131 Final Project: Dementia Classification

Keon Dibley

Contents

In this document, I will be working with data I found on Kaggle, compiled by user Fatemeh Mehrparvar. The original dataset information can be found at the link below. The original data was collected by researchers who relayed their findings in a research paper, which I have also cited below.

Data set link: https://www.kaggle.com/datasets/fatemehmehrparvar/dementia/data

Research paper citation:

- Amin Al Olama, Ali et al. "Simple MRI score aids prediction of dementia in cerebral small vessel disease." Neurology vol. 94,12 (2020): e1294-e1302. doi:10.1212/WNL.0000000000009141/

Introduction

In this classification project, we will be identifying whether or not a patient has dementia using different medical and personal data. Dementia is not a specific disease, but rather a general term that refers to a decline in mental ability. Suffering from dementia can impact one's ability to remember, think, and make decisions. For example, the most common cause of dementia is Alzheimer's disease, which is characterized by the death of brain cell connections. Some of the predictors of dementia which we will explore are white matter brain damage, presence of lacunes in the brain, and presence of microbleeds in the brain. More information on these predictors and what they mean can be found in the codebook section of this document.

Also, the data used in making this model was collected from three different studies, labeled "scans", "ASPS", and "rundmc" in the data set. The "ASPS" study is split into two different cohorts in this data set, which are labeled "ASPS-family" and "ASPS-elderly". Any information surrounding the original data collection that was referenced in this document can be found at the following link to the original study: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7274929/. The goals of this project are to find the optimal model to predict dementia presence in patients and to find the most important features and things to look out for when predicting dementia. After completing this project, I hope to have a better understanding of dementia risk factors along with a model that can be make accurate predictions on new data.

