CMDA3654 Final Project

Brian Lee

Honor code:

"I have neither given nor received unauthorized assistance on this assignment." B.L.

I receive help from no one and give help to Cara Dunnavant.

Part I

Data description

The energy efficiency dataset contains data on the energy efficiency of 768 building shapes derived from 12 building shapes by simulation of various building characteristics such as orientation, glazing, and size. Energy efficiency is measured by the buildings' heating and cooling load requirements. This energy efficiency level is recorded with relevant data on each building, including two-dimensional and three-dimensional physical attributes.

This dataset is obtained from University of California Irvine's Machine Learning Repository. It was created by Angeliki Xifara and processed by Athanasios Tsanas.

```
## # A tibble: 6 x 10
##      X1    X2    X3    X4    X5    X6    X7    X8    Y1    Y2
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  0.98  514.  294   110.     7     2     0     0  15.6  21.3
## 2  0.98  514.  294   110.     7     3     0     0  15.6  21.3
## 3  0.98  514.  294   110.     7     4     0     0  15.6  21.3
## 4  0.98  514.  294   110.     7     5     0     0  15.6  21.3
## 5  0.9   564.  318.  122.     7     2     0     0  20.8  28.3
## 6  0.9   564.  318.  122.     7     3     0     0  21.5  25.4
```

In this situation, there are two response variables, the heating and cooling load on the buildings. There are eight explanatory variables. These explanatory variables include compactness of the building, building height, building surface area that of overall space, walls, and the roof. They also have the area and distribution of glazing on the building. Here is a preliminary look at the data frame.

Data visualization

In our exploratory data analysis, it is a good idea to visualize many variables of interest to assess the makeups of the data, both predictors and response, might have. This way, we gain insight to a variety of subjects of interest. By visualizing data, we can detect outliers on our predictors not for removal, but for awareness of which observations might be influencing the regression. We can also view the distributions of our data so we can generally see what makes a "typical" value for a predictor.

We can start by visualizing the value of every predictor against each predictor as well as response variables to scout out for multicollinearity or possible significant relationships between predictors and responses.

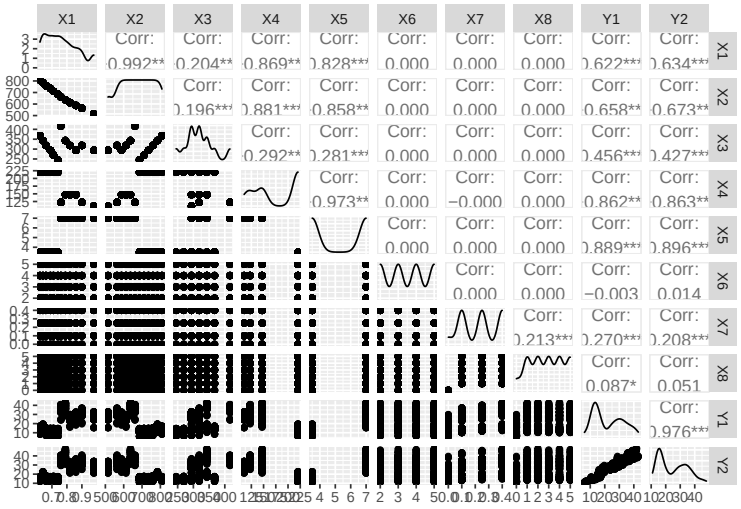

Figure 1: Matrix Plot of All Variables

Such a matrix plot can possibly be difficult to read. We can also visualize the distribution of variables of interest. We want to assess the general shape of the data to identify possible relationships. This will also help us detect outliers in our data.

