**MSDS 6306: Doing Data Science**

**Case Study 02**

**Due: Saturday, April 21st 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 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.

You have been given a dataset (**CaseStudy2-data.csv on AWS S3 in the msdsds6306 bucket**) 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). 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. Finally, you will develop an RShiny App to visualize some of the relationships or lack thereof. Details are below.

This is an individual project.

**Deliverables:**

**UNIT 14 and 15 Live Sessions:**

The due date for videoed submission is Saturday, April 21st 11:59pm CST (Week 15). We will meet for Live Session 15 at the beginning for a DSNOW and to answer any general questions although there will not be a live presentation component … use this time finish up your project. I will be available to meet during the Live Session Time for optional / voluntary meetings… I will have sign up times on the [Google Doc](https://docs.google.com/document/d/1unptcaTgPwcWKVF54HpNRn0K_t2HITK9Aw2lyQP73vU/edit?usp=sharing). Make it a great recording and end the semester with a fantastic presentation / analysis! Make this your MASTERPIECE!

We will also, meet for Live Session 14. I will answer any questions about the project that develop by that time and we will have a DSNOW!

**Further Details:**

Similar to Case Study 1, 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 of their time. 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!

GOOGLE DOC:

<https://docs.google.com/document/d/1unptcaTgPwcWKVF54HpNRn0K_t2HITK9Aw2lyQP73vU/edit?usp=sharing>

I provided an additional data set of 300 observations (also on AWS S3) 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 “**CaseStudy2CompSet 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 **“Case2PredictionsXXXX Attrition.csv”.** XXXX is your last name. (Example: Case2PredictionsSadler Attrition.csv” would be mine.) An example submission file can be found on AWS S3 in the msdsds6306 bucket: **Case2PredictionsClassifyEXAMPLE.csv**.

I have also provided an additional data set of 300 observations that do not have the Monthly Incomes. This data is in the file “**CaseStudy2CompSet 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 a RMSE < $3000 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.) An example submission file can be found on AWS S3 in the msdsds6306 bucket: **Case2PredictionsRegressEXAMPLE.csv**.

**Notes on models to fit**: ***IMPORTANT:*** **Many of you may have worked with other models, (Random Forest, logistic regression, neural networks, etc.). These will be covered in depth in the ML1 class and Stat 2. For this project, we want to deeply explore the usefulness of KNN, Naïve Bayes and Linear Regression and thus all classifications and predictions must be produced from these models (the models we studied in this class.). However, if you would like to submit a separate csv file of prediction from other models I would more than happy, thrilled actually, to also evaluate those results in the leaderboard I produce.**

Create a GitHub repository named **CaseStudy2DDS** 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 02 page in 2DS. Finally, make sure and put the link to the YouTube / Zoom video on the Google Doc found on 2DS.

You will also provide an RShiny app that displays and allow the user to explore at least one of the relationships you studied.

Finally, create a Knit file out of your RMD and display it on your GitHub (with Jekyll!) Site you created in Unit 12. Include the link to your YouTube / Zoom video as well as the link to your RShiny App.

**Due Dates:**

Saturday, April 21st 11:59pm CST: Rmd, Powerpoint, RShiny App, Github Jekyll Page 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!

The data scientist with the lowest RMSE on the regression validation set will win the Bonus: 3 extra points and bragging rights!

**Rubric:**

10% RMarkdown File

25% Final Video Presentation and analysis(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

7-minute time limit

Voice inflection

Roll in suggestions from project 1

Creativity

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

everyone to have a good handle on at least t-tests, KNN and Naïve Bayes classification and KNN and linear regression to this point.

10% RShiny App

25% Analysis

Correct interpretation

Appropriate analysis (tests, methods, descriptions)

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

10% Validation Requirement for Salary(RMSE < $3000)

10% Knit RMD and YouTube link on a tab on your GitHub Site … or smoothly integrated into your Github page that uses a Jekyll theme.

**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'