EDM Final Assignment

Deji Huaqu, Huancheng Xu, Guotai Sun, Siyuan Gu, Yuetong Lyu

HUDK 4050

Professor Lukas Liu

December 21, 2021

In recent decades, the world has undergone earth-shaking changes and the global economy has developed rapidly. At the same time, there has been obvious progress in the investment and development of education by the U.S government.

Additionally, with the wide application of information technology in the field of education, educational data is also explosively increasing. Data scientists find new characteristics and rules of education to improve decision-making level and educational quality by exploring and analyzing educational data through data mining technology.

As one of the indexes of higher education, graduation rates are considered to be highly related to education equality. According to the American Council on Education, the graduation rate reflects the quantity and fundamental quality of qualified students in a certain period of education or school based on a specified target and within the specified time limit. Dropout rates and repeat rates in a given year are inversely proportional to on-time graduation rates; the efficiency of using educational resources and the on-time graduation rate is inversely proportional to the number of educational resources per student in a given period of time. Therefore, it is a paramount index to measure the equality of education and the efficiency of using educational resources in a region or a school.

In this project, we will use several data mining methods to trace back the major causes of graduation rate disparity and reveal the educational policy shortcomings behind them.

For the data selection part. More affluent families will have more learning planned for their children, and they will have access to more educational resources and opportunities to practice. So we need to improve the overall education level of all people based on analyzing big data to compensate for the inequality of education resources. Compared with the traditional education data, the collection of education big data has stronger real-time, coherence,

comprehensiveness, and naturalness, it is the more complex and diverse analysis and processing, and in-depth applications. For more equitable distribution of educational resources, we need diagnostic, personalized learning analytics based on big data. Based on big data we can improve the quality of education for most people. Big data can collect and analyze all aspects of behavior records of administrators, parents, teachers, and students. And it can provide better services for learners, teachers, parents, and others. The comprehensive collection, accurate analysis, and rational use of big data in education have become a driving force for schools to improve their service capabilities, form data to speak, make decisions, and manage with data, and use data to carry out accurate services. The data structure of education big data is more Complex. Conventional structured data (such as grades, academic records, employment rates, attendance records) will remain important, but unstructured data (such as the variety of classroom activities, videos, lesson plans, teaching software, learning games) will increasingly dominate. So we not only collect the structured data, but we also need to collect more unstructured data.

For data manipulation. For more equitable distribution of educational resources, we need diagnostic, personalized learning analytics based on big data. Based on big data we can improve the quality of education for most people and the quality of education. The data collection phase will establish the association of different concepts in the learning content, then integrate the categories, learning objectives, and student interactions. And then the data will be processed by the model calculation Algorithm. (We will talk about it later) The inference phase analyzes the collected data through a testing Algorithm, then to a feedback report phase, and the results of the analysis are provided to the recommendation phase for personalized learning recommendations. The suggestion phase provides learning recommendations for teachers and students through the suggestion Algorithm and predictive analytic Algorithm.

Model 1: Group--- K means clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. A cluster refers to a collection of data points aggregated together because of certain similarities. You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. Mathematically, given a set of observations (x_1, x_2, ..., x_n), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* ($\leq n$) sets S = {S_1, S_2, ..., S_k} so as to minimize the within-cluster sum of squares. Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculations to optimize the positions of the centroids until the centroids have stabilized — there is no change in their values because the clustering has been successful. The basic steps are:

- Find potential influential factors which could increase/decrease college graduation rates
- Data clean - make a new table with those essential factors

- Decide the number of K, it always starts from 1 and then performs iterative calculations until the positions of centroids are optimized with the nearest mean.
- Report the optimal K-value and make a conclusion

Firstly, if a college has a lower ADM_RATE, it only picks the greatest students, who will have higher graduation rates in the future. Also, if students get a higher SAT grade, which proves their hard work, they will also be more likely to learn better in their college, resulting in a higher graduation rate. What's more, CCBASIC, CCSIZSET, I believe these two variables stand for the class management of a campus, which is very related to the quality of education, and proves the graduation rates.

Secondly, data clean. After that, we pick ADM_RATE, SAT_AVG_All, CCBASIC, and CCSIZSET as four potential influencers, and then produce a new cleaned data set. new_CollegeScorecard <- select(CollegeScorecard, `UNITID`,`INSTNM`, `AD M_RATE` ,`SAT_AVG_ALL`,`CCBASIC`,`CCSIZSET`) CollegeClean <- new_CollegeScorecard %>% na.omit() %>% select(-UNITED,-INSTNM)

Then, run K-means clustering. K - means clustering, with k = 1,2,3,...etc. CollegeKCL <- kmeans(CollegeClean, centers = K) disKCL = dist(CollegeClean)^2 silKCL = silhouette(CollegeKCL$cluster, disKCL) fviz_silhouette(silKCL)

The goal of the plan is to help researchers do a further study based on those factors, which put colleges into high graduation rate groups.

Model 2: Logic classification--- Naïve Bayes' model

The second model is called the Naïve Bayes classification model. It is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. It has high accuracy and speed on large binary datasets.

Here Naïve means that we have two assumptions on our dataset.

Independent---We assume that no pair of features are dependent. It means that our elements do not correlate with each other. Hence, the features are assumed to be independent. Equal---Secondly, each feature is given the same weight (or importance). None of the attributes is irrelevant and assumed to be contributing equally to the outcome.

