Tossed Bumpers and Scrambled Lanes: Predicted Peak-Hour Collisions in Seattle

By Claudia Panarello

September 14th, 2020

Introduction

Motor vehicle collisions have always been untimely and unfortunate circumstances, not only for the driver or drivers involved, but for the countless bystanders left with no choice to reroute, detour, and delay their arrival to their destinations. Regardless of their severity, collisions oblige the intervention of law enforcement and sometimes first responders and towing contractors. While it might be unlikely to prevent them altogether, it would be wise to ensure that enough resources are available and accessible if one were to occur. More automobiles are in circulation during rush hour, which might suggest that the mere presence of more vehicles would result in more collisions. It would be imperative to schedule enough traffic officers, first responders, and towing contractors to work during these peak hours if that is true. Not only would this increased assistance ensure accessibility and prompt attention to the event, its resolution would be relatively timelier, leaving fewer bystanders impacted.

Data

The dataset is from the Seattle Department of Transportation [1]and gathered by the Seattle Police Department's traffic records. This dataset is updated weekly, and its samples date as far back as 2004. Within this dataset, a particular focus will be on the timestamp of recorded collisions and weather, light, and road conditions for additional insight and determining a classification prediction. Peak rush hour times fall between 6:00 am-9:00 am and 3:00 pm-6:00 pm.

While the basis of this research is mainly temporal, it would have been useful to acquire more geographical data from external sources, segment the samples into boroughs, and then predict the frequency of collisions in each area. Hopefully, this would encourage dispatching the right amount of assistance to the areas with a higher likelihood of reported collisions. While some geographical data are present, only two attributes are provided in the dataset, and the number of outliers is much too large.

[1] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The dataset is also missing a significant amount of driver-related and demographic data. It would be beneficial to predict the likelihood of collisions during certain times of the day if drivers reported being distracted, fatigued, or driving under the influence of alcohol or narcotics.

The following attributes are dropped from the dataset due to irrelevancy and/or a remarkable absence of sample data:

Column name	**dtype**
X	float64
Y	float64
PERSONCOUNT	int64
VEHCOUNT	Int64
LOCATION	object
PEDCOUNT	int64
PEDCYLCOUNT	int64
STATUS	object
PEDROWNOTGRNT	object
SPEEDING	object
SEGLANEKEY	int64
CROSSWALKKEY	int64
HITPARKEDCAR	object
ST_COLDESC	object
SDOTCOLNUM	float64
OBJECTID	int64
INCKEY	int64
COLDETKEY	int64
INTKEY	float64
EXCEPTRSNCODE	object
EXCEPTRSNDESC	object
SEVERITYCODE.1	int64
SEVERITYDESC	object
COLLISIONTYPE	object
INCDATE	object
SDOT_COLDESC	object
INATTENTIONIND	object
UNDERINFL	object
ADDRTYPE	object
JUNCTIONTYPE	object
SDOT_COLCODE	int64
ST_COLCODE	Int64

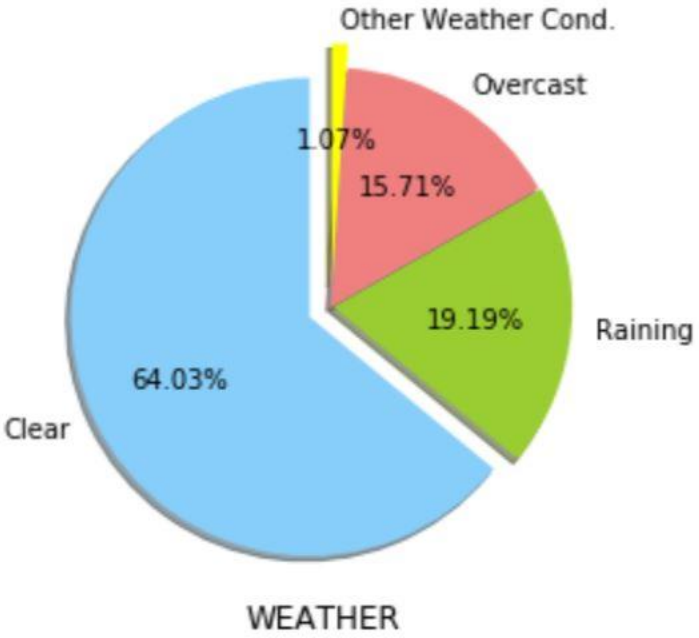
The net number of samples used in this research is 147,698.

