PPOL670 FINAL PROJECT

Predicting Depression and Anxiety using Supervised Learning Models

Jagir Pipalia

Introduction

In 2015, over 43 million adults had a mental illness, and nearly 10 million had a serious mental illness, such as depression, bipolar disorder, or schizophrenia . People with mental health conditions often have chronic medical conditions, significant health care services utilization, and barriers to employment, and are frequently involved with the criminal justice system.[1]

In this paper I used data from the National Health Interview Survey to generate a supervised learning model that predicts the occurrence of depression, anxiety, or emotional problems among respondents.

I utilize the National Health Interview Survey to identify datasets that contain variables that are relevant in predicting the occurence of depression and anxiety. I explored the datasets to determine the variables that may predict the occurence of depression and anxiety. I import the variables in R to create a R dataset. I than take a random sample from the dataset for building a supervised learning models to predict the occurence of depression and anxiety. I compared the models to find the model with the best predictive ability.

Problem Statement and Background

Depression is among the leading causes of disability in industrialized countries.To effectively target interventions for patients at risk for a worse long-term clinical outcome, there is a need to identify predictors of chronicity and remission at an early stage. This could allow a quicker escalation of treatment for patients with a low long-term chance of recovery, thus potentially avoiding initial treatment resistance.[2]

[1]KFF. (2019). Facilitating Access to Mental Health Services: A Look at Medicaid, Private Insurance, and the Uninsured. [online] Available at: https://www.kff.org/medicaid/fact-sheet/facilitating-access-to-mental-health-services-a-look-at-medicaid-private-insurance-and-the-uninsured/:~:targetText=The%2 [Accessed 12 Nov. 2019].

[2]Dinga, R., Marquand, A. F., Veltman, D. J., Beekman, A., Schoevers, R. A., van Hemert, A. M., ... Schmaal, L. (2018). Predicting the naturalistic course of depression from a wide range of clinical,

Neuroscientists and clinicians around the world are using machine learning to develop treatment plans for patients and to identify some of the key markers for mental health disorders before they may set in. One of the benefits is that machine learning helps clinicians predict who may be at risk of a particular disorder.There is so much data available that we are now able to compile data for mental health professionals so they may do their job better. What makes machine learning so helpful today is that in the past, understanding of diagnoses were based off group averages and statistics over populations. Machine learning gives clinicians the opportunity to personalize.[3]

David and colleagues use a layered, hierarchical model for translating raw sensor data into markers of behaviors and states related to mental health.[4]

Through this project, I aim to create a supervised learning model that can accurately predict if a person is at risk of experiencing depression or anxiety. I take health variables on several diseases such as Hypertension, Cholestrol, COPD, Diebetes, etc and combine it with a person's age, weight, race, and region they combine along with other variables to predict the occurence of depression or anxiety.

Data

The data for the project comes from The National Health Interview Survey (NHIS). NHIS has monitored the health of the nation since 1957. NHIS data on a broad range of health topics, including mental health conditions, are collected through personal household interviews. The U.S. Census Bureau has been the data collection agent for the National Health Interview Survey. Survey results have been instrumental in providing data to track health status, health care access, and progress toward achieving national health objectives.

The unit of observation is survey participant. Each participant is an observation. I selected 30 variables from the NHIS datasets that seemed relevant to the question. There were 104874 observations in total. The variable of interest was AFLHCA17 which is marker for if depression, anxiety, or emotional problems caused difficulty with activity.

Most of the variables are categorical values. Some variables appear to be continues variables. There is a lot of missing data, so I will impute the missing data.

---

psychological, and biological data: a machine learning approach. Translational psychiatry, 8(1), 241. doi: 10.1038/s41398-018-0289-1

[3]Abbas, N. (2019, August 29). Machine Learning and Mental Health Can Big Data help the Mental Health field? Retrieved December 13, 2019, from https://towardsdatascience.com/machine-learning-and-mental-health-7981a6001bd5.

[4]Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology, 13(1), 23–47. doi: 10.1146/annurev-clinpsy-032816-044949

Data Wrangling

I first imported the datasets 'dataset_adult' and 'dataset_person'. I then selected the following variables using the tidyverse package:

