Seattle Traffic Accident Severity Prediction

Chi Zhang

September 30, 2020

1. Introduction

1.1 Background

Traffic accidents have risen to the 3rd main reason for mortality among countries by 2020, which also hurts the economy society. Citizens feel a lack of security when they drive on roads, cross streets, or even walk along pavements. Vehicle manufacturers have invested a large number of professionals and funds in improving the quality of vehicles, but all these efforts have less contribution to traffic rushes. It is time for the government to take some action to find out the leading causes of traffic accidents.

1.2 Business Problem

The purpose of this project is to analyze the collision dataset for the city of Seattle and find patterns and determinate the key factors such as weather, light and road conditions, drug or alcohol influence, driver inattention to provide the best traffic accident severity prediction. It will use various analytical techniques and machine learning classification algorithms such as k-nearest-neighbors, decision tree analysis, support vector machine, logistic regression, etc.

1.3 Target Audience

This study can mainly help transportation governments improve traffic policies or update public facilities such as street lamps, speed bumps at proper positions. Car rental or insurance companies are also among the target groups of this analysis because they can classify potential customers and design different service content based on customers driving habits.

2. Data

2.1 Data Source

Seattle Department of Transportation provides traffic accident cases for almost 15 years to discover the reasons behind these collisions. The dataset contains all kinds of collisions in Seattle from 2004 to 2020. The full dataset can be found here. The metadata can be found here.

To predict the damage level of road accidents, the indicator "SEVERITYCODE" is chosen as the dependent variable. The degree of collision climbs up from property damage only collision to an injury collision. Among dozens of attributes, this project concentrates on both nature and human factors of car accidents. Nature factors are made up of "ADDRTYPE", "COLLISIONTYPE", "WEATHER", "ROADCOND", and "LIGHTCOND" which represents location, contact parts, weather, road, and view circumstances separately. On the other hand, human factors usually reflect the status of drivers such as "INCDTTM", "INATTENTIONIND", "UNDERINFL", and "SPEEDING", which shows the time of accidents, the concentration of drivers mind, drug or alcohol influence, and the speeding driving. All attributes involved in this project are shown below (Table 1).

Table 1. Indicators involved in analysis

Indicator	Data type, length	Description
ADDRTYPE	Text,1	A description of the address type of the collision.
COLLISIONTYPE	Text,300	A description of the collision type.
WEATHER	Text,300	A description of the weather conditions during the time of the collision.
ROADCOND	Text,300	The condition of the road during the collision.
LIGHTCOND	Text,300	The light conditions during the collision.
INCDTTM	Text,30	The date and time of the incident.
INATTENTIONIND	Text,1	Whether or not collision was due to inattention.(Y/N)

UNDERINFL	Text,10	Whether or not a driver involved was under the influence of drugs or alcohol.
SPEEDING	Text,1	Whether or not speeding was a factor in the collision.(Y/N)

2.2 Data Cleansing

The whole dataset has 194673 records of car accidents with 37 indicators that describe characteristics of each accident in many aspects. The procedure of data cleansing proceeded step by step as follows: data check, indicator customization, feature selection, and classification simplify.

Firstly, I checked the whole dataset for a follow-up feature selection. There are mixed types of data such as numerical and categorical indicators. On the other hand, from a data integrity perspective, nearly half of the indicators have missing values to a variable extent.

Secondly, I translated incident timestamps to an additional feature - “WEEKEND” - whether the misfortune occurred at weekends that also means from Friday to Sunday. The translation followed the steps below: splitting “INCDTTM” into two columns including “DATE” and “TIME”, transforming “DATE” to the day of a week, and finally converting to whether the date was among weekends.

Thirdly, I conducted feature selection to only focus on the 9 categorical indicators which may be relative with severity degree. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. To predict the damage level of road accidents, the indicator “SEVERITYCODE” is chosen as the dependent variable. For independent variables, this program does not use the columns represented keys in many series such as “OBJECTID”, “COLDETKEY”, “INTKEY”, “SEGLANEKEY” and “CROSSWALKKEY”. Among dozens of attributes, this project concentrates on both nature and human factors which may lead to car accidents. Nature factors are made up of “ADDRTYPE”, “COLLISIONTYPE”, “WEATHER”, “ROADCOND”, and “LIGHTCOND” which represents location, contact parts, weather, road and view circumstances separately. On the other hand, human factors usually reflect the status of drivers such as “ WEEKEND”, “INATTENTIONIND”, “UNDERINFL”, and “SPEEDING” which shows the time of accidents, the concentration of drivers mind, drug or alcohol influence, and the speeding driving.

Fourthly, preprocessing the categorical values of 8 chosen features can interpret each column with simple and clear categories, because some of these features

even defined 10 or 11 different class names which described similar conditions. 5 features were simplified by combining close categories into new terms to reduce the number of categories while making them more typical (Table 2). In particular, “Unknown” was treated as missing value and “Other” was regarded as a separate class. Moreover, “INATTENTIONIND” and “SPEEDING” only have less than 30 thousand items in the whole dataset, because they only show the “Y” which means yes while ignoring the “N” for no. As a result, I filled in the missing data with “No”.

Table 2. Indicators simplified with less categories

Indicator	Before	After
UNDERINFL	N,0,Y,1	No, Yes
WEATHER	Clear, Raining, Overcast, Unknown, Snowing, Other, Fog/Smog/Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy	Clear, Overcast and Cloudy, Windy, Rain and Snow, Other
ROADCOND	Dry, Wet, Unknown, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt, Oil	Dry, Mushy, Wet, Other
LIGHTCOND	Daylight, Dark - Street Lights On, Unknown, Dusk, Dawn, Dark - No Street Lights, Dark - Street Lights Off, Other, Dark - Unknown Lighting	Bright, Medium, Dark, Other
COLLISIONTYPE	Parked Car, Angles, Rear Ended, Other, Sideswipe, Left Turn, Pedestrian, Cycles, Right Turn, Head On	Not Running Cars, Corners, Sides, Front and Back, Other

3. Methodology

3.1 Exploratory Data Analysis

Exploratory Data Analysis refers to a set of techniques originally developed by John Tukey to display data in such a way that interesting features will become apparent. Unlike classical methods which usually begin with an assumed model for the data, EDA techniques are used to encourage the data to suggest models that might be appropriate.

