CAR ACCIDENT SEVERITY

CAPSTONE PROJECT

SUMEET GREWAL

1 INTRODUCTION

One of the most dangerous and complex activities we engage in on a daily basis is driving a vehicle. Increasing population density in major cities such as Seattle increases the propensity for traffic incidents and collisions. According to a dataset on collisions in the city of Seattle, there have been over 190 000 incidents since 2004. Traffic collisions have an enormous impact, most notably in terms of physical injury and in property damage.

It is possible that certain environmental factors and driver behaviours increase the severity of collisions. Some of these conditions could be specific days and times, weather and road conditions, driver inattention, and speeding.

The objective of this project is to understand which factors have the greatest impact on the severity of collisions, and identify which mitigation tactics may be effective. The key audience is the City of Seattle as they would ideally undertake initiatives to reduce the number and severity of traffic collisions.

2 DATA

The selected dataset contains all collisions provided by the Seattle Police Department (SPD) from 2004 - Present. It contains over 194000 records and 37 fields which are described in detail in the associated metadata file. The objective is to develop a model to accurately predict the 'severitycode' label. This field currently either contains a value of '1' indicating property damage only or a value of '2' indicating physical injury from the collision.

There are many columns that describe the various details of the collision. The focus of this study will be on identifying the key contributors to the collision. Some of the fields that will be assessed for their impact on severity are:

ATTRIBUTE	DATA TYPE	DESCRIPTION
INCDTTM	Text	Incident Date and Time
WEATHER	Text	The weather conditions at the time of the collision

ROADCOND	Text	The conditions of the road at the time of the collision
LIGHTCOND	Text	The light conditions at the time of the collision
INATTENTIONIND	Text (Y/N)	Whether or not the collision was due to inattention by the driver
SPEEDING	Text (Y/N)	Whether or not the collision was due to speeding by the driver
PEDROWNOTGRNT	Text (Y/N)	Whether or not the pedestrian right of way was not granted by the driver
UNDERINFL	Text (Y/N)	Whether or not the driver was under the influence of drugs or alcohol

Once the data has been prepared, the model will be trained on this data to predict the severity. This will help to identify which factors are most significant in the severity of a collision. As the objective is to determine the class label of a categorical target attribute, a classification model is most appropriate. As the data is binary and we are looking to understand the impact of various features, a Logistic Regression model is likely the best choice. Other models such as Support Vector Machine and Decision Trees may also be evaluated for their accuracy.

3 METHODOLOGY

3.1 EXPLORATORY DATA ANALYSIS

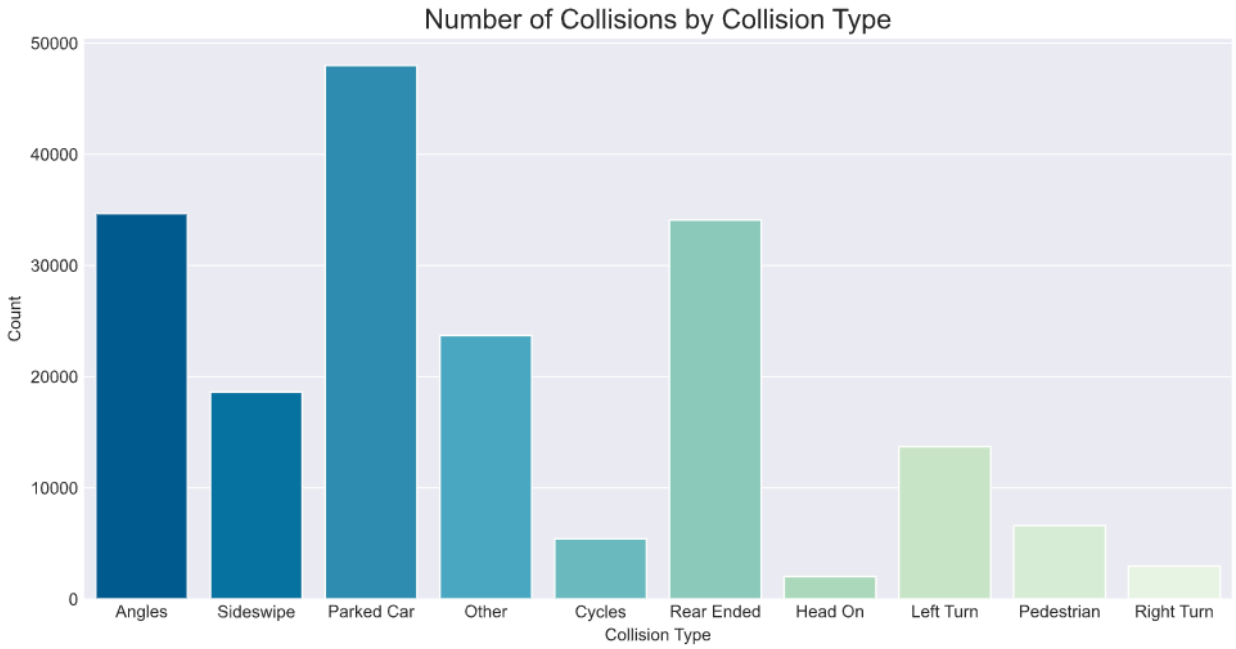
There are a number of columns that are used to describe the outcome and severity of the collision. The column 'SEVERITYCODE' is the only column that will be used to measure the severity so the other columns can be dropped. These include 'PEDCOUNT', 'PERSONCOUNT', 'INJURIES', and so on. The column 'SEVERITYDESC' is redundant information to 'SEVERITYCODE' so it can also be removed.

