO and J48) on ***original1*** to verifies that imbalanced data processing can improve the performance of classification model and selects the best performing model. The values of F1-measure are shown in Fig 9 and Fig 10. In this experiment, ***smote1***, ***cost-sensitive1, subsample1 and resample1*** represent the datasets that were generated by applying SMOTE, Cost-sensitive learning, Subsampling and Resampling on ***original1***, respectively.

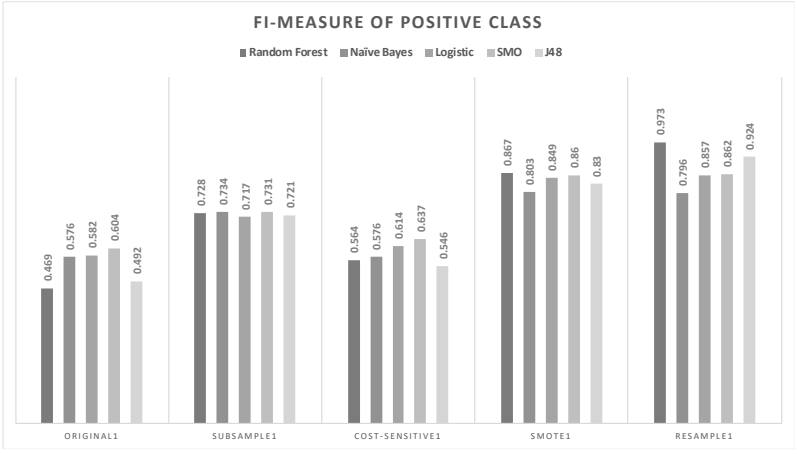

Fig 9 Values of PS student's F1-measure by four imbalanced data processing methods and five classification methods carried on ***original***

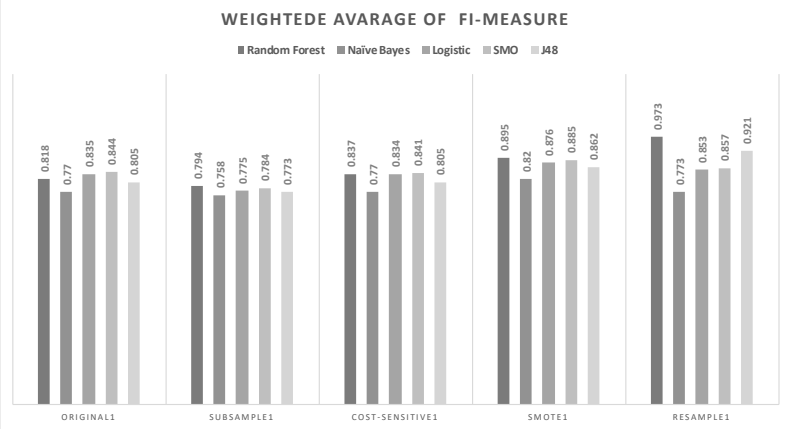

Fig 10 weighted average values of F1-measure by four imbalanced data processing methods and five classification methods carried on ***original***

The results of Experiment #2 show that the model based on Resampling and Random Forest performed best among other three imbalanced data processing methods and four classification algorithms. The value of PS student's F1-measure and the weighted average of the F1-measure in ***resample1*** are 0.973, as shown in Figures 9 and 10, respectively. Therefore, we choose the model trained on resample1 by applying Random Forest as the final classification model.

4.4.3 Experimental results and analysis of model stability verification

Experiment #3 applied five classification algorithms (Random Forest, Naïve Bayes, Logistic Regression, SMO and J48) on six datasets to verify the stability of the chosen model in 4.4.2. The values of F1-measure are shown in Fig 11 and Fig 12. In this experiment, ***resample 1, resample2*** and ***resample3*** represent the datasets that were generated by applying Resampling on ***original1, original2*** and ***original3***, respectively.

