IST 707: Homework 6

**Introduction**

Technology has accelerated the global economy in the past several decades. One of the most revolutionary advances has been the advent of eCommerce. Consumers no longer need to visit brick and mortar locations to obtain essential items, opting instead to order items online.

The lack of necessity for visiting physical locations across the majority industries has presented new challenges for businesses. Those most impacted by the growth of eCommerce have been those that previously relied on superior customer service for competitive advantage. Establishments such as local banks and credit unions previously relied on the inviting, ‘home-town’ feel to attract customers; however, now customers are more prone to choose larger banks for their increased accessibility online.

In order to stay competitive with the major banks, the Syracuse Family Credit Union (SFCU) is seeking to improve its mobile check deposit process. They rushed the feature to market to stay relevant, which led to the use of two full time employees to examine mobile check deposits and verify the handwritten information and signatures. They are hoping to use a machine learning method to automate the process—improving efficiency and reducing customer wait time for mobile funds availability. They have sponsored a study to help them decide which model is best for recognizing handwritten information.

**Analysis and Models**

**About the Data**

The data for this study consisted of images of handwritten digits (numbers) from Kaggle (https://www.kaggle.com/c/digit-recognizer/data). This data was divided into two separate sets for training and testing. The training data set consisted of 42000 digits and the testing data set contained 28000 digits.

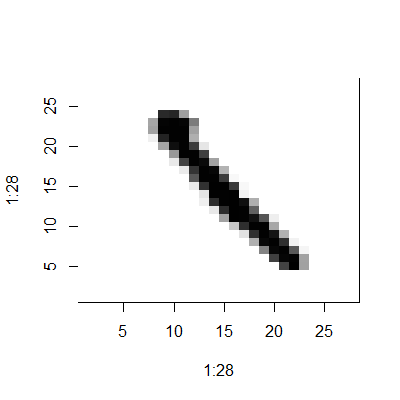
For both the train and test data set, each row of data represented a vectorized version of a handwritten digit. Each digit had 28x28pixels (784 total pixels) that comprised the image, and each row had the individual pixels as attributes. The values for each pixel represented the darkness of the pixel.

The training dataset contained a label at the beginning of each row denoting what number was depicted in the image. For the purpose of validating models, the test dataset did not contain such a label.

To enable modeling on the data, all the pixel values were converted to numeric values and all the labels for the training dataset were converted to a factor ranging from 0-9.

Figure 1 (below) depicts the first row from the training dataset rendered as an image. The correct label for this image is the number ‘1’.

**Figure 1: Image 1 from the Training Digit Data (Number ‘1’)**



**Decision Tree Model**

Initially, a decision tree model was applied to the digits data. This model used the training data subset to inform its decision tree creation and the test data to check its prediction accuracy. This model attempts to make splits based on pixel values to determine the most likely label for the digit. This model uses a complexity parameter to prune its decision trees. For the purpose of this study, .01 was used which causes the model to prune all splits that do not improve the overall R-squared value of the model by at least .01. Once trained, this model was used to predict the digit labels of the test data.

**Naïve Bayes Model**

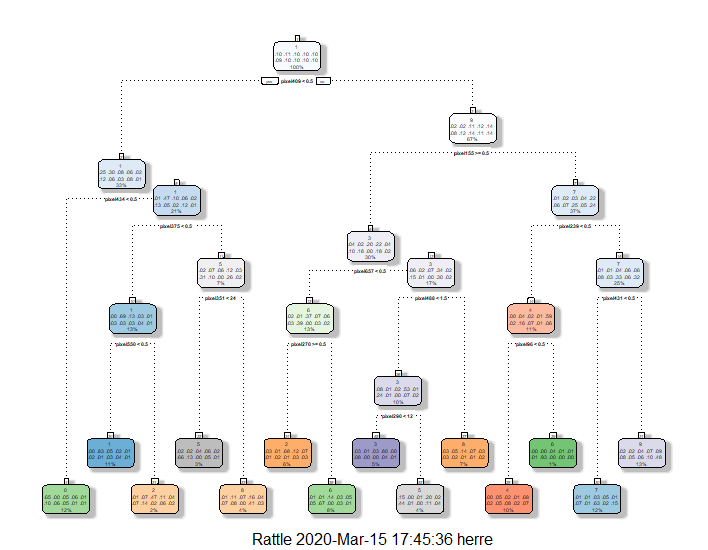
Furthermore, a Naïve Bayes Model was applied to the handwritten digit data. This model uses the learned past probabilities of pixel values associated with certain classes (digit labels) to predict the most likely label for the current digit. Once this model was trained, it was also used to predict the label for the handwritten test digit data.

