**Introduction to Data Science**

**Data Science for All Capstone Project Draft**

**July 25, 2019**

**Submitted by:**

**Tarji R. Kinsey**

**Clark Atlanta University**

**(404) 880-8082**

**tkinsey@cau.edu**

**Approval:**

**Date:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_

**(Instructor’s Signature)**

***Problem:*** Clark Atlanta University School of Business would like to have an idea of what to expect regarding undergraduate and graduate enrollment for the upcoming Fall 2019 term. More specifically, they would like to predict:

* New first-time student population
* International student population
* Undergraduate and Graduate student headcount
* Accounting major headcount
* Economics major headcount

Analyzing multiple rounds of enrollment data is a reasonably accurate way of determining whether there will be an increase or decrease in undergraduate and graduate level headcount. Additionally, when these data are compared to previously captured data (multiple rounds), additional information can be gleaned such as how many new students are enrolling, who is no longer enrolling, what majors most students are choosing, and is there an increase and decrease happening within the School of Business overall population.

***Dataset:***

* Clark Atlanta University Fall 2019 Preliminary Enrollment Data Set

The dataset “enroll\_capstone” gives a preliminary view of what the School of Business can expect regarding their student demographics and headcount for the upcoming Fall 2019 term. This file holds variables such: student id, pre-registrated indicator, name and state particulars, primary major, gender, international status and country of origin, student type and admissions admitted term, total credit hours, overall gpa, classification, etc.

There are no limitation as all variables and prior year data is available.

The intended prediction model will use the “ROUNDS” label with values (1, 2, 3) to indicate the data from prior years.

***Data Wrangling Steps:*** Several data wrangling techniques were performed in R on the data frame (enroll) to transform it into a format that can be analyzed. The techniques used, but were not limited to were: Basic Data Manipulation and Dealing with Missing Values. These techniques allowed me to remove or add variables and columns, change column names, modify blank or null data fields, and include functions for term hours.

**Remove or Add Variables and Columns:**

The original data frame included 60 variables. Some of the variables were not necessary to perform enrollment prediction techniques as they did not provide relevant information for the projection model. Removing 30 variables reduced the data frame to 30 variables. Some of the variables needed partnering variables to transition easily into the model. Two new columns were added to captures partnering variables.

**Changing Columns Names:** There are five variables included in the data frame that underwent a name change. When the data was imported into R Studio, some variables names were too long, and others did not accurately describe the particular data. Therefore, the names of the variables were changed to shorter, concise names that accurately describe the data in the column.

**Add Values to Blank or “NULL” fields:** Several columns displayed data that was either blank or the word “NULL” was found. A treatment column and a column that displayed standardized variables for the final outcome was created. In the treatment column, an if-else statement was written to place a 1 for the observation that was treated and a 0 otherwise. This information helped to accurately and easily determine the control and treatment groups.

**Data Function:** One column contained a function written to normalize the data for student term hours. Adding this column increased the data frame to 30 variables. It represents the students load status (FT or PT). Values assigned to this new column are FT = Full time Student and PT = Part time Student.

***Applying Machine Learning:*** The main question behind the problem is framed to look for patterns between data and labels that can be expressed mathematically as functions. Given an input feature, I have told the system what the expected output label is in hopes that the system will learn patterns based on this labeled data.

This is a supervised problem based on numerical values that is used in conjuction with the Linear Regression technique to predict various characteristics about the students expected to enroll Fall 2019 in the School of Business. The dataset (enroll) includes labels (UG or GR), values (0 or 1), and include the following predictors: Major, International Indicator, Student Type, Registration (“ROUNDS”), and Term Hours.

Cross-validation will be used to evaluate the success of Linear Regression model. Several rounds of cross-validation will involve separating out a sample of the dataset into a subset, performing analysis on each subset (training set), and then validating the analysis on the other subset (testing set).

***Key Findings & Recommendation:***

**Recommendation:**It is important that the School of Business is aware of the changes and current conditions of their student body as this will assist greatly with their future planning.

**Key findings for the School of Business are:**

**Finding:** The most heavily populated major within the School of Business for the upcoming Fall 2019 will be Business Administration; when overall headcounts are compared to previous Fall terms, it appears this major is continuously increasing by 3% rate while Economics continuously fluctuates.

**Finding:** Graduate level headcount within the School of Business for the upcoming Fall 2019 is expected to be 51 with 15 new first-time in college graduate students. Comparative to previous Fall terms, graduate level headcounts appear to continuously decreased by 20-23% rate.

**Finding:** Less than 4% of the student population expected to enroll for Fall 2019 are international students; comparative to prior Fall terms, the School of Business is expecting a **major** decrease (-40%) in this student population.

**Finding:** Many undergraduate students within the School of Business having the highest Fall 2019 term hours, have overall grade point averages ranging from 2.33 to 3.83.