IST 707 – HOMEWORK 6
HANDWRITING RECOGNITION

1 Introduction 2

2 Analysis & Models 3

2.1 About the Data 3

2.2 Data Retrieval 3

2.3 Models 5

3 Results 6

4 Conclusions 12

1 Introduction

The U.S. Constitution mandates that a census be taken every 10 years to count all people both citizens and noncitizens living in the United States. Responding to the census is mandatory because getting a complete and accurate count of the population is critically important. An accurate count of the population serves as the basis for fair political representation and plays a vital role in many areas of public life.

The census tells us who we are and where we are going as a nation, and helps our communities determine where to build everything from schools to supermarkets, and from homes to hospitals. It helps the government decide how to distribute funds and assistance to states and localities. It is also used to draw the lines of legislative districts and reapportion the seats each State holds in Congress. (Our Censuses, n.d.)

The biggest part of census is collecting the census data and counting the responses, specifically documenting the handwritten census surveys. The survey processing is a humongous task which bring an opportunity for computers to read and recognize the human text. We will be exploring statistical methods which can solve

2 Analysis & Models

2.1 About the Data

We will be using an image data base called MNIST. (MNIST DATABASE, n.d.) The MNIST database was constructed from National Institute of Standards and Technology’s Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore, it was necessary to build a new database by mixing NIST's datasets.

The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels and translating the image so as to position this point at the center of the 28x28 field.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

2.2 Data Processing

The training data set has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. We will remove the “pixel” from the column name, so it will be easy to plot them.

Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as x = i * 28 + j, where i and j are integers between 0 and 27, inclusive. Then pixelx is located on row i and column j of a 28 x 28 matrix, (indexing by zero). We will convert the long form pixel to 28x28 format to measure the distance from center.

2.3 Exploratory Data Analysis

We will first check how many samples we have each class of digit. Figure 1 show the class distribution, where all of the digits have 4000 samples on average each. We have a good classification set on our hand, where each class is well represented, and no one class is underrepresented.

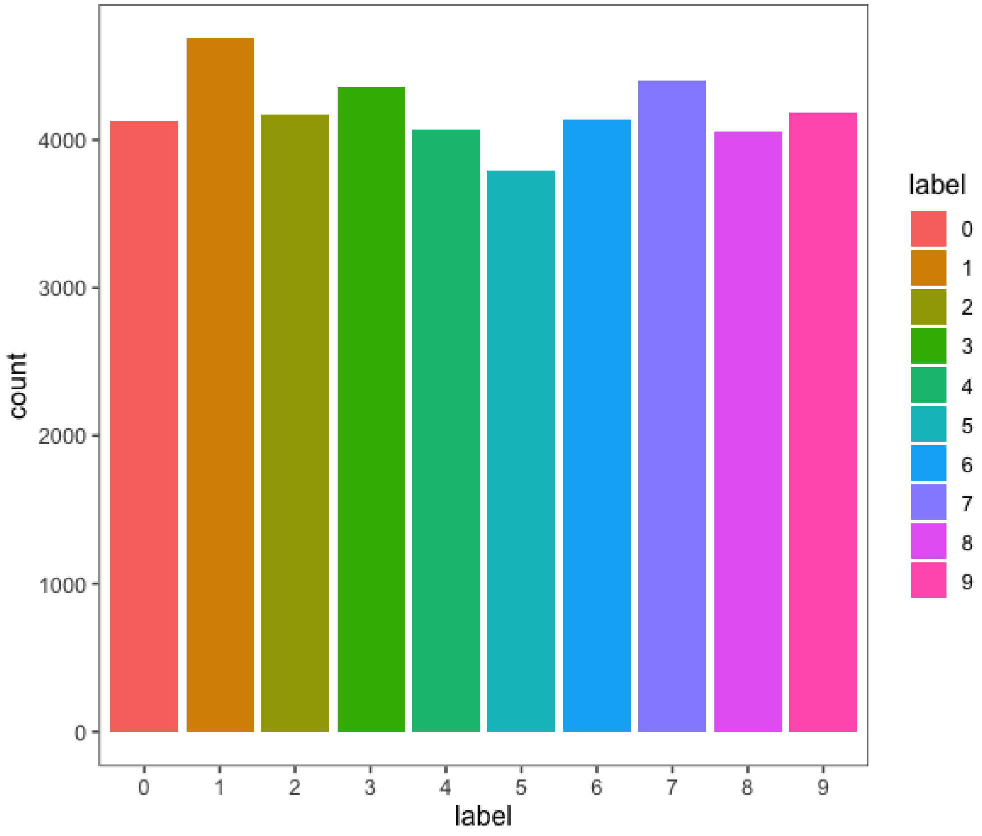


Figure 1

Next, we will look at the pixels itself, we will group by the color of the pixel. Figure 2 show the frequency of the colors. Zero indicates blank cell and 255 indicates black, and in between numbers show the intensity of the grey. More than 80 percent of the cells are blank and only 20% of the cells are occupied. That gives us roughly 150 pixels to work with on average.