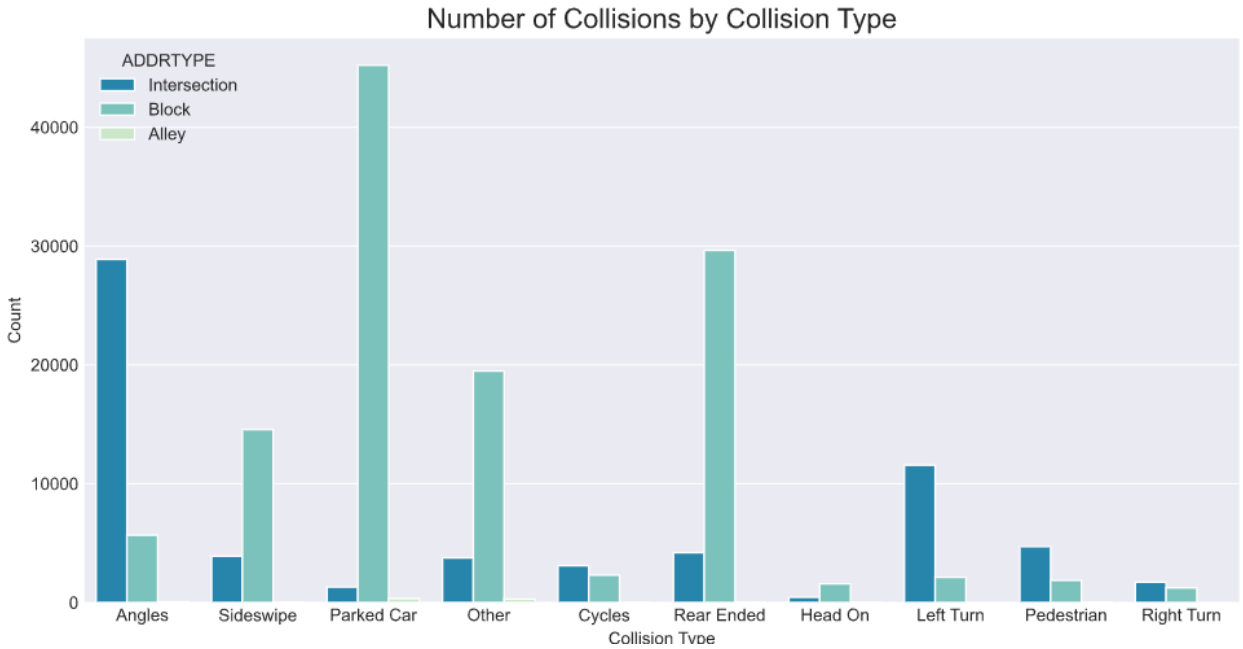
In addition, several of the attributes are primarily for identification and tracking purposes so they won't be useful for the model and can be dropped as well.

The remaining columns are: 'SEVERITYCODE', 'ADDRTYPE', 'INTKEY', 'COLLISIONTYPE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', ‘SPEEDING'

A surprisingly significant number of collisions have been recorded where one of the vehicles involved is a parked car. The two next highest collision types are rear-end collisions and collisions at angles. The second graph below illustrates the address type of each location. From this we can see which types of collisions are most common on ‘Blocks’ versus ‘Intersections’.

