Machine Learning Part 1: Brainy Smurf's Model

Mauro Vella and Ignacio Bruzone

October 24, 2024

1 Preprocessing

Firstly, we kept the 1000 and 500 split of the data because it's a good split as the data set is limited and if we reduce the test data the results would be more incorrect. Instead of using label encoding for all categorical variables, we used OneHotEncoder for the profession categorical variables. Also, For the variables with ordinal values such as high, moderate, and low, we used a mapping from 1-5 for the categories to preserve the scale. Additionally, we also scaled all the data after preprocessing because for lineal model its essential for data to be scaled. 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 8 (see Figure 1). This number is higher than the original amount because the OneHotEncoder adds columns for different professions.

3 Model Selection

We have tried some models with Gradient decent and Stochastic Gradient Descent but as the data set is not that big, we agreed with brainy and used the LinearRegressor class for sklearn.

4 Result Analysis

Using the test data, we obtained an RMSE score of 07788299804065363, 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

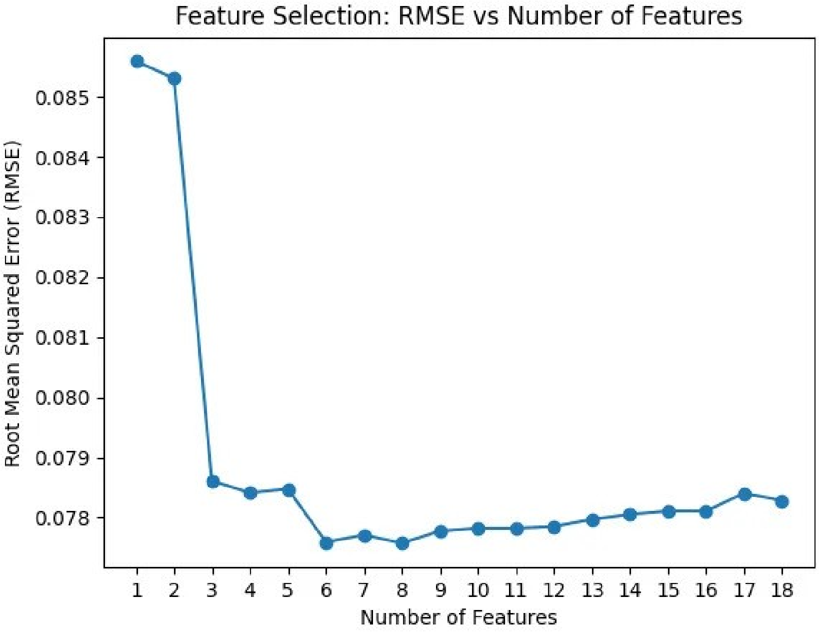

Figure 1: Number of features and their cross-validation performance.

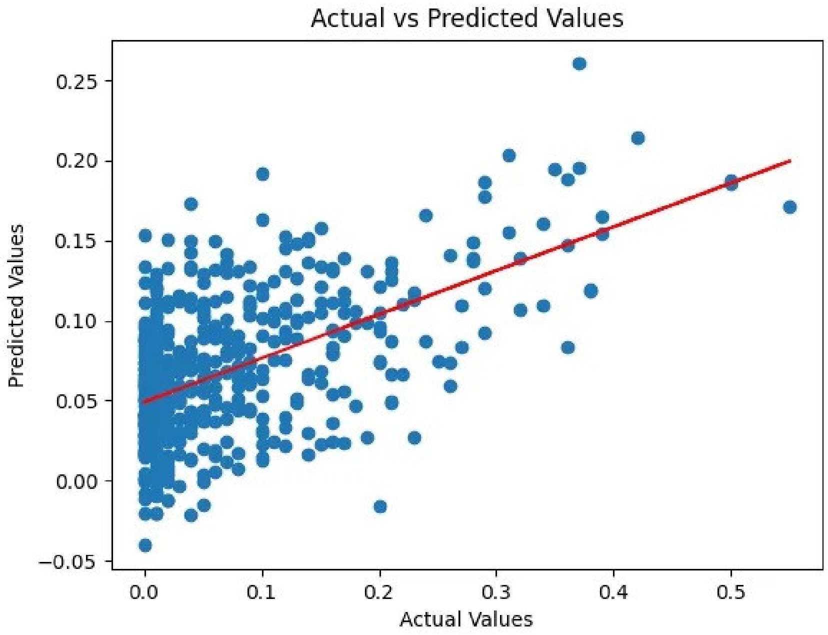

Figure 2: Actual versus predicted risk. The red line represents perfect predictions.