Predicting Severe Traffic Outcomes: Recommendations for Proactive Traffic Accident Mitigation

Stefan Banducci
September 2020

I. Introduction / Business Problem

Road traffic accidents are a serious problem in the United States and worldwide. The most severe traffic accidents each year are responsible for serious injuries as well as loss of life. According to the National Safety Council, in 2019, an estimated 38,800 people lost their lives to car crashes in the United States. In the same year, an additional ~4.4 million people in the US were injured seriously enough to require medical attention in crashes. The cost associated with a traffic accident death is often immeasurable - a lost love one, a child who grows up without a loving mom or dad. Additionally, those that suffer serious injuries in traffic accidents often bear significant costs as well. Traffic-accident-driven serious injuries can lead to job losses, financial hardships, as well as psychological impacts for both the injured party as well as his/her family. Given the high societal, family, and individual costs associated with accidents, addressing traffic accident fatalities and serious injuries represents a significant opportunity.

At the highest level, severe traffic accidents can be addressed proactively and/or reactively. A proactive approach involves trying to stop severe accidents from happening before they occur. This approach necessarily involves a solid understanding of the variables that contribute to severe accident outcomes. Knowing what variables drive severe accident outcomes (e.g., speeding, drunk driving, weather, etc.) can enable development of specific strategies for accident prevention. For example, if speeding is found to be a key driver of severe traffic accidents in an area, this understanding can inform appropriate speed limits and enforcement strategies.

In addition to proactive strategies, it's also possible to mitigate the cost of traffic accidents by improving the way that society reacts when accidents are first reported. Today, emergency response resources are not always deployed optimally when accidents occur. Being able to predict whether a newly-reported accident is likely to be relatively severe vs. minor could be very helpful in ensuring that severe accidents are responded to quickly and appropriately. This could in turn save lives and/or reduce the severity of injuries sustained in accidents.

This study takes an analytical, machine-learning approach primarily focused on proactive approaches. Specifically, this study aims to understand the key drivers of fatal and serious injury traffic accidents so as to inform proactive strategies for mitigating accidents before they happen.

The target audience for this study is city government. City government will benefit from the increased understanding of the key drivers of severe traffic accidents and can leverage this data to help implement policies, reforms, infrastructure changes, and laws/ordinances designed to decrease the incidence of severe traffic accidents.

II. Data Overview and Preparation

A. Data Overview

This study leverages historical collisions data made available by the City of Seattle via their open data portal (https://data-seattlecitygis.opendata.arcgis.com/). The dataset contains information on ~220,000 accidents that occurred in the City of Seattle between 10/06/2003 and 9/5/2020. The data is collected by the Seattle Police Department (SPD) and is updated weekly. Accident category / severity data is available for ~200,000 of the accidents in the database. Accidents within the data set are classified into one of 4 severity categories:

Table1: Seattle GeoData Collisions Data

Severity Category	Description	% of Accidents
1	Property Damage Only Collision	68.9%
2	Injury Collision	29.4%
2b	Serious Injury Collison	1.6%
3	Fatality Collision	0.2%

The City of Seattle collisions database also contains a rich number of fields that describe each accident. Amongst other data, information on accident date, location, and type (head on, sideswipe, etc.) are available. Additionally, environmental information relevant to each accident is also provided including weather and road conditions. A summary of select key fields included in the data is provided in Table 2.

Table2: Seattle GeoData Collisions Data – Select Data Fields

Field	Description
INCDATE	Date of the accident
INCDTTM	Time of the accident
ADDRTYPE	Collision Address Type (Block, Intersection, or Alley)
LOCATION	Description of the general location of accident
COLLISIONTYPE	Type of collision e.g., Parked Car, Sideswipe, etc.
JUNCTIONTYPE	Category of junction at which incident took place
UNDERINFL	Whether or not driver was under influence of alcohol/drugs
WEATHER	Weather conditions during the accident
ROADCOND	Road conditions during the accident
LIGHTCOND	Light conditions during the accident
SPEEDING	Whether or not speeding was a factor
HITPARKEDCAR	Whether or not collision involved a parked car
PERSONCOUNT	Number of people involved in the accident
VEHCOUNT	Number of vehicles involved in the accident

B. Data Preparation and Cleaning

Given the focus of this study on mitigating serious and fatal collisions, 'Serious Injury Collisions' and 'Fatality Collisions' were grouped together as 'Major' Accidents. Additionally, 'Property Damage Only Collisions' and 'Injury Collisions' were analyzed collectively as 'Minor' Accidents.

