BYGB/ISGB 7967 (Fall 2017)

DATA MINING FOR BUSINESS

FINAL REPORT

The relationship between future income and the university attendance

Group Members: Yu Zheng

Tiecheng Zhao

Chang Su

Fang Zhou

Date: Dec. 5

Contents

[Abstract 2](#_Toc500255276)

[Introduction 2](#_Toc500255277)

[Problem Statement 3](#_Toc500255278)

[Data description 4](#_Toc500255279)

[Methodology 5](#_Toc500255280)

[Result and Discussion 7](#_Toc500255281)

[Conclusions 9](#_Toc500255282)

[Appendix 1:Contributions of each group member 10](#_Toc500255283)

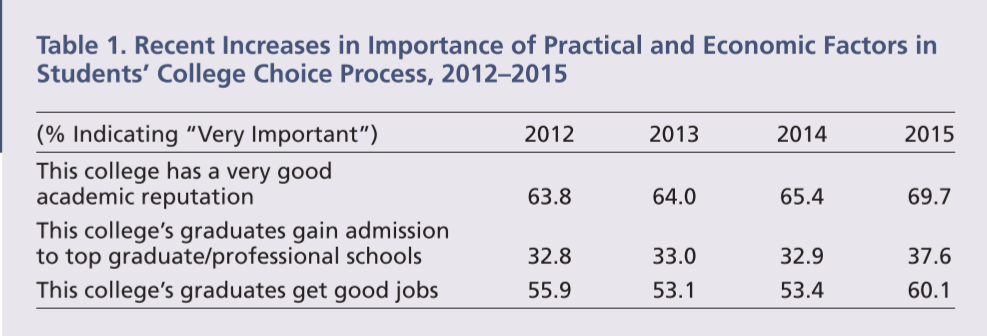
[Appendix 2: Back-up 10](#_Toc500255284)

Abstract

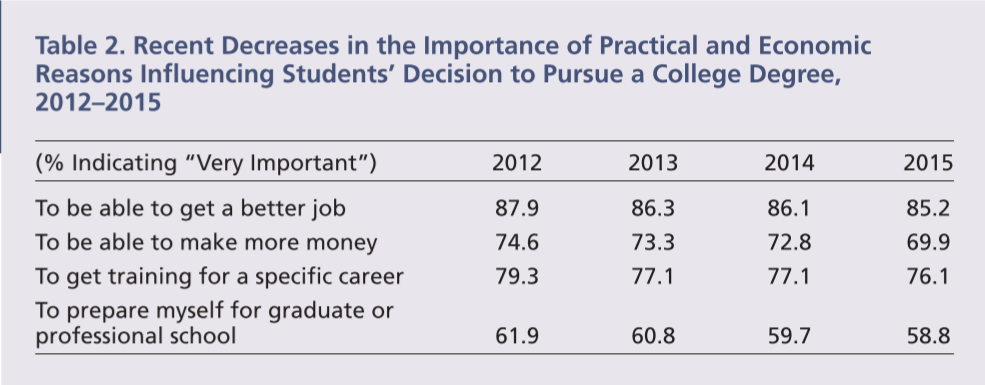
It is always known that the majority of the college students graduates with debts, which are far more than their future incomes level. There is also a common sense that students graduating from elite colleges tend to receive relatively higher earnings than those who are from less prestigious schools, however, the relationship between the college student's future income and types of the university attendance does not appears to be quite clear. Thus, our project is conducted to investigate what schools would help students to earn a higher income in the future, and what factors students should be aware of if they take the future earnings as the essential consideration.

Introduction

In 2015, UCLA did a survey about college choice among 1.5 million freshmen in America. In recent years, the percentage of students reporting economic and practical factors as “very important” in their choice of where to go to college has increased. Specifically, students now give more weight to post-college opportunities in their consideration of a specific college (see Table 1). The importance that students place on graduates’ ability to get good jobs and graduates’ admission to top graduate or professional schools has increased substantially.



As for the reason why students would go to college, the result showed that 85% students are for getting a better job, 70% students are for making more money 76% students are for getting training for a specific career.



So obviously most of the students hope to get better jobs or make more money after college education. But now, the real relationships between future income and university attendance are not very clear. 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.

Problem Statement

With the primary interest to understand what kind of university can cultivate the students who will earn more or less in the future, in this project, we’d like to use data mining tools to discuss what kind of universities could have the graduates with higher future income and which factors will influence the future income of graduates.

Data description

The data used in this study was obtained from US Department of Education, via the website of Kaggle. The US Department of Education has matched the information from the student financial aid system with federal tax returns to create this dataset. It initially contained thousands of attributes, describing the features of colleges and students. Since the objective of the project is to provide feasible suggestions to those who are determining to attend colleges with the future income as a consideration, the variable “means earning of student in 10 years” has been selected to be the target field. Additionally, a couple of relevant attributes, considered as the main potential factors which will impact the student’s future income, were added to the data analysis process.

The target variable is mean earnings of student in 10 years, appearing as the continuous variable in the file. Once the methods of decision tree, neural network and clustering are going to be applied, target variable should be converted into a nominal variable. As the consequence of this fact, the variable of mean earnings has been manually classified into three categories, low, medium and high, based on the lowest 25%, middle 50%, and top 25%. In addition to the target variable, there are nine continuous variables and two nominal variables, measuring various aspects of college and college students. The variables are admission rate, average SAT equivalent score of students admitted, enrollment of undergraduate degree-seeking students, tuition and fees for in-state and out-of-state students, proportion of full-time faculty, percent of all undergraduate student receiving a federal student loan, percent withdrawn from original institution within 2 years, percentage of first-generation students. There are also two nominal variables included in the file. One is predominant degree awarded with the contents of not classified(PREDDEG=0), predominantly certificate-degree granting(PREDDEG=1), predominantly associate-degree granting(PREDDEG=2), bachelor degree-granting(PREDDEG=3) and graduate-degree granting(PREDDEG=4). The other nominal one is control of institution, which is classified as public institution( Control=1), private non-profit institution(Control=2) and private for-profit institution(Control=3) . Before processing the model with the data, the step of data cleaning has been conducted to remove all the missing values and errors.

Methodology

**1.Data collection**

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

**2.Data preprocessing**

No noisy, inconsistent or intentional data is found in the original document, but there are lots of missing values. However, these tuples cannot be filled in manually because the true values are unknown. For the accuracy of the prediction, all the tuples that have missing values are ignored and after data preprocessing, there are 1310 tuples available left.

However, in the dataset, income cannot be set as the target directly since the variable type of it is continuous from 12300 to 250000. So, to specify the range of the income, it’s necessary to treat the first 25% of the observations as low-income group, the middle 50% as intermediate income group and the top 25% as the high-income group. As consequence of this operation, which group of universities can cultivate students who will earn the high income or relatively less income in the future can be presented.

**3.Variable selection**

Though there are too many variables in the dataset and many of them are redundant and repetitive, only 11 meaningful and interesting variables from the data set. They include 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 the target is Mean earnings of students working and not enrolled 10 years after entry(mn\_earn\_wne\_p10).

