Imputation of SEER Data using Artificial Neural Networks

A comparison of technique performance

Matthew Prest

COLUMBIA | COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

CISNET

Outline

Background

- Imputation
- Neural networks
- SEER selection criteria

Performance Analysis

- Overfitting indicators
- Cross validation scores
- Train/Test distribution comparison

Next Steps/Discussion & Questions

- NOS reclassification
- More comparative techniques
- Alternative architectures

Background

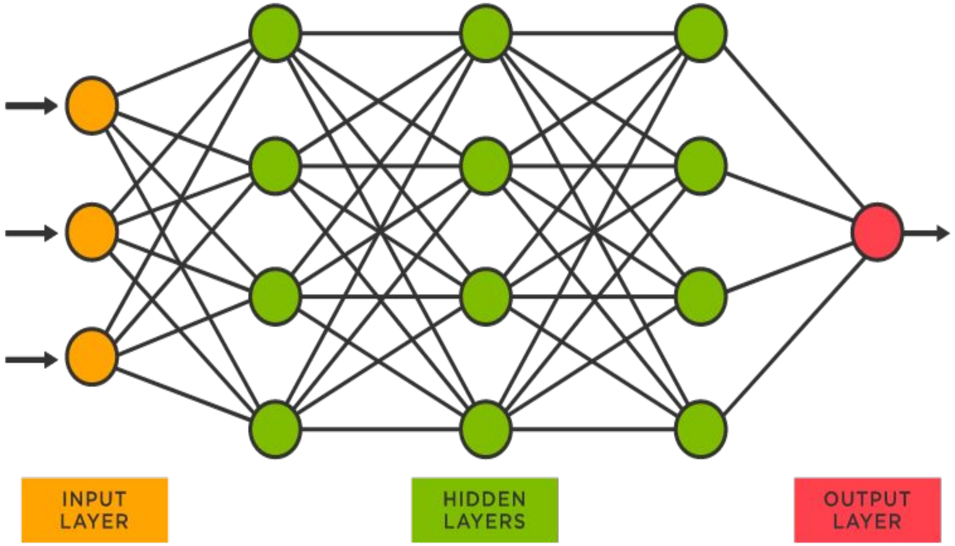
Imputation:

- The process of replacing missing data using various techniques
- Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)
- Assumptions, bias compounding, and effects on statistical power
- Techniques, nested imputation, and multiple imputation

Background

Artificial Neural Networks (ANNs) and Multilayer Perceptrons (MLPs):

- Input data (features) with normalization and embedding.
- Artificial Neurons containing non-linear activation function with weighted input and output edges.
- Output layer maps to categorical classification
- Loss function to train against
- Gradient descent optimization with learning rate factor
- Dropout probability to guard against overfitting



Parameters of MLP used in this project:

- 43 Predictors/features, 15 numerical, 28 categorical embedded with dimensions = ceiling(n/2)
- ReLU activation function, dropout p = 0.4, hidden layer dimensions [200, 100, 50], output layer size = 4, learning rate = 0.001

Background

SEER Stage Variables:

- “The Staging Over Time Project.”
- AJCC 3rd (1988-2004), AJCC 6th (2004-2015), AJCC 7th (2016-2017), AJCC 8th (2018+)
- Target is time merged AJCC with 6th > 7th > 8th hierarchy excluding substages, created by NCI

Selection Criteria (N = 1, 947, 359):

- SEER 18 Incidence (Not delay adjusted)
- 2004-2018
- Non-Cardia Gastric cases
  - Intestinal histologies: 8140, 8143-8144, 8210-8211, 8221, 8260-8263
  - Diffuse histologies: 8141-8142, 8145, 8490
- Known age, race, diagnostic confirmation
-

Background

Predictor Variable Groupings and Stage Missingness:

- Demographic (Age, Sex, Race/Origin, Year, n = 4)
- AJCC Stage (AJCC Editions, n = 4)
- Histology (Intestinal vs Diffuse, n = 1)
- Summary Stages (L/R/D, n = 8)
- Extent of Disease (T/N/M, n = 19)
- CS Variables (Tumor size, extension etc, n = 7)

