Applied Data Science Capstone Project Report – Car Accident Severity

BY IRINA REBROVA

Problem Statement

TO REDUCE THE FREQUENCY OF CAR COLLISIONS IN A COMMUNITY, AN ALGORITHM IS RECOMMENDED TO PREDICT THE SEVERITY OF AN ACCIDENT GIVEN THE KEY PARAMETERS:

- CURRENT WEATHER
- ROAD CONDITION
- VISIBILITY CONDITIONS

WHEN CONDITIONS ARE BAD, THIS MODEL WILL ALERT DRIVERS TO REMIND THEM TO BE MORE CAREFUL.

Data Understanding

- THE DATA WAS COLLECTED BY THE SEATTLE POLICE DEPARTMENT AND ACCIDENT TRAFFIC RECORDS DEPARTMENT FROM 2004 TO PRESENT

- PREDICTOR:
  - SEVERITYCODE

  IT IS USED TO MEASURE THE SEVERITY OF AN ACCIDENT FROM 0 TO 4 WITHIN THE DATASET.
- ATTRIBUTES:
  - WEATHER
  - ROADCOND
  - LIGHTCOND

  THE ATTRIBUTES ARE USED TO WEIGH THE SEVERITY OF AN ACCIDENT

Data Preprocessing

- DATA TYPE ENCODING

```
SEVERITYCODE      int64
WEATHER           int64
ROADCOND          int64
LIGHTCOND         int64
```

- BALANCING DATASET

```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

Methodology

ML MODELS USED:

1. **K-NEAREST NEIGHBOR (KNN)**

   KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance

2. **DECISION TREE**

   A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions

3. **LOGISTIC REGRESSION**

   Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is good to use with logistic regression

Results

ML Model	Jaccard Score	F1 Score	Log Loss
KNN	0.52	0.52	
Decision Tree	0.56	0.49	
Linear Regression	0.54	0.53	0.68

Results Understanding

- WHILE **KNN AND DECISION TREE** MODELS WORKED WELL FOR THIS PROJECT, **LOGISTIC REGRESSION** MADE MOST SENSE BECAUSE OF ITS BINARY NATURE.
- EVALUATION METRICS USED TO TEST THE ACCURACY OF OUR MODELS:
  - Jaccard index
  - f-1 score
  - Logloss (for logistic regression)
- MODEL ACCURACY IMPROVEMENT OPTIONS – CHOOSING DIFFERENT:
  - K
  - Max Depth
  - Hyparameter C

Conclusion

BASED ON HISTORICAL DATA FROM WEATHER CONDITIONS POINTING TO CERTAIN CLASSES, WE CAN CONCLUDE THAT PARTICULAR WEATHER CONDITIONS HAVE A SOMEWHAT IMPACT ON WHETHER OR NOT TRAVEL COULD RESULT IN

1. PROPERTY DAMAGE (CLASS 1) OR
2. INJURY (CLASS 2).