IBM Applied Data Science Capstone

PREDICTING CAR ACCIDENT SEVERITY REPORT

Table of Contents

Table of Figures

Introduction (Problem and Background Discussion)

The number of motor vehicles have increased immensely over the years, as have the resulting car injuries and deaths per year rates. 1.35 million people die in road accidents worldwide every year that is 3,700 deaths a day. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities [1]. Also, fatal car wrecks and road accidents cost countries worldwide about 3% of their GDP [2].

There are several factors that influence the occurrence and severity of car accidents. These can be broadly categorized into behavioral, environmental and situational factors. Behavioral factors include careless driving, speeding and driving under the influence of drugs and alcohol. In fact, distracted driving is a leading cause of accidents, causing 25-50% of all crashes. Environmental factors are related to road, weather and visibility conditions. Situational factors include sudden malfunctioning of the vehicles.

Majority of accidents can be prevented as most result from poor and reckless driving and increase in specific environmental conditions. What if we can relate the occurrence and severity of car accidents to specific conditions such as road, weather and visibility conditions? Wouldn’t we be able to alert drivers to be more careful and hence increase their attentive driving? Wouldn’t such info also help healthcare institutions and governments such as the police and transportation departments by allowing them to maintain sufficient staff and alertness to adequately address any upcoming and unfortunate accidents?

Due to the availability of large amount of accident related data generated by different government bodies, it would be possible to build machine learning models to predict the occurrence and severity of car accidents given specific conditions. This would eventually help in both reducing the occurrence and severity of such incidents.

Many institutions including local governments, police departments, healthcare institutions, rescue groups, insurance companies may benefit from such predictive models.

Description of the Data

For this assignment, collision data collected by the Seattle Police Department (SPD) and recorded by the Traffic Records team will be used. The data was downloaded from Seattle's government site (link provided below). "Seattle City GIS Program" is the data owner.

https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

The data was last updated on September 22, 2020 and covers 214,050 collision records spanning from October 2003 to September 2020.

The data has 40 attributes including the dependent variable "SEVERITYCODE" which represents the severity of the collision represented by the following codes and their respective severity:

- 3 for fatality
- 2b for serious injury
- 2 for injury
- 1 for prop damage
- 0 for unknown

Following "Data Understanding", only a subset of the attributes that may be related to the occurrence and severity of the collisions will be selected in this study and the rest non-relevant attributes or variables will be ignored. The following attributes of interest will initially be considered to train the model. As part of the "Data Preparation", balancing unbalance data and cleaning the dataset will be performed if required.

Attribute	**Description**
ADDRTYPE	Collision address type: Alley, Block, or Intersection
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision
PEDCYLCOUNT	The number of bicycles involved in the collision
VEHCOUNT	The number of vehicles involved in the collision
INJURIES	The number of total injuries in the collision
FATALITIES	The number of fatalities in the collision
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
INATTENTIONIND	Whether or not collision was due to inattention (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
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
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N)

The following machine learning classification algorithms will be employed to build the prediction model:

- K Nearest Neighbors (KNN)
- Decision Tree
- Logistic Regression
- Support Vector Machine (SVM)

Approach to Tackle the Problem

Supervised machine learning models were built and evaluated. K-Nearest Neighbors, Decision Tree, Support Vector Machines and Logistic Regression were implemented.

Cross-Industry standard process for Data Mining (CRISP-DM) methodology was followed throughout the process as indicated in figure 1 below.

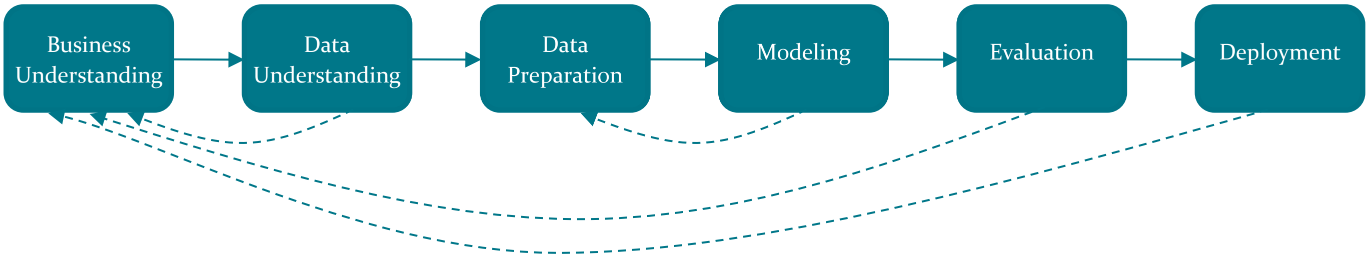


Figure 1 Cross-Industry standard process for Data Mining (CRISP-DM)

Data Visualization

Comprehensive data visualization was conducted on the dataset. Data was visualized to check whether there was any correlation between the number of accidents and the time of the day, week of the day, month of the year and whether there was any trend over the years. One of the observations was that a very high percentage of the accidents occur around midnight and as expected during rush hours as shown in figure 2.