Missing and unknown values were identified throughout the dataset and records with missing values for 'SEVERITY', 'WEATHER', 'LIGHTCOND', 'ROADCOND', 'UNDERINFL', 'JUNCTIONTYPE', 'ADDRTYPE', and/or 'COLLISIONTYPE' were purged from the dataset. After purging these records with missing values, ~173,000 records were available for further analysis.

The 'INCDATE' and 'INCTDTTM' date/time fields were leveraged to generate several potential features including a day of the week feature as well as a time of day feature (e.g., Late night, evening, afternoon commute, etc).

Categorical variables were identified and converted to dummy variables for the 'HITPARKEDCAR', 'SPEEDING', 'UNDERINFL', and 'INATTENTIONIND', 'WEATHER', 'LIGHTCOND', 'ROADCOND', 'ADDRTYPE', 'WEEKDAY', 'COLLSIONTYPE', 'JUNCTIONTYPE', and 'TIMEOFDAY' features. One hot encoding was used as necessary for select features.

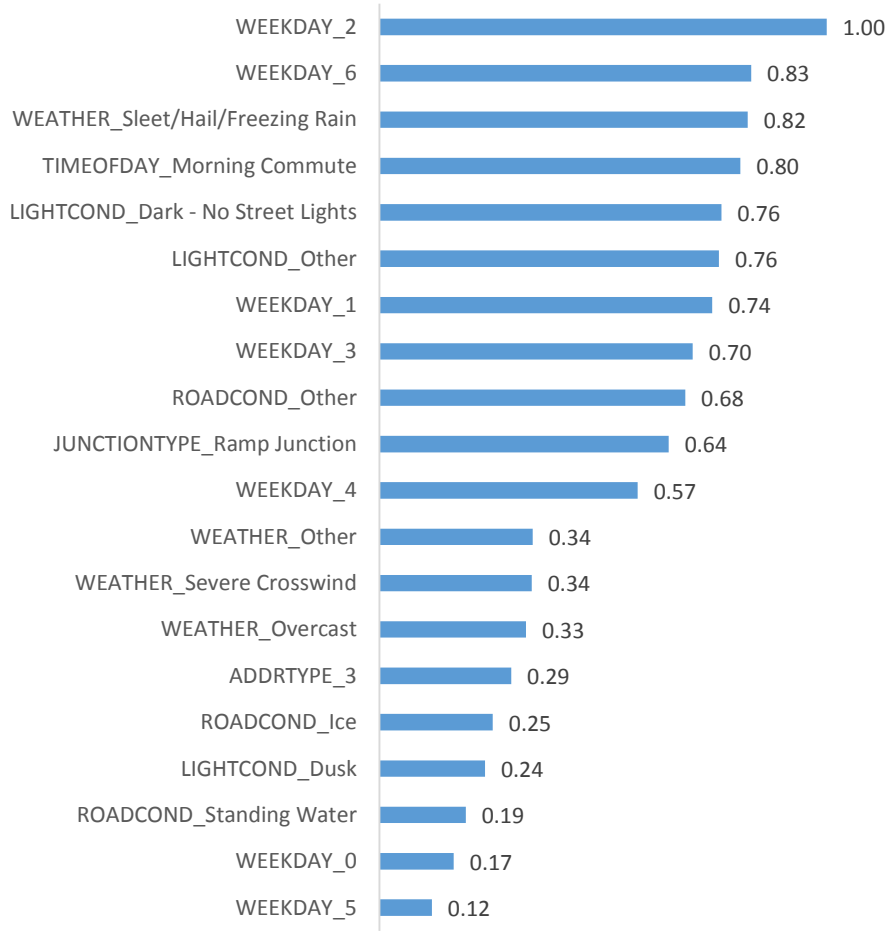
Additionally the data set was balanced so that roughly equal numbers of 'Minor' and 'Major' accidents were included in the analysis. Approximately 9,500 'Minor' accidents were sampled from the broader data set. Given that only ~3,300 clean 'Major' accidents records were available in the data, the 'Major' accident records were oversampled to achieve balance for modeling. 'Major' accident records were triplicated in the dataset resulting in ~9,900 'Major' accident records used for modeling.

