Coursera Capstone Project: Applied Data Science Capstone

Nikhil Namburi

Predicting Accident Severity in Seattle Given Certain Conditions and Accident Types

1. Introduction

1.1 Background

Home to some of the world's largest companies like Boeing and Microsoft, as well as cultural staples like Pike Place Market and the Space Needle, Seattle has established itself as one of the most prominent cities on the American west coast. In just the last forty years, the population has doubled to nearly 4 million people. Along with this tremendous growth arrived the various challenges that confront large cities, including the issue of traffic. As the 14th most congested city in the United States, Seattle requires its citizens to spend hundreds of hours each year on the road. It is imperative that Seatlleites feel safe while driving in their home city.

1.2 Problem

In order to help create a safer driving environment in Seattle, drivers and local officials should be able to identify conditions and types of accidents with a greater incidence of injury. By predicting the severity of accidents under different weather, visibility, and time conditions, Seattle's residents can avoid adverse conditions and reduce the likelihood of falling victim to a severe accident. In addition, drivers can exercise extra caution to avoid certain types of accidents with particularly high injury rates.

1.3 Interest

Any resident of the Seattle Metropolitan area would benefit from this information, as it would allow them to identify the safest conditions and accident types for driving. Public officials could also use this information to prepare for the likelihood of severe accidents given certain conditions and accident types. Lastly, insurance companies could better predict the likelihood of a severe claim for individuals that more regularly drive in adverse conditions.

2. Data

2.1 Data Sources

This project's dataset was gathered by the Seattle Police Department and Traffic Records Department since 2004 to identify factors that cause accidents. The dataset lists 65,534 accidents and 37 variables in those accidents. These variables can provide insight into the various factors that contribute to accidents in the Seattle area.

2.2 Feature Selection

I have selected features that provide information about specific driver conditions and intersections. 'ST_COLDESC' were used to identify specific types of accidents. 'WEATHER', 'ROADCOND', and 'LIGHTCOND' were used to identify weather, road conditions, and light conditions, respectively. Of course, 'SEVERITY CODE' would be included, as it is the dependent or target variable that is being predicted in this project.

2.3 Data Cleaning

All the features not mentioned in the Feature Selection (2.2) section will be removed. Then, any accidents with a null value (among the remaining features) will be removed from the dataset. In doing so, the predictive model will have the necessary information to make predictions about the severity of an accident.

