Imputing Missing Values for Race and Gender in Emergency Room Records

Bradley Katcher and Elizabeth Driskill

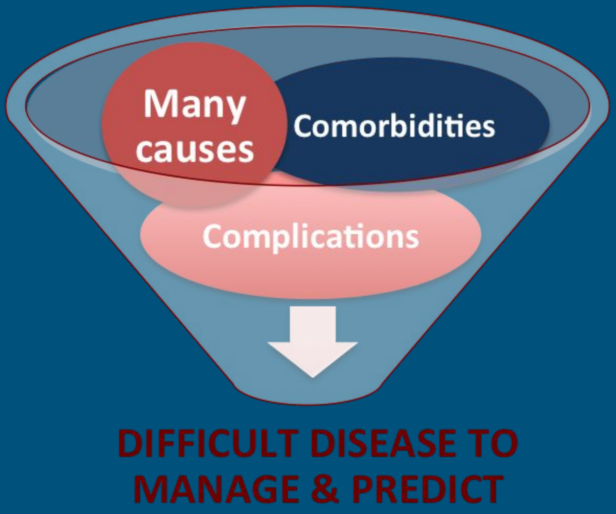
Problem

- Missing values for race and gender within our dataset
- Difficult to analyze demographic characteristics with missing data

Goal

- Fill in missing values so that future researchers have access to additional data
- Hospitals can gain a greater understanding of certain patient populations and be better equipped to provide care

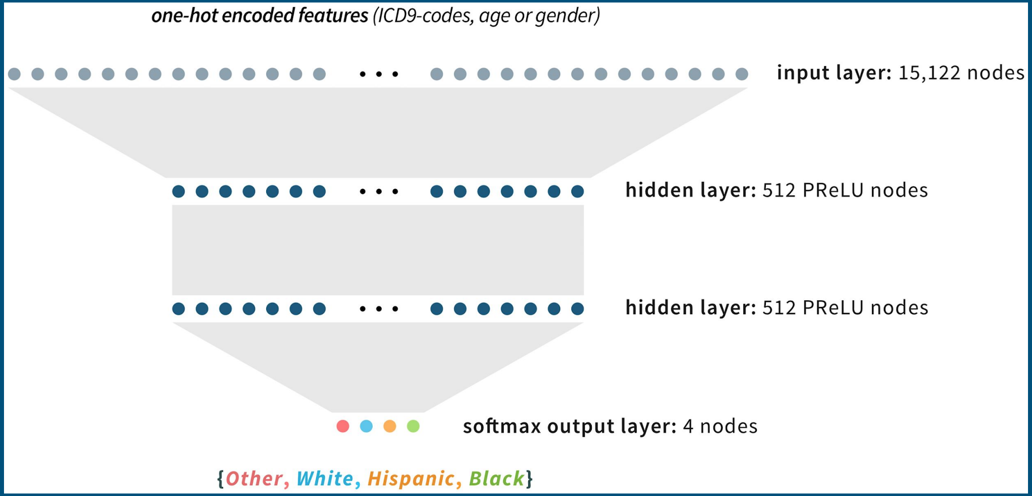
Background



- More than 30.3 Americans have diabetes, and another 84.1 million are prediabetic (National Center for Chronic Disease Prevention and Health Promotion, 2017)
- Nearly 1 in every 10 Virginians has diabetes
- Certain groups of people are disproportionately affected by diabetes (Virginia Department of Health, 2019)
  - Some groups are also more likely to present to the emergency room for diabetes care
    - NOT IDEAL
- Can we identify which patient populations are most at risk of presenting to the emergency room with diabetes-related symptoms?
  - Need more data - project motivation

What has been done in the past?

- RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning
  - Logistic regression used to be primary way to impute race/ethnicity
  - RIDDLE uses a relatively simple multilayer perceptron (MLP), a type of neural network architecture that is a directed acyclic graph to impute race and ethnicity



Kim et. al. (2018)

Data

- Emergency room records from seven hospitals within Central Virginia
- Demographic characteristics:
  - Age
  - Race
  - Gender
  - Ethnicity
- Climate variables:
  - Apparent temperature
  - Precipitation
  - Wind speed
- Other: visit date, patient zip code, health insurance provider, principal diagnosis

Determine Missingness

- What was missing?
- None includes unknown and patient refused

UVA

W	268927	65.81%
B	110033	26.93%
O	17969	4.40%
H	4468	1.09%
A	3739	0.91%
I	178	0.04%
None	3323	0.81%

Carilion

White	1141838	82.26%
Black	184013	13.26%
Hispanic	29833	2.15%
Biracial	15593	1.12%
Asian	7170	0.52%
Other	5275	0.38%
Am Indian	1043	0.08%
Pac Islander	460	0.03%
None	2791	0.20%

Females	764982	55.11%
Males	620725	44.72%
None	2309	0.17%

Data Preprocessing

- Reduce data to remove redundant variables
  - Weather at one time of day
  - One variable for chief patient complaint
- Remove non-pertinent columns with significant missingness
- Characterize missing values in variables other than race and gender
  - Categorical - replace with ‘None’
  - Numeric - replace with the average
- Massage data-types (important issue in administrative data)
- Create indicators for missing variables
- Scale numeric values (between 0 and 1)
- One-hot encode categorical variables

What’s Left

- Facility
- Gender
- Age
- Ethnicity
- Station
- Financial Class
- Diagnosis **(MANY)**
- Zip Code **(MANY)**
- Weather Variables
- Race

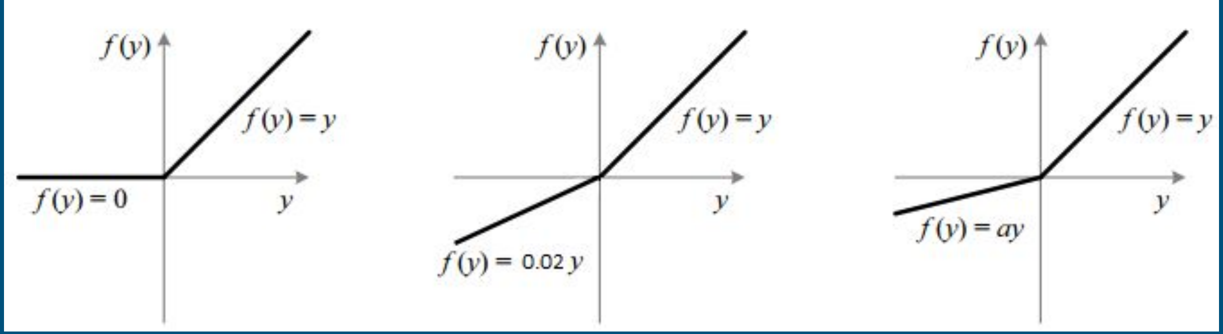
Dealing with dimensionality:

- 3 digit zip codes
- Prefixes of ICD-9 and ICD-10 codes (look at a higher classification of an illness)

Methods

- Train-validation split (80/20)
- Build neural networks using keras
  - 2-3 hidden layers using ‘relu’ activation function
  - 2-3 hidden layers using ‘prelu’ activation function
  - 5 hidden layers using ‘relu’ activation function
  - 5 hidden layers using ‘prelu’ activation function
- Hidden layers included 64 or 128 nodes
- Dropout layers incorporated to adjust for overfitting
- Four neural networks built to predict missing values for:
  - Carilion race - 8 output nodes, ‘softmax’ activation function
  - Carilion gender - 1 output node, ‘sigmoid’ activation function
  - UVA race - 6 output nodes, ‘softmax’ activation function
- Plot training and validation loss and accuracy curves
- Adam optimizer

What is PReLU?