C. Feature Selection

Once the data was cleaned, feature selection was completed. As a first step in feature selection, some features from which it would be difficult to draw tangible or useful conclusions were excluded. Attributes like 'VEHCOUNT' and 'PERSONCOUNT' were eliminated from consideration in this step. Despite the fact that these variables have a significant relationship with the 'SEVERITY' target variable, understanding that the number of vehicles or people involved in an accident contributes to the level of severity of an accident is of limited practical use for proactive accident mitigation.

After eliminating select features based on their usefulness for drawing conclusions, the remaining categorical features were analyzed statistically using the chi-square test to understand potential relationships between the variables and accident severity. Features with p-values less than 0.1 were removed from the data set. Figure 1 below summarizes the features that were dropped based on this criteria.

Figure 1: Chi-Square Test p-values > .1 for Feature Selection (Features Removed For Modeling)

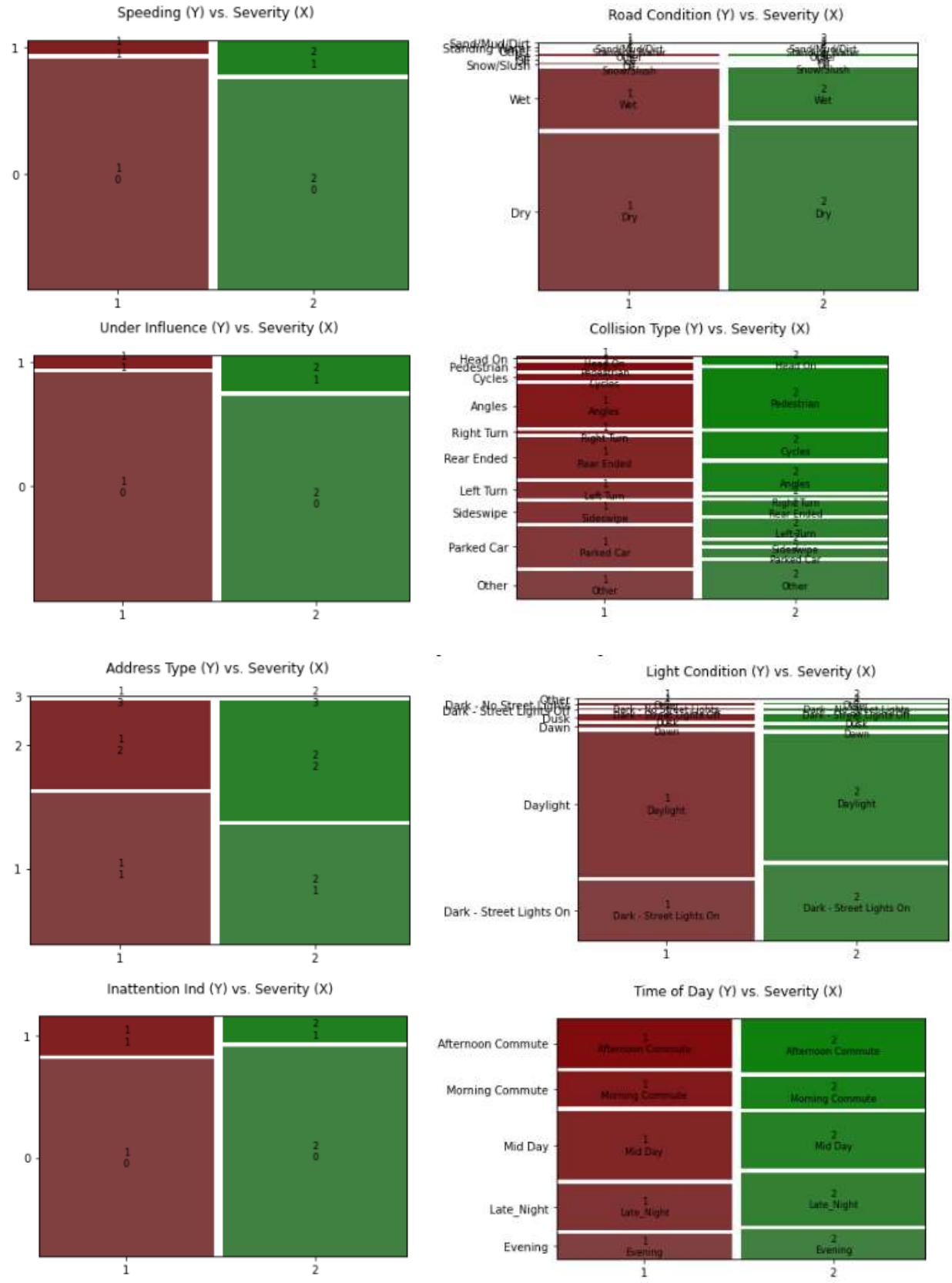


After the removal of select features based on the above criteria, 41 features were included in the proactive accident mitigation model.

