Coursera Capstone Project

November 17th, 2020

Introduction

Seattle ranks as one of the nation's worst cities to drive in and own a car, according to a newly released study. With the 10th worst score in the U.S. across all criteria and has the nation's third-highest auto maintenance costs. This problem has become increasingly worse over the years and has even deterred potential residents from moving there. The issue is widely realized, and the city of Seattle is willing to invest in the road system to help with these issues. They just need help finding where their money can be an impactful solution.

Data

Data source from SDOT Traffic Management Division, Traffic Records Group. Seattle collisions provided by SPD and recoded by Traffic Records. 37 attributes from 2014 – 2020 including coordinates.

Using data such as location, weather, car speeding, light conditions, road conditions, etc. I will be analyzing leading causes in vehicle damage in the Seattle area. In hopes to point the city in the right direction. It could be a weather app that directs people to different roads during a storm, junctions that are prone to crashes, or even poor lighting on a road. I will work to find correlations and hopefully a solution to improve these conditions.

Methodology

Using Logistic Regressing, K Means and Decision Trees I will find the best method to get relationships between a set of input variables and a categorical output variables.

Let's select some features for modeling.

SEVERITYCODE - A code that corresponds to the severity of the collision: 2—injury, 1—prop damage

ADDRTYPE - Collision address type: Alley, Block,

COLLISIONTYPE - Collision type

VEHCOUNT - The number of vehicles involved in the collision. This is entered by the state.

JUNCTIONTYPE - Category of junction at which collision took place

INATTENTIONIND - Whether collision was due to inattention. (Y/N)

UNDERINFL - Whether a driver involved was under the influence of drugs or alcohol.

WEATHER - A description of the weather conditions during the time of the collision.

ROADCOND - The condition of the road during the collision.

LIGHTCOND - The light conditions during the collision.

SPEEDING - Whether speeding was a factor in the collision. (Y/N)

HITPARKEDCAR - Whether the collision involved hitting a parked car. (Y/N)

```
df = df[['SEVERITYCODE','ADDRTYPE','COLLISIONTYPE','VEHCOUNT','JUNCTIONTYPE','INATTENTIONIND','UNDERINFL','WEATHER','ROADCOND','LIGHTCOND','SPEEDING','HITPARKEDCAR']]
df.head()
```

	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	VEHCOUNT	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	HITPARKEDCAR
0	2	Intersection	Angles	2	At Intersection (intersection related)	NaN	N	Overcast	Wet	Daylight	NaN	N
1	1	Block	Sideswipe	2	Mid-Block (not related to intersection)	NaN	0	Raining	Wet	Dark - Street Lights On	NaN	N
2	1	Block	Parked Car	3	Mid-Block (not related to intersection)	NaN	0	Overcast	Dry	Daylight	NaN	N
3	1	Block	Other	3	Mid-Block (not related to intersection)	NaN	N	Clear	Dry	Daylight	NaN	N
4	2	Intersection	Angles	2	At Intersection (intersection related)	NaN	0	Raining	Wet	Daylight	NaN	N

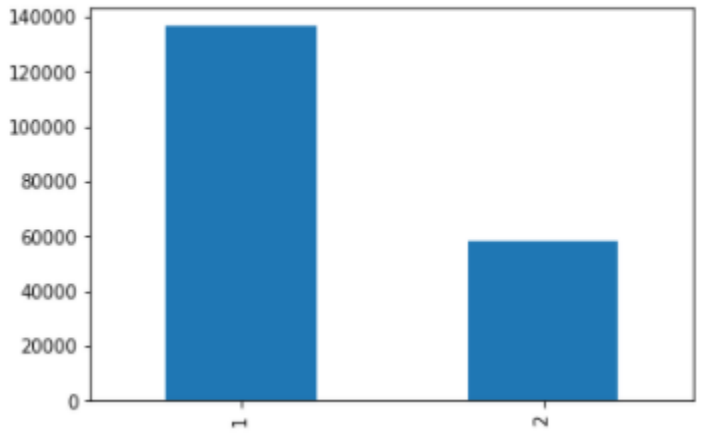
```
df.dtypes.sample(10)
```

```
VEHCOUNT          int64
ADDRTYPE         object
HITPARKEDCAR     object
LIGHTCOND        object
SEVERITYCODE      int64
COLLISIONTYPE    object
WEATHER          object
SPEEDING         object
INATTENTIONIND   object
UNDERINFL        object
dtype: object
```

Balance Dataset

Looking at the data it is easy to see that the data for Severity = 1 and Severity = 2 are unbalanced, see below.

