Week 11 Assignment

Note: Late homework assignments are not accepted because the solutions are discussed at the start of every class on Tuesday. Your solution to this assignment must be uploaded on eCampus as a PDF file. A common approach is to put the solution into a word document and then save that into a PDF file. Please only submit PDF files.

Assignment: You can complete one of the two parts to this assignment, or both. If you do both parts, you can obtain extra points for this assignment. You are expected to complete at least one part successfully.

Data Files: Excel File <*GMC_Complaints.xlsx*> -> An excel file containing 2,375 complaints about specific GMC vechicles submitted to the National Highway Safety and Traffic Administration (NHTSA). The complaints are in the column labeled 'description'. The full data dictionary is:

nthsa_id:	A number unique for each complaint (an ID)
Year:	The car year – 2003 thru 2011
make:	The make of the car – CHEVROLET, PONTIAC & SATURN
model:	The car model – COBALT, G5, HHR, ION, SKY & SOLSTICE
description:	The actual complaint(text)
crashed:	A binary attribute – 'N' for no and 'Y' for yes
abs:	Anti-Brake System – 'N' for no and 'Y' for yes
mileage:	The miles on the car at the time of the accident – 0 – 200,000

There are outliers and missing values in these data.

Part 1: Create a SAS EM project names "Week 11 Homework". In that project read this data file using the file import node from the text mining tab in SAS EM.

1. Conduct a text classification analysis of these data following the analytical process described in class.
2. Use POS, stop words and stemming for building the term/doc matrix.

3. Use TF-IDF weighting for the term/doc matrix.
4. Use text topic node to develop 8 clusters for the complaints.
5. Develop a Regression Model that predicts the probability of a crash (binary). (Use stepwise regression to find the best combination of predictors. The possible predictors are the other attributes: the topic probabilities, year, make, model, abs & mileage.
6. Use a 70/30 split to evaluate the model and report the confusion matrix along with the binary metrics.

REPORT:

1. Report a screen shot of the diagram and the property windows for the parse, filter, text topic and regression nodes.
2. A description of the 8 topics.
3. The confusion matrix
4. The accuracy, precision, and F1 for the validation data.

Part 2: Do the same assignment as Part 1 using Python, except instead of stepwise regression use 10-fold cross validation to select the regularization parameter 'C'.

Follow the process described in the week 10 notes. Use pandas to read the file, NLTK for tokenization, POS, stop word removal and stemming. Use 'sci-learn' for LDA to identify the top 8 topics.

Analysis:

1. Conduct a text classification analysis of these data following the analytical process described in class.
2. Use POS, stop words and stemming for building the term/doc matrix.
3. Use TF-IDF weighting for the term/doc matrix.
4. Use LDA to develop exactly 8 clusters for the wine reviews.
5. Develop a Regression Model that predicts the probability of a crash (binary). (Use stepwise regression to find the best combination of predictors. The possible predictors are the other attributes: the topic probabilities, year, make, model, abs & mileage.

6. Use a 70/30 split to evaluate the model and report the confusion matrix along with the binary metrics.

REPORT:

1. The Python code (.py file).
2. The logistic regression model found to predict the probability of a crash.
3. The confusion matrix and associated metrics;