AFLHCA17,RECTYPE,REGION,SEX,HISPAN_I,RACERPI2,AGE_P,R_MARITL,DOINGLWA,HYPY
PHSTAT

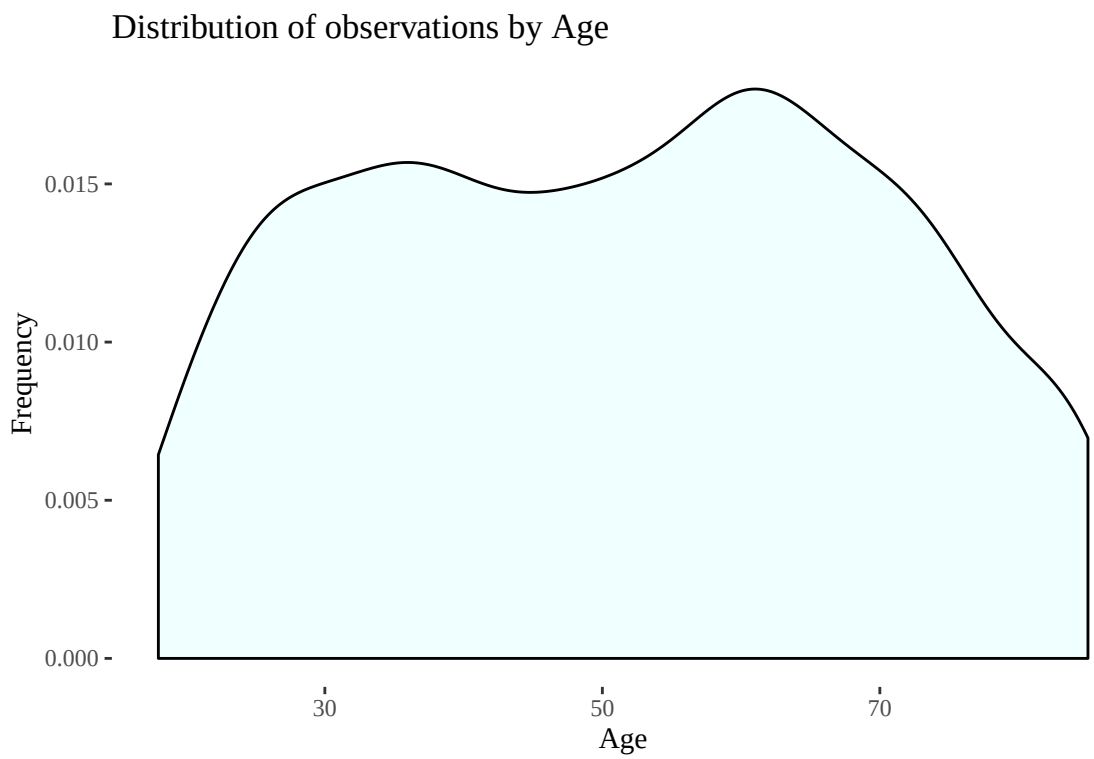
I then merged the two datasets using 'full_join' function to form the 'final_dataset_raw'.In order to minimize the time to run the machine learning models, I took a random sample of 10,000 observations from 'final_dataset_raw'.The sample was named 'final_dataset'. I divided the 'final_dataset' into training data and test data. Training data has 75% of the observations while test data has the remaining 25% of the data.

Vizualizations for Independent Variables

Following are vizualizations for some of the indepedent variables. As we can see from the density graph for Age variable. Majority of participants seem to be above 50 years of age.

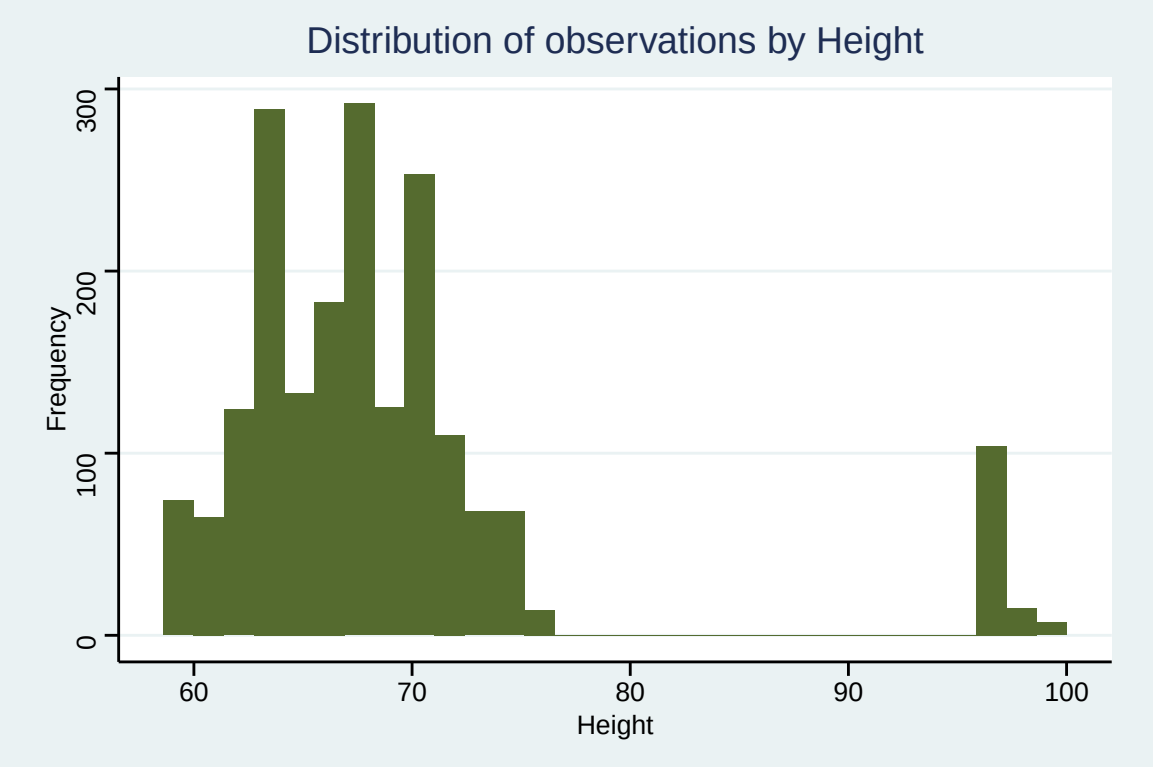
```
## Warning: Ignoring unknown parameters: bins
```

```
## Warning: Removed 5576 rows containing non-finite values (stat_density).
```