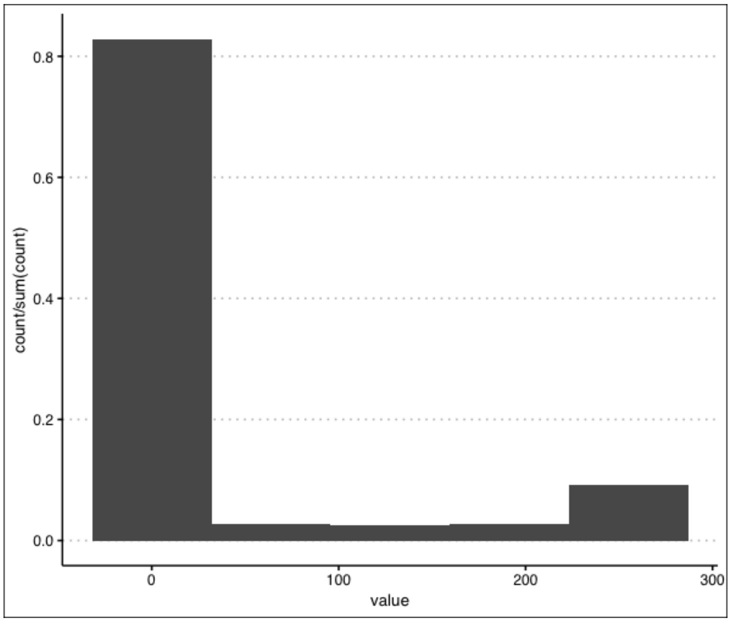


FIGURE 2

Next, we will check the variance in each digit. We will compute the Euclidean distance for each class and look at the variance. Except digit one, all of them has good variance, indicating a diverse writing styles for each digit. And all of them has lot of outliers, so we can't drop the outliers for our analysis. And four and nine are going to be problematic, they have too many outliers.

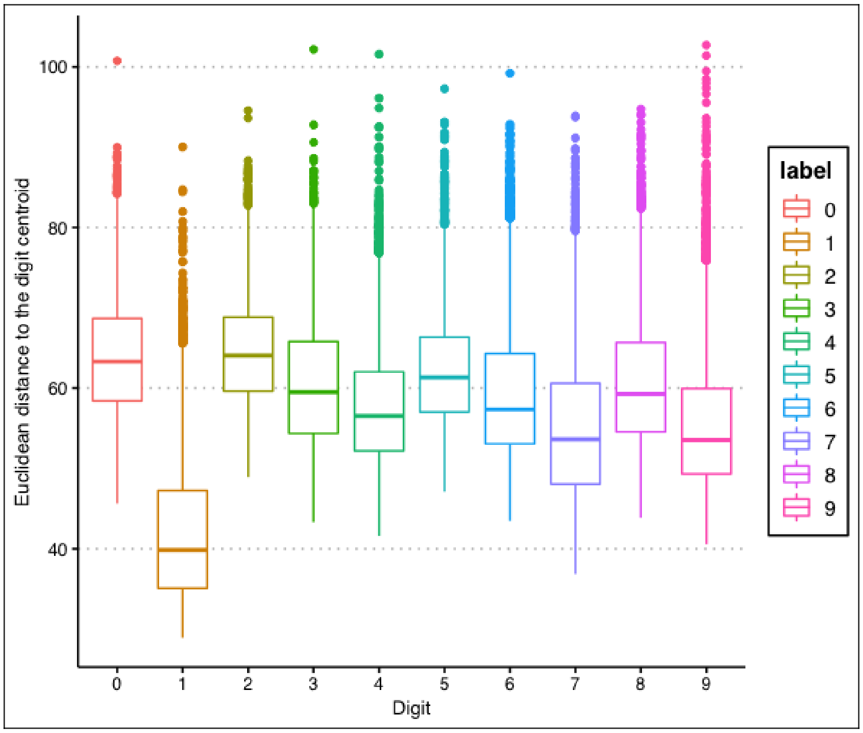
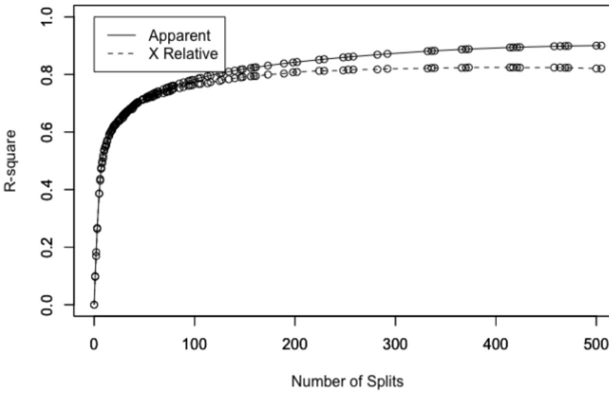


