BrainStation Data Science – Sprint 0

Problem area:

Improving the design of clinical trials

Often times, many resources are spent on designing clinical trials, analyzing them, and actually reaching a proper conclusion. In order to save time and expenses, the aim is to identify certain factors or parameters that cause clinical trials to fail. Some of these factors could include, not enough funding, not enough people to conduct the study, inadequate study design, ethical and scientific issues, loss of staff members etc.

The user:

Many large pharmaceutical, investors, and other research companies would benefit by understanding what parameters to focus on when designing a clinical trial, and whether or not their clinical trial would be terminated, or perhaps what factors would cause their clinical trial to fail.

The big data:

There was a similar study conducted by: https://www.nature.com/articles/s41598-021-82840-x#Tab2 in which they used machine learning to determine terminated clinical trials, specifically, feature engineering and embedded learning. This is essentially a categorical prediction approach, where we will be predicting whether the outcome of the clinical trial will fail. The previous study also used models similar to NLP, where they would match similar keywords in the titles and descriptions for aggregated calculations and statistical analysis in similar trials. For example, the word ‘cancer’ and ‘oncology’ would be matched together, and then we would compare all the trials that have ‘cancer’ and ‘oncology’ with their successful or failed outcomes.

The impact:

By creating this analysis and model, we can potentially help save millions of dollars in funding by sponsors, investors, and the government. Additionally, by helping medical companies better understand why some trials may fail, we can improve the design of clinical trials which will save a lot of time and resources during the planning process. Lastly, terminated trials due to lack of funding or lack of participants results in a loss of scientific contribution in the community, which is why creating better designs of clinical trials is critically important.

The data:

I’ve looked into datasets on the Canadian government website and the American government website, and they sort of pool in their datasets into one big database called ClinicalTrials.gov (https://clinicaltrials.gov/). This dataset is also largely used by other studies which performed this same analysis. Additionally, this dataset is enormous

(483,077 studies), which includes the Study title, ID, status, brief summary, sponsors, conditions, patient information and other important information that will help us in the analysis. I believe the most important column to consider is the ‘Study Status’ as this determines whether the study is still in progress, completed, interrupted or terminated. There will definitely be a lot of data cleaning to do, and we need to pick out the important information in order to have a better scope for the solution.

The alternative:

There is a huge problem in the housing market where often times, fresh grads cannot find places to live especially in the GTA. It would be very interesting to provide some insight into why houses are so expensive. We can create a Real Estate prediction model where you can predict the price of your home with factors including location, number of rooms, square feet, etc. I would probably look into cities with lots of data on current homes, how much they are selling for and look at what causes homes to be so expensive.