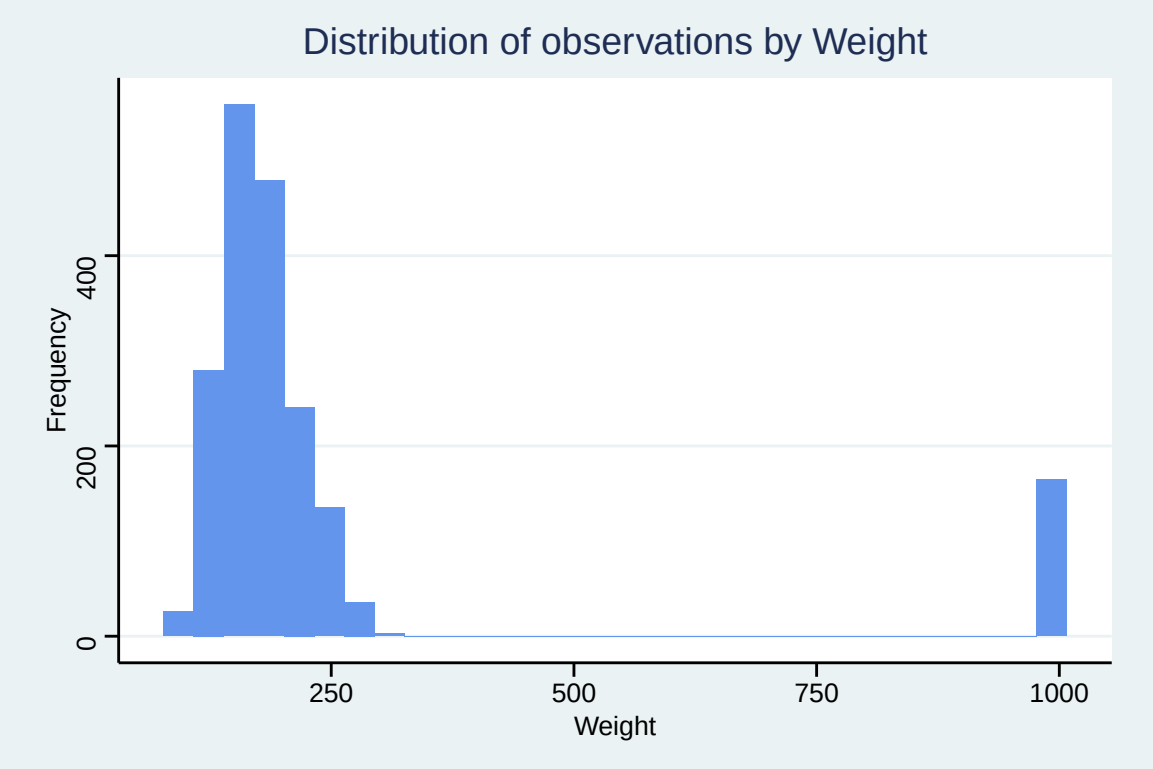
From the bar graph below for Height variable, we can see that there are some outliers.

```
## Warning: Removed 5576 rows containing non-finite values (stat_bin).
```

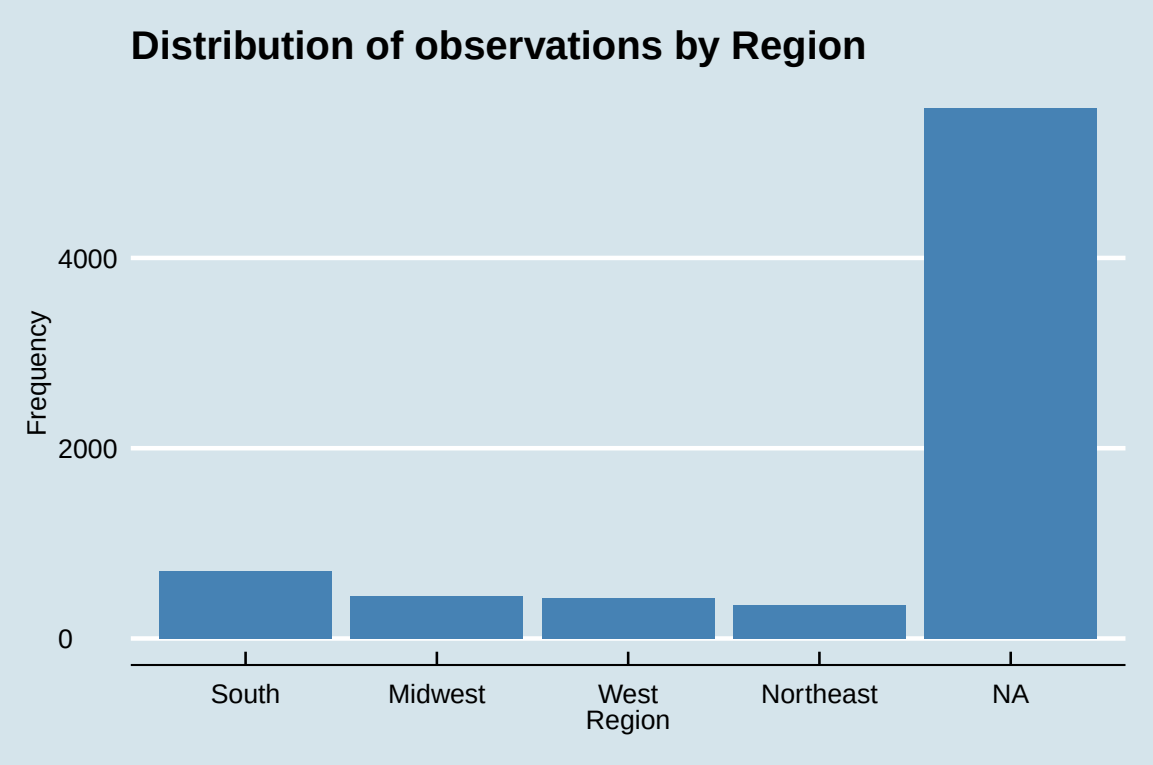


Similarly, as evident in the histogram below, the Weight variable also has some outliers:

