Road Accident Severity Prediction

Introduction:

Due to the urbanization process around the globe, traffic accidents have been on a tremendous rise, causing significant life and property losses. Predicting traffic accidents is critically important in improving transportation, public safety and as well as safe routing. The objective was to focus on analyzing the factors which could help predict the severity of the accident. Since most of the predictor variables in the dataset were categorical, categorical variables were recorded. 11 models were built, evaluated for complexity and accuracy, and compared to conclude which model is the best fit for predicting accident severity.

Spot Checking technique was used to fit the 11 models to determine which models would predict the accident severity with the highest accuracy. We also performed feature engineering to enrich our dataset Hyperparameter tuning and pipelining the best performing model helped to improve the performance of the model by making accurate predictions.

Data Description

The dataset on road traffic accident analysis has about 2 million rows and 63 columns. The predictor variables in the dataset had information geographical locations, weather conditions, type of vehicle, number of casualties and vehicle maneuvers. The dataset had missing values which were dropped.

During the exploratory data analysis part, w some compelling details were found like the age group between 26 years to 35 years was more frequent with accidents than any other age group. It was also explored that most of the accidents happened on Friday. It was surprising to see that weather conditions did not play much role at the time of accident. Most of the accidents happened during the afternoon rush that is between 3.00 pm to 7.00 pm. This data exploration gave a clear insights that road conditions, rush hours, weekdays, age group, age of vehicle, and junction details played a major role in road accidents.

Multiclass target variable was classified into 2 classes i.e. slight as ’0’ and Fatal or severe as ‘1’. The original dataset had only 13% rows for severe or

Fatal accident severity, this was a highly unbalanced dataset. In order to balance it for the modeling, a sampling technique was used to balance out the Slight and Severe or Fatal accident severity.

Model Description

The final modeling dataset of 381K rows and 62 columns was split into training data with 70% and test data with 30%. Spot-checking technique was implemented to determine the machine learning model which was best suited to predict the accident severity. Machine learning models which was used are Logistic Regression, Random Forest Classifier, K Neighbors Classifier, Gaussian NB, Perceptron, SGD Classifier, Decision Tree Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis, Extra Trees Classifier, and Bagging Classifier.

Based on the accuracy, from 11 different models the data was narrowed down to 6 best models whose accuracy is more than 60%.

Further, permutation testing was implemented for feature importance for our best model i.e. Gradient Boosting. Here it was seen that the features like Engine capacity, Number of vehicles, speed limit, age group of drivers are some of the features which were believed to be important in making predictions. While it was also seen that features like pedestrian crossing human control, vehicle location restricted lane, road surface condition code, weather condition code, were comparatively less important features.

Results

I used spot checking with 11 different models to find out which all models are the best suitable for predicting the results. After this step, hyper parameter tuning for all the 6 models was done. With this step I got an insight about the best parameters for each algorithm. After using hyperparameter tuning and based on the accuracy score, it was decided that the best model was Gradient Boosting Classifier.

Discussion

Road accidents are a serious problem in our societies across the globe. The world health organization estimated that 1.25 million deaths were related to road traffic injuries in the year 2010. Transport authorities worldwide have been striving to implement strategies to minimize the road traffic accidents by introducing safety regulations. There were many strategies implemented for road traffic accidents reduction but has proven to be an elusive goal as these measures have hitherto failed to make a considerable reduction in the frequency of the road traffic accidents. Accidents are influenced by many measurable factors such as driving speed, road condition, weather condition, light condition and so on. Therefore, many researchers have come together to understand the dynamics of road traffic accidents.

Katannya Kapeli and Meraldo Antonio (2019) researched that weather conditions had no role in severe or fatal accidents. This is quite evident because there were multiple factors that affected the road traffic accidents, they considered junction, time, origin and destination played a vital role in predicting the road traffic accidents. They used a negative sampling technique using the several hundreds of accident hot spots and they classified their target variable with accident and no accident. They used classification machine learning models and their best model was random forest with only numerical predictors.