Assignment Week 3:

Clustering and Measures of Accuracy

For this week's assignment, write your answers, code, explanation, and interpretation of the results in R Markdown. Upload the final document as a PDF file to *Canvas* on or before 9.00 a.m. in the morning of next Wednesday (April 19, 2023).

Applied

The following applied case uses data on the characteristics of top-performing songs on *Spotify*. The data are taken from the *Kaggle* platform where you can find more information about the dataset and the variables contained in the datasets.[1] We will use the 9 variables on songs' sonic characteristics (*bpm*, *nrgy*, *dnce*, *dB*, *live*, *val*, *dur*, *acous*, *spch*) to run some cluster and predictive analysis.

[Q1] Load and examine the dataset *SpotifyTop10s.csv* from *Canvas*. Run *k*-means clustering with 2, 3, and 4 clusters and each time examine using Cluster plots the overlap between clusters. Explain what looks likely to be a good choice for the number of clusters based on these plots. (1.5 points)

[Q2] Run *k*-means clustering with three clusters and assign a cluster to each song. Explore the clusters to find an interpretation of the clusters using the techniques and plots from the lecture. Come up with a meaningful name for each cluster that characterizes the songs in each cluster, e.g. 'fast dance songs' or 'slow acoustic songs.' Explain your choice of names.[2] (2 points)

The following questions are based on the dataset contained in *SpotifyTop50country_prepared.csv* on *Canvas*. The columns with the country and region names indicate the level of popularity of a song in the specified regions and

[1] The first dataset in the file *SpotifyTop10s.csv* is from `https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year` and the second dateset in the file *SpotifyTop50country.csv* is from `https://www.kaggle.com/leonardopena/top-50-spotify-songs-by-each-country`.

[2] It might help to listen to a few songs from different clusters on *Youtube* or *Spotify* to get an impression of their common characteristics.

countries of the world. Missing values indicate that a song was not popular in the particular region or country. The variable *n.countries.hit* indicates the number of countries or regions that a song was popular in. The categorical variable *hit* indicates whether a song was popular in more than one country or region. Your goal in the following exercises is to predict whether a song is likely to be an international hit as defined by the variable *hit* using the nine variables of a song's sonic characteristics.

[Q3] Use a logit model, a random forest model and a K-nearest neighbors algorithm with $K = 5$ to predict whether a song is likely to be an international hit using its nine sonic characteristics as attributes. Use a random split of the data into a train-set of size 70% of the original data to train the models and the remaining 30% of the original data to compute their accuracy (where applicable use a probability threshold of 50% to classify an observation as a positive case). Compute the accuracy of each of the models on the test set as well as of the "model" that predicts for every song that it is not going to be a hit. (2 points)

[Q4] By examining the confusion matrix of each of the four models on the test set explain why the "model" that always predicts that a song is not going to be a hit performs similarly in terms of accuracy compared to the other three models. (1 point)

[Q5] Explain whether one of the three alternative measures (*specificity*, *sensitivity* or *precision*) might be preferable in this case. (1 point)

[Q6] Plot the ROC curve for the logit and the random forest model and compare their performance based on the plot. Also compute the area below the ROC curve in both cases. (1 point)

[Q7] A newly founded data-driven record label is considering to use a random forest model to decide whether a new song that is being proposed by an artist is likely to be an international hit (as previously defined) and is thus worthy of being produced by the label. Assume that an international hit generates revenues of €1,300,000 for the label while producing and promoting a song costs €500,000. A song that is not an international hit does not generate any revenues but still costs €500,000 if the label has decided to produce and promote the song. Would you advise the label to use the random forrest in such a manner? Explain your answer using an expected value calculation using the previously estimated model and test set. The baseline scenario for comparison is to not produce and promote any songs. (1.5 points)