Methodology

Weather

One would expect that weather and time of day are weakly correlated, mainly since weather events can occur at any moment. Based on a frequency count of weather conditions, a majority of 64.03% of collisions occurred when it was reported to be clear outside. This finding is not unsurprising; one can suggest that clear skies generally tend to be the most common weather occurrence. Not to mention, individuals often feel incentivized to make plans to leave the house when the weather is pleasant, and as such, an increased volume of automobiles could result in a higher collision count. The remaining weather conditions include rain at 19.19% of all collisions, then overcast at 15.71%. The remaining conditions comprise the last 1.07% of all collisions:

- snow
- fog, smog, smoke
- sleet, hail, freezing rain
- blowing sand and/or dirt
- severe crosswind
- partly cloudy
- other unspecified conditions in the data

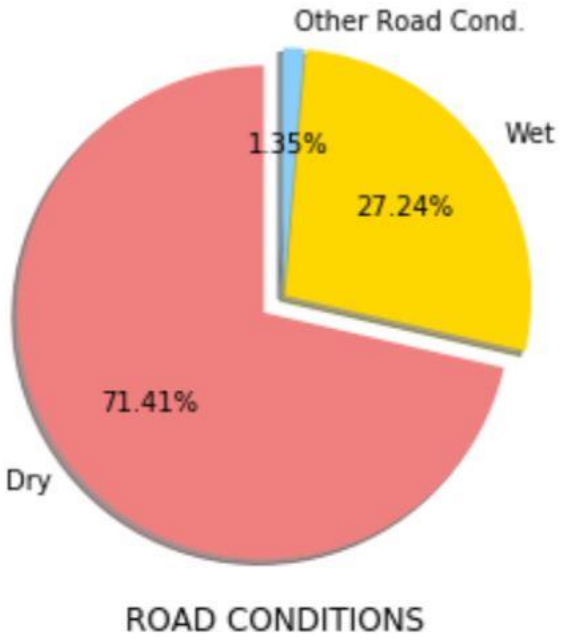


Road Conditions

Concerning road conditions, surprisingly, the majority of all collisions occurred on dry roads at 71.41%. Many individuals share a fear that terrible road conditions will increase their susceptibility to collision. This dataset tallied only 27.24% of all collisions for wet and 1.35% for other road conditions, including:

- ice
- snow/slush
- standing water
- oil
- sand, mud and dirt
- other unreported conditions

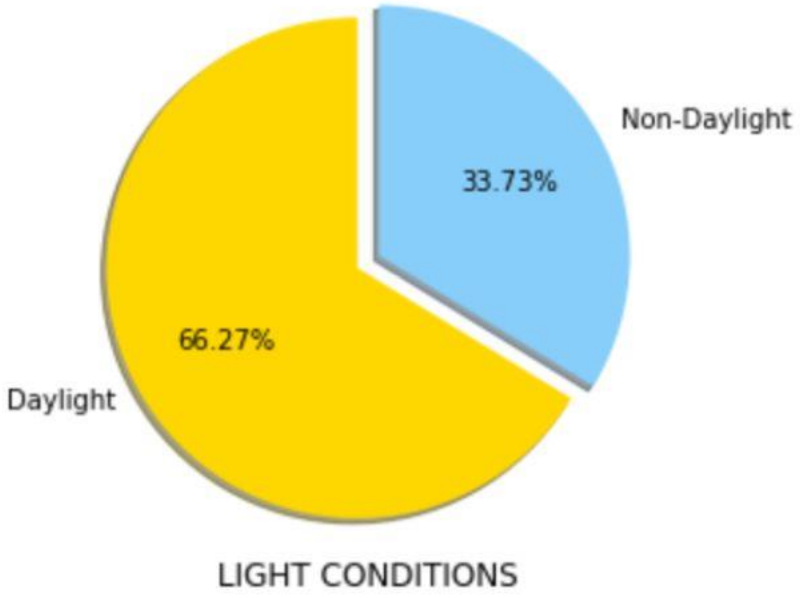
Perhaps this tally might not be surprising for some. In one sense, it might be due to the impetus mentioned above of commuting when the weather is nice, and roads are expected to be dry, resulting in more vehicles on the road. In another sense, the frequency count could be explained by the fact roads are dry most of the time, and poor road conditions are uncommon occurrences, albeit still a nuisance nevertheless.