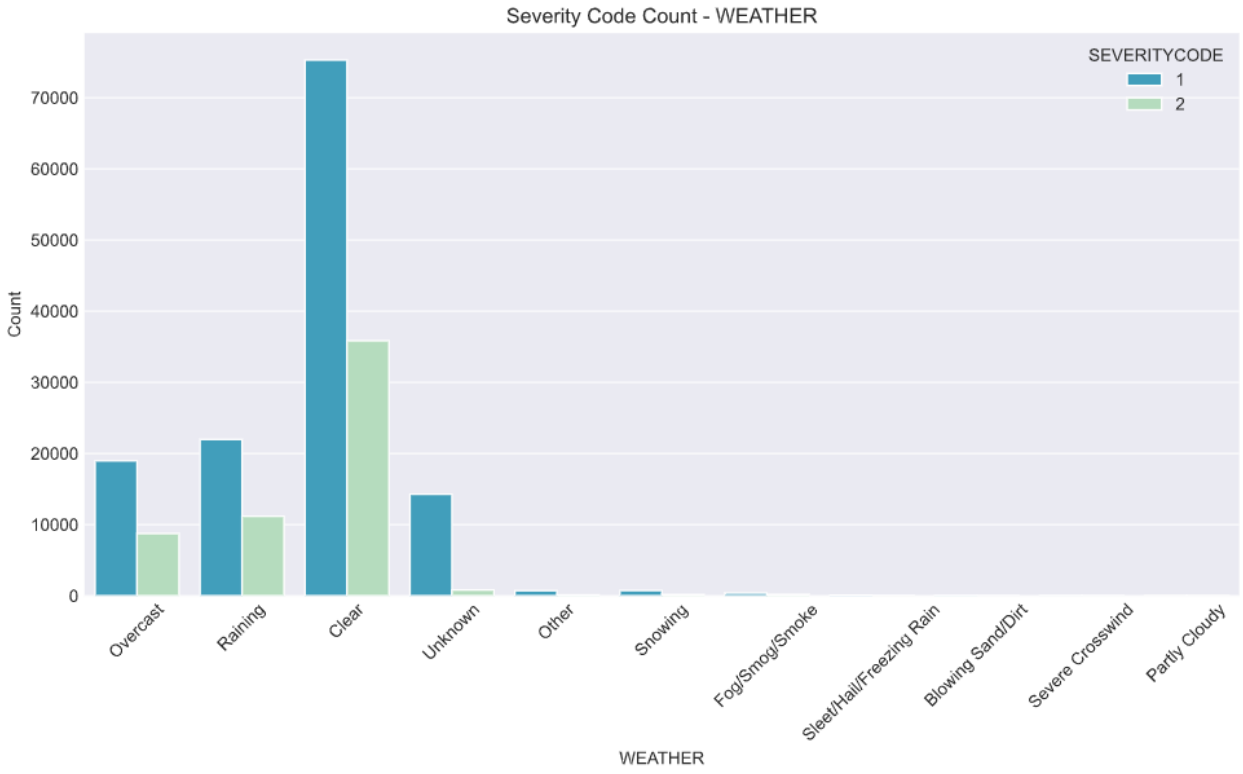
We can tell that most intersection collisions involve angles and left turns. Pedestrians are also more involved in intersection collisions. On the other hand, block collisions are mostly rear-end, parked car, or sideswipe collisions.