I depicted the relationship between each feature and the severity degree in bar charts to determine which features attribute to the injury collisions significantly. The histogram and probability density function of these injury collisions that took place at intersections can also help us observe the rules behind the scene of incidents and make some efforts to change the status quo.

3.2 Predictive Modelling

Classification Modelling can predict output by giving several inputs into trained models. I conducted dataset splitting, label encoding, and oversampling techniques before formal modeling.

The steps began with dropping missing values from the features since the amount of NAN within tolerance. The dataset scale decreased by 12.8% from 194673 to 169764 items. Then I split the inputs and the output into the train and test subset before label encoding to avoid information leakage. 20% of non-null items were separated as test subset, in other words, the train set and test set contained 135811 and 33953 items separately. Because the inputs were all categorical values, I chose the One Hot Encoding technique to convert each category value of the nominal variables into a new column and assigned a 1 or 0 (True/False) value to the column. This has the benefit of not weighing a value improperly but does have the downside of adding more columns to the data set. The number of vectors increased to 29 from 9 original features depending on the number of categories for original features. Finally, one of the popular oversampling techniques (SMOTE) was implemented to the data, because the output was imbalanced which would influence the predicting outcomes. Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in the dataset in a balanced way. The module worked by generating new instances from existing minority cases that I supplied as input. This implementation increased minority cases from 44556 to 91255 without changing the number of majority cases. The new instances were not just copies of existing minority cases. Instead, the algorithm took samples of the feature space for each target class and its nearest neighbors. The algorithm then generated new examples that combine features of the target case with features of its neighbors. This approach increased the features available to each class and made the samples more general. As a result, the final train set had 182510 items with commensurable cases in binary categories.

4. Results

4.1 Exploratory Data Analysis

How did the chosen features affect collision severity? Did the human factors from drivers themselves surpass the reason of environmental factors? Which one was

the most influential cause? The exploratory data analysis provided an opportunity for us to abstract knowledge from a mass of data.

4.1.1 Different Feature Impacts

I explored the relations between accident severity with these 9 attributes separately, which shows below.

Inattention drivers were 5.9 percentile much likely to come across serious collisions than ordinary ones (Figure 1). Among accident drivers, 34.8% of distracted drivers were involved in injury collisions, while 28.9% of focused drivers suffer the same misfortune. Some people cannot discipline themselves to avoid making and receiving calls, drinking, or eating when they are driving along the streets. These distractions can slow down the driver"s reaction time that may be just mere seconds. It is the seconds of hesitation that always causes irremediable loss.

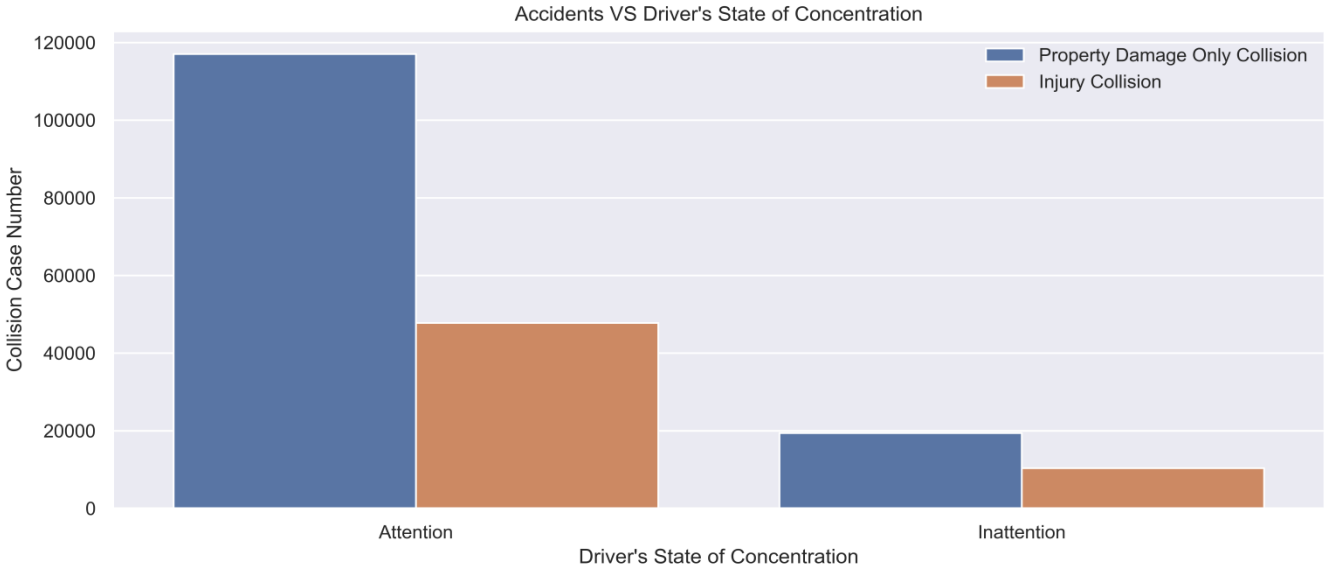

Figure 1. The relation between accidents and drivers" state of concentration

The possibility of serious car accidents caused by drunk or drug drivers is 9.4 percentage points higher than self-disciplined people generally (Figure 2). Nearly 40% of drunk driving or drugged driving caused injury collisions. Alcohol and hard drugs can paralyze the nerves and even create illusions, which have negative impacts on driving safety.