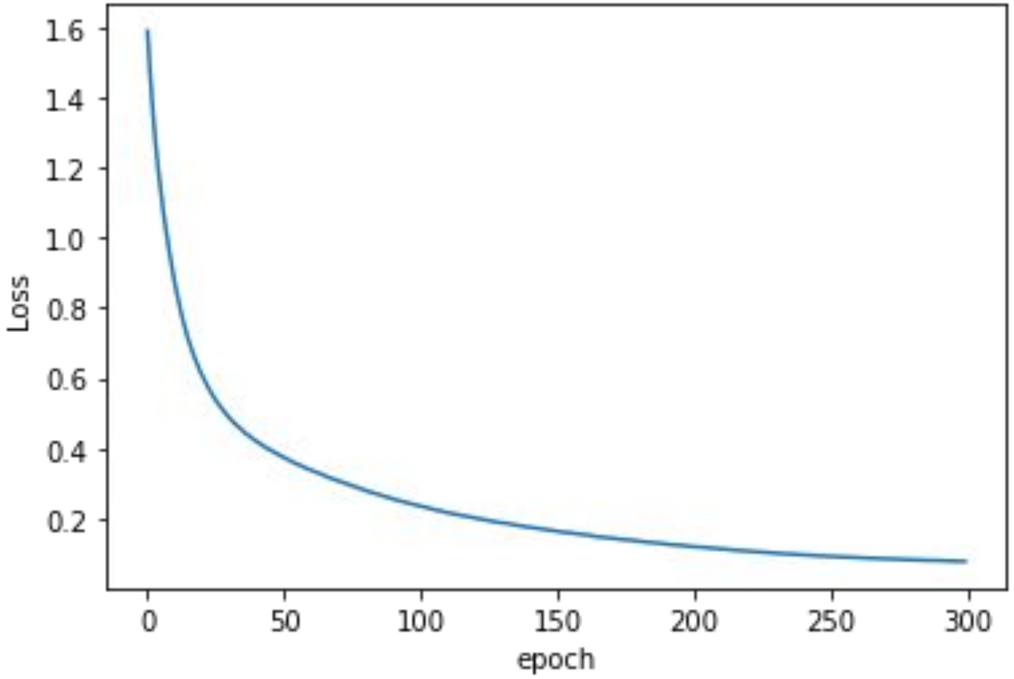
Group	Staged Count	Unstaged Count	% Missing
Total	1771115	176244	9.05
Race			
NH White	1234824	117585	8.69
AAPI	122566	13384	9.84
AI/AN	8997	1000	10.00
Hispanic	176240	23396	11.72
NH Black	228488	20879	8.37
Sex			
Female	525981	65870	11.13
Male	1245134	110374	8.14
Histology			
Intestinal	1743261	171850	8.97
Diffuse	27854	4394	13.63

Year	Staged Count	Unstaged Count	% Missing
2004	109140	11724	9.70
2005	108701	10754	9.00
2006	115619	10697	8.47
2007	122315	11462	8.57
2008	121148	11096	8.39
2009	122818	10976	8.20
2010	122087	10032	7.59
2011	124073	10655	7.91
2012	117458	10152	7.96
2013	117226	9921	7.80
2014	116121	9871	7.83
2015	119458	9789	7.57
2016	118719	14183	10.67
2017	120399	15164	11.19
2018	115833	19768	14.58

Performance Analysis

Cross Entropy Loss:

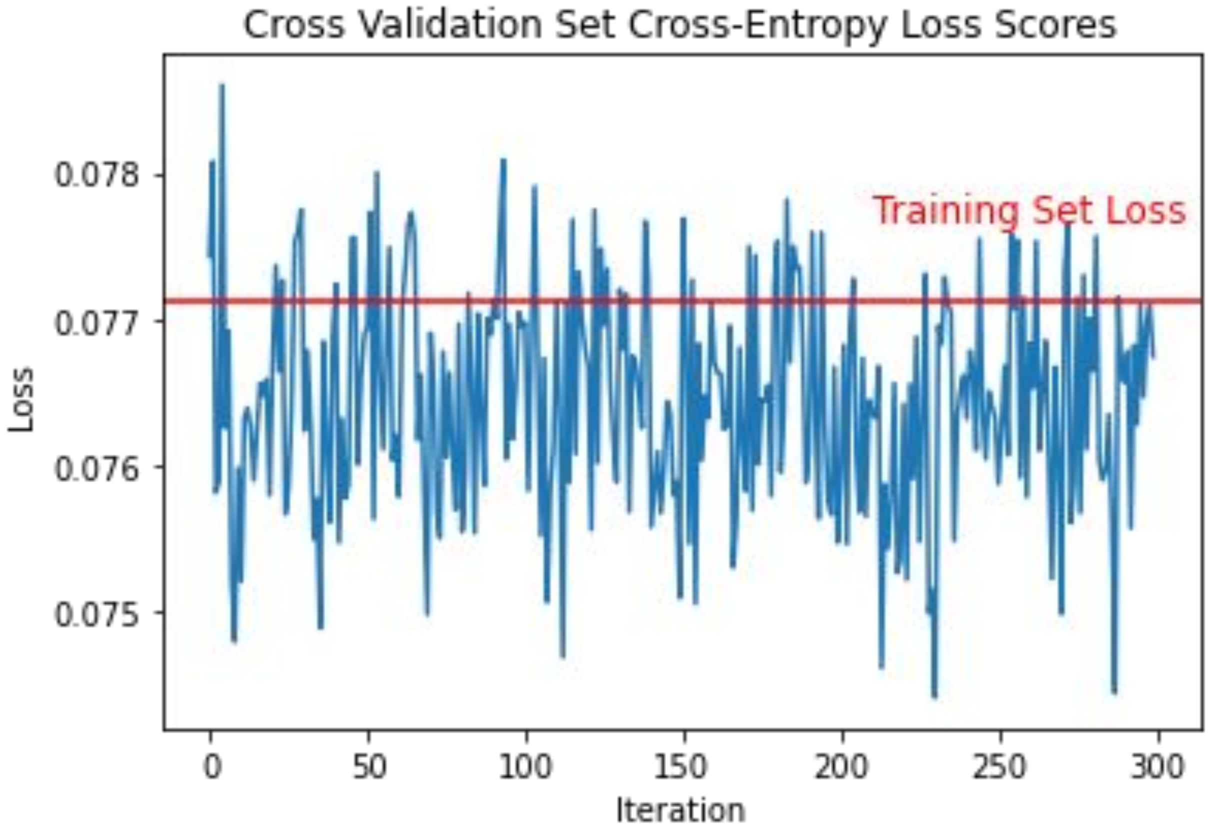
- Cost function, measure of performance
- Training set Cross Entropy = 0.07712 bits after 300 epochs
- Noisy training curve tail indicates overfitting



Performance Analysis

Cross Validation Performance

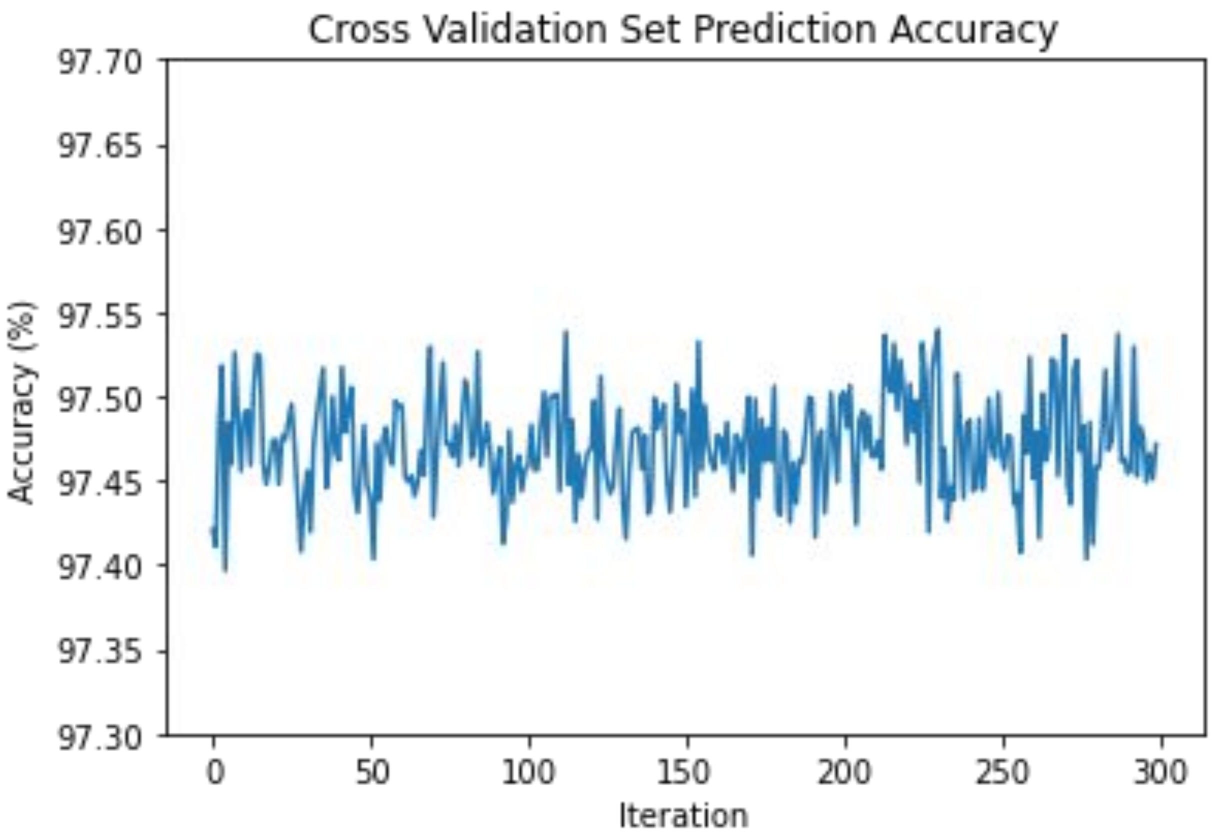
- Higher Cross Entropy loss indicates overfitting
- Every iteration with the same cross validation set (20% of Train Set)