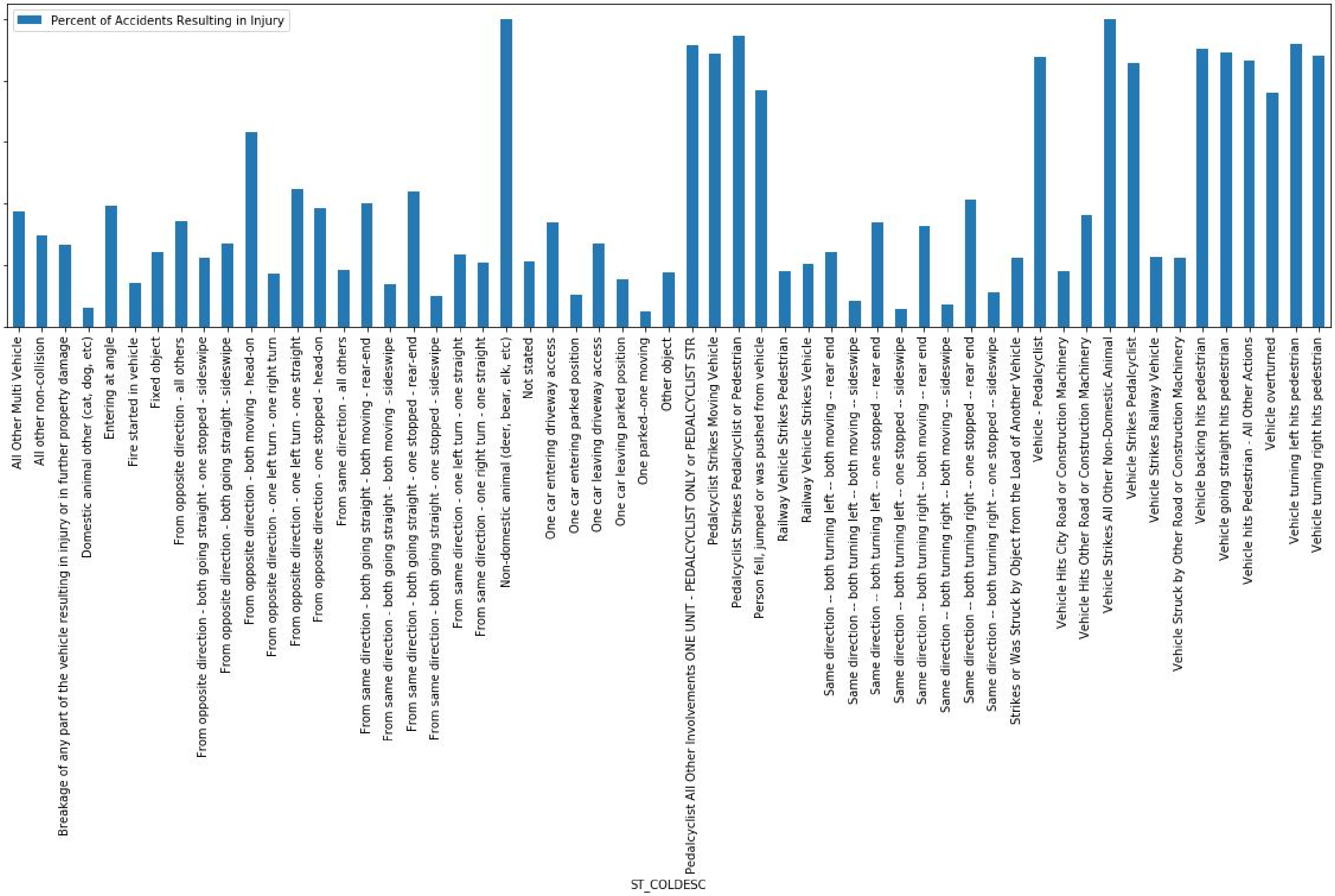
3. Methodology

I will employ exploratory data analysis to uncover insights about the impact of each accident type, weather condition, road condition, and light condition in predicting the severity of an accident. Then, by identifying the most impactful variables, I will include those variables into a balanced logistic regression that classifies each accident as “severe” or “not severe.” With this information, local officials can determine when and where to prepare resources needed to handle severe accidents.

4. Exploratory Data Analysis

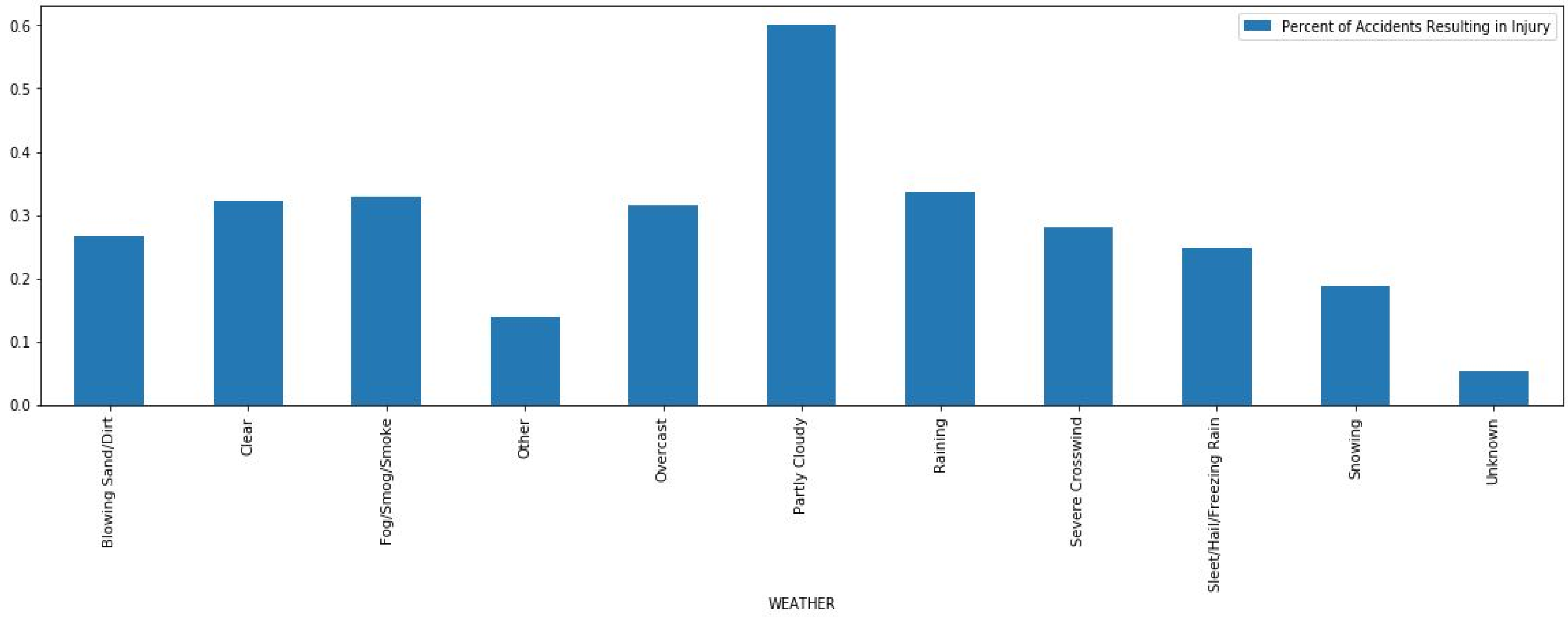
4.1 Accident Type as a Predictor of Accident Severity