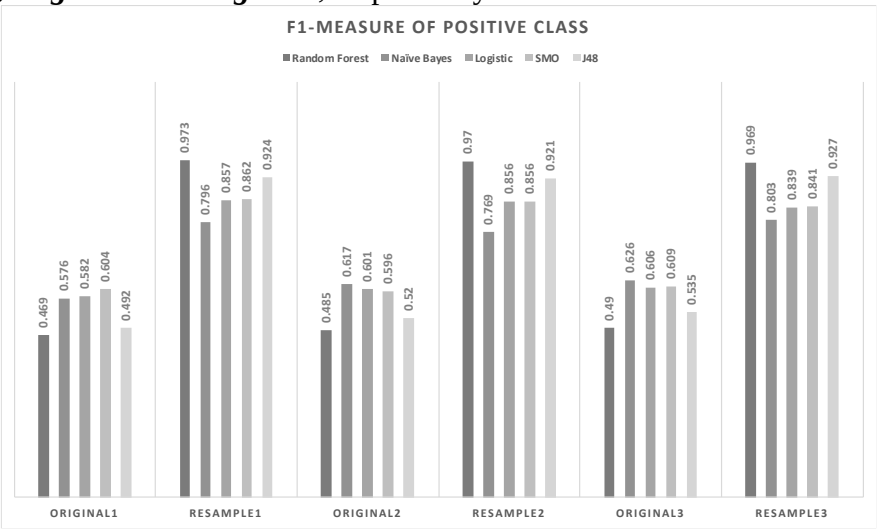

Fig 11 Values of PS student's F1-measure by five classification methods carried on ***original1,2,3*** and ***resample1,2,3***

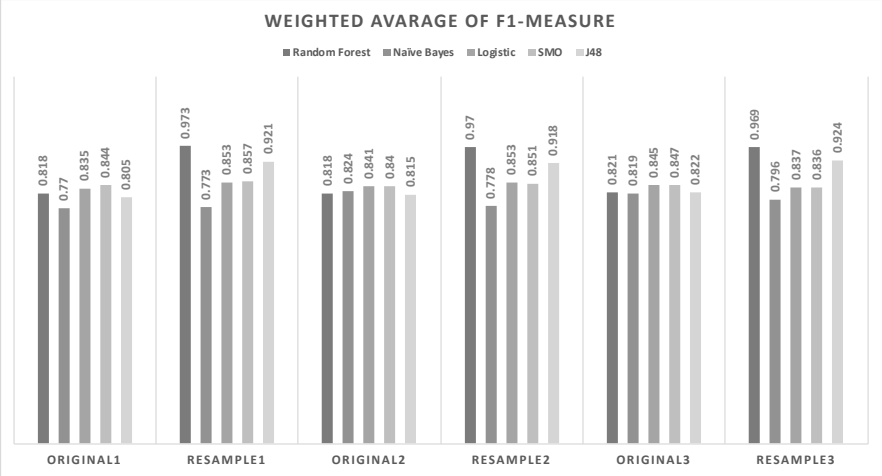

Fig 12 weighted average values of F1-measure by five classification methods carried on ***original1,2,3*** and ***resample1,2,3***

The results of Experiment #3 show that the classification models based on Resampling and Random Forest performed stably and best among other four classification algorithms. The values of F1-measure of PS students and the weighted average of the F1-measure are above 0.96, as shown in Figures 11 and 12, respectively.

Brief introduction to our big data platform and Support and analysis system for Student development.

The big data platform for real-time monitoring of university teaching quality and students learning activities has been introduced to improve university management and teachers' teaching performance. This platform collects data of 2,244 courses, 209,066 classes and 203 specialties. It also collects data about 21,216 students in relation to their class work for four-year undergraduate programs. Since 2011, 520 million records have been stored, including 149GB structure data, such as those of teaching management, teaching process and students' learning, and 200TB of non-structure data, such as audio-visual recordings for classroom teaching, and teaching assessment texts, etc. A 3-step (categorical-assessment, multi-criterion-ranking and results-fusion) assessment model and algorithm has been developed to achieve precision assessment of class-teaching quality. Driven by technology, the big data platform has innovated university management. In particular, a teaching supervision system driven by precision identification of problems and a student development assistant analysis subsystem have been developed. The methods proposed in this paper are integrated in the support and analysis system for student development.

Based on the data collected by the big data platform, the support and analysis system for student development employs a classification model of PS students, an abnormal behavior detection model for students, and a social relationship discovery model, which provide effective support for early warning of abnormal behavior and personalized support. The subsystem realizes five functions, including student's electronic archives, comprehensive early warning, behavior portrait, accurate financial supporting and tracking, and target group analysis. The five functions are described as subsections.

