Predicting Car Accident Severity

Seattle Traffic Collisions from 2004 - 2020

Rohan Lewis

2021.02.11

Predicting the severity of automobile collisions is invaluable for emergency services.

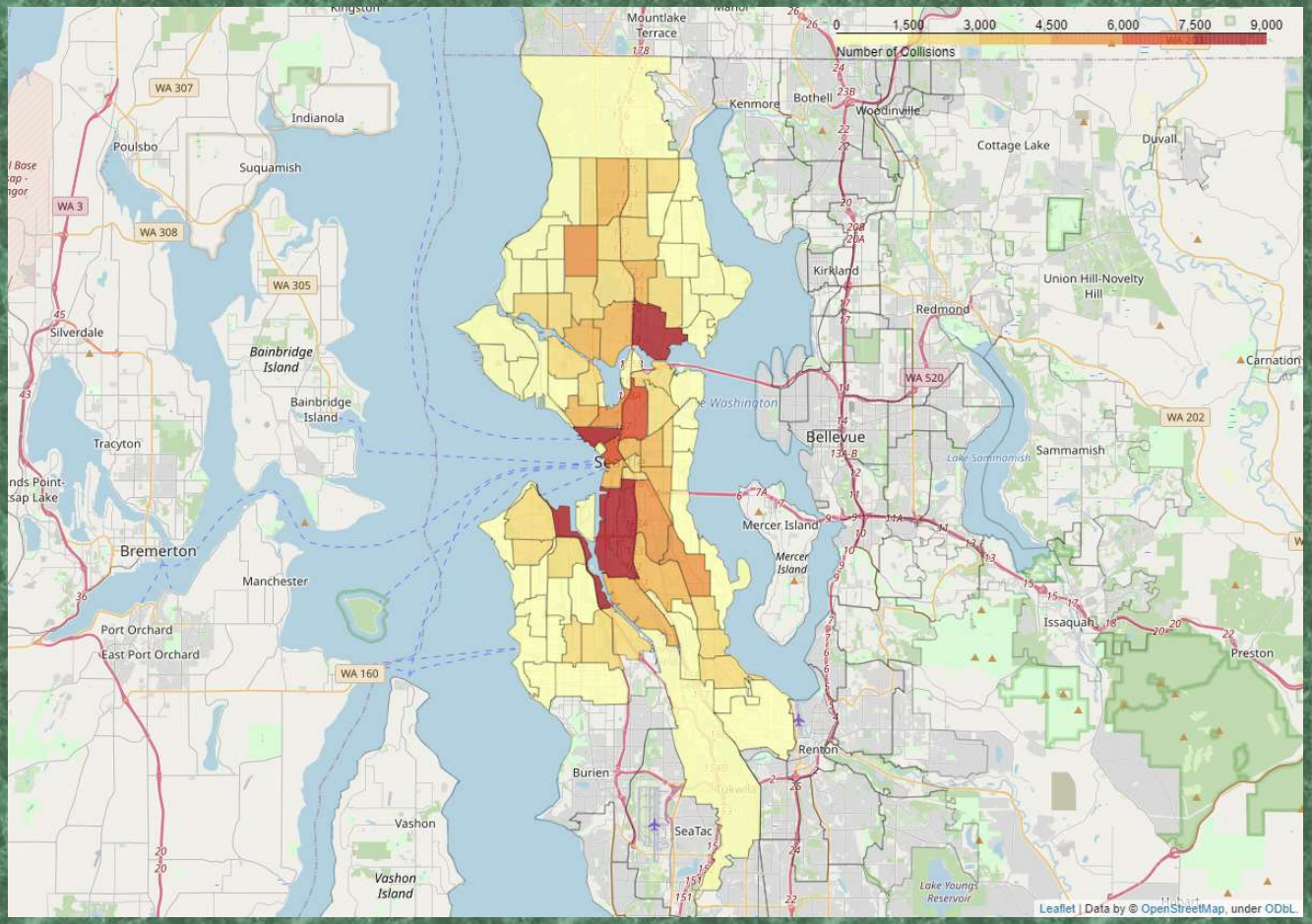
- Traffic accidents are a common occurrence of our daily life, commuting to and from work, the grocery store, and gym. They can be in a rural or urban location. They can affect pedestrians, bicyclists, commuters, public transportation, and truck drivers. They occur throughout the year in all weather conditions.
- “In 2010, there were 32,999 people killed, 3.9 million were injured, and 24 million vehicles were damaged in motor vehicle crashes in the United States. The economic costs of these crashes totaled $242 billion.”(NHTSA)
- Accidents can result in property damage or injury.
- Predicting the severity of these accidents would be beneficial to allocate emergency response teams more efficiently at particular locations and specific times.