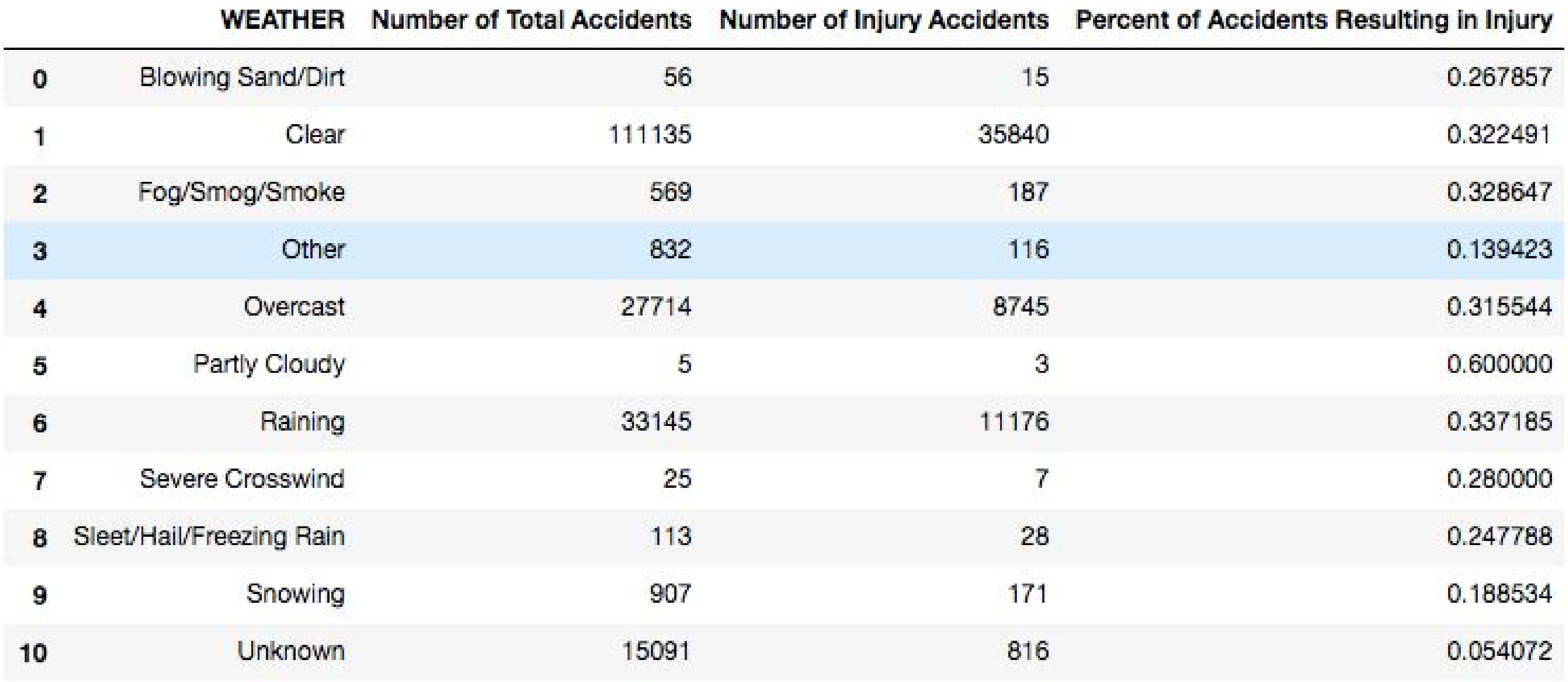
In order to identify certain accident types that had a high incidence rate of severe accidents, I created a table and graph that determined the total number of accidents in each accident type, as well as a table and graph that categorized severe accidents according to accident type. Then, I merged the tables, calculated the percentage of accidents that were severe at each given accident type, and presented the information as a graph.