III. Methodology and Results

The data set was initially analyzed to explore potential relationships between accident severity and individual features. Mosaics were created for select attributes in the balanced data set and through this process strong relationships to severity became apparent for several variables including (1) speeding, (2) under the influence, (3) select collision types, (4) select junction types, (5) select light conditions, and (6) select address types. Figure 2 below details the mosaics that were created for various attributes.

Figure 2: Mosaics to explore relationships between Severity and Select Features



Four machine learning models were initially built and compared in an effort to appropriately classify accident severity into either 'Major' or 'Minor' categories. These included (1) K Nearest Neighbors, (2) Support Vector Machine, (3) Decision Tree, and (4) Logistic Regression models.

A1. K-Nearest Neighbor Modeling

The K-Nearest Neighbor (KNN) model was run at several values of K and achieved best results at N=6 (See Figure 3). A confusion matrix was also generated for this model and is provided in Figure 4. While the KNN model provides similar levels of accuracy to the other approaches, it's of somewhat limited value in terms of understanding the underlying key drivers of accident severity.

Figure 3: K-Nearest Neighbor Model Accuracy for Different Values of K

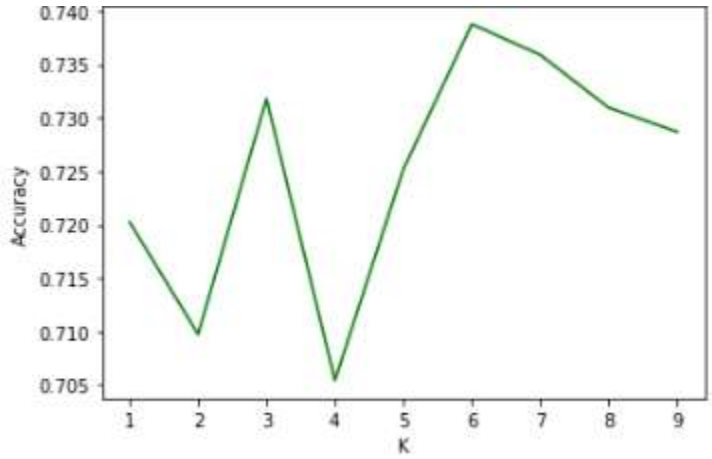
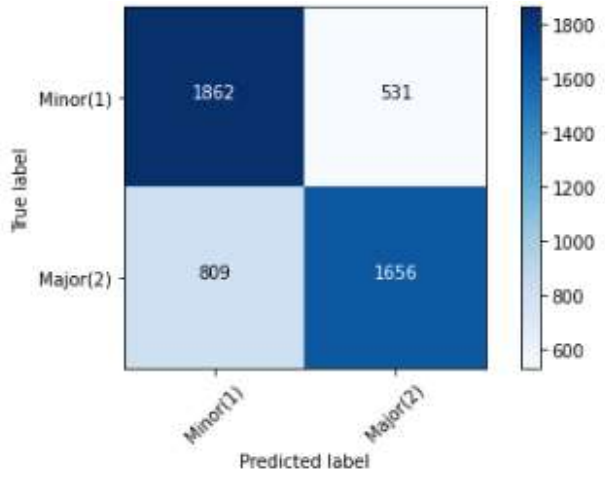


