Digital Recognition Data Set:

Number of observation is: 3822  
Number of attributes is: 63

Target Variable is a number between: [0-9]

Mean for each variable:

Figure Mean/Stdv distribution among all attributes

Labels distribution over the dataset:

|  |  |  |
| --- | --- | --- |
| Label | Count | Percentage |
| 3 | 389 | 10.18% |
| 1 | 389 | 10.18% |
| 7 | 387 | 10.13% |
| 4 | 387 | 10.13% |
| 9 | 382 | 9.99% |
| 8 | 380 | 9.94% |
| 2 | 380 | 9.94% |
| 6 | 377 | 9.86% |
| 5 | 376 | 9.84% |
| 0 | 375 | 9.81% |

Table Label distribution

Data Preprocessing:

Looking at the mean and standard deviation attribute number 38 is zero all the time which can be dropped safely without effecting the accuracy of the model either on the test set or training set.

Building Decision Tree:

The process of building the tree was iterative to select the best parameters.

1. Splitting Criteria: “Gini” vs “Entropy”

Entropy proven to provide better accuracy for this data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Time(seconds) | Accuracy/Train | Accuracy/Test |
| Entropy | 0.05 | 0.996860282575 | 0.874721603563 |
| Gini | 0.04 | 0.994505494505 | 0.856347438753 |

Table Effect of using Entropy or Gini on Accuracy

So “Entropy” will be used for the rest of analysis.

1. Tree depth:

Tree depth was a major factor that effects the accuracy. Different values for tree depth tested on the training and testing set.

|  |  |  |  |
| --- | --- | --- | --- |
| Depth | Time(sec) | Accuracy/Training | Accuracy/Testing |
| 1 | 0.00628686 | 0.199110413 | 0.19376392 |
| 2 | 0.011609077 | 0.360282575 | 0.33518931 |
| 3 | 0.015888214 | 0.580847724 | 0.534521158 |
| 4 | 0.021426916 | 0.748822606 | 0.682628062 |
| 5 | 0.033725023 | 0.829408687 | 0.767817372 |
| 6 | 0.045710802 | 0.893511251 | 0.823496659 |
| 7 | 0.046647072 | 0.937205651 | 0.842427617 |
| 8 | 0.054705858 | 0.968341183 | 0.865256125 |
| 9 | 0.056745052 | 0.984563056 | 0.871937639 |
| 10 | 0.056041002 | 0.994767138 | 0.877505568 |
| 11 | 0.064783096 | 0.996336996 | 0.872494432 |
| 12 | 0.060129166 | 0.996860283 | 0.874721604 |
| 13 | 0.066601038 | 0.996860283 | 0.874721604 |
| 14 | 0.062906981 | 0.996860283 | 0.874721604 |
| 15 | 0.060199976 | 0.996860283 | 0.874721604 |
| 16 | 0.055397034 | 0.996860283 | 0.874721604 |
| 17 | 0.073102951 | 0.996860283 | 0.874721604 |
| 18 | 0.066876888 | 0.996860283 | 0.874721604 |
| 19 | 0.06619978 | 0.996860283 | 0.874721604 |

Table Different Training/Testing Accuracy over different Depth

Figure Training/Testing Accuracy over different depth values

Based on the above graph. Tree Depth of 11 is enough to produce 99% accuracy on the training set and 87% on test set.

1. Cross Validation:

To test cross validation as a way to prevent overfitting.

|  |  |
| --- | --- |
| Depth | Cross Validation |
| 1 | 0.19910476 |
| 2 | 0.357384792 |
| 3 | 0.568015093 |
| 4 | 0.730968908 |
| 5 | 0.80006758 |
| 6 | 0.849799618 |
| 7 | 0.877578068 |
| 8 | 0.888056141 |
| 9 | 0.892207654 |
| 10 | 0.894572073 |
| 11 | 0.897181118 |
| 12 | 0.897718442 |
| 13 | 0.897718442 |
| 14 | 0.897718442 |
| 15 | 0.897718442 |
| 16 | 0.897718442 |
| 17 | 0.897718442 |
| 18 | 0.897718442 |
| 19 | 0.897718442 |

Table Cross Validation Mean Accuracy

Figure Accuracy with cross validation

As shown above, the values of 10 folds cross validation are less than the values tested on the train set and much closer to the real values. In addition to that, increasing the tree depth will not increase the performance either on training set or in the cross validation test.