ReLU Leaky ReLU PReLU

$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases}$$

- PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk (He et. al. 2015)
- Increase learning speed by not deactivating certain neurons

Results: Carilion Race

Type of Network	Final Training Accuracy	Final Training Loss	Final Validation Accuracy	Final Validation Loss
64 relu, 128 relu	0.8447	0.4592	0.8443	0.4632
64 prelu, 64 prelu, 64 prelu	0.8451	0.4537	0.8441	0.4599
64 relu, 64 relu, 64 relu, 64 relu, 64 relu	0.8444	0.4621	0.8443	0.4648
64 prelu, 64 prelu, 64 prelu, 64 prelu, 64 prelu	0.8412	1.246	0.8368	0.5390

3 Layers of PReLU with 64 nodes

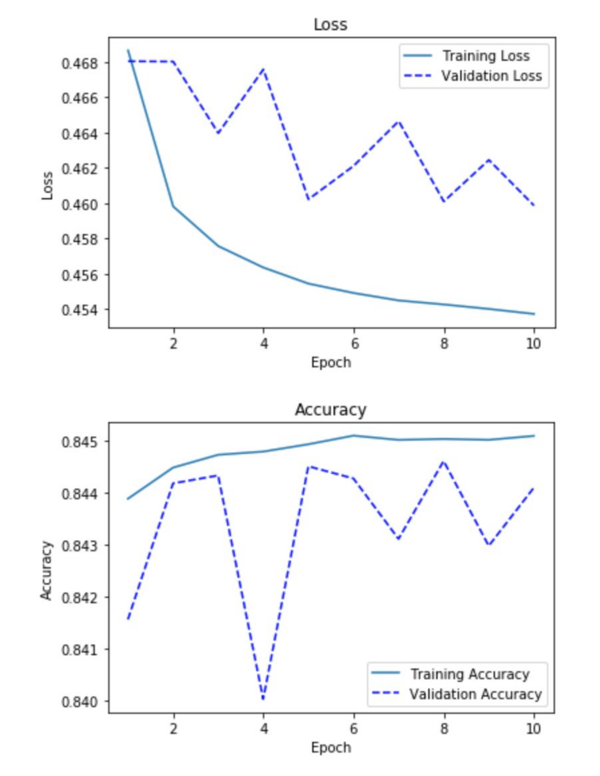
Predictions on Validation Set:

White	Black	Hispanic	Biracial	Asian
262,452	8,225	6,351	8	9

Validation set did not predict all races (excluded “Other”, “Pacific Islander,” and “American Indian”)

Predictions on Unknown set:

White	Black	Hispanic
2,741	3	47



Results: Carilion Gender

Type of Network	Final Training Accuracy	Final Training Loss	Final Validation Accuracy	Final Validation Loss
64 relu, 0.3 dropout, 64 relu, 0.3 dropout, 64 relu	0.6220	0.6288	0.6196	0.6311
64 prelu, 0.2 dropout, 64 prelu, 0.2 dropout, 128 prelu	0.6011	0.6448	0.6100	0.6392
64 relu, 64 relu, 64 relu, 0.3 dropout, 64 relu, 0.3 dropout, 64 relu	0.6055	0.6437	0.6024	0.6459
64 prelu, 64 prelu, 64 prelu, 64 prelu, 128 prelu	0.6305	0.6219	0.6214	0.6306

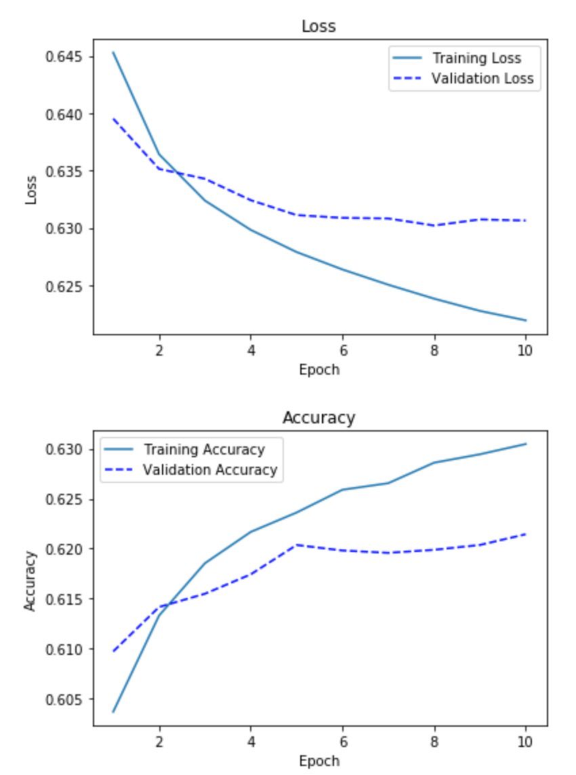
4 layers of PReLU with 64 nodes, 1 layer of PReLU with 128 nodes

