**MSDS 6306: Doing Data Science**

**Case Study 1**

**Due: Sunday, March 9th 11:59pm CST**

**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 has been hired by Frito Lay to identify factors related to employee attrition. They aim to predict employee attrition in order to identify those employees who may be more likely to leave the company as well as identifying factors that are related to attrition so that they can best reduce the probability of attrition where desired.

The company has also indicated that they would like for you to build a model(s) to predict attrition and be able to measure the cost or savings impact of your model(s). Specifically, they have identified, based on [this article](https://mgrworkforce.com/employers/employee-retention-costs/#:~:text=The%20Society%20for%20Human%20Resource%20Management%20(SHRM),training%2C%20reduced%20productivity%2C%20et%20cetera%2C%20were%20considered.) and prior research that it costs between 50% and 400% of the employee’s salary to recruit a replacement for someone who has left the company. Additionally, they have estimated that if they give extra attention or incentives to employees that are about to leave they may be able to keep them from leaving. Frito Lay has estimated the cost of these incentives to be $200 per employee. These numbers can be used to help you estimate how much money a particular model could help save the company.

Frito Lay has provided a dataset (**CaseStudy1-data.csv**)to build your models and to do a data analysis to identify factors that lead to attrition. You should identify the top three factors that contribute to turnover (backed up by evidence provided by analysis). Note: your model(s) may have more than three variables / features, we just want to specifically identify the top three most important. You may be able to / find it useful to create derived attributes/variables/features.

The business is also interested in learning about any other interesting trends and/or 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.

This is an individual project.

**UNIT 8 and 9 Live Sessions:**

The due date for videoed submission is Sunday, March 9th at 11:59pm CST (Week 9). We will meet for Live Session 8 at the beginning for a DSNOW and a special topic.

We will also, meet for Live Session 9. I will answer any questions about the project that develop by that time and we will have a DSNOW!

**Deliverables:**

Unit 9 Live Session:

You will present an EDA. This is the [sign-up](ohttps://docs.google.com/document/d/1dDC4CQvQhbnCWZ7KjpwmNLrzZkqDkQ7XxvMpD66wNgU/edit?usp=sharing) sheet for a particular time. We will meet for Unit 9 Live session. We will start with a Q&A and then we will start the one-on-one presentations. For you presentation you should have:

* + A PowerPoint deck with well-designed and carefully created presentation quality slides. (No typos, labeled plots, consistent font, well organized slides.)
  + Include a summary of the number of attritions and non-attritions in the data set (percents and totals/counts.)
  + Visualizations that reflect potential importance / correlation and lack of importance / correlation of a variety of the features.
  + At least one KNN and one NB model with corresponding specificity, sensitivity and accuracy. Note that these models do not have to have at least 60% sensitivity and specificity at this point.
  + Additional analysis and/or plots and/or information that you have found in your exploration.
  + Be prepared to answer questions about your EDA including the interpretation of sensitivity and specify to the context of this case study.

Sunday March 9:

* **PowerPoint:** Create a PowerPoint deck to display your findings.
* **RMD File:** Format your code in an easy to read, well documented RMD File.
* **Video Submission:**

One of the final deliverables is a YouTube/Zoom video in which you present your findings. You will need to record and upload to YouTube a **7-minute** presentation or provide the link to your Zoom recording. 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 7 minutes. 20% 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 put the link to the YouTube / Zoom video in the Google Doc. The links will be available for a week at which time you may take your video off of YouTube / Zoom if you wish. Please make sure and check out at least 3 of your peer’s presentations! Make this your masterpiece. ☺

* **GOOGLE DOC:**
* <https://docs.google.com/document/d/1dDC4CQvQhbnCWZ7KjpwmNLrzZkqDkQ7XxvMpD66wNgU/edit?usp=sharing>
* You will also deliver predictions for the hold out (Validation) set as described below.

**Competition Portion:**

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 “Competition Set” and is in the file “**CaseStudy1CompSet 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 (60 each = 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 **“Case1PredictionsXXXX Attrition.csv”.** XXXX is your last name. (Example: Case1PredictionsSadler Attrition.csv” would be mine.) An example submission file can be found on Github in the Unit 8 and 9 folder: **Case1PredictionsClassifyEXAMPLE.csv**.

