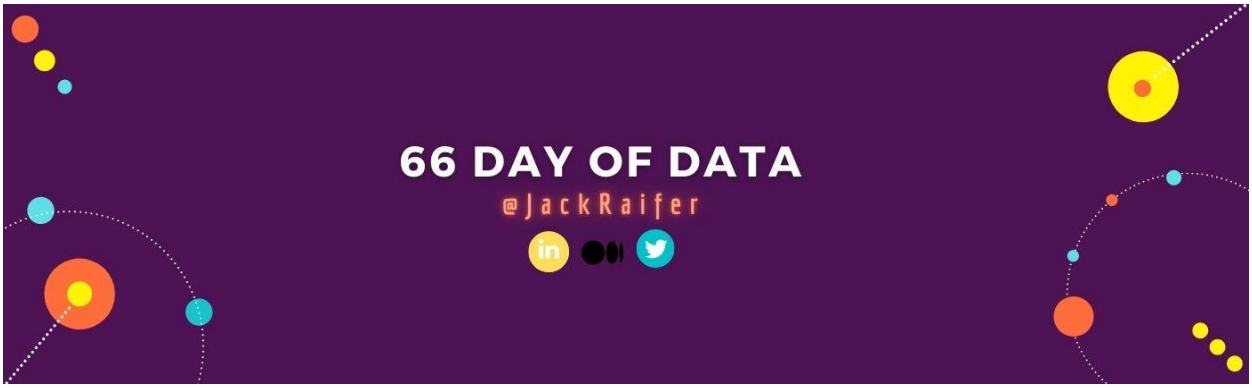

#66DaysOfData – Days 10 & 11: Never reinvent the wheel and Biostatistics to the Rescue

Last time I felt a bit disappointed about the data I had found, but this was to be expected, when you set out to answer a question, there will be many disappointments along the way. The important thing is that it got me moving, and it reminded me of an old common saying: *"Never reinvent the wheel."* This seems a bit counterintuitive when it comes to building new things and innovating, but this is a misconception.

Not reinventing the wheel means you should not strive to create something in a vacuum, after all, the reason we have all the technology we do is because someone, somewhere took what had been invited before and improved on it. Sometimes, this meant a small but important improvement, other times it led people to create completely new things. Nonetheless the important lesson here is to first try and build something on top of something else that already works instead of trying to invent it.

For the attrition prediction model this got me thinking: Where have they used specific inputs to generate predictive probabilities of specific outcomes? At least one answer is biostatistics. This field has spent decades improving their methods to create amazing probabilistic models for disease. For example, using a set of variables such as age, history of previous disease, family medical history and behavior such as smoking, there are particularly good models for predicting the current probability of someone developing lung cancer in the next year. This is called risk.

The same way, there are great biostatistics models used in car insurance that take into account the age of the driver, years driving, previous accidents, the type of car, and believe it or not, the vehicle color (silver cars have 10% higher risk of crashing and red cars are 7%). These models are used by insurance companies to set your premiums. If you want to save money, buy a white car.

Since we have a clear methodological approach to calculating risk, what can be done with our model? We can use the results from the studies we have and create a biostatistical formula to calculate the risk of attrition at 3, 6, and 12 months. This will allow us to build a base model that we can then test out.

Another great piece of news is that for the Turnover Dataset we got some new information about the contents. It turns out that the in the HR industry there is quite a bit of a never talk about it debate about the meaning of certain words, specifically, attrition and turnover. For me, attrition has always been people quitting the company, for whatever reason, but for others it means when people leave on amicable terms, even more confusing, there is also a dictionary definition that just specifies the decrease of a workforce by not replacing people who leave. On the other hand, turnover has been used to mean the flow of people coming and going from an organization, to people who leave in bad terms. Now, before we untangle all of this, the good news is that the Turnover Dataset does consider the term turnover as people leaving the company (my original definition of attrition), and that the 5 personality columns are actually OCEAN results with different names.

This means that we DO have a dataset in which to create a model. The only downside to this is that the dataset is tiny, only 1129 observations.

So finally, here are our options:

1. Build a model with the Turnover Dataset, even though it is small, it could be insightful and we can deal with the small dataset using Naives Bayes methods to lower the bias inherent in them.

2. From the results of the studies we found on our initial investigation we can use Biostatistical methods to build a theoretical algorithm.

3. We could use either or both of the above to create synthetic data on our other datasets. This could improve our model or develop into a useless model.

My take on it is, since I have some, we still have a bit over 50 days to build this project (it is always good to have a deadline), is to do all 3, and more if we have the time. We should start by using the small but complete dataset to build a predictive model, and see where that leads us,

Next Time – Regression, Decision Trees, Random Forests, choosing an algorithm.

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.**