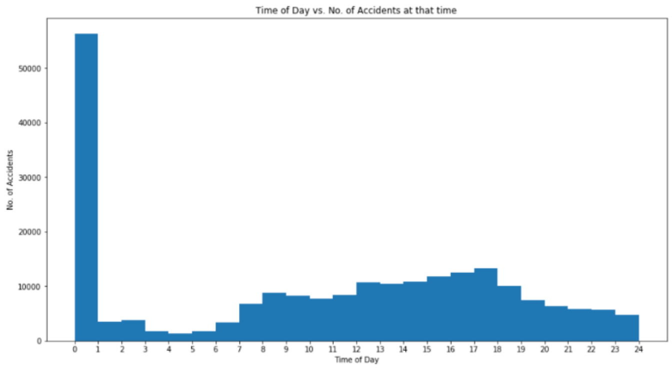


Figure 2 Number of Accidents at Different Times of the Day

Data was also visualized to plot the number of vehicles involved in accidents. As figure 3 shows, in majority of accidents 2-4 vehicles were involved.

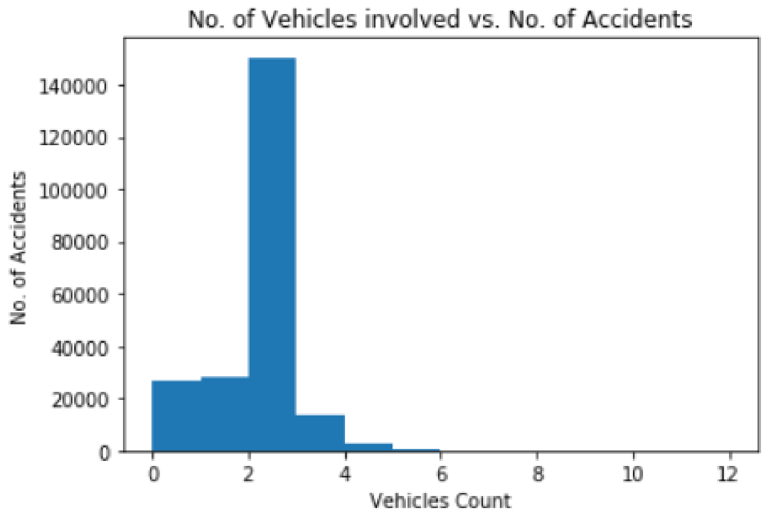


Figure 3 Number of Vehicles Involved in Car Accidents

The number and severity of collisions whilst driving under the influence of alcohol and drugs was also plotted. However, the volume and the severity of the accidents were low as indicated in figure 4.

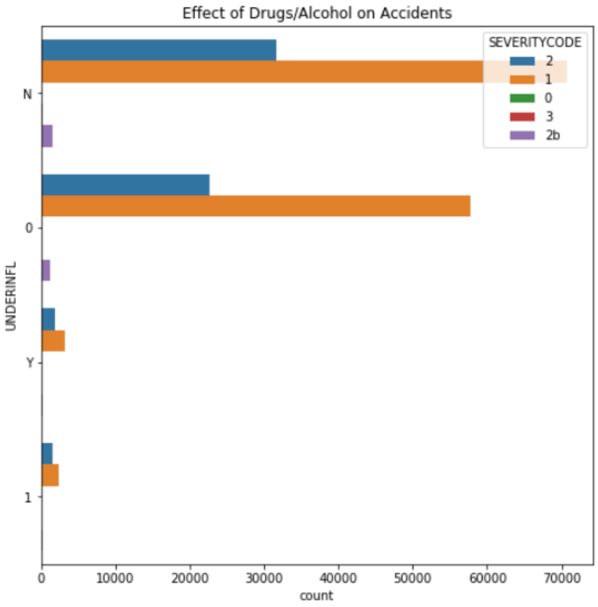


Figure 4 Effect of Driving Under Influence of Drugs and Alcohol on Accidents

The number and severity of collisions were also plotted against the weather conditions. Interestingly, accident volume and severity were not hugely impacted by weather conditions as show in figure 5.