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```



Before diving into exploratory analysis this needs to be adjusted.

```
#  It is not balanced so we will need to downsample Severity Code = 1
df_majority = df[df.SEVERITYCODE == 1]
df_minority = df[df.SEVERITYCODE == 2]
df_majority_downsampled = resample(df_majority,
                                   replace=False,
                                   n_samples=58188,
                                   random_state=145470)
#this df is now downsampled
df = pd.concat([df_majority_downsampled, df_minority])
df.SEVERITYCODE.value_counts()
```

```
2    58188
1    58188
Name: SEVERITYCODE, dtype: int64
```

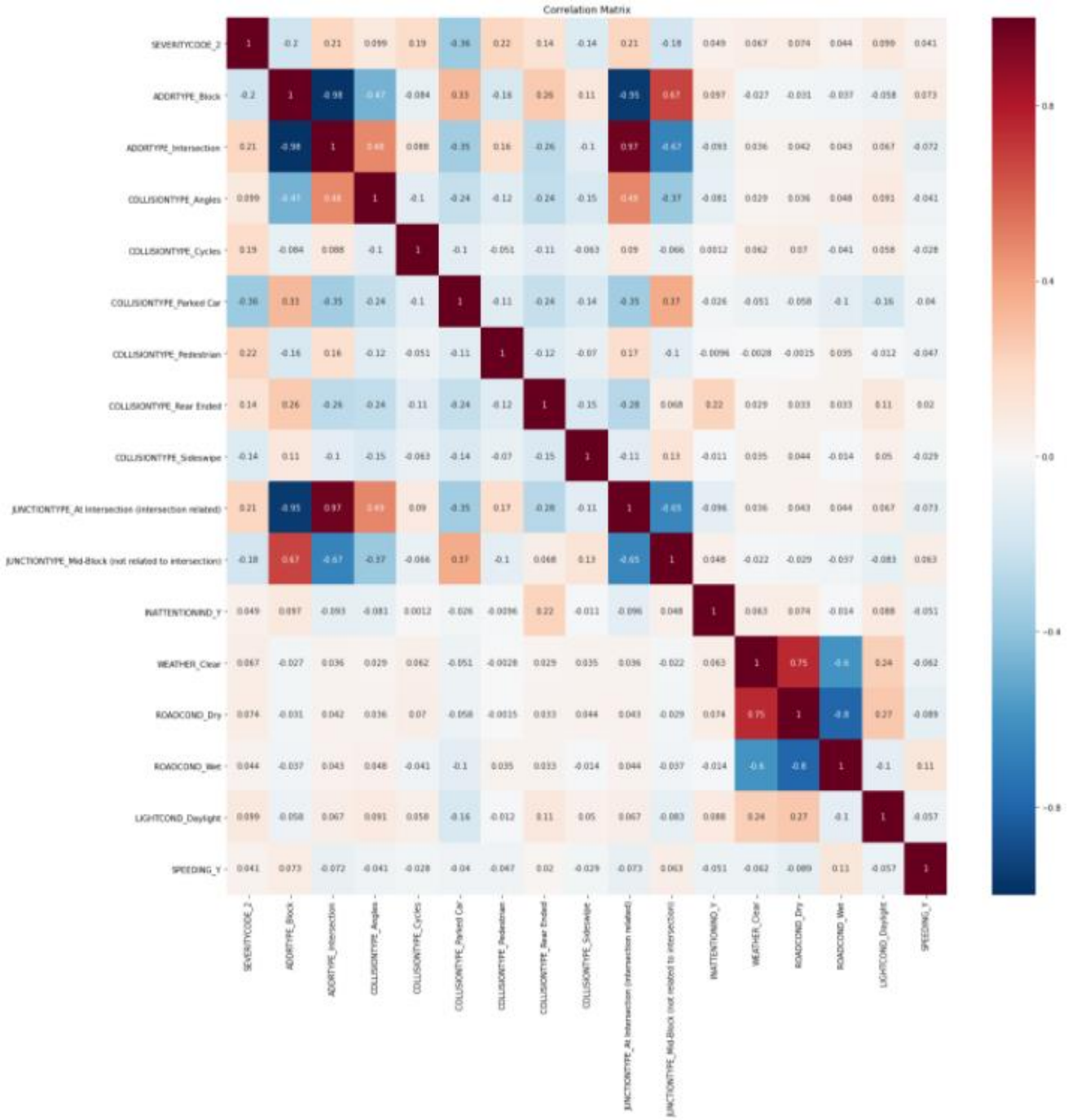
Now that the table is balanced, I then used One-hot encoding to convert these text objects into binary numbers, making it more manageable when using Logistic Regression, KNN, and Decision Trees. Here is an example of that dataset, below.