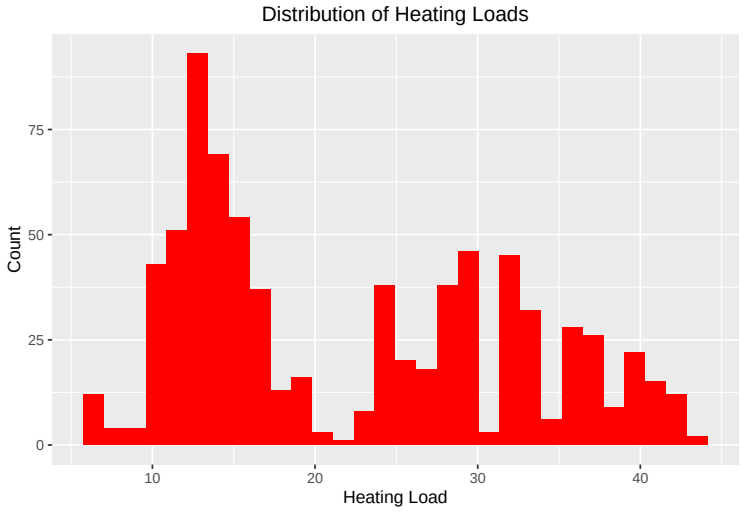

Figure 2: Various Predictors against other Predictors and Response Variables

We can comment on some insights gained from our data visualization.

- Examining both response variables, both the heating load and cooling load seem to favor lower values and seem far from evenly distributed.
- There is a higher mean response of heating load when building height, or glaze area increases, but there is a lower mean response of heating load when overall surface area increases. From this, we can reason that building size statistics does not necessarily correspond positively heating and cooling loads. In other words, one should not assume that a bigger building means higher heating or cooling

