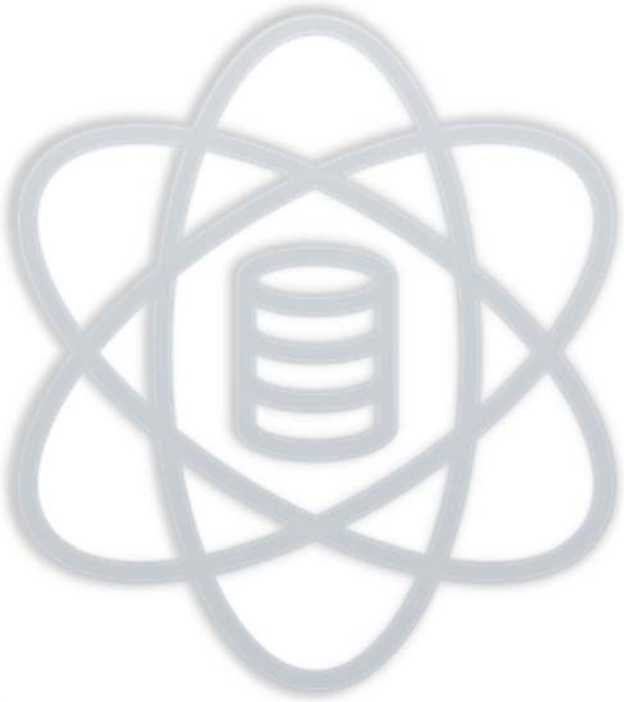
# PREDICTING CAR ACCIDENT SEVERITY

## Machine learning approaches to the problem



**Kauvin Lucas**

September 2020

## Introduction

Car road accidents are the number one problem in US road transportation, accounting for 99% of transportation injuries and representing an economic cost of USD 150 billion annually [1]. Deaths and injuries from the accidents are some of the main issues concerned by insurance companies, healthcare institutions and governments.

Due to the large amount of accident reports generated by the city and state government officials, it's possible to identify predictive patterns from the data that will ultimately aid in the decision making of the concerned institutions.

This project will use a sample dataset from the Seattle Department of Transportation (SWOT) to employ classification algorithms to predict accident severity rates in seaport city of Seattle and compare the accuracy of each built model by using several evaluation approaches.

## Business problem

The seaport city of Seattle is one of the busiest cities in the Washington state, and cars accidents are very common in the place. The city administrators are making an effort to improve the road conditions, but serious injuries from the decreasing accidents rate are still a concern. The city has a goal of eliminating all traffic-related deaths by the year of 2030, and city officials are employing a data-driven approach to accomplish the goal.

Those who survive in a car accident can face hefty medical bills and thousand dollars in property damage. Discovering patterns in the data and making predictions may aid the decision-making processes of healthcare and insurance companies.

## Analytic approach

Predicting car accidents severity is one of the biggest challenges faced by many actors. No single factor can help explain the severity of an accident, and the relationship between these factors, although intuitively positive, are mostly unclear from a statistical point of view.

But while each accident may be unique, accumulating insights from each accident may show macro trends and thus allow us to make accurate predictions. Taking these several seemly uncorrelated variables into a single accurate prediction model is a long process that requires finding generalizable predictive patterns. That can only be achieved by using machine learning algorithms.

Machine learning classification approaches and data science methodology will be employed to achieve the goal of this project. The following classification techniques will be used to predict and evaluate the model:

1. K Nearest Neighbors (KNN)
2. Decision Tree
3. Logistic Regression
4. Support Vector Machine (SVM)

## Data collection

The data used was provided by the Washington State Department of Transportation (WSDOT) in a csv file. It's a public data that describes the accidents occurred between 2004 and 2020 in the city of Seattle. The information contained in the data goes through a month-long process that involves city and state transportation officials reviewing, comparing and analyzing reports from the Seattle Police Department.

## Data understanding

The data had 37 features and 194,673 rows. We are going to predict the severity code of the accidents, which is labeled in the dataset as numbers between 0 and 3, from least to more severe:

- 0: unknown
- 1: property damage only
- 2: injury
- 2b: serious injury
- 3: fatality

But when analyzing this feature, it only has two categories of data: property damage only and injury, so it follows a binomial distribution.

Before the Feature Selection stage, one more feature was added: "ISHOLIDAY". The data was reduced to include the following features before next step (Table 1):

**Table 1 – Features selected**

| Feature name | Description | Comments |
|---|---|---|
| SEVERITYCODE | A code that corresponds to the severity of the collision: | This will be the output variable |
| ADDRTYPE | A description of the collision address type (intersection/alley/block) | Crash type may affect collision severity |

| | | |
|---|---|---|
| INATTENTIONIND | Whether the person was not paying attention | Inattention and distraction may affect collision severity |
| UNDERINFL | Whether the person was driving under the influence of alcohol | Alcohol or drug impaired drivers may cause more severe collisions |
| WEATHER | A description of the weather conditions during the time of the collision | Poor weather conditions may affect driver visibility, and lead to more severe accidents |
| ROADCOND | The condition of the road during the collision | Poor road conditions may affect collision severity |
| LIGHTCOND | The light conditions during the collision | Poor road light conditions may increase accident frequency and fatality |
| SPEEDING | Whether or not speeding was a factor in the collision | Speeding vehicles may represent greater risk of huge material damage and bigger number of injuries |
| ISHOLIDAY | Whether or not collision has occurred in holiday | It's known that holidays have an impact in the accident frequency and severity numbers |

## Data preparation

Before moving to the data modelling stage, missing values were dropped from the data. Some of the missing were assumed to contain a certain category of answers, so these missing values were replaced and put to the corresponding category before dropping any remaining rows with missing data. Also, some duplicated values in the column "LIGHTCOND" were merged.