**4.Model Building**

The first step of model building is partition. To better build up the model, 50% of the records are chosen to go training and the order half go testing. Those training data in the modes will show what the predictors are and what the important level of the predictors is. Since the objective is to see how the income will be impacted, the field income will be apparently selected as the target in the Type nod, and those predictors such as SAT scores, withdraw rate and tuition fee are selected as input. In order to have an understanding of the inner interaction among these variables, and to get a straightforward view to the relationship with respect to the attendance and income, firstly, decision tree model neural networks model are built respectively with C5.0 algorithm and Neural Net node, and then histograms of SAT and tuition fee and plot of undergraduate students’ number are drawn to explore the reasoning behind the predictors in neural networks model. In addition, another clustering model with four clusters is shown as supplement and the distribution of K-means is able to analyze the model deeply and precisely. All of these three models are considered as appropriate methods applied to the model building process.

**5.Evaluation**

For the models constructed, it’s always needed to see how accurately the model can predict the data. So, confusion matrices below for training and testing data are made to evaluate the performance accuracy of the Decision Tree and Neural Networks models. As a silhouette measure of cohesion and separation, cluster quality is shown as fair in model summary of clustering.

|  |  |  |  |
| --- | --- | --- | --- |
|  | High | Low | Medium |
| Decision Tree  (Training) | 88.820% | 78.125% | 95.268% |
| Decision Tree  (Testing) | 69.880% | 43.114% | 70.030% |
| Neural Network  (Training) | 65.839% | 65.265% | 82.965% |
| Neural Network  (Testing) | 67.740% | 52.096% | 77.151% |

**6. Optimization**

In order to purchase more specific information of universities which have the greatest well-earning graduates, the level of income is reset and the top 10% is defined as “very high” income group.After optimization, the new outcomes of the models above present more precise details and the conclusions are updated accordingly.

Result and Discussion

**1.Decision tree model**

By observing the decision tree model, it could be learned in specific what kind of universities will have graduates with relatively low, medium, or high income in the future. The most important predictor here is the tuition school charge for out of state student, followed by average SAT admission score and tuition for in-state students. For example, If a school’s average SAT admission score is greater than 1,172, the graduates’ average future income will be all in the medium or high class as we define. If the average SAT admission score is less than 1,172, but the tuition the school charges for out of state student is greater than 31,010, then the school will still be predicted as the medium or high class.

**2. Neural Network model**

For Neural Network, the most important predictors are number of undergraduate students, tuition for out of state student and for in-state student, and the average SAT score for admitted students. According to the histogram of SAT\_AVG, schools with average SAT score of student admitted greater than 1400 are all in high class, school with SAT smaller than 770 are all in low class.The higher the SAT score is, the higher income the school’s student will earn in the future. According to the plot of class vs.UGDS, most of schools in low group tend to have less than 2000 undergraduate students. Compared to those in medium and high groups, low group students tend to be in a smaller size.

**3. Cluster Model**

The cluster model is fair quality. The most important predictors in this model are the tuition school charge for in-state student and the ownership of school. Ranking by the size, cluster 2 is 40.2%, cluster 1 is 39.4%, cluster 4 is 19.7% and cluster is 0.7%. The cluster 4 is the most interesting because it has the most high-income schools and least low-income schools. Those schools have following features in common.They are all Private nonprofit school. Compared to other 3 clusters, they charge the highest tuition, have least percentage of first-generation students, have highest SAT admission score for students, and least withdraw rate.

**4.Discussion**

Combining the results of three models, there are 4 important predictors: SAT admission score, tuition fee, number of undergraduate students, and ownership of the institution. Besides, three trends could be driven by those models. First, university with high SAT admission score tends to have high future income graduates. Second, universities charges high tuition usually are confident that they worth it. On the financial side. Third, schools with average low future income graduates tend to have a smaller size in the number of undergraduate students.

**5. Top 10% very high group**

In decision tree model, if a school’s average SAT admission score is greater than 1,250 and its admission rate is less 0.52. If SAT score is between 1,141 and 1,250, percent withdrawn from original institution within 2 years is less than 0.112, undergraduates number is greater than 2069, tuition for out of state student is greater than 25,050 and the school is private for profit, then the school will be in the very high income group.

In neural network model, the very high group shows strong relationships with SAT score and with tuition. With average SAT admission score higher than 1,400 are all very high group school. With graduate having highest average future income tend to charge the highest tuition. Very few of them charge a relatively low tuition.

In cluster model, the very high group shows similar characteristic with the high group in the previous classification.

Conclusions

After using three model to analyze the relationship between future income and university attendance, we could have a general impression on what factors about the school a student who hopes to earn a high income after graduation should consider when he or she making a decision about college choice: SAT, tuition, school size, and ownership. Thus students could have a clear idea about how to achieve the goal.

Appendix 1:Contributions of each group member

Fang Zhou: responsible for result, discussion and conclusion.

Yu Zheng: responsible for methodology, especially including model building, evaluation and optimization.

Chang Su: responsible for introduction and problem statement, and information composing.

Tiecheng Zhao: responsible for data processing, specifically for variable adjustment and data cleaning.

Appendix 2: Back-up

1. Data Description

(1)Data Source: US Dept of Education: College Scorecard <https://www.kaggle.com/kaggle/college-scorecard>

(2)Original data: 7676 records, 1720 attributes (school, academics, student, cost, completion, admission, aid, earning, repayment and root.)

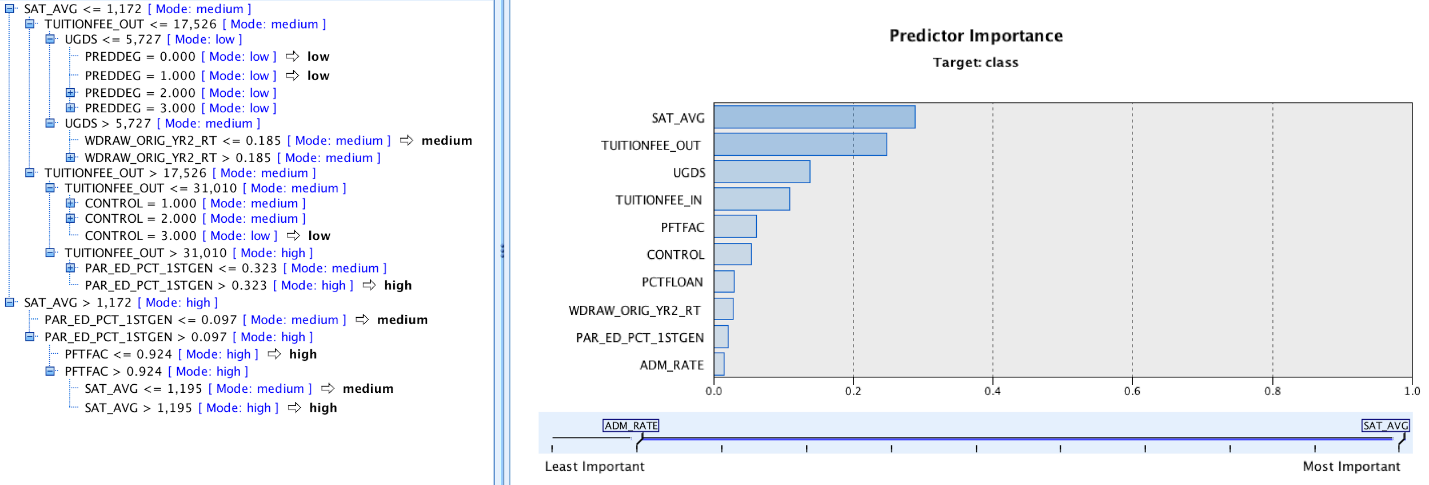
(3)Processed data: 1308 records,11 attributes, 1 target.

