Project 4 Feedback

Team ADM : Avraham Adler, Donny Lofland, Michael Yampol
http://rpubs.com/myampol/DT607_Fall2019_Project4_TeamADM
Great strategy. Splitting the sample in 50% training, 25% validation, and 25% testing groups. Testing and comparing performance of a number of prediction models : logistic regression, GLM, random forest, neural network. "Naive Bayes performed quite poorly while the remaining models all did quite well, but the winner is the gradient boosted model, with the highest F-score and fewest miscategorized emails of any type". Insightful analysis of the data set and prediction techniques.

In the word cloud visualization, the parameter « random order » equal to « FALSE » would position the words in order of frequency, going from center, as it is traditionally the case in modern clouds. Smth of the sort :
wordcloud(hamCorp, scale=c(3,0.5), max.words=80, random.order=FALSE, rot.per=0.35, use.r.layout=FALSE, colors=brewer.pal(8, "Dark2"))
title("Ham Wordcloud",col.main = "grey14")

Mael Illien
http://rpubs.com/maelillien/data607project4
Correct differentiation between the classes of Ham and SPAM top words, using Naive Bayes.
Interesting suggestions in conclusions : Equalizing the number of training instances by sampling so that spam and ham emails are represented equally. Cross-validation of training data. Using different models for performance comparisons (logistic regression, SVM, random forest, etc.).

James Mundy, Samuel Kigamba, Alaine T Kuiete, Banu Boopalan
http://rpubs.com/BanuB/550426
http://rpubs.com/BanuB/550501
Great collaborative project using two models: Naive Bayes and SVM, exploratory analysis including word clouds and sentiment analysis.
Amazing accuracy.

Devanshu Mehrotra
http://rpubs.com/DevanshuMehrotra/551002
Great job importing spams and hams into data frames, creating and cleaning the corpus, producing a Document Term Matrix, and training a Decision Tree Classifier. High precision of prediction algorithm.

Uliana Plotnikova
http://rpubs.com/uplotnik/548829
Concise correct solution : unpacking the data, creating and cleaning corpora, building word clouds, Document Term Matrix, training prediction model (Naive Bayes Classifier) with high accuracy.

Jai Jeffryes, Tamiko Jenkins, Nicholas Chung
http://rpubs.com/PnoJai/550986
Great strategy and step by step description of considerations. Bravo for tenting more than one technique. High precision of prediction using SVM algorithm.

Word cloud would look in a more traditional way if using a similar code:
wordcloud(hamCorp, scale=c(3,0.5), max.words=80, random.order=FALSE, rot.per=0.35, use.r.layout=FALSE, colors=brewer.pal(8, "Dark2"))
title("Ham Wordcloud",col.main = "grey14")

Sufian http://rpubs.com/ssufian/546815
Correct understanding of the task, unpacking & cleaning of the data, preparation of the corpus and Term Document Matrix, drawing word clouds, using Naive Bayes classifier with high accuracy. Bravo for checking for the imbalance of the original dataset.

Habib Khan
http://rpubs.com/habibkhan89/550957
Nice work creating TermDocument Matrix, word clouds, and performing sentiment analysis.
In order to predict which email will be spam and which ham, one could train a Naive Bayes classifier or an SVM model.

Jayraman Ramalingam
http://rpubs.com/jey1987/551048
Nice job creating data corpus and Document Term Matrix.

Very interesting attempt to train numerous different models: SVM, random forests, max entropy, decision tree, logit boost, bagging.
However, producing a confusion matrix would give prediction accuracy.

Word cloud would visualize the comparison of top words in hams and spams.