Each feature was encoded by using one hot encoding for the variables having some order associated with them and binary encoding for the rest.

## Methodology

A supervised, classification machine learning approach was applied to the data, since the output variable contains categorical data. The following classification models were used and evaluated by using Jaccard's score, F1 score and log loss:

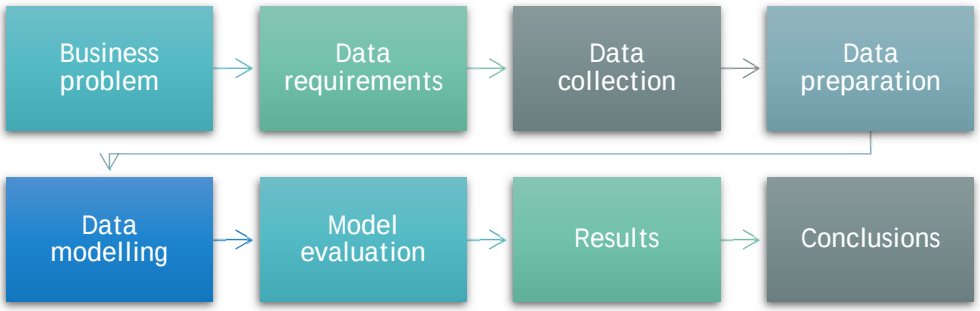
- K Nearest Neighbors
- Decision Tree
- Support Vector Machine
- Logistic Regression

It was noticed that the input data was too large to be processed, specially to the k-nearest neighbors algorithm. The data was then reduced to a much smaller, random sample of 15,000 before transforming the data into set of features.

The data was normalized to improve the results of lazy learning algorithms. Each trained model was evaluated for subset accuracy.

In summary, the methodology used in this project to model and predict the data is represented by the following graph:

**Image 1 – The project methodology**




## Results

The choice of k for the k-nearest neighbors algorithm largely depends on the data, and it may change on each evaluation because of the randomly sampled data. To always choose the best k, a heuristic technique called hyperparameter optimization was employed. The training has returned an accuracy score of 65,82% for the training set, and 65,47% for the test set.

The decision tree model had an accuracy score of 65,93%, while the support vector machine had an accuracy of 66,23% and logistic regression an 61,93%.

The models were evaluated and reported by using Jaccard score, F1 score and log loss. The results are displayed in the table 2.

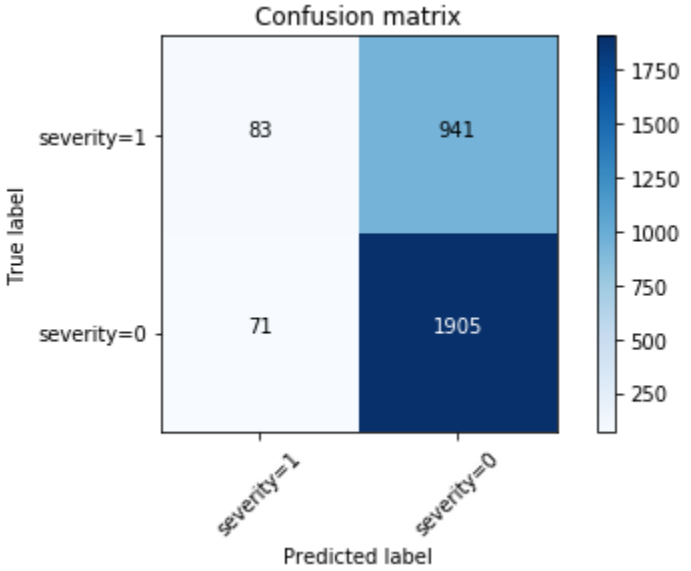
**Table 2 – final report**

| Algorithm | Jaccard | F1-score | LogLoss |
|---|---|---|---|
| **KNN** | 0.65 | 0.61 | NA |
| **Decision Tree** | 0.65 | 0.54 | NA |
| **SVM** | 0.66 | 0.53 | NA |
| **LogisticRegression** | 0.66 | 0.61 | 0.63 |

By the scores shown in the table above, the logistic regression model is the one that returns a higher accuracy scores compared to the rest. This is expected, since the algorithm is best suited for estimating a logistic model of a variable represented by one or zero.

According to the confusion matrix (image 2), which was calculated to evaluate the logistic regression classifier, has correctly predicted a property damage only accident in 1,905 of 2,846, with a noticeable presence of type I errors.

**Image 2 – Confusion matrix**




## Conclusions

In this project, we analyzed the dataset of car accident severity to predict the severity of the accident by employing supervised machine learning models. After carefully preprocessing the dataset, which included sampling and feature selection and extraction, the data was modeled by using 4 models: k-nearest neighbors, decision trees, support vector machine and logistic regression.

The outcome variable (severity code) can take 6 different categories, but it was treated as a binary dependent variable since it only has two categories (property damage only and injuries). It's good to remind that this variable can take more than 2 categories, but the model most suitable for predicting the car accident severity rates with the current dataset is the logistic regression.

## References

[1] NHTSA 2020 Report - https://one.nhtsa.gov/nhtsa/whatis/planning/2020Report/2020report.html