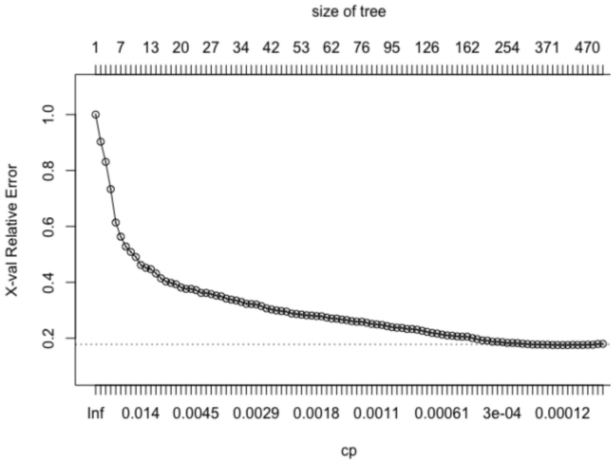
FIGURE 3

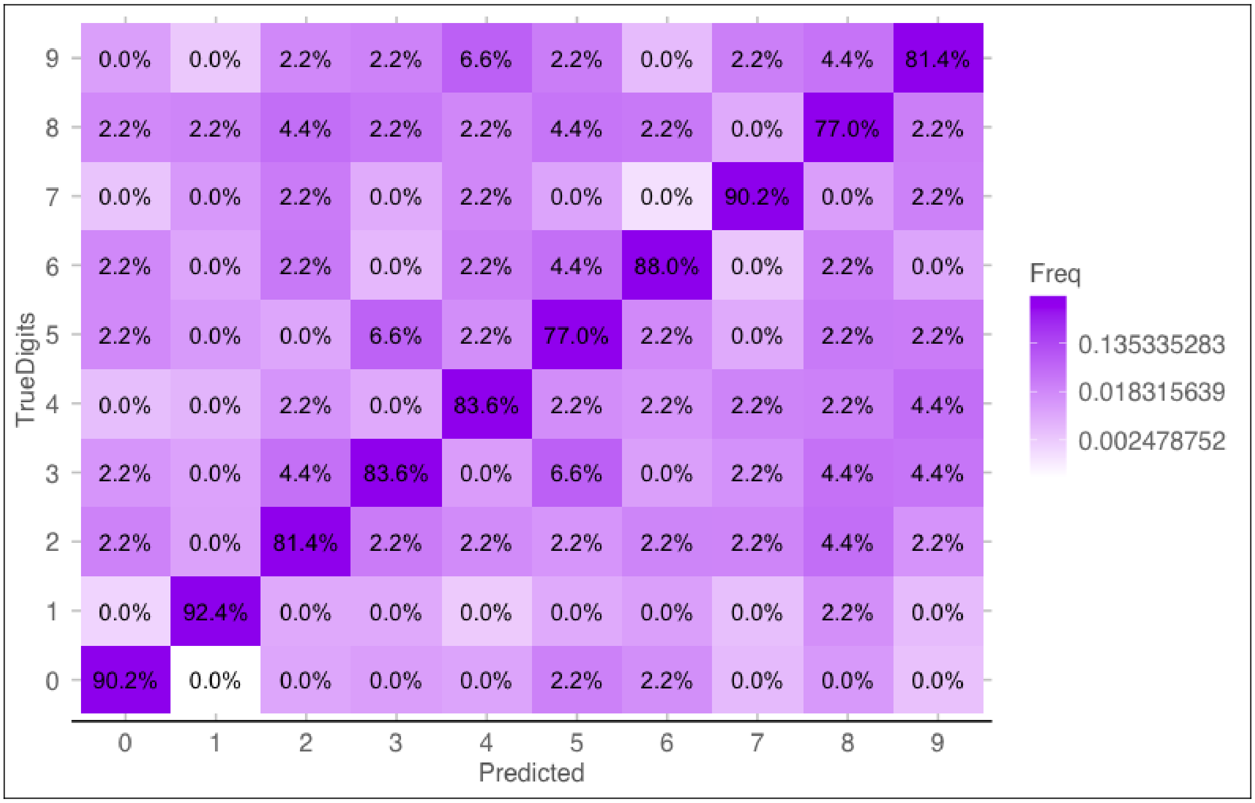
2.4 Models

3 Results

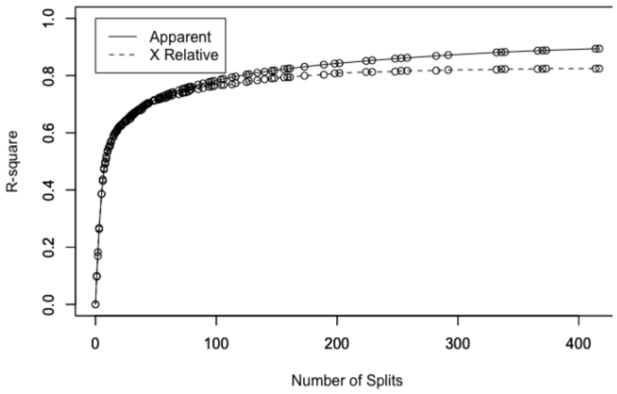
3.1.1 Decision Tree Models

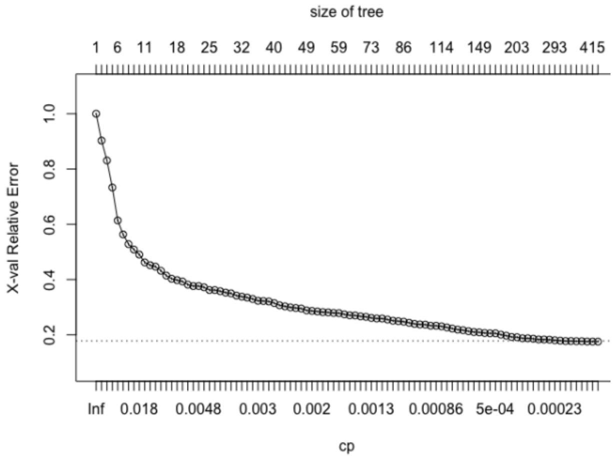


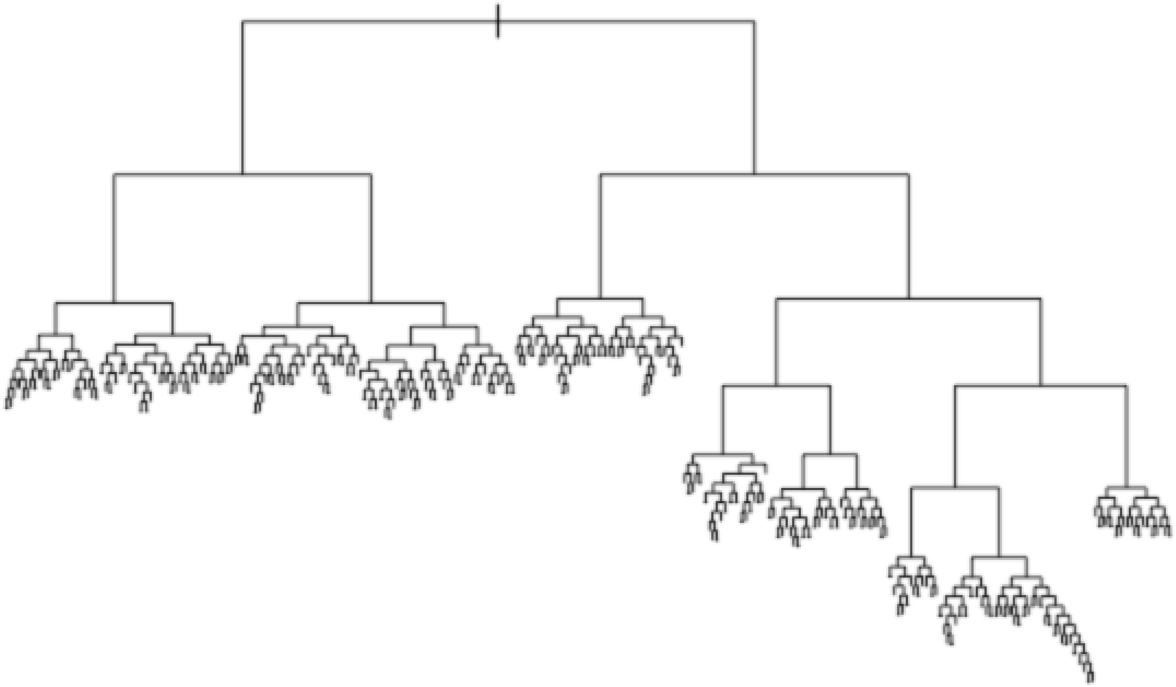


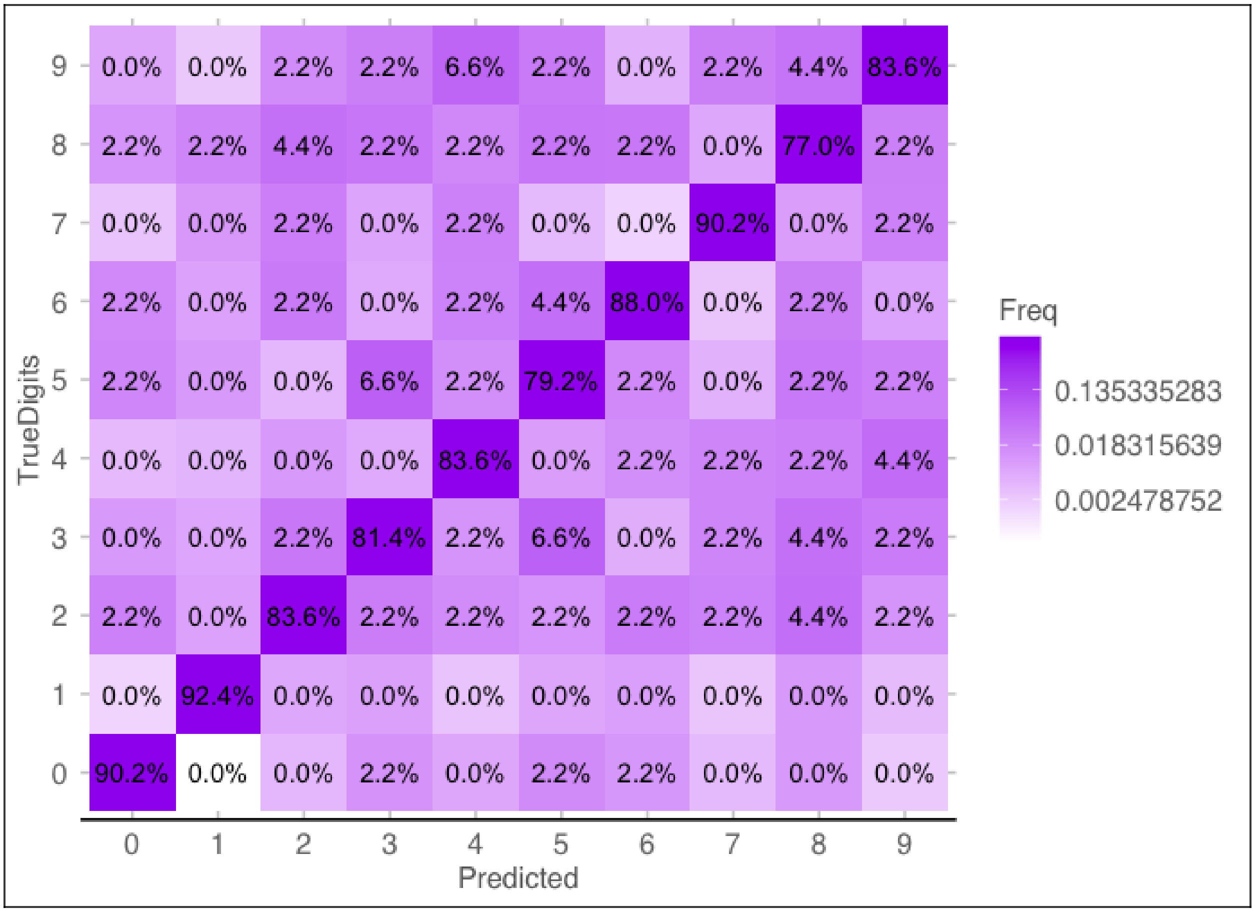


	0	1	2	3	4	5	6	7	8	9
9	0.0%	0.0%	2.2%	2.2%	6.6%	2.2%	0.0%	2.2%	4.4%	81.4%
8	2.2%	2.2%	4.4%	2.2%	2.2%	4.4%	2.2%	0.0%	77.0%	2.2%
7	0.0%	0.0%	2.2%	0.0%	2.2%	0.0%	0.0%	90.2%	0.0%	2.2%
6	2.2%	0.0%	2.2%	0.0%	2.2%	4.4%	88.0%	0.0%	2.2%	0.0%
5	2.2%	0.0%	0.0%	6.6%	2.2%	77.0%	2.2%	0.0%	2.2%	2.2%
4	0.0%	0.0%	2.2%	0.0%	83.6%	2.2%	2.2%	2.2%	2.2%	4.4%
3	2.2%	0.0%	4.4%	83.6%	0.0%	6.6%	0.0%	2.2%	4.4%	4.4%
2	2.2%	0.0%	81.4%	2.2%	2.2%	2.2%	2.2%	2.2%	4.4%	2.2%
1	0.0%	92.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.2%	0.0%
0	90.2%	0.0%	0.0%	0.0%	0.0%	2.2%	2.2%	0.0%	0.0%	0.0%