Percent of Accidents Resulting in Injury
1.0
0.8
0.6
0.4
0.2
0.0
All Other Multi Vehicle
All other non-collision
Breakage of any part of the vehicle resulting in injury or in further property damage
Domestic animal other (cat, dog, etc)
Entering at angle
Fire started in vehicle
Fixed object
From opposite direction - all others
From opposite direction - both going straight - one stopped - sideswipe
From opposite direction - both going straight - sideswipe
From opposite direction - both moving - head-on
From opposite direction - one left turn - one right turn
From opposite direction - one left turn - one straight
From opposite direction - one stopped - head-on
From same direction - all others
From same direction - both going straight - both moving - rear-end
From same direction - both going straight - both moving - sideswipe
From same direction - both going straight - one stopped - rear-end
From same direction - both going straight - one stopped - sideswipe
From same direction - one left turn - one straight
From same direction - one right turn - one straight
Non-domestic animal (deer, bear, elk, etc)
Not stated
One car entering driveway access
One car entering parked position
One car leaving driveway access
One car leaving parked position
One parked--one moving
Other object
Pedalcyclist All Other Involvements ONE UNIT - PEDALCYCLIST ONLY or PEDALCYCLIST STR
Pedalcyclist Strikes Moving Vehicle
Pedalcyclist Strikes Pedalcyclist or Pedestrian
Person fell, jumped or was pushed from vehicle
Railway Vehicle Strikes Pedestrian
Railway Vehicle Strikes Vehicle
Same direction -- both turning left -- both moving -- rear end
Same direction -- both turning left -- both moving -- sideswipe
Same direction -- both turning left -- one stopped -- rear end
Same direction -- both turning left -- one stopped -- sideswipe
Same direction -- both turning right -- both moving -- rear end
Same direction -- both turning right -- both moving -- sideswipe
Same direction -- both turning right -- one stopped -- rear end
Same direction -- both turning right -- one stopped -- sideswipe
Strikes or Was Struck by Object from the Load of Another Vehicle
Vehicle - Pedalcyclist
Vehicle Hits City Road or Construction Machinery
Vehicle Hits Other Road or Construction Machinery
Vehicle Strikes All Other Non-Domestic Animal
Vehicle Strikes Pedalcyclist
Vehicle Strikes Railway Vehicle
Vehicle Struck by Other Road or Construction Machinery
Vehicle backing hits pedestrian
Vehicle going straight hits pedestrian
Vehicle hits Pedestrian - All Other Actions
Vehicle overturned
Vehicle turning left hits pedestrian
Vehicle turning right hits pedestrian
ST_COLDESC

4.2 Weather Condition as a Predictor of Accident Severity

In order to identify certain weather conditions that had a high incidence rate of severe accidents, I created a table and graph that determined the total number of accidents with certain weather conditions, as well as a table and graph that categorized severe accidents according to weather conditions. Then, I merged the tables, calculated the percentage of accidents that were severe under each given weather condition, and presented the information as a graph.



