Minimizing Risk Of Car Accidents

Geoffrey Gacuca

1. Introduction

1.1 Background

Road Safety is an incredibly important issue that is undervalued in society today. Almost 40,000 Americans die each year due to car accidents and according to the Center for Disease Control nearly 1.35 million people around the world die due to car and road accidents each year. That is roughly 3,700 people per day. This is clearly an enormous issue that not only affects individuals, but can also hamper economies and create significant damage if proper precautions are not taken. Understanding where, when and how Road Traffic Accidents (RTAs) happen is a game changer. It enables communities to properly understand how to deal with the common causes and issues that can happen as a result of car accidents and create the necessary rules, regulations and infrastructure to prevent future cases from reoccurring.

1.2 Problem

There are multiple factors that contribute to RTAs and because of the large amount of information on accidents it can be difficult to determine at times which data is important and has a strong relationship to the severity of a car accident and which data is simply extra information. The objective of this project is to look at many of these factors such as location, road conditions, weather, speeding, number of people and determine which are key to discovering the severity of an accident.

1.3 Interest

Being able to precisely determine the causes of RTAs is of great interest to the local, state and federal governments Departments of Transportation (DOT) that are tasked with understanding the causation of RTAs and providing solutions to those issues. Whether it is developing new safety signs or creating driver awareness when it comes to certain RTA causes such as texting while driving or drunk driving. They would benefit from data driven information on road accidents.

2. Data Acquisition & Cleaning

2.1 Data Sources

The dataset used in this project comes from the Seattle Department of Transportation and provides RTA data from 1st January 2004 and 20th May 2020. The dataset has a total of 194,673 collision incidents with 37 different features such as COLLISIONTYPE , PERSONCOUNT, WEATHER, ROADCOND, LIGHTCOND and more.The data did have some missing columns and some of the information presented did not appear on the metadata.For instance , the SEVERITYCODE column only had three values out of a total of five possible values according to metadata. This was less than ideal, because the SEVERITYCODE is a very key element to finding out the severity of an accident. It was still manageable to deal with and there was still plenty of data to dissect and pull useful information from. Also there are columns listed in the metadata that are not provided in actual dataset, which became a constraint when I began to examine the dataset.

2.2 Data Wrangling

The data was analyzed and cleansed based on the requirements for this project. I checked for duplicate rows. There are no duplicate rows, however there was a duplicate column, SEVERITYCODE.1 was the same information as SEVERITYCODE so i dropped it from the dataset.

Some fields need to be cleaned up so that Y, N become 0s and 1s. I used one hot encoding to transform several variables such as WEATHER, LIGHTCOND & ROADCOND. I removed or did not use the following columns for reasons listed below:

- OBJECTID & SHAPE - while these are a unique identifier for the dataset, it does not provide valuable information for analysis
- INCKEY, COLDETKEY & COLLISIONTYPE - keys that represents each collision for each incident, however the types of collisions was not my focus, I was more interested in how dangerous the collisions were
- INCDATE, JUNCTIONTYPE, INCDTTM - provided the date & time of each incident, yet my focus was not on when the incident occurred but on the impact of the incidents
- SDOT_COLDESC,ST_COLDESC, SEGLANEKEY, CROSSWALKKEY - a code used by the local government representing the collision and its description, however the code covered a very broad spectrum of accidents and issues that went beyond the scope of what I was looking for

2.3 Feature Selection

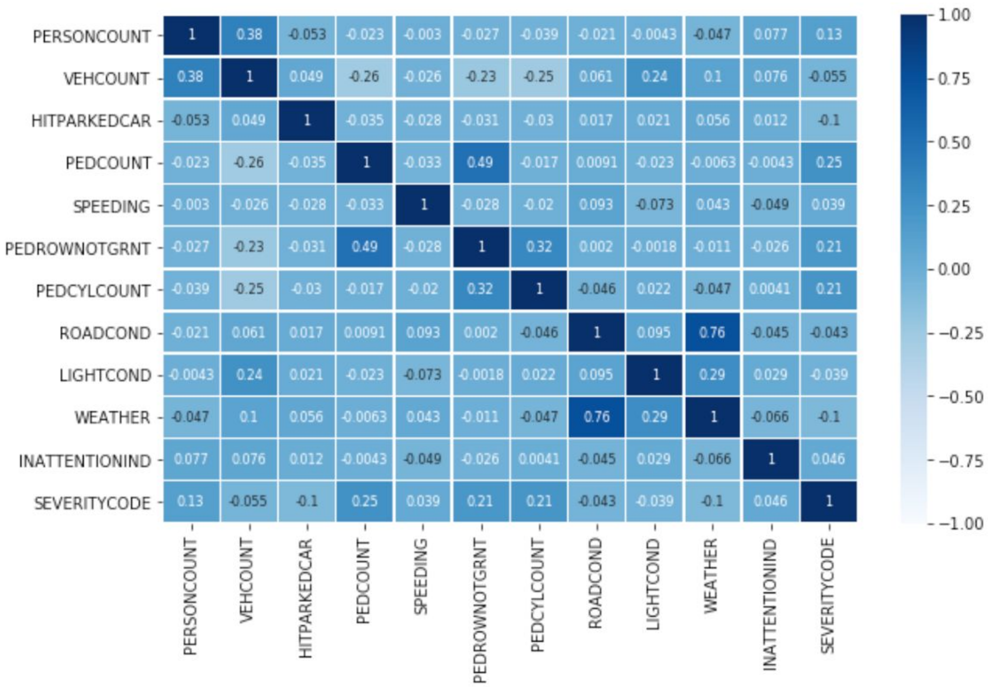
The following columns were chosen as features:

- LOCATION (X,Y)
- PERSONCOUNT
- VEHCOUNT
- HITPARKEDCAR
- PEDCOUNT
- UNDERINFL
- SPEEDING
- PEDROWNOTGRNT
- PEDCYLCOUNT
- ROADCOND
- LIGHTCOND
- WEATHER
- INATTENTIONIND
- SEVERITYCODE

3. Exploratory Data Analysis

3.1 Relationships between variables

After going through the process of data wrangling and feature selection, I produced a hotmap to visualize the relationship between variable and see which variables had a strong correlation and would be beneficial to continue analyzing. Weather (WEATHER) & Road Conditions (ROADCOND) showed a strong positive correlation (0.76). As well as the number of pedestrians (PEDCOUNT) involved in accidents and those pedestrians not being given right of way during a traffic accident(PEDROWNOTGRNT) (0.49).



	PERSONCOUNT	VEHCOUNT	HITPARKEDCAR	PEDCOUNT	SPEEDING	PEDROWNOTGRNT	PEDCYLCOUNT	ROADCOND	LIGHTCOND	WEATHER	INATTENTIONIND	SEVERITYCODE
PERSONCOUNT	1	0.38	-0.053	-0.023	-0.003	-0.027	-0.039	-0.021	-0.0043	-0.047	0.077	0.13
VEHCOUNT	0.38	1	0.049	-0.26	-0.026	-0.23	-0.25	0.061	0.24	0.1	0.076	-0.055
HITPARKEDCAR	-0.053	0.049	1	-0.035	-0.028	-0.031	-0.03	0.017	0.021	0.056	0.012	-0.1
PEDCOUNT	-0.023	-0.26	-0.035	1	-0.033	0.49	-0.017	0.0091	-0.023	-0.0063	-0.0043	0.25
SPEEDING	-0.003	-0.026	-0.028	-0.033	1	-0.028	-0.02	0.093	-0.073	0.043	-0.049	0.039
PEDROWNOTGRNT	-0.027	-0.23	-0.031	0.49	-0.028	1	0.32	0.002	-0.0018	-0.011	-0.026	0.21
PEDCYLCOUNT	-0.039	-0.25	-0.03	-0.017	-0.02	0.32	1	-0.046	0.022	-0.047	0.0041	0.21
ROADCOND	-0.021	0.061	0.017	0.0091	0.093	0.002	-0.046	1	0.095	0.76	-0.045	-0.043
LIGHTCOND	-0.0043	0.24	0.021	-0.023	-0.073	-0.0018	0.022	0.095	1	0.29	0.029	-0.039
WEATHER	-0.047	0.1	0.056	-0.0063	0.043	-0.011	-0.047	0.76	0.29	1	-0.066	-0.1
INATTENTIONIND	0.077	0.076	0.012	-0.0043	-0.049	-0.026	0.0041	-0.045	0.029	-0.066	1	0.046
SEVERITYCODE	0.13	-0.055	-0.1	0.25	0.039	0.21	0.21	-0.043	-0.039	-0.1	0.046	1

Fig. 1 is a visualization of the variables chosen for feature selection and their correlation to one another. Darker color indicates a stronger correlation.

