DATA MINING GROUP PROJECT

THE RELATIONSHIP BETWEEN FUTURE INCOME AND THE UNIVERSITY ATTENDANCE

Group Members: Yu Zheng

Tiecheng Zhao

Chang Su

Fang Zhou

Phase 1: DESIGN PROPOSAL

Abstract:

With the primary interest to understand what kind of university can cultivate the students who will earn more or less in the future, our group set the mean earnings of students in 10 years as target and other key information of university as input. Then we use the C5.0 node in SPSS to build up the decision tree model and classify the universities into several categories.

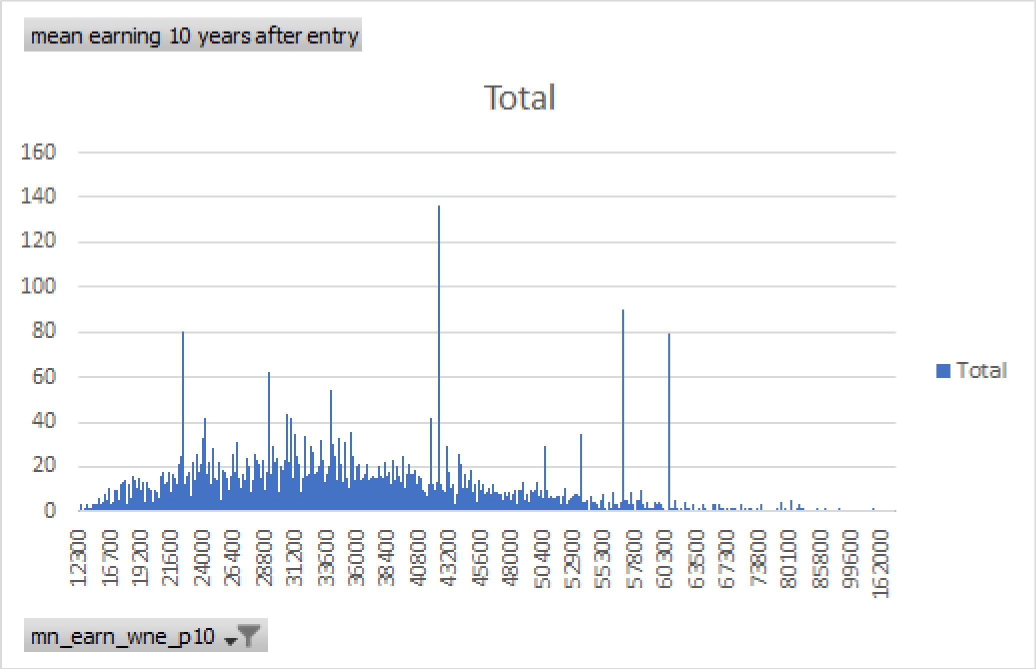
Introduction:

It's no secret that US university students often graduate with debt repayment obligations that far outstrip their employment and income prospects. While it‘s understood that students from elite colleges tend to earn more than graduates from less prestigious universities, the finer relationships between future income and university attended are quite murky. In an effort to make educational investments less speculative, we are trying to identify what kind of universities can cultivate the students who will make a decent income in the future after graduation. The US Department of Education has matched information from the student financial aid system with federal tax returns to create the College Scorecard dataset. With the College Scorecard dataset, we are able to access the information we need, so as to do the project and finally make the returns on higher education more transparent.

Data Description:

The dataset we use is College Scorecard dataset, which is from Kaggle.com and originally created by the US Education Department. The complete dataset includes data from thousands of school in the past twenty years and it’s attached with a data dictionary containing the description of the 1720 dimensions. The object of the dataset is the school, and the 1720 attributes include information about the school, academics, student, cost, completion, admission, aid, earning, repayment and root. For instance, Unit ID, name, zip code and carnegie classification of the school, percentage of degrees awarded in Education, average SAT equivalent score of students admitted, total share of enrollment of undergraduate students who are Hispanic, average cost of attendance, completion rate for first-time, full-time students at four-year institutions, admission rate, share of students who received a federal loan while in school, number of students working and not enrolled 6 years after entry.

About the average earning of student working and not enrolled 10 years after entry in 2011, the minimum earning is 12300, maximum earning:250000, mean earning is 37062, mode is 42300. The meaning earnings of students 10 years after entry is distributed as below:



Problem Statement:

We are interested in the relationship between future income and the university attended. We would like to predict the future income of students from the university attributes.

Methodology:

1. Data collection

To realize the level of graduates' earning and the condition of universities, we use College Scorecard dataset, which is from Kaggle.com and originally created by the US Education Department. This dataset contains data from about 7000 universities and 1720 attributes in the aspects of school, student, cost, academics, admission, aid, earning, repayment and root.

2. Data preprocessing

We found on noisy or inconsistent or intentional data in the original document after checking, but there are bounds of missing values. We can't fill in the missing value manually because we don't know its' true value and for the accuracy of the prediction, we ignore all the tuples that have missing values. After clean, there are 1310 tuples available.

3. Variable selection

Because there are too many variables in the dataset and many of them are redundant and repetitive, we select 11 variables that we are interest in and meaningful, including Admission rate (ADM_RATE), Average SAT equivalent score of students admitted(SAT_AVG), Predominant degree awarded(PREDDEG), Enrollment of undergraduate degree-seeking students(UGDS), In-state tuition and fees(TUITIONFEE_IN), Out-of-state tuition and fees(TUITIONFEE_OUT), Proportion of faculty that is full-time(PFTFAC), Percent of all federal undergraduate students receiving a federal student loan (PCTFLOAN), Percent withdrawn from original institution within 2 years (WDRAW_ORIG_YR2_RT), Percentage first-generation students (PAR_ED_PCT_1STGEN), Control of institution(CONTROL). And target is Mean earnings of students working and not enrolled 10 years after entry(mn_earn_wne_p10).

4. Model building

In order to have a better understanding of the inner interaction among these variables, and to get a straightforward view to the relationship with respect to the attendance and income, a decision tree would be an appropriate method applied on the model building process. Since our objective is to see how the income will be impacted, the field income will be apparently selected as the target value. In the dataset, we can’t set income as the target directly since the variable type of it is continuous from 12300 to 250000. To specify the range of the income, we treat the first 33% of the observations as low income group, the middle 33% as intermediate income group and the top 33% as the high income group. As consequence of this operation, we will be able to see which group of students will earn the high income and which students will earn relatively less income in the future.

In the decision tree model, it is always necessary to divide the data sample into two groups, training and testing respectively. To better build up the model, the division with 50% of training data and 50% of testing data will be used in our decision tree construction. We are going to definitely deal with the training data in the model to see what the predictors are and what the important level of the predictors is. After obtaining the information of predictor importance, we would have an overall evaluation about which attributes will impact student’ s future income, and subsequently, we will be able to ignore those factors which do not influence the income or maybe those which produce tiny impacts.

5. Evaluation

For each model we construct, we would always need to see how accurately the model can predict the data. By introducing two matrices for training and testing data, we will be aware of the accuracy of the model by comparing them together.

Expectation:

Ideally, we expect the decision tree to give us the information that students who are the first generation and who have the high SAT scores will have a high income in the future. For those students whose institution have a higher withdrawal rate may not have a good income.