Jina Lee

Problem and Discussion of Background:

Looking at the target field of collision severity description, there are 4 levels excluding the 0-Unknown. However, looking at the dataset given, we see that this is an imbalanced dataset with actually only two types of collision severity of 2-injury vs 1-prop damage. Using this dataset, we can look to solve the task to predict and classify a given collision into “1-property damage only” or “2-injury” groups based on this given dataset. Being able to predict the collision severity ahead of time would be valuable in preparing for the amount of damage ahead of time or applying new measures to reduce accidents. It would also be valuable to determine which variables have the most influence on collision severity. Therefore, I have defined my objective to produce a machine learning model that can classify the severity of a collision in Seattle with high accuracy based the collision data recorded by Traffic Records and additional features created during the course of this project.

Description of data and how to solve:

The dataset that will be used in this study is collected and recorded by the Seattle police department with 194673 rows and 37 columns excluding the target severity code column. Each row is a collision record in the city of Seattle from 2004 to present with information about the location, number of injuries, number of vehicles involved, and the weather condition during the time of collision. There are undoubtedly more features than are really necessary here, and some of the dimensions such as severity description is useless. Some will be cleaned out using PCA or other feature selection methods to keep only those with the most impact on the classification.

Then, the data will need to be cleaned and transformed into a usable format for analysis. With this step, feature engineering will also be considered to see if derived features can be used to better predict the severity of the collision. Pipelines will be created to apply on numerical features and categorical features. For numerical features, we will implement imputers with several different missing value handling strategies. A scaler will also be used as part of the pipeline for preprocessing of the numerical data. For categorical features, we will use Sklearn's OneHotEncoder or LabelEncoder. Some steps for feature engineering and dimension reduction will need to be conducted separately outside the pipeline. Then, the pipeline will union the final transformed, scaled numeric and categorical features that will be used in predicting the target variable.

This is a classification problem where the target variable has a value of 1 or 2. Therefore, standard algorithms that are fit for classification problems will be considered. Some of the algorithms that will be assessed in this project are logistic regression, K-nearest neighbors, support vector machines, naive Bayes, random forests, and finally, deep learning. After processing the data, a designed pipeline will try this list of classifiers during the grid search. The pipeline will also try different combinations of parameters for each classifier. Cross validation will be used to evaluate estimator performance. The estimator with the best cross-validation training accuracy will be used for analysis. Other metrics such as Receiver Operating Characteristic Area Under Curve (ROC AUC) and p-value of the models will also be evaluated and discussed as part of the results. The deliverable will be a working ML model capable of predicting the collision severity with high precision and accuracy based on other information about the collision.