#66DaysOfData – Day 2: Never Start from Scratch

Cooking is an art, and many times you can and should start from scratch with fresh ingredients. As for data science, you probably do NOT want to do the same. Odds are that whatever question you are trying to answer, someone has already asked it, or at least asked something similar. This means that after you have your initial question (which again, will change many times before the end of the project), it is time to do some research. Learning more about your question, and about other people that might have insight on it, will help you understand the problem a lot better and, if it exists, lead you to some data that might be useful.

For me, one of the main roles of a Data Scientist is to investigate, understand the problem deeply so that we can determine if we have the data needed or what can be done with the available data. As in any iterative process, if we want to discover and create valuable models, we need to search and read, A LOT. It is very, very rare when we get to work on a project where we decide what kind of data we need and help design how we are going to get it. So, we need to make do with what is available, sort of like turning dirty clay into great pottery. The more you understand the clay, the better pottery you can create.

Since I already have my main question:

Can we predict if a new employee will stay for at least 3 years at a Tech company by using OCEAN/HEXACO personality traits model?

Today was the day to start investigating. And it turns out, I am NOT (obviously) the first person to try and find a relation between OCEAN/HEXACO personality traits and attrition (quitting). I found a few interesting papers:

[Can personality predict longitudinal study attrition? Evidence from a population-based sample of older adults.](#)

[Identifying Personality Traits Associated With Attrition in Systematic Training for Effective Parenting Groups](#)

[A person-centric investigation of personality types, job performance, and attrition](#)

A couple of these papers are pay only, so I cannot share them, but let me give you a ludicrously brief summary:

There seems to be positive, independent positive relation between personality traits of extraversion and neuroticism, and facets such as overcontrol or lack of control with attrition. There is also a negative independent relation between agreeableness trait and resilience facets.

In short, extraversion, neuroticism and high and low levels of control push people towards quitting, and agreeableness and resilience push people towards not quitting.

In other words, there have been some studies on linking OCEAN personality traits and attrition, and there are several interesting relations we can study. I also found some interesting results, which lead me to think of a few interesting options when I inevitably run into the reality that there is no such thing as the perfect dataset to build the model in my head.

If you would like to know more about OCEAN/HEXACO personality traits, please let me know and I will be happy to write a bit about it in the next few days. But do know that I will try to explain the important parts either here in the articles, or directly on the Jupyter Notebooks once I start working with them.

Next Time – Where to find raw data to play with?

Jack Raifer Baruch

Follow me on Twitter: @JackRaifer

Follow me on LinkedIN: jackraifer

About the Road to Data Science - #66DaysOfData Series

Road to Data Science series began after I experienced the first round of Ken Jee´s #66DaysOfData challenge back in 2020. Since we are starting the second round of the challenge, I thought it would be a good idea to add small articles every day where I can comment my progress.

I will be sharing all the notebooks, articles and data I can on GitHub: https://github.com/jackraifer/66DaysOfData-Road-to-Data-Science

Please do understand I might have to withhold some information, including code, data, visualizations and/or models, because of confidentiality regards. But I will try to share as much as possible.

Want to follow the #66DaysOfDataChallenge?

Just follow Ken Jee on twitter **@KenJee_DS** and join the #66DaysOfData challenge.

You can also reach out to me at any time through **LinkedIN** or **Twitter.**