Another parameter that could be tune to overcome overfitting is the max leaf nodes. Which helps to specify the number of leaf nodes for each tree. Specifying the number will force the decision tree to do pruning at early stages to overcome the overfitting issue. The results after applying different values from 10 to 160 with 10 increments each time was as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| # Max Leaf Node | Accuracy/Train | Accuracy/Test | Cross Validation Mean |
| 10 | 0.697278912 | 0.654231626 | 0.682344846 |
| 20 | 0.813710099 | 0.764476615 | 0.783817657 |
| 30 | 0.851125065 | 0.795657016 | 0.821027129 |
| 40 | 0.871794872 | 0.826280624 | 0.834123701 |
| 50 | 0.892203035 | 0.829621381 | 0.848999244 |
| 60 | 0.908948195 | 0.830178174 | 0.861325668 |
| 70 | 0.922291994 | 0.841314031 | 0.86996342 |
| 80 | 0.930141287 | 0.844097996 | 0.875472696 |
| 90 | 0.941653585 | 0.852449889 | 0.883587635 |
| 100 | 0.950549451 | 0.853006682 | 0.888315791 |
| 110 | 0.958398744 | 0.85467706 | 0.889110777 |
| 120 | 0.962585034 | 0.859688196 | 0.89198086 |
| 130 | 0.968341183 | 0.864142539 | 0.894837158 |
| 140 | 0.972789116 | 0.861358575 | 0.897173374 |
| 150 | 0.976975406 | 0.863028953 | 0.897429676 |
| 160 | 0.97985348 | 0.866926503 | 0.898219186 |

Figure Accuracy for Train/Test/CV-Mean over different Maximum leaf node

Looking at the results, the cross-validation accuracy change is very small after 100 leaf node. Even though, the accuracy on the test is higher but the cross validation mean accuracy is more realistic in this case.

The execution time was less than 0.05s which is not a metric to consider in terms of performance.

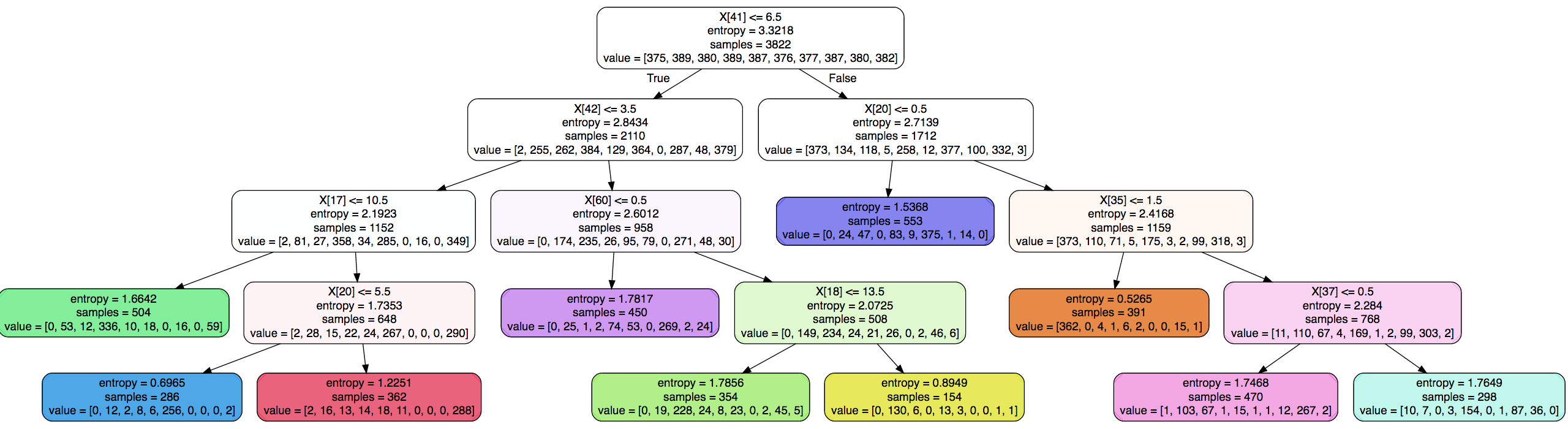
A visualization for the tree with maximum leaf nodes 11 is shown below.