Codebook

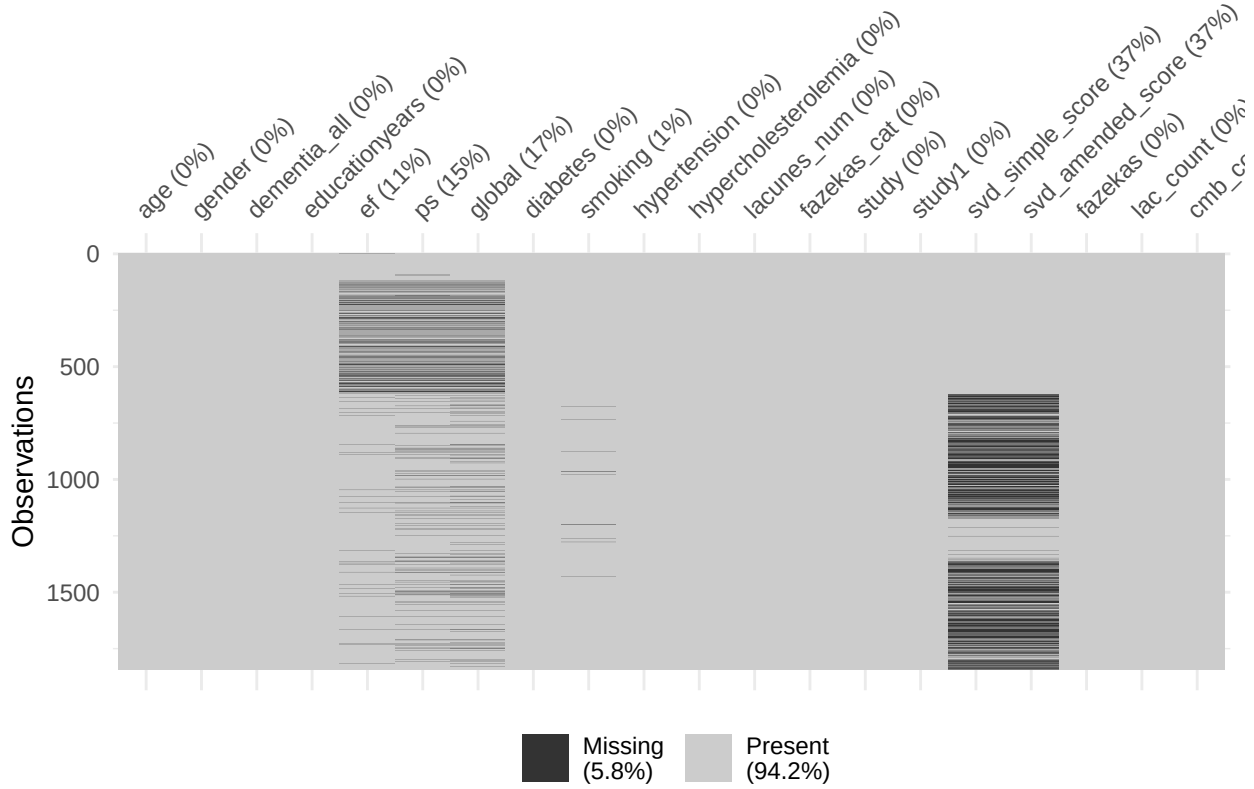
- age - Age of patient
- gender - Patient's gender
- dementia_all - Presence of dementia, 1=dementia (excluding missing values, response variable)
- educationyears - Length of education (in years)
- EF - Executive function (numeric variable, describes the level of mental skills, can be assessed through neurological tests)
- PS - Processing speed (numeric variable, evaluates the brain's ability to process information, it affects one's ability to use executive functions)
- Global - Global cognitive score (assesses overall status of cognition)
- diabetes - Presence of diabetes (1 - yes, 0 - no)
- smoking - Smoking status (categorical, describes "current smokers", "non-smokers", "ex-smokers"
- hypertension - Presence of hypertension (high blood pressure) in the patient ("yes"/"no)
- hypercholesterolemia - Hypercholesterolemia (presence of high cholesterol levels in the blood, "yes"/"no")
- lacunes_num - Number of lacunes (categorical, binary, "zero"/"more-than-zero"), lacunes are small cavities in the brain which can be indicative of diabetes/cognitive impairment
- fazekas_cat - Indicates level of white matter brain damage, uses Fazekas scale: 0 = absent, 1 = "caps" or pencil-thin lining, 2 = smooth "halo", 3 = irregular periventricular signal extending into deep white matter, information found outside of dataset: https://radiopaedia.org/articles/fazekas-scale-for-white-matter-lesions?lang=us categorizes scores into "0 to 1" and "2 to 3".
- study1 - indicates which study the observation originated from ("scans", "rundmc", "ASPS")
- study - same as study1, but splits "ASPS" into "ASPS-elderly" and "ASPS-family", indicates family and elderly cohorts of ASPS study. Information found about the studies wasn't in the dataset, but at this link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7274929/
- SVD Simple Score - SVD (Small Vessel Disease) Simple Score, measures the signs of SVD, includes factors like lacunes, fazekas score, and microbleeds
- SVD Amended Score - SVD Amended Score, similar to Simple Score, but with a wider range, weighting factors more heavily.
- Fazekas - numerical variable, has the same meaning as fazekas_cat, gives discrete value ranging from 0 to 3.
- lac_count - Same meaning as lacunes_num, but describes the category "more-than-zero" in more detail: "1 to 2", "3 to 5", and ">5"
- CMB_count - categorical variable with two values: 0 and at least one, describes cerebral microbleeds, which are small brain hemorrhages.

Exploratory Data Analysis

First, I read in the data, loaded relevant packages, and cleaned the predictor names to make it easier to subset the data. Also, I removed the unnecessary variables *ID* and *dementia* (duplicate variable).

Missingness

Before modeling my data, I wanted to examine missing values in the data. To do so, I created a plot to visualize missingness by missing predictor value, which is shown below:



In this plot, we can see that there is missing data for the variables `ef`, `ps`, `global`, `svd_simple_score`, and `svd_amended_score`. For these variables, a majority of the data is still present, so it makes sense to use imputation to generate reasonable values for these missing values.