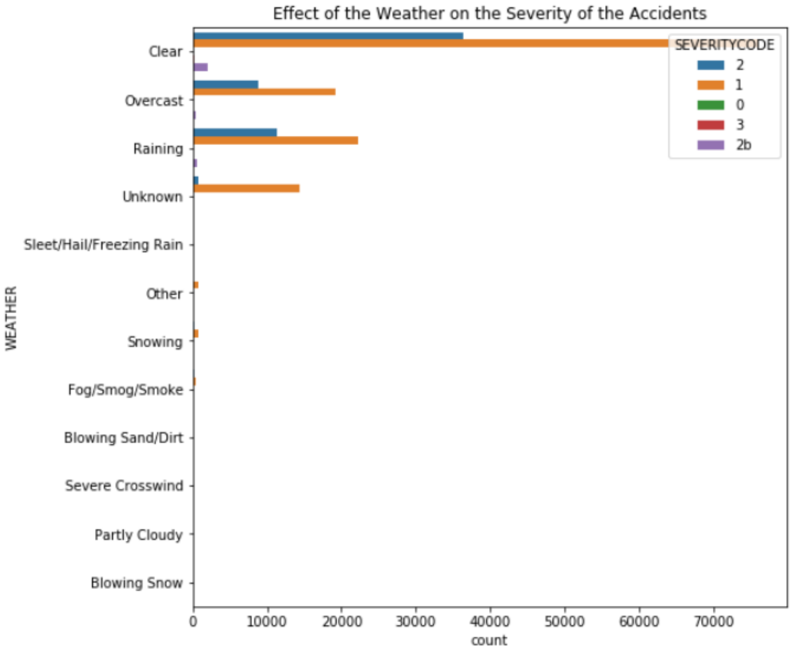


Figure 5 Effect of Weather Conditions on Accidents

Data Preparation and Cleaning

Data cleaning was conducted. To do so, the unique values of the variables were looked at. Several variables had unique values as “Unknown”. All records that has “Unknown” as a value in any of the variables were removed from the data set. These included variables such as “WEATHER”, “LIGHTCOND”, “ROADCOND” and “JUNCTIONTYPE”.

Inconsistencies in data values were also taken care of. For example, “UNDERINFL” variable had 0, 1, Y, and N unique values. Y and N were set as 1 and 0 respectively. Similar unique values of variables such as “LIGHTCOND” were combined.

Missing data in different variables was looked at thoroughly. It was assumed that all the missing data in the “'UNDERINFL'”, “INATTENTIONIND” and “SPEEDING” were supposed to reflect a negation, hence the respective missing data were replaced with the numeric value of zero. The rest of records with missing values or NA were all dropped from the dataset.

The dataset was checked and it was observed that the data related to the attribute SEVERITYCODE was imbalanced. Records with code “1” was over double the rest of all other codes. As this would influence the predictive model, balancing of the dataset was conducted by down sampling of the majority class.

Binary Encoding and One Hot Encoding of the relevant data was implemented as algorithms don’t work with categorical data. Binary encoding to attributes such as “SEVERITYCODE” and “HITPARKEDCAR” were conducted. One Hot Encoding process was conducted to categorical attributes such as “ADDRTYPE”, “WEATHER”, ROADCOND” and “LIGHTCOND” so the machine learning algorithms do a better job in prediction

It was important to normalize the dataset and rescale the values of the attributes to the range of 0 to 1. This helps avoid biases in the machine learning algorithm.

Modeling and Evaluation

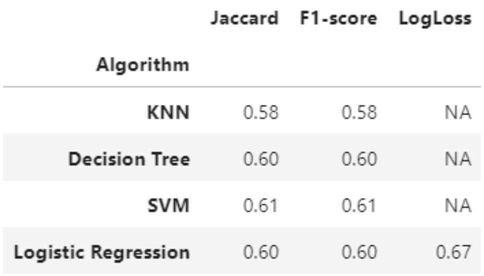
Before training the model, the dataset was split into training and test sets with a split ratio of 0.8/0.2. This helps in evaluating the models effectively using unseen data and hence helps in preventing of overfitting.

K-Nearest Neighbors (K-NN), Decision Tree Classifier, SVM classifier, and Logistic Regression models were used for training. The trained models were evaluated and reported by using the Jaccard Score, F1 Score and the Log Loss performance metrics.

Below are the accuracy scores for the 4 models that were trained.

- K-Nearest Neighbors (KNN): Accuracy score of 0.57
- Decision Tree: Accuracy of 0.6
- Support Vector Machine: Accuracy of 0.6
- Logistic Regression: Accuracy of 0.6

The following table summarizes the results. Based on these results, all the four models have similar performance.



	Jaccard	F1-score	LogLoss
Algorithm			
KNN	0.58	0.58	NA
Decision Tree	0.60	0.60	NA
SVM	0.61	0.61	NA
Logistic Regression	0.60	0.60	0.67

Figure 6 Performance Metrics of the Machine Learning Models

Conclusion

In this project, collision dataset collected by the Seattle Police Department (SPD) and recorded by the Traffic Records team was analysed and used to predict the severity accidents by employing supervised machine learning models. The dataset was first cleaned and then processed and finally modeled by using four supervised machine learning models: K-Nearest Neighbors, Decision Trees, Support vector Machine and Logistic Regression.

There were couple of interesting observations following the analysis of the dataset. Very high percentage of the accidents occurred around midnight. Majority of accidents involved 2-4 vehicles. Number and severity of collisions whilst driving under the influence of alcohol and drugs were low. Number and severity of collisions were not hugely impacted by weather conditions. Further detailed analysis would be required to dig into these observations to provide recommendations that would eventually reduce the occurrence and severity of accidents.

All four of the machine learning models that were built to predict the severity of car accidents had good accuracy of prediction close to 0.6 with good performance metrics. It's worth to mention that some key assumptions were made related to the data. Missing data in different variables was looked at thoroughly. It was assumed that all the Missing data in the "'UNDERINFL'", "INATTENTIONIND" and "SPEEDING" variables all considered to reflect a negation and were treated accordingly.

Although there is room to improve the accuracy predicting the severity of accidents, the models that were built provided a fair prediction accuracy in the context of this study.

References

[1] https://www.asirt.org/safe-travel/road-safety-facts/

[2] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries