Car Accident Severity Capstone Project (ACP)

Introduction

- where you discuss the business problem and who would be interested in this project.

Road accident is most unwanted thing to happen to a road user, though they happen quite often. The most unfortunate thing is that we do not learn from our mistakes on road. Most of the road users are quite 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.

Solutions that can help decrease accidents are categorized into three systems: active (often preventative measures, such as driver support or warning systems), passive (such as the roll cage in the vehicle) and behavioral (these include education, together with behaviors such as wearing a seatbelt).

The target audience of the capstone is the local government, police, rescue groups. The machine learning model and its results are going to provide some perspective for the target audience to make calculated decisions in reducing the number of accidents and injuries in the locality.

Data

- where you describe the data that will be used to solve the problem and the source of the data.

We will be using all collisions provided by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. The data will be first cleaned (i.e. all the unnecessary columns will be removed) in order for the data set to be fit for analysis. All of this will be discussed further in Week 2.

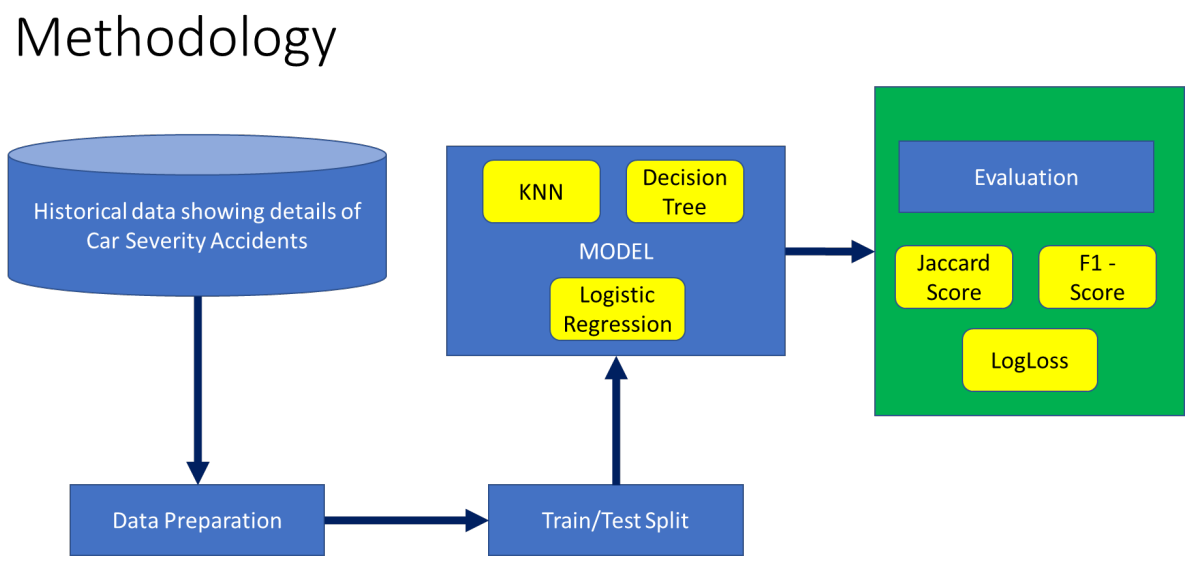
The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Little to no Probability (Clear Conditions)
1: Very Low Probability — Chance or Property Damage
2: Low Probability — Chance of Injury
3: Mild Probability — Chance of Serious Injury
4: High Probability — Chance of Fatality
Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

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.



Results

- section where you discuss the results.

We will be using the ff: Machine Learning Models

- K-Nearest Neighbor
    - KNN will assist us with anticipating the severity code of a result by finding the most like information point inside k separation.
- Decision Tree
    - A choice tree model gives us a format of every single imaginable result so we can completely examine the consequences of a choice. It setting, the choice tree watches all potential results of various climate conditions.
- Logistic Regression
    - Since our dataset just gives us two severity code results, our model will just anticipate one of those two classes. This makes our information twofold, which is flawless to use with calculated relapse.

	Jaccard-Similarity Score	F1 Score	LogLoss
KNN	0.564	0.540	N/A
Decision Tree	0.566	0.545	N/A
Logistic Regression	0.526	0.511	0.68495

Discussion

- section where you discuss any observations you noted and any recommendations you can make based on the results.
-

The objective of this project is to create a machine learning model in order to reduce the number of accidents in the locality. The target audience of the capstone is the local government, police, rescue groups. The machine learning model and its results are going to provide some perspective for the target audience to make calculated decisions in reducing the number of accidents and injuries in the locality.

The data will be first cleaned (i.e. all the unnecessary columns will be removed) for the data set to be fit for analysis.

We will be using the ff: Machine Learning Models: 1. K-Nearest Neighbor, 2. Decision Tree and 3. Logistic Regression

Assessment measurements used to test the precision of our models were jaccard list, f-1 score and logloss for calculated relapse. Picking distinctive k, max profundity and hyparameter C esteems assisted with improving our exactness to be the most ideal.

Conclusion

- section where you conclude the report.

It can therefore be concluded that an impact due to certain weather conditions that would result to injury – class 2 or property damage – class 1.

It can also be recommended that a complete data set would result into a more accurate prediction.