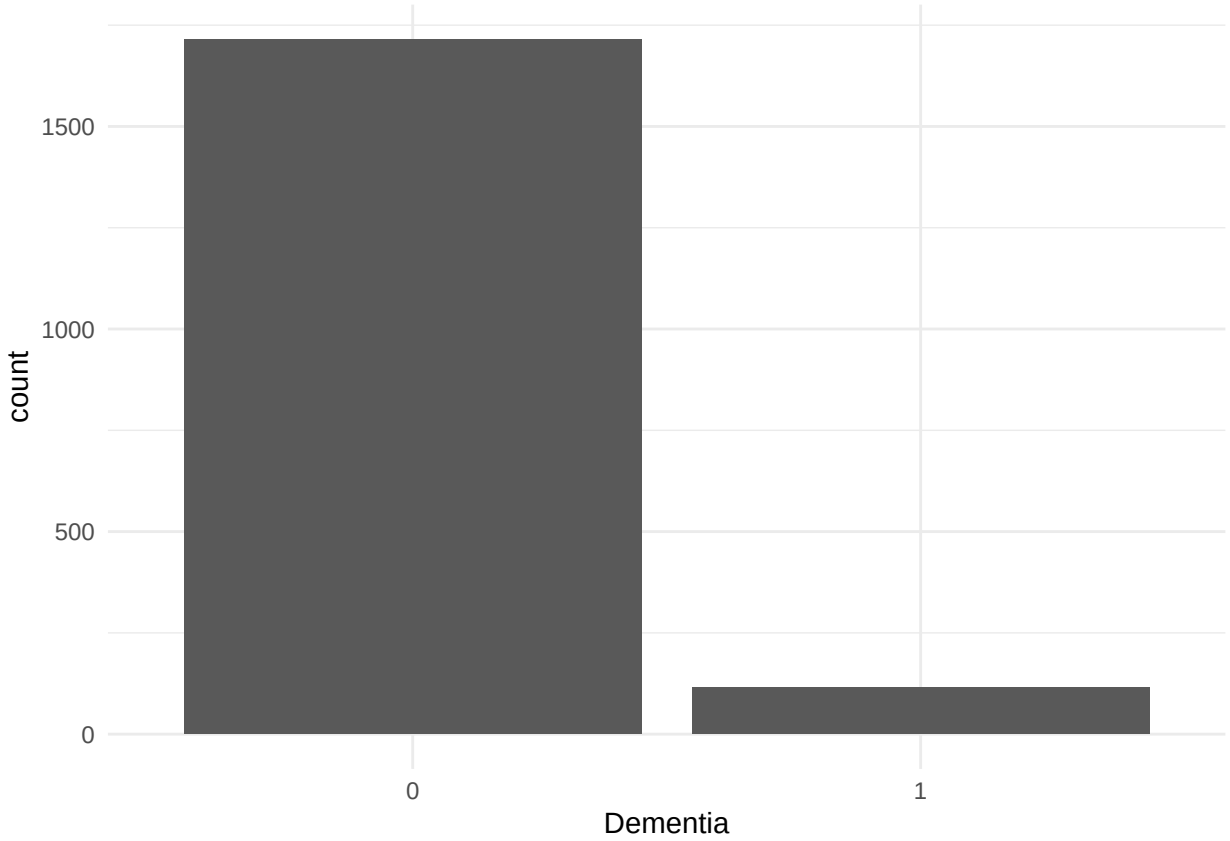
Through investigation of the original research investigation, I discovered that there is a pattern to the SVD score missing data. For the ASPS study, many participants didn't have complete MRI data available, which lead to these particular missing values. Since the column with the largest proportion of missing values is only 37%, we will try to use imputation for all missing values to avoid bias in our model, which could be caused by dropping these values. We will touch on this again when creating our recipe.

Since `smoking` only has 1% missing data, we will just drop these rows, for simplicity's sake.

After dealing with missing data, I factored the categorical variables in my data set, so that they could be included in my model.