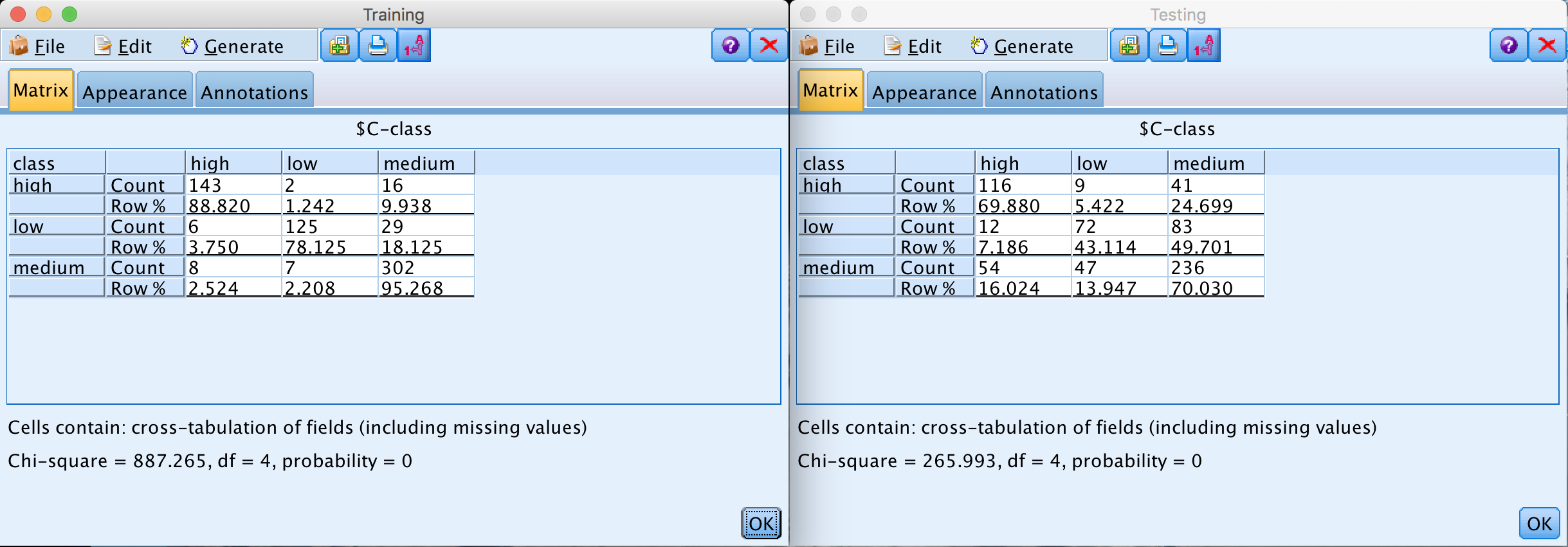
(4) 11 attributes selected

|  |  |
| --- | --- |
| Row name | Description |
| ADM\_RATE | Admission Rate |
| SAT\_AVG | Average SAT equivalent score of students admitted |
| PREDDEG | Predominant degree awarded:  0: Not classified  1: Predominantly certificate-degree granting  2: Predominantly associate-degree granting  3: Predominantly bachelor’s-degree granting  4: Entirely graduate-degree granting |
| UGDS | Enrollment of undergraduate degree-seeking students |
| TUITIONFEE\_IN | In-state tuition and fees |
| TUITIONFEE\_OUT | Out-of-state tuition and fees |
| PFTFAC | Proportion of faculty that is full-time |

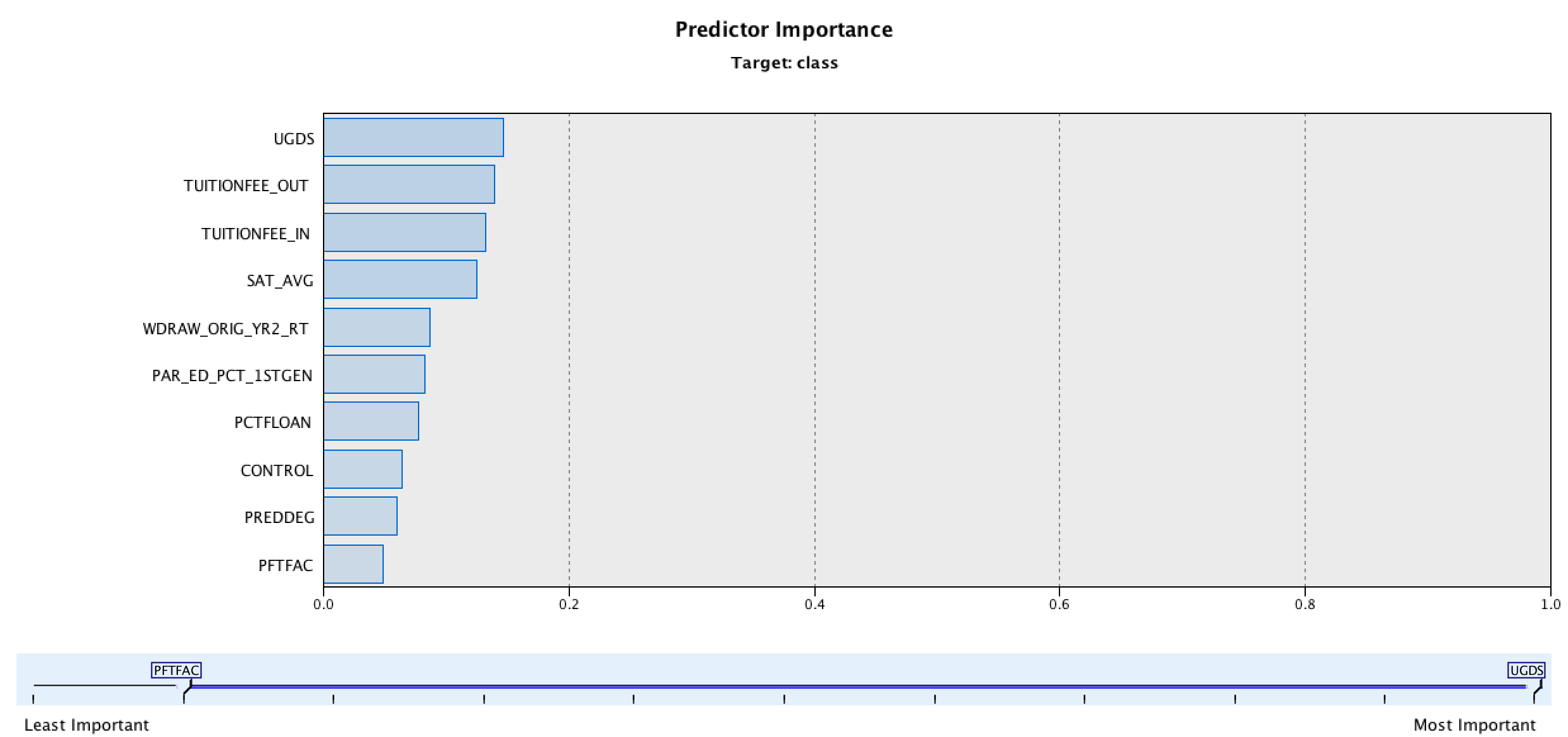
|  |  |
| --- | --- |
| PCTFLOAN | Percent of all federal undergraduate students receiving a federal student loan |
| WDRAW\_\_ORIG\_YEAR2\_RT | Percent withdrawn from original institution within 2 years |
| PAR\_ED\_PCT\_1STGEN | Percentage first-generation students |
| CONTROL | Control of institution  1: Public  2: Private nonprofit  3: Private for-profit |