Understandably, one would expect dry road conditions and clear weather to be highly correlated. In this dataset, its correlation is approximately 72.34%. Similarly, the correlation between wet road conditions and rain is approximately 76.83%, equally unsurprising.

Light Conditions

Regarding light conditions, the data are split into two categories for simplicity: daylight and non-daylight. Based on frequency, 66.27% of collisions were reported during the day, which might surprise some, since there is a preconception that lower visibility might raise the likelihood of collision. Some might not be as shocked since most drivers commute during the day, and this relatively larger volume of drivers could be responsible for this higher proportion of collisions. While one might assume that daylight and peak rush hour times are highly correlated, the dataset cites only an approximate 22.48% correlation.

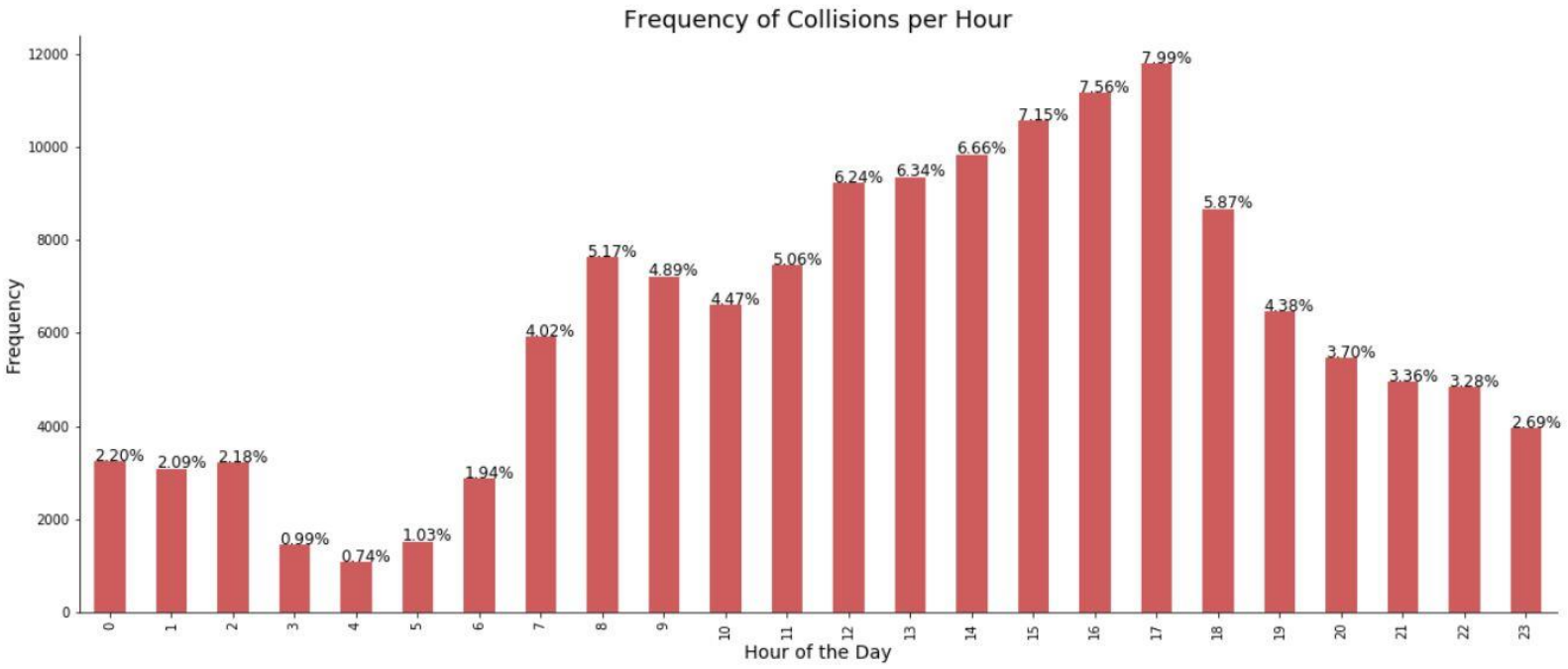


Time

Time is the most crucial exploratory variable since it is the crux of the peak vs. non-peak hours experiment. Based on all collisions' frequency, 7.99% of all samples fall between 5:00 pm and 6:00 pm,

and this is the largest proportion of any given hour. The second- and third-highest proportions are 4:00pm-5:00pm and 3:00pm-4:00pm, respectively. In other words, the top three proportions fall under the afternoon rush hour period.

Considering that the peak-hour periods in Seattle fall between 6:00 am-9:00 am and 3:00 pm-6:00 pm, which only consists of six hours of the day, note that their proportion of collisions is approximately 33.83%. In other words, over one-third of reported collisions occur during one-quarter of the day.



Severity Code

