Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/5971
DOI URL: http://dx.doi.org/10.21474/IJAR01/5971

RESEARCH ARTICLE

EXAMINING STUDENT RETENTION WITH DATA ANALYTICS.

Charles E. Downing. Ph. D.
Dean's Distinguished Professor, Presidential Teaching Professor, Operations Management and Information Systems Department. College of Business, Northern Illinois University.

Manuscript Info

Manuscript History

Received: 05 October 2017
Final Accepted: 07 November 2017
Published: December 2017

Key words:-
Student Retention, Data Analytics, R and RStudio.

Abstract

Student retention is a critical issue for institutions today. As students have increasing options for educational and career opportunities, institutions need to engage and retain students so they complete their degrees. Simultaneously, the need for Data Analytics knowledge and talent is exploding in the Information Systems field. This paper aims to utilize prevalent Data Analytics techniques using the open source software packages R and RStudio to examine characteristics important for student retention. K-means clustering and Decision Tree analyses were conducted to tell the story of student retention. Data used was collected in a large Freshmen class at a public institution. Results show that the perceived atmosphere at the institution, a student's GPA, and lack of financial pressure are critical factors for retention. Other factors are also important and discussed. Suggestions are offered to increase student retention.

Introduction:-

As institutions compete with community colleges for student enrollment, and potential employers compete with both, student retention continues to be a hot topic for higher education (Borgan, 2016; Braxton & Hirschy, 2005; Cabrera, Nora, & Castenada, 1992; Downing, Spears, & Holtz, 2014; Leppel, 2001; Meeuwisse et al., 2010; Pascarella & Terenzini, 2005; Tinto, 2006; Turner & Carriveau, 2010). Xu (2015) points out that retention has been one of the most studied topics in higher education over the past four decades, but despite the attention the college attrition rate has remained alarmingly stable at around 45 percent over the past 100 years (NCES Digest of Educational Statistics, 2014). If students are not retained, enrollment decreases, and if enrollment decreases, budgets are reduced and quality and availability of services is compromised. Administrators and policy makers are scrambling to learn what factors can increase student retention.

Simultaneously, a hot topic in the Information Systems field is "Big Data Analytics". McKinsey and Co. have predicted that the talent gap for qualified data scientists will reach over 1.5 million by 2018 (Manyika et al., 2011). Colleges and institutions have taken note, and courses for Big Data Analytics and related areas are rapidly proliferating.

The paper will explore both of these areas, by demonstrating two mainstream analytics techniques on student survey data on the topic of student retention. The methods of the analytics techniques will be demonstrated, with the additional benefit of the results concerning student retention being discussed.

Corresponding Author:- Charles E. Downing.
Address:- Dean's Distinguished Professor, Presidential Teaching Professor, Operations Management and Information Systems Department. College of Business, Northern Illinois University.

Methodology:-

The research question is an exploration of student retention over multiple potential factors. The goal is to use multiple data analytics techniques on multiple potential retention factors to build a story of, and guidelines for, student retention. Potential factors used come from Xu's study in 2015 (Xu, 2015). To explore the research question, surveys were made available electronically, on a voluntary basis, to 302 students in a large business information systems course at a public institution. The goal of this research was not to create and validate a new instrument, so questions from Xu's study on student retention were used (Xu, 2015). The Appendix lists the questions used for this study. The questions were administered during a normal class period. Extra credit was given for completing the survey questions which were asked using a student response system. Students were informed that all responses were anonymous, participation was voluntary, and all collected results would be reported in aggregate only. Numbers of responses varied by one or a couple of students per question (students might leave to use the restroom or various other reasons to not answer one or more questions), and 252 students out of the 302 enrolled responded to the survey (83% response rate). The 252 responses were made up of 162 freshman, 44 sophomores, 34 juniors, 1 senior, 4 "other", and 7 students did not identify their year. All students were enrolled in the College of Business. (IRB Protocol # HS15-0346 "Examining student retention with data analytics", reviewed by Institutional Review Board #2 on 11-Nov-2015 and it was determined that it meets the criteria for exemption, as defined by the U. S. Department of Health and Human Services Regulations for the Protection of Human Subjects, 45 CFR 46.101(b), 2.).