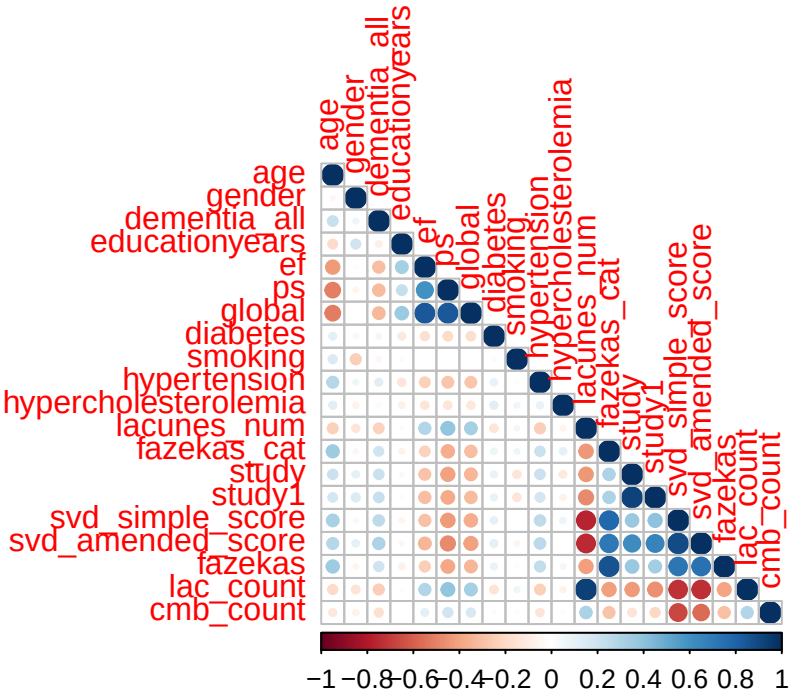
Data Visualization

First, I wanted to visualize the distribution of the response variable, `dementia_all`, to get a good idea of what I would be predicting:



Clearly, we are dealing with an imbalanced response variable. This could cause complications in predicting patients who *do* have dementia, so we will deal with this by upsampling when creating our recipe.

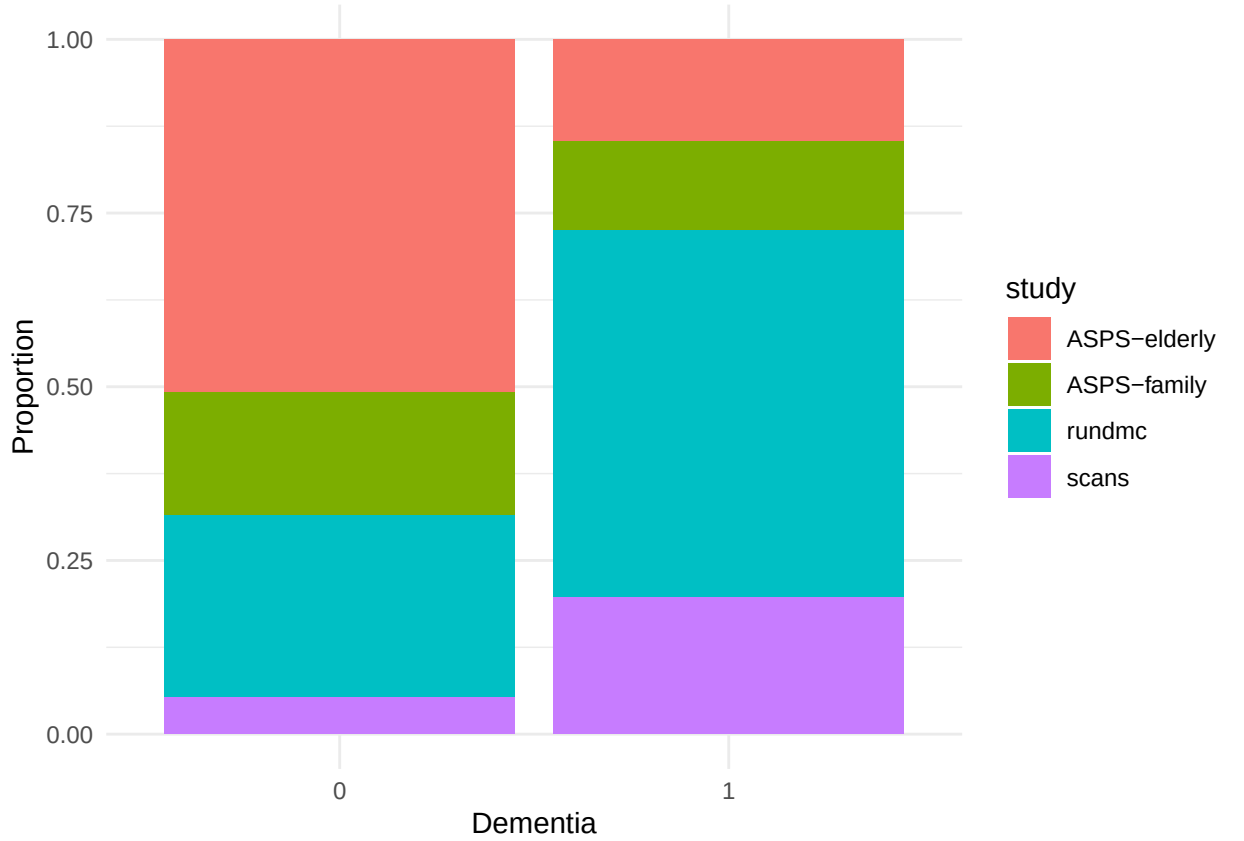
Next, here is a correlation plot of the relevant variables:



From this plot, we can see that our response variable, `dementia_all`, has strong negative correlations with `ef` (Executive Function), `ps` (Processing Speed) and `global` (Global cognitive score). Also, `dementia_all` has a positive correlation with the SVD (Small Vessel Disease) variables, as well as with the the variables that are used to compute this score (lac_count, fazekas, etc.).