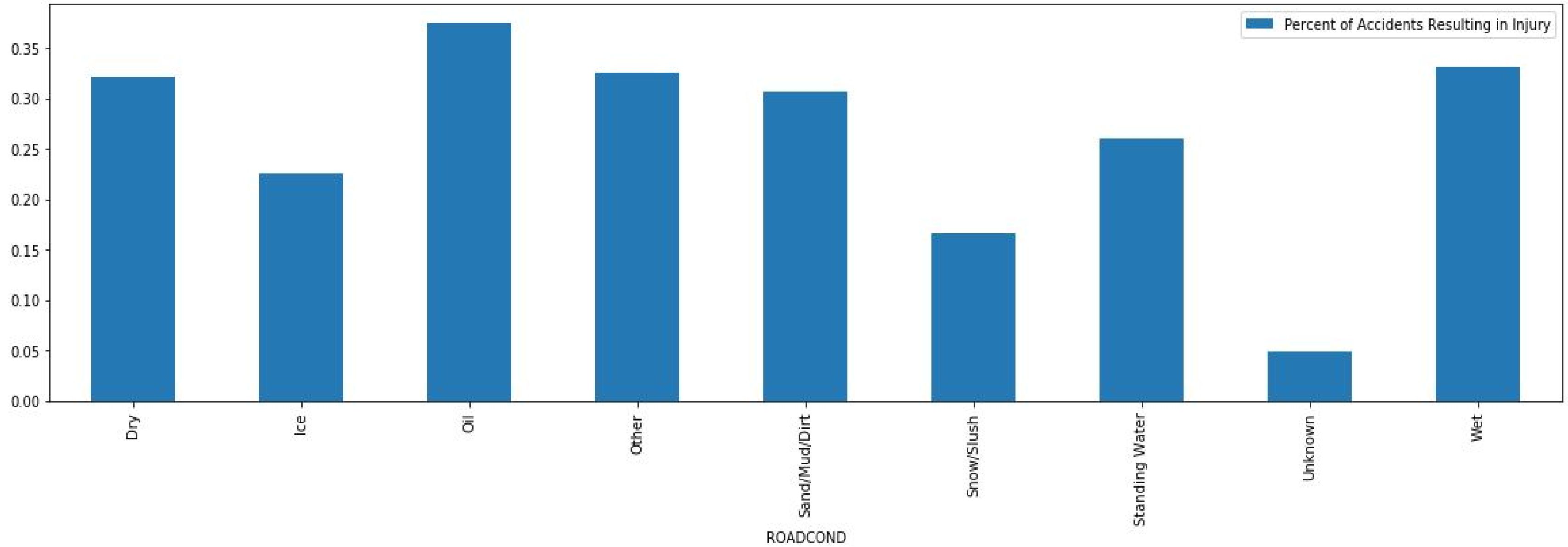


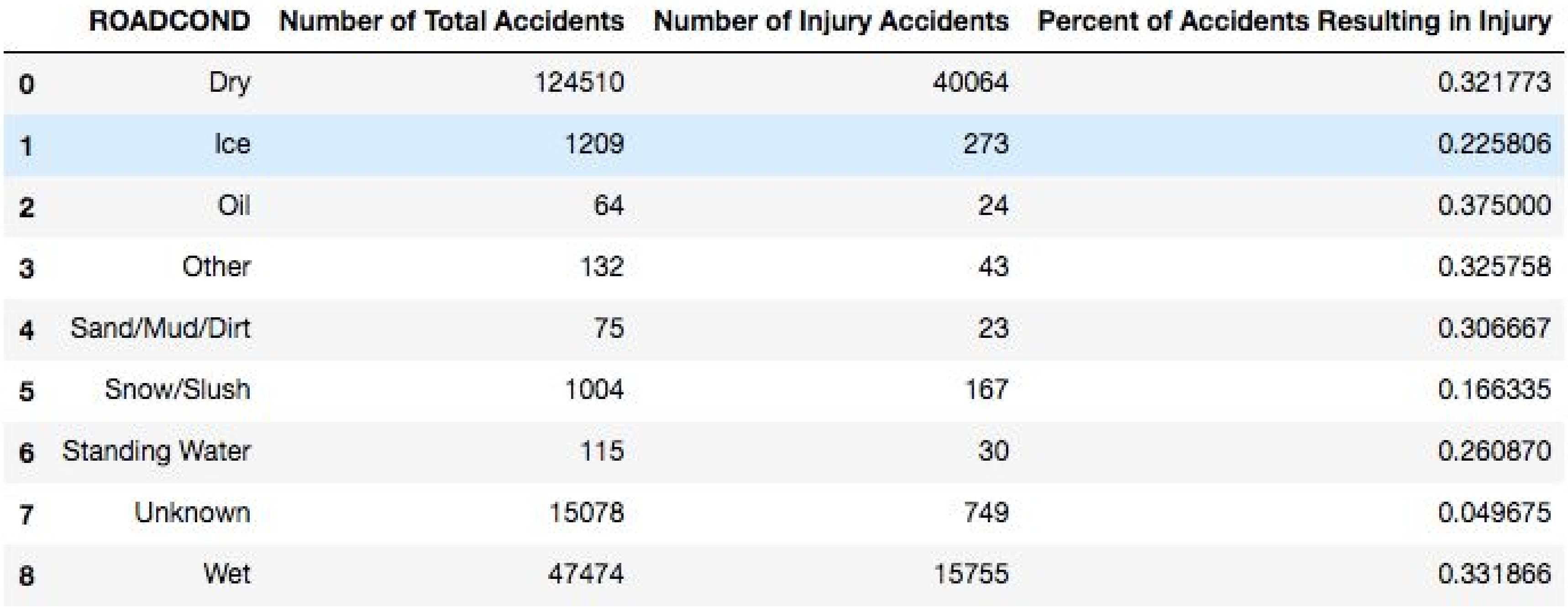
	WEATHER	Number of Total Accidents	Number of Injury Accidents	Percent of Accidents Resulting in Injury
0	Blowing Sand/Dirt	56	15	0.267857
1	Clear	111135	35840	0.322491
2	Fog/Smog/Smoke	569	187	0.328647
3	Other	832	116	0.139423
4	Overcast	27714	8745	0.315544
5	Partly Cloudy	5	3	0.600000
6	Raining	33145	11176	0.337185
7	Severe Crosswind	25	7	0.280000
8	Sleet/Hail/Freezing Rain	113	28	0.247788
9	Snowing	907	171	0.188534
10	Unknown	15091	816	0.054072

The graph shows that there is little variation between the different types of weather conditions. The only condition with a particularly different rate of severe accident is “Partly Cloudy.” After examining the total number of accidents that happened under “Partly Cloudy” conditions, I discovered that there were only 5 such accidents, making the sample size small enough to make the rate of severity inconsequential. Counterintuitively, weather conditions that are typically viewed as more dangerous have relatively low rates of severe accidents. Weather such as “snow” and “sleet/hail/freezing rain” actually had the lowest rate of severe accidents among all defined weather conditions. Perhaps drivers were more cautious in poor weather conditions, and that lowered the incidence of injury in accidents.

4.3 Road Condition as a Predictor of Accident Severity

In order to identify certain road conditions that had a high incidence rate of severe accidents, I created a table and graph that determined the total number of accidents with certain road conditions, as well as a table and graph that categorized severe accidents according to road conditions. Then, I merged the tables, calculated the percentage of accidents that were severe under each given road condition, and presented the information as a graph.