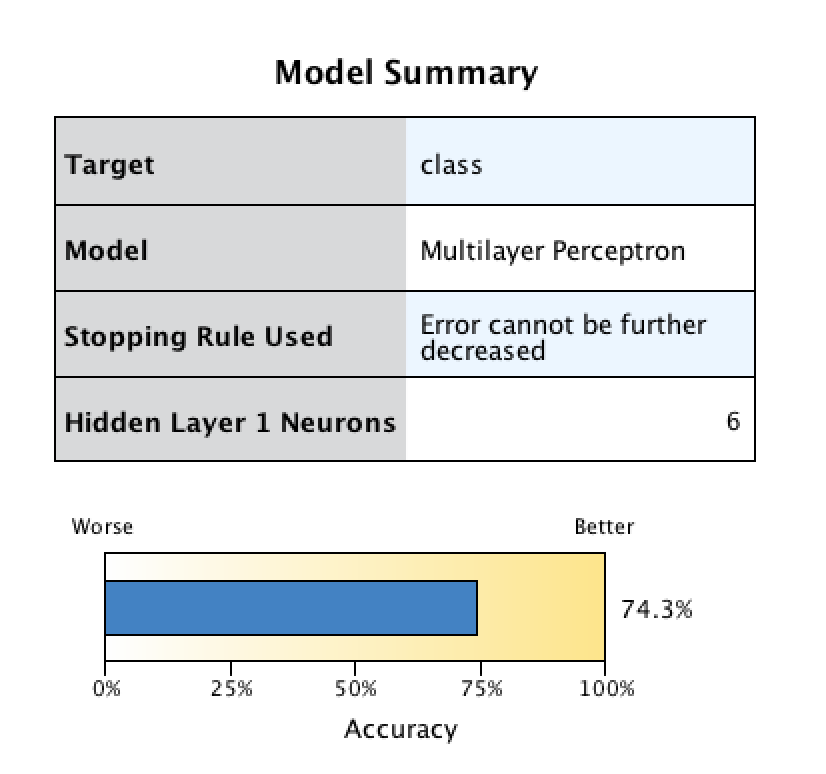
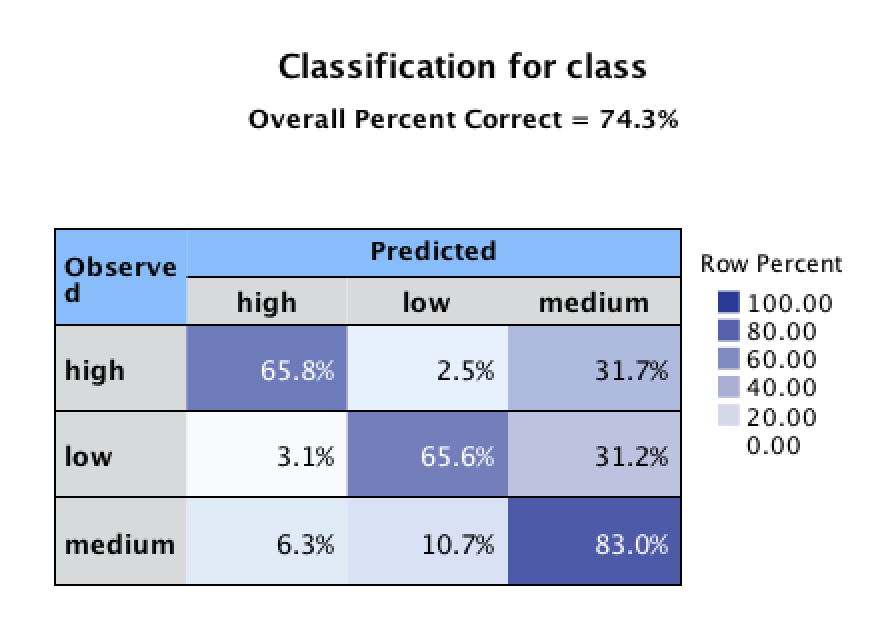
2.Decision Tree Model & Confusion Matrix