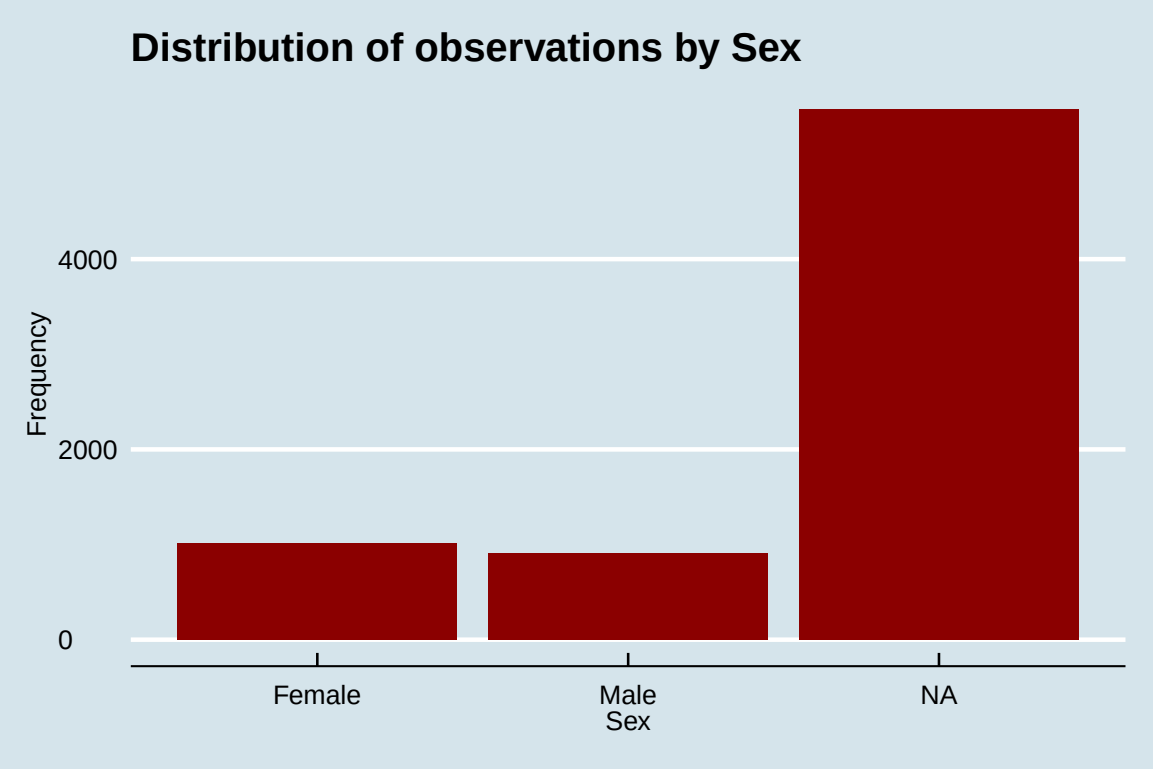
```
## Warning: Removed 5576 rows containing non-finite values (stat_bin).
```



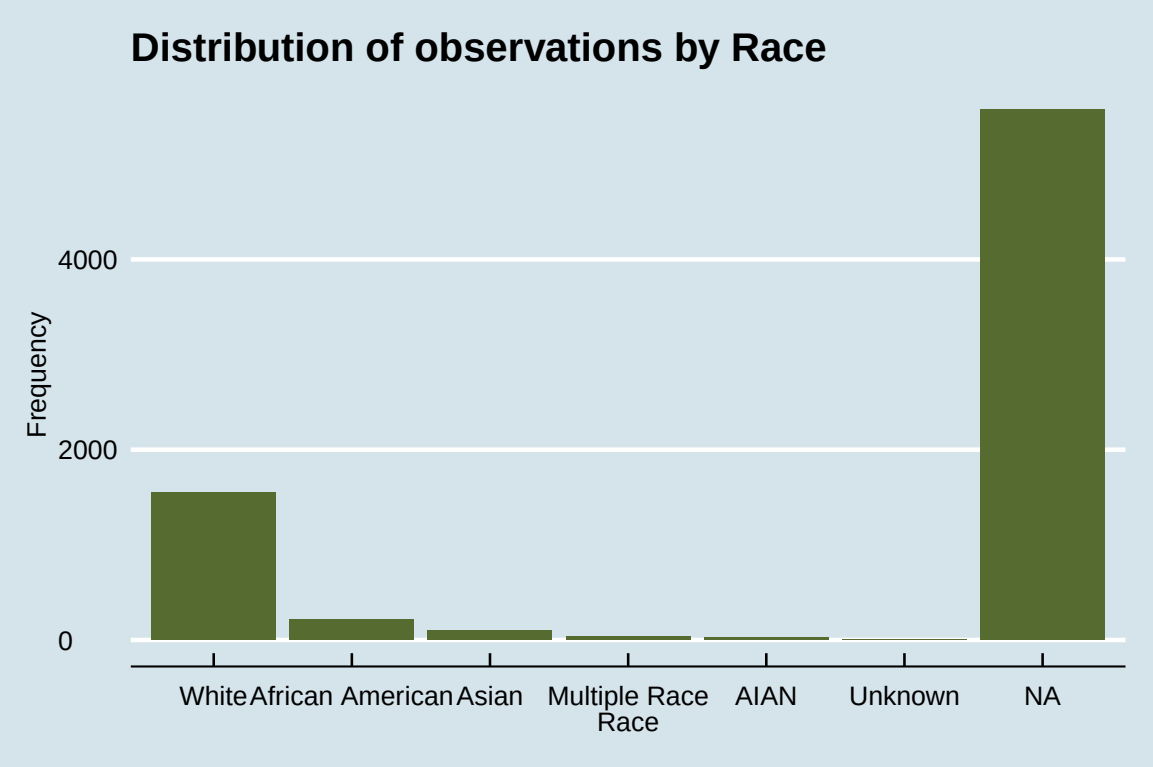
Following is a bar graph for Region variable. The highest number of observations were from the South region.