	VEHCOUNT	SEVERITYCODE_2	ADDRTYPE_Alley	ADDRTYPE_Block	ADDRTYPE_Intersection	COLLISIONTYPE_Angles	COLLISIONTYPE_Cycles	COLLISIONTYPE_Head On	COLLISIONTYPE_Left Turn
72120	2	0	0	1	0	0	0	0	0
184848	2	0	0	0	1	0	0	0	1
179921	2	0	0	0	1	1	0	0	0
42749	2	0	0	1	0	0	0	0	0
191968	3	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
194663	2	1	0	1	0	1	0	0	0
194666	2	1	0	1	0	1	0	0	0
194668	2	1	0	1	0	0	0	1	0
194670	2	1	0	0	1	0	0	0	1
194671	1	1	0	0	1	0	1	0	0

Then I defined x and y and normalized the dataset by changing the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Now the data is ready for exploratory analysis.

Exploratory Analysis

I will start with a correlation matrix of the balanced dataset to see which variables most affect Severity.



A lot to take in with this graph but there are some obvious outcomes I.e. the weather is clear when the road conditions are dry and the collision type = Angles when the junction type is at an intersection.

While this is a great starting point for some broad assumptions, we need to take a deeper dive.

Train/Test Dataset

Before testing the accuracy of several methods, it is important to split the data into two sets: a training set and a testing set. We will use 20% of our data for testing and 80% for training. After splitting the data, we will use three different methods to test accuracy:

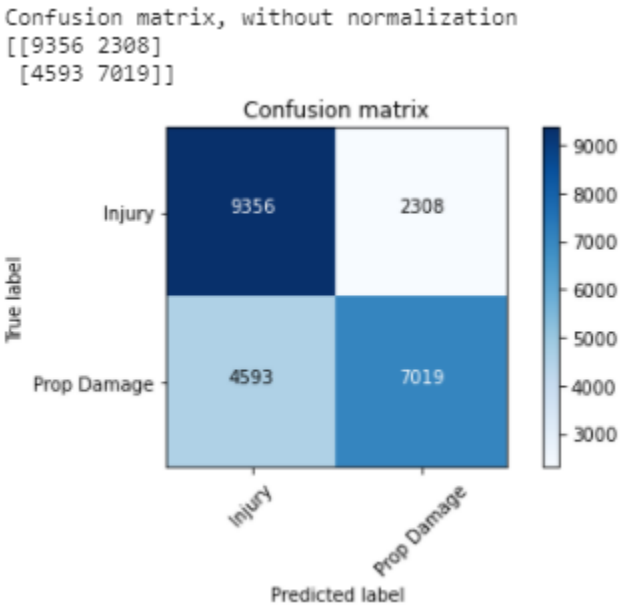
1. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.
2. Decision trees can help organizations structure and automate (complex) information. Decision trees are decision models that answer a specific question based on a question structure and certain conditions.
3. The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows.

After running the three methods above and taking into the type and size of the dataset, Logistic regression has the best fit and accuracy. So, we will use this moving forward and we can use this method to forecast accident severity in Seattle with 70.35% accuracy.

Logistic Regression Accuracy Score: 0.7035143495445952
Decision Tree Accuracy Score: 0.699948444749957
K-Nearest Neighbors Accuracy Score: 0.6895514693246262