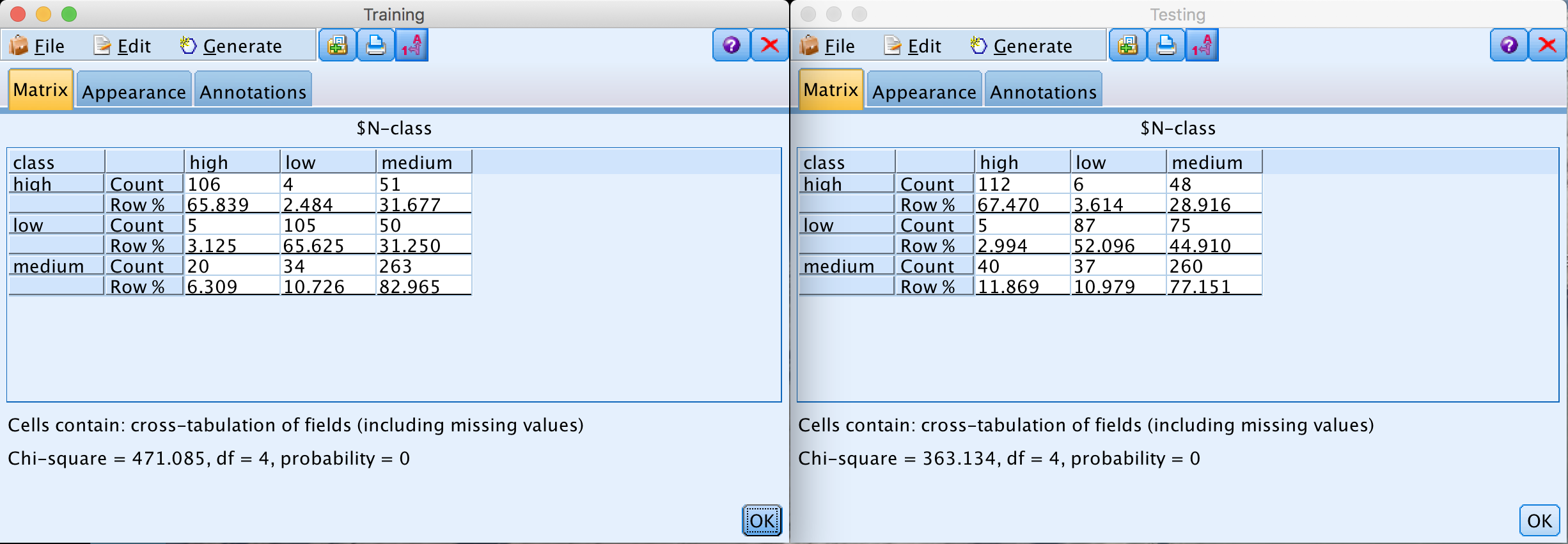
****

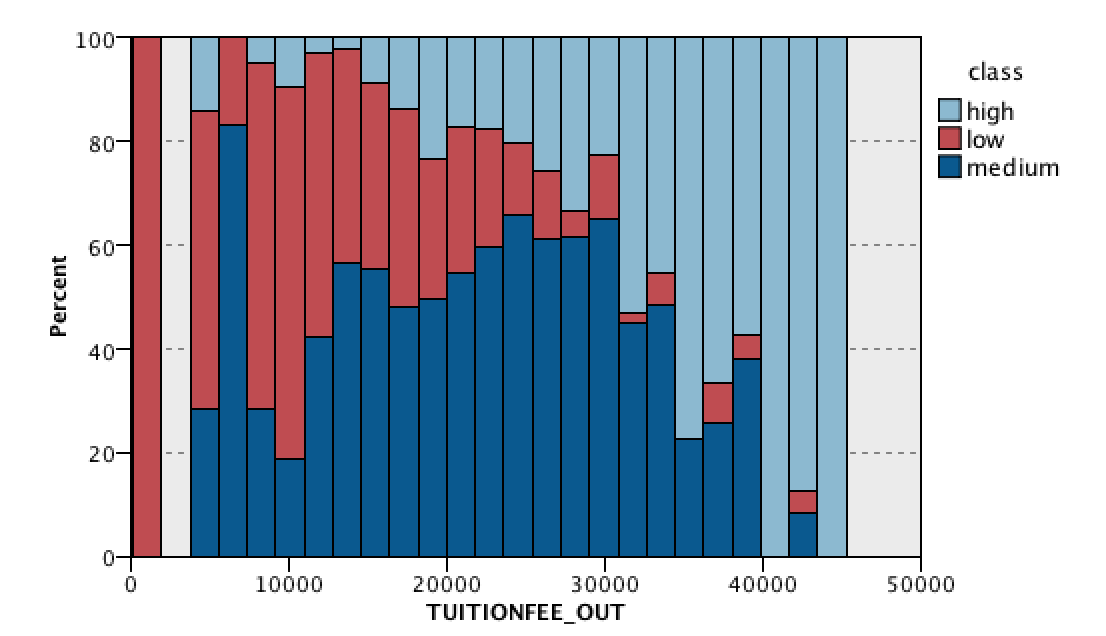
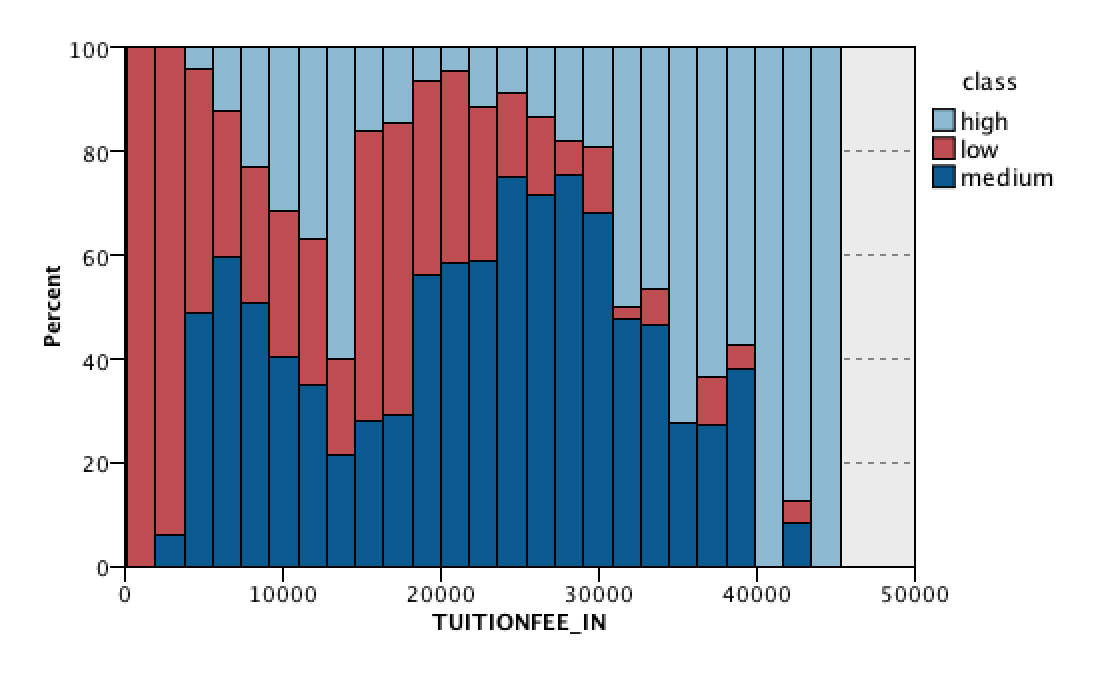
