Machine Learning Part 1: Brainy Smurf's Model

Mauro Vella and Ignacio Bruzone

October 21, 2024

1 Preprocessing

Firstly, we kept the 1000 and 500 split of the data because it didn't significantly change the score of the model. Instead of using label encoding for all categorical variables, we used OneHotEncoder for the profession categorical variables. For the variables with ordinal values such as high, moderate, and low, we used label encoding to preserve the scale. Additionally, we kept the same standardization for the numerical variables. Lastly, we removed the image file column as it was not deemed necessary.

2 Feature Selection

Brainy's feature selection considers correlation between variables, which means the selected features may not directly influence heart failure. Using mutual information would help capture non-linear relationships between variables. We employed mutual information and, through cross-validation with a k-fold of 5 to have a good balance between the variance and bias, determined that the optimal number of selected features was 12 (see Figure 1). This number is higher than the original amount because the OneHotEncoder adds columns for different professions. After further testing, we decided to use a forward search wrapper with 6 features to improve the model's precision.

3 Model Selection

We used Stochastic Gradient Descent (SGD) to optimize the model and tested hyperparameters using GridSearchCV, also with a kfold of 5, to find the most optimal values for the model's performance. The best results were obtained with a learning rate of 0.01 and 1000 iterations with RMSE.

4 Result Analysis

Using the test data, we obtained an RMSE score of 0.07984, representing a slight improvement over Brainy's model. Figure 2 shows that our model performs well at predicting smaller values but struggles with larger ones. This may be due to the higher concentration of data between 0.0 and 0.2 compared to higher values. We conclude that a linear model may not be ideal for predicting heart failure risk, as it is heavily influenced by the lower values and under performs at the higher end.

5 Graphs

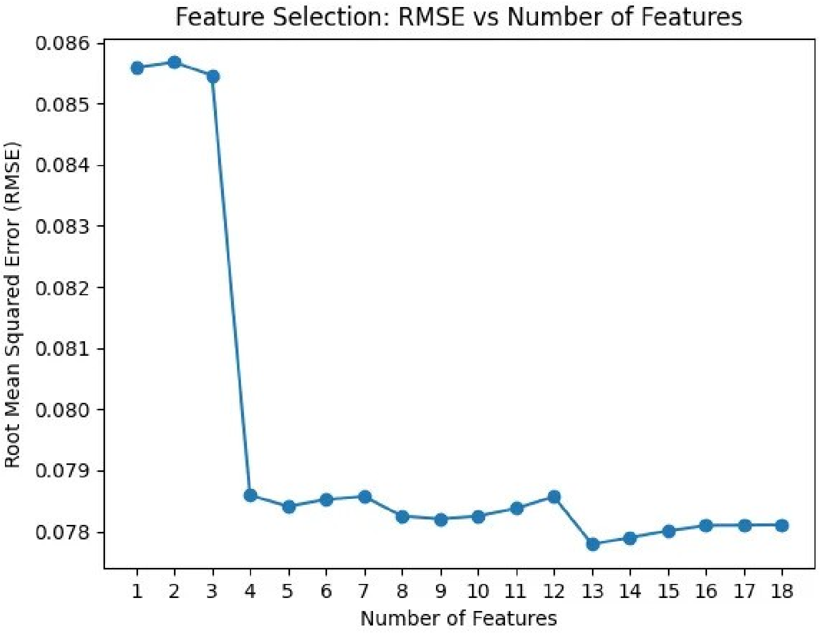

Figure 1: Number of features and their cross-validation performance.

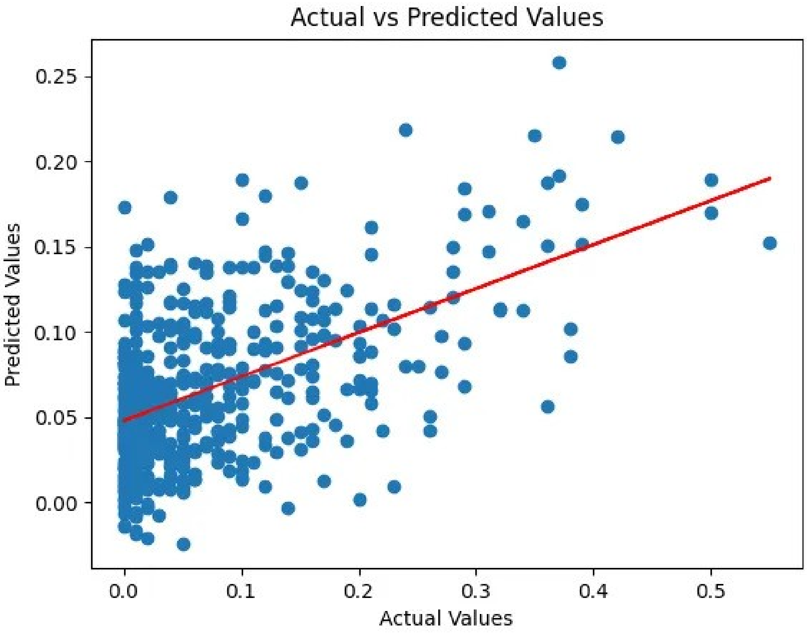

Figure 2: Actual versus predicted risk. The red line represents perfect predictions.