Data Collection and Analysis:-

A five-point Likert-scale was used for each question in Xu's (2015) survey (see Appendix) administered to the 252 students. Answers ranged from "Strongly Agree" (numeric value 1) to "Strongly Disagree" (numeric value 5). The open source software packages R and RStudio were used to examine the data to determine characteristics important for student retention. K-means clustering and Decision Tree analyses were conducted to tell the story of student retention.

The first analysis performed was K-means clustering. According to EMC (2015), K-means should be used when the data scientist wants to group items by similarity, and wants to find structure (commonalities) in the data. In that sense, K-means is often an exploratory technique, used as a prelude to classification. K-means uses Euclidian distance to categorize ordered pairs that are closest to one other. Individual data points then get assigned to their closest cluster. This is an excellent technique to use to start to tell the story of student retention… the idea is to cluster student answers into similar groupings. Most notably, if students indicated they were considering dropping out or leaving the institution, what other survey answers were grouped in that cluster? What traits showed an inclination to "be or not be retained"? The plot of the "Within Groups Sum of Squares" of the potential number of clusters is used to determine K, the final number of clusters. For this data set, K was chosen to be 8 (8 clusters). Table 1 shows the survey response means for the 8 clusters, with the clusters appearing as rows. The larger bold-faced means are of note. Questions 24 and 26 (see Appendix) were used for the output variable of Retention. Cluster 1 has very low means (1.68 and 1.47) for questions 24 ("I have seriously considered dropping out of college") and 26 ("I have seriously considered leaving my institution for another school"), with "1" being "Strongly Agree". Therefore, Cluster 1 could be considered the "retention risk" cluster. These results will be discussed in the following section.

Table 1:- K-Means Cluster Means For The 8 Clusters For Student Retention

```
Cluster means:
      Year      GPA StudyGroup3 FacultyInteraction5   Greek6 ResHall7 Cultural8 CommServ9 Friends11
1 2.052632 4.052632    3.000000            3.315789 3.842105 3.578947  3.578947  3.000000  3.210526
2 1.583333 4.708333    2.750000            2.375000 4.583333 4.583333  3.916667  2.333333  2.875000
3 1.617647 4.382353    2.000000            2.500000 1.705882 3.264706  2.911765  1.970588  1.470588
4 1.242424 4.757576    1.818182            2.454545 1.242424 3.575758  3.757576  1.878788  1.181818
5 1.555556 4.500000    3.777778            3.111111 2.277778 3.555556  3.555556  4.111111  2.444444
6 1.562500 4.781250    2.843750            2.531250 4.250000 3.656250  4.437500  3.718750  1.812500
7 1.147059 4.794118    2.088235            2.823529 3.235294 2.088235  3.235294  2.558824  2.294118
8 1.950000 4.500000    2.600000            3.400000 4.150000 4.200000  4.450000  3.150000  2.300000
  SocialLife12 CoursesInter13 AcademicDevel14 FinSupportFamily15 FinEase16r AcademicGood18
1     3.421053       3.210526        3.315789           4.000000   4.105263       1.894737
2     2.791667       2.625000        2.458333           1.666667   2.000000       2.000000
3     1.352941       2.794118        2.088235           1.647059   3.264706       1.558824
4     1.151515       2.151515        1.575758           1.575758   2.181818       1.393939
5     2.388889       4.055556        2.833333           1.722222   2.888889       2.111111
6     1.687500       2.625000        1.937500           1.562500   2.593750       1.562500
7     2.470588       2.558824        1.794118           2.029412   2.176471       1.647059
8     2.050000       2.250000        1.700000           4.350000   4.050000       1.650000
  AtmosphereGood19 AdvisingSufficent21r AccessFaculty22 ImpFinish623  DropOut24   LeaveU26
1         2.894737             3.000000        2.210526     1.368421    1.68421   1.473684
2         3.083333             3.041667        2.458333     1.000000   4.083333   2.583333
3         1.676471             2.617647        2.000000     1.117647   2.058824   3.058824
4         1.575758             1.969697        1.818182     1.090909   4.757576   3.939394
5         2.611111             2.666667        2.388889     1.555556   2.833333   2.388889
6         1.906250             2.281250        1.968750     1.250000   4.625000   4.281250
7         1.970588             2.235294        2.088235     1.205882   4.470588   3.147059
8         2.050000             3.050000        2.100000     1.400000   3.900000   3.600000
```