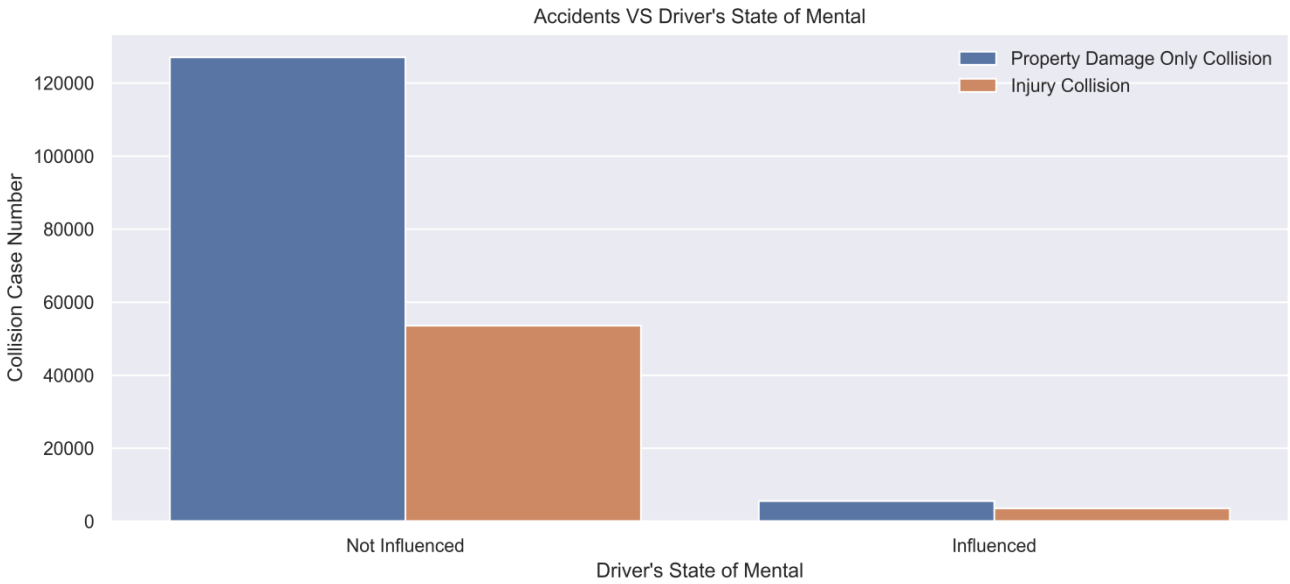

Figure 2. The relation between accidents and drivers" state of mental

Overspeeding also led to an 8.3 percentage point higher proportion of harmful collisions compared to traveling at normal speeds (Figure 3). 37.8% of drivers who exceeded the speed limits caused bad crashes. In general, speeding driving means less emergency response time, longer brake stopping distance, and more intensive collision strength. These reasons will aggravate the severity of car crashes and increase the possibility of personal injury.

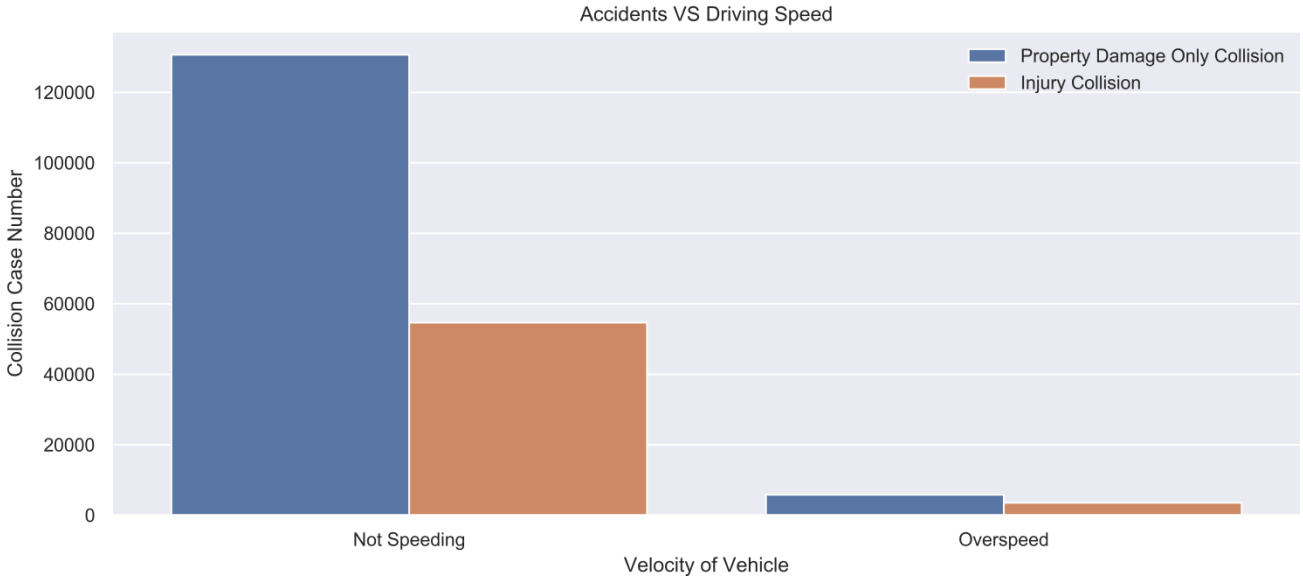

Figure 3. The relation between accidents and velocity of vehicles

Whether the day was among weekends, from Friday to Sunday, had nuances to the severity of crashes, but they were nearly the same at the 30% level (Figure 4). The subtle difference between workdays and weekends remained within 1.5 percentile.

