PROJECT TITLE: Predicting Opioid Addiction Likelihood Using Neural Nets

PROJECT SUMMARY: Utilize artificial and convolutional neural networks in TensorFlow to classify individuals at risk for opioid addiction based off socioeconomic, familial, and personal health data.  Datasets include US Census Data, Center for Disease Control, and National Institute on Drug Abuse.

MILESTONES:

* Data collection - DONE
* Data exploration – IN PROGRESS
* Feature Exploration - TO BE DONE
* Build models: ANN, CNN - TO BE DONE
* Model Tuning – TO BE DONE
* Result visualization - TO BE DONE

PROPOSED ‘TO DO’ FROM THE LAST WEEK (Copy & Paste from your previous week’s TO DO)

*This week I need to focus on finalizing the datasets that I actually intend to use, and continuing to research the opioid epidemic as well to get a better understanding of what sort of features contribute to the likelihood of addiction. I hope to have all the data compiled and stored in AWS by the end of this coming Week in order to start the exploration*

THIS WEEK’S PROGRESS

This week I started to have concerns that the data the I needed to complete my practicum may be non-existent, due to HIPPA regulations. Essentially, I know that I need one file that represents an individual and measures related to their demographics, socio-economic status, family history, etc. I would also need a response variable that indicates if they are an opioid addict. Medical record information is notoriously difficult to find, understandably so. However, after MUCH searching, I found a database that contains information on individuals that were admitted to a hospital for some sort of substance abuse. This includes 17 million+ records. I finished this week with some basic visualizations.

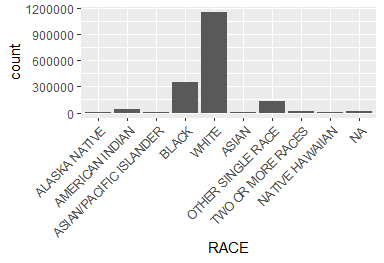
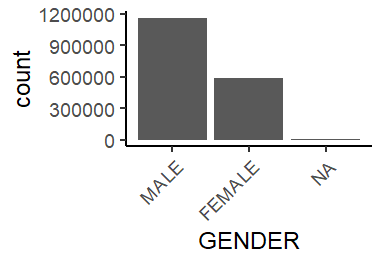
ISSUES AND DISCUSSION:

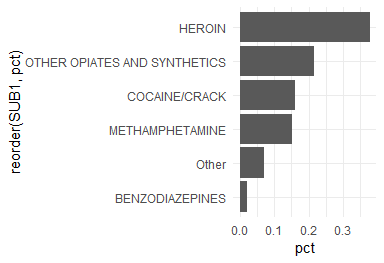
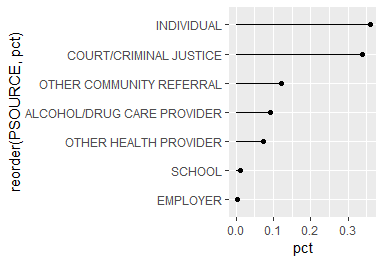
A large part of this data is not normalized, and nearly every row has at least one missing value. Each data point is a categorical variable, so I will likely need to do one-hot encoding on these variables in order to feed them to the model appropriately.

TO DO:

By the end of this week I intend on loading all my data into an AWS S3 bucket, and exploring how I may be able to take advantage of their RDS services in order to store the data as a table and connect it directly to R in order to query it. I also intend to complete the data cleansing process in order to complete my overall exploration.

 Visualizations:





We can see from the beginning of the basic exploration, what sorts of individuals are being sent to the hospital, for what substance, and by who. Ultimately, the biggest demographic consists of white males, with heroin, or other opiates in their system, who are admitted by wither themselves, or by the criminal justice system.

RESOURCE (Optional: list resource or links you want to share with me):

<https://www.cdc.gov/drugoverdose/epidemic/index.html>

<https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis>

<https://www.hhs.gov/opioids/about-the-epidemic/index.html>