Figure 4: K-Nearest Neighbor Confusion Matrix

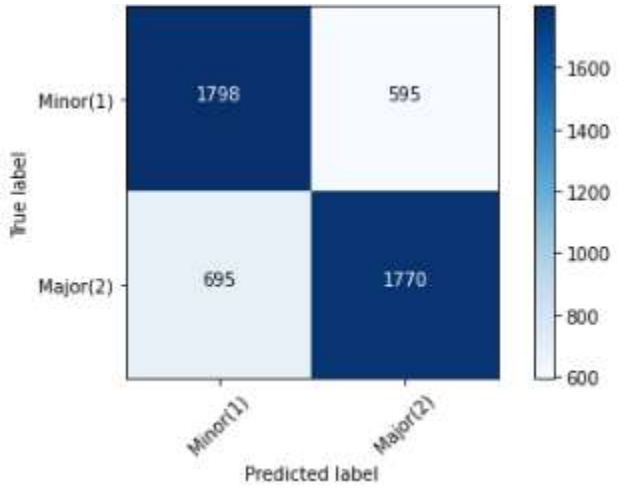


A2. Support Vector Machine

The Support Vector Machine (SVM) model was run using 4 different kernels and achieved best results with the 'rbf' kernel. A confusion matrix was also generated for this model and is provided in Figure 5.

Similar to the KNN model, it is somewhat challenging to interpret the underlying drivers of the support vector machine model (i.e., which features are most significantly driving level of accident severity).

Figure 5: Support Vector Machine Confusion Matrix



A3. Decision Tree

A decision tree model was also created to classify accident severity. Model performance leveled off quickly with minimal improvement in performance with a depth larger than 3. The confusion matrix for the decision tree model is provided in Figure 6 below. The decision tree model is easier to interpret than the KNN or SVM models given the ability to visualize the tree. Figure 7 below details the decision tree that was generated for accident severity. Based on the visualized tree, it's possible to understand that pedestrian involvement, bicycle or motorcycle involvement, and speeding are highly predictive of a major / serious accident.

Figure 6: Decision Tree Confusion Matrix

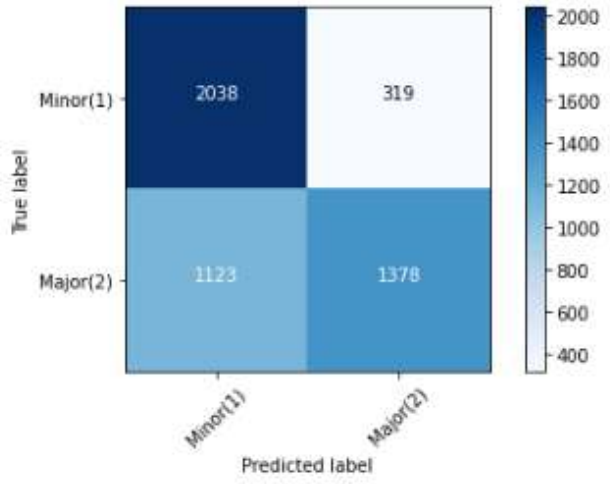
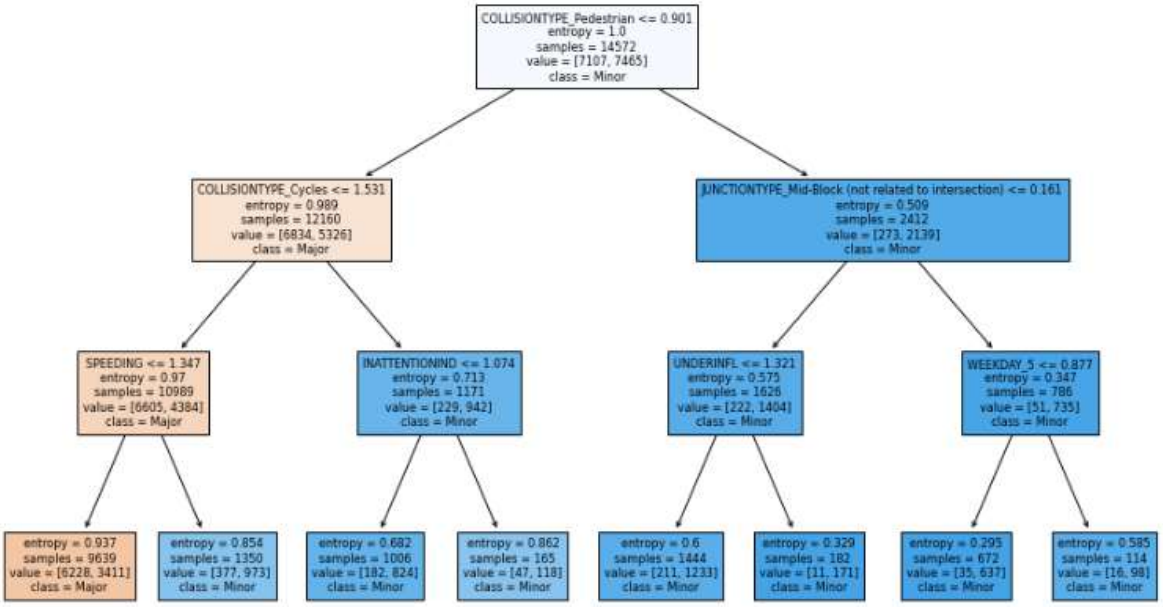