This result may be inconsistent with our subjective cognizance that traffic congestion is worse during weekends. One explanation for less serious injury rate at weekends maybe is that drivers were in a hurry to get to work in rush hours during workdays while they would follow the principle of comity when they have leisure time.

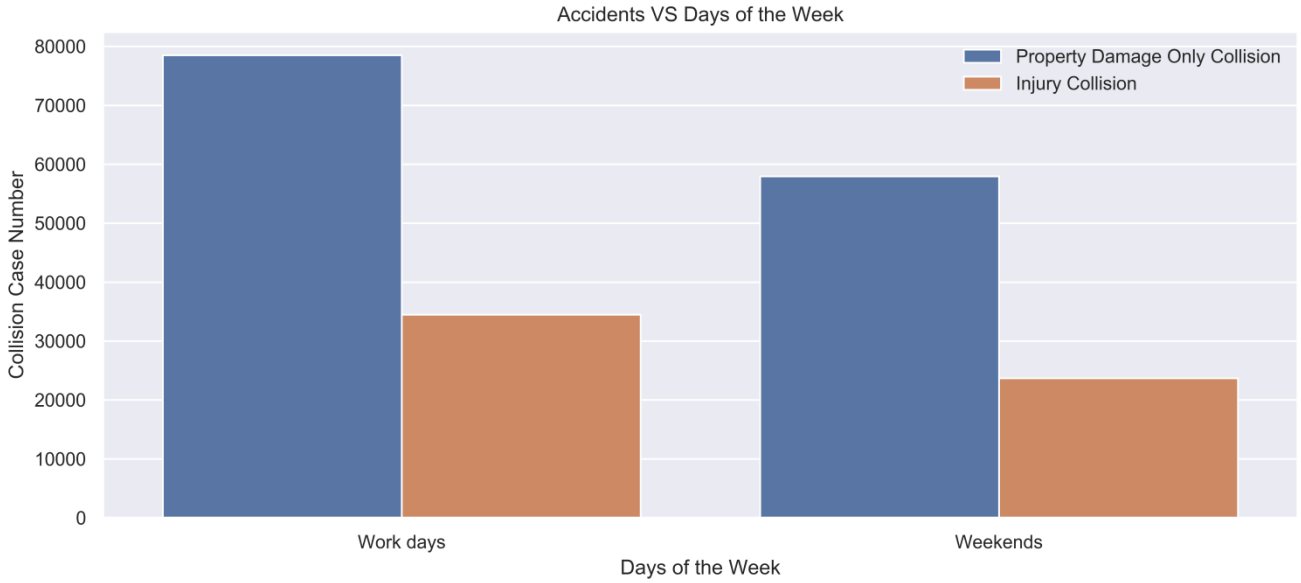

Figure 4. The relation between accidents and days of the week

The weather types have dramatically different influences on the severity of car accidents (Figure 5). Even though sunny, cloudy, windy, and rainy days have close serious accident rates, some other kinds of weather have a much lower rate at 13.9%.

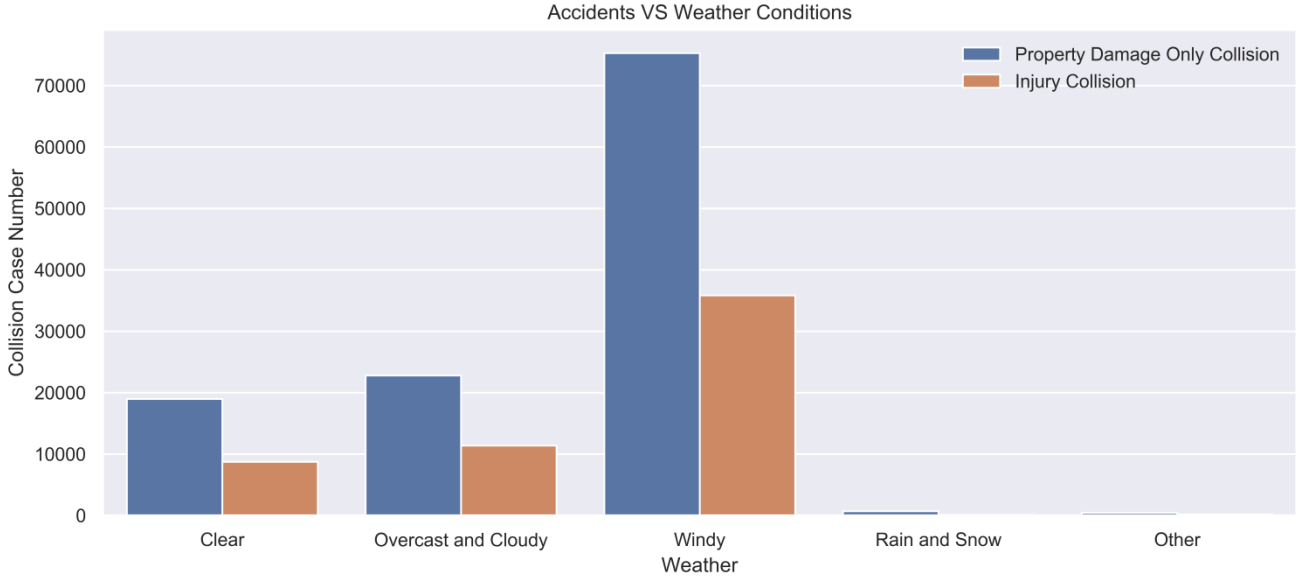

Figure 5. The relation between accidents and weather conditions

The serious accident rate was less than 20% under mushy road condition, which was 14.2 percentile below other kinds of roads (Figure 6). The wet road had the highest injury accident rate at 32.9%. The brake stopping distance is being drastically expanded because of the decrease of resistance on the slippery road. As a result, the vehicle would crash into an unpredictable direction out of control.

