Employee Attrition - Prediction

# **Vinay Nagaraj DSC680 - T302 Applied Data Science (2217-1) https://github.com/vinaynagaraj88/vinaynagaraj88.github.io**

**Proposal Document** - <https://github.com/vinaynagaraj88/vinaynagaraj88.github.io/blob/main/P1%20-%20Employee%20Attrition/Documentation/Project%201_Proposal.docx>

# Week 2 check-in

Any surprises from your domain from these data?

* I have not come across any challenges or surprises with respect to the domain. The data so far is looking good and I am currently working on reviewing the data for trends or data classification based on different features. I have a plan to review some of the references mentioned to see how a ML model can be integrated into my dataset and try to come up with predictions about Employee Attrition. I am planning to spend the next week on completing EDA on the dataset and see what is the best way to use a ML model in our prediction.

The dataset is what you thought it was?

* The dataset is close to my expectations. I see a lot of records with Attrition feature = ‘No’ as compared to ‘Yes’. So there is a slight imbalance in the dataset and I will probably need to look around for oversampling or under sampling options to balance the dataset.
* One other thing I have noticed during data clean-up is that certain features in the dataset have only one valid value for all records and I plan on eliminating those features before I start working on my feature selection and models.

Have you had to adjust your approach or research questions?

* At this moment, I am going ahead with my plan. I am following the CRISP-DM methodology which I have kind of used in all my previous projects.
* As I mentioned above, there is a slight imbalance in the dataset and that is something I found out during my data understanding. So I will need to look at options of oversampling or under sampling. That wasn’t something I was aware of.
* Also as per the reference I was using, I have seen that they have compared models build with different methods, that looks interesting to me. So this has given me a lot of options and earlier I was just going to build one model, I might go for building different models and compare their performance.

Is your method working?

* So far, my approach seems to be providing me the expected results. After going through the references, it has helped me through which I might go ahead and do a little bit more than initially planned, by building multiple models rather than just building one like initially planned.
* I have almost completed the data understanding and performing some graph analysis. Once I am comfortable with my graphs, then I will proceed further with feature selection & building models.
* I have planned to use Python notebook and haven’t had any issues with it so far. Everything has been going as per the initial plan.

What challenges are you having?

* I haven’t run into any challenges yet either with technology or with data. Everything has been going pretty smooth so far for this project.
* I have spent more time on understanding the GitHub pages when compared to the time I have spent on my project. So, this week I would like to concentrate more on my project and hope to complete some model comparisons by the end of next week.
* There is a lot of insight I am getting on the features and I have an idea which might be my most important features which can determine the attrition rate. Monthly Income, Job satisfaction and worklifebalance might be some important features in my modeling.
* So, at this time everything is going smooth for me, but next week might be interesting.