Predictions on Validation Set:

Male	Female
179,306	97,836

Predictions on Unknown set:

Male	Female
2309	0



Results: UVA Race

Type of Network	Final Training Accuracy	Final Training Loss	Final Validation Accuracy	Final Validation Loss
64 relu, 64 relu, 64 relu	0.6936	0.6947	0.6813	0.7346
64 prelu, 64 prelu, 64 prelu	0.7020	0.6741	0.6836	0.7358
64 relu, 64 relu, 64 relu, 64 relu, 64 relu	0.6929	0.6982	0.6812	0.7412
64 prelu, 64 prelu, 64 prelu, 64 prelu, 64 prelu	0.7015	0.6744	0.6855	0.7323

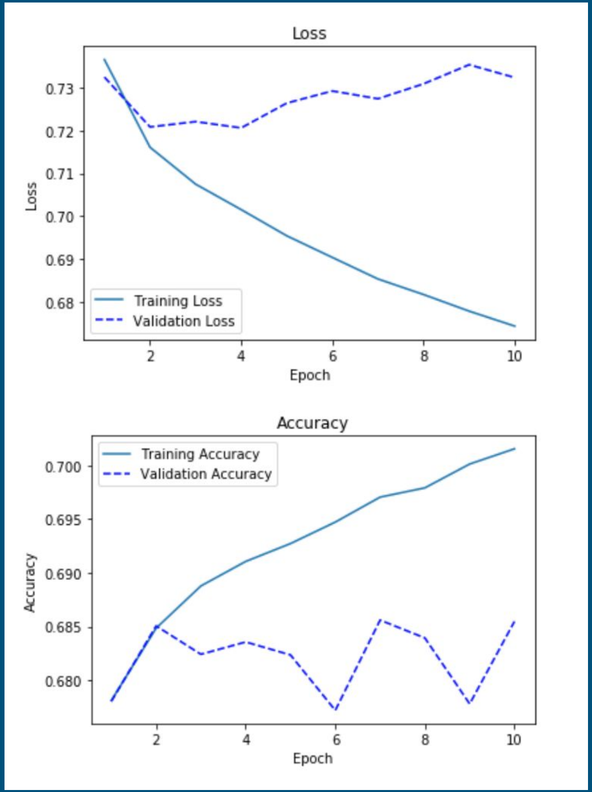
5 layers of PReLU with 64 nodes

Predictions on Validation Set:

W	B	O	H	A	I
69,461	9,089	2,042	455	15	1

Predictions on Unknown set:

White	Other
3,322	1



Conclusions

- Deep neural networks were successful in predicting missing race values
- Do these models need to train longer?
- Not successful in predicting missing gender values
  - Explore alternative methods to hopefully produce more accurate results
- Future work:
  - These methods could be applied to predict missing values for other categorical variables
  - Experiment with:
    - Different number of hidden layers
    - Different number of nodes in each hidden layer
    - Different activation function
    - Add or remove dropout layers
- Future research involving different patient populations and social determinants of health can hopefully benefit from additional demographic information

References

He, Kaiming, et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” *ArXiv:1502.01852 [Cs]*, Feb. 2015. *arXiv.org*, http://arxiv.org/abs/1502.01852.

Kim JS, Gao X, Rzhetsky A (2018) RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. PLOS Computational Biology 14(4): e1006106. https://doi.org/10.1371/journal.pcbi.1006106

National Center for Chronic Disease Prevention and Health Promotion. (2017). *National Diabetes Statistics Report, 2017* (Estimates of Diabetes and Its Burden in the United States, p. 20). Center for Disease Control and Prevention: Division of Diabetes Translation. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

Virginia Department of Health. (2019). *Diabetes and Prediabetes*. Data. http://www.vdh.virginia.gov/diabetes/data/

QUESTIONS?