Project 3 Summary - Breast Cancer Prediction
Classification Project
Flora Xinru Cheng
Oct. 30, 2019

TL;DR
(Summary of the summary, since I tend to get too wordy for my blog)

For this project, I intentionally chose a simpler dataset and a clearly-defined binary classification problem. I made this choice in order to analyze the problem more thoroughly, and to have more time to practice presenting the results than I was able to do for previous projects (see Project 2 summary here). My prediction results were very good (AUC ~ 0.99 for top 4 models), and the natural next step would be to expand the problem to include image classification on biopsy images before classification.

The main lesson learned is to start thinking about the story and specific business use case from the design phase, and focus my analysis on aspects that would help the story.

Data
Before and after training and testing my classification models, I searched for additional datasets to complement this dataset, without much success. I probably could have foreseen that when I found out how popular this dataset is (#6 on UCIrvine Machine Learning Repository). I would've liked to have more realistic features such as patient age and race, instead of just the features of the tumors. Part of the challenge is my lack of domain knowledge in this field, as well as information about patients are likely confidential.

The biggest issue with the dataset is the multicollinearity of the features. All the features except for the target were geometric features calculated from biopsy images of tumors, and many of these features have mathematical relationships with each other, such as radius_mean and area_mean, concavity_mean and concave_points_mean. Upon exploratory data analysis, I decided to only use the first set of unique features (mean values), and discarded the features that could be leaky variables, and don't provide new information in terms of which features are the most important in predicting the diagnosis.

Further Work
To continue working on this and expand it to include image classification, I would first look into the papers that cited this dataset (linked on the data source), to see what image data they used. I would also check those papers to see if there are more advanced models used to calculate these features from biopsy images, before trying to repeat the calculation to get geometric features similar to the ones I used in this dataset.

Lessons Learned

I achieved the goals I set for myself for this project, among them a smaller, more manageable scope in the design phase, better file organization, and improved presentation delivery. However, I decided not to do a Tableau dashboard for feature importance towards the end of the project. This was because I realized when making my slides that having that information would not significantly enhance my data story. Learning from this experience, I need to start thinking about the story and the business use case from the project design phase, and target my analysis of modelling results accordingly. Also, spending a significant amount of time on a result does not justify including it in my presentation, and I'm glad I took it out of my presentation slides and spent time focusing on my story and practicing the talk.
I was really pleased that taking the initiative to start a presentation_practice channel on Slack and actively inviting my peers to practice together paid off. We all agreed that, compared to practicing on our own, getting feedback from our peers was incredibly helpful. I plan to continue

I started on the SQL challenge early, because I knew there was not a whole lot of data cleaning or preprocessing required for this dataset, which worked out nicely.

Using the cookiecutter template and JupyterNotebook Extensions (such as table of contents and execution time) really helped with file organization and efficiency, so I'll continue using those. For the next project, I want to force myself to commit and push to GitHub remote more frequently from the command line, not just at the end. Also because the dataset I used for this project is small, I didn't work with AWS or Google Colab, which I would probably need for unsupervised learning projects. For my next project, I also want to discuss ideas and issues with my peers more.