SYS 6016 Final Project Report
Bradley Katcher and Elizabeth Driskill

Introduction

Our capstone project involved dealing with patient-level emergency room records from several hospitals across Central Virginia within the Carilion and University of Virginia health systems. The goal was first to analyze the demographic composition of patients presenting to the emergency room with diabetes-related symptoms and then to explore the relationship between climate variables and the proportion of diabetes patients presenting to the ER. While working on our capstone project, the main challenge we faced was deciding how to handle missingness within our dataset. For climate variables, we were able to impute the majority of the missing data, but for many of the categorical variables, like race and gender, we were unable to do this, forcing us to exclude patients with missing data for these variables from our analysis.

For our deep learning project, we wanted to implement deep neural networks to predict the missing values for race and gender within our dataset. Although we are finished working on our capstone project, the research will continue and hopefully be passed along to another group in the future, so we think it would be helpful to fill in some of this missing data so that future researchers will be able to incorporate it into their analysis. With these additional predicted values for race and gender included, we hope that researchers will be able to gain a better understanding of the demographic characteristics of the population of patients presenting to the emergency room with diabetes-related symptoms. Also, if this dataset is used for other purposes or to explore different diseases, filling in missing values for race and gender can still be beneficial to provide additional data that can be used in the exploration of the demographic composition of other patient populations.

Literature Review

Diabetes is an epidemic both nationally and in the Commonwealth of Virginia, with more than 30.3 million Americans living with diabetes, and another 84.1 million living with prediabetes (National Center for Chronic Disease Prevention and Health Promotion, 2017). Nearly one in every ten Virginians (approximately 631,000 individuals) has diabetes (Virginia Department of Health, 2019), with the Eastern Shore of Virginia most severely impacted with a prevalence rate of approximately 20.8% (Virginia Department of Health, 2017).

Diabetes patients without access to a primary care physician or who come from a lower socioeconomic background may have higher emergency room utilization for diabetes management (Ford et al., 2013). In 2010, there were approximately 12.1 million diabetes-related ER visits for adults, comprising almost 10% of all ER visits for adults. Roughly 42% of diabetes-related ER visits resulted in hospitalization, compared to roughly 15% of all ER adult visits (both with and without diabetes). These visit rates were greatest among patients aged 65 and older, from lower-income communities, and predominantly from individuals with government insurance (Washington et al., 2013).

Our dataset contained a significant amount of missing variables. Administrative data, specifically, administrative health records, as a whole often has the problem of lacking critical information. Logistic regression has often been used to impute missing data for categorical

variables (Sentas and Angelis, 2006). However, with the strong predictive classification power of deep, feedforward neural networks, there is strong potential to use deep learning for imputing missing values.

Race and ethnicity are often missing variables in important administrative health data sets, providing a bottleneck on their usage ability (Xue et. al., 2019). Yet, current methods such as multiple imputation lead to bias and loss of information and create multiple imputed datasets (Sterne et. al. 2009). The goal in most imputation is determining the distribution of mutually-exclusive categories (for race or gender), given a set of features.

RIDDLE (Race and ethnicity Imputation from Disease history with Deep LEarning) is a newly pioneered multi-layer perceptron (MLP) network containing two hidden layers of either ReLU or PReLU nodes, with inputs being the set of binary encoded features regarding age, gender, and diagnoses (via ICD-9 codes), and outputs comprised of the set of probability estimates for each of the four race and ethnicity classes, categorized via softmax (Kim et. al., 2018). We seek to apply the methods of RIDDLE to our dataset.

We hope that, by using deep neural networks to impute missing values for race and gender, future researchers will be able to have a more holistic view of the demographic composition of diabetes patients in Central Virginia (or other relevant patient populations) and thus be better equipped to provide these patients with quality healthcare.

Data

The data was shared with us through our capstone sponsor, Dr. Wendy Novicoff, and it can only be accessed on the Ivy Virtual Machine due to patient confidentiality restrictions. It consists of nearly 2 million emergency room records from seven different hospitals across central Virginia between the years of 2010-2017 via two different hospital systems, UVA and Carilion. Throughout this project, we worked with two separate csv files, one for the UVA emergency room, and one for the six emergency rooms within the Carilion health system. We limited the scope of this analysis to only include some of the variables present within our dataset, and these will be outlined below.