**Results**

**Decision Tree Model Results**

The resulting decision tree from the decision tree model can be seen below (Figure 2).

**Figure 2: Handwritten Digits Decision Tree**



This model used 27 nodes to decide the most important pixels and respective values in determining a handwritten digit label. This model was able to correctly predict the train data digit label with a 63.72% accuracy. The confusion matrix (with correct predictions in green) from using the model to predict the training data is:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual Digit** | | | | | | | | | | |
| **Predicted Digit** |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 3397 | 13 | 261 | 307 | 52 | 513 | 290 | 275 | 50 | 62 |
| 1 | 2 | 3805 | 236 | 70 | 42 | 76 | 23 | 153 | 163 | 27 |
| 2 | 73 | 106 | 2063 | 389 | 199 | 84 | 174 | 38 | 128 | 95 |
| 3 | 61 | 20 | 68 | 1829 | 8 | 176 | 6 | 7 | 90 | 20 |
| 4 | 12 | 194 | 78 | 60 | 2791 | 88 | 211 | 343 | 63 | 280 |
| 5 | 297 | 26 | 75 | 456 | 54 | 1673 | 192 | 4 | 269 | 87 |
| 6 | 23 | 22 | 460 | 109 | 175 | 169 | 2725 | 4 | 86 | 42 |
| 7 | 50 | 66 | 175 | 235 | 73 | 373 | 27 | 3232 | 108 | 771 |
| 8 | 115 | 345 | 540 | 511 | 158 | 203 | 201 | 27 | 2569 | 127 |
| 9 | 102 | 87 | 221 | 385 | 520 | 440 | 288 | 318 | 537 | 2677 |

This model appears to be quite accurate, however, it will still incorrectly identify digits 1/3 of the time. This model would certainly eliminate the need for SFCU to have two full time employees assigned to mobile deposit verification, but it is likely part of one of the employee’s time would need to be spent verifying the misidentifications of this model.

**Naïve Bayes Model Results**

This model was relatively less successful than the Decision Tree Model in that it was only able to successfully label the digits 52.98% of the time. The confusion matrix (with correct predictions in green) from using the model to predict the training data is:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual Digit** | | | | | | | | | | |
| **Predicted Digit** |  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 3583 | 6 | 9 | 10 | 3 | 8 | 272 | 0 | 158 | 83 |
| 1 | 1 | 4485 | 2 | 10 | 1 | 5 | 58 | 1 | 86 | 35 |
| 2 | 435 | 119 | 702 | 444 | 14 | 36 | 1364 | 12 | 981 | 70 |
| 3 | 220 | 252 | 24 | 1386 | 0 | 5 | 390 | 19 | 1628 | 427 |
| 4 | 111 | 40 | 15 | 19 | 429 | 17 | 532 | 14 | 741 | 2154 |
| 5 | 390 | 120 | 16 | 72 | 15 | 130 | 349 | 5 | 2283 | 415 |
| 6 | 39 | 58 | 4 | 2 | 4 | 13 | 3943 | 3 | 58 | 13 |
| 7 | 30 | 60 | 2 | 24 | 21 | 2 | 45 | 1139 | 207 | 2871 |
| 8 | 60 | 578 | 6 | 35 | 11 | 12 | 122 | 5 | 2517 | 717 |
| 9 | 24 | 56 | 5 | 4 | 15 | 4 | 18 | 25 | 98 | 3939 |

If SFCU were to opt for using the Naïve Bayes Model to verify its mobile check deposits, it would still necessitate one full-time employee for this process. The employee would have to verify the half of deposits that would have digits incorrectly labeled by this model (resulting in mobile deposit error).

**Conclusions**

In summation, technology has contributed exponentially to the global economy as of late. Perhaps the most significant of these technological contributions has been the rapid growth of eCommerce. Consumers no longer need face-to-face interactions in physical locations to obtain the goods and services they need.

Formerly customer service centric industries such as banking have had to adapt their business strategy to be more accessible to customers in online platforms. The Syracuse Family Credit Union (SFCU) rushed the launch of its mobile check depositing system to keep pace with the bigger banks in industry. As a result, they are inefficiently using two full-time employees to validate the handwriting on mobile deposits. In order to help automate this process, SFCU sponsored a study to determine the best machine learning method to employ.

Both a Decision Tree Model and a Naïve Bayes model were tested to see which would more aptly predict handwritten image values for online check deposit verification. The Decision Tree Model proved to be the best option for SFCU, as it would reduce the workload to only a portion of one employee’s time spent verifying misidentified check deposits.