****

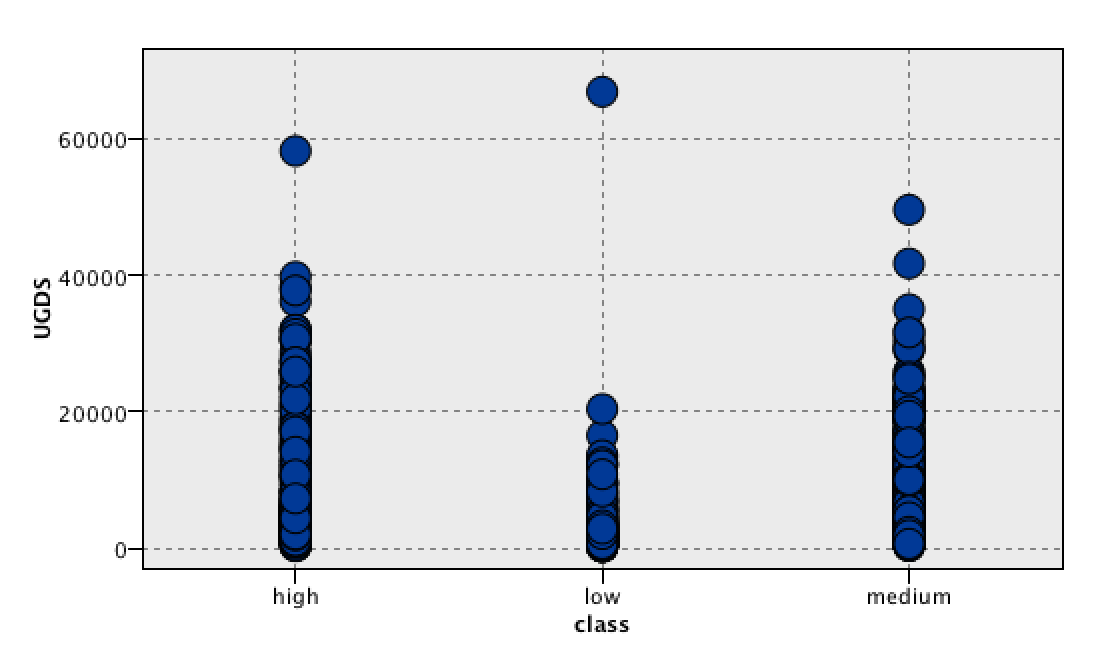
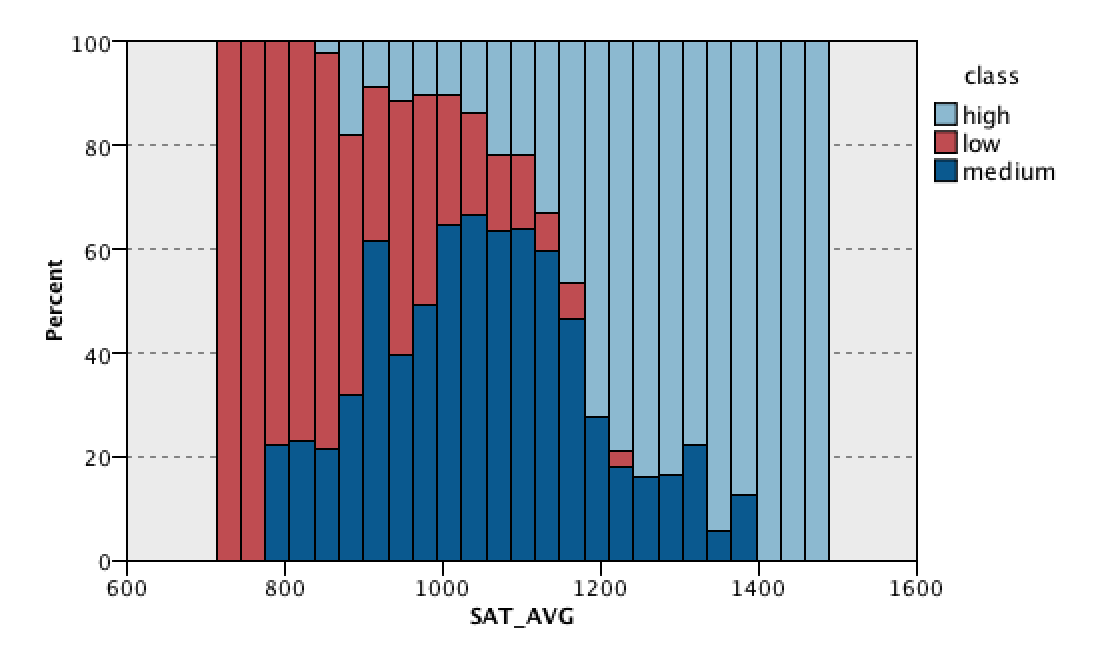
3.Neural Network, Confusion Matrix & Graphs



