Unsupervised Learning Capstone
Clustering Vs. Modeling

This article conducts a comparative analysis using clustering and modeling to try and predict the authors of reviews. The purpose is to show competency in nlp topics using clustering and modeling techniques. The data covers 10 years, including 500,000 reviews up to October 2012. Reviews include product and user information, ratings, and plain text reviews. The end result was that the modeling methods outperformed the clustering.

I. DATA

The data includes Amazon fine food reviews from 1999-2012. There are 568,454 reviews, 256,059 users, and 74,258 products. There are 260 users with greater than 50 reviews. The data was published by the Stanford Network Analysis Project (SNAP). It appears SNAPs original use for the data was to predict users preferences based on their expertise; finding products that users will buy now, as well as projecting out what their tastes will be as they gain an 'acquired taste'.

The main packages used are nltk and pandas; as well as sklearn′s feature extraction, preprocessing, clustering, and metrics packages. The data is first cleaned , dropping unnecessary columns. Extras are removed such as dashes, white spaces, hyperlinks, and uppercase letters.

II. FEATURE GENERATION

The target variable that we are trying to predict is theauthor of any given Amazon item. The features to be used will mainly be the text from each review. Individual punctuation marks will be separated and used as features as well.

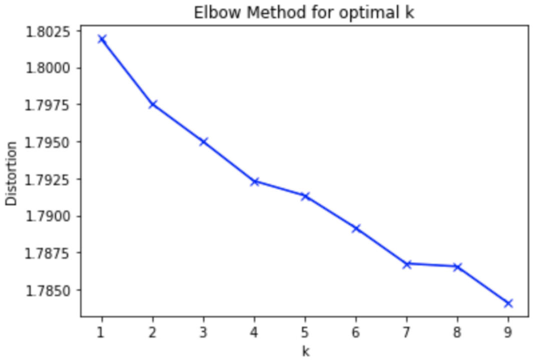
A. Vectorization and Dimensionality reduction

As a part of our feature generation we need to get our text into something that the computer can understand. Sklearn provides a few functions that allow us to be able to turn our text into numbers, which allow the computer to work with it. This process is known as vectorization. Vectorization takes the text data and turns into digit values, it then places it into a vector object like an array or tuple. A few different methods of vectorization are Bag of Words and Tf-idf. BoW finds the most common words. Tf-idf was used for this example. Tf-idf encodes the reviews; this is done with L2 norm - aka euclidean norm. Tf-idf allows the use of rare words to have greater weight than more common words. Now that the computer can use the data we can move on to the next step.

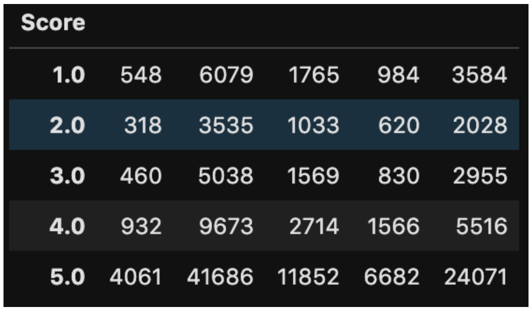
The feature space has 70 thousand features. In order to get this down we need to use some calculus via dimensionality reduction. Traditionally we use a PCA but are unable to do so due to the Tf-idf function. Instead we can use a Truncated SVD, which is a linear dimensionality reduction by means of truncated singular value decomposition. This will allow us to bring the feature space down to 15 thousand count. Now that we have vectorized and reduced our features to a workable amount we can move on to clustering and modeling.

III. CLUSTERING

Clustering is the process of dividing the population of data points into smaller groups so that the groups are similar to the other points within each group. It can be used to find the intrinsic grouping of unlabeled data points.