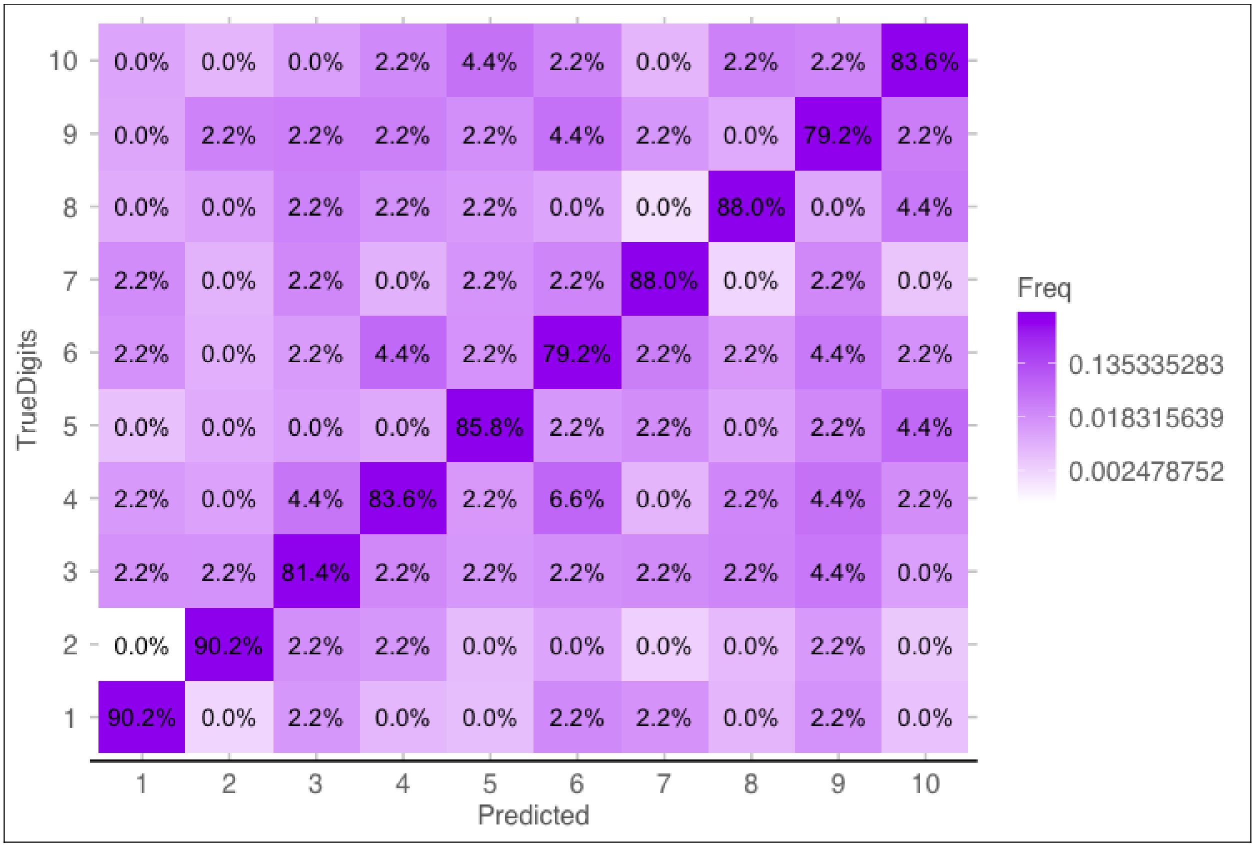






TrueDigits	0	1	2	3	4	5	6	7	8	9
9	0.0%	0.0%	2.2%	2.2%	6.6%	2.2%	0.0%	2.2%	4.4%	83.6%
8	2.2%	2.2%	4.4%	2.2%	2.2%	2.2%	2.2%	0.0%	77.0%	2.2%
7	0.0%	0.0%	2.2%	0.0%	2.2%	0.0%	0.0%	90.2%	0.0%	2.2%
6	2.2%	0.0%	2.2%	0.0%	2.2%	4.4%	88.0%	0.0%	2.2%	0.0%
5	2.2%	0.0%	0.0%	6.6%	2.2%	79.2%	2.2%	0.0%	2.2%	2.2%
4	0.0%	0.0%	0.0%	0.0%	83.6%	0.0%	2.2%	2.2%	2.2%	4.4%
3	0.0%	0.0%	2.2%	81.4%	2.2%	6.6%	0.0%	2.2%	4.4%	2.2%
2	2.2%	0.0%	83.6%	2.2%	2.2%	2.2%	2.2%	2.2%	4.4%	2.2%
1	0.0%	92.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0	90.2%	0.0%	0.0%	2.2%	0.0%	2.2%	2.2%	0.0%	0.0%	0.0%

