Data Science Capstone

October 13, 2020

Introduction/Business Problem

- Vehicular accidents are common on roads across the world.
- Accidents vary in severity.
- What if it was possible to predict the severity of an accident occurring given current conditions?
- Drivers across the world would benefit from this information.

Data

- Seattle Department of Transportation (SDOT) dataset selected.
- SDOT dataset includes entries for nearly 195,000 accidents from 2004 to the present.
- The severity of each accident is categorized with multiple features to choose from for modeling.
- A few examples of the features:
  - Location of Collision and Collision Type
  - Number of people, pedestrians, cyclists, and vehicles involved in the collision
  - Number of fatalities
  - Weather, Road, & Lighting conditions
  - And more
- The investigation is focused on environmental driving conditions and will use to following features:
  - Weather Conditions (WEATHER)
  - Road Conditions (ROADCOND)
  - Light Conditions (LIGHTCOND)

Methodology

- Create a clean dataframe:
  - Import desired data into a dataframe
  - Remove any rows with missing or NAN values
  - Drop any rows containing “Other” or “Unknown” in the feature columns
  - Consolidate similar feature values into a single value type
    - For example, the LIGHTCOND feature had 4 types of “Dark” that were consolidated into a single “Dark” value

	WEATHER	ROADCOND	LIGHTCOND	SEVERITYCODE
0	Overcast	Wet	Daylight	2
1	Raining	Wet	Dark	1
2	Overcast	Dry	Daylight	1
3	Clear	Dry	Daylight	1
4	Raining	Wet	Daylight	2

Methodology continued

- Create a new dataframe that is a numeric representation of the clean dataframe:

	WEATHER	ROADCOND	LIGHTCOND	SEVERITYCODE
0	2	1	0	2
1	1	1	1	1
2	2	0	0	1
3	0	0	0	1
4	1	1	0	2

Methodology continued

- Supervised Machine Learning with Classification Models selected
- K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and Support Vector Machine (SVM) models to be evaluated
- Prepare data:
  - Create dataframe “X” for features
  - Create array “y” for labels
  - Split X & y into training and testing data with 20% for testing
  - Normalize X train and test data
- Create models and fit the training data:
  - Use for loop to determine best value of k for KNN model
  - Create Decision Tree using the entropy criterion and a max depth of 4
  - Create a Logistic Regression model using the default lbfgs solver and C = 0.01
  - Create an SVM model using the default rbf kernel

Results

- Predict a “yhat” for each model using the test data
- Generate the Accuracy and F1 scores for each model
- The Decision Tree and Logistic Regression models have the highest and equal values for both the Accuracy and F1 scores
- Either model may be considered the “best” out of the four models created

Model	Accuracy	F1-Score
KNN	0.670243	0.802569
Decision Tree	0.673305	0.804761
Logistic Regression	0.673305	0.804761
SVM	0.673247	0.804719

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
- Models are not recommended for use until improved.

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.