Predicting Severity of Car Accidents

The Value of Predicting Car Accidents

- The automobile has been around since 1885 and has been one the greatest inventions as it changed the lives of many in terms of travel. However the automobile has also change the lives of many for the worst. It is reported that on average 1.35 million people lose there lives to car accidents a year, that then totals to 3700 people losing their lives everyday on the road. In terms of accidents the biggest factor is the severity as it can be used to determine whether certain car characteristics put you at less or further risk. It could also be used to determine whether if car accidents are occurring less throughout the years and to go even determine certain levels of severity is increasing of decreasing. The main aim of this research is to evaluate and investigate the effect of risk factors on the injury severity of car accidents.

The Problem

- We will use a distinct data set on collisions useing machine learning principles, this project aim is to predict the level of severity of a accident. Data that might contribute to determining the severity of a collision location, address type, collsion type,person count, severity type, vechile count, date, weather conditions, road conditions, and lighting.

Why Would Someone Need This Data

- Obviously, Local and State Highway and Road Departments as well Traffic Departments would be interested in accurate prediction of car collsions, for the advantage and business values of making the roads safer. Others who are interested in this model would be car manufactuers to see if they can make there vechiles even safer.

Data Aqustition and Data Cleaning:

Data Sources

- The data utlized for this report was provided by IBM, it includes collsion data dating from 2004 all the way to the present.

Data Cleaning & Feature Selection

- The data I choose to drop included rows with missing or null values, as it would not benefit in predicting the model. Additonally Rows containing values that may be considered ambigous values.

- The feature selection I choose to work with that will predict severity includes: (Others omitted as it was not necessary for the prediction.)
  - ADDRESS TYPE (ADDRTYPE) : Collsion Location i.e Intersection, Block, Alley
  - VECHILE COUNT (VEHCOUNT) : # of Vechiles in Accident
  - LIGHT CONDITION (LIGHTCOND) : Type of Light ie Daylight, Nighttime
  - ROAD CONDITION (ROADCOND) : i.e Wet or Dry
  - WEATHER (WEATHER): I.E Snow, Rain, Sunny
  - Person Count (PersonCount) : # of people invoved in accident
  - Severity: Rated on a index between 1 and 2 with 2 being the most sever

3. Methodology:

Models and Results

- In this experiment, we constructed four binary classification models to see how well each of them can perform on the dataset we input (Figure 1). We focused on four types of model: Logistic Regression, Neural Networks, K Nearest Neighbors Classifier, and the Perceptron Learning Algorithm. (Lowest accuracy to highest accuracy)

Model	Accuracy (%)
Logistic Regression	64 %
Perceptron Learning Algorithm (PLA)	67 %
K Nearest Neighbors	67 %
Neural Networks	74 %

Logistic Regression

- Logistic Regression is a form of classification analysis when a dependent variable is considered binary. It is a form of predictive analysis, in which it can describe data and its relationship between a dependent binary variable. To create this model, I used the sklearn library and imported Logistic Regression. Then declared the y prediction equal to logistic predicting the input of x_testset. I was able to obtain an accuracy score of 64 %.

Perceptron Learning Algorithim (PLA)

- Perceptron is a supervised learning algorithm that is used to train inputs to determine which class they belong to. To create this model, I used the sklearn library and imported Perceptron. The iteration that was chosen was set 1000 times with an eta of 0.15. Then, we did predictions on X_test_std against the y_pred. I discovered that it was only 67 % accurate. However, when it was decreased to 10 iterations the accuracy rate decreased to 60%. Even when I change the number of iterations from 10000, 100000, and 1000000 it remained the same showing that when the number of iterations past 1000 increased accuracy score will remain the same. Proving that the max iterations for this model to get the best accuracy score is 1000. The eta in this model does not affect the accuracy score.

K Nearest Neighbors Classifier:

- The K Nearest Neighbors Classifier is useful as it can do both classification and regression, in this case, we choose the former. Input consists of the k closet training examples in a space, and output is dependent on it is classification or regression. As stated before we are focusing on classification so an object is classified by the vote of its neighbors and that same object being assigned to a class that is common among the nearest neighbors. For example, if k=1, then the object is simply assigned to the class of that single nearest neighbor. In this model, I manipulated the number of neighbors to get a better accuracy score. I was able to obtain an accuracy score of 67% with the number of neighbors being 1 if the number of neighbors was to increase the accuracy score became worse over time.

Neural Networks

- Neural Networks are a combination of algorithms that can recognize relationships in a set of data that can mimic the way the human brain operates. Neural Networks references a system of neurons that can be both organic and artificial. The Neural Network can adapt to a change in input, allowing for the network to generate the best result without having to redesign the output criteria. We used this library to train our training sets by predicting both x_trainset and x_testset. We found that it was 74 % accurate, making it the most accurate model out of the four we attempted.

Results:

Results

- Based on the four models I used in this experiment, I was able to get an average score of 68 %. Out of the four models, Neural Networks got the highest accuracy rate, while Logistic Regression performed the worst in terms of accuracy score. I would also like to add that the change in controllable parameters played no significant effect in terms of accuracy score for each model. The accuracy scores stayed with the same range and this could be due to many factors such as an imbalanced dataset where the models focused a certain main class than the other class.

Discussion/Conclusions:

- In this paper, I tried to create models to classify Car Accidents based on its severity score which range from 1 to 2 and other contributing features. I wasn't able to obtain the best results after performing models on the same dataset I can see how the amount of variable can play a factor in our dataset being imbalanced. This shows how the models that I created using Perceptron, Logistic Regression, K Nearest Neighbors, and Neural Networks could have possibly been influenced by the models focusing too much on a certain main class rather than the other classes.