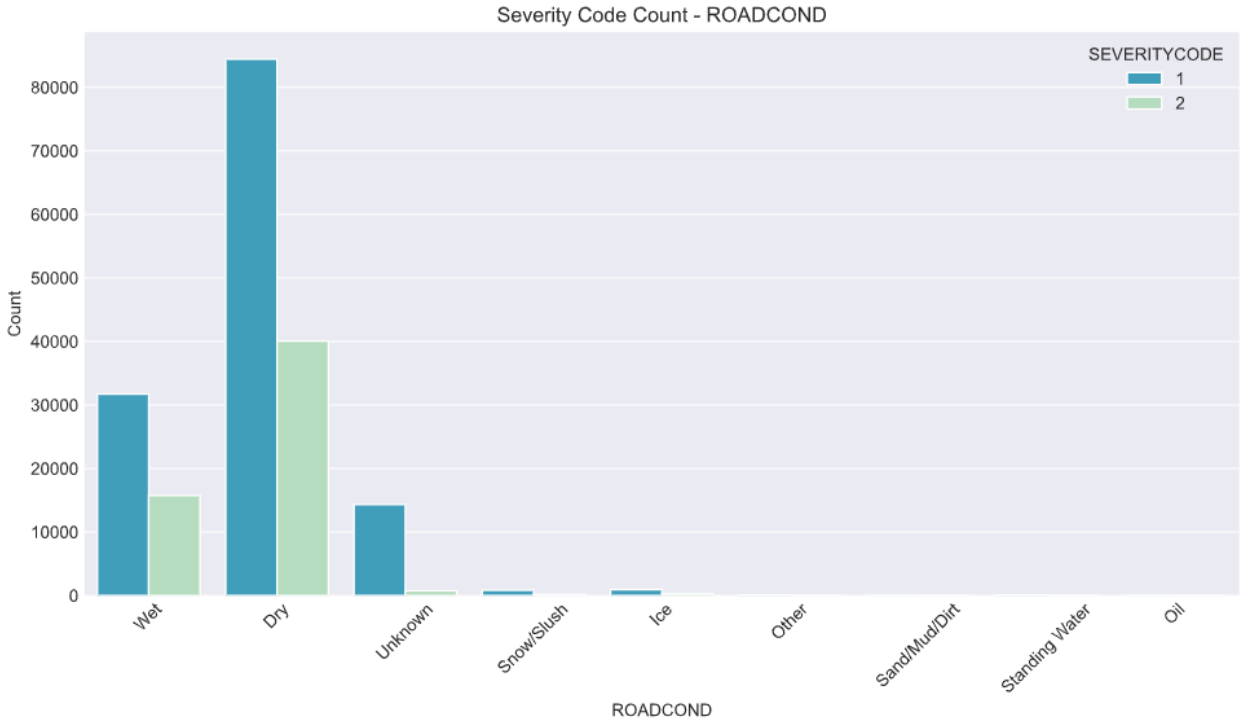
Looking at the weather conditions for the collisions, it is evident that most accidents occur in Clear conditions. An approximately equal proportion of the gross amount of collisions occur during Raining or Overcast collisions.

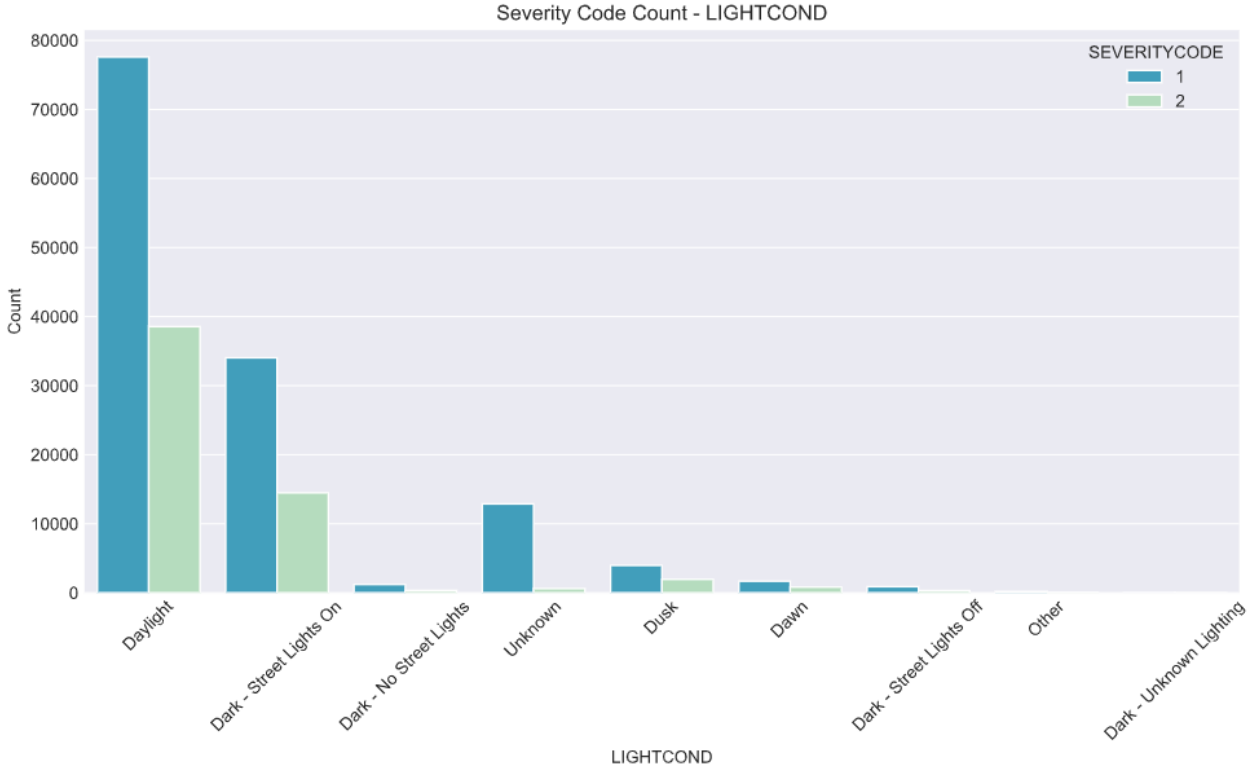
This distribution is likely skewed to reflect the overall incidence of each of these weather conditions. Far more days out of the year are likely to be Clear, so this information is not particularly indicative of increased risk. To improve the relevance of this information, the data for the weather attribute should be normalized against the average distribution of weather conditions in Seattle.