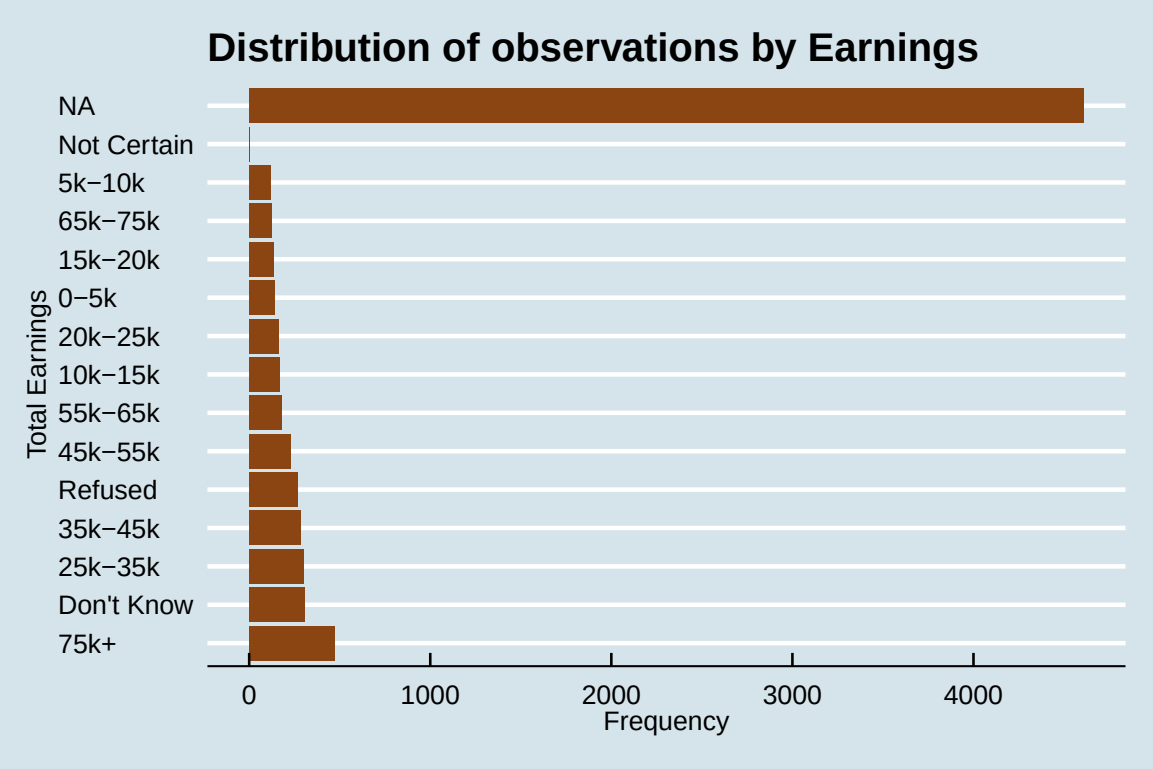
Following is a bar graph for Sex variable. There are more females in the dataset than males.



Following is a bar graph for Race variable. The highest number of observations are White.



Following is a bar graph for Earnings variable. There are significant number of observations that earn more than $75,000.



Pre-Processing

As it can be seen from the above graphs, the sample has a high number of missing values. I generated a recipe using the recipes package to impute the missing data and scale the data. I also created a dummy variable for each level of the factor variables for each factor variable.

After processing the data using the recipe generated, the data was ready to be used in supervised learning models.

