1.1 Feature Selection

The following features are used to train the model:

TODO: Add features

1.2 Mapping the Features

We followed a very standard approach for mapping the features. For sex, we mapped female to 0, male to 1. For chronic\_disease\_binary, we mapped false to 0, true to 1. We used to required mapping for outcome\_group.

TODO: Add writeup for Country if we end up using that feature

1.3 Balancing the Classes in the Training Dataset

We balanced the training data by randomly sampling from the minority class with replacement. The value N of this random sampling is the amount of non-minority class records divided by 2 (i.e., the mean of the amount of non-minority class records). This approach prevents the dataset from being skewed, allowing the data mining models that we use to optimize for a metric that is applicable and useful with respect to the dataset. By randomly sampling in order to make the minority class’s number of records comparable in magnitude to the other classes, we make it so a large amount of algorithms and optimization metrics are now applicable to the training dataset (since they assume a non-skewed class distribution for the records). We can thus experiment and compare the performance of these algorithms since we have more to choose from, and ultimately come up with a better model.

Amount of records per each class prior to balancing

0 997

1 13241

2 2974

Amount of records per each class post balancing

0 8107

1 13241

2 2974

TODO: Modify this if we end up sampling duplicates from the second minority class as well

1.4 Building models and hyperparameter tuning

We selected k = 5, as this allows each cross validation and training dataset to be large enough to be a sample of the entire dataset. We noted that the training datasets end up having 15565 rows, and the cross-validation sets have 1946 rows. Experimental data from data scientists over the years shows that k = 5 or 10 provides a good tradeoff between bias and variance, when it comes to tuning hyperparameters based on the results of cross-validation training (<https://machinelearningmastery.com/k-fold-cross-validation/>)

References:

<https://elitedatascience.com/imbalanced-classes>

<https://machinelearningmastery.com/k-fold-cross-validation/>

<https://scikit-learn.org/>