**MSDS 6306: Doing Data Science**

**Case Study 02**

**Due: Sunday, December 9 11:59pm.**

**Description**: DDSAnalytics is an analytics company that specializes in talent management solutions for Fortune 1000 companies. Talent management is defined as the iterative process of developing and retaining employees. It may include workforce planning, employee training programs, identifying high-potential employees and reducing/preventing voluntary employee turnover (attrition). To gain a competitive edge over its competition, DDSAnalytics is planning to leverage data science for talent management. The executive leadership has identified predicting employee turnover as its first application of data science for talent management. Before the business green lights the project, they have tasked your data science team to conduct an analysis of existing employee data.

Your team has been given a dataset (**CaseStudy2-data.csv**) to conduct exploratory data analysis (EDA) to determine factors that lead to attrition. You should identify (at least) the top three factors that contribute to turnover. There may or may not be a need to create derived attributes/variables. The business is also interested in learning about any job role specific trends that may exist in the data set (e.g., “Data Scientists have the highest job satisfaction”). You can also provide any other interesting trends and observations from your analysis. The analysis should be backed up by robust experimentation and (where applicable) appropriate visualization. Experiments and analysis must be conducted in R. You will also be asked to build a model to predict attrition. Details are below.

**Deliverables:**

This is a group project so it’s the responsibility of the group members to collaborate accordingly. Team work is important.

The due date for submission is Sunday December 9 at 11:59pm. During live session 15, each team will sign up for a 15-minute time period in which each group member will present their presentation live to me. Grading for this assignment will be 5% of the overall grade and will be graded as a 100% if my grade is over 80%. If the presentation is assessed to be less than an 80%, the assigned grade will be recorded. To be clear, if a presentation receives an 85%, a 100% will be recorded. However, if a presentation receives a 75% a 75% will be recorded. As a matter of reference, every student that presented for Case Study 1 would have had a 100% recorded.

Similar to Case Study 1, each team member will need to record and upload to YouTube a **5-minute** presentation. To do this you can download Jing which is a free video software available at <https://www.techsmith.com/jing-tool.html> or use your preferred screen capture software. You can assume that your audience is the CEO and CFO of Budweiser (your client) and that they only have had one class in statistics and have indicated that you cannot take more than 5 minutes of their time. 35% of your grade will be based on the presentation. The presentation slides that include a link to your video should be in the Case Study Github repo before the start of the session. The goal is to communicate the findings of the project in a clear, concise and scientific manner. I will make the link available to everyone in the class so that your peers can benefit from your work. The links will be available for a week at which time you may take your video off of YouTube if you wish. Finally, include the link in your RMarkdown file.

I provided an additional data set of 300 observations that do not have the labels (attrition or not attrition). We will refer to this data set as the “validation set” and is in the file “**CaseStudy2Validation.csv**”. I have the real labels and will thus assess the the accuracy rate of your best classification model. 10% of your grade will depend on the accuracy rate of your “best” classification model for identifying attrition. You must provide a model that will attain at least 60% accuracy for the training and the validation set. Therefore, you must provide the labels (ordered by ID) in a csv file. Please include this in your GitHub repository and call the file **“Case2PredictionsXXXX.csv”.** XXXX is your last name. (Example: Case2PredictionsSadler.csv” would be mine.)

Create a GitHub repository with an RMarkdown file containing an executive summary, introduction to the project, all supporting code and analysis, and the group presentation. The repository should also include your prediction csv file and don’t forget to put the link to the youtube video in the RMarkdown file. Submit a link to the GitHub repository via the space provided for the Case Study 02 page in 2DS.

**BONUS:**

The team with the highest accuracy rate on the validation set will win the Bonus: 5 extra points and bragging rights!

**Rubric:**

50% RMarkdown File

35% Final Video Presentation

5% Initial Live Session Presentation (Unit 15).

10% Validation Requirement (Accuracy > 60%)