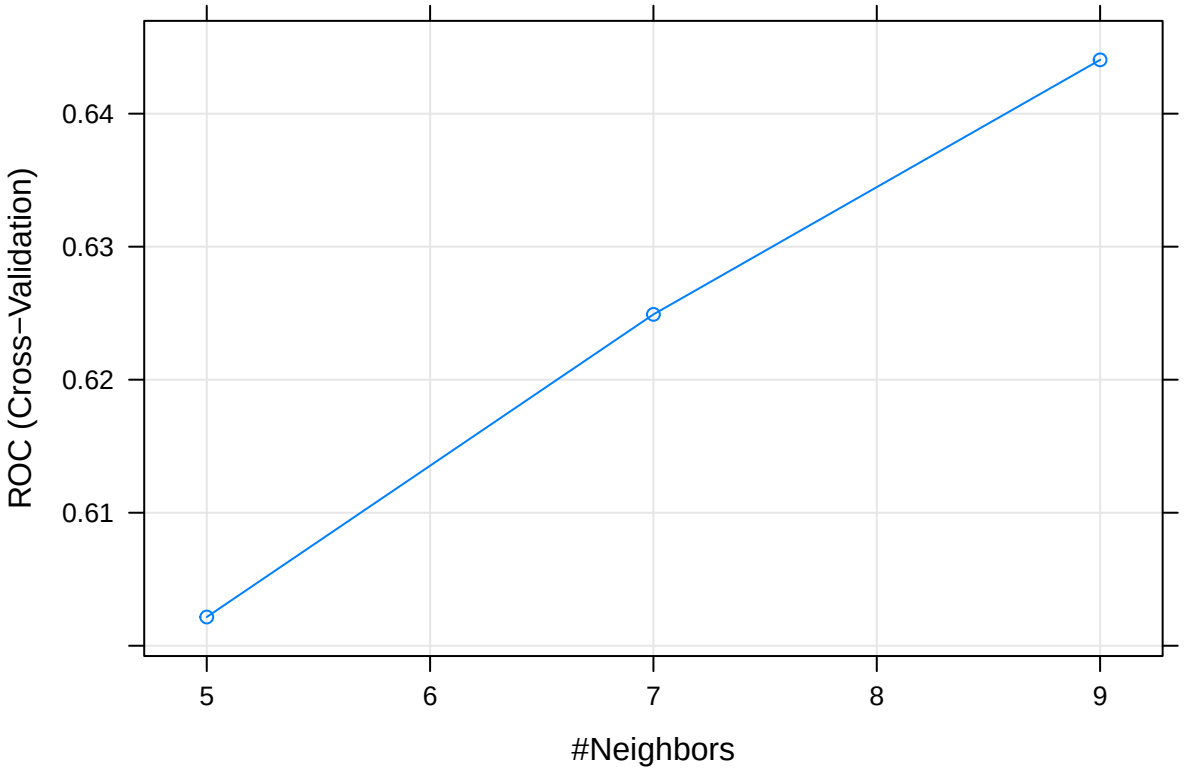
Analysis

I started by setting cross-validation settings. I used k-fold validation. I used 10 folds as the training data contains 7500 observations. Each fold was about 750 observations. The dependent variable, occurence, in my analysis is a dummy variable which takes the value 1 when the observation experienced depression or anxiety. It takes the value 0 when the observation did not report experiencing depression or anxiety.

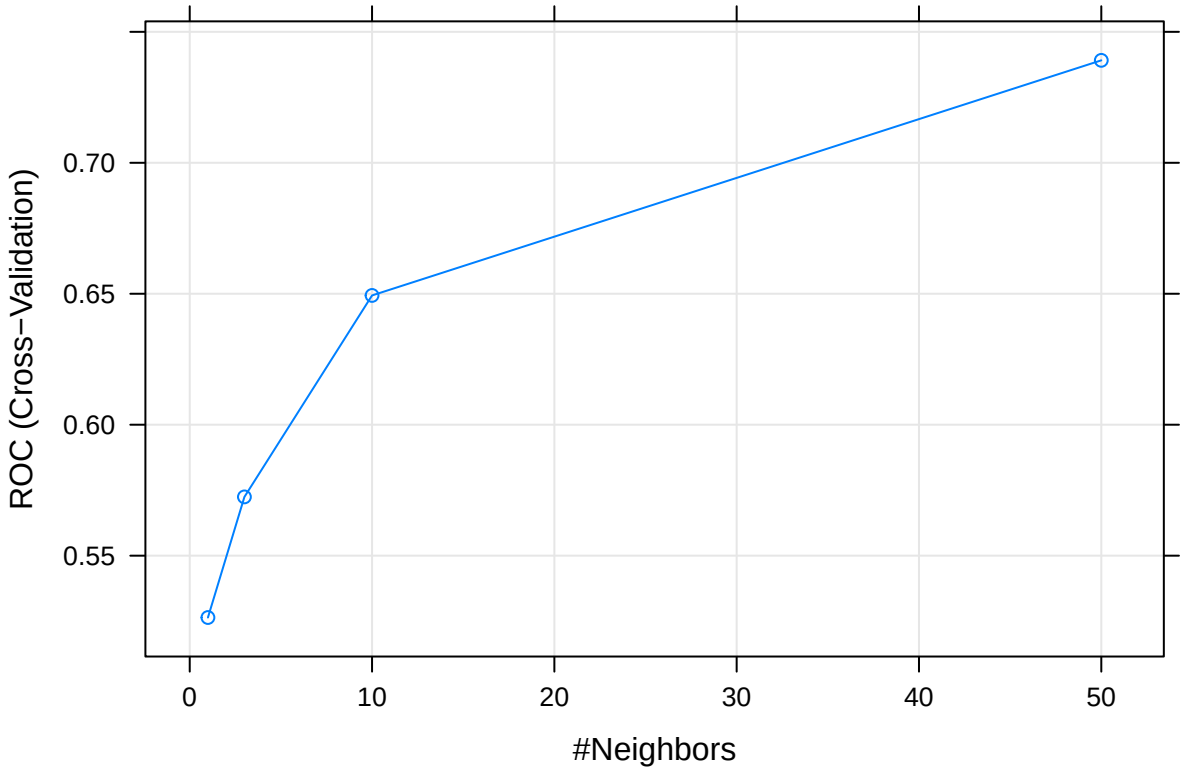
I began with my first model which was Logistic Regression, I set the metric to be 'ROC' as the dependent variable is a dummy variable. As we can see from the output, the ROC from the Logistic Regression is close 0.5, which is almost as good as a coin-toss. Therefore, I decided to use KNN model next.

```
## Generalized Linear Model
##
## 7500 samples
##  104 predictor
##    2 classes: 'Not_occur', 'occur'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 750, 750, 750, 750, 750, 751, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.4960952  0.9719496  0.2550505
```