5.1 Student's Electronic profile

Students' electronic profile include students' personal information, course information, academic achievement, book borrowing, individual awards, campus activities and internships. It covers almost all aspects of student's campus life.

5.2 Comprehensive Early Warning

The comprehensive early warning includes four parts: emergency situation early warning, behavior early warning, academic early warning and behavior mutation early warning. Its interface is shown in Fig13.

Fig 13 Interface of Comprehensive Early Warning in student development assistant analysis subsystem

In addition, psychological abnormality early warning of PS students is a part of behavior mutation early warning. By analyzing students' behavior, this paper proposes several indicators to evaluate the psychological state of students and gives an early warning of psychologically abnormal student in the system. The features and evaluation indicators are shown in Table 9.

Table 9 Evaluation indicators of Psychological state assessment

Subclass	**Assessment content**	**indicators**
life	diet	weekly/monthly diet frequency
		weekly/monthly expense in canteen
	exercise	weekly/monthly number of entry and exit in sports venues
study	degrees of learning diligence	weekly/monthly number of entry and exit in library
		Class check-in Number

As shown in Table 8, if more than 60% features' values (e.g. diet and exercise, etc.) have a significant change in a subclass (i.e. life and study), it will be defined as subclass exception. When more than one of the subclasses becomes exception, the class of social functions is defined as abnormal, and if the abnormal state lasts for two weeks, then a warning is issued.

5.3 Student's behavior portrait

Behavior portrait includes individual behavior portrait and student group portrait, it describes the characteristics of each student or student groups in terms of academic performance, pattern of learning, resting, social interaction, and consumption behavior. The features extracted in this paper (consumption, campus activity, location and price preference and consumption of diverse items) lay a foundation for the implementation of this function. Fig 14 shows the interface of this function.

Fig 14 Interface of Behavioral Portraits in student development assistant analysis subsystem

5.4 Accurate financial supporting and tracking

Accurate financial supporting and tracking includes identification and tracking of PS students and individualized quantification of the amount of subsidy. The classification model of PS students proposed in this paper supports this function. In addition, a method of quantization of subsidy is proposed. The method divides students into different groups according to their grade, gender, major and then calculates the average monthly spending of ordinary students in each group, and the PS students in the group will be subsidized by a dynamic appropriate amount, according to the average spending in the group. Fig 15 depicts the interface of this function.

Fig 15 the interface of Accurate financial supporting and tracking in student development assistant analysis subsystem

5.5 Target Group Analysis

This function analyzes the behavior and academic performance of students to identify the outstanding students, isolated students and problem students, and provide relevant information to support counselor's work.

The student development assistant analysis subsystem has served about 17,000 students. The system has provided more than 2,000 consulting services and conducted 124 crisis

interventions until now. Moreover, this subsystem and its analysis results provide us some new insights.

- From the perspective of identifying and supporting PS students this paper puts forward some new features to help identify PS students. These features also provide college personnel with a new perspective of PS students, such as PS students spend more time on campus during vacation compared to ordinary students, ordinary students have more diversified purchasing choices in school than PS students, and PS students seldom buy dessert and fruit. According to these findings, college administrators have launched targeted funding activities, such as the distribution of holiday canteen subsidy vouchers and fruit vouchers to students. At the same time, the data and results analysis of other functions in the platform also provide new ways for identifying PS students. For example, based on the function of the Student's behavior portrait, we can learn the relationship between the characteristics of students' social networks and their economic conditions. We will focus more on these topics in the future study.
- From the perspective of reducing workload and improving work quality. The way of identifying PS students is changed from manual examination and verifying based on proof materials and written applications submitted by students to the combination of big data analysis suggestions and students' online applications, which eliminates the tedious process of written material examination and approval. If the suggestion of big data platform is consistent with students' application, it will automatically approve the application, and only the application with inconsistent results will be manually checked, which improves the work efficiency. After the application of the platform, the number of applications to be audited by staff has been reduced from about 4,200 per year to about 630 per year, so the workload has been reduced by about 80%, and the working time has been reduced by about two weeks at least. In addition, under the condition of keeping the number of staff unchanged, this platform meets the target of expanding the coverage of student supporting work from the students who submitted applications (about 22% of all students) to all students, and achieves full coverage tracking and analysis of the finance situation of all students. It has improved the efficiency and quality of student management and reduced the workload of college staff.

Conclusion

In presence of the excellent campus network, extensive application of intelligent terminals and the big data platform at Xi'an Jiaotong University, this paper reports on an attempt of making use of big data collected from smart campus to study three problems related to PS students and overcome the difficulties in identifying and supporting PS students.

- Five kinds of features are extracted to identify PS students, including personal and family situation, consumption, campus activity, location and price preference and consumption of diverse items. The effectiveness of these features has been proved by experimental result obtained.
- Four imbalanced data processing methods (Subsampling, Resampling, Cost-sensitive learning and SMOTE) have been applied to produce four different experimental datasets, and five classification algorithms (Random Forest, J48, Naïve Bayes, SMO, Logistic regression) are used to train the classification model on each dataset correspondingly. According to experimental results, the model based on Resampling and Random Forest achieves the best performance in the F1-measure of PS students, among the combinations of four imbalanced processing methods and five classification algorithms.

- Based on the results of pedagogy and psychology research for students with family financial difficulties, this research systematically utilizes all kinds of students' campus big data, and extracts implicit features related to students' financial status and psychological state to overcome the problem of low recognition performance caused by the difficulty of using massive recorded data and the lack of outside of school consumption data. The results of 10-fold cross-validation show that the proposed new features can increase the F1-measure of PS students by 8.26%. At the same time, the imbalanced data processing methods effectively improve the performance of classification model for PS students. The model obtained by using Resampling method and Random Forest algorithm can achieve optimal performance, as the F1-measure of PS students is stable at 0.96, which is effectively improved.
- A method of quantization of subsidy, and strategies of early warning and counseling for students have been presented. Based on the above methods, a system was developed, which meets the individualized needs and provides diversified support for PS students.

The methods and system proposed in this paper have been applied to the support and analysis system for student development of big data platform for real-time monitoring of university teaching quality. The developed system severed more than 17,000 students, and its application has also been reported in the media [26].

Acknowledgement
This work is supported by National Key Research and Development Program of China (2018YFB1004500), National Nature Science Foundation of China (61877048，61472315), Innovative Research Group of the National Natural Science Foundation of China (61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Center for Engineering Science and Technology. Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China".

Reference:
[1] General Framework of Smart Campus, China National Standard,2018.6
[2] Conijn R, Snijders C, Kleingeld A, et al. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS[J]. IEEE Transactions on Learning Technologies, 2017, 10(1):17-29.
[3] Ashwin S, Gayathri R, Mining Student data by Ensemble Classification and Clustering for Profiling and Prediction of Student Academic Performance[C]. 2016 ASEE Mid-Atlantic Section Conference
[4] Hu Y H, Lo C L, Shih S P. Developing early warning systems to predict students' online learning performance[J]. Computers in Human Behavior, 2014, 36:469-478.
[5] Qu S, Li K, Zhang S, et al. Predicting Achievement of Students in Smart Campus[J]. IEEE Access, 2018:1-1.
[6] Cao Y, Gao J, Lian D, et al. Orderliness predicts academic performance: behavioral analysis on campus lifestyle[J]. Journal of The Royal Society Interface, 2018, 15(146).
[7] Bordea G, Shahiri A M, Husain W, et al. A Review on Predicting Student's Performance Using Data Mining Techniques[J]. Procedia Computer Science, 2015, 72:414-422.
[8] Ding D, Li J, Wang H, et al. Student Behavior Clustering Method Based on Campus Big Data[C]// International Conference on Computational Intelligence & Security. IEEE Computer Society, 2017.

[9] Liu T, Yang L, Liu S, et al. Inferring and analysis of social networks using RFID check-in data in China[J]. Plos One, 2017, 12(6): e0178492.
[10] Liu Y, Hu M, Lu X. Social Frequency Analysis of University Students via Digital Campus Cards[C]// International Conference on Intelligent Human-machine Systems & Cybernetics. IEEE, 2016.
[11] Fan S, Li P, Liu T, et al. Population Behavior Analysis of Chinese University Students via Digital Campus Cards[C]// 2015 IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, 2016.
[12] China Student Aid Development Report 2016.
http://www.moe.edu.cn/jyb_xwfb/xw_fbh/moe_2069/xwfbh_2017n/xwfb_170228/170228_sfcl/201702/t20170228_297543.html
[13] Holmes G, Donkin A, Witten I H. WEKA: a machine learning workbench[C]// Conference on Intelligent Information Systems. 2002.
[14] Wenjuan W. Study on the College Students' Consumption technical route and instance Based on the Information from Smart Card[D]. Dalian Medical University,2013.
[15] Hongwei M. The mining and analysis of consumption data of campus card[D]. Shanxi Normal University, 2016.
[16] Suo L, Gong J. Identification of University Poor Students Based on Data Mining[C]// International Conference on Intelligent Computation Technology & Automation. IEEE, 2016.
[17] Jiang, Tongjun & Cao, Jianmei & Su, Dan & Yang, Xianglai. (2017). Analysis and Data Mining of Students' Consumption Behavior Based on a Campus Card System. 58-60. 10.1109/ICSCSE.2017.22.
[18] Lin Z. A Design of Poor Students Identification System in Higher Learning Institutions Based on Differential Privacy[J]. Computing Technology and Automation, 2017(3).
[19] Mingjun L. Research on the method of determining Poor Scholarship on data mining [D]. Center China Normal University,2017.
[20] Huaifeng Q. How to Identify the Needy College Students against the Background of Big Data [J]. Journal of College Advisor, 2016, 8(5):74-77.
[21] Fuqiang G. On Current Conditions and Solutions for Consumption Behavior and Consumption Viewpoint of Poor Students of Universities [J]. JOURNAL OF CHINA YOUTH UNIVERSITY FOR POLITICAL SCIENCES, 2007, 26(1):20-24.
[22] Chaowen W, Jing D, Yannan S. Research on the Targeted Poverty Reduction Model of the Needy Undergraduates in the Big Data Environment[J]. Heilongjiang Researches on Higher Education, 2016(12):41-44.
[23] Yan W, Chunping W, Yanchao S, et al. A research on the related factors affecting the subjective well-being of the poor medical university students[J]. CHINESE JOURNAL OF BEHAVIORAL MEDICAL SCIENCE, 2006, 15(1):67-69.
[24] Desheng K. A study of the relationship between the personality and academic achievements of the impoverished college students[J]. CHINESE JOURNAL OF BEHAVIORAL MEDICAL SCIENCE, 2005, 14(2):126-127.
[25] Guidance on Identifying poverty-stricken students from Six Departments, such as the Ministry of Education.
http://www.moe.gov.cn/srcsite/A05/s7505/201811/t20181106_353764.html
[26] http://tv.cctv.com/2018/08/17/VIDEBGl5rFTNS8prvA7YP0d2180817.shtml
[27]Han J., Kamber M., Data Mining Concept and Techniques, The Morgan Kaufmann. 2th Edition, 2006.
[28] Feng T, Fan W, Xiang F, Nazaraf S, Qinghua Z, Yuanyuan W. Improveing generalization ability of instance-transfer based imbalanced sentiment classification of turn-level interactive Chinese texts, Service Oriented Computing and Application, DOI:10.1007/s11761-019-00264-y,(accepted)

Appendix

Table I. Examples of Original Records of Campus Consumption Data

Student ID	Name	TIME	Operation	Expend	Balance	Consumption type	POS No.	POS location
**	*	2016-02-21 17:10:13	Consumption	-600	17250	1	24	**
**	*	2016-02-21 17:10:15	Consumption	-100	18928	2	20	**
**	*	2016-02-21 20:32:42	Consumption	-250	3740	1	9	**
**	*	2016-02-21 20:33:25	Consumption	-1270	27	1	9	**
**	*	2016-02-21 20:33:35	Consumption	-400	4448	1	2	**
**	*	2016-02-21 20:33:38	Consumption	-1550	16313	3	9	**
**	*	2016-02-21 20:33:51	Consumption	-670	20490	2	11	**
**	*	2016-02-21 20:34:11	Consumption	-800	5421	2	24	**
**	*	2016-02-21 20:34:20	Consumption	-7990	21982	1	11	**
**	*	2016-02-21 18:32:14	Consumption	-2100	6172	1	9	**