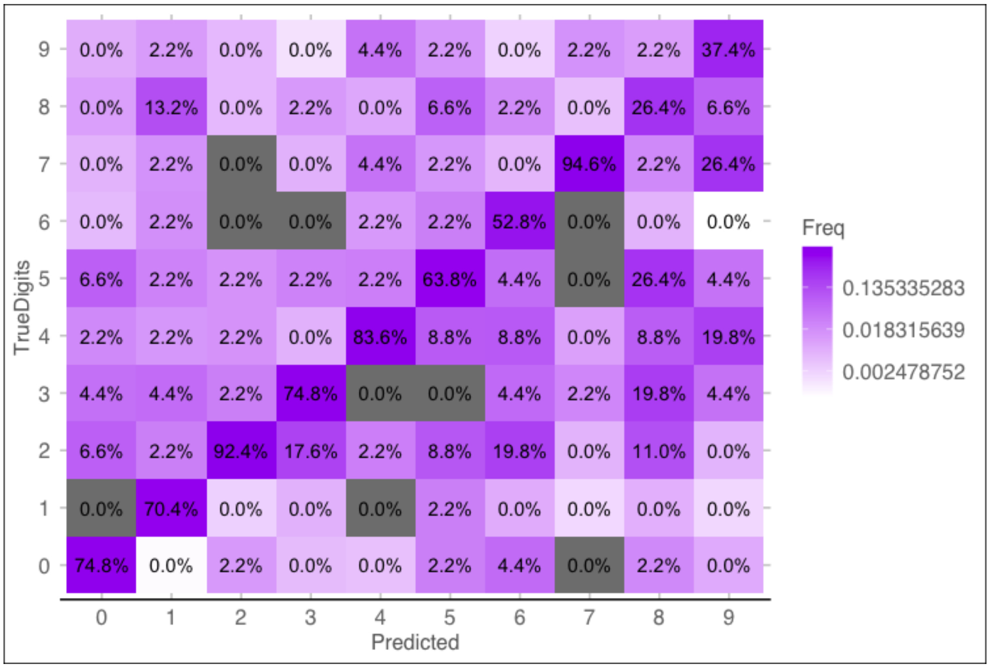
Overall Accuracy : 85.09%



TrueDigits	1	2	3	4	5	6	7	8	9	10
10	0.0%	0.0%	0.0%	2.2%	4.4%	2.2%	0.0%	2.2%	2.2%	83.6%
9	0.0%	2.2%	2.2%	2.2%	2.2%	4.4%	2.2%	0.0%	79.2%	2.2%
8	0.0%	0.0%	2.2%	2.2%	2.2%	0.0%	0.0%	88.0%	0.0%	4.4%
7	2.2%	0.0%	2.2%	0.0%	2.2%	2.2%	88.0%	0.0%	2.2%	0.0%
6	2.2%	0.0%	2.2%	4.4%	2.2%	79.2%	2.2%	2.2%	4.4%	2.2%
5	0.0%	0.0%	0.0%	0.0%	85.8%	2.2%	2.2%	0.0%	2.2%	4.4%
4	2.2%	0.0%	4.4%	83.6%	2.2%	6.6%	0.0%	2.2%	4.4%	2.2%
3	2.2%	2.2%	81.4%	2.2%	2.2%	2.2%	2.2%	2.2%	4.4%	0.0%
2	0.0%	90.2%	2.2%	2.2%	0.0%	0.0%	0.0%	0.0%	2.2%	0.0%
1	90.2%	0.0%	2.2%	0.0%	0.0%	2.2%	2.2%	0.0%	2.2%	0.0%

3.1.2 Naïve Bayes Models

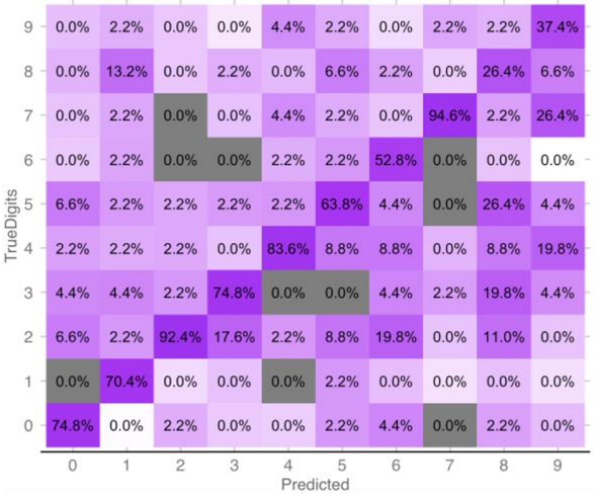
All 784 Features, Accuracy: 52.26%



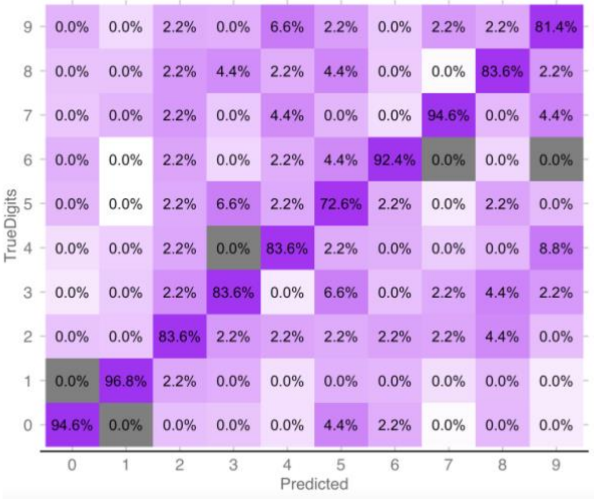
	0	1	2	3	4	5	6	7	8	9
9	0.0%	2.2%	0.0%	0.0%	4.4%	2.2%	0.0%	2.2%	2.2%	37.4%
8	0.0%	13.2%	0.0%	2.2%	0.0%	6.6%	2.2%	0.0%	26.4%	6.6%
7	0.0%	2.2%	0.0%	0.0%	4.4%	2.2%	0.0%	94.6%	2.2%	26.4%
6	0.0%	2.2%	0.0%	0.0%	2.2%	2.2%	52.8%	0.0%	0.0%	0.0%
5	6.6%	2.2%	2.2%	2.2%	2.2%	63.8%	4.4%	0.0%	26.4%	4.4%
4	2.2%	2.2%	2.2%	0.0%	83.6%	8.8%	8.8%	0.0%	8.8%	19.8%
3	4.4%	4.4%	2.2%	74.8%	0.0%	0.0%	4.4%	2.2%	19.8%	4.4%
2	6.6%	2.2%	92.4%	17.6%	2.2%	8.8%	19.8%	0.0%	11.0%	0.0%
1	0.0%	70.4%	0.0%	0.0%	0.0%	2.2%	0.0%	0.0%	0.0%	0.0%
0	74.8%	0.0%	2.2%	0.0%	0.0%	2.2%	4.4%	0.0%	2.2%	0.0%