With regards to severity codes, only two types are numerically classified in the entire dataset:

1 – property damage

2 – injury

SEVERITYCODE	peak	
1	NON-PEAK	0.6741
	PEAK	0.3259
2	NON-PEAK	0.6364
	PEAK	0.3636

Only 32.98% of all sampled collisions resulted in injury, and the remaining majority of 67.02% consisted of property damage. When segmented, 67.41% of the property damage-related collisions occurred during non-peak hours, and the remaining 32.59% occurred during rush hour. Meanwhile, 63.64% of the injury-related collisions occurred outside peak hours, while 36.36% of this segment occurred during peak hours. Although these figures might seem roughly congruent, the proportion of property damage during peak hours denotes a relative difference of -1.81%. During these two 3-hour periods, the proportion of injuries denotes a relative difference of approximately 10.25% from its global estimate. That said, one might suggest that the proportion of injury is higher during peak hours, while property damage is slightly lower.

Machine-Learning Approach

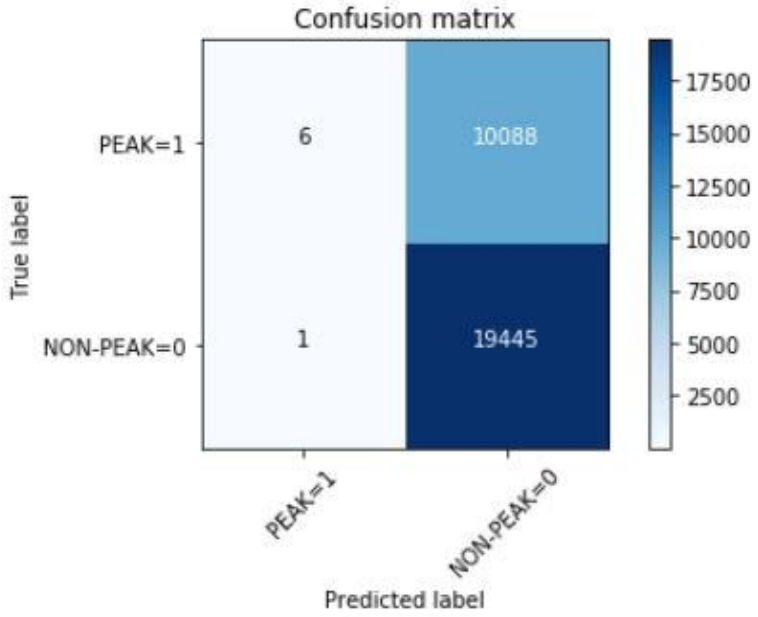
Due to the nature of the data being purely categorical, a classification machine-learning algorithm is used to predict the likelihood of collisions occurring during peak rush-hour times of day. Logistic regression is used on the testing and training data for said predictions since its target is binary: peak vs. non-peak. The test size consisted of 20%, while the remainder is used to train the model, resulting in 118,158 samples for training and 29,540 samples for testing. The C-parameter or the inverse of regularization strength is 0.02. The predictor attributes consist of the daylight, weather, road condition observations, and the severity code categories.