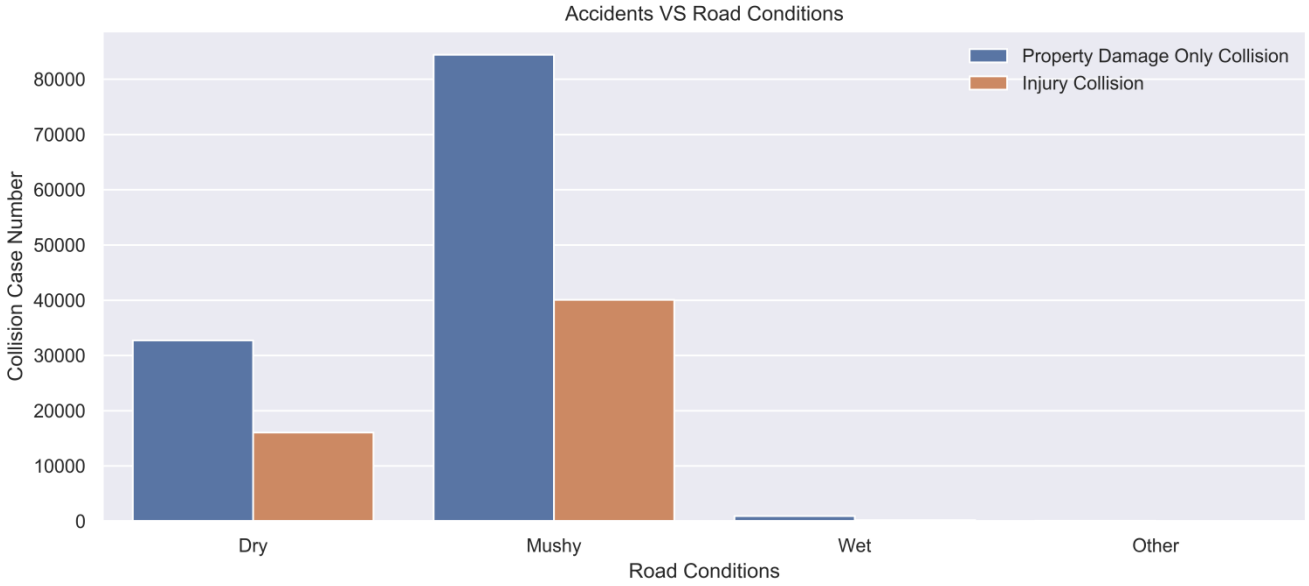

Figure 6. The relation between accidents and road conditions

The serious accident rates among light conditions sound beyond expectation (Figure 7). It is the dark condition at 23.8% of injury accident rate that much safer than bright and medium circumstances. In other words, the bright light condition was at the highest risk for drivers.

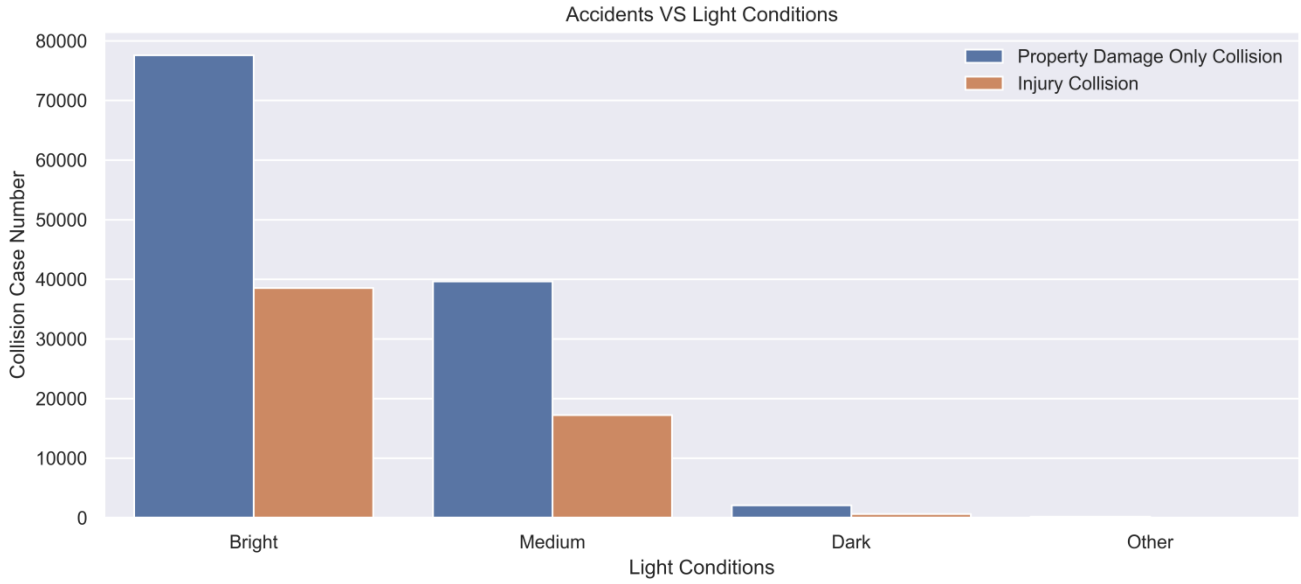

Figure 7. The relation between accidents and light conditions

Three address types had a remarkable influence on crash severity (Table 8). Up to 42.7% of intersection collisions were classified into serious accidents, while the block counterpart was 23.7% and the alley counterpart was only 10.9%. As a consequence, intersections without a doubt were the most dangerous sites for drivers to pay much attention. Cars, engineering vehicles, bicycles, and pedestrians meet together in chaos at intersections.

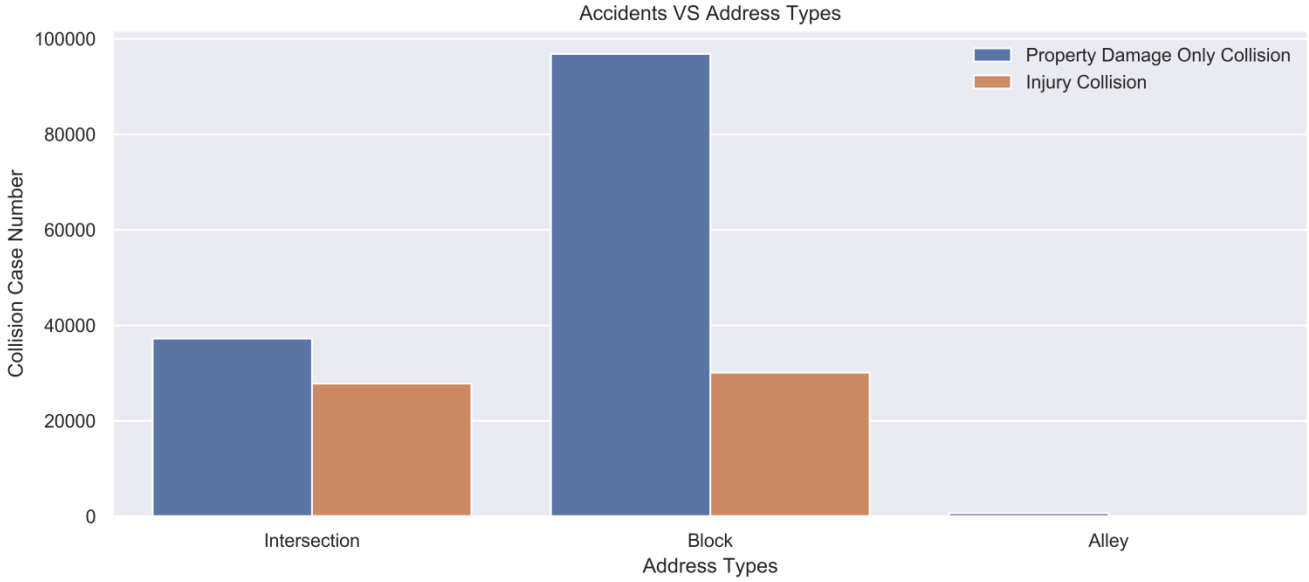

Figure 8. The relation between accidents and address types

Face to face collisions led to a high proportion of injury accidents at 43.0% (Table 9). By contrast, crashes from sides only included 13.5% of personal injury. A head-on collision can cause compressional deformation on the car body, which means that passengers in the car would be squeezed at the same time.

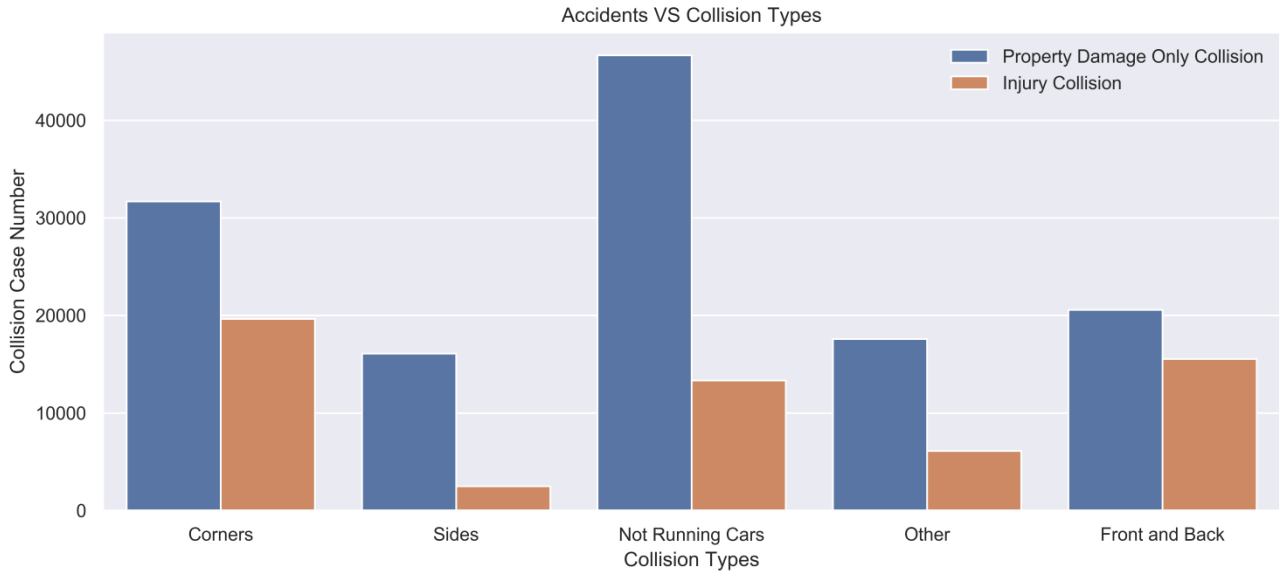

Figure 9. The relation between accidents and collision types

In conclusion, natural factors, including address, collision types, weather, road, and light conditions surpassed human factors related to drunk driving, speeding, inattention, and weekend relaxation on the impacts of car accidents (Figure 10). I measured the fluctuation of serious accident rates within each feature to explore the prominent reason for fatal misfortune. In the order, the top 5 remained to natural factors. As our common sense, bad circumstances properly contribute to casualties in road accidents. The address was the most important reason among these features, while weekends were the least one. All human factors had an influence on the severity of less than 10 percentile fluctuation.

