Data Mining Final Project Individual Report

Jenny Tsai
April 28, 2020

I. Introduction

Our project is to explore weather and road conditions that might impact the car accident severity. Below is the project assignment for each team member:

- Aziz – Preprocessing & EDA
- Jenny – Modeling (plus a bit preprocessing prior to modeling)
- Mojahid – GUI (PyQT5)

II. Description of My Work

- Find the dataset from Kaggle
- 2nd stage preprocessing prior to modeling
    - Explore ways to address imbalanced data issue (e.g., resample)
    - Run simple EDA (e.g., frequency, descriptive statistics) and pick variables that are useful for modeling (e.g., drop location and time and drop variables with too many NaNs after pre-processing)
    - Drop Nans
    - Recode binary variables to 0 and 1
- Try different models: Logistic, Random Forest, AdaBoost
- Perform grid search and cross validation for RF and AdaBoost to find best parameters
- Write up introduction, modeling, results, summary, and reference section in the group report
- Use QT designer to build modeling tabs layout as reference for Mojahid
- Create a subsample file of 2,000 records for GUI demo

Please reference the py.files in the folder for my codes.

- The modeling codes of Random Forest and Logistic Regression are mainly from Professor Amir's lecture codes.
- The original codes are grid search and CV, some pre-processing (e.g., frequency table loop), resample codes, and AdaBoost (from sklearn documentation).
- Approximately 50% are original.

III. Results & Summary

- Selected 19 variables for modeling after cleaning & EDA
- Resampled results were either underfitting (when undersampled) or overfitting (when oversampled), so we used the subset pulling high severity cases from 2018 instead
- Logistic regression didn't perform well, so only kept RF and AdaBoost in the final presentation and report