Data Acquisition and Cleaning

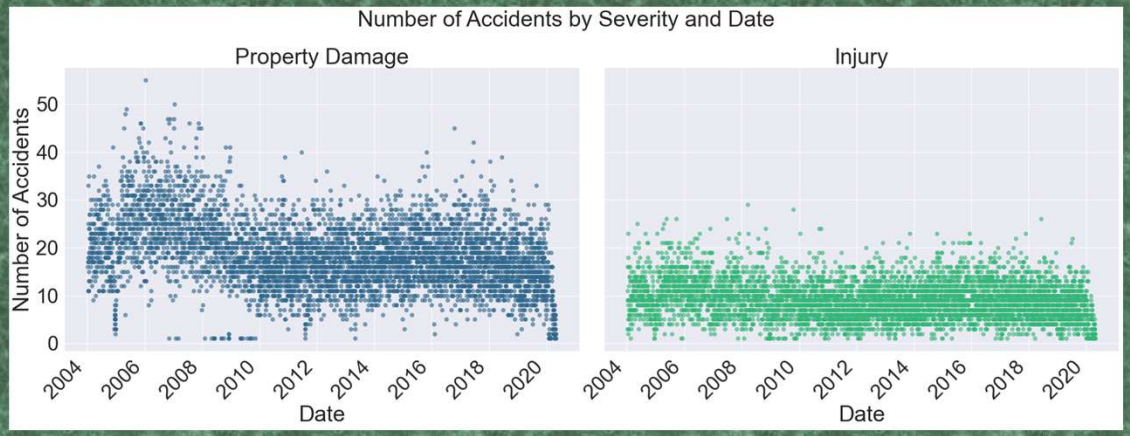
- The dataset provided is a collaborative effort by Seattle Police Depart and Seattle Department of Transportation. It has approximately 195,000 collisions of various types from Jan 2004 - May 2020.
- Seattle Neighborhood boundaries were acquired from GitHub.
- In order to predict accident severity, I eliminated variables not resulting from the accident. I kept variables related to the setting. These were location (neighborhood and address type), date and time, light conditions, road conditions, and weather conditions.
- 48,318 collisions, or approximately 24.8% of the original 194,673 collisions in the dataset, had missing or 'Unknown' values, and were removed. 146,355 collisions remain for analysis.
- 98,273 accidents, or 67.1%, caused property damage.
- 48,082 accidents, or 32.9%, caused injury.