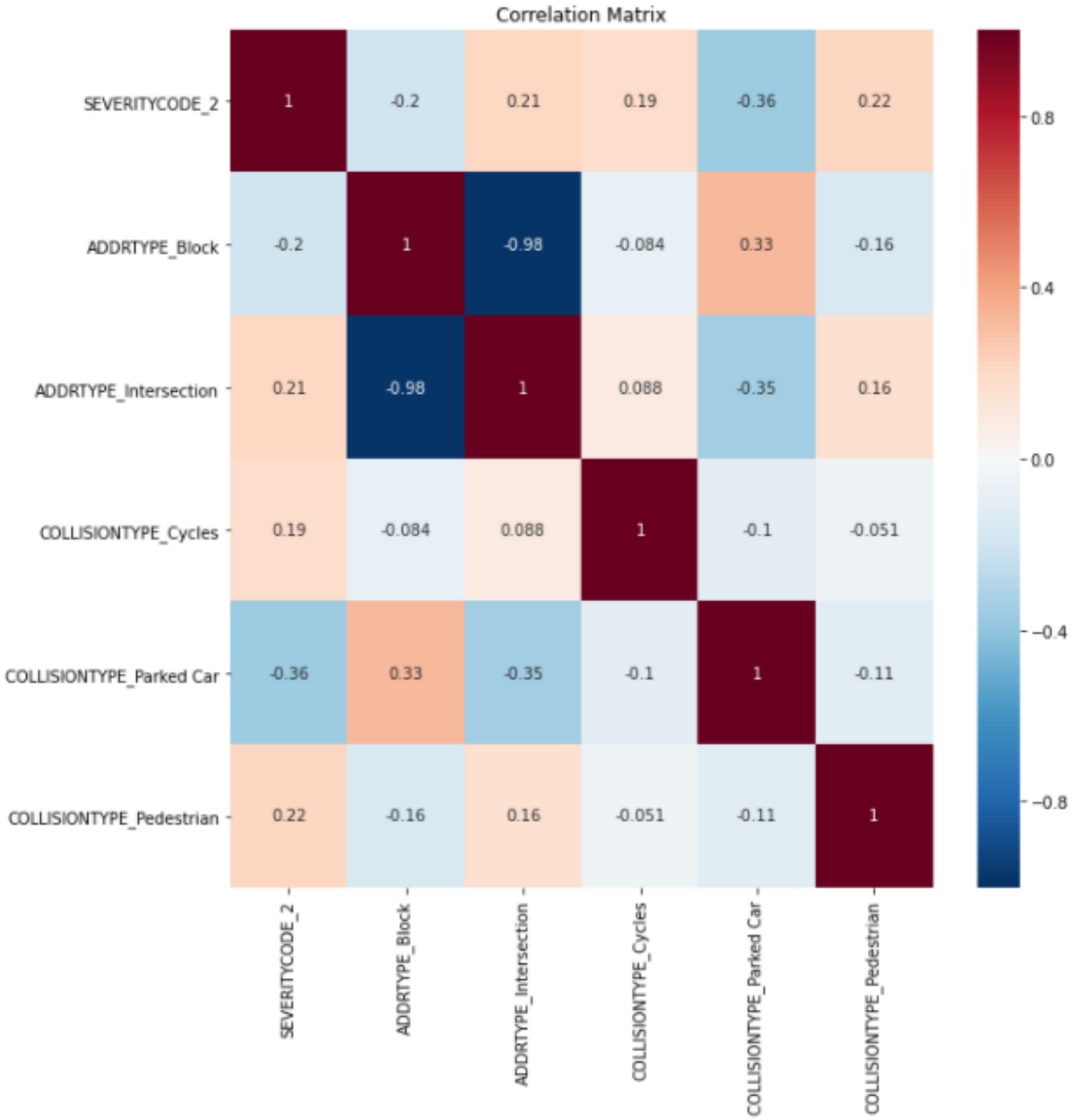
Results

Using Logistic Regression, we can produce the confusion matrix, below.



This tells us that out of 23,276 total instances, 11,664 result in an injury and 11,612 result in prop damage. Note that this is a balanced dataset. Out of the 11,612 instances of prop damage, the classifier correctly predicted 7,019 as prop damage and 4,593 an injury. Looking at the top row for Injury we can see that it has done a better job in predicting injury. This is because out of 11,664 instances the classifier correctly predicted 9,356 as injury and incorrectly predicted 2,308. Correct 80.2% while prop damage is correct 60.4% for those instances.