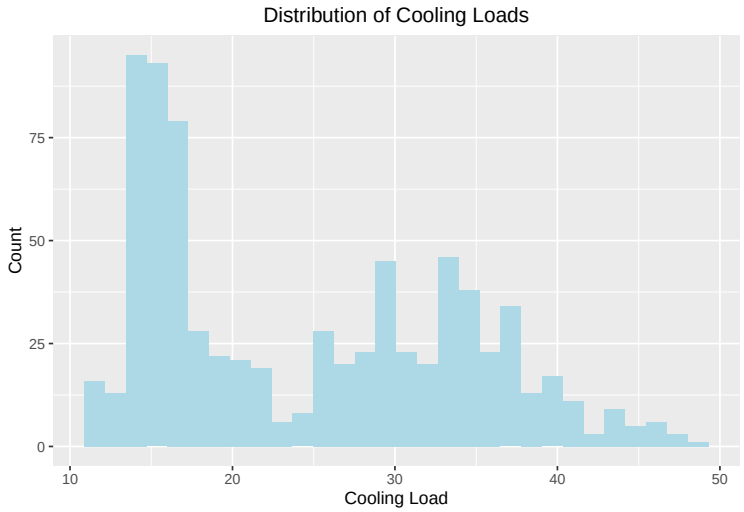

Figure 3: Various Predictors against other Predictors and Response Variables

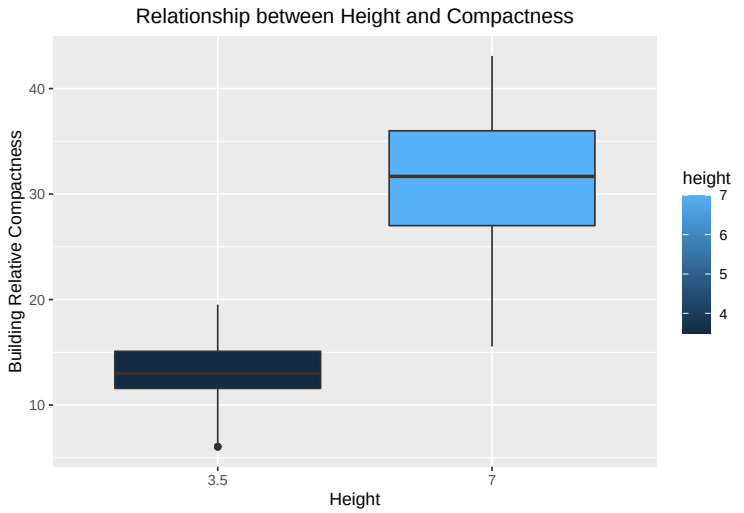

Figure 4: Various Predictors against other Predictors and Response Variables

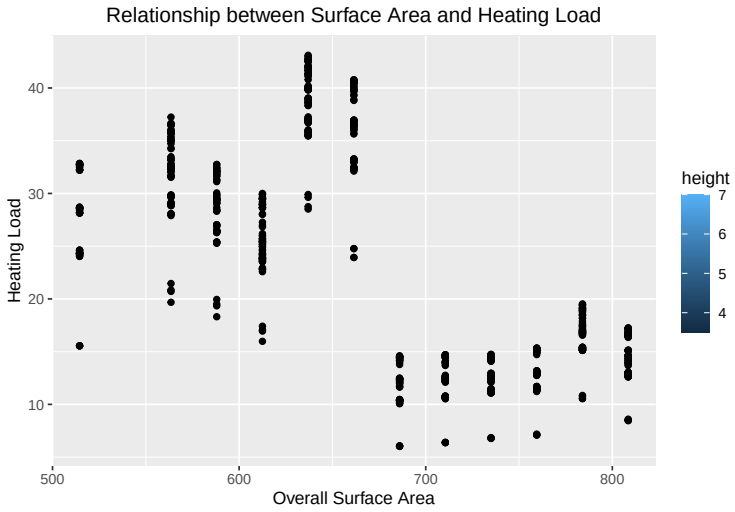

Figure 5: Various Predictors against other Predictors and Response Variables

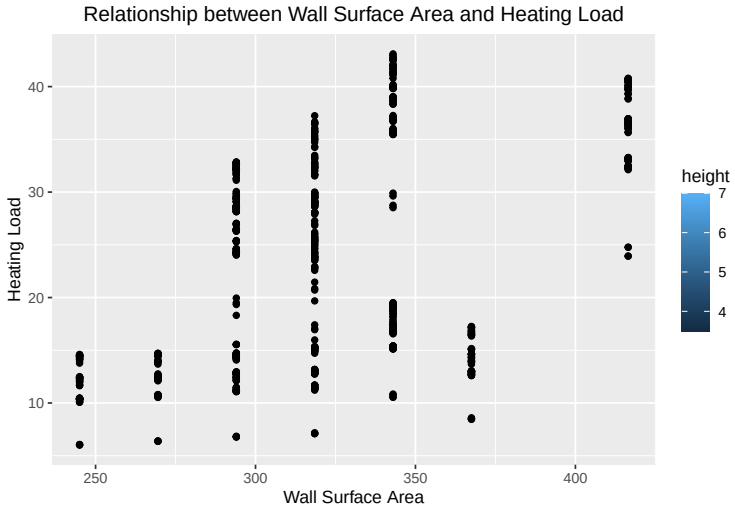

Figure 6: Various Predictors against other Predictors and Response Variables

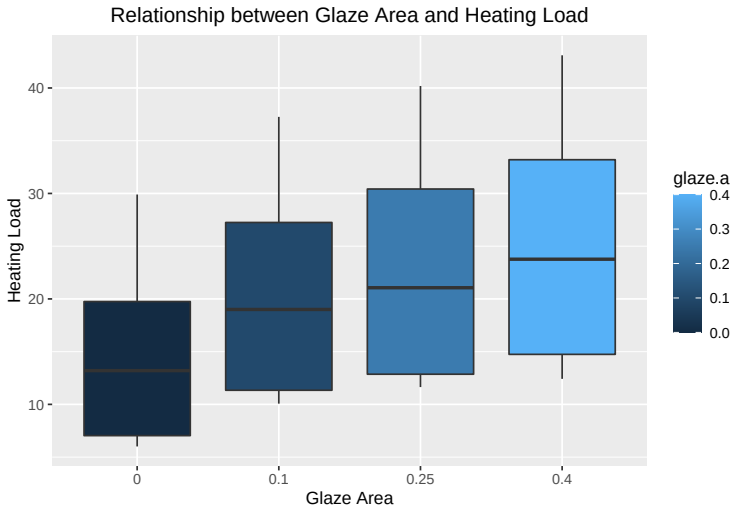

Figure 7: Various Predictors against other Predictors and Response Variables

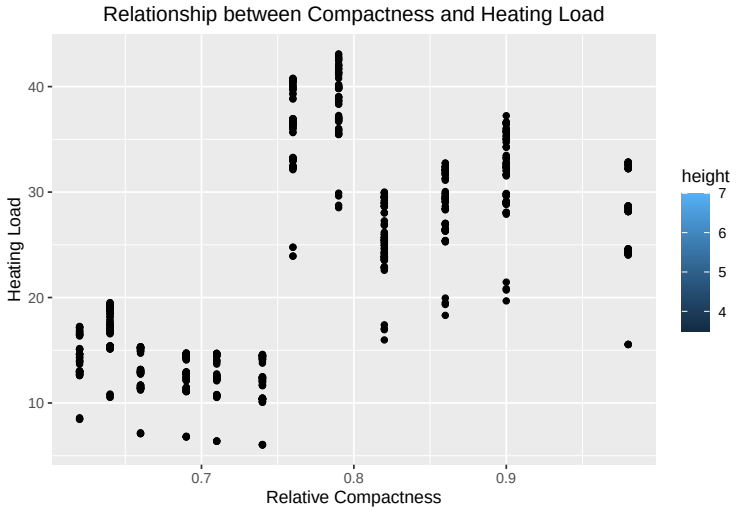

Figure 8: Various Predictors against other Predictors and Response Variables

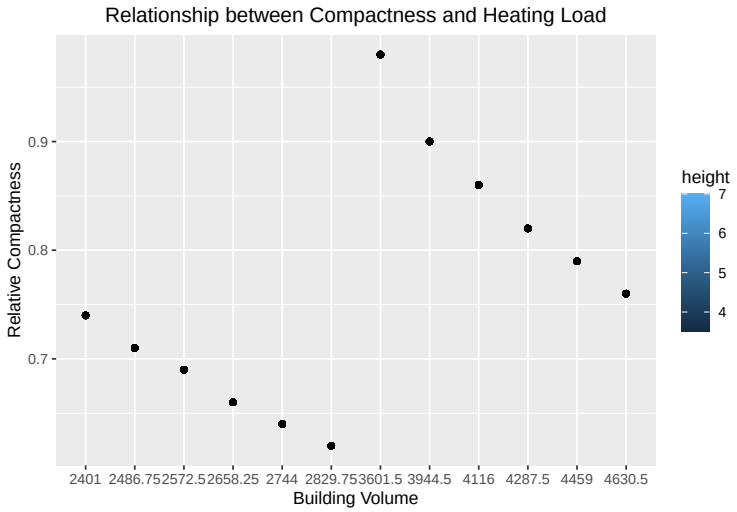

Figure 9: Various Predictors against other Predictors and Response Variables

load requirements. Perhaps heating and cooling load requirements, measures of energy efficiency, are dependent on physical efficiency as well.

- Compactness could possibly be an interesting variable to examine, as it takes into account both volume and overall size.

The goal, then is to construct models that can effectively predict the heating and cooling load scores required for each building. Given certain values for each of the building's potential explanatory variables, it should produce a value the resembles what a typical value for heating or cooling load should be. We can do this by constructing a multiple linear regression model. This is because area is a continuous response variable.

Because response variable cooling load and heating load both have similar relationships to the explanatory variables, only a multiple linear regression model on heating load will be constructed.