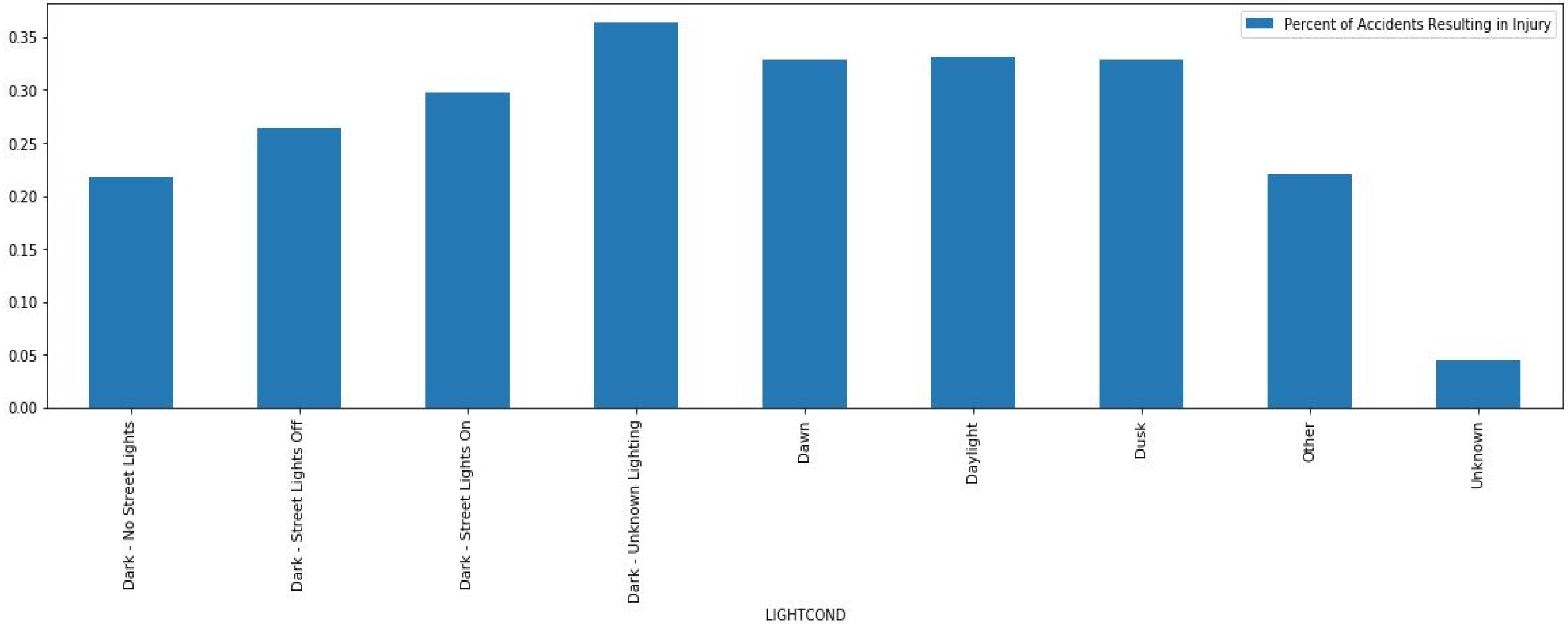


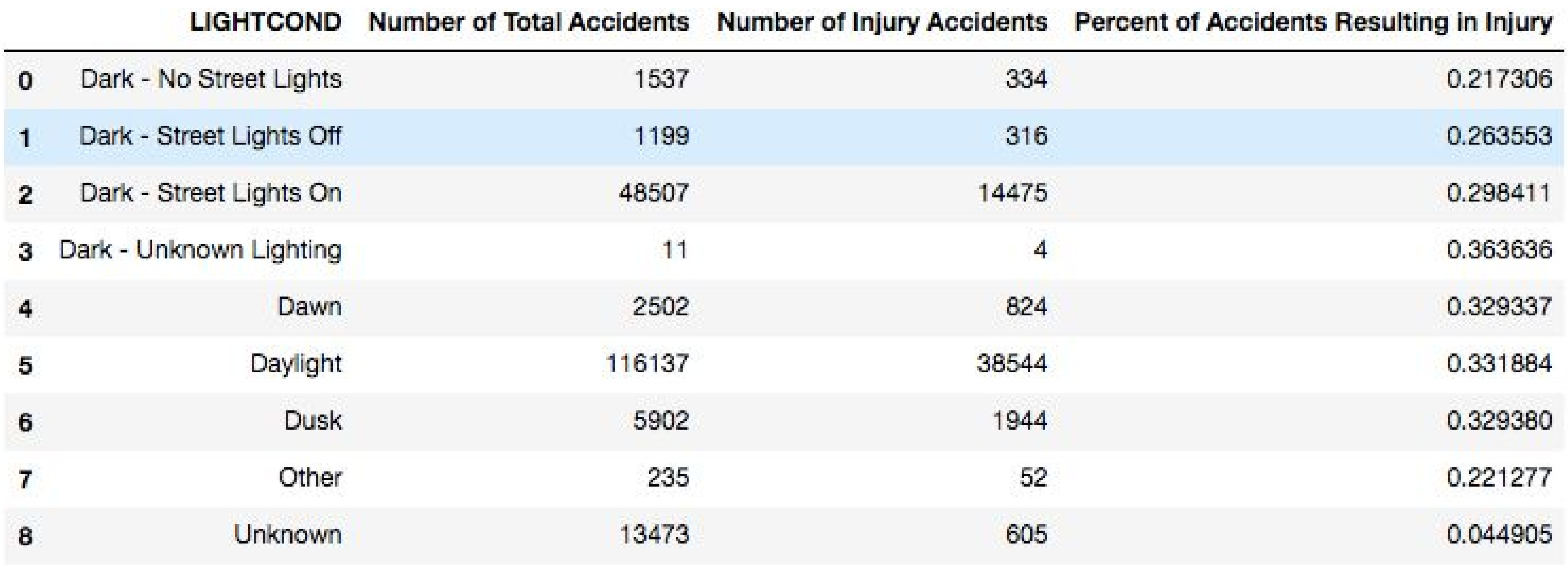
	ROADCOND	Number of Total Accidents	Number of Injury Accidents	Percent of Accidents Resulting in Injury
0	Dry	124510	40064	0.321773
1	Ice	1209	273	0.225806
2	Oil	64	24	0.375000
3	Other	132	43	0.325758
4	Sand/Mud/Dirt	75	23	0.306667
5	Snow/Slush	1004	167	0.166335
6	Standing Water	115	30	0.260870
7	Unknown	15078	749	0.049675
8	Wet	47474	15755	0.331866

Again, it appears that the severity rate of accidents remained fairly constant across different road conditions. The most common road conditions, “Dry” and “Wet”, both had approximately a third of accidents resulting in injuries. The high severity rate of the most dangerous condition, “Oil”, can be attributed to a low number of total accidents, meaning that a few severe accidents would greatly impact the total severity rate. Similar to the analysis of accidents under given weather conditions, I discovered that there was a relatively (and significantly) low severity rate during icy and snowy road conditions. When the road conditions were listed as “Ice”, the accident severity rate was close to 23%, while the accident severity rate was 17% for roads with “Snow/Slush.” The exact