A very similar trend is identifiable in the following graphs of Road Conditions (ROADCOND) and Light Conditions (LIGHTCOND). Most accidents occur in Dry Road Conditions and Daylight Light Conditions. Again, this is in line with the Weather graph above. One interesting feature of the Light Conditions graph is that more collisions occur at Dusk than at Dawn.

It is evident from these graphs that worse weather, road, and light conditions are not likely to have a significant impact on the Severity of the collisions, at least according to our severity code measurement system. In both Raining and Clear conditions, a severity code value of 2 is approximately half as likely as a value of 1. This is also true for the comparison between Wet and Dry conditions, and the comparison between Daylight and Dark.


Severity Code Count - ROADCOND
SEVERITYCODE
1
2
Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Wet
Dry
Unknown
Snow/Slush
Ice
Other
Sand/Mud/Dirt
Standing Water
Oil
ROADCOND


Severity Code Count - LIGHTCOND
SEVERITYCODE
1
2
Count
0
10000
20000
30000
40000
50000
60000
70000
80000
Daylight
Dark - Street Lights On
Dark - No Street Lights
Unknown
Dusk
Dawn
Dark - Street Lights Off
Other
Dark - Unknown Lighting
LIGHTCOND

3.2 DATA CLEANING

Each of the remaining data columns is examined to understand what preparation and cleaning is required. To begin, the INTKEY columns is converted to type INT and null values were replaced with 0. There are only about 700 records with no value for ADDRTYPE, so these are dropped.

The original attributes of SPEEDING, PEDROWNOTGRNT, INATTENTIONIND, and UNDERINFL are all indicator values, though there are some unknown values and some inconsistent labelling. For example the value Y or 1 is used to represent True and an empty value or 0 is used to represent False. These values are all standardized so that all False values are coded as 0 and all True records are coded as 1.

Inattentive driver: 16.83 % of records
Speeding driver: 5.34 % of records
Driver under the influence: 5.29 % of records

Moving on to the WEATHER, ROADCOND, and LIGHTCOND attributes, the records with empty, 'Unknown', or 'Other' values are all dropped as they do not provide any useful information. Additionally, if the value for light condition is 'Dark – Street Lights Off', it is converted to 'Dark – No Street Lights' as these two descriptions are functionally identical for our purposes. Approximately 100 records are dropped where LIGHTCOND is equal to 'Dark – Unknown Lighting'. There are now just under 170 000 records remaining

Finally, 'One-Hot-Encoding' is used for each of these columns to convert them to indicator values. A unique feature is created for each of the possible values and it is given a value of 1 for a record if it was the original value.

WEATHER		ROADCOND		LIGHTCOND	
Clear	108822	Dry	121487	Daylight	112618
Raining	32644	Wet	46319	Dark – Street Lights On	46748
Overcast	26922	Ice	1080	Dusk	5648
Snowing	825	Snow/Slush	833	Dark – No Street Lights	2522
Fog/Smog/Smoke	553	Standing Water	105	Dawn	2413
Sleet/Hail/Freezing Rain	107	Sand/Mud/Dirt	65		
Blowing Sand/Dirt	46	Oil	60		
Severe Crosswind	25				
Partly Cloudy	5				

3.3 DATA BALANCING

The dataset is now unbalanced in terms of the class variable we are trying to predict. There are approximately twice as many records with SEVERITYCODE = 1 than SEVERITYCODE = 2 in our cleaned dataset.

