Severity of Accidents
Becky Gribben
10/15/2020

Introduction

Problem definition

Car accidents occur every day with a range of severity from minor property damage to fatalities. Many factors are involved including road conditions, weather, lighting conditions, human error, and speed just to touch on a few. Using machine learning data with Seattle's accident information we can bring to light some of these factors to educate drivers with.

Interest
Driver educators, insurance companies and municipalities can use this information to educate drivers. It could reduce the number of accidents and therefore the overall cost of accidents. It can assist municipalities in controlling congestion when certain conditions exist.

Data used is Seattle's car accident history (csv file) consisting of 194827 rows of data. With this data non relevant rows were removed. Utilizing more relevant data to gather specific information, see screenshot.

```
# Select wanted columns
df1=df1[['SEVERITYCODE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','ROADCOND','LIGHTCOND','WEATHER']]
df1.head(9)
```

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	ROADCOND	LIGHTCOND	WEATHER
221737	2	3	0	0	2	1	0	1
149856	2b	1	0	0	1	1	1	2
96238	1	3	0	0	2	1	0	2
95902	1	4	0	0	2	1	1	1
96271	1	3	0	0	2	1	1	1
96244	1	4	0	0	2	1	1	1
149863	2	4	0	0	3	1	0	1
149862	1	2	0	0	2	1	1	1
149860	3	4	0	0	2	1	1	1

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

```
#Testing basemodel with Logistic Regression

#Set test training
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7)

#function
def run_model(X_train, X_test, y_train, y_test):
    clf_base = LogisticRegression(C=1.0,penalty='l2',random_state=1,solver="newton-cg")
    clf_base.fit(X_train, y_train)
    return clf_base
 #executing basemodel
model = run_model(X_train, X_test, y_train, y_test)

#define work
def showresults(y_test, pred_y):
    conf_matrix = confusion_matrix(y_test, pred_y)
    plt.figure(figsize=(8, 8))
    sns.heatmap(conf_matrix, xticklabels=LABELS, yticklabels=LABELS, annot=True, fmt="d");
    plt.title("Confusion matrix")
    plt.ylabel('True SeverityCode')
    plt.xlabel('Predicted SeverityCode')
    plt.show()
    print (classification_report(y_test, pred_y))

pred_y = model.predict(X_test)
showresults(y_test, pred_y)
```

	precision	recall	f1-score	support
1	0.74	0.98	0.84	40121
2	0.74	0.22	0.34	17328
2b	0.00	0.00	0.00	905
3	0.00	0.00	0.00	95
micro avg	0.74	0.74	0.74	58449
macro avg	0.37	0.30	0.29	58449
weighted avg	0.72	0.74	0.68	58449

40,121 accidents are a severity one.

The results bring to light there are many factors and combinations of factors that could cause an accident. This data can be used to educate people on how to react but in a given situation a driver may not know all the factors surrounding them.

In conclusion it seems bringing data like this out to the public may reduce the number of fatalities. Even if you don't know all the factors you may know enough to prevent an accident from being worse. For example if it is raining and you slow down and still have an accident your slow down may change the outcome for future statistics.