Table II. Examples of Original Records of Access Record Data

Student ID	NAME	visit time	Gate No.	Operation
**********	***	2008-01-02 08:34:57	02	exit
**********	***	2008-01-02 08:35:00	01	exit
**********	***	2008-01-02 08:35:01	02	exit
**********	***	2008-01-02 08:35:05	06	exit
**********	***	2008-01-02 08:35:05	02	exit
**********	***	2008-01-02 14:59:22	05	enter
**********	***	2008-01-02 14:59:29	06	enter
**********	***	2008-01-02 14:59:33	02	enter
**********	***	2008-01-02 14:59:34	06	enter
**********	***	2008-01-02 08:59:03	01	exit

Table III. Data items of Students' basic Information Data

student ID	name	province	subject type	gender	birthday	institute name	Major	Dormitory	class	political stat	Ethnic group	Examinee ty	Home address	County and c	Enrollment Y
2********0	Ge**	Shanghai	Science and Engine	male	1998/1/21	College of Electrical	Electrical	Nanyang Shuyuan	Electrical 68	Communist Y	Han	urban	********	Jinshan Distr	2016
2********1	Liu**	Tianjin	Science and Engine	femle	1998/3/29	College of Electrical	Electrical	Nanyang Shuyuan	Electrical 68	Communist Y	Han	urban	********	Jinghai Distri	2016
2********9	Zhang**	Shandong	Liberal arts	femle	1999/2/27	College of Humanities	Literature	Nanyang Shuyuan	Literature Ex	Communist Y	Han	rural	********	Changle Dist	2016

Table IV. Data items of Students' Family Situation Questionnaires Data

student ID	name	family members (und	family members (1	family meml	family meml	PARENT_ALIVE	DIVORCE	FATHER_HOSPITAL	FATHER_ILL	MATHER_HO	MOTHER_ILL	RELATIVE_HI	FATHER_EDUCATION	FATHER_JOI	MATHER_EDI	MATHER_JOB	
2********0	Ge**	1	5	6	2	both	no	yes, once	no	no	no	yes,once	High School/Technical Seco	Individual bu	middle schol	Individual businessmen	
2********1	Liu**	0	3	3	0	both	no	no	no	no	no	yes, more th	Undergraduate	Cadres of Sta	Undergradua	Private entrepreneurs	
2********2	Xie**	1	2	3	0	both	no	no	no	no	no	no	postgraduate	Cadres at or	Undergradua	famer	
		SIBILLING	SIBLING_QUANTITY	LEADER	TECHNICIAN	MANAGER	BOSS	NONE	LEADER1	TECHNICIAN	MANAGER1	BOSS1	NONE1	SAVING	DEBT	CAR	ECONOMIC
		no	0	no	no	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	Middle and l
		yes	0	yes	no	no	no	yes	yes	yes	no	no	yes	yes	no	yes	Middle and l
		yes	0	no	no	no	no	no	no	no	yes	no	yes	yes	no	yes	Middle and l