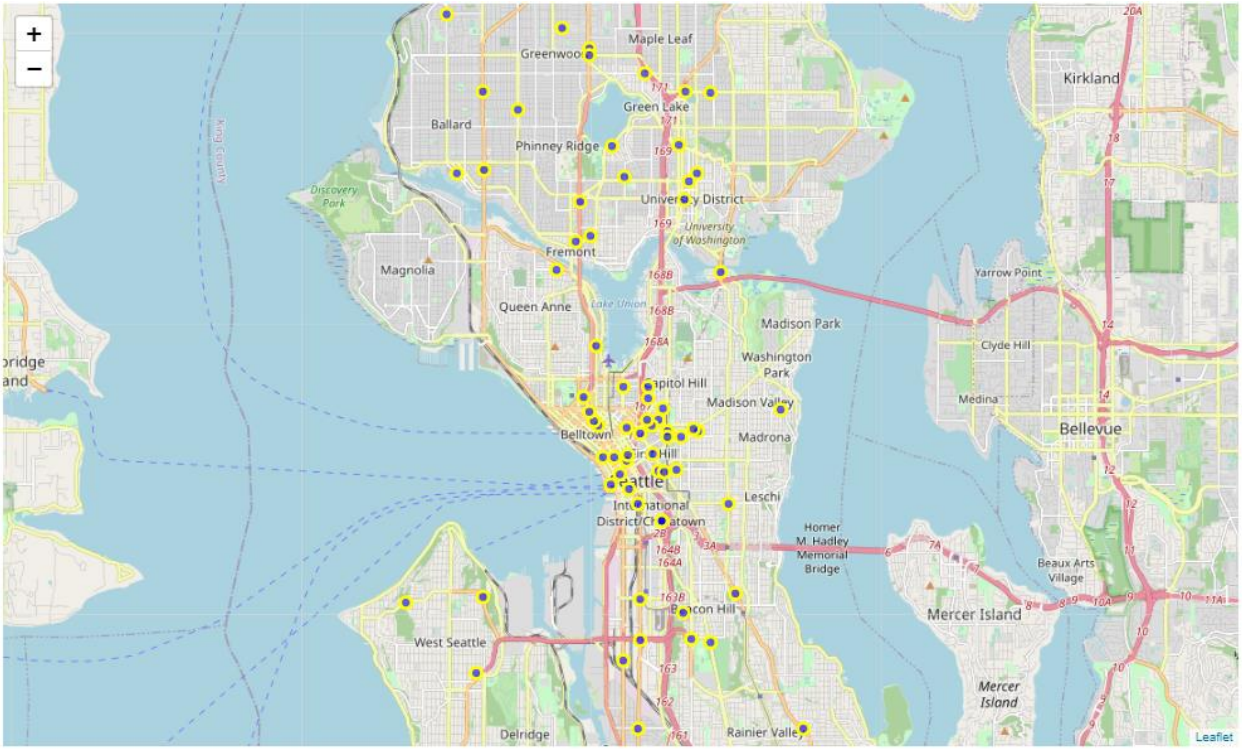
This tells us that by using Logistic Regression we can predict with higher accuracy an injury rather than prop damage. So, with that said we will look at "SEVERITYCODE_2" which is defined as an injury. Using the Correlation Matrix below we can see that the highest positive correlation for an injury was a collision with a pedestrian, bicyclists, and at an intersection. While the highest negative correlation to an injury in an accident is accidents happening with a block and within a parked car.