While K-means clustering provides good guidance for determining what factors help retain (or not) students, in data analytics a more sophisticated classification technique is generally used after clustering. For this paper, that technique is Decision Trees. Table 2 shows the R code for fitting a Decision Tree to the retention data, and plotting and summarizing the tree. R uses Information Gain and Entropy to determine the most appropriate nodes to make a decision, in this case a prediction of Retention ("Student will be retained"/"Student will not be retained"). As mentioned, Questions 24 and 26 (see Appendix) were used for the output variable of Retention, and for the training data considered for this paper if a student answered "Strongly Agree" (1) or "Agree" (2) to either Question 24 or 26, the student was labelled a "NO" Retention. The second check was whether the student answered "Neutral" (3) to either Question 24 or 26, and if so the student was labelled a "MAYBE" Retention. The third and final check was whether the student answered "Disagree" (4) or "Strongly Disagree" (5) to either Question 24 or 26, and if so the student was labelled a "YES" Retention. "MAYBE" students were discarded, and the training data set was left with 115 NOs and 87 YESs. Table 3 shows the R Summary Output after fitting a Decision Tree to the retention training data. As EMC (2015) notes, "The output produced by the summary is difficult to read and comprehend" (EMC Education Services, 2015, pp. 209). Nonetheless, the "Variable Importance" provides good guidelines for building the retention story. Figure 1 displays the RStudio produced Decision Tree itself, adding more information.

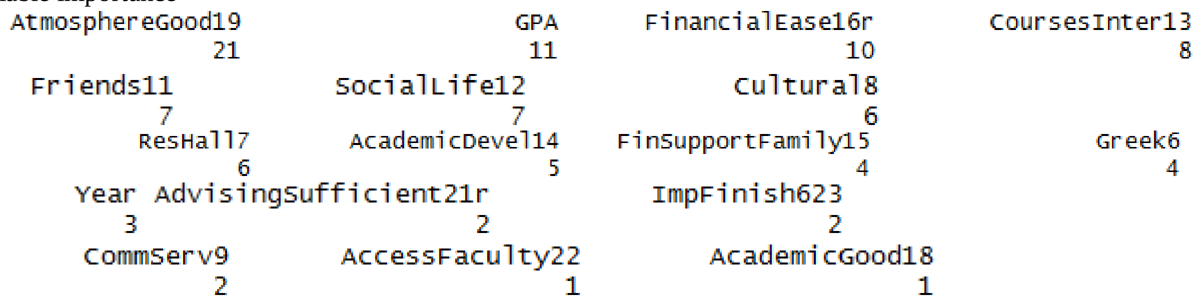
Table 2:- R code for fitting a Decision Tree to the retention data, and plotting and summarizing the tree

```
> install.packages("rpart.plot")
> library(rpart)
> library(rpart.plot)
> Retain_decision <- read.table("DataforRBin.csv", header=TRUE,sep=",")
> fit <- rpart(Retain ~ Year + GPA + StudyGroup3 + FacultyInteraction5 + Greek6 + ResHall7 + Cultural8 +
CommServ9 + Friends11 + SocialLife12 + CoursesInter13 + AcademicDevel14 + FinSupportFamily15 +
FinancialEase16r + AcademicGood18 + AtmosphereGood19 + AdvisingSufficient21r + AccessFaculty22 +
ImpFinish623, method="class", data=Retain_decision)
> rpart.plot(fit, type=4,extra=1)
> summary(fit)
```