Table V. List of all 135 extracted features in 4.2

No.	Personal/Family feature	Consumption feature		Consumption of diverse items features	Location and price preference features	Campus activity features
1	gender	monthly expense in canteen	**mean-daily breakfast expense**	**monthly consume location num**	**morning top1 item type**	**median-store consume time**
2	number of family members(4)	monthly expense in store	**var-daily breakfast expense**	**monthly consumption item type num**	**price preference of breakfast**	**var-store consume time**
3	father's education level	monthly consume times in canteen	**max-daily breakfast expense**	**monthly breakfast location num**	**noon top1 item type**	**median-morning consume time**
4	mother's education level	monthly consume times in store	**min-daily breakfast expense**	**monthly breakfast type num**	**price preference of lunch**	**var-morning consume time**
5	father's job type	mean-every purchase expense in canteen	**mean-daily lunch expense**	**monthly lunch location num**	**afternoon top1 item type**	**median-noon consume time**
6	mother's job type	var-every purchase expense in canteen	**var-daily lunch expense**	**monthly lunch type num**	**price preference in afternoon**	**var-noon consume time**
7	family deposit	mean-every purchase expense in store	**max-daily lunch expense**	**monthly location num of consumption in afternoon**	**night top1 item type**	**median-afternoon consume time**
8	family debt	var-every purchase expense in store	**min-daily lunch expense**	**monthly type num of consumption in afternoon**	**price preference of dinner**	**var-afternoon consume time**
9	family economic level	mean-daily expense in canteen	**mean-daily afternoon expense in canteen**	**monthly dinner location num**	**school aceeee day num**	**median-night consume time**
10	**parent alive**	mean-daily expense in store	**var-daily afternoon expense in canteen**	**monthly dinner type num**	**library day num**	**var-night consume time**
11	**parent divorce**	var-daily expense in canteen	**max-daily afternoon expense in canteen**	**median-daily breakfast location num**	**monthly library count**	**number of weekdays consumption in the campus**
12	**physical condition of parents(**	var-daily expense in store	**min-daily afternoon expense in canteen**	**median-daily breakfast type num**	**monthly canteen price tendency**	**weekday store consume day num**
13	**sibling condition(2)**	max-every purchase expense in canteen	**mean-daily dinner expense**	**median-daily lunch location num**	**morning canteen price tendency**	**number of weekends consumption in the campus**
14	**family positon level(10)**	min-every purchase expense in canteen	**var-daily dinner expense**	**median-daily lunch type num**	**noon canteen price tendency**	**weekend store consume day num**
15	**household car**	max-daily expense in canteen	**max-daily dinner expense**	**median-daily location num of consumption in afternoo**	**afternoon canteen price tendency**	**median-weekday store consume time**
16	**province**	min-daily expense in canteen	**min-daily dinner expense**	**median-daily type num of consumption in afternoon**	**night canteen price tendency**	**var-weekday store consume time**
17	**subject**	max-daily expense in store	**mean-weekday daily expense in canteen**	**median-daily dinner location num**	**weekday accedd day num**	**median-weekend store consume time**
18	**major**	min-daily expense in store	**var-weekday daily expense in canteen**	**median-daily dinner type num**	**weekday library day num**	**var-weekend store consume time**
19	**ethnic groups**	**median-daily consume times in morning**	**max-weekday daily expense in canteen**		**weekend accedd day num**	**median-weekday morning canteen consume time**
20	**graduation type**	**median-daily consume times in noon**	**min-weekday daily expense in canteen**		**weekend library day num**	**var-weekday morning canteen consume time**
21		**median-daily consume times in afternoon**	**mean-weekend daily expense in canteen**		**weekday canteen price tendency**	**median-weekend morning canteen consume time**
22		**median-daily consume times in night**	**var-weekend daily expense in canteen**		**weekend canteen price tendency**	**var-weekend morning canteen consume time**
23		**min-weekend every purchase expense in store**	**max-weekend daily expense in canteen**		**tendency of food type (1)**	**median-weekday noon canteen consume time**
24		**mean-weekend daily expense in store**	**min-weekend daily expense in canteen**			**var-weekday noon canteen consume time**
25		**var-weekend daily expense in store**	**var-weekday every purchase expense in canteen**			**median-weekend noon canteen consume time**
26		**mean-weekday daily expense in store**	**var-weekend every purchase expense in canteen**			**var-weekend mooncanteen consume time**
27		**var-weekday daily expense in store**	**var-weekday every purchase expense in store**			**median-weekday afternoon canteen consume time**
28		**var-weekend every purchase expense in store**	**max-weekday every purchase expense in store**			**var-weekday afternoon canteen consume time**
29		**max-weekend every purchase expense in stor**	**min-weekday every purchase expense in store**			**median-weekend afternoon canteen consume time**
30						**var-weekend afternoon canteen consume time**
31						**median-weekday night canteen consume time**
32						**var-weekday night canteen consume time**
33						**median-weekend night canteen consume time**
34						**var-weekend night canteen consume time**
35						**weekday_weekend canteen single cost mean compar**
36						**weekday_weekend canteen daily cost mean compare**
37						**weekday_weekend store single cost mean compare**
38						**weekday_weekend store daily cost mean compare**

* the newly proposed features in this paper are in bold, and the features in plain text has been taken from Refs. [16-20]