Figure 7: Decision Tree (max depth = 3) Visualized



A4. Logistic Regression

A logistic regression was also created to classify 'Major' vs. 'Minor' accidents. The confusion matrix associated with the logistic regression is summarized in Figure 8. Key features that predict accident severity can also be directly interpreted using logistic regression through an analysis of the regression coefficients. Figure 9 below details the six features with the largest positive coefficient values as well as the six features with the largest negative coefficient values in the logistic regression. The features with large positive coefficient values increase the odds of a 'Major' accident and include 1) pedestrian involvement, 2) cycle involvement, 3) speeding, 4) under the influence / drunk driving, 5) head on collisions, and 6) clear weather. The six features with the largest negative coefficient values increase the likelihood that the accident is relatively 'Minor'. These include (1) parked car collisions, (2) sideswipes, (3) rear ends, (4) angles collisions, (5) mid block junctions, and (6) driveway junctions.

Figure 8: Logistic Regression Confusion Matrix

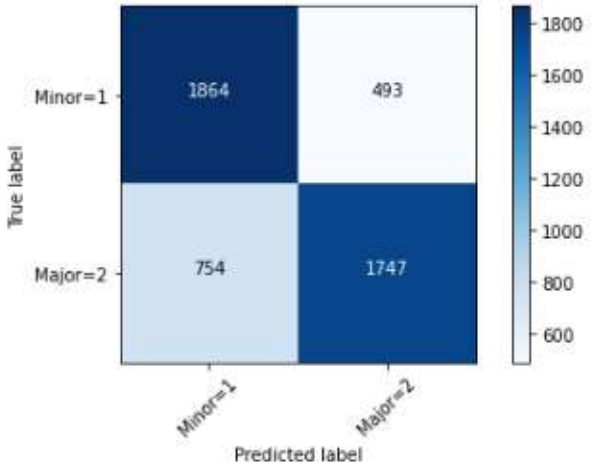
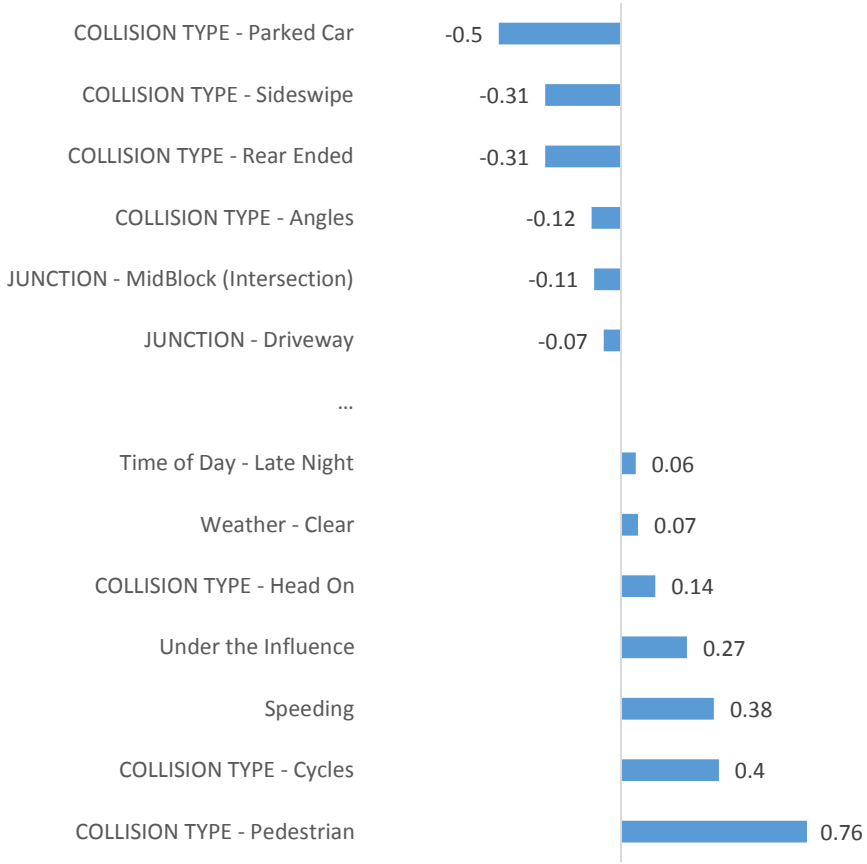