Table 3:- R Summary Output after fitting a Decision Tree to the retention data

```
Call:
rpart(formula = Retain ~ Year + GPA + StudyGroup3 + FacultyInteraction5 +
    Greek6 + ResHall7 + Cultural8 + CommServ9 + Friends11 + SocialLife12 +
    CoursesInter13 + AcademicDevel14 + FinSupportFamily15 + FinancialEase16r +
    AcademicGood18 + AtmosphereGood19 + AdvisingSufficient21r +
    AccessFaculty22 + ImpFinish623, data = Retain_decision, method = "class")
  n= 202

Variable importance
```



```
   AtmosphereGood19                GPA     FinancialEase16r        CoursesInter13
                 21                 11                   10                     8
     Friends11          SocialLife12            Cultural8
             7                     7                    6
             ResHall7      AcademicDevel14   FinSupportFamily15              Greek6
                    6                    5                    4                  4
        Year  AdvisingSufficient21r          ImpFinish623
           3                      2                     2
        CommServ9       AccessFaculty22          AcademicGood18
                2                     1                       1
```

```
Node number 1: 202 observations,    complexity param=0.137931
  predicted class=No   expected loss=0.4306931  P(node) =1
    class counts:   115    87
   probabilities: 0.569 0.431
  left son=2 (54 obs) right son=3 (148 obs)
  Primary splits:
      AtmosphereGood19 < 2.5 to the right, improve=10.519300, (4 missing)
      Friends11        < 2.5 to the right, improve= 7.289174, (3 missing)
      GPA              splits as  RLLR,    improve= 6.873189, (0 missing)
      SocialLife12     < 2.5 to the right, improve= 6.253316, (1 missing)
      AcademicDevel14  < 2.5 to the right, improve= 6.244695, (1 missing)
```