cause of the lower injury rate cannot be confirmed, but I referred to extra driver caution as a possible cause in the Weather Condition section (3.2), and similar reasoning can be applied to road conditions.

4.4 Light Condition as a Predictor of Accident Severity

In order to identify certain light conditions that had a high incidence rate of severe accidents, I created a table and graph that determined the total number of accidents with certain light conditions, as well as a table and graph that categorized severe accidents according to light conditions. Then, I merged the tables, calculated the percentage of accidents that were severe under each given light condition, and presented the information as a graph.



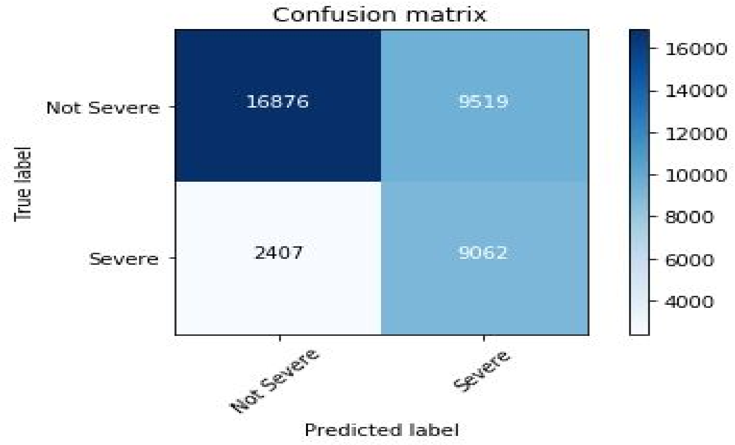


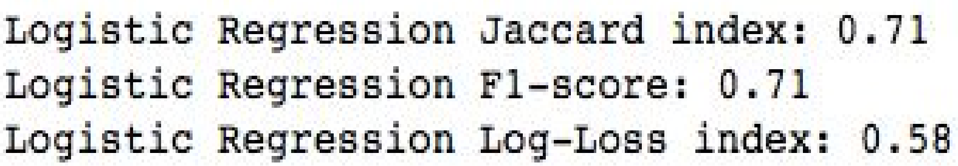
	LIGHTCOND	Number of Total Accidents	Number of Injury Accidents	Percent of Accidents Resulting in Injury
0	Dark - No Street Lights	1537	334	0.217306
1	Dark - Street Lights Off	1199	316	0.263553
2	Dark - Street Lights On	48507	14475	0.298411
3	Dark - Unknown Lighting	11	4	0.363636
4	Dawn	2502	824	0.329337
5	Daylight	116137	38544	0.331884
6	Dusk	5902	1944	0.329380
7	Other	235	52	0.221277
8	Unknown	13473	605	0.044905