Then why do we use Naïve Bayes' model? Because they are extremely fast for both training and prediction and they provide a straightforward probabilistic prediction. They are often very easy to explain. Finally, they have very few (if any) tunable parameters.

Here is a basic formula of Naïve Bayes' theorem

$$P(L \mid features)=P(features \mid L)*P(L)/P(features)$$

- P(L): the probability of hypothesis L being true (regardless of the data). This is known as the prior probability of L.
- P(features): the probability of the data (regardless of the hypothesis). This is known as the prior probability.
- P(L|features): the probability of hypothesis h given the data D. This is known as posterior probability.

- P(features|L): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

Our procedure could be described as below:

1. Grab data from each college and transfer the graduation rate into binary variables (for example, we set a value to classify the rate as high or low, let's say 0.5, then if the rate is greater than 0.5, we identify it as HIGH graduation rate; if it's <0.5, we say it's LOW graduation rate)

2. Pick potential factors that may influence the graduation rate, such as crime rate, average SAT score, whether located in a big city and so on. Then do the binary transformation like step 1.

3. Import the package and train the model by fitting dependent and independent variables.

4. Compare the predicted outcome to the actual dataset to evaluate our model.

5. Check the accuracy of each factor, pick those with relatively high accurate predictions to make a further move (like applying for a more complex prediction model)

Our goal is to wipe out the low- or non-correlated elements to our target outcome (graduation rate). So, we don't care about the accuracy ratio but just use them to do the identification(classification) for more complex correlation prediction. The key is that we aim to do a quick estimation here.

To get deeper, we may also find out the potential correlation between possible dependent variables and independent variables other than graduation rate which would like to cause education equality. Such as the tuition fee per semester, the overall rating of educational

facilities, the average salary of graduates, and so on. If we got a relatively high coefficient between the DVs and RVs, we should regard it as a causal effect and add that causation into our output table. Once we finish analyzing all possible causation of independent variables, we can draw some conclusions about key factors that influence education equality.

Model 3: PCA & Logistic Regression

Now we want to try to use Principal Component Analysis (PCA) combined with logistic regression to approach this problem. PCA is a fast and flexible unsupervised method for dimensionality reduction in data; It tries to preserve the essential parts that have more variation of the data and remove the non-essential parts with fewer variation. PCA serves many purposes in data analysis, but in this part we mainly utilize PCA to decrease the number of variables and make further analysis simpler. Logistic Regression is a classification algorithm that is used to predict the probability of a categorical dependent variable. Here we choose to utilize it since it is relatively fast and uncomplicated, quite convenient to interpret the results and can be applied to multiclass problems. What follows is the detailed procedure:

- Preprocess data:

Data preparation is the first and also the most important step. We need to go through four sub-steps: (1) Select the variables that may help us identify the inner relationships between educational resources and graduate rate, such as average faculty salary(AVGFACSAL), the average annual cost of attendance for academic year institutions (COSTT4_A), the percentage of undergraduate students whose family incomes range is relatively low (INC_PCT_LO), etc; (2) Get rid of all missing data. As we tend to choose public reports datasets, some selected information are withheld or removed to protect the identities, privacy, and personal information of individual students, teachers, or administrators. It's vital to get rid of all incomplete

information to make sure our dataset is clean; (3) Transfer categorical variables into continuous variables; (4)Scale the data. The larger the variable, the more possibilities there could be. Also, as PCA transformation is sensitive to the relative scaling of the original variables so that our data ranges need to be scaled before applying PCA.

- Reduce the dimensionality of variables and decide the number of PCA components based on the explained variance:

As a result of the PCA transformation, the first principal component has the largest possible variance; each succeeding component has the highest possible variance under the constraint that it is orthogonal to the preceding components. Keeping only the first m < n components reduces the data dimensionality while retaining most of the data information in the data.

- Create and define the classification model:

The logistic regression function:

$$y = 1 / (1 + \exp(-f(\mathbf{x}))$$

Predict variable:
y — whether a student would drop out from college? (binary: "1", means "Yes", "0" means "No")

- Model training: determining the coefficients b_0, b_1, …, b_r that correspond to the best value of the cost function
- Select the most appropriate model
- Model Evaluation — Confusion Matrix

True positives (TP): Students predicted to drop out of college, that actually drop out of college.

True negatives (TN): Students predicted to not drop out of college, that actually not drop out of college.

False positives (FP): Students predicted to drop out of college, that actually not drop out of college.

False negatives (FN): Students predicted not to drop out of college, that actually drop out of college.

Overall, this approach can be used to predict whether a student would drop out of college based on current educational resources. And we can put forward more specific suggestions and improvement measures to ensure students have a complete and meaningful college experience.

We searched for similar cases and one was very close to our idea. And they used a more suitable method, and we can refine our project based on the data and methods they provided. The researchers implemented the Center for Educational Leadership's (CEL) 5 Dimensions of Teaching and Learning™ instructional framework and the accompanying 5D+™ Rubric, a teaching model and framework that helps schools and districts successfully develop high-quality education (Tack, 2017). This model not only develops strategies for critical thinking, literacy, and math skills but also prepares for future career paths and fundamentally addresses some of the inequalities caused by poverty. In conclusion, no matter what model we use, the ultimate goal of data mining is to build better educational policies based on the results to benefit society. More specifically, high quality teaching, learning and equity for all is the mission.

Cite

“How my school district improved graduation rates with equity” Mary B. T.

From

https://blog.k-12leadership.org/instructional-leadership-in-action/how-kelso-school-district-improved-graduation-rates-with-equity