Something I find interesting, yet makes sense, is the negative correlations between `ef`, `ps`, and `global`, with the `age` variable. This indicates that brain function generally worsens as one ages, which is supported by the positive correlation shown between `age` and `dementia_all`.

Additionally, I notice that `svd_amended_score` and `svd_simple_score` have positive correlations with the `study` and `study1` variables. This indicates some sort of significant difference between the results of the different studies included in our data. I explored these differences in the chart below:



Here, we see that certain studies included in the dataset yielded a higher proportion of positive dementia results than others. For example, the ASPS study had many patients who didn't have dementia, while the rundmc and scans studies had higher proportions of positive dementia patients. I am unsure of the cause of this discrepancy in the original studies, so I decided to not use `study` or `study1` as predictor variables in my recipe, as I find it irrelevant in predicting a patient's medical diagnosis.

Now that I've performed some initial data analysis, I feel ready to start building my model.

Model Setup

First, we want to organize our training and testing data through the use of stratified sampling and k-fold cross validation:

Sampling

For the split, I chose a proportion of 70% of our data to be in the training set, with 30% in the testing set. I stratified on the response variable, `dementia_all`, set up cross validation with 5 folds, and got the following split between training and testing data.

```
## [1] 1281
```

```
## [1] 550
```

Recipe Setup

Time to get cooking! I created a recipe for the different models that I will fit, which included 17 of the original 19 predictor variables, excluding `study` and `study1` due to reasons previously stated. Next, I dummy-coded all categorical variables, and then used imputation on all predictors to generate values for missing data. Also, I upsampled to increase the ratio between sample patients who have dementia and those who don't to 0.5. Finally, I centered and scaled the predictors so that they could be used in prediction.

Model Fitting

I fit four models to the cross-validated training data:

- Logistic Regression
- Elastic Net
- Random Forest
- Gradient Boosted Trees

To fit these models, I first specified the type of model and workflow for the model. I then fit the models to the cross-validated training data, tuning different parameters for Elastic Net, Random Forest, and Gradient Boosted Trees. Finally, I saved the models to external files so that I wouldn't have to fit them every time I ran the document.

Model Results

After fitting the models, I selected the four best models, and compared them to each other below. I selected one model from each type of model that I fit, those being Logistic Regression, Random Forest, Elastic Net, and Gradient Boosted Trees. Below are the roc_auc values for the four models, which lead me to choose the best performing model with the highest value.

