DSCI 4520 – Introduction to Data Mining

Fall 2022

**Term Project Description**

**Introduction:**

Student retention (sometimes called student persistence) is defined as the number of students who re-enroll from one year to the next in the same institute. Retention is a very important metric of a university success, particularly for “first-time in college” (FTIC) students. Universities spend a great deal of money, time, planning, and other resources to improve their FTIC retention rates. Studies have shown several factors influence student retention such as academic performance of students, financial concerns, faculty support, and community life. If universities can predict who is more likely to not retain next year, they can optimize their plans and resources to focus on those students. Therefore, there is a great benefit for higher education business to be able to predict student retention. In fact, there is a huge market for such products and services. Currently several companies have developed products to predict and improve college student retention that are commercially available.

Read more: [1](https://en.wikipedia.org/wiki/University_student_retention), [2](https://msmhighered.com/blog_posts/what-is-student-retention-in-higher-education/#:~:text=The%20definition%20of%20retention%20in,is%20performing%20for%20its%20students), [3](https://blog.campusgroups.com/campusgroups/2021/5/25/strategies-to-improve-student-retention)

**Objective:**

You are a team of business analytics experts that are hired by the Bright Future University (BFU) to help them increase their FTIC student retention. Your method is to use machine learning tools to predict who will not re-enroll next fall semester. You have collected three years of student data that includes their academic, demographic, and financial information. The goal of this project is to develop a classification model to predict students who will not re-enroll next fall. The BFU administrators expect you to deliver the model (R codes) and your recommendations and suggestions based on your data mining work.

**Data Set:**

The data set includes information of 9660 unique students who enrolled in BFU for the first time in the fall semester. Data records were collected for three academic years (three fall semester). Each student is represented by an ID and 22 attributes (columns) including their demographic, financial, and academic information. The last column (“Fall\_Not\_Retained”) shows if the student re-enrolled next fall semester or not. If it is 1 then student DID NOT re-enroll. This is your target variable that your classification model should predict. A data dictionary is also provided to explain all columns.

**Deliverables:**

The project requires two deliverables:

1. A detailed report that should be turned in in fours phases (four due dates)
2. All R programs that you used for data manipulation, visualization, and modeling

Project report should be written in the template provided on the Canvas.

The detailed report should have four chapters:

1. *Business Case*: this chapter includes an introduction, a definition of the business problem, a description of your approach to solve the problem, and performance metrics that you choose to evaluate your model based on the business context.

Questions that you should think about and address in this chapters are:

* What is the definition of FTIC student retention and why is it important for universities?
* What is the overall FTIC retention rate at Bright Future University (BFU)? You can use excel or R to answer this question. Retention rate is defined as:
* How much it will cost for BFU to lose one FTIC student after the first year? Note that BFU is an imaginary university, however, to answer this question you can assume BFU is a state university in Texas. Then do a web search to estimate the average tuition fee of public universities in Texas. Then estimate the cost for BFU. You don’t need to come up with exact numbers, but your method of estimation is important. You should describe it clearly.
* What is the business problem your team is hired to solve?
* How is your team planning to help BFU to solve the problem?
* What other universities are doing to improve their retention rate? Use google and write at least one real example
* What are the real-life applications of machine learning tools or data mining predictive programs to improve retention rates? Use google and write at least one example
* What are the metrics you want to use to evaluate the performance of your prediction algorithm? Note: In the provided data set, “not retained” is equal to 1. This means “not retained” is the label for the positive class. So, look at the performance evaluation metrics and pick one or two that serves your purpose in this project.
* A brief description of the report chapters.

1. *Data Exploration*: In this chapter you present results of your exploratory data analysis that includes data visualization and explanation of what data shows (such as interesting patterns, trends, etc.)
   * All variables in the student retention data set are categorical. Therefore, to explore the relationship between the target variable (“Fall\_Not\_Retained”) and the rest of the variables you can start by plotting the percentage of each retention group (“1” is not retained, “0” is retained) across levels of categorical variables using bar charts.
   * You should plot bar charts for ALL variables and select those that are showing interesting patterns and copy/paste them in your report. If a bar chart does not show an interesting pattern or trend, you should not include it in your report.
   * Each bar chart that is in the report must have a few lines of description and insightful information. The main goal of this chapter is to extract some insights from the data visualization output. Without a good description, the plot itself is not enough.
   * In addition to the bar charts, heatmaps can also be plotted to explore student retention patterns, across two variables at the same time.
2. *Classification Model*: in this chapter you explain the method and model you used for predicting student retention and you present the results of your model performance evaluation. You should show that your model works and can be used to answer the business problem that you presented in the first chapter
   * The goal is to build classification models to predict student drop-out. The models should be built in R with the project data and should be evaluated by the metrics that you have chosen in the first chapter.
   * In this chapter, you are expected to present the results of three different classification models. You can pick from these models: logistic regression, Naïve Bayes, decision tree, random forest, gradient boosting, or an ensemble of other models.
   * For each model, the data set should be split into a train (70%) and a validation set (30%). The confusion matrix of each model should be calculated and presented in the report. Also, the AUC ROC graph should be created and presented in the report.
   * The confusion matrix and all the metrics should be reported for training and validation sets. You should also show your models are not under or over-fitted.
   * Each model should be trained and tested independently.
   * Models should be compared based on performance metrics. You should explain which model is the best and why you choose it.
   * Write and save all R codes in a clean and understandable way. You should submit them at the end of the semester with your final report.
3. *Conclusion*: In this chapter you present a summary of your findings and explain how your analysis can be used to improve BFU business.

Several sample reports on other business problems can be found [on this page](https://mays.tamu.edu/humana-tamu-analytics/2019-finalist-submissions/). This [report](http://mays.tamu.edu/humana-tamu-analytics/wp-content/uploads/sites/113/2021/04/Finalist_2020_3.pdf) is a good example for chapter layouts.

**Teamwork:**

This is a group project, and all group members should participate equally. Submission of the report and codes can be done by any of the team members, but only people who have really participated will get credits. At the end of the semester, there will be a survey for each student to rank and score the participation of their groupmates. If a student does not earn at least half of the ranking points by his/her groupmates, he/she will not receive any credit (this means more than half of your group believe you didn’t participate enough.)

**Due Dates and Grading:**

Project has four dues dates. By each due date a group member should submit the following documents:

1. Sunday, October 2nd, 11:59 pm. First chapter of the report
2. Sunday, October 23rd, 11:59 pm. Second chapter of the report
3. Sunday November 20th, 11:59. Third chapter of the report
4. Sunday December 11th , 11:59. Final report (all four chapters) and all R codes.

After you turn in a chapter, I’ll review it and give you feedback so you can revise and modify it. Therefore, it is very important to submit chapters by the due dates. I will not grade your first, second, and third chapters individually, however, for each chapter that is not submitted on time your group will lose a 10% of the final project score. The final score (300 points) will be based on your final report and codes.