Model Setup

First, a linear regression model on heat load requirement is constructed by including all explanatory variables, X1 through X8.

```
##
## Call:
## lm(formula = Y1 ~ . - Y2, data = enb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8965 -1.3196 -0.0252  1.3532  7.7052
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.013418  19.033613   4.414 1.16e-05 ***
## X1          -64.773432  10.289448  -6.295 5.19e-10 ***
## X2           -0.087289   0.017075  -5.112 4.04e-07 ***
```

```
## X3            0.060813   0.006648   9.148  < 2e-16 ***
## X4                  NA         NA      NA       NA
## X5            4.169954   0.337990  12.338  < 2e-16 ***
## X6           -0.023330   0.094705  -0.246  0.80548
## X7           19.932736   0.813986  24.488  < 2e-16 ***
## X8            0.203777   0.069918   2.915  0.00367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 760 degrees of freedom
## Multiple R-squared:  0.9162, Adjusted R-squared:  0.9154
## F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

Next, we utilize Akaike's Information Criterion method (AIC), which will iterate until it a minimum AIC score is reached. In this way an optimal set of explanatory variables is obtained. These are very often statistically significant as well. As can be seen, the optimal set of predictors to predict heat load is relative compactness, surface area, wall area, height, glazing area, and distribution of glazing.

We can see if we can tune our model to be even more optimal with the addition of interaction terms or higher order terms as predictors. This can be done through trial and error.

Result

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2 + X3 + X5 + X7 + X8 + X3:X2 + X7:X8 +
##     X1:X2 + X1:X5, data = enb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2156 -1.2059  0.3352  1.6056  5.1556
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.913e+03  1.204e+02 -15.888  < 2e-16 ***
## X1           1.226e+03  8.933e+01  13.726  < 2e-16 ***
## X2           1.688e+00  1.179e-01  14.314  < 2e-16 ***
## X3           1.045e+00  7.498e-02  13.938  < 2e-16 ***
## X5           6.680e+01  4.602e+00  14.514  < 2e-16 ***
## X7           2.706e+01  1.192e+00  22.695  < 2e-16 ***
## X8           7.883e-01  9.927e-02   7.940 7.28e-15 ***
## X2:X3       -1.672e-03  1.236e-04 -13.521  < 2e-16 ***
## X7:X8       -2.860e+00  3.944e-01  -7.251 1.02e-12 ***
## X1:X2       -3.410e-01  8.272e-02  -4.122 4.17e-05 ***
## X1:X5       -7.072e+01  5.425e+00 -13.037  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.432 on 757 degrees of freedom
## Multiple R-squared:  0.9427, Adjusted R-squared:  0.9419
## F-statistic:  1245 on 10 and 757 DF,  p-value: < 2.2e-16
```

As a result, our regression equation is given by the following:

$$\hat{y} = -5.622 + 9.553X1 + 0.3756X2 + 0.4836X3 + 7.041X5 + 2.706X7 +$$

$$0.7883X8 - 0.0006714X2X3 - 2.86X7X8 + 0.4082X1X2 - 7.072X1X5 \quad (1)$$

It should be explained as to why certain terms were included or not included after running the Information Criteria. As for the interaction terms, these were included on the basis of how they would interact with each other in a physical interpretation. For example, perhaps a better model might be found by including an interaction to represent the real-life interaction between the compactness of a building and its surface area.

- An interaction between surface area and wall area was included because it makes sense that the overall surface area changes when one surface changes in its area.
- An interaction term between glazing area and glazing area distribution was included because how glaze is distributed must be accounted for, not just its total area.
- An interaction term between compactness and surface area is included because a more compact building would have less surface area.
- An interaction term between compactness and height is included because a building that is very tall is less compact.

It was decided not to include any higher-order and nonlinear terms in the data because upon viewing the matrix plot where each predictor is plotted against the response variable, heating load, it is not at all clear if any one predictor acts in a nonlinear fashion against the response variable.

This model is quite satisfactory because not only is every predictor's coefficient statistically significant, we have an even higher R^2 and R^2adj value of 94.27% and 94.19%, respectively. In other words, around 94.27% of the variability in the data is explained by the model created.

