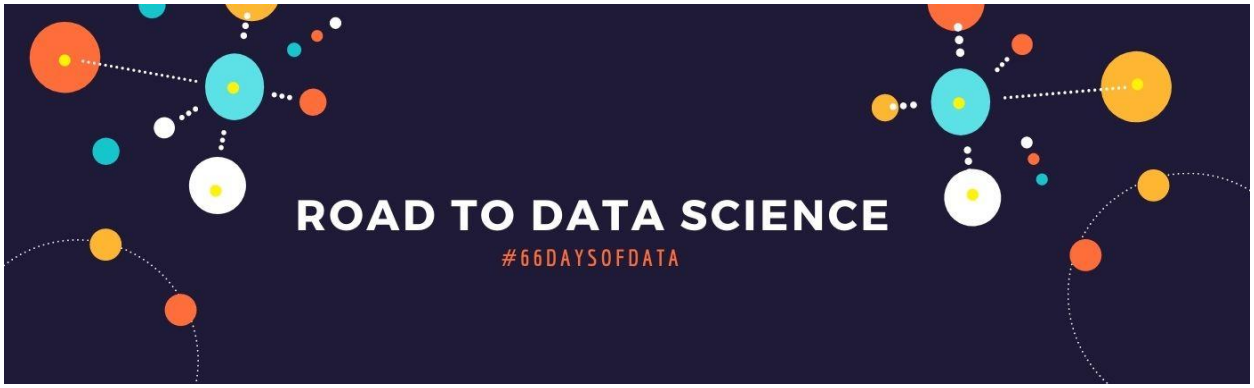

#66DaysOfData – Day 3: Digging Deeper and Finding Data

Yesterday I found some interesting information about personality and attrition which I describe in my Day 2 Article here. Reading a couple of papers gave me some ideas about the interesting possibilities of creating a model that can predict the probability of attrition at an X time interval. So, I decided to take some time to dig deeper and analyze some interesting business questions.

Let us remember that when building projects, we must always look at the endgame, meaning, how can this become a product that people what to use. Taking advantage of the fact that I have a background and experience in organizational psychology, I called a few acquaintances that work in HR recruitment with one simple question: How much value do you see on a system that can predict attrition? All of them agreed it would be valuable, but the intention of my call was to understand the specifics of why that would be the case.

Turns out that companies have specific costs for talent searching, recruitment, hiring and onboarding. Others also must contemplate initial training. This is mostly important for entry level positions, which is where the biggest numbers of people are hired and where turnover and attrition tend to be the highest. Another brilliant tidbit is that most of them already have a good idea of the cost and the return on investment (in this case, how long a person must stay on the job) for the costs to be offset. Turns out that for some, when training is 2 weeks or less, they offset the cost at 3 months and consider it a victory when they reach 6 months. For companies with longer training, more than 2 weeks, they need people to stay 6 months to offset and 12 months to consider it a good investment. Let us consider that most of the people I talked where from the call center (support and customer service) industry.

Knowing the probability of attrition at these points in time, can help them in many ways, from improving the hiring process to better projecting the costs of the operation.

Once again, we follow the evidence, and I will change the question we are chasing once again, this time to: **Can we predict the probability of attrition at 3, 6 and/or 12 months in call centers through OCEAN personality traits?**

Each time, our question becomes more specific, more interesting and, mainly, more operational (and do not worry, it will keep changing and that is completely normal and expected).

Since we do know that there is some clear evidence that this is possible, our next step is to find some interesting datasets. I was not expecting to find any data that combined OCEAN personality and specific information on hiring and attrition, and I was unfortunately right (at least to the extent of my search on google).

But like always we keep moving forward (with quick jumps backwards), and I did find some interesting sets we can use, if at least to build a theoretical model to start with.

A dataset from an OCEAN project by automoto on Github with answers from 307,313 people, including other information like age, biological sex and geographic location.

Raw datasets of OCEAN inventories from OpenPsychometrics with 19,719 responses.

A Kaggle OCEAN dataset with 1,015,32 questionnaire answers.

Since we now have a good idea of where we are going, we have some interesting data to investigate and we do know that our project has some merit, both theoretically and businesswise, we can now move on to setting up our environment for EDA (Exploratory Data Analysis), which we will do in Jupyter Notebooks (the why also on the next chapter).

For those of you interested in learning more about OCEAN personality model, here is a short basic intro on YouTube.

Next Time – Starting a Jupyter Notebook for the Project

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.**