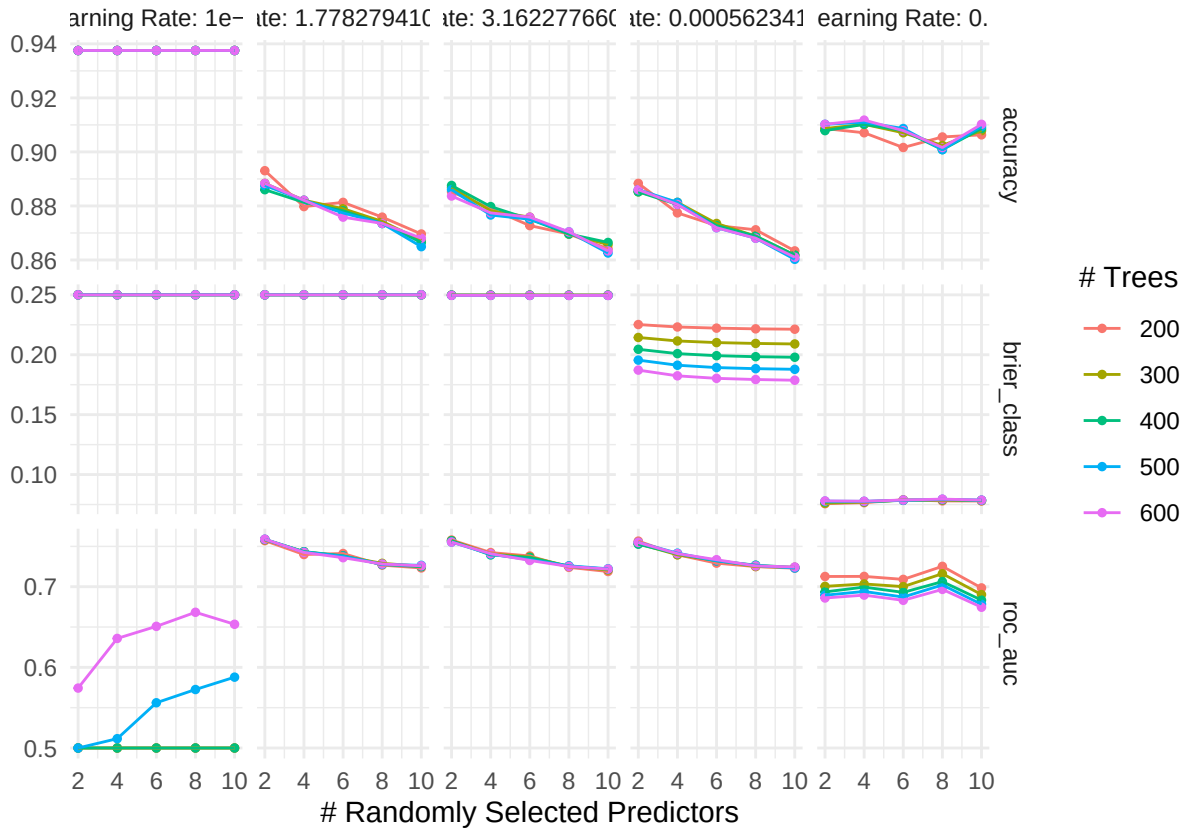
```
## # A tibble: 1 x 6
##   .metric .estimator  mean     n std_err .config
##   <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 roc_auc binary     0.787     5  0.0366 Preprocessor1_Model1
```

```
## # A tibble: 1 x 8
##   penalty mixture .metric .estimator  mean     n std_err .config
##     <dbl>   <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1   0.342   0.333 roc_auc binary     0.785     5  0.0219 Preprocessor1_Model032
```

```
## # A tibble: 1 x 9
##    mtry trees min_n .metric .estimator  mean     n std_err .config
##   <int> <int> <int> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1     2   300    20 roc_auc binary     0.743     5  0.0347 Preprocessor1_Model1~
```

```
## # A tibble: 1 x 9
##    mtry trees   learn_rate .metric .estimator  mean     n std_err .config
##   <int> <int>        <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1     2   600 0.0000000178 roc_auc binary     0.759     5  0.0350 Preprocessor1~
```

Surprisingly, the model with the highest roc_auc value was actually logistic regression! Typically, tree and forest models tend to perform better, but that didn't seem to be the case here. To make predictions on testing data, I will use this logistic regression model, along with the boosted tree that performed best, which had hyperparameter values of `mtry` = 2, `trees` = 600, and `learn_rate` = 1.778279e-08. Below I will show a plot of the boosted tree models fit that visualizes the optimization of these hyperparameters.



Predictions

First, we finalize our logistic regression and boosted tree models, and use them to make predictions. Here are predictions and metrics for the logistic regression model first:

```
## # A tibble: 6 x 4
##   dementia_all .pred_class .pred_0 .pred_1
##   <fct>        <fct>         <dbl>   <dbl>
## 1 0            0             0.757   0.243
## 2 1            1             0.326   0.674
## 3 0            0             0.684   0.316
## 4 0            0             0.824   0.176
## 5 1            1             0.164   0.836
## 6 0            0             0.776   0.224
```