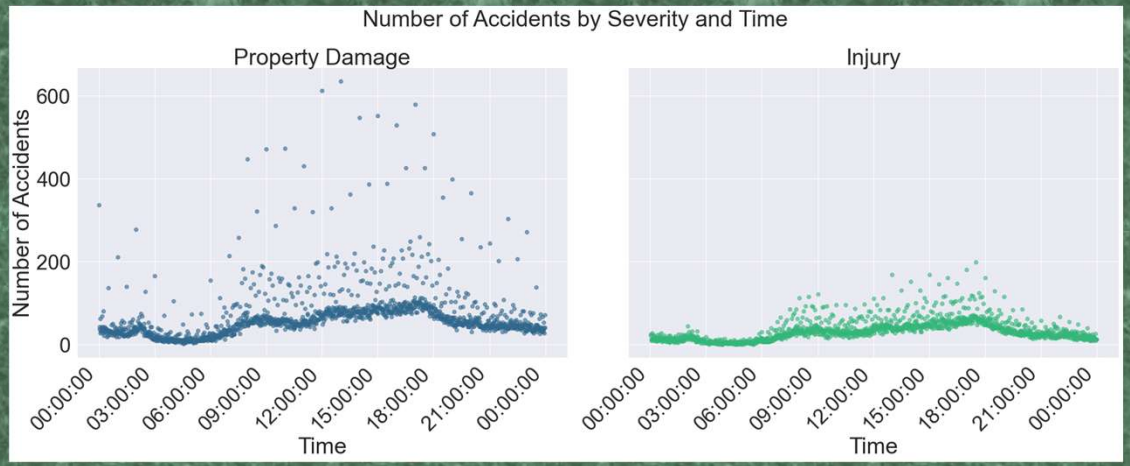
Car Accidents by Neighborhood

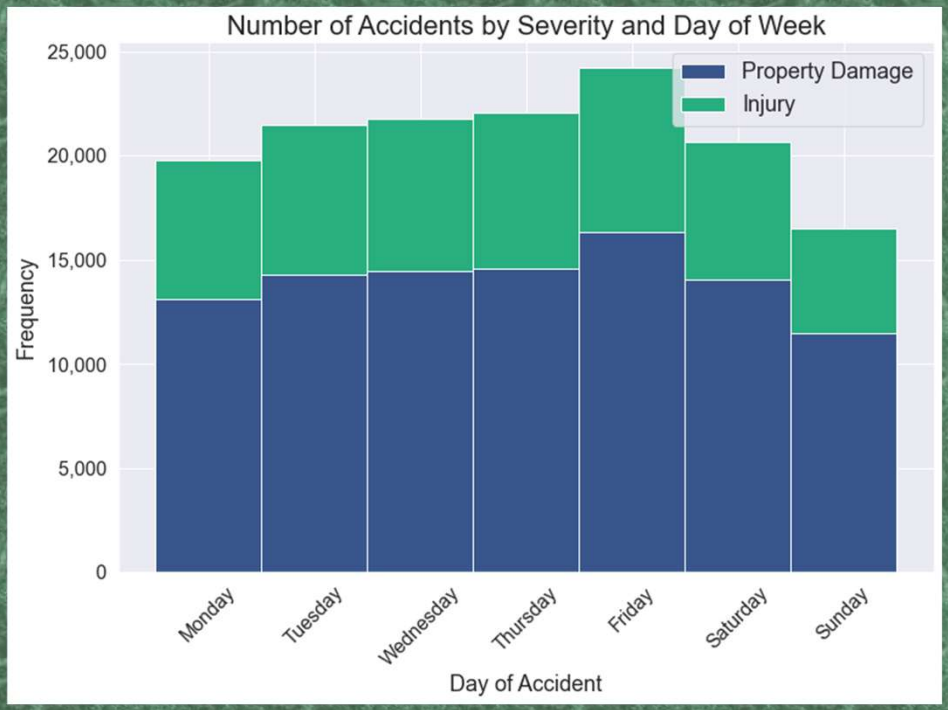
(Click for external html interactive)



