Week 9 Assignment

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 File: Excel File <*HondaComplaints.xlsx*> -> contains over 5K consumer complaints submitted to the NHTSA. The complaints are related to the Honda and Acura models 2001-2003.
Data Dictionary: PDF File <*Honda_Data_Dictionary.pdf* > ->
There should be no missing values in these data. However, the information on "State" is unreliable and should be ignored.

Part 1: Create a SAS EM project named "Week 9 Homework". The objective of this project is to build the best supervised learning classifier for classifying accidents based on the attribute "crash". The predictors for these models will be the 8 topics identified using the text cluster node.

Diagram Points

1. Impute any missing values and outliers.
2. For the Text Analysis, use the default parse properties. For Text Filter, use TFIDF weighting on the term counts (frequency). Identify any needed additional synonyms and stop words.
3. The Text Cluster node should produce exactly 8 topics.
4. Follow the structure described in class for fitting and validating logistic regression, decision tree, random forest and ensemble models for classifying "crash." Neural Networks is not being used since it does not provide direct information importance of the predictors on the response.
5. For validation, use a 70/30 validation approach.
6. For random forests and decision trees, explore maximum depths between 5 and 20.
7. For logistic regression, use stepwise variable selection.

REPORT:

1. Report a screen shot of the diagram and the final attribute setting shown in the metadata variable table.
2. The fit statistics for each of the models you build.
3. For random forests and decision trees, show variable importance
4. For the logistic regression, show the table of coefficients
5. Summarize you analysis in a few paragraphs describing the relative performance of your models. Select the best model and describe what it says about the relationship of "crash" to the other attributes.

Part 2: Do the same assignment as Part 1 using Python.

Follow the process described in the week 9 notes. Use pandas to read the file. Use LDA with TFIDF to identify the top 8 topics. Build the three models described in part 1 – random forest, decision tree, logistic regression and ensemble. Use 70/30 cross validation to identify the best depths for the trees and the best logistic regression model.

Analysis: Try using 5-fold cross validation to identify the best parameters for the tree and regularization models. If this does not work well, resort to using a simple 70/30 partition of the data to evaluate the model parameters.

For trees, explore depths from 5 to 20. For the forest, use n_estimators = 100 (number of trees) to reduce the computer runtime.

For logistic regression, use stepwise for feature selection. Also use logistic regression with Elastic-Net regularization. Finally, you can build an ensemble of all models.

Hint - Templates: Use the Week 9 GMC example as a template for conducting the text analysis and model. This code allows you to customize the text analysis. The biggest modification is reading the data file. In the end, it adds the identified topics and their probabilities to your original DataFrame. This is stored in a pickle file.

Once the pickle file is created, you can open it and build your model for the binary target "crash." The models are there, but the hyperparameter optimization is not. You are asked to add this as a challenge. This will be discussed at the Q&A.

You will need to produce the Python attribute map and resort to using ReplaceImputeEncode. This is described in the data dictionary.

Use one-hot encoding for the class variables, and do not drop the last nominal column (drop=False). This creates a colinear relationship in X, but this is only a problem if you try to fit all predictors in a logistic regression model. Use Stepwise regression should avoid this problem.

Best of luck with this coding. I plan on answering questions about this assignment at the Monday Q&A, July 20.

REPORT:

1. The Python code (.py file) and its output
2. Your description of the differences between the models. Select the best model and describe its implications for the relationship of the predictors for classifying "crash."