From the graph showing the accident severity rate as a function of light conditions, there appears to be no light condition that significantly increases or decreases the likelihood of an accident resulting in injury. The likelihood of a severe accident is about one in three for common light conditions. When most drivers are on the road with good lighting conditions at “Dawn”, “Daylight”, and “Dusk”, the accident severity rate remains constant among the three conditions. However, the likelihood of an accident resulting in injury diminishes considerably when lighting is “Dark” with either “Street Lights Off” or “No Street Lights”. This trend plays into the earlier narrative of increased driver caution in poor conditions, which can be found in sections (3.2) and (3.3).

5. Results

5.1 Logistic Regression Model and Accuracy Measurements





```
Logistic Regression Jaccard index: 0.71
Logistic Regression F1-score: 0.71
Logistic Regression Log-Loss index: 0.58
```

5.2. Model Summary

A logistic regression model was applied first to use each variable as a predictor of accident severity. As expected, weather, road, and light conditions were not good predictors of an accident's severity, and the models for these variables just predicted every accident to be "not severe". After balancing the data, the models had the opposite problem, predicting every accident to be "severe". As a result, I decided that those conditions would be ultimately inconsequential.

Conversely, the type of accident was fairly consequential in the severity of a given accident. A balanced logistic regression model for accident type as a predictor created a confusion matrix similar to the one above. From this model, I inferred that accident type would be the most important feature in predicting the severity of an accident. So I developed a model that took into account weather conditions, light conditions, road conditions, and accident type as predictors. This model was created with the anticipation that accident type would definitely be the most important predicting feature.

The model did a fairly good job in its predictions, with a particularly low rate of False Negatives (predictions that an accident was not severe when it was). By balancing the Logistic Regression model, the model shifted from an extraordinarily high rate of False Negatives and very few False Positives to a fairly moderate rate of each. Especially in a context that emphasizes predicting severe accidents, the increased rate of False Positives is justifiable because it is accompanied by a significant increase in True Positives. Without balancing the data, the model classified nearly every single accident as "not severe."

6. Discussion

The exploratory data analysis and logistic regression models reveal surprising and helpful insights to both city officials and drivers in avoiding severe accidents. From the information gathered in the exploratory data analysis of the impact of weather, road, and light conditions on accident severity, the actual rate of severe accidents remained largely unbothered by the conditions of an accident. If anything, the conditions actually had an opposite impact than expected, with a lower accident severity rate occurring in "dangerous" conditions like snow, slush, ice, and dark lighting. Although the exact cause of the lower severity rate is uncertain, I suggest that the lower rate may be attributed to greater caution from the drivers. Further information can be inferred by using the "INATTENTIONIND", "UNDERINFL", "SPEEDING" features, which refer to the inattention, drunkenness, and traveling speed of the driver. If more information was collected about driver behavior, it would be a worthwhile effort to correlate certain behavior to the severity of accidents.

Drivers can use the information about the severity rates of certain accident types to be more cautious when performing certain driving maneuvers. In particular, the severity rate of accidents involving pedestrians and pedalcyclists are alarmingly (and expectedly) high. The information from this project can be used as an educational tool for drivers, pedestrians, and pedalcyclists alike to exercise greater caution when interacting with one another. Local officials could plan for the severe nature of these accidents by constructing city and transportation policies that prevent these accidents. Similar extensions could be made to any other type of accident that had a particularly high rate of incidence and injury.

7. Conclusion

In this study, I observed the impact of weather conditions, road conditions, light conditions, and accident type on the severity of accidents in the Seattle area. Using a dataset compiled since 2004 by the Seattle Police Department and Traffic Records Department, I extracted the aforementioned information from nearly 200,000 accidents and analyzed those accidents for patterns that were correlated with severe accidents. With that exploratory data analysis, I concluded that only the type of accident seemed to have a significantly meaningful impact on the severity of accidents. Then, I built and deployed a logistic regression machine learning model that incorporated weather conditions, light conditions, road conditions, and accident type to predict whether a given accident is severe or not severe. Similar exploratory data analyses and logistic regression models can be replicated for other cities and can take into account other factors that might impact the severity of an accident. The project could also be expanded to compare the number of accidents to the total number of cars on the road under given conditions. The models in this project contribute important information to the ultimate goal of improving driver and passenger safety on the streets of Seattle and across the world.