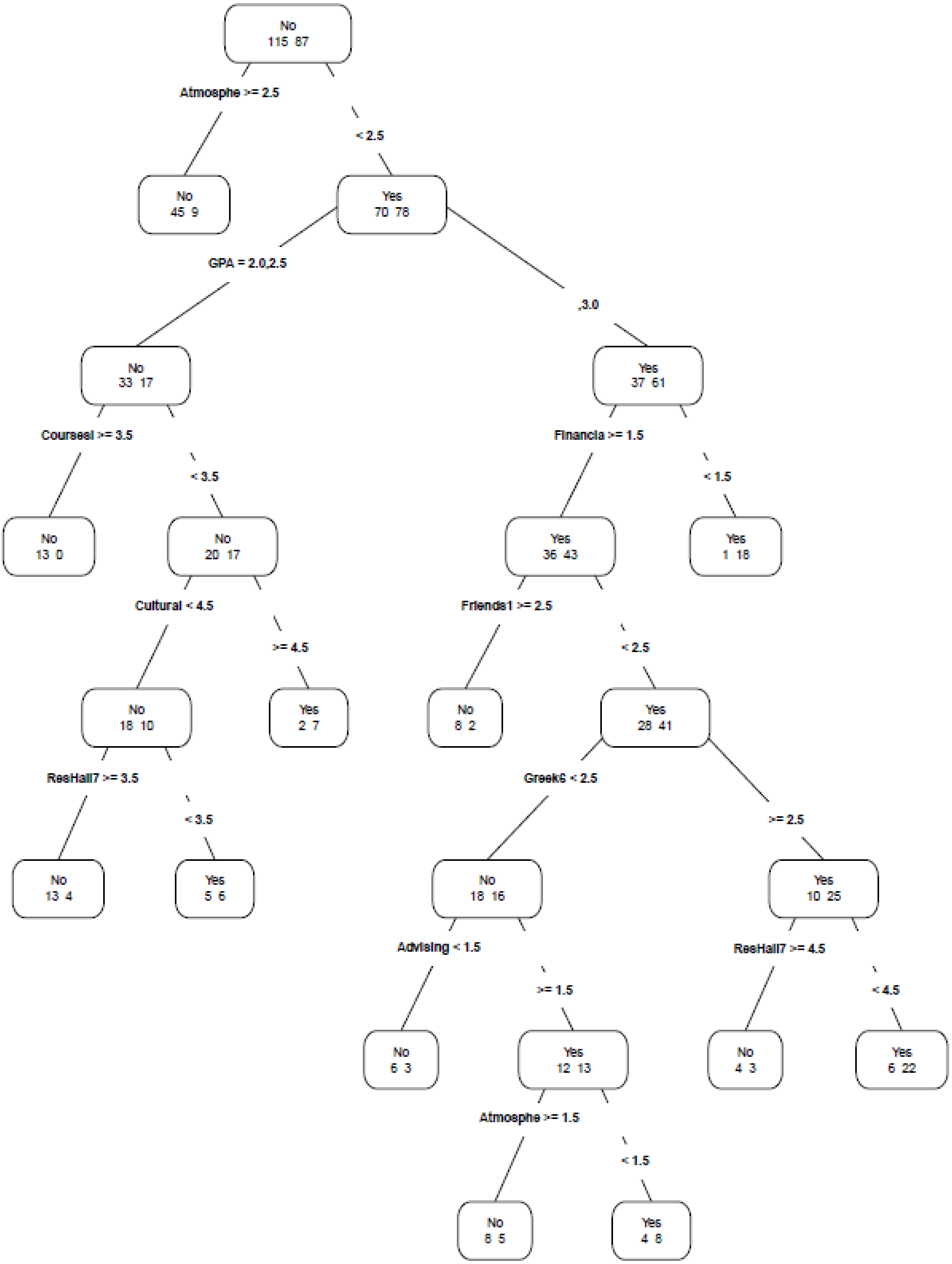


Figure 1:- RStudio Decision Tree output for binary case with default pruning

Discussion and Conclusions:-

Looking first at the K-means clustering output in Table 1, Cluster 1 has low means for Questions 24 and 26, the "Retention" questions, and it is important to note how the students in this cluster responded to the other questions. Question 2, the GPA question, was coded from left to right as 1 to 6. So a lower survey response average indicates a lower GPA. The Cluster 1 respondents have an average response of 4.05, which translates to a GPA of 2.5-3.0. While not failing by any measure, this GPA is still easily the lowest of the 8 clusters. This could indicate the intuitive result that as a student's GPA goes down, so does the likelihood of retaining that student. Additionally, Question 12 ("I am satisfied with my social life on campus"), Question 14 ("I am satisfied with my academic development at this institution"), Question 15 ("I have financial support from my family members for completing my college degree"), Question 16 ("Financial pressure is distracting me from my college coursework") and Question 19 ("The atmosphere at the institution which I attend is good") showed to be important questions in the Retention story, as shown in Table 1. All of these factors showed up in clustering as being important in the Retention story, and need to be "fed in" to the following technique, the Decision Tree.

The Decision Tree shown in Figure 1 demonstrates that the top node, the Root Node, branches on atmosphere (Question 19: "The atmosphere at the institution which I attend is good."). If students choose a response greater than 2.5 (meaning at least at the "Neutral" response, value 3, and extending to "Disagree", value 4, and "Strongly Disagree", value 5) then they are highly likely to not be retained (at 45 NO, or "Student Not Retained", and 9 YES). Note that, with default pruning, R returned the tree in Figure 1 which does not have all leaf nodes with explicit decisions. Meaning, the tree shows that if a student chooses "Strongly Disagree", "Disagree", or "Neutral" to the atmosphere question (#19) that student is 83% (45/54) likely to not be retained, but this determination is not at 100%. A tree with forced 100% leaf nodes was also constructed, but this tree contained 113 nodes as compared to the 23 nodes in the tree shown in Figure 1. Such a tree, completing every decision at 100%, is often referred to in data analytics as "over-fit".

Next in importance after the atmosphere at the institution is GPA (Question 2), not surprisingly. Of the 148 students who felt the atmosphere was good, 50 had low GPAs (less than 3.0) and were 66% (33/50) likely to not be retained. Those in the "atmosphere is good" group with GPAs above 3.0 were 62% (61/98) likely to be retained. Question 16: "Financial pressure is distracting me from my college coursework" (scale reversed prior to input for R), was the next most important factor. Similar guidelines can be obtained by following down the tree and reviewing the summary results from R. For example, next in retention importance is Question 13: "I find my courses to be generally interesting." The three most important factors, Atmosphere, GPA, and Financial Pressure, are consistent with the K-means clustering results, providing confidence that these factors are critical to student retention.