Figure Decision tree with maximum leaf node of 11 (example of pruning)

Amazon Review Baby Product:

Data Exploration:

Number of observation is: 146824 before dropping empty cells, 145927 after dropping empty cells  
Number of attributes is: 3

Target Variable is a number between: [1-5]

Mean for each variable: Name and Review attributes are text. For rating:

Mean is 4.1204, standard deviation is 1.2850

Labels distribution over the dataset:

|  |  |  |
| --- | --- | --- |
| Rating | Numbers | Percentage |
| 1 | 12081 | 9.03% |
| 2 | 9008 | 6.73% |
| 3 | 13309 | 9.94% |
| 4 | 26387 | 19.71% |
| 5 | 85142 | 63.61% |

Table Label distriubtion

Text Preprocessing:

The data cleaning and preprocessing step include:

1. Drop all rows with missing name or review
2. Combine the title and review columns
3. Take the stem for each word in each row/observation
4. Remove stop words and punctuation

Feature Extraction:

In the feature extraction step the count for each stemmed word will be counted. This will create a document-term matrix with the count for the number of appearance of each word.

Training Step:

The document-term matrix has a count for every single appearance for the words. This matrix fed to the Decision Tree Classifier to predict new observation. The results for different runs over different tree depths as follows:

|  |  |  |
| --- | --- | --- |
| Depth | Accuracy/Train | Accuracy/Test |
| 1 | 0.583456112 | 0.579394904 |
| 2 | 0.593927101 | 0.590750747 |
| 3 | 0.600286445 | 0.596126944 |
| 4 | 0.607666847 | 0.602737472 |
| 5 | 0.611819608 | 0.605864443 |
| 6 | 0.614533294 | 0.607619936 |
| 7 | 0.617116778 | 0.608909126 |
| 8 | 0.618720319 | 0.608717119 |
| 9 | 0.622468769 | 0.608333105 |
| 10 | 0.626463917 | 0.609649724 |
| 11 | 0.630657795 | 0.61058233 |
| 12 | 0.63472147 | 0.609073703 |
| 13 | 0.639216869 | 0.608963985 |
| 14 | 0.64556936 | 0.6105549 |
| 15 | 0.651510687 | 0.610445182 |
| 16 | 0.657116229 | 0.609183422 |
| 17 | 0.663160347 | 0.607537647 |
| 18 | 0.669458017 | 0.60731821 |
| 19 | 0.67636558 | 0.606358175 |

Table Accuracy for training and test set over different values of tree depth

Figure Accuracy for training and test set over different values of tree depth

From the graph above, increasing tree depth will not increase the accuracy over the test set after depth 14 which is the maximum. The accuracy for the training set still increase while no increase over the test set. In terms of execution time, the results were as follow:

|  |  |
| --- | --- |
| Depth | Time |
| 1 | 0.732275009 |
| 2 | 0.831966877 |
| 3 | 1.259785891 |
| 4 | 1.754051924 |
| 5 | 2.679045916 |
| 6 | 3.852375031 |
| 7 | 5.427010059 |
| 8 | 7.55112505 |
| 9 | 9.922681808 |
| 10 | 13.16235304 |
| 11 | 16.99228191 |
| 12 | 21.36358094 |
| 13 | 26.12269688 |
| 14 | 31.32825208 |
| 15 | 37.21623611 |
| 16 | 460.6179769 |
| 17 | 56.52627206 |
| 18 | 60.804708 |
| 19 | 68.41602898 |

Figure Execution time in ms with different tree depth

Figure : Execution time (ms) with different tree depth

In general, as expected, the execution time increases while the tree depth increases. There is a spike at the tree depth 16 which might happen due to other processes in the OS that might has higher priority over the python script.