An undersampling technique is used to balance the dataset. There are 55 539 records with value of 2 for the predictor. The records with value of 1 are isolated in a separate dataframe and shuffled. Then an equal number of records is randomly sampled from the shuffled dataframe. The sampled dataset is randomly combined with the records with a value of 2 to create a balanced dataframe.

The dataframe is now well balanced and is ready for the model.

3.4 MODELING

The first step is to define the dependent and independent variables. X will include all of the indicator columns that have been prepared and Y will be SEVERITYCODE.

The dataset is then divided into mutually exclusive training testing sets. This technique should improve the out-of-sample accuracy of the model. 20% of the records will be reserved for the test set.

Various metrics are used to evaluate and tune the Logistic Regression model. These include F1-Score, Jaccard Score, and LogLoss metrics.

The logistic regression model is trained using the training dataset. To tune the model, we will first try different solvers to identify which one produces the maximum accuracy and then different values for regularization strength.

Solver	liblinear				
C	0.01	0.05	0.1	0.5	1
Jaccard Score	0.4692	0.4693	0.4692	0.4693	0.4693
LogLoss	0.6710	0.6700	0.6699	0.6699	0.6699

Solver	newton-cg				
C	0.01	0.05	0.1	0.5	1
Jaccard Score	0.4694	0.4693	0.4693	0.4693	0.4693
LogLoss	0.6710	0.6700	0.6699	0.6699	0.6699

Solver	lbfgs				
C	0.01	0.05	0.1	0.5	1
Jaccard Score	0.4694	0.4693	0.4693	0.4693	0.4693
LogLoss	0.6710	0.6700	0.6699	0.6699	0.6699

Based on these results, the best parameters for our model will be C = 0.01 and solver = newton-cg or lbfgs.

4 RESULTS

From this classification report, we can see that the model is more accurate at predicting records with severity of 1 and not that great at labelling severity code 2. The weighted average f1-score

indicates that the average accuracy of our model is approximately 0.52. From this we may conclude that a number of the predictor variables that we defined are likely inconsequential in predicting the severitycode.

	precision	recall	f1-score	support
1	0.53	0.80	0.64	11083
2	0.60	0.31	0.41	11133
accuracy			0.55	22216
macro avg	0.57	0.55	0.52	22216
weighted avg	0.57	0.55	0.52	22216

5 DISCUSSION

Based on the analysis in the report, intersection collisions are far more likely to be defined as being a result of angles, left turns, or pedestrians whereas 'block' collisions are more likely to involve a parked car or rear ending. The analysis indicates that the type of collision is not very correlated with the severity of an accident.

One key area where correlation does exist is between severity code 2 (persons injured) and the indicator flag, PEDROWNOTGRNT (Pedestrian Right of Way Not Granted). This makes logical sense as if a pedestrian is not given their right of way, they are more likely to be involved in the collision.

One interesting difference is that more collisions occur at Dusk than at Dawn. However, most accidents in our dataset occurred in Dry Road Conditions and Daylight. This model doesn't take into account the various distributions of weather. An opportunity to improve the model is to normalize these values according to the actual average distribution of road, light, and weather conditions.

From a mitigation perspective, it would be unwise to apply resources uniformly across the city in order to address a specific issue. Mitigation efforts should be concentrated where they would have the greatest impact. It would thus be constructive to identify which intersections are the most dangerous. This way, the City of Seattle can optimally allocate their resources and mitigation tactics towards addressing key factors in key intersections.

6 CONCLUSION

A significant number of factors may determine whether a collision occurs and whether or not property is damaged or persons are injured. These include human behaviours such as distracted driving or speeding, and environmental factors such as weather, road, and light conditions.

The objective of this report was to identify if any such factors are accurately able to predict the severity of an accident. Severity was classified as a 1 if only property was damaged or as a 2 if injuries occurred as well. After preparing, cleaning, and balancing the dataset, a logistic regression model was built to predict the severity code. The dataset was split into 80% for training and 20% for training, in order to improve the out of sample accuracy.

The model itself was able to somewhat accurately predict the class value of 1 but was not very effective at accurately predicting the class value of 2. The results of the analysis determined that the identified factors are not significant indicators of the severity code measurement.