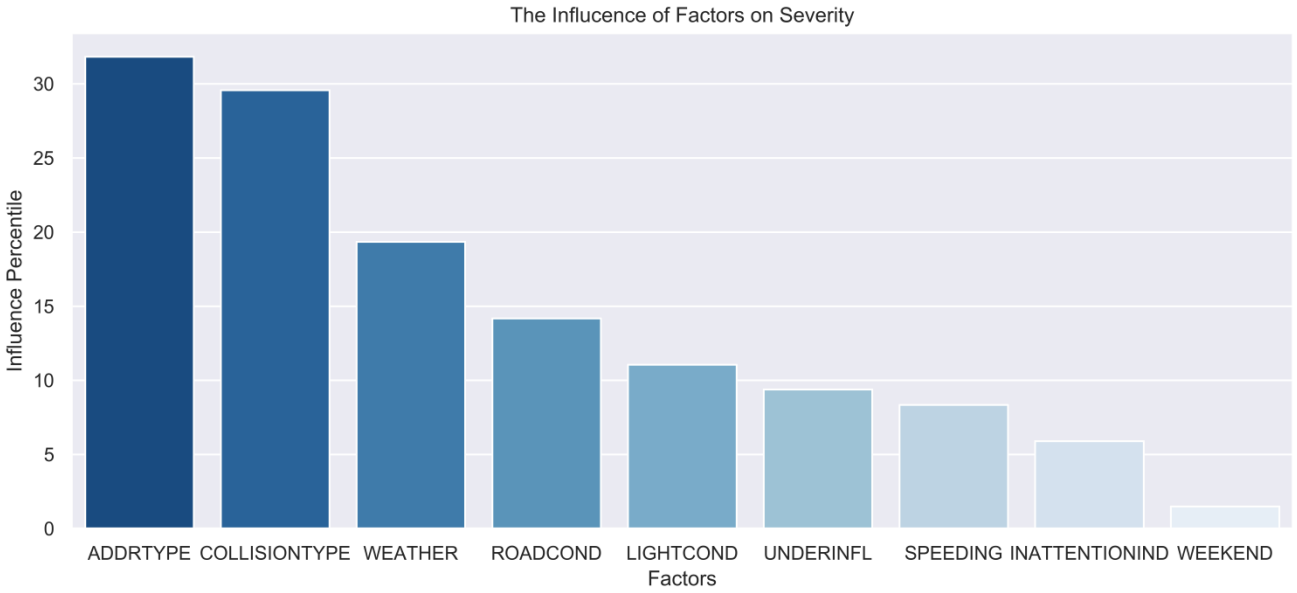

Figure 10. The relation between accidents and collision types

4.1.2 Map Illustration

I did further research to explore the serious accidents that happened at intersections because the address type was the most influential factor to the severity and within the address categories intersection represented the highest serious accident rate.

27718 cases of serious accidents occurred at intersections with latitude and longitude information in detail. By analyzing the frequency of serious accidents that took place at the same intersection, I found that the median number of frequency was twice in these years. The majority of the frequency of serious accidents" occurrence at the same place was less than 20 times (Figure 11), but the distribution of the frequency had a long tail on the right side. 26.7% of serious intersection collisions happened at where the above frequency was more than 20 times, in other words, high-risk positions.

For example, the most dangerous site may be a place where 107 serious crashes occurred in 15 years.

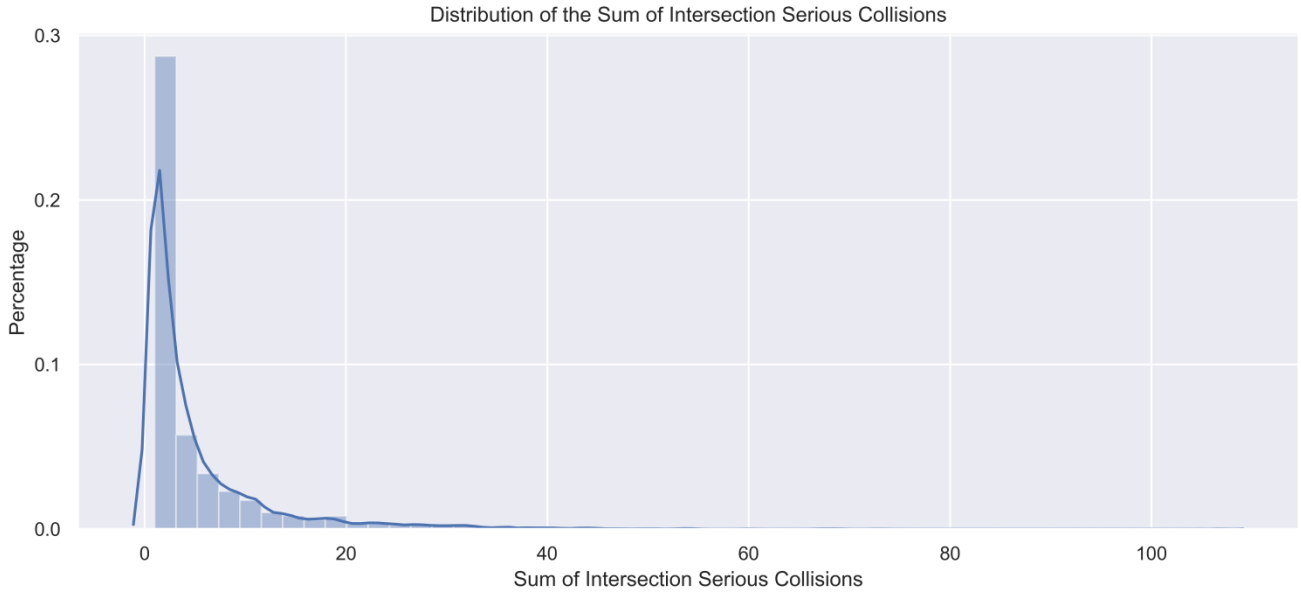

Figure 11. The relation between accidents and collision types

Depicting the top 1000 serious intersection accident sites on the Seattle map can help stakeholders make some efforts in the future (Figure 12). Every car mark on the map shows the position and the number of collisions that happened.

