Aaron Morales
BME 205
Fall 2025
11/6/25

Assignment Explanation

Data Preprocessing

Preliminary data cleaning was to first assess the various distributions of our PRS scores, ensure there are no N/A entries, and determine variable types. Next, the only non-numerical feature was ancestry, and this had four unique data types(AMR, AFR, EUR, A). After leveraging sklearn's OneHotEncoder to drop the original column and concatenate our OHE columns. Lastly, used StandardScaler on our feature columns(despite already being normalized) and made sure to separate features into the X df and labels into the y df.

Machine Learning Model(s)

My goal was to make use of a linear and non-linear ML model. Firstly, logistic regression models say the probability of disease using a logistic function to model the relationship between independent variables and the probability of a binary outcome. This was an obvious choice for its interpretability and continuous classification (probability). Secondly, random forest builds a large number of individual, randomized decision trees and then combines their outputs to make a more accurate prediction. I made this choice to potentially capture the non-linear relationships between numerous features(if they exist in the data). Though I initially ran into a hiccup attempting to fit the model to 5k+ feature space and required a work around.

Instead of training on the entire feature space, I selected top features based on individual ROC AUC scores (threshold ≥ 0.52) to reduce noise and dimensionality. This new 144 space of the most deterministic features allowed for detectable signal in a lower dimensional and computationally taxing environment.

Next, I used crossvalidation and gridsearch to determine the optimal hyperparameters for my models as well as avoid overfitting between training and testing datasets. I came to the conclusion that even with a large number of trees and increased maximum depth, random forest was rarely beating logistic regression even when obviously overfitting. The best results are depicted below

```
Logistic Regression:
  AUC mean ± std: 0.7806 ± 0.0048
  Tested with top 144 features (5-fold CV)
```

Closing Remarks

Feature selection along the lines of individual ROC AUC scores seems like too naive of an approach for this task. Within a biological context, those at risk of developing breast cancer is due to a (not always linear) combination of various lifestyle, demographic, and mutational traits.

Aaron Morales
BME 205
Fall 2025
11/6/25

In order to better understand the potential underlying relationships between PRS features, I spent a lot of time working with principal component analysis.

On the raw feature space, I determined 90% variance at 1500 comps, 85% variance at 1100 comps, and 80% variance at 830 comps. I figure these ‘elbows’ in variance, component space could translate to some predictive power, but that is simply not the case. PCA maximizes variance, not necessarily class separability(0 or 1). Even when covering 95% variance, the AUC values were close but not better than selecting features individually for ROC AUC.