4. Predictive Modeling

The models used to predict the level of car accident severity are K- Nearest Neighbor (KNN), Decision Trees (DT) and Logistic Regression (LR). Each model has its advantages and disadvantages in predicting the severity of car accident likelihood. KNN is a classification algorithm that can do a good job of predicting the severity of a car accident based on using the features and best K number to take data from a number of variables and then predicting the severity of the accident. Logistic Regression is another classification algorithm that can examine the relationship between the level of car accident severity (SEVERITYCODE) and compare it to the other variables listed in feature selection

4.1 K - Nearest Neighbors

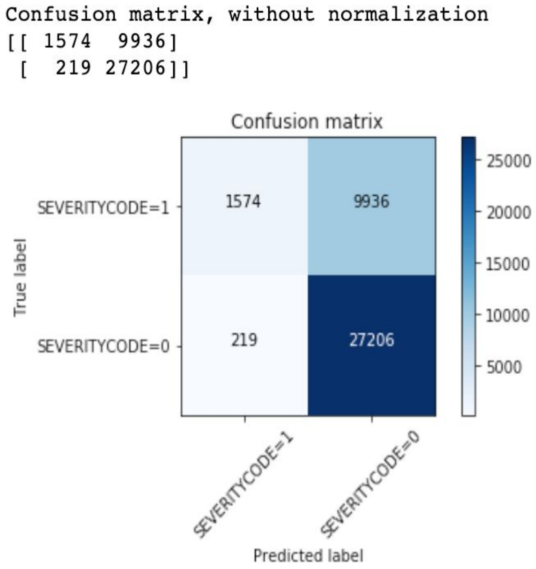
Based on the results of the model, the best K for KNN is K = 8, which was found to have an accuracy of 72.12% . This means the model looks at the 8 nearest pieces of information in the dataset in order to accurately determine the level of severity of the accident. The model is decently accurate in being able to determine severity, with an F-1 score of 67.9%.

4.2 Decision Trees

Based on the results of the model, decision trees also have relatively decent accuracy of 73.9% when it comes to accurately predicting the severity level of a car accident, however while the model boasts a higher accuracy level than KNN, it has a lower F-1 score of 66.3%.

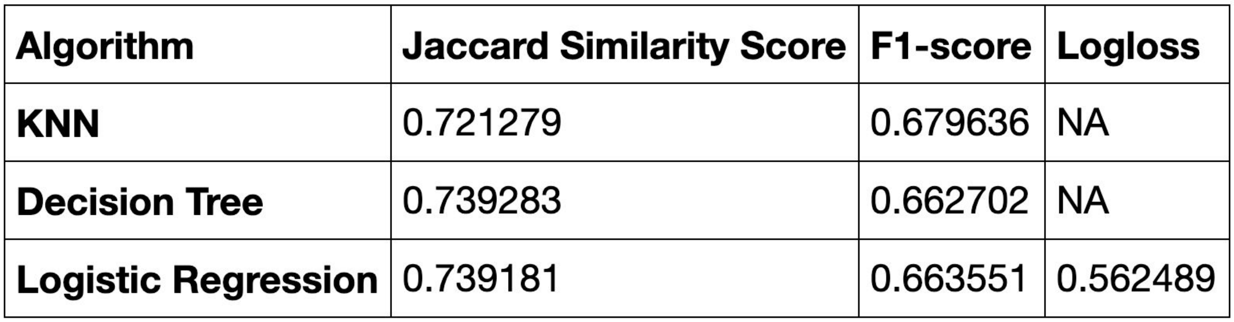
4.3 Logistic Regression

LR performs very similarly to DT achieving an accuracy level of 73.9% and an F-1 score of 66.4% and a LogLoss of 56.2%. The confusion matrix below shows the two possible predicted classes as well as the true and false positives and negatives of the matrix.



```
Confusion matrix, without normalization
[[ 1574  9936]
 [  219 27206]]
```

Fig. 2 is a visualization of the Confusion Matrix



Algorithm	Jaccard Similarity Score	F1-score	Logloss
KNN	0.721279	0.679636	NA
Decision Tree	0.739283	0.662702	NA
Logistic Regression	0.739181	0.663551	0.562489

Fig. 3 shows the accuracy scores of the three models used for this project

5. Conclusions

In this study, I analyzed the relationship between the level of severity in car accidents compared to several features such as weather, light conditions & road conditions. I identified weather and road conditions to be two of the most important features that affect the severity level of car accidents. I built three classification models, KNN, DT & LR to analyze and predict how greatly features can impact car accidents. These models can be very helpful to DOTs in helping them determine where they need to put resources in order to minimize car accidents. For example, road conditions were a key factor in many accidents. DOTs could use this information to lobby for more resources to be put towards regularly maintaining roads so that they are in good condition in order to minimize the costs associated with car accidents.

6. Future Directions

Models in this study mainly focused on how the drivers and pedestrians environment affected the severity level of accidents. However,