***PALASH GOYAL*** [**palashgoyal1@gmail.com**](mailto:palashgoyal1@gmail.com)Listing as many use cases for the dataset as possible. - Predicting Prices for the given Car/Truck  
 - Classifying Symboling for the given Car/Truck : To figure out the Car/Truck is safe or not  
 - Finding the Normalized losses for any given Car/Truck, i.e., average loss on each car every year  
 - Optimizating the best configuration of cars to yield minimum loss

Picking one of the use cases listed above and describing how building a statistical model based on the dataset could best be used to improve it  
 - I have picked up the Symboling classification problem  
 - The assigned symbol to the car has been partitioned in 2 categories, 0 and 1.   
 - 0 for cars with riskiness (i.e., Symboling value greater than 1), and 1 for cars with lesser riskiness (i.e., Symboling value lesser or equal to 0)   
 - Various features given have been segmented as continuous and categorical variables, and are used for the classification model  
 - XGBoost model has been used finally for classification approach

Implementing the model described above in Python. The code is retrieving the data, train and test a statistical model, and reporting relevant performance criteria.   
 - Model described above has been implemented in Python, in 3 separate files

Explaining each and every design choices (e.g., preprocessing, model selection, hyper-parameters, evaluation criteria). Comparing and contrasting choices with alternative methodologies.  
 - All the steps have been explained in detail along with the code. Below is a snapshot of the same  
 - The code has been divided in 3 files :  
 ***Data\_Preprocessing\_1.ipynb*** :   
 - Data Loading  
 - Segmentation of Continuous and Categorical variables  
 - Imputation for Continuous and Categorical variables   
 - Categorical variables split into multiple cols  
 - Train data and Target variable are saved   
 ***Feature\_Selection\_2.ipynb :*** - VIF checks are made for all variables  
 - Variable Importance is recorded for all variables  
 - ANOVA F-value for Continuous variables, and Chi2 P-value for Categorical variables used to select variables  
 - Above stats for all variables are stored in *auto\_data.csv* (or, *auto\_data\_xls.xls*) file, and variables are shortlisted  
 - Shortlisted Continuous and Categorical variables are saved   
 ***Model\_Tuning\_Minimizing\_CV\_LogLoss\_3.ipynb :*** - Above saved Train, Target, Shortlisted Categorical, Shortlisted Continuous variables are loaded   
 - XGBoost model with default parameters is fitted on the above dataset   
 - Model parameters are tuned with aim of Minimizing LogLoss  
 - all hyper parameters are stored simultaneously in separate files  
 - Final model with best hyper-parameters is saved  
 - Model is tested on Training data itself, and shows promising results across all the top decile

Improving the model made above if you had more time.  
 - Using other classification models, like Logisitic, RandomForest, Ensemble Approach  
 - Creating more features for the given set of Car/Truck  
- Using better techniques for the imputation of the Normalized Losses, Price and other variables