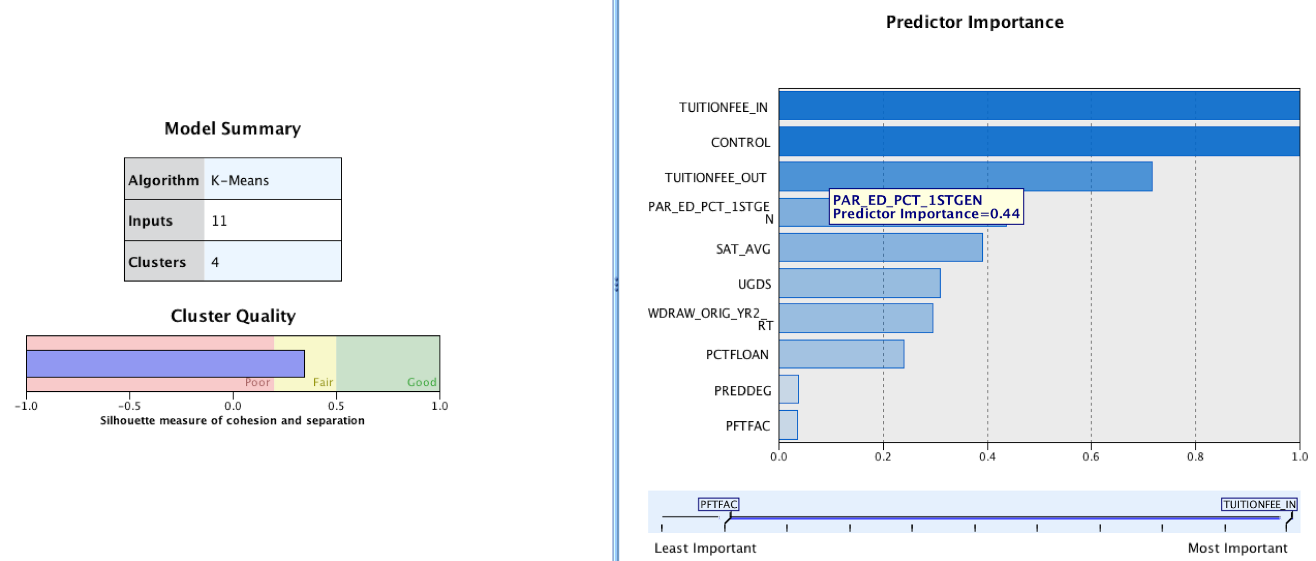


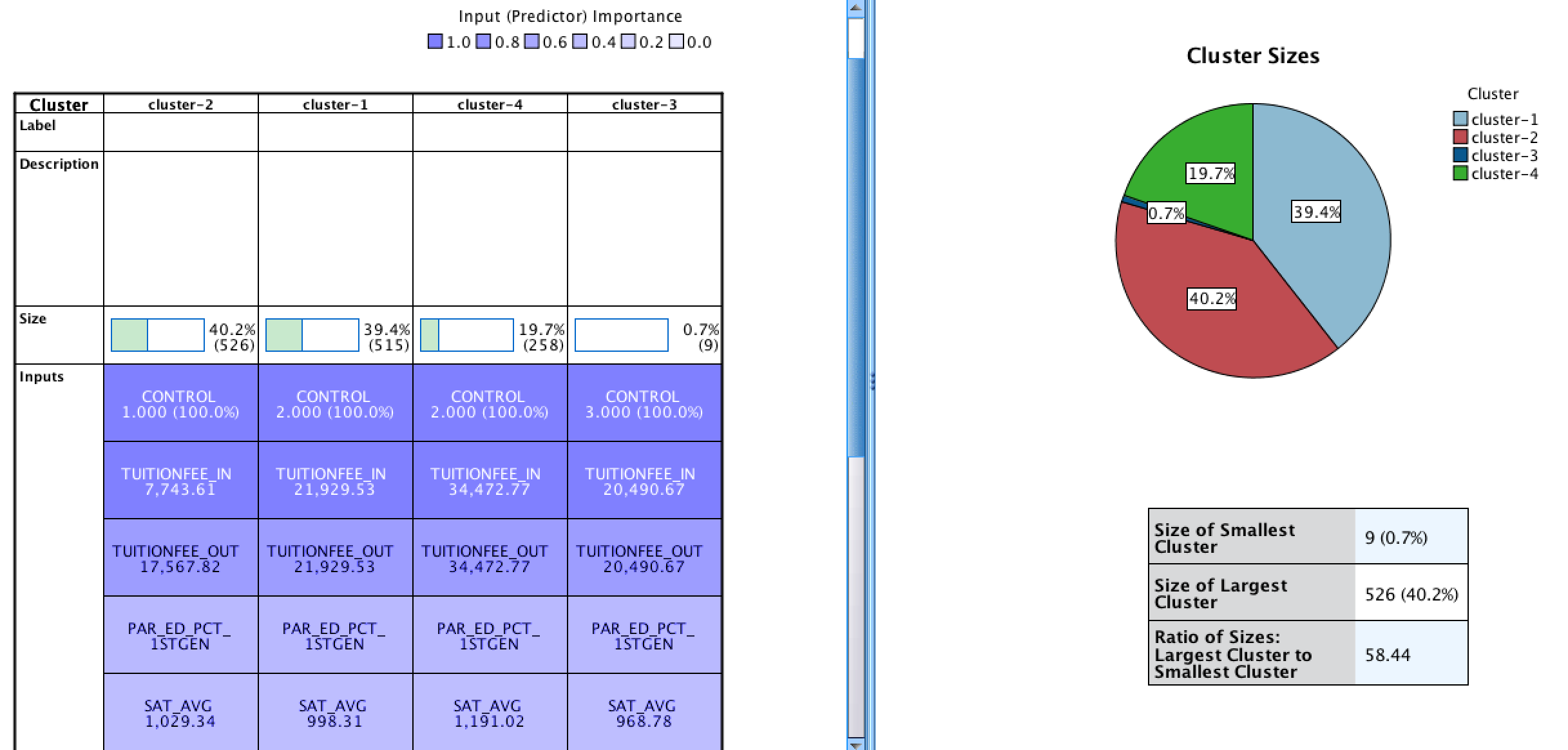


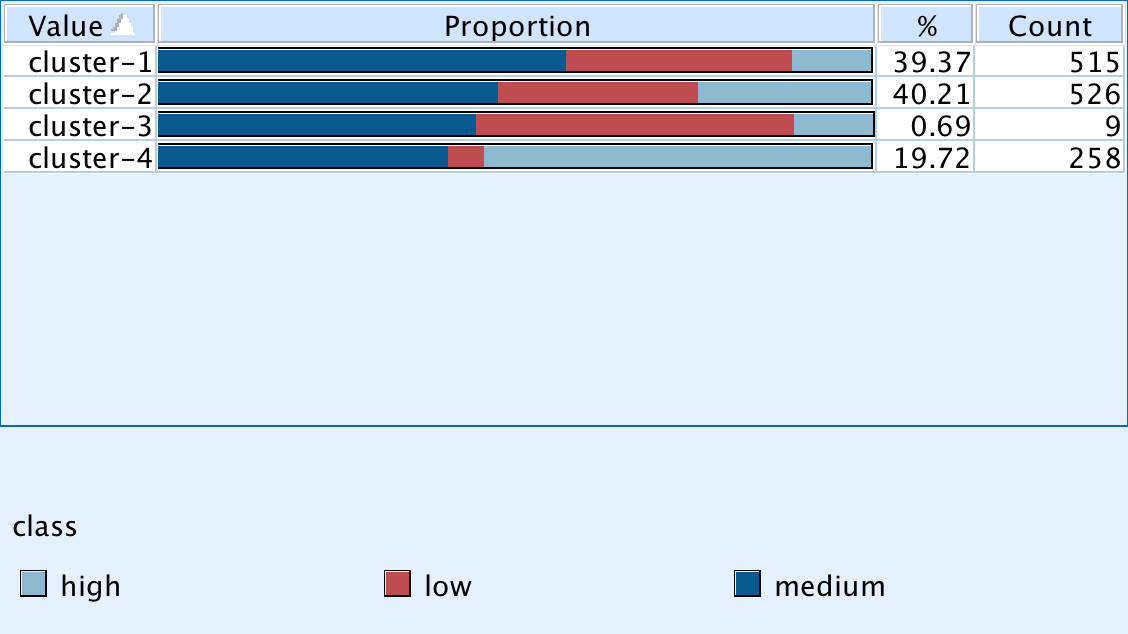




4. Cluster Model & Cluster Distribution Graph

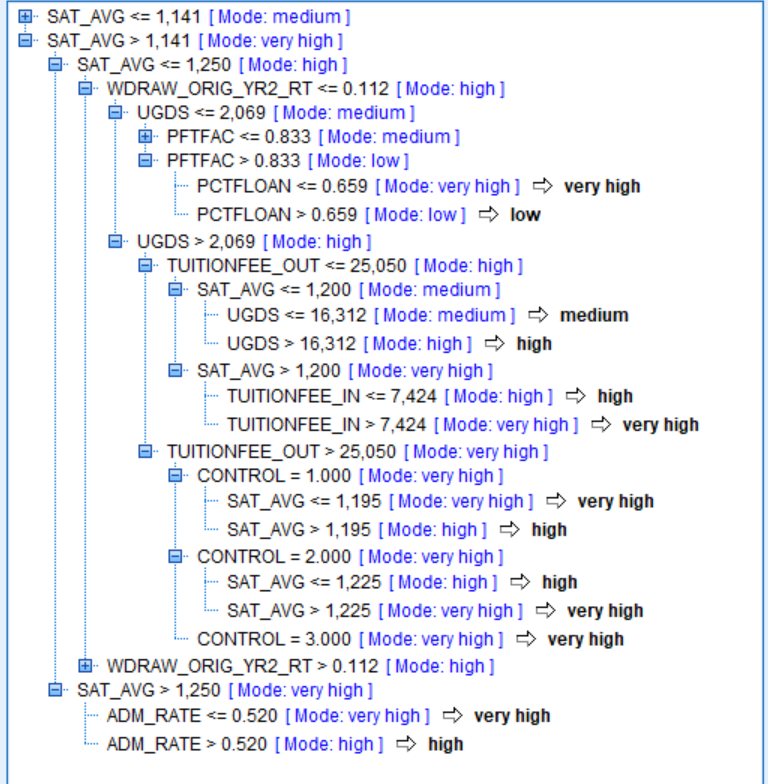
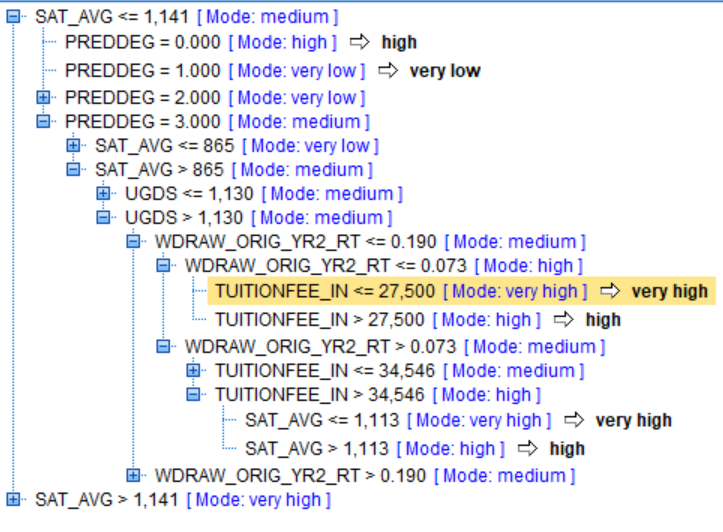


