**Agenda:**

* Linear and logistic regression to justify use of machine learning
* Do pairwise correlation analysis of features
  + Potentially drop those that are unnecessary (i.e. very high magnitude correlation values)
* Test models on test set with setting excluded
* Look at intersection of test features
* Look at ensemble model predictions for each category of models, and then models combined
* Mape -> mean relative error

How to show linearity in MMR?

I’ve done some reading of the literature to find sources that help me justify choice of ML models

* Does this justification go into the ‘background research’, ‘literature review’, or ‘methods’ sections?

To produce ensemble-based models (both voting & stacking, for both train/test splits):

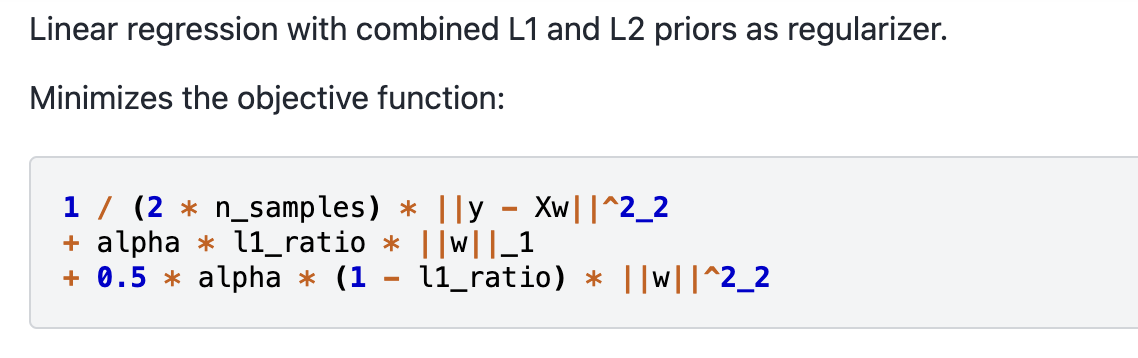
* There are 20 base estimators
  + 10 base estimators are LightGBM Regressors
  + 10 base estimators are Random Forest Regressors
  + 5 each base estimators were trained on the full train dataset (1 for each of the train/test folds)
  + 5 each base estimators were trained on the train dataset with a 95% missing data threshold (1 for each of the train/test folds)
* Xgboost base estimators and all estimators trained on the dataset with the 85% missing data threshold were not included in the ensemble model due to their consistent relatively poor performance.
* Train all base estimators using their optimal hyperparameters (from previous fine-tuning experiments) and their respective train set from their respective train/validation cross-validation fold
* Use each base estimator to predict using whole 90% training data (with columns removed that it was not trained on)
* Then train the model with the best set of hyperparameters on the full training dataset

Voting:

* Using 300 Optuna trials to experiment with relative weightings of different base estimators’ predictions in calculating the weighted average prediction score

Stacked

* Past study finding random forest and xgboost as meta-estimator in stacking outperformed base estimators
  + They found XGBoost performed the best
  + However, given XGBoost’s poor performance in this study, and that Random Forest is one of the base estimators, potentially being too similar (?), we instead represented the decision tree based meta-estimator approach using AdaBoost
* Linear regression (elastic net - > other papers used this)



A screenshot of a computer

AI-generated content may be incorrect.

* Linreg
* MLP
  + Not, because not enough data?
* SVM

Mlp overtrained, seen by much lower validation error than testing error

* Even though best hyperparameters were at the ends of their range, they were not increased to try to mitigate overfitting, as already there were 7 hidden layers with high numbers of nodes per layer

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

I have been playing around with the features, and I have a few updates:

* There are 584 pairs of features whose absolute pairwise correlation is greater than 0.95 and 923 pairs of features whose absolute correlations are greater than 0.90. I have included a histogram showing the distribution of correlations in this message.
  + Would you recommend excluding the variable in the pair with more missing data?

Feature Correlation Pairs:

* There are 584 pairs of features whose absolute pairwise correlation is greater than 0.95 and 923 pairs of features whose absolute correlations are greater than 0.90.

A graph of a graph

AI-generated content may be incorrect.

* I took the example of feature pairings with absolute correlation coefficient greater than 0.9 and filtered by pairings whose absolute difference in their proportion of missing data was greater than 0.05 (to make it worth replacing the member of the pair with less data with the one with more data), there were only 129 of the original 923 pairs remaining.
* When I repeated this with all pairs whose absolute correlation coefficient was greater than .80 (n=2944) and filtered for pairs whose missing data proportions differed by more than 20%, the code returned 785 pairs.
  + The following table shows the number of times different indicators appear in a pairing.

|  | **count** |
| --- | --- |
| Mortality rate, infant, male (per 1,000 live births) | 57 |
| Mortality rate, infant (per 1,000 live births) | 57 |
| Mortality rate, under-5, male (per 1,000 live births) | 52 |
| Mortality rate, infant, female (per 1,000 live births) | 52 |
| Mortality rate, under-5 (per 1,000 live births) | 49 |
| ... | ... |
| Women and girls who participate in activities during menstrual period (% of women and girls ages 15-49 who had a menstrual period within the last year) | 1 |
| Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant) | 1 |
| Prevalence of underweight, weight for age, female (% of children under 5) | 1 |
| 1.D.14 Intestinal nematode infections prevalence (age standardized) (per 100 000 population) female | 1 |
| Unemployment with intermediate education (% of total labor force with intermediate education) male | 1 |