Part II

Data description

The MNIST dataset is a database of 60,000 images of handwritten digits from 0 to 9, derived from the NIST Special Database. To clarify, each image exists with a size of 28x28 pixels, and each pixel takes on a value between 0.0 to 1.0 as the value on the grayscale.

The MNIST database of handwritten digits was constructed by Yann LeCun, Corinna Cortes, and Christopher J.C. Burges.

The high-dimensionality of the MNIST dataset can be visualized by constructing a t-SNE plot that gives an intuitive idea as to how the data is arranged. t-SNE reduces this high dimensionality to produce this plot.

```
## Performing PCA
## Read the 60000 x 50 data matrix successfully!
## OpenMP is working. 1 threads.
## Using no_dims = 2, perplexity = 30.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
##  - point 10000 of 60000
##  - point 20000 of 60000
##  - point 30000 of 60000
##  - point 40000 of 60000
##  - point 50000 of 60000
```

```
##  - point 60000 of 60000
## Done in 501.27 seconds (sparsity = 0.002086)!
## Learning embedding...
## Iteration 50: error is 118.896811 (50 iterations in 18.55 seconds)
## Iteration 100: error is 118.896811 (50 iterations in 19.85 seconds)
## Iteration 150: error is 118.833437 (50 iterations in 28.87 seconds)
## Iteration 200: error is 108.762441 (50 iterations in 24.39 seconds)
## Fitting performed in 91.67 seconds.
```

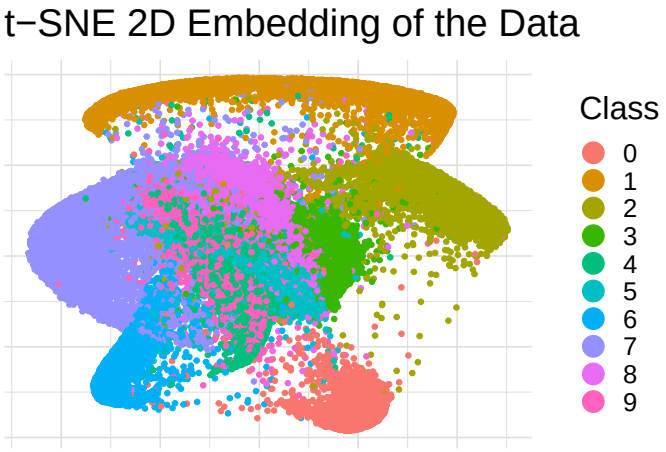

Figure 10: t-SNE Plot of Data Embedding, 2D

In this dataset, the response variable would be a categorical variable of digits 0 through 9. In other words, these would be what the guesses and conclusions our models arrive at on which digit is being resembled. The predictors are the grayscale values of each pixel in the 784 pixel image. This is because the values of the pixels determine how the model categorizes the input image.

The goal with this dataset, then, is to build a model that can accurately classify what digit has been handwritten in the image. With the understanding that individuals have variability in their handwriting, it would be helpful to create a model that can accommodate those differences when determining what digit is handwritten in the image. If we are able to construct such a model that can classify which digit is which, we have tremendous utility in real-life settings that may require a checking of handwriting, such as auto-grading for academia or text parsing.

Dimension Reduction and LDA Setup

Because the problem presented by the MNIST dataset is one of classification, one classification method that can be constructed is through dimension reduction and subsequent linear discrimination analysis. First, the training and testing data for the predictor must be reshaped into a matrix and scaled by 255 in order that the values in said matrix are between 0 and 1.

After reshaping and rescaling the data, we perform principle component analysis to reduce the dimensional space. This is done using the training dataset. In addition, through making a scree plot, the number of principle components is determined.