As we can see from the output below, the KNN model with k=9, perfomed better than other knn models and also better than Logistic Regression.The ROC was close to 0.6. In order to see if the ROC would increase with a higher k, I run a tuned KNN model with the highest k equal to 50.

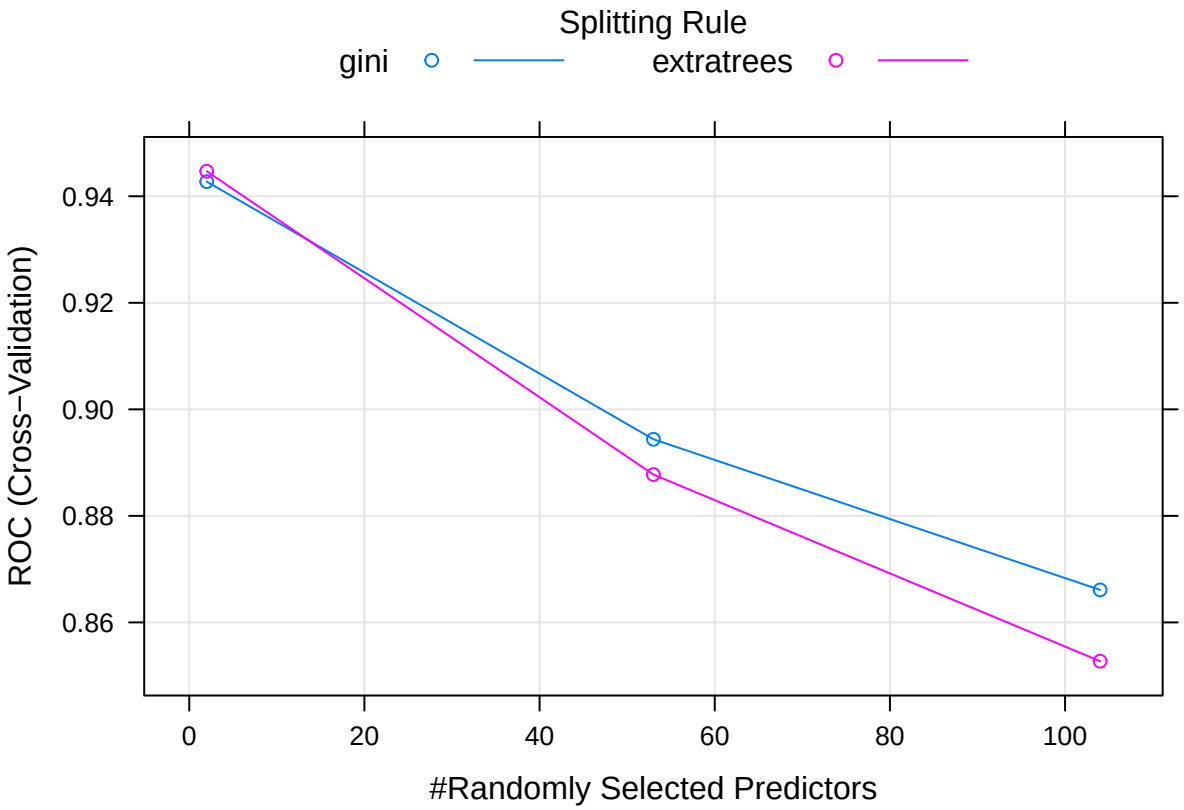


As we can see, the ROC continues to improve with higher ks and reaches close to 0.75 when k=50. I run the CART and Random Forest models next to see how well they do in predicting