Using the decision tree can provide many opportunities for administrators and policy makers in higher education. Knowing that the student response to "The atmosphere at the institution which I attend is good" was the most important branching variable in the decision tree is critical. What makes an atmosphere "good" is of course another discussion. But one example of a possible action taken would be that surveys could be conducted with multiple potentials factors for what makes an atmosphere good. Many of these factors undoubtedly would be easier to "solve" or implement than something like financial pressure or a low GPA. Similar direction can be obtained from the decision tree, and R and data analytics have begun to build a story of student retention.

References:-

1. Borgen, S. T., & Borgen, N. T. (2016). Student retention in higher education: Folk high schools and educational decisions. Higher Education, 71(4), 505-523.
2. Braxton, J. M., & Hirschy, A. S. (2005). Theoretical developments in the study of college student departure. In A. Seidman (Ed.), College student retention: Formula for student success. Westport, CT: ACE/Praeger.
3. Cabrera, A. F., Nora, A., & Casta-eda, M. B. (1992). The role of finances in the persistence process: A structural model. Research in Higher Education, 33(5), 571-593.
4. Downing, C. E., Spears, J., & Holtz, M. (2014). Transforming a Course to Blended Learning for Student Engagement. Education Research International, 2014.
5. EMC Education Services. (2015). Data Science and Big Data Analytics. Indianapolis, IN: John Wiley & Sons, Inc.
6. Leppel, K. (2001). The impact of major on college persistence among freshmen. Higher Education, 41, 327-342.

7. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.
8. Meeuwisse, M., Severiens, S. E., & Born, M. Ph. (2010). Learning environment, interaction, sense of belonging and study success in ethnically diverse student groups. Research in Higher Education, 51, 528-545.
9. National Center for Education Statistics (2014). Digest of Educational Statistics, 2014, Washington, DC: National Center for Education Statistics.
10. Pascarella, E. T., & Terenzini, P. T. (2005). How college affects students: A third decade of research. San Francisco, CA:Jossey-Bass.
11. Tinto, V. (2006). Research and practice of student retention: What next? Journal of College Student Retention, 8, 1-19.
12. Xu, Y. J. (2015). Attention to Retention: Exploring and Addressing the Needs of College Students in STEM Majors. Journal of Education and Training Studies,4(2), 67-76.

Appendix A – Survey Instrument

1. My year in college is: Freshman Sophomore Junior Senior Other

2. After this semester, I expect my cumulative GPA to be:

Below 1.0 1.0-2.0 Above 2.0 but less than 2.5 2.5-3.0 Above 3.0 but less than 3.5 3.5-4.0 Other

Questions 3-26 were all answered on the following five-point Likert Scale:
Strongly Agree
Agree
Neutral
Disagree
Strongly Disagree

3. I participate in organized academic activities with peers (e.g. study groups).
4. I work with other students on school work outside of class.
5. I interact with faculty outside of the classroom concerning coursework.
6. I participate in events sponsored by a fraternity or sorority.
7. I participate in residence hall activities.
8. I participate in social or cultural events hosted by groups reflecting my own cultural heritage.
9. I participate in community service activities.
10. I find it difficult to get in touch with other students in my academic area.
11. I have plenty of friends amongst my fellow students.
12. I am satisfied with my social life on campus.
13. I find my courses to be generally interesting.
14. I am satisfied with my academic development at this institution.
15. I have financial support from my family members for completing my college degree.
16. Financial pressure is distracting me from my college coursework.
17. I have to work over 20 hours a week in order to fund my college education.
18. My academic program is of good quality.
19. The atmosphere at the institution which I attend is good.
20. I am satisfied with my interactions with faculty members.
21. I have found the academic advising offered to students to be insufficient.
22. I have access to faculty members for discussion and to receive advice.
23. It is important for me to complete my degree within six or fewer years.
24. I have seriously considered dropping out of college.
25. I may drop out of college if there are good-paying jobs available.
26. I have seriously considered leaving my institution for another school.