MSDS 6306: Doing Data Science

Case Study 02

Due: Saturday, April 16th 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.

You 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 (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. 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 recognize 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, just shoot me an email.

UNIT 14 and 15 Live Sessions:

The due date for videoed submission is Saturday, Dec 5th at 11:59pm (Week 15). We will **not** meet for Live Session 15 and there will not be a live presentation component … use this time finish up your project. You know what I am looking for now and in reality, you will only get one shot at a great presentation. With that said, I will be available to meet during the Live Session Time for optional / voluntary meetings… I will have sign up times on the wall as usual. Make it a great recording and end the semester with a fantastic presentation / analysis!

However, we will meet for Live Session 14. Live Session 14 will be for any Data Science News of the Weeks that needs to be completed and I will answer any questions about the projects that develop by that time (asked via the Wall.) With that said, our Unit 14 class is currently on Thanksgiving so I will need to reschedule that one. I will hold that class on **Sunday evening Dec 1 at 7pm**. Please try and make the class but if you cannot, it is required to watch the video.

Further Details:
Similar to Case Study 1, you will need to record and upload to YouTube a **7-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 7 minutes of their time. 30% 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 video in the Google Doc. 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 "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 GitHub: **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 GitHub: **Case2PredictionsRegressEXAMPLE.csv**.

Notes on models to fit: ***IMPORTANT:*** **First and foremost, you may use outside models that we have not covered but it must be in addition to models that we have covered in class. This means for classifying attrition, you must use either k-NN or naive Bayes but may also use other models (logistic regression, random forest, LDA, SVM, etc) as long as you compare the results between the two or more models. You may then use any of the models to fulfill the 60/60**

sensitivity/specificity requirement. This goes for regression as well; you must use linear regression but may include additional models for comparison and use in the competition (LASSO, random forest, ensemble models, etc.).

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 slides for the 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 put the link to the YouTube video on the Google Doc.

Finally, create a Knit file out of your RMD and display it on your GitHub Site you created in Unit 12. Include the link to your Youtube video (and a link to your RShiny app too if there is one!)

Bonus: Up to 3 points: Create an RShiny App to help visualize you results. The amount of bonus points awarded will be based on correctness, creativeness, effectiveness of the visualization / app.

Due Dates:

April 16th (Saturday) at 11:59pm: Rmd, Powerpoint, and Final videoed submission due.

BONUS:

The data scientist with the highest sensitivity + specificity (both at least 60%) on the classification validation set will win the Bonus: 5 extra points and bragging rights! This bonus is between all 4 sections. 1 prize between all 40+ students.

The data scientist with the lowest RMSE on the regression validation set will win the Bonus: 5 extra points and bragging rights! This bonus is between all 4 sections. 1 prize between all 40+ students.

Rubric:

30% RMarkdown File
30% Final 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
- 7-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, KNN and Naïve Bayes classification and KNN and linear regression to this point.

10% Validation Requirement for Attrition(Sensitivity > 60% and Specificity > 60%)
10% Validation Requirement for Salary(RMSE < $3000)
10% Creativity and completeness in presentation and analysis.
10% Knit RMD and YouTube link on a tab on your GitHub Site.

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: In 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 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 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.