Evaluating Local Feature Explanation Algorithms on Germline Mutation Pathogenicity Classification Model

Team members: Taykhoom Dalal

Problem:
Classifying the pathogenicity of germline mutations in cancer patients is an important task, allowing doctors to prescribe drugs and medications that have been proven to effectively target these mutations and hopefully shut down the cancer's ability to replicate and metastasize. However, classifying the pathogenicity of germline mutations is extremely difficult, as it is often unclear how germline mutations interact with the cancer, and thus it is typically up to clinical geneticists to use their own knowledge, relevant literature, public databases, functional studies, and other resources to ascribe pathogenicity to these germline mutations. By creating a machine learning model that can aggregate this information automatically and use it to predict the pathogenicity of the mutations, we could potentially cut down the time and effort required to get this important information to clinicians, allowing them to prescribe drugs faster and alleviate issues caused by the cancer sooner. However, building such a model is not enough, it is important for researchers to understand exactly which features of a mutation contributed the most to the prediction of pathogenicity for that mutation. This is where my project comes in.

Approach:
There are several prominent feature explanation algorithms that have cropped over the years, and thus I plan to test these various algorithms on my classifier to see what features they predict contributed how much weight to the end prediction of a variant's pathogenicity. I plan to test the following algorithms: LIME[1] (Local Interpretable Model-Agnostic Explanations), SHAP[2] (SHapley Additive exPlanations), MAPLE[3] (Model Agnostic Supervised Local Explanations), TCXP[4] (Tree Classifier eXPlanation), and maybe some others (anchor Explanations[5]) depending on if I have time. I may also try to change the model from a RandomForestClassifier (which is what it is currently) to another model, such as a boosting method (although TCXP won't work).

Evaluation of Approach:
One method I will use to evaluate my approach will be to take a look at the features that each algorithm predicts to have a large effect on the prediction, and then remove that feature and see how much the prediction changes (my model outputs the prediction as well as a prediction probability and thus I can measure the change in this value). Another method I can use is based more on the concept of "human-grounded evaluation" where I can investigate which features are given how much weight towards certain predictions and try to evaluate based on outside research whether this makes sense or not, although this is limited by my lack of expertise.

Dataset:
I will be using a training dataset containing 12678 germline mutations to train my classifier, and a test data set of 844 mutations for which we have labels (all of them are labeled as pathogenic).

[1] https://github.com/marcotcr/lime
[2] https://github.com/slundberg/shap
[3] https://github.com/GDPlumb/MAPLE
[4] https://towardsdatascience.com/why-did-your-model-predict-that-4f7ed3526397
[5] https://github.com/marcotcr/anchor