Create a GitHub repository named **CaseStudy1DDS** with a RMarkdown file containing an executive summary (in the Readme.md), introduction to the project, all supporting code and analysis, and the slides for the presentation. The repository should also include your prediction csv file and don’t forget to put the link to the YouTube / Zoom video in the RMarkdown file. Submit a link to the GitHub repository via the space provided for the Case Study 01 page in Canvas. Finally, make sure and put the link to the YouTube / Zoom video on the Google Doc.

The executive summary should include:

* overall question of interest and setting
* the factors you identified that were significantly related to attrition
* the performance of your model(s)
* any additional inference or finding that you came across in your study

**Due Dates:**

Sunday, March 9th at 11:59pm CST: Rmd, Powerpoint, Predictions and Final videoed submission due.

**BONUS:**

The individual with the highest sensitivity + specificity (both at least 60%) on the classification validation set will win the Bonus: 3 extra points and bragging rights!

**Rubric:**

10% RMarkdown File

30% Final PowerPoint and Video Presentation (15% slide content, 15% presentation)

**Minimal Stumbles / mis statements / etc.** if you trip up more than a couple of times, reshoot the video. It will be much better with the practice!

**Labeled Plots with readable labels**

**7-minute time limit**

**Voice inflection**

**Creativity**

40% Analysis

Correct interpretation

Appropriate analysis (tests, methods, descriptions)

**Robust analysis** – this means supporting your findings with plots, charts, confusion matrices, sensitivity and specificity, F1, and sound logic that ties your points together. An example of this “logic” is to evaluate any differences between your model making a false positive and a false negative.

20% Validation Requirement for Attrition (Sensitivity > 60% and Specificity > 60%)

**FAQ and Comments:**

**1. Question: In the dataset, what does Relationship Satisfaction mean...(relationship to manager, to peers)**

Relationship satisfaction with manager.

2. Advice: Don't eliminate variables simply because they have a high correlation with one another.  This is an indication that they do share some information although the information they don't share may be correlated with the response individually.

3. Advice: When plotting and exploring attrition, the percentage of those who left is probably more useful than the count.

4. Question: Is the dataset, is the distance from home in miles or kilos?

We don't have that information (however we do know whether its high or low)

5.  **Question: In the dataset: what is the definition of pay rates: Hourly, Daily & Monthly.  These values to not seem to relate to each other or the Monthly Salary (which is different than Monthly Rate).**

We don't have that information (however we do know whether they are high or low). They may or may not relate to each other or the monthly salary (this is for the student to infer and decide whether theres any correlation or whether this is a useful feature for attrition)

6**. Question: In the dataset: we do see that Job Levels go from 1-5 and assume that 1 may symbolize a lower level employee, but this is not defined.  Though this level does have evidence of a positive linear relationship with Monthly Income, it does not seem to correlate well with the Job Titles. in other words someone with a Director can be a 2-5, and manager a 3-5.**

Yes we can assume 1 is a lower job level than 5.

**7. Question: In the dataset, does overtime mean Hourly vs. Salaried worker?**

We can assume that people with overtime are non-exempt / hourly employees.

**8 Question: In the dataset, Performance Ratings are only 3 & 4, is there a mistake?  Unless a corrupted system, hard to imagine ratings consistently high, even as 2 still means "good".**

**It is self-reported data, think about why the employees may only answer 3 and 4**

No this is the only data we have, there is no mistake.

**9 Question: In the dataset, does Training times mean: hours, weeks, or instances and over what period?**

Training times last year means number of training sessions attended by the employee.

10. Question: Do we have any information on the other columns?

Yes…. Here is the break down for a few columns. Other columns don’t have any additional information but may be useful in predicting attrition. This is often the case and can be a feature rather than a bug in that it may prevent biased decisions. You may plausibly speculate on their meaning (just make sure it is clear that it is speculation) or you can simply report that no information was given and report what you are able to about the relationship and what it might mean. Here is some additional information on a few of the columns:

Education  
1 'Below College'  
2 'College'  
3 'Bachelor'  
4 'Master'  
5 'Doctor'

EnvironmentSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

JobInvolvement  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

JobSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

PerformanceRating  
1 'Low'  
2 'Good'  
3 'Excellent'  
4 'Outstanding'

RelationshipSatisfaction  
1 'Low'  
2 'Medium'  
3 'High'  
4 'Very High'

WorkLifeBalance  
1 'Bad'  
2 'Good'  
3 'Better'  
4 'Best'