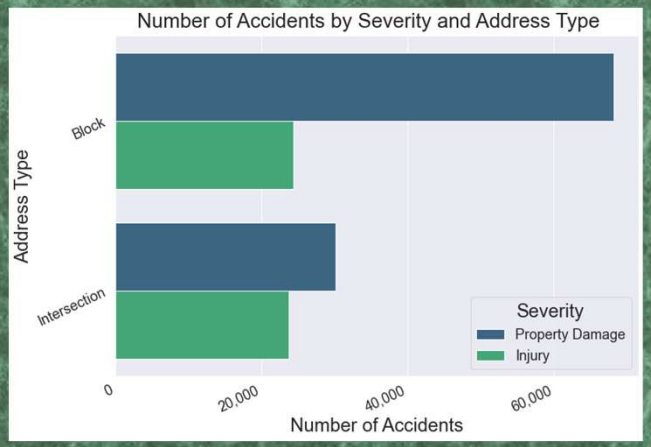
Accidents by Date and Time

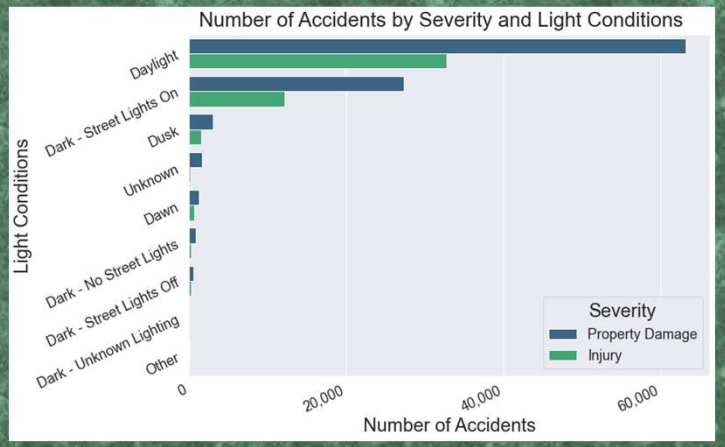


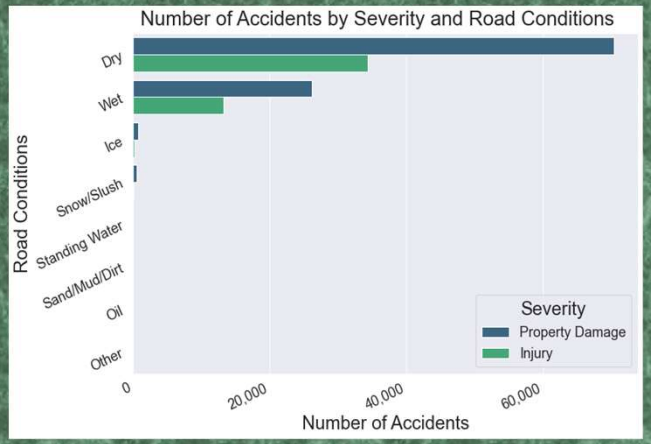


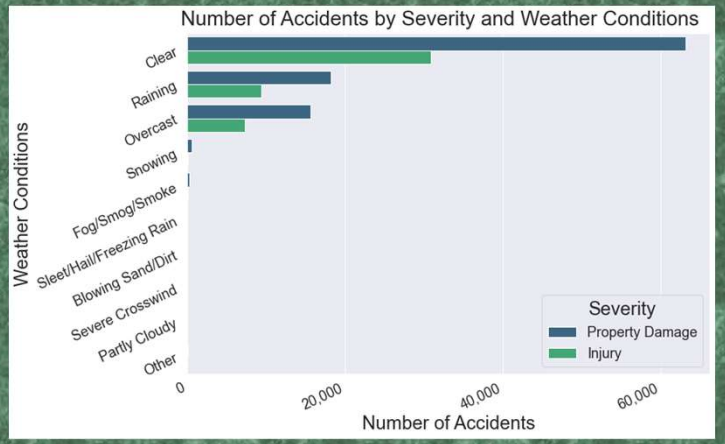


Accidents by Address and Light, Road, and Weather Conditions









Classification Models and Results

Training Set Results

Algorithm	Jaccard	F1-Score	Log Loss
Decision Tree/Random Forest (Latitude/Longitude)	0.504928	0.671033	NA
Decision Tree/Random Forest (Neighborhood)	0.504928	0.671033	NA
K-Nearest Neighbor (Latitude/Longitude)	0.506079	0.672048	NA
K-Nearest Neighbor (Neighborhood)	0.505183	0.671258	NA
Logistic Regression (Latitude/Longitude)	0.504928	0.671033	0.615137
Logistic Regression (Neighborhood)	0.504928	0.671033	0.615592
Support Vector Machine (Latitude/Longitude)	0.504928	0.671033	NA
Support Vector Machine (Neighborhood)	0.505161	0.671238	NA

Test Set Results

Algorithm	Jaccard	F1-Score	Log Loss
Decision Tree/Random Forest (Latitude/Longitude)	0.506580	0.672490	NA
Decision Tree/Random Forest (Neighborhood)	0.506580	0.672490	NA
K-Nearest Neighbor (Latitude/Longitude)	0.506683	0.672581	NA
K-Nearest Neighbor (Neighborhood)	0.506786	0.672672	NA
Logistic Regression (Latitude/Longitude)	0.506580	0.672490	0.613581
Logistic Regression (Neighborhood)	0.506580	0.672490	0.614233
Support Vector Machine (Latitude/Longitude)	0.506580	0.672490	NA
Support Vector Machine (Neighborhood)	0.506709	0.672603	NA

Conclusion

- Accuracy is around 67% for all models because they are predicting 95%-99% of the observations as Property Damage, and 67% of all accidents are actually Property Damage!
- In general, Seattle emergency responders should plan for a 2:1 ratio of Property Damage to Injury for accidents.
- More variables could improve accuracy, however, acquiring those variables at the scene of the accident would allow responders to immediately deduce the severity of the accident.
- Variable interaction with more computing power could possibly improve the model.