Each row in the dataset includes the date that the patient presented to the emergency room, demographic characteristics such as age, race, gender, and ethnicity, the principal diagnosis of the patient, and various climate features measured from the weather station closest to the patient's zip code on the day they presented. We also brought in data to characterize each patient's health insurance provider. Our goal was to predict missing values for race and gender, but many of the other variables within our dataset had missing values as well. In these cases, missing values were either encoded to other values like "None" for categorical variables or filled in with the average (the preferred choice from our project sponsor) for numeric variables so that they could still be incorporated into our analysis.

After removing redundant variables and non-pertinent features with significant missingness, we needed to preprocess the data for analysis. This included scaling quantitative variables to between 0 and 1, and one-hot encoding categorical variables. Additionally, by one-hot encoding zips and ICD-9 and ICD-10 codes, we would encounter a significant dimensionality problem. To deal with this, we used 3-digit zip-codes and ICD-9 and ICD-10 prefixes, so while we may not have the same level of detail, we still capture essential

components of these features. This left us with input nodes of 1,999 for the Carilion race, 2,006 input nodes for the Carilion gender, and 1,753 input nodes for the UVA race. The breakdown of races and genders for each of the hospital systems can be found below:

UVA Race		
W	268,927	65.81%
B	110,033	26.93%
O	17,969	4.40%
H	4,468	1.09%
A	3,739	0.91%
I	178	0.04%
None	3,323	0.81%

Carilion Race		
White	114,1838	82.26%
Black	184,013	13.26%
Hispanic	29,833	2.15%
Biracial	15,593	1.12%
Asian	7,170	0.52%
Other	5,275	0.38%
Am Indian	1,043	0.08%
Pac Islander	460	0.03%
None	2,791	0.20%

Carilion Gender:

Females	764,982	55.11%
Males	620,725	44.72%
None	2,309	0.17%

Methods/Analysis

We implemented feedforward deep neural networks to impute missing values for both race and gender within the Carilion dataset and for race within the UVA dataset (this dataset was not missing any values for gender). We first encoded missing values in variables other than race and gender to ‘None’ for categorical variables and the mean of the column for numeric variables. Then, we scaled the numeric variables and used one-hot encoding to create dummy columns for the categorical variables. Next, we performed a train-validation split for each dataset.

Using keras, we built four different neural networks for each of the categories of missing values (Carilion race, Carilion gender, UVA race), summing up to 12 neural networks in total. The four neural networks were as follows: 2-3 hidden layers using ‘relu’ activation function, 3 hidden layers using ‘prelu’ activation function, 5 hidden layers using ‘relu’ activation function, and 5 hidden layers using ‘prelu’ activation function. Each hidden layer consisted of either 64 or 128 nodes, and the final output layer used the ‘softmax’ activation function when predicting race and the ‘sigmoid’ activation function when predicting gender. We also incorporated dropout layers to adjust for overfitting. The number of output nodes varied depending on the number of unique categories for race and gender. We plotted the training and validation loss and accuracy

curves for each neural network and then predicted the missing values for the test data (either race or gender). The exact networks are as follows:

Carilion Race: (all had softmax output with 8 nodes)

- 64 ReLU nodes, 128 ReLU nodes
- 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes
- 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes
- 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes

Carilion Gender: (all had sigmoid output with 1 node)

- 64 ReLU nodes, 0.3 dropout, 64 ReLU nodes, 0.3 dropout, 64 ReLU
- 64 PReLU nodes, 0.2 dropout, 64 PReLU nodes, 0.2 dropout, 128 PReLU nodes
- 64 ReLU, 64 ReLU, 64 ReLU, 0.3 dropout, 64 ReLU, 0.3 dropout, 64 ReLU
- 64 PReLU, 64 PReLU, 64 PReLU, 64 PReLU, 128 PReLU

UVA Race: (all had softmax output with 6 nodes)

