HEART FAILURE ON THE RISE IN THE SMURF SOCIETY (PART 1)

We obtained linear model baseline using 4 steps: categorical data transformation, feature scaling, feature selection and model selection.

There are four categorical features in the smurf dataset. *Consumption of sarsaparilla*, *smurfberry liquor* and *smurfin donuts*, were mapped from 0 to 4 (0 being very low and 4 being very high), since the categories have a natural sense of order

Jobs are not ordered. Therefore, we one-hot encode them by creating six distinct features (one for each job type) with value 1 if the smurf has that job and 0 if not.

Since different features have different data ranges, we scaled the data to avoid certain features having more weight during training. We decided to standardize (using Scikit-Learn's StandardScaler) because it is more numerically stable with polynomial regression. Creating polynomial Features, it make interaction terms and powers of features, which can have wildly different magnitudes if scaled differently. We also tried normalizing, but it yielded worse results.

Model and feature selection were performed together, rather than sequentially, since the best set of features for one model might not be the same for another.

For model selection, we assume *linear models* also include polynomial regression. If not, then no model selection is needed since linear regression has no hyperparameters. Polynomial regression's sole hyperparameter is the degree.

In practice, we use Scikit-Learn's PolynomialFeatures to create new features that capture all relationships between features ($(X_1)^3, (X_1)^2X_2, X_1X_2X_3$...)

We performed feature selection using greedy search (Scikit-Learn's SequentialFeatureSelector). We initially wanted to use backwards search with an uncapped number of possible features. However, with polynomial regression, this was too computationally expensive. We then capped the number of features to select to 15 and used forward greedy search. Instead of using a dedicated validation set for model and feature selection, we perform k-fold cross validation, with k set to 5 (arbitrarily).

We find that for plain linear regression, *calcium*, *smurfberry liquor*, *smurfin donuts*, *Vitamin D*, and all jobs except *administration and governance* have negligible impact on RMSE.

With each model trained using its best (according forward greedy search) 15 features (at most), we observe significantly better results in 2nd degree regression ($RMSE_2 = 0.0419$) over 1st degree regression ($RMSE_1 = 0.0556$). 3rd and 4th degree regression offer worse results ($RMSE_3 = 0.0426, RMSE_4 = 0.0440$), likely due to overfitting.

Therefore, our final baseline model is a 2nd degree linear regression with the following features: *age*, *blood pressure*, *cholesterol*, *hemoglobin*, *height*, *potassium*, *sarsaparilla*, *weight*, *blood pressure*^{2}, *cholesterol*height*, *cholesterol*weight*, *hemoglobin*^{2}, *height*weight*, *potassium*^{2}, and *weight*^{2}

One quirk is that this model sometimes yields negative predictions, which does not make sense for a probability of heart failure. This could be fixed by using a different model or bounding the output, but we will leave it as-is for this baseline