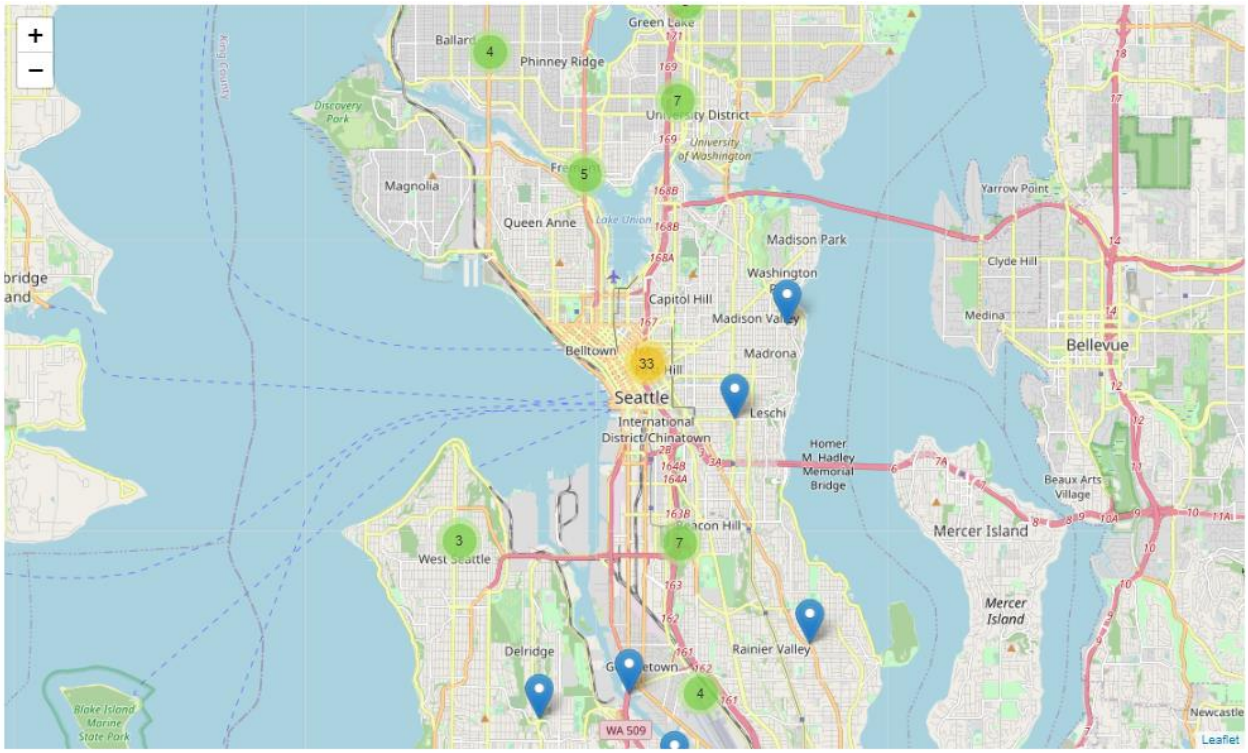


	precision	recall	f1-score	support
0	0.75	0.60	0.67	11612
1	0.67	0.80	0.73	11664
micro avg	0.70	0.70	0.70	23276
macro avg	0.71	0.70	0.70	23276
weighted avg	0.71	0.70	0.70	23276

Taking this a step further using the folium package we can graph the incidents where there was an injury at an intersection from the filtered dataset below.

	SEVERITYCODE	X	Y	ADDRTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
0	2	-122.323148	47.703140	Intersection	2	0	0	2
4	2	-122.306426	47.545739	Intersection	2	0	0	2
7	2	-122.320780	47.614076	Intersection	3	0	1	1
9	2	-122.384700	47.528475	Intersection	2	0	0	2
16	2	-122.344539	47.692012	Intersection	3	0	0	2
...	...	...	...	...	...	...	...	...
194649	2	-122.302382	47.626759	Intersection	6	0	0	3
194655	2	-122.380016	47.664879	Intersection	4	0	0	0
194656	2	-122.340474	47.614496	Intersection	2	1	0	1
194670	2	-122.306689	47.683047	Intersection	3	0	0	2
194671	2	-122.355317	47.678734	Intersection	2	0	1	1


Maple Leaf
Greenwood
Green Lake
Ballard
Phinney Ridge
Kirkland
Discovery Park
University District
University of Washington
Fremont
Magnolia
Queen Anne
Lake Union
Madison Park
Yarrow Point
Clyde Hill
Washington Park
Capitol Hill
Madison Valley
Medina
Belltown
Madrona
Bellevue
Seattle
Leschi
International District/Chinatown
Homer M. Hadley Memorial Bridge
Beaux Arts Village
Mercer Island
West Seattle
Delridge
Rainier Valley
Leaflet


Green Lake
Phinney Ridge
Discovery Park
University District
University of Washington
Magnolia
Queen Anne
Lake Union
Madison Park
Yarrow Point
Clyde Hill
Washington Park
Capitol Hill
Madison Valley
Medina
Belltown
Madrona
Bellevue
Seattle
Leschi
International District/Chinatown
Homer M. Hadley Memorial Bridge
Beaux Arts Village
Mercer Island
West Seattle
Delridge
Rainier Valley
Blake Island Marine State Park
WA 509
Newcastle
Leaflet

Discussion

As stated above, using Logistic Regression at 70.4% accuracy, we come up with several assumptions about accidents involving an injury. The highest positive correlation to an injury was a collision with a pedestrian, bicyclists, and at an intersection. While the highest negative correlation to an injury in an accident is accidents happens within a block and with a parked car.

Weather, speed, and lighting had to be taken out because either there were too many NULLs or the correlation was irrelevant.

Conclusion

While there are many results to be drawn from this data. After balancing the data set and removing the variables with the large unknown values. My advice to the city of Seattle would be to invest in more bike/pedestrian friendly efforts. Brighter flashing crosswalk lights, more bike lanes, longer walk times for crosswalk lights, and longer times between red light to crosswalk walk light. These efforts should bring more awareness to bikers, and pedestrian around intersections which as we know are a large cause of accidents in Seattle that result in injuries. By implementing some of these new features we can hope to decrease traffic injuries in Seattle.