**BAT 3305 Admission Project**

**1. Introduction**

The Office of Admissions at Trinity University is responsible for managing each applicant’s admission process and materials. Applicants that have completed the required materials may be rejected from the University, accepted to the University, or deferred for application at a later date.

The goal of this project is to construct a classification model to determine whether an accepted applicant will decide to attend the University, based on information collected regarding the University’s recently accepted applicants, ranging in entry term from Fall 2017, Fall 2018, Fall 2019, Fall 2020, and Fall 2021. The classification model you construct and insights obtain will allow admissions employees to follow up and increase the yield (accepted / admitted).

Please refer to the “Description of Variables” file to get yourself familiar with each variable. The response variable is “Decision”, with 1 representing accepted the offer and 0 representing not accepted the offer.

**This is an individual project; absolutely no collaboration is allowed**.

**2. Deliverables**

This project requires two deliverables: an error-free, clear, and well-commented R scrip and a brief report.

**Part I – R script**

The R scrip must include:

1. Perform data cleaning on the whole data set. Data cleaning process normally involves:

* Handling missing values
* Correcting typos or measurement errors
* Recoding and/or augmenting existing variables
* Dealing with outliers
* Transformations
* Creating new variables
* Removing variables

*Note*: I have posted my data cleaning for the first 46 columns to Tlearn to offload some burden from student shoulders. If you would like to use some or all of my code directly, that is fine since the cleaning is performed on the entire data set not just the training set. If you have different thoughts on the cleaning of certain variables, feel free to modify my code. If you would like to develop your own way of data cleaning, you are more than welcome to do so.

If you decide to use some or all of my code, please pay attention to the following two things:

1. Please make sure you carefully read the code line by line. This is because I only commented on whether a feature may be considered for removal, but I didn’t do the removal on the data set. It is your responsibility to remove unnecessary predictors based on my comments and your own judgement.
2. There are several categorical variables with lots of levels (e.g., 50-60 levels!!). I merged similar levels together based on my understanding of the levels. However, my understanding might be inaccurate, especially for the “religion” variables. If your understanding of some levels of a categorical variable is different from mine, please modify the merge according to yours.
3. Once data is cleaned, split the combined data set into a training set and a test set aligned with the requirements posted on Tlearn.
4. Implement EACH of the following classification methods with all or the most features you can use to predict “Decision” for training and test sets and compute Kappa scores.
5. Logistic regression (You may use 0.5 as the initial cut-off probability and write a loop to try a range of values to determine the optimal probability).
6. KNN (Need to run a loop to select the optimal k).
7. Simple classification tree
8. Tree Pruning
9. Classification trees with Bagging
10. Classification trees with Random Forests (Need to run a loop to select the optimal mtry)
11. Classification trees with boosting (You may use 0.5 as the initial cut-off probability and write a loop to try a range of values to determine the optimal probability)
12. Support vector machine with Linear kernel (i.e., support vector classifier or soft margin classifier)
13. Support vector machine with Polynomial kernel
14. Support vector machine with Radial kernel

Please report the training and the test Kappa scores for each method in the comment.

**Part II – A brief report in PDF format**

Write a brief report describing what you did. You decide the length and content, and you can use your own formatting and structure, but consider the following as guidance:

1. Introduction and description of the problem and why it is important
2. Brief description of the original data set
3. Briefly discuss how you cleaned your dataset in terms of

* Handling missing values if any
* Correcting typos or measurement errors if any
* Recoding and/or augmenting existing variables if needed
* Dealing with outliers if any
* Transformations if needed
* Creating new variables if needed
* Removing variables if needed

1. Compare and contrast the results you get using various analytical techniques required above and determine which method gives you the best classification performance. Please mention explicitly your best overall run and best Kappa score.
2. Final conclusion and personal reflection including

* What you learned
* Obstacles you encountered and how you solved them (or why not)
* Overall reflection on the experience

**3. Grading Rubric**

The assignment will be graded using the following guidelines:

1. Data cleaning: 35%.

I will look for appropriate use of the following methods for data cleaning.

* Handling missing values
* Correcting typos or measurement errors (if any)
* Recoding and/or augmenting existing variables
* Dealing with outliers
* Transformations
* Creating new variables
* Removing variables

1. Implementation of all the required analytical techniques: 40%

* Quality of the code
* The code runs perfectly without errors or warnings

Note: Your script must run from the original, unmodified Admission data. All modifications must be done inside your script.

1. Report quality and insightfulness: 25%

**4. Milestones**

Although I helped you clean the first 46 columns, this is still a TIME-CONSUMNING project. To facilitate students to successfully complete this project, I set up four milestones below. Each time you reach a milestone, you need to let me know. To demonstrate you reach a milestone, you need to come to my office hours to briefly explain what you have done, and I will give you feedback if needed. Your preparedness for these short meetings with me will highly influence your participation grade.

**Milestone #1**

* Understand all fields in the dataset
* Know how and why you are going to add, delete, or modify features.
* Finish data cleaning.

**Milestone #2**

* About half of the analytical techniques required are implemented.

**Milestone #3**

* All the analytical techniques required are implemented.

**Milestone #4**

* Select the final model (i.e., the best model).
* Report write-up draft complete.