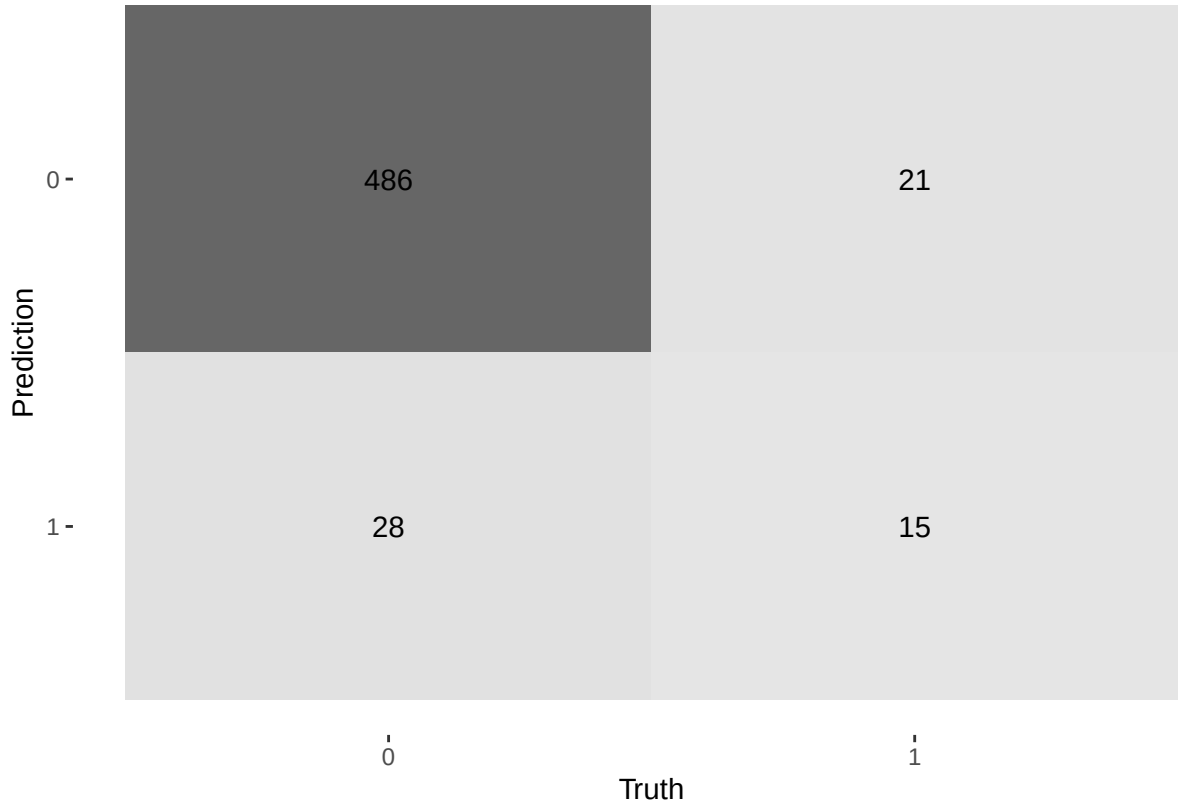
```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.832
```

Now our boosted tree model:

```
## # A tibble: 6 x 4
##   dementia_all .pred_class .pred_0 .pred_1
##   <fct>        <fct>         <dbl>   <dbl>
## 1 0            0             0.500   0.500
## 2 1            1             0.500   0.500
## 3 0            0             0.500   0.500
## 4 0            0             0.500   0.500
## 5 1            1             0.500   0.500
## 6 0            0             0.500   0.500
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.834
```

After testing our two best models, I found that the boosted tree had a slightly higher ROC AUC than the logistic regression model. Both of these models have high testing accuracy, with over 0.8 ROC AUC respectively. To measure error rate among the imbalanced categories of the categorical response variable, I also included a confusion matrix heat map below for our final boosted tree model:



As we can see from this graph, the model predicted a negative diagnosis very accurately, but struggled a bit more when a patient did have dementia. This is to be expected, as this type of issue is common with imbalanced responses, but my model still performed well, and made some correct predictions of patients having dementia.

Conclusion

Throughout this paper, we explored the initial dataset and its relevant features, created a recipe for a model, fit that recipe to four different model types, and used those models to make predictions of dementia data. We identified important correlations between variables, and their likelihood to be a dementia risk factor. Some of these risk factors can be age, high blood pressure and microbleeds in the brain. The bulk of the work that went into this paper was in the model fitting process, in which we fit a logistic regression model, an elastic net model, a random forest model and a gradient boosted trees model. Our logistic regression and gradient boosted trees models performed best, and upon further investigation, we found that the boosted trees model performed best on the testing set.

In the future, I hope to return to this project to refine my models, and perhaps fit more in search of more accurate results. Also, I want to examine the data more closely to investigate why certain models worked better than others. For example, I was surprised by how well my logistic regression model performed, and I think the answer to this curiosity can be found somewhere deeper in the original dataset. In the future, I

hope researchers continue to investigate dementia risk factors, as dementia-related illnesses are shrouded in mystery, with their direct causes being somewhat vague. I hope that with this new research may come new, interesting data that can help improve models like the ones I've showcased in this paper.