Seeing as that there are more than a binomial outcome space for our classification, linear discriminant analysis, or LDA, should be used as a classifier after the dimension reduction. And looking at the screeplot,

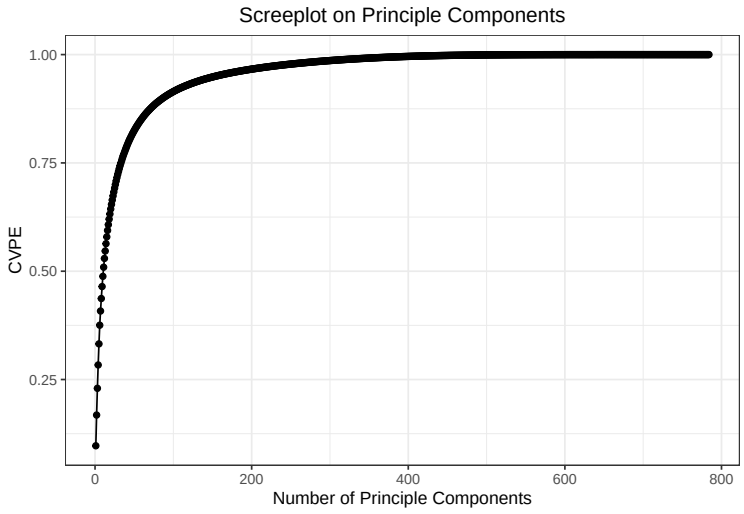

Figure 11: PCA Screeplot

the leveling of cumulative proportion of variance explained occurs around 150 to 175 principle components, so the number of principle compenents to use should exist somewhere within that range.

New predictors must be obtained in order to construct a new classifier. The scores from a table of principle component analysis scores after performing the dimension reduction will be used. They are stored in a new training dataset and serve as the new predictors for the LDA classifier, which is then fit to said training data.

After constructing a classification method through applying LDA to the training dataset of PCA scores, we run it against the testing data to make predictions and also see how satisfactory the classifer is.

Dimension Reduction and QDA Result

A receiver operating characteristic curve, or ROC curve, can assess the effectiveness of the classifier. It is the measure of the true positive versus the false positive rate. In this case, the ROC curve is constructed on the linear discriminant classifier.

As can be seen, the area under the ROC curve is 0.9997, suggesting that the linear discriminant analysis applied to the dimension reduction is very effective.

Shrinkage Regression Setup

For the MNIST dataset, it would also make sense to run a shrinkage regression. This shrinkage regression can perform

In this case, the LASSO shrinkage regression will be used because it can perform variable selection and regularization to arrive at enhanced predictions. The shrinkage regressions are then plotted.

We are also to choose a tuning parameter through cross-validation. Selecting a good tuning parameter is essential to fit a model that can classify the data.

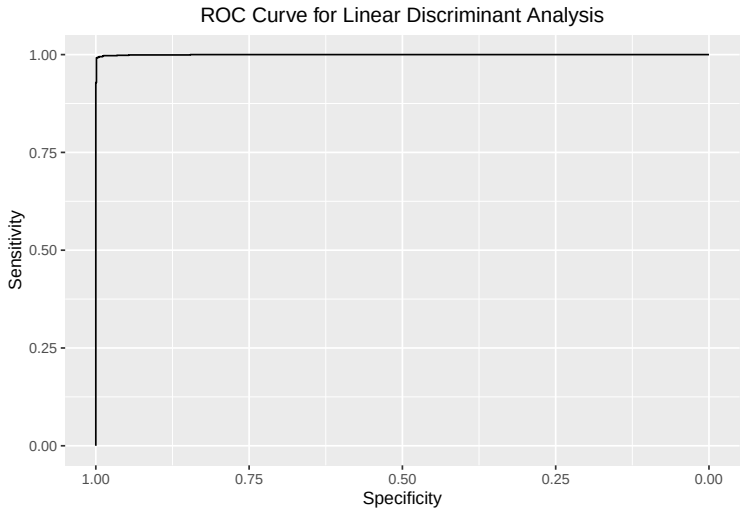

Figure 12: ROC Curve on LDA

Shrinkage Regression Result

A receiver operating characteristic curve, or ROC curve, can assess the effectiveness of the classifier. It is the measure of the true positive versus the false positive rate. In this case, the ROC curve is constructed on the LASSO shrinkage regression model we constructed.