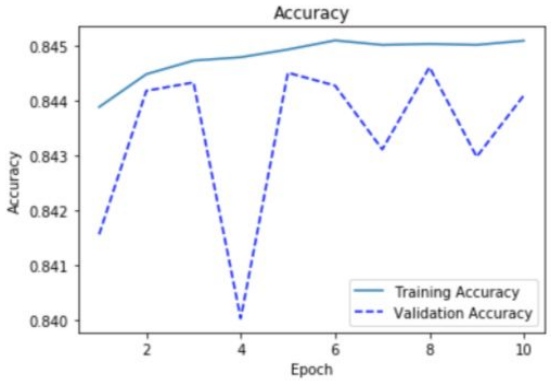
- 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes
- 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes
- 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes, 64 ReLU nodes
- 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes, 64 PReLU nodes

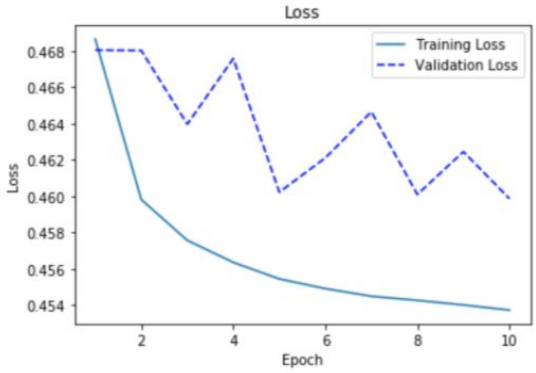
Results

The deep neural networks we implemented were able to predict missing values for race and gender. For predicting race in the Carilion dataset, the neural network with 3 hidden layers of 64 nodes and the PReLU activation function resulted in the lowest loss and highest accuracy. For predicting race in the UVA dataset, the neural network with 5 hidden layers of 64 nodes and the PReLU activation function resulted in the lowest loss and highest accuracy. Plots of training and validation loss and accuracy for the best performing neural network in each case can be seen on the next page. For Carilion, of the 2,791 missing races, our neural network predicted 2,741 white individuals, 3 black individuals, and 47 hispanic individuals. It is worth noting that the predictions on the validation set failed to predict any individuals as “Other,” “Am Indian,” or “Pac Islander.” For UVA race, all but one individual of the 3,323 were predicted to be white, however, the predictions on the neural network did contain all races in the sample, indicating that the network likely learned all of the races.

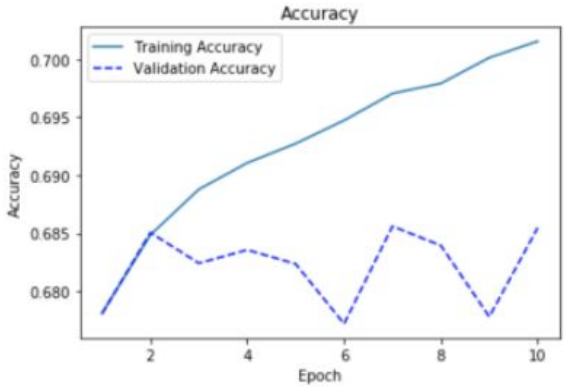
When predicting gender in the Carilion dataset, we found that, in each of the four neural networks, the returned predicted values were all male, except for one observation. There are two possibilities to explain these results: 1. The data itself is biased such that the observations that have gender excluded are predominantly males, or 2. The methods we applied are not adequate for predicting gender, and alternative approaches should be explored. Regardless, there were not very many missing values for gender in the first place, so it will not make a huge difference if these observations are excluded in future analysis.

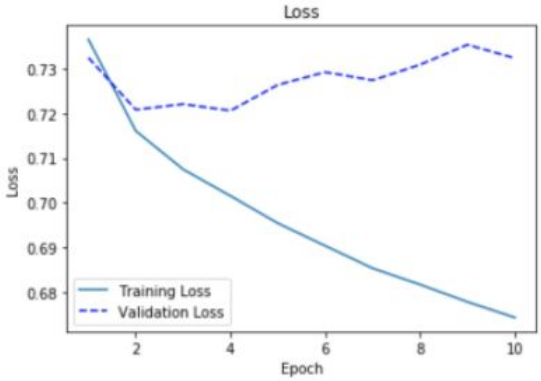
Carilion Race:



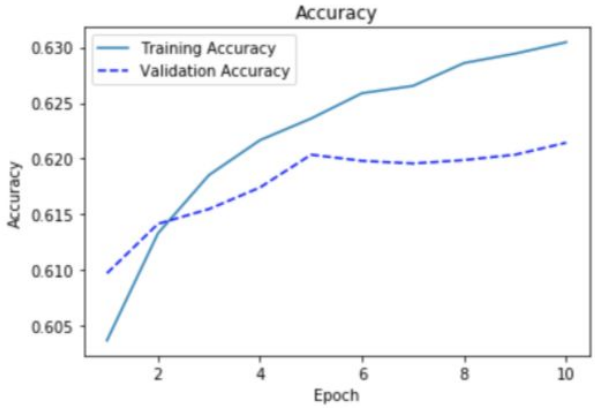


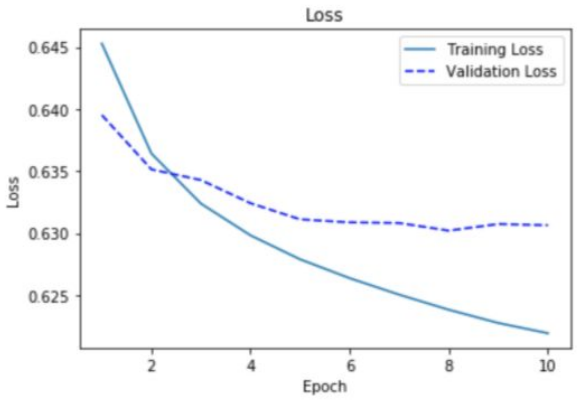
UVA Race:





Carilion Gender:





Conclusion

While deep neural networks did not do the best job of predicting missing values for gender, they were successful in predicting missing values for race in both the Carilion and UVA datasets. We also noticed that the PReLU activation function slightly outperformed the ReLU activation function, suggesting that this should be retained in future models and future research. We wonder if these models perhaps needed to train longer in order to be able to distinguish more factors, we simply ran for 10 epochs due to computational limitations and large training times. Given that the datasets were so "white-heavy" (making up 65-82% of the total sample) to begin with, we may have experienced some issues in accurately predicting race and might want to explore a loss function that punishes misclassification more intensely.

We hope that this additional demographic information will assist future researchers in their analysis of different patient populations and social determinants of health. These methods could also be applied to impute other missing values for categorical variables and could provide a better alternative than simply removing rows with missing values from the dataset.

In the future, we would like to experiment with using different numbers of hidden layers, different sizes of hidden layers, adding or removing more dropout layers, and using different activation functions to see how these changes may produce better or worse results. We also think it would be useful to explore different methods to predict missing values for gender, possibly incorporating different types of neural networks.

References

Ford, W., Self, W. H., Slovis, C., & McNaughton, C. D. (2013). Diabetes in the Emergency Department and Hospital: Acute Care of Diabetes Patients. *Current Emergency and Hospital Medicine Reports*, *1*(1), 1–9. https://doi.org/10.1007/s40138-012-0007-x

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

Sentas, Panagiotis, and Lefteris Angelis. "Categorical Missing Data Imputation for Software Cost Estimation by Multinomial Logistic Regression." Journal of Systems and Software, vol. 79, no. 3, Mar. 2006, pp. 404–14. DOI.org (Crossref), doi:10.1016/j.jss.2005.02.026.

Sterne, Jonathan A. C., et al. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls." BMJ, vol. 338, June 2009. www.bmj.com, doi:10.1136/bmj.b2393.

Virginia Department of Health. (2017). *Diabetes Burden in Virginia* (Diabetes Burden Report). http://www.vdh.virginia.gov/content/uploads/sites/25/2016/05/Diabetes-in-Virginia-2017_final_7_17.pdf

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

Washington, R., Andrews, R., & Mutter, R. (2013). *Emergency Department Visits for Adults with Diabetes, 2010* (Statistical Brief No. 167; Healthcare Cost and Utilization Project). Agency for Healthcare Research and Quality. https://www.hcup-us.ahrq.gov/reports/statbriefs/sb167.jsp

Xue, Yishu, et al. “Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data.” 2019 13th International Conference on Sampling Theory and Applications (SampTA), 2019, pp. 1–4. IEEE Xplore, doi:10.1109/SampTA45681.2019.9030977.