Figure 9: Logistic Regression Coefficient Values (7 Largest Positive and 6 Most Negative)



IV. Results

At the highest level, all models achieved relatively similar levels of accuracy with F1-scores ranging from 0.7 – 0.75. Table 3 below summarizes the Jaccard Similarity Scores and F1-scores associated with each model.

Table3: Proactive Accident Mitigation Models – Jaccard and F1-Scores

Model	Jaccard Similarity Score	F1 Score
K-Nearest Neighbors (n=6)	0.74	0.74
Support Vector Machines (kernel = 'rbf')	0.75	0.75
Decision Tree (max depth = 3)	0.70	0.70
Logistic Regression	0.74	0.74

Based on the models generated, the logistic regression and decision tree models generated the most compelling and *interpretable* results. The logistic regression model, in particular, generated F1-scores and Jaccard Similarity scores that were on par with the SVM and KNN models. Additionally, this model provided valuable insight into the specific features that drive classification into 'major' vs. 'minor' severity. The model suggests that pedestrian and cycle involvement, speeding, and drunk

driving, and head-on collisions are strong predictors of major accidents. Similarly, accidents involving sideswipes, rear ends, and parked cars are more likely to be 'minor'. Having a solid understanding of features that drive 'major' vs. 'minor' accidents is particularly useful for developing and gauging potential strategies to proactively mitigate future accidents.

V. Discussion

The modeling results suggest that city government may be able to prioritize several key areas as they seek to reduce severe traffic accident outcomes. Given that pedestrian and/or bicycle involvement in a crash is significantly more likely to drive serious injuries and/or death, efforts aimed at better protecting pedestrians and bicyclists could be funded. Specific ideas for consideration include (1) limiting traffic in select high risk areas, (2) building out barrier protected bike lanes in high risk areas, and (3) funding enforcement of bicycle and pedestrian safety rules (e.g., helmet laws / jaywalking fines). Another potential area of focus for city government could involve better communicating and enforcing speeding laws in select high risk areas as speeding plays a significant role in driving accident severity. Additionally, implementing efforts designed to reduce driving under the influence could also be helpful. Programs designed to provide at cost, no-questions-asked rides to impaired drivers could be considered. Furthermore, any efforts to reduce head-on collisions (e.g., installation of median barriers, reflectors, etc.) might also be worthwhile to consider given the high risk of injury / death associated with these accidents.

It's also important to note that while the modeling provides an understanding of key factors that contribute to serious/fatal accidents, it also highlights features that signal relatively more minor accidents. To the extent that the goal of city government is to reduce severe accidents, it's possible that dollars currently being allocated to mitigate less severe accidents could be reallocated / reprioritized to focus on more serious/fatal accidents. For example any spend currently focused on reducing the number of sideswipe or rear-end accidents could be re-prioritized to focus on mitigation of events that typically have more serious outcomes.

VI. Conclusion and Further Research Opportunities

Severe accidents represent a major issue in the US and lead to both significant loss of life and many serious injuries each year. Careful analysis of the Seattle accident data suggests that there are several proactive strategies city government can take to decrease the incidence of 'major' / severe accidents. Implementing strategies and regulations that seek to (1) better protect pedestrians and bicyclists, (2) curb under the influence driving, and (3) reduce speeding could have a material impact on the incidence of severe accidents.

In addition to implementing proactive strategies to reduce the number of 'major' accidents, it may also be interesting to explore if there are better *reactive* approaches that can be developed to address accidents immediately after they occur. Would it be possible to leverage a machine learning model to predict an accident's severity immediately after it is reported? Based on a few key inputs related to the accident, emergency response dispatchers could potentially deploy their limited

resources more effectively and thereby save lives and/or reduce accident severity. Further exploration and research in this area would be interesting to consider.