Top 10 Features. Accuracy: 76.15

Top 50 Features. Accuracy: 86.89%



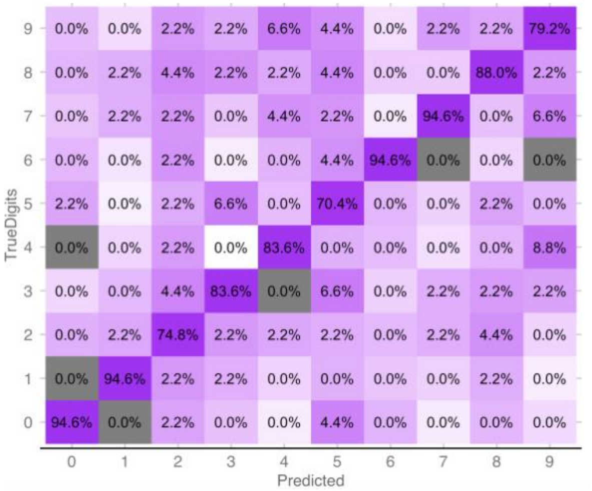
	0	1	2	3	4	5	6	7	8	9
9	0.0%	2.2%	0.0%	0.0%	4.4%	2.2%	0.0%	2.2%	2.2%	37.4%
8	0.0%	13.2%	0.0%	2.2%	0.0%	6.6%	2.2%	0.0%	26.4%	6.6%
7	0.0%	2.2%	0.0%	0.0%	4.4%	2.2%	0.0%	94.6%	2.2%	26.4%
6	0.0%	2.2%	0.0%	0.0%	2.2%	2.2%	52.8%	0.0%	0.0%	0.0%
5	6.6%	2.2%	2.2%	2.2%	2.2%	63.8%	4.4%	0.0%	26.4%	4.4%
4	2.2%	2.2%	2.2%	0.0%	83.6%	8.8%	8.8%	0.0%	8.8%	19.8%
3	4.4%	4.4%	2.2%	74.8%	0.0%	0.0%	4.4%	2.2%	19.8%	4.4%
2	6.6%	2.2%	92.4%	17.6%	2.2%	8.8%	19.8%	0.0%	11.0%	0.0%
1	0.0%	70.4%	0.0%	0.0%	0.0%	2.2%	0.0%	0.0%	0.0%	0.0%
0	74.8%	0.0%	2.2%	0.0%	0.0%	2.2%	4.4%	0.0%	2.2%	0.0%

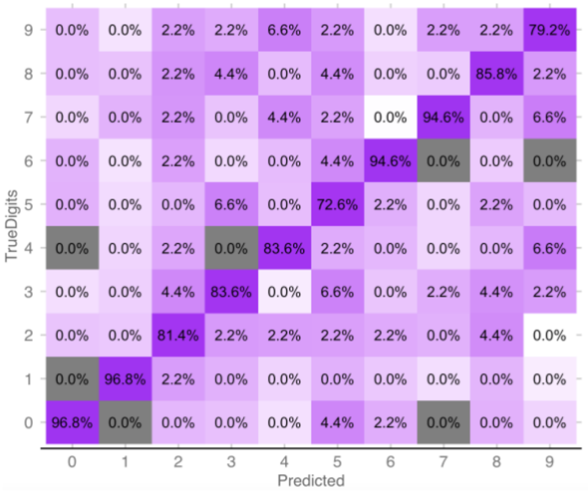


	0	1	2	3	4	5	6	7	8	9
9	0.0%	0.0%	2.2%	0.0%	6.6%	2.2%	0.0%	2.2%	2.2%	81.4%
8	0.0%	0.0%	2.2%	4.4%	2.2%	4.4%	0.0%	0.0%	83.6%	2.2%
7	0.0%	0.0%	2.2%	0.0%	4.4%	0.0%	0.0%	94.6%	0.0%	4.4%
6	0.0%	0.0%	2.2%	0.0%	2.2%	4.4%	92.4%	0.0%	0.0%	0.0%
5	0.0%	0.0%	2.2%	6.6%	2.2%	72.6%	2.2%	0.0%	2.2%	0.0%
4	0.0%	0.0%	2.2%	0.0%	83.6%	2.2%	0.0%	0.0%	0.0%	8.8%
3	0.0%	0.0%	2.2%	83.6%	0.0%	6.6%	0.0%	2.2%	4.4%	2.2%
2	0.0%	0.0%	83.6%	2.2%	2.2%	2.2%	2.2%	2.2%	4.4%	0.0%
1	0.0%	96.8%	2.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0	94.6%	0.0%	0.0%	0.0%	0.0%	4.4%	2.2%	0.0%	0.0%	0.0%

