Data Science

Introduction/Business Problem

Vehicular accidents are common on roads across the world.

1. Accidents vary in severity.
2. What if it was possible to predict the severity of an accident occurring given current conditions?
3. Drivers across the world would benefit from this information.

Data

4. Seattle Department of Transportation (SDOT) dataset selected.
5. SDOT dataset includes entries for nearly 195,000 accidents from 2004 to the present.

6. The severity of each accident is categorized with multiple features to choose from for modeling.
7. A few examples of the features:
    a. Location of Collision and Collision Type
    b. Number of people, pedestrians, cyclists, and vehicles involved in the collision
    c. Number of fatalities
    d. Weather, Road, & Lighting conditions
    e. And more
8. The investigation is focused on environmental driving conditions and will use to following features:
    a. Weather Conditions (WEATHER)
    b. Road Conditions (ROADCOND)
    c. Light Conditions (LIGHTCOND)

Methodology :

1. Imported the required dataset using pandas read_csv method.
2. Then gone through the dataset
3. Then removed the rows having nan or empty values using dropna method
4. Then dropped the rows having ‘other’ or any ‘unkown’ values in it
5. The dataset had categorical features and ML model cannot understand it so i converted then to numerical values
6. Then splited data into X and y
7. Then X and y was splited in X_train,X_test,y_train,y_test and 20% data was used for testing
8. Then used supervised machine learning models like decision tree and logistic regression
9. Using decision tree we got accuracy of 67%
10. Using logistic regression we got 67% accuracy

Discussion

▸ Models are inaccurate.

▸ Environmental factors are not enough on their own. ▸

Root cause is believed to be the lack of features. ▸

Examples of features to improve the model include: ▸

Time of day

  ▸ Location

  ▸ Clusters of inattention or DUI related causes

Conclusion

▸ Overall, the results of this investigation are disappointing. ▸ The hope was to create models based only on environmental factors. ▸ Created models are inaccurate.

▸ Not enough features were selected from the dataset.

- Recommendation is to learn from the results and not use the created models.
- Future models with more features may yield better results.