As can be seen, the area under the ROC curve is 0.8850, suggesting that the LASSO shrinkage regression applied to the data is quite effective.

Convolutional Neural Network Setup

The MNIST dataset contains training and testing datasets for the predictor and categorical response variables. In the case of images, it makes the most sense to use convolutional neural network which view the 28x28 pixel as a convolutional layer on which we apply filters to. Therefore, the training and testing data must be reshaped to be 28x28 layers.

For the MNIST dataset, it would also make sense to construct a neural network. The neural network should be able to take in input neurons and iterate through various hidden layers to arrive at a conclusion of what digit is presented by the MNIST data. Here, each neuron in the input layer corresponds to a pixel in the 28x28 image. This means that we would have 784 neurons in the input layer, structured in a square layer of 28x28 neurons.

```
## Model: "sequential"
## _________________________________________________________________________
## Layer (type)                 Output Shape              Param #
## =========================================================================
## conv2d_1 (Conv2D)            (None, 27, 27, 32)        160
## _________________________________________________________________________
## max_pooling2d_1 (MaxPooling2D)  (None, 13, 13, 32)     0
## _________________________________________________________________________
## conv2d (Conv2D)              (None, 12, 12, 64)        8256
## _________________________________________________________________________
## dropout_2 (Dropout)          (None, 12, 12, 64)        0
## _________________________________________________________________________
```

```
## max_pooling2d (MaxPooling2D)        (None, 6, 6, 64)              0
## _________________________________________________________________________
## dropout_1 (Dropout)                 (None, 6, 6, 64)              0
## _________________________________________________________________________
## flatten (Flatten)                   (None, 2304)                  0
## _________________________________________________________________________
## dense_1 (Dense)                     (None, 98)                    225890
## _________________________________________________________________________
## dropout (Dropout)                   (None, 98)                    0
## _________________________________________________________________________
## dense (Dense)                       (None, 10)                    990
## =========================================================================
## Total params: 235,296
## Trainable params: 235,296
## Non-trainable params: 0
## _________________________________________________________________________
```

The neural network is then compiled and trained using the training datasets on predictor and response variables in order to assign fair weights and bias values to each hidden layer.

Neural Network Result

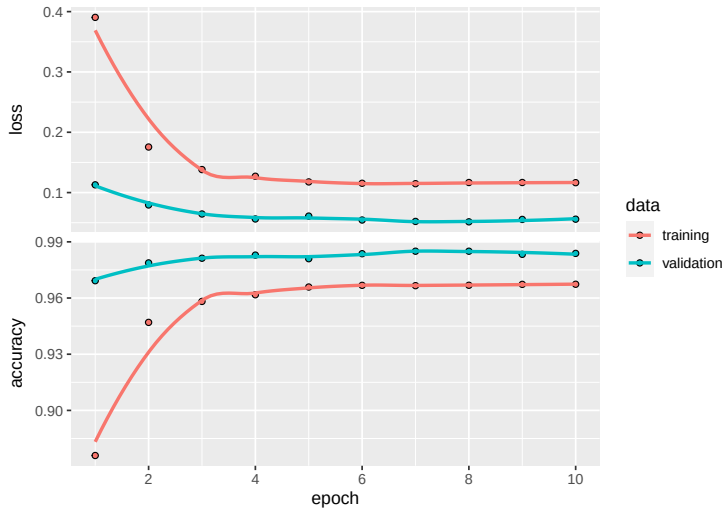

Figure 13: Accuracy and Loss of Convolutional Neural Network

```
## Test loss: 0.04952027
```

```
## Test accuracy: 0.9845
```

We can see that the accuracy is around 98%, and the loss is around 4%. Although this is most likely a solid model that can adequately classify the images in the dataset, this raises questions of over-fitting. Such questions can be explored by fine-tuning the network.