**MSDS 6306: Doing Data Science**

**Case Study 02**

**Due: Thursday, April 18 11:59pm.**

**Description**: DDSAnalytics is an analytics company that specializes in talent management solutions for Fortune 100 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 have been given a dataset (**CaseStudy2-data.csv**) to do a data analysis to identify factors that lead to attrition. You should identify the top three factors that contribute to turnover. There may or may not be a need to create derived attributes/variables/features. 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 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 an individual project. You all are a tight cohort which is so valuable; I want to harness the power that that represents while maintaining the individual-ness of the project. The general rules (dos and don’ts) that each student will be on their honor to abide by are:

Dos:

1. Do get excited about the project!

2. Do do all the coding.

3. Do discuss specific questions with your peers … i.e. if you want to add a heatmap

but don’t know how to code it … you may ask a peer to show you how they did theirs or ask them to point you to a good resource. You are encouraged to ask me as well. However, be sure to the Don’ts below!

Don’ts:

1. Don’t ever email/give your code / write up / materials related to your approach to the

problem or presentation to any student in the class.

2. Don’t code anything for anyone else.

In general, when deciding between two courses of action, if it feels like it is in the grey area, it is probably in the “Don’t” column. If you feel like you need clarity or a ruling on something you are not sure is appropriate or not, just shoot me an email.

The due date for videoed submission is Thursday, April 18 at 11:59pm (Live Session 15). I will meet with each student individually for 10 minutes during the time for Live Session 14. This is a mandatory meeting in which you will present your EDA and ask any questions that you might have developed. I will be available from 7 to 9pm on Tuesday April 16 for a voluntary 10ish-minute meeting where you can ask specific questions pertaining to your project. I am also available of course on email to answer specific questions. You will have the time reserved for live session to put the final touches on your project; we will **not** meet for Live Session 15 (April 18) and there will not be a live presentation component … only the video component. You know what I am looking for now and in reality, you will only get one shot at a great presentation. Make it a great recording and end the semester with a fantastic presentation / analysis.

Similar to Case Study 1, you 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 Frito Lay (your client). It is a diverse audience; the CEO is a statistician and the CFO has had only one class in statistics. They have indicated that you cannot take more than 5 minutes of their time. 40% of your grade will be based on the presentation. The goal is to communicate the findings of the project in a clear, concise and scientific manner. Finally, include the link in your RMarkdown file. Finally, finally make sure to email me the link to the YouTube video I will manually put them on the Google Doc which will be available on Friday April 19. The links will be available for a week at which time you may take your video off of YouTube if you wish. Please make sure and check out at least 3 of your peer’s presentations!

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 No Attrition.csv**”. I have the real labels and will thus assess the accuracy rate of your best classification model. 10% of your grade will depend on the sensitivity and specificity rate of your “best” classification model for identifying attrition. You must provide a model that will attain at least 60% sensitivity and specificity (120 total) 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 Attrition.csv”.** XXXX is your last name. (Example: Case2PredictionsSadler Attrition.csv” would be mine.)

I have also provided an additional data set of 300 observations that do not have the Monthly Incomes. This data is in the file “**CaseStudy2Validation No Salary.csv**”. I have the real monthly incomes (salaries) and will thus assess the RMSE regression model. 10% of your grade will depend on the RMSE (Root Mean square error) of your final model. You must provide a model that will attain at RMSE < .3 for the training and the validation set. Therefore, you must provide the predicted salaries (ordered by ID) in a csv file. Please include this in your GitHub repository and call the file **“Case2PredictionsXXXX Salary.csv”.** XXXX is your last name. (Example: Case2PredictionsSadler Salary.csv” would be mine.)

Create a GitHub repository named **CaseStudy2DDS** 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. Finally, make sure to email me the link to the YouTube video and I will manually put them on the Google Doc which will be available on Friday, April 19. Please make sure and check out at least 3 of your peer’s presentations!

**Due Dates:**

April 11 (Thursday) Live Session 14: Sign up on the Wall for a 10-minute spot:

Exploratory Data Analysis and Questions

April 18 (Thursday) at 11:59pm: Rmd, Powerpoint, and Final videoed submission due.

You may sign up for a voluntary meeting during the live

Session time.

**BONUS:**

The team with the highest sensitivity + specificity on the validation set will win the Bonus: 5 extra points and bragging rights!

**Rubric:**

35% RMarkdown File

35% Final Video Presentation (25% slide content, 25% presentation)

Minimal Stumbles / mis statements / etc. if you trip up more than a couple of times, reshoot the video.

Labeled Plots

5-minute time limit

Correct interpretation

Complete analysis – this means adding pvalues and conducting tests where appropriate (I expect

everyone to have a good handle on at least t-tests and linear regression to this point.

10% Validation Requirement for Attrition(Sensitivity and Specificity > 60%)

10% Validation Requirement for Salary(RMSE < .3)

10% Creativity and completeness in presentation and analysis.