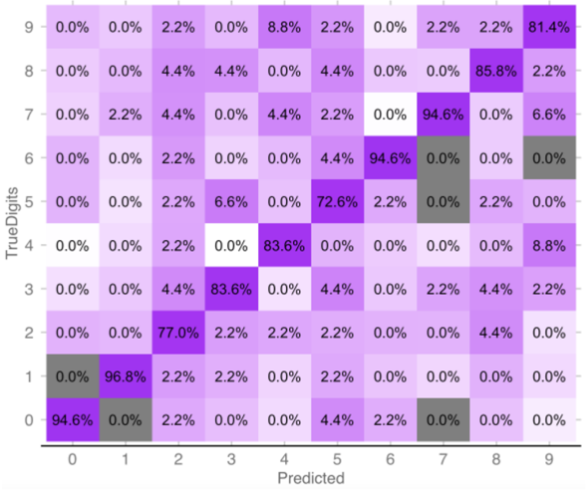
Top 100 Features, Accuracy: 85.16%

Top 75 Features ,Accuracy: 86.4%

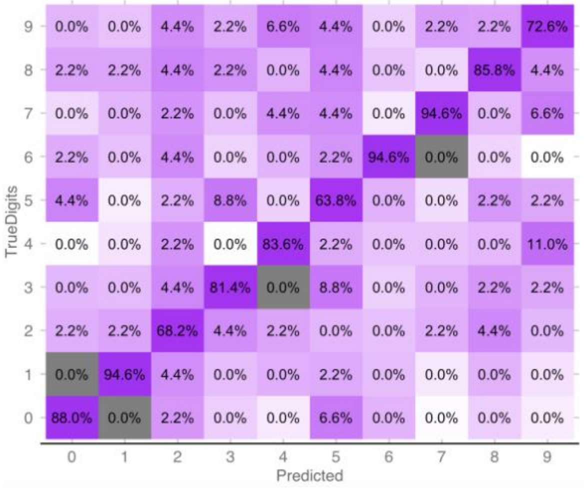




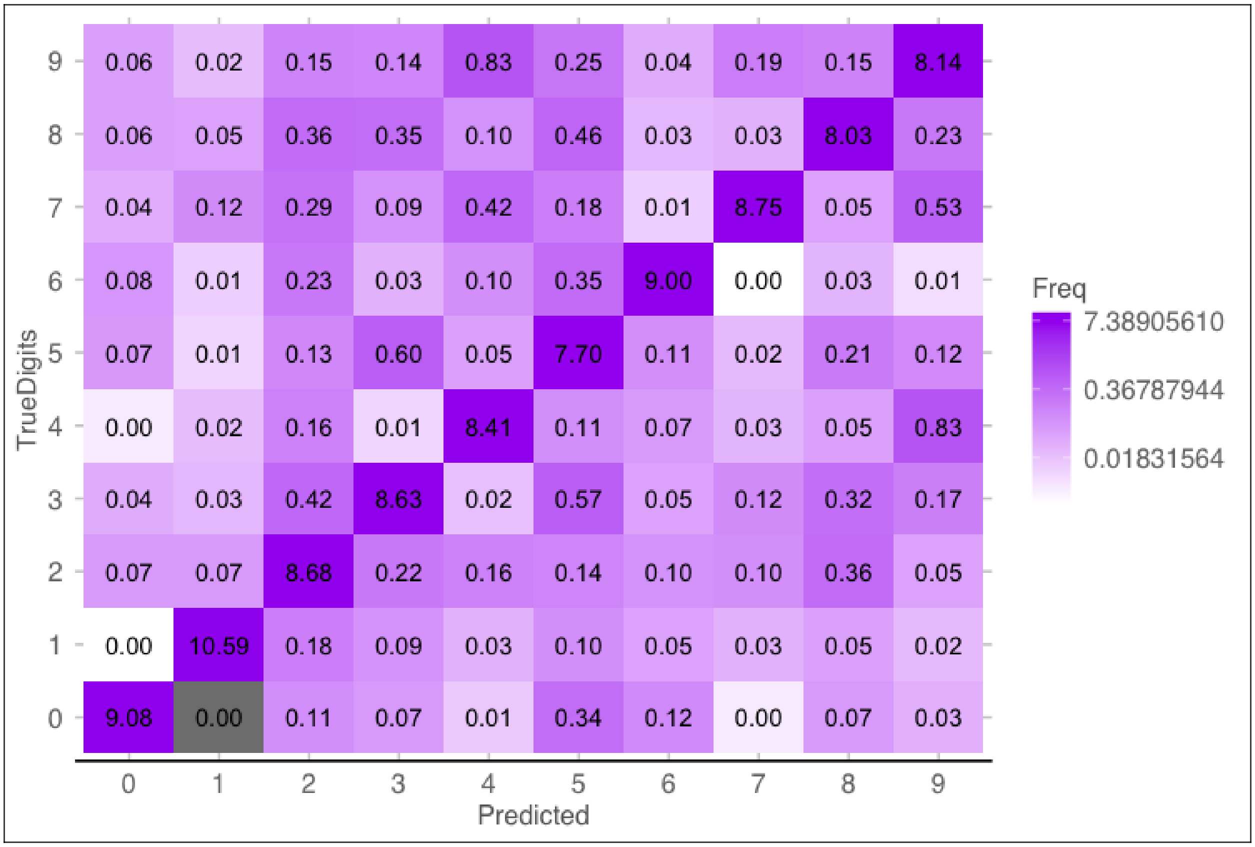
Top 80 Features, Accuracy: 86.15%



Top 150 Features, Accuracy: 82.17%



Cross-Validated 3 fold Confusion Matrix



	0	1	2	3	4	5	6	7	8	9
9	0.06	0.02	0.15	0.14	0.83	0.25	0.04	0.19	0.15	8.14
8	0.06	0.05	0.36	0.35	0.10	0.46	0.03	0.03	8.03	0.23
7	0.04	0.12	0.29	0.09	0.42	0.18	0.01	8.75	0.05	0.53
6	0.08	0.01	0.23	0.03	0.10	0.35	9.00	0.00	0.03	0.01
5	0.07	0.01	0.13	0.60	0.05	7.70	0.11	0.02	0.21	0.12
4	0.00	0.02	0.16	0.01	8.41	0.11	0.07	0.03	0.05	0.83
3	0.04	0.03	0.42	8.63	0.02	0.57	0.05	0.12	0.32	0.17
2	0.07	0.07	8.68	0.22	0.16	0.14	0.10	0.10	0.36	0.05
1	0.00	10.59	0.18	0.09	0.03	0.10	0.05	0.03	0.05	0.02
0	9.08	0.00	0.11	0.07	0.01	0.34	0.12	0.00	0.07	0.03

4 Conclusions