Results

From the machine-learning logistic regression algorithm, the model results in a Jaccard similarity score of approximately 0.6585. In other words, the testing data and the predicted value are 65.85% alike. Additionally, the model generates a weighted average F1-score of 52.31%. While the precision for 'NON-PEAK' and 'PEAK' is approximately 66% and 86%, respectively, its weighted average is roughly 73%. Furthermore, the recall of the model results in a weighted average of approximately 66%.

	precision	recall	f1-score	support
NON-PEAK	0.66	1.00	0.79	19446
PEAK	0.86	0.00	0.00	10094
micro avg	0.66	0.66	0.66	29540
macro avg	0.76	0.50	0.40	29540
weighted avg	0.73	0.66	0.52	29540

A confusion matrix is presented below to denote the true vs. predicted attribute:



Based on the findings, out of 29,540 testing samples, only 10,094 occur during peak hours, the remaining 19,446 do not occur during rush hour. The top-left and bottom-right quadrants denote the model's correct predictions, which are six occurrences of a collision during rush hour, and 19,445 collisions occurring outside of peak hours. However, the top-right incorrect predictions during peak hours denote an overwhelmingly low number of true-positive values. In other words, this quadrant depicts an extremely high number of Type I or false-positive errors. Meanwhile, the bottom-left quadrant characterizes the Type II, or false-negative, errors in the model. Fortunately, only one was detected. Lastly, this logistic regression model's log-loss score is 0.6151, which translates to a percentage accuracy of 24.26%.

Discussion

Overall, this logistic regression model is inadequate in predicting whether collisions are likely to occur during peak hours. Not only does this model present a weak percentage accuracy of approximately 24%, its abundance of Type I errors renders it unacceptable as a classifier. This error type might be due to the lack of explanatory power from its predictors.

However, while this model fails to measure the number of true positives in the data adequately, its ability to detect true negatives (i.e., collisions occurring during non-peak hours) is impressive. The model's Jaccard similarity score and its individual and average weighted precision are satisfactory, but the model's F1-score is unremarkable.

Nevertheless, the model fails to support the hypothesis that the proportion of collisions during rush hour times is predicted by environmental factors and the collisions' severity. Unfortunately, this model cannot be used to demand increased presence from law enforcement, first responders, or towing contractors during peak hours.

Conclusion

This research can predict the likelihood of vehicle collision during rush hour times of 6:00 am-9:00 am and 3:00 pm-6:00 pm in Seattle by 24.26% via a machine-learning classification algorithm, specifically

logistic regression. Road and weather conditions, the hours of daylight, time of day, and severity code, were used for additional insight. However, this model and its predictors failed to confirm the likelihood of collisions during peak traffic hours. While this research's primary goal is to provide a suggestion to first responders, traffic police, and towing contractors to improve commuters' well-being, its second goal is to encourage others to keep investigating. In matters of public safety, there could never be enough to research.