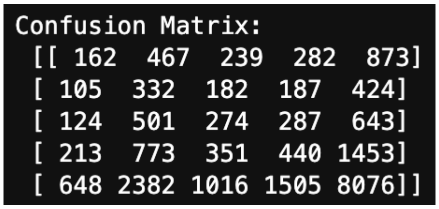


The above image is known a depiction of the Elbow method which shows consistency within cluster analysis. When viewing the above graph, there is no clear drop off point of diminishing return; in other words there is no clear elbow. Mid-point is a fair place to stop for the number of clusters to be used. We also have a few other means of looking at our cluster data. The silhouette score is how similar a point is to its' group; the computed score is -.019; with 1 being a perfect match and -1 being a complete mismatch. The computed mean Silhouette Coefficient of all samples given 5 clusters is slightly skewed to being negatively aligned.



Score					
1.0	548	6079	1765	984	3584
2.0	318	3535	1033	620	2028
3.0	460	5038	1569	830	2955
4.0	932	9673	2714	1566	5516
5.0	4061	41686	11852	6682	24071

The training data confusion matrix above shows that the accuracy isn't very good. The adjusted rand score shows the similarity between a specified number of groups. For 5 clusters the rand score is .0006; this shows that the similarity is next to nothing. When looking at the accuracy for a random forest classifier, the result of 60 is given again. Now that we have reviewed our data via clustering let′s move on to modeling. The adjusted rand score(ari) can be likened unto that of accuracy.

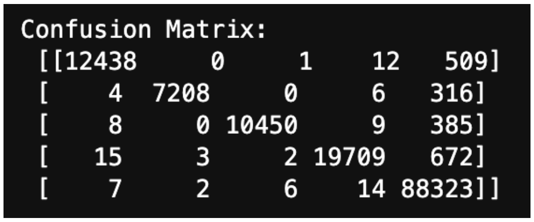


```
Confusion Matrix:
 [[ 162  467  239  282  873]
 [ 105  332  182  187  424]
 [ 124  501  274  287  643]
 [ 213  773  351  440 1453]
 [ 648 2382 1016 1505 8076]]
```

The testing data confusion matrix above shows that our training data isn't quite getting to the point that we want it to.

IV. MODELING

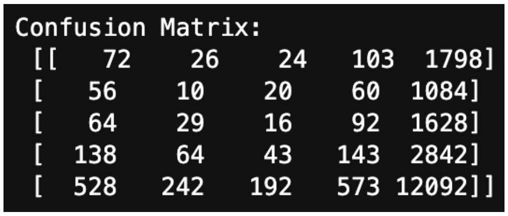
The modeling method of choice is the random forest classifier. This should reduce over-fitting and keep the accuracy high.



```
Confusion Matrix:
 [[12438     0     1    12   509]
 [    4  7208     0     6   316]
 [    8     0 10450     9   385]
 [   15     3     2 19709   672]
 [    7     2     6    14 88323]]
```

A quick look at the above training data confusion matrix will show that the accuracy is better than that of our clustering method. We can see when printing out the accuracy score it is 60 percent. Still not what we're looking for but it's an improvement.

When looking at the cumulative sum of the percent variance captured by all components, we can see that there is an 55 percent score. This shows that there is barely a slant towards the positive, which probably isn't significant. After a dimensionality reduction is performed the adjusted rand score is .0009 and the silhouette score is -.02 percent. The logistic regression is from this is 42 percent.



```
Confusion Matrix:
 [[   72    26    24   103  1798]
 [   56    10    20    60  1084]
 [   64    29    16    92  1628]
 [  138    64    43   143  2842]
 [  528   242   192   573 12092]]
```

We can see that the training data is over fitting by viewing the confusion matrix for the testing set; as presented above.

V. CONCLUSION

We can see that in this scenario, our models appeared to work better than our clusters. When comparing our confusion matrices it can be noted that there seems to be a higher rate of accuracy for the random forest model.

A. Citations and References

http://i.stanford.edu/ (tilde) julian/pdfs/www13.pdf
https://www.kaggle.com/snap/amazon-fine-food-reviews
Code can be found at:
https://github.com/blazecolby/Thinkful/blob/master/Unit%204.%20Unsupervised%20Learning/Unsupervised_Learning_Capstone.ipynb