Table VI. The description of 81 selected features

Category	Feature	Feature Description
Personal/Family feature	gender	Male/Female
	number of family members (4)	Age structure of family members (four features: under 18, 18 to 60, 60 to 65, upper 65).
	father's education level	feature values: not attending school, primary school or below, junior high school, senior high school/secondary school/technical school, college, undergraduate, graduate
	mother's education level	feature values: not attending school, primary school or below, junior high school, senior high school/secondary school/technical school, college, undergraduate, graduate
	father's job type	There are 11 job types.
	mother's job type	There are 11 job types.
	family deposit	Yes/No
	family debt	Yes/No
	family economic level	There are 5 levels.
	parent alive	feature values: they are all alive, the father is dead, the mother is dead, both dead.
	parent divorced	whether the parents divorced
	physical condition of parents (5)	physical condition of parents
	sibling condition (2)	the children's number in the family
	family position level (10)	family member's work position
	household car	whether the family has car
	province	/
	subject	/
	major	/
	ethnic groups	/
	graduation type	/
Consumption features	monthly expense in canteen	the total cost that student spend in canteen during a month
	monthly expense in store	the total cost that student spend in store during a month
	monthly consume times in canteen	number of times that students spend in canteen during a month
	monthly consume times in store	number of times that students spend in canteen during a month
	mean-every purchase expense in canteen	the average of the cost of every spend in canteen
	mean-daily expense in canteen	the average of daily cost in canteen
	var-every purchase expense in canteen	the variance of the cost of every spend in canteen
	var-daily expense in canteen	the variance of daily cost in canteen
	max-daily expense in canteen	the maximum of daily cost in canteen during a month
	mean-daily breakfast expense	the average of the cost of breakfast during a month
	mean-daily lunch expense	the average of the cost of lunch during a month
	mean-daily afternoon expense in canteen	the average of the cost of in afternoon during a month
	mean-daily dinner expense	the average of the cost of dinner during a month
	max-daily dinner expense	the maximum of the cost of dinner during a month
	var-daily dinner expense	the variance of the cost of dinner during a month
	var-weekday daily expense in canteen	the variance of the total cost in weekday in canteen
	max-weekday daily expense in canteen	the maximum of the total cost in weekday in canteen
	max-weekday every purchase expense in canteen	the maximum of the cost of every spend in canteen during weekday
	var-weekday every purchase expense in canteen	the variance of the cost of every spend in canteen during weekday

Consumption of diverse items features	**monthly consumption item type num**	the number of item types that a student buying during a month
	median-daily lunch location num	the median of number of locations that a student buying during lunch during a month
	median-daily lunch type num	the median of number of item types that a student buying during lunch during a month
	median-daily breakfast location num	the median of number of locations that a student buying during breakfast during a month
Location and price preference features	**tendency of food type (1)**	the food preference during a month
	noon top1 postype	the food type preference of lunch
	price preference of lunch	the food price preference of lunch
	afternoon top1 item type	the food type preference in the afternoon
	night top1 item type	the food type preference at night
	price preference of dinner	the food price preference of dinner
	monthly canteen price tendency	food price preference during a month
	noon canteen price tendency	the food price preference at noon
	weekday canteen price tendency	food price preference in weekdays
Campus activity features	**monthly canteen consume day num**	the number of days that student consumes in canteen during a month
	monthly store consume day num	the number of days that student consumes in store during a month
	monthly morning consume day num	the number of days that student consumes in the campus in the morning during a month
	monthly noon consume day num	the number of days that student consumes in the campus at noon during a month
	monthly afternoon consume day num	the number of days that student consumes in the campus in the afternoon during a month
	monthly night consume day num	the number of days that student consumes in the campus at night during a month
	weekday_weekend canteen single cost mean compare	the comparation between every cost in canteen in weekday and weekend
	weekday_weekend canteen daily cost mean compare	the comparation between daily cost in canteen in weekday and weekend
	weekday_weekend store single cost mean compare	the comparation between every cost in store in weekday and weekend
	weekday_weekend store daily cost mean compare	the comparation between daily cost in store in weekday and weekend
	number of weekdays consumption in the campus	the number of weekdays that student consumes in the campus during a month
	number of weekends consumption in the campus	the number of weekends that student consumes in the campus during a month