CAR ACCIDENT SEVERITY PREDICTOR

**1. Introduction/Business Problem**

**Background**

Traffic accidents cause serious threat to the human life worldwide. In order to take necessary actions to control this ever-growing problem extensive research has been carried out into the prediction of traffic accidents in both developed and developing countries using various statistical techniques. A number of factors contribute to the risk of car accidents. Some of the important ones are road conditions, weather, type of junction and light conditions. These conditions also influence the severity of the accidents. In this paper, we aim to predict the severity of car accidents through machine learning models.

**Business Problem**  
There are several factors that can contribute to an accident. In this study, I would delve into the three main factors namely the weather condition, the road condition and the light conditions during the collision to predict the severity of car accidents so that drivers of vehicles would be given a notice as to the best conditions to drive in.

**Interest**  
The SDOT Traffic Management Division can use a car accident severity predictor in making policies on road safety. This makes them a good target audience for this project.

The results of the project would be particularly helpful for them to plan and improve the road conditions in the city and placement of street lights. They would also be curious to know the influence of other factors like weather and visibility conditions on the occurrence of accidents and the level of severity. Also, the model can be generalized to other areas and cities aside Seattle helping drivers to understand the conditions to which driving can be risky and take precaution. This would go a long way to prevent accidents from occurring.

**2. Data**

The dataset used in this project includes all collisions from Seattle from 2004 to present. This dataset is provided by SPD and recorded by Traffic Records. The dataset has 38 columns and 194,673 entries. The target variable from the dataset is “SEVERITYCODE”. It is made up of numbers, which correspond to different levels of severity:

0: no/negligible chance  
1: Chance of car damage  
2: Chance of injury and/or car damage  
3: Chance of serious injury and car damage

4: Fatal and car damage

Out of the 37 independent features, the ones that would be of interest in this study are the following:

“WEATHER”: A description of the weather conditions during the time of the collision.

“ROADCOND”: The condition of the road during the collision.

“LIGHTCOND”: The light conditions during the collision.

These attributes were selected because they directly affect the severity of road accidents and they also had a large number of entries that can make a reasonable dataset unlike the other independent features.

**2.1 Data preprocessing**

The first step in the data preprocessing phase is to create a dataframe with the columns that are relevant to the study. After this, all null values from the dataset are removed.

Table

Description automatically generated

The current data type of the feature columns being an object would not work well with some machine learning models. So, we convert them to strings and then perform label encoding on the features to get a set of features in a format that can be used by machine learning models.

Table

Description automatically generated

**2.2 Balancing the dataset**

**Chart, bar chart

Description automatically generated**

From the count of values above we see that the dataset is unbalanced, and this can affect how our model works. To balance this dataset, we need to make sure that the data pertaining to both severity codes are equal. To do this, data with severity code of 1 is down sampled to be equal to that of entries with severity code of 2.

Chart, bar chart

Description automatically generated

**3. Methodology**

The data has been preprocessed and balanced, and it is time to train the data. The data has predetermined classes for severity codes hence it is a typical example of a classification problem. For this study, we would use multiple classification models. The models that we would use in this study are:

1. K Nearest Neighbors
2. Support Vector Machines
3. Decision Trees

**3.1 Selecting Features and Target Variables**

For this project, the feature set is the weather, the road conditions and the light conditions for all entries and the target variable is the severity code.

**3.2 Normalizing the feature set**

The next step is to normalize the feature vector into a scale range normally between -3 to 3. This is important because it prevents any bias that might exist in the model when training due to extremely high or extremely low values.

**3.3 Split data into training and test sets**

It is good practice in machine learning to split the dataset into train and test sets to help in evaluating the models effectively using unseen data and prevent overfitting as well as get a more generalized model. 75% of the dataset was used as the training set and 25% was used for testing.

**3.4 K Nearest Neighbors (KNN)**

The K Nearest Neighbors model is one whose efficiency depends very much on the value k selected for the model. Taking this into consideration, I built several models with k values ranging from 1 to 20 and chose the one with the best accuracy score. This occurred for the KNN model with a k value of 17.

Graphical user interface, text, application

Description automatically generated

**3.5 Support Vector Machines**

For support vector machines, the efficiency of the model depends on the type of kernel that is selected. A general rule of thumb is that a linear kernel is used for linear problems and the radial basis function kernel for non-linear problems. Since the problem is non-linear in nature, the radial basis function kernel was used.

Graphical user interface, text, application, email

Description automatically generated

**3.6 Decision Trees**

A decision tree of max depth equal to 10 and criterion equal to entropy was built and trained on the training set. Different values of the depth were experimented with and the model corresponding to the highest accuracy was chosen.

Graphical user interface, text, application, Word

Description automatically generated

**4. Results**

To know the right model to use for actual deployment or production we need to evaluate the model using some evaluation metrics. Since this is a classification problem, the evaluation metrics used for this project are the Jaccard similarity score and the F1 score. The results of evaluation for the different approaches are shown in this section.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**5. Discussion**

For this project, an accuracy of 55.6% was achieved by the SVM model. There is however a significant variance that could not be predicted by the models in this study. I believe features like speed or uninterrupted time of traveling could be used to predict a more accurate classification. These are characteristics that may be impossible to know right now, but at the incredible pace that technology is evolving nowadays, soon cars will be able to track them so that the emergency services could use them.

The next step on this problem could be to add an accident prediction model able to not just predict the accuracy but also the critical time and spots where potential accidents can occur in advance.

**6. Conclusion**

In this study, I analyzed the relationship between severity of car accidents and some characteristics which describe the situation that involved the accident. I considered features such as weather condition, the road condition and the light conditions during the collision. I built and compared 3 different classification models to predict whether an accident would have a high or low severity. These models can have multiple application in real life. For instance, imagine that emergency services have an application with some default features such as date, time and department/municipality and then with the information given by the witness calling to inform on the accident they could predict the severity of the accident before getting there and so alert nearby hospitals and prepare with the necessary equipment and staff. Also, by identifying the features that favor the most the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.