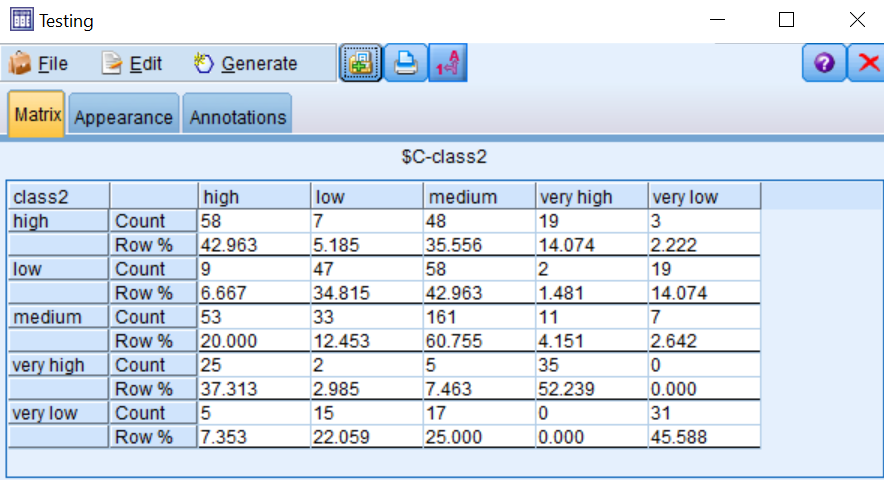
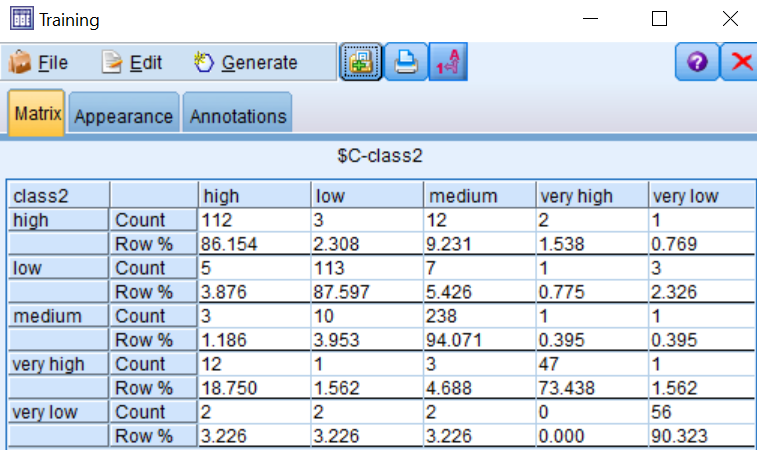




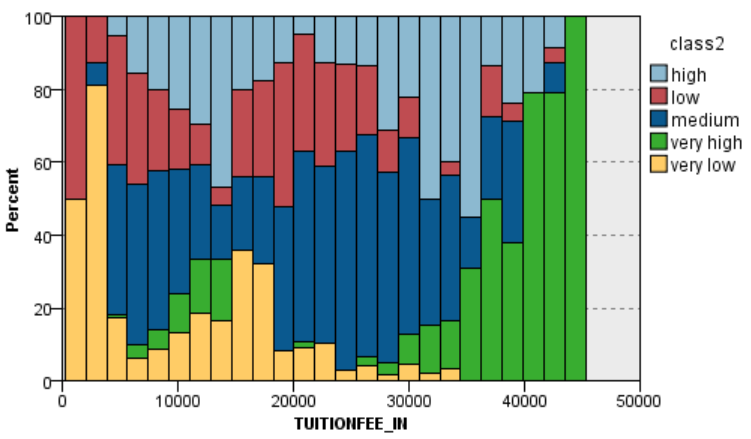
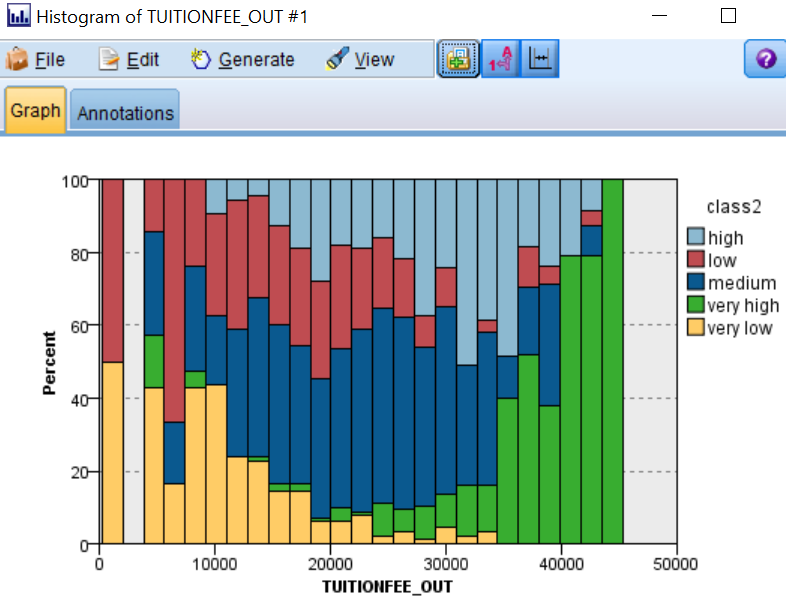
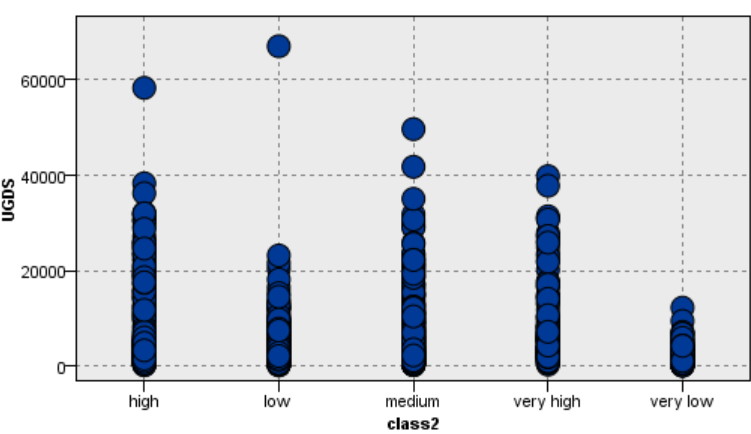
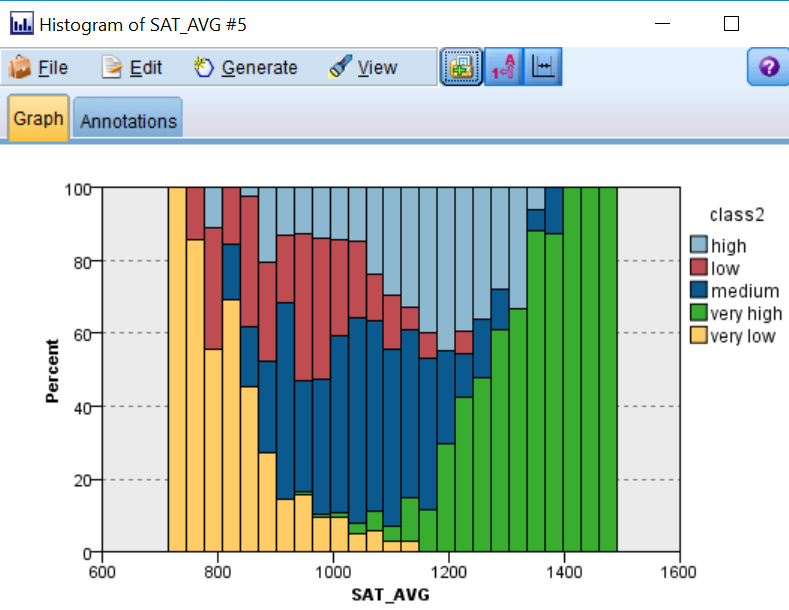
5. Top 10% Very High Group

(1)Decision Tree Model





(2) Neural Network Model and Graphs



(3) Cluster Model: Distribution Graph