Performance Analysis

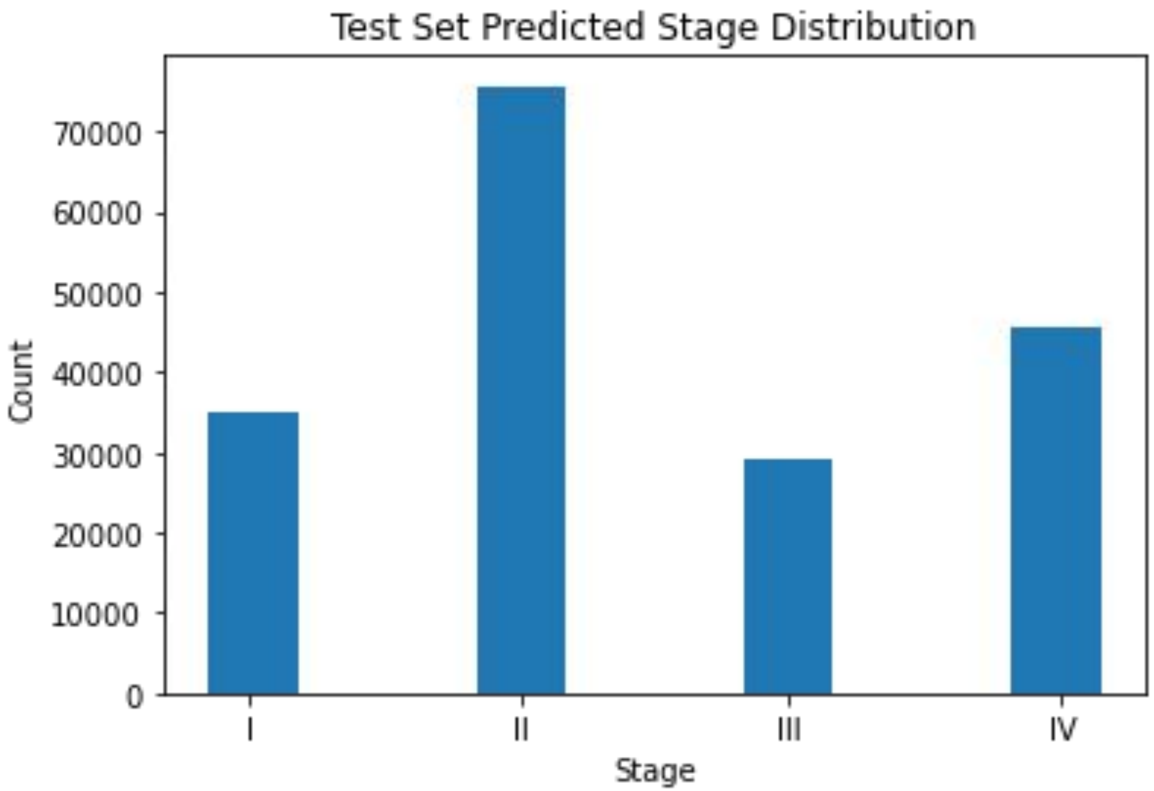
Cross Validation Performance

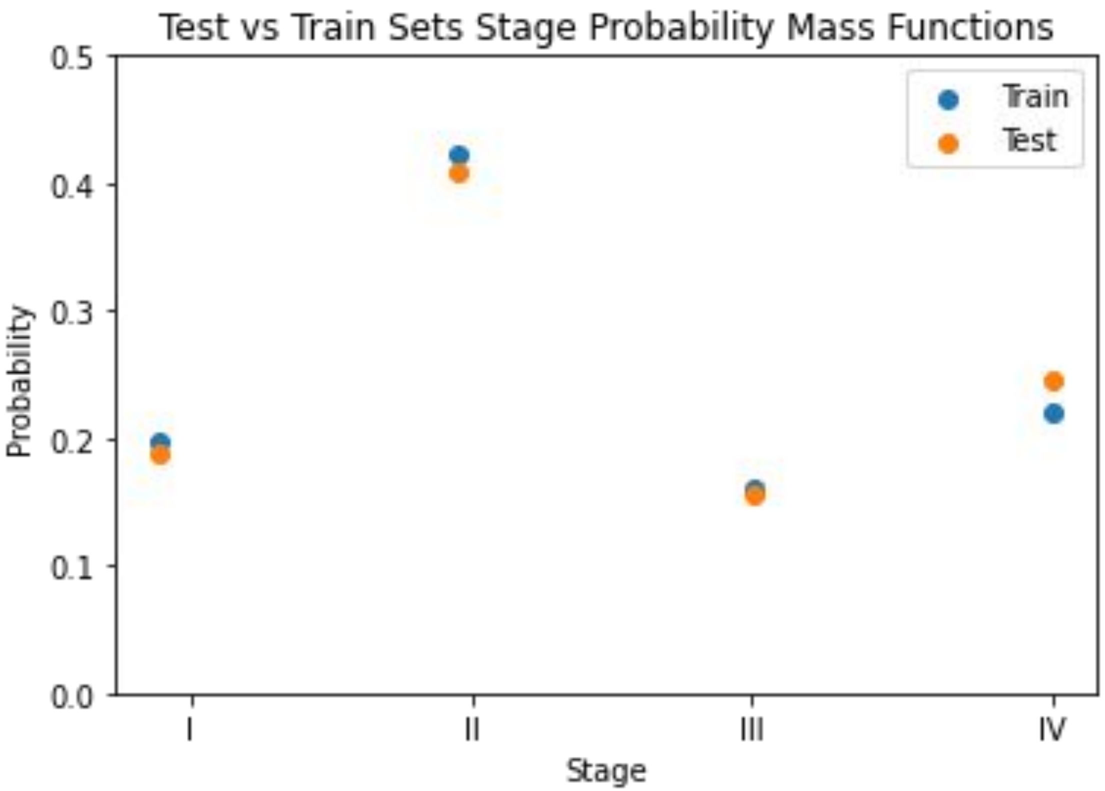
- Accuracy records the percentage of correctly labelled observations in the Cross Validation set



Performance Analysis

Test Prediction Distribution





Next Steps

- New Model Architectures
  - Properly encode the time validity of certain predictors
  - Finetune hyperparameters
- Sensitivity Analysis
  - Dimension reduction, pruning predictors
  - Multifold CV scores
- Further Applications
  - Other sites
  - NOS histology reclassification

Closing Discussion

Is this too much of a “black box” technique?

Is this too dangerous given that we are forcing changes to our evidence?

How much uncertainty will this introduce to modelling based on the imputed data?

Additional Questions from the Audience

Citations

1. Mandi Yu, Eric J. Feuer, Kathleen A. Cronin, and Neil E. Caporaso, Use of Multiple Imputation to Correct for Bias in Lung Cancer Incidence Trends by Histologic Subtype, Cancer Epidemiology, Biomarkers & Prevention 2014.
2. Chung-Yuan Cheng, Wan-Ling Tseng, Ching-Fen Chang, Chuan-Hsiung Chang and Susan Shur-Fen Gau, A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder, Frontiers in Psychiatry 2020.
3. Marek Smieja, Łukasz Struski, Jacek Tabor, Bartosz Zielinski. Przemysław Spurek , Processing of missing data by neural networks, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada.
4. Akshaj Verma, PyTorch[Tabular]-Multiclass Classification, How to Train Your Neural Net, Towards Data Science, Mar 18, 2020
5. Schafer, Joseph L. "Multiple imputation: a primer." Statistical methods in medical research 8.1 (1999): 3-15.
6. Royston, Patrick. "Multiple imputation of missing values." The Stata Journal 4.3 (2004): 227-241.
7. Nordbotten, Svein. "Neural network imputation applied to the Norwegian 1990 population census data." JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM- 12 (1996): 385-402.
8. Nishanth, Kancherla Jonah, and Vadlamani Ravi. "Probabilistic neural network based categorical data imputation." Neurocomputing 218 (2016): 17-25.