Figure 12. The relation between accidents and collision types

4.2 Predictive Modelling

To predict accident severity based on given features, I applied 5 predictive algorithms, including KNN, Decision Tree, SVM, Logistic Regression, and Random Forest. By adjusting the best hyper-parameter for each algorithm, I built 5 alternative supervised machine learning models. Some metrics, such as the accuracy, the precision, the recall, and the F1-score are calculated from predictive results based on upper models to select the best one (Table 3).

Table 3. Metrics for predictive models

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
KNN	67.00	50.00	21.00	30.00
Decision Tree	62.89	45.53	70.32	55.28
SVM	63.00	45.68	71.18	55.65
Logistic Regression	62.35	45.14	71.80	55.44
Random Forest	62.57	45.48	70.73	55.36

Four models had similar metric results at around 63% on the accuracy, at 45% on the precision, at 71% on the recall, and at 55% on the F1-score. Only the KNN algorithm had the lowest evaluation results in the 5 preparatory models. Because the KNN assumes that similar things exist in close proximity. In other words, similar things are near to each other. But traffic accidents happened randomly, so the KNN is not suitable in this situation.

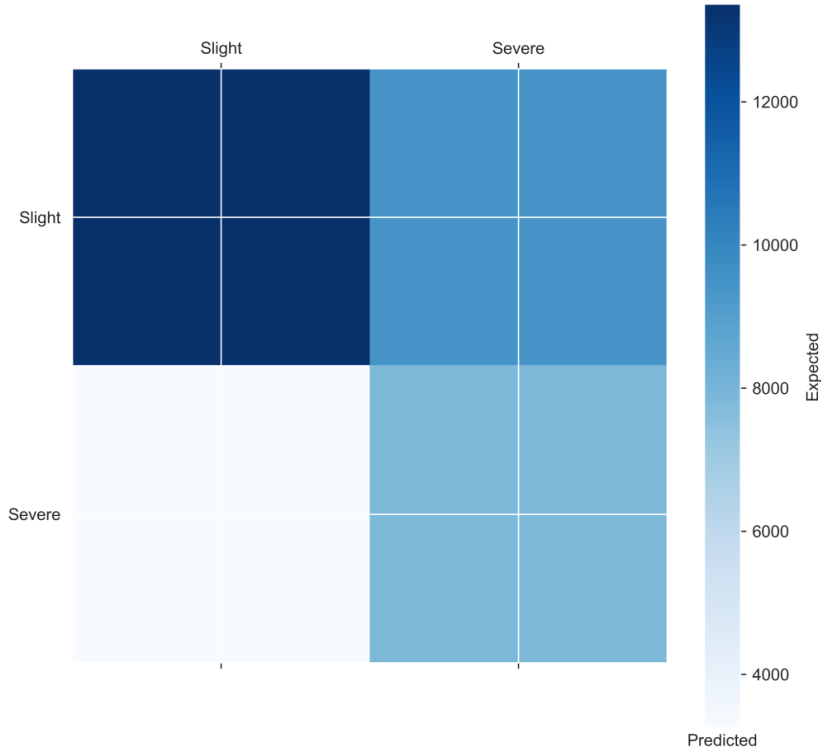

Figure 13. The confusion matrix of the Random Forest model results

Even though the accuracy scores are around 63%, but recall scores are much higher at 71%, for example, the Random Forest model (Figure 13). 70.73% of actual serious cases were predicted correctly, while only 45.48% of predicted serious cases were really serious. Some property only collisions were predicted wrong to serious ones. Since the purpose of this prediction is to build a pre-waring system for injury collisions, the stakeholders are eager to pick out every serious case among all accidents. They can ignore the wrong classification for the serious case instead of non-serious one because the strategy is "Catching every bad guy". As a result, models except the KNN were accepted because of the high recall.

I choose the Random Forest model (n=100) finally because it had the least processing time. If stakeholders want to predict other cases later, the Random Forest model will help them tackle the issue in a short time. Other models took hours when I calculated the results on known data.

5. Discussion

I was able to achieve ~65% accuracy in the classification problem. However, there was still significant misclassification by the models in this study. I think the models could use more improvements to reduce false positive predictions. In further research, I will try to adjust the hyper-parameter of existed models or apply new models, such as the Xgboost model.

Moreover, models in this study mainly focused on categorical features. I will try to combine some numerical variables into predictive models to boost accuracy. Before the involvement of numerical variables into models, I need to preprocess them with standard scale transformation.

6. Conclusion

Firstly, governments should update facilities to warn and protect drivers and walkers, because natural factors surpassed human factors on the impacts of car accidents. Bad circumstances properly contribute to casualties in road accidents. For example, the construction of speed bumps at crossings may decrease the fatal accident rate. If governments are under a financial strain, they can concentrate on high-risk regions first from the map from this study.

Secondly, insurance companies can customize service according to this study. They could set higher vehicle insurance premiums on people who have a behavior

record of drunk driving, under drug driving, or speeding driving. The optimal scheme can help insurance companies reduce risks and save money.

Thirdly, vehicle manufacturers should organize some research about improving the mechanical structure of cars because the front and end part of cars are vulnerable. The innovation of car body materials or buffer structures could alleviate hurts on passengers.