Another factor needed to tune is the minimum number of word counts. For the above approach the minimum number set to 1. The next analysis conducted on a tree depth 11 based on previous analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Min Count | Accuracy/Train | Accuracy/Test | Time/Sec | # Of Features |
| 10 | 0.630698911 | 0.610692048 | 16.67487097 | 11075 |
| 20 | 0.630650942 | 0.610472612 | 15.24012899 | 8101 |
| 30 | 0.630692058 | 0.610637189 | 14.9332571 | 6738 |
| 40 | 0.630685206 | 0.610527471 | 14.90692687 | 5921 |
| 50 | 0.630657795 | 0.610884055 | 15.03190708 | 5328 |
| 60 | 0.630678353 | 0.610692048 | 15.35181904 | 4889 |
| 70 | 0.630685206 | 0.610719478 | 15.10243201 | 4560 |
| 80 | 0.6306715 | 0.610829196 | 15.91423392 | 4274 |
| 90 | 0.630685206 | 0.610856626 | 16.48785496 | 4041 |
| 100 | 0.6306715 | 0.610884055 | 16.09994888 | 3856 |

Table Different accuracy over different number of features

There is no difference between different runs while reducing the number of features. In addition to that, the execution time is almost the same. So, in terms of memory usage and dimension reduction choosing a higher number will be better.

The full tree is very huge given the big and sparse document-term matrix. The figure below visualizes part of the tree NOT all of it.

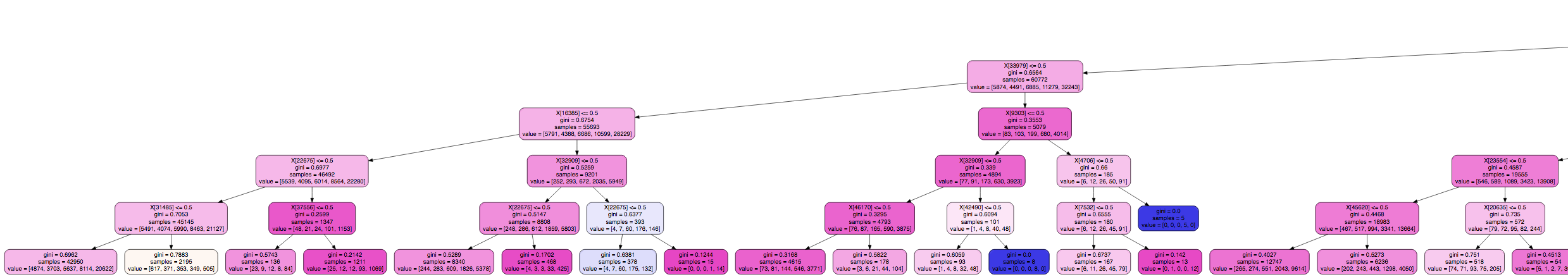


Figure Decsion Tree predict rating. Tree depth is 11; Min feature count is 10

The cross validation tested for the first 6 iteration only since the execution time is very high. The results stayed 3% less which is acceptable as show in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Depth | Accuracy/Train | Accuracy/Test | Cross Validation/Mean | Execution Time/min |
| 1 | 0.583456112 | 0.579394904 | 0.55 | 3 |
| 2 | 0.593927101 | 0.590750747 | 0.56 | 4 |
| 3 | 0.600286445 | 0.596126944 | 0.57 | 5.5 |
| 4 | 0.607666847 | 0.602737472 | 0.57 | 8 |
| 5 | 0.611819608 | 0.605864443 | 0.57 | 12 |
| 6 | 0.614533294 | 0.607619936 | 0.68 | 15 |

Table Cross validation mean accuracy and execution time for each CV

References:

http://scikit-learn.org/stable/modules/tree.html