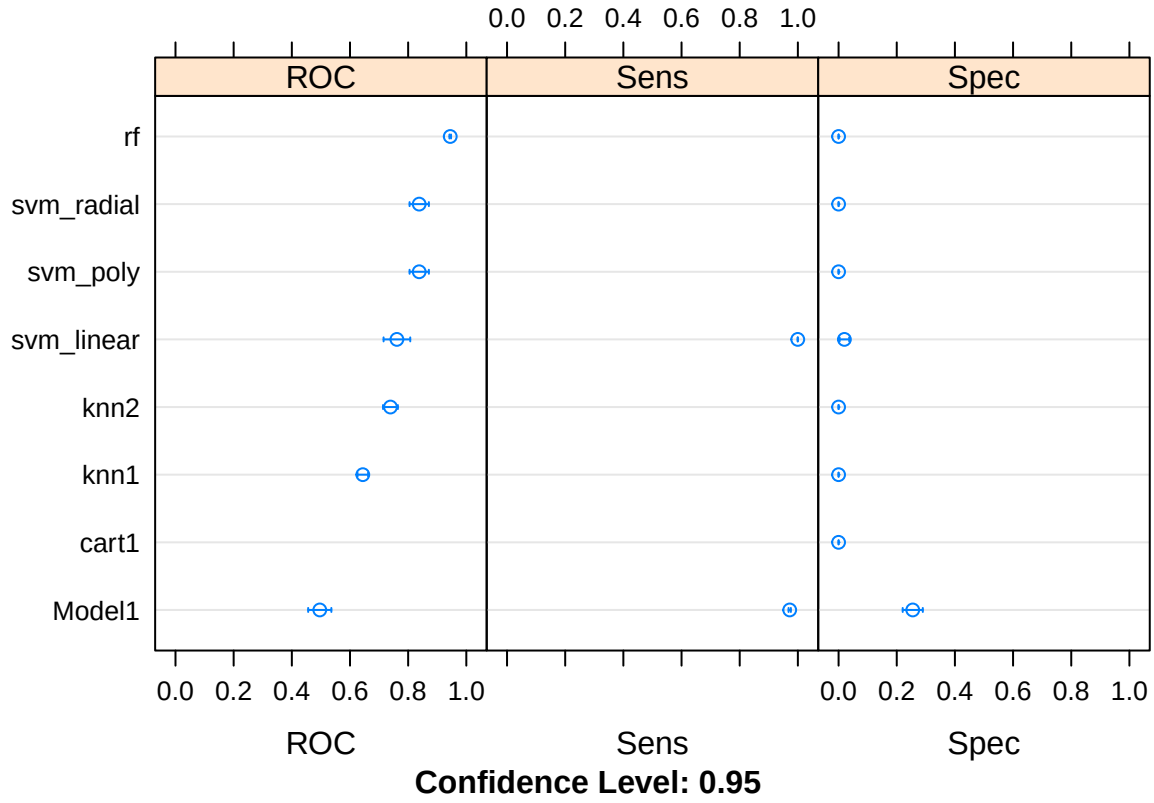
occurrence.

The CART model had an ROC of 0.5. The random forest model had a much higher ROC of 0.94 with 2 randomly selected variables and using gini splitting rule. I then run SVM models followed by a Polynomial Boundary and Radial boundary model. The SVM Model has a ROC close to 0.7 while the Polynomial Boundary and Radial Boundary models have ROC close to 0.81.

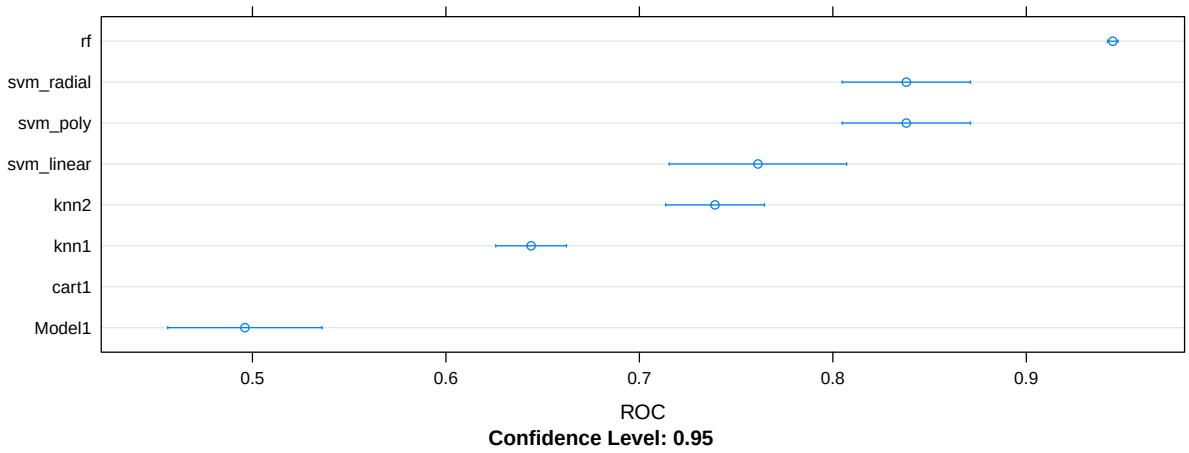


Results

We ran Logistic Regression, KNN, KNN with tunning parameters, CART, Random Forest, SVM Linear, SVM Polynomial Boundary, and SVM Radial Boundary models. Following dotplot compares the performace of models with each other. As we can observe the random forest model performed the best in terms of ROC. The SVM models also performed very well.



The following plot gives a closer look at comparision of ROC accross models



Since, the random forest model performed the best, we will use to predict the occurrence of dpression or anxiety in the test dataset. As we can see from the results below, the model predicts 0 occurence of depression or anxiety in tre test dataset.

```
## Not_occur     occur
##      2495         0
```

Discussion

Although the Random model and SVM models had a high ROC, we did not see a good prediction for occurrence from the test dataset. I think the project was partially successful as it was able to create a good predictive model for occurrence of depression or anxiety as it had high ROC but it was not good in actually predicting the occurrence from the test dataset.

I considered using a larger sample and variables with more variation, but due to time and computational constraints, I was not able do it. In the future, I plan to use a bigger sample along with more suitable variables which have a lower level of missing vlaues.