Exploring the Lung Transplant System with the MICE imputation methods and ML modelling

Kaya Panchalingam SEAS 2023
GRASP Labs School of Engineering and Applied Science

Motivation

- Lung transplants have one of the highest 1-year mortality rates among all major organ transplants in the country
- Using data about lung recipients and donors, we are hoping to improve the matching process
- This is done by using the lung allocation score (LAS) which is used to prioritize waiting list candidates based on a combination of waitlist urgency and post-transplant survival.
- The available data disproportionately contains certain race-gender groups of the population so we are hoping to leverage Machine Learning concepts to more accurately model the LAS scores of minority groups of the population

Findings

ORIGINAL DATA	MICE DATA
Fold 1	**Fold 1**
training loss 5.552	training loss 1.449
training loss 1.239	training loss 0.839
training loss 0.866	training loss 0.642
training loss 0.797	training loss 0.426
training loss 0.777	training loss 0.378
Loss for fold 1: 0.725	Loss for fold 1: 0.297
Fold 2	**Fold 2**
training loss 5.521	training loss 2.764
training loss 1.124	training loss 0.897
training loss 0.855	training loss 0.806
training loss 0.763	training loss 0.623
training loss 0.745	training loss 0.430
Loss for fold 2: 0.708	Loss for fold 2: 0.370
Fold 3	**Fold 3**
training loss 3.943	training loss 1.644
training loss 1.050	training loss 0.822
training loss 0.817	training loss 0.621
training loss 0.791	training loss 0.444
training loss 0.767	training loss 0.408
Loss for fold 3: 0.770	Loss for fold 3: 0.341
Fold 4	**Fold 4**
training loss 4.548	training loss 1.281
training loss 1.314	training loss 0.819
training loss 0.931	training loss 0.618
training loss 0.782	training loss 0.371
training loss 0.758	training loss 0.400
Loss for fold 4: 0.708	Loss for fold 4: 0.344
Fold 5	**Fold 5**
training loss 2.773	training loss 1.466
training loss 0.945	training loss 0.834
training loss 0.766	training loss 0.564
training loss 0.735	training loss 0.418
training loss 0.701	training loss 0.372
Loss for fold 5: 0.888	Loss for fold 5: 0.332

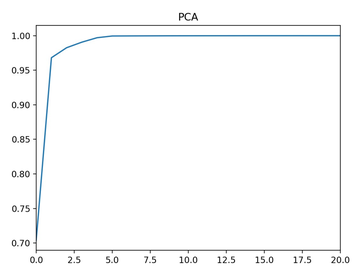
The results on the left are the loss for each fold for the complete data from the original dataset without imputation and on the left are the results of the model on the imputed data.

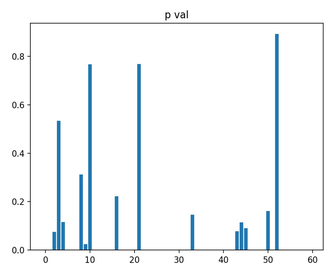
Calculating the Lung Allocation Score

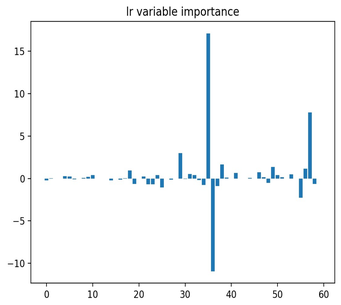
The steps include:

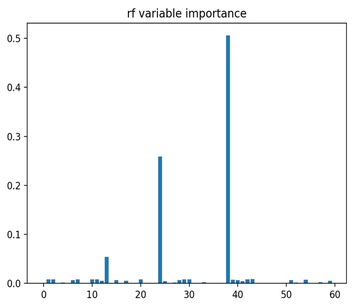
- Calculate the waiting list survival probability during the next year
- Calculate the waitlist urgency measure
- Calculate the post-transplant survival probability during the first post-transplant year
- Calculate the post-transplant survival measure
- Calculate the raw allocation score
- Normalize the raw allocation score to obtain the LAS.

Results









Methodology

- •Exploration of data using Variable Importance. This measured the statistical significance of each variable in the data with respect to its effects on a model.
- •Ran MICE (multivariate Imputation by Chained Equations) on the incomplete data set to impute values that were missing. This method will first set placeholder values for missing values in the columns (using the mean value of the column for instance). Then it will remove all the placeholder values for one column but keep them for all other columns and regress the missing values for the certain feature. We repeat this process for each feature and then repeat this whole process a few times until the values stabilize.
- •Ran PCA on the imputed data set to capture most of the variance of the dataset with fewer dimensions.
- •Ran a fully connected neural network (FFN) with 5-fold cross-validation to regress the LAS. We have 60 features and used to fully connected layers with 120 perceptrons and an output of 1 value which is the regressed LAS score. We used ReLU as the activation function.

Conclusion and Next Steps

- •The model managed to predict the score with a high degree of accuracy
- •Next, we are looking to implement few-shot learning methods for subgroups of the data based on the age group to see if the model works better. This would be exploring if retraining the model on a subgroup of the data, would lead to higher accuracy in predicting the LAS for that subgroup

References

1. https://unos.org/wp-content/uploads/unos/Lung_Calculation.pdf
2. http://d2l.ai/, chapters 4-7
3. https://arxiv.org/abs/2103.12857, Embracing Disharmony in Medical Imaging: A simple and Effective Framework for Domain Adaption
4. https://arxiv.org/pdf/1909.02729.pdf, A Baseline for Few-Shot Image Classification