PREDICTING THE SEVERITY OF CAR ACCIDENTS

IBM DATA SCIENCE CAPSTONE PROJECT

BUSINESS PROBLEM

Road accident is most unwanted thing to happen to a road user, though they happen quite often. The most unfortunate thing is that we don't learn from our mistakes on road. Most of the road users are quite well aware of the general rules and safety measures while using roads but it is only the laxity on part of road users, which cause accidents and crashes. Main cause of accidents and crashes are due to human errors. We are elaborating some of the common behaviour of humans which results in accident. How different factors of Roads contribute in Accidents:

1. Drivers: Over-speeding, rash driving, violation of rules, failure to understand signs, fatigue, alcohol.
2. Pedestrian: Carelessness, illiteracy, crossing at wrong places moving on carriageway, Jaywalkers.
3. Passengers: Projecting their body outside vehicle, by talking to drivers, alighting and boarding vehicle from wrong side travelling on footboards, catching a running bus etc.
4. Vehicles: Failure of brakes or steering, tyre burst, insufficient headlights, overloading, projecting loads.
5. Road Conditions: Potholes, damaged road, eroded road merging of rural roads with highways, diversions, illegal speed breakers.
6. Weather conditions: Fog, snow, heavy rainfall, wind storms, hail storms.

DATA

We chose the unbalanced dataset provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state (whether under influence or not), collision type. Hence, we think it is a good generalized dataset which will help us in creating an accurate predictive model.
The unbalance with respect to the severity code in the dataset is as follows.

SEVERITY CODE	Count
1	136485
2	58188

METHODOLOGY

- Data Cleaning
  - Remove unrelated attributes
    - Location coordinates, Incident Key etc.
  - Remove attributes vulnerable to significant data missing
    - Speeding – Over 95% of values missing
- Exploratory Data analysis
  - Determine impact of features to severity
- Data Preparation
  - Convert categorical variables to numerical variables
    - One-hot encoding
- Modelling and Prediction
  - Logistic Regression, KNN, SVM, Decision-tree etc.

CONCLUSION

None of the algorithms implemented above gave an accuracy score equal to or greater than 0.7, they all ranged from 0.6 to 0.7. Meaning, these models can predict the severity code of an accident with an accuracy equalling 60-70%. A bar plot is plotted below with the bars representing the accuracy of each model in descending order respectively.

THANK YOU!