Glioblastoma regressor

The features were classified according to whether they were naturally binary, ordinals or cardinals and further whether they were dates or not.

A first stab at feature importance was made using a ttest\_ind for binary features and pearsonr for cardinal features. I also checked for the number of missing samples.

I then selected the most important features pvalue (cut-off between 0.01 and 0.05) with the fewest missing samples (between 100 and 200 max) using a grid search to test for different parameters. Missing features were treated either by setting them to the median or by dropping the relevant samples.

The samples were repeatedly shuffled and split 80%/10%/10% to provide training, testing and validation data.

A random forest regressor was then fitted to the data and the R2 score for the predictions found. The R2 score is defined as 1-u/v where v is sum of (y\_real - y\_mean)\*\*2 and u is the sum of (y\_pred - y\_real)\*\*2. Using this definition, a perfect score is 1, predicting the average each time gives a score of 0 and arbitrary negative scores are possible.

Intention

Section to put in characteristics of disease (eg presenting headache etc.) and then should be able to provide you with prognoses for different next steps eg adjuvant vs not.