Predicting Car Accident Severity

Applied Data Science Capstone

Introduction

Background

- New drivers are frequently involved in accidents due to their inexperience. Predicting the severity and type of accidents based on weather and road conditions.

Problem Description

- The results of this project will enable driving instructors to adjust their training lessons to better prepare new drivers to deal with these challenges from the start. By training new drivers how to effectively deal with these situations at the beginning of their driving career, the frequency of these types of accidents will decrease preventing injuries and property damage.

Data Description

- The Seattle area collisions data will be used to solve this problem including the following key variables:
  - Weather conditions (ie. overcast, raining, clear)
  - Road conditions (ie. wet, dry)
  - Light conditions (ie. daylight, dark)
  - Address type (ie. intersection, block, alley)
  - Severity of accident (ie. injury, property damage only)

Methodology

For the purposes of this project, 5 key variables were analyzed and compared to determine their impact on the severity of the accident. These variables include:

- Weather
- Road Conditions
- Light Conditions
- Address Type
- Speeding

The data was cleaned to ensure all formatting aligned and rows with blank values were removed to ensure the data would not be skewed.

Heat maps were used to visualize the relationships between two variables and the severity of the accident. This was used to identify key combinations of characteristics that impact accident severity.

Methodology

The correlation between the variables weather, road conditions, light conditions, speeding, and address type with severity was examined to determine significance of the relationship. The following characteristics were used:

- The Pearson Correlation was used to measure the linear dependence between two variables X and Y. The resulting coefficients were between -1 and 1, where 1 = total positive linear correlation, 0 = no linear correlation (the two variables most likely do not affect each other), and -1 = total negative linear correlation.
- The P-value was used to determine the probability that the correlation between these two variables is statistically significant. A significance level of 0.05 was used indicating 95% confidence in the correlation variables.
  - When p-value <0.001, there is a strong evidence for a significant correlation.
  - When p-value <0.05, there is moderate evidence that the correlation is significant
  - When p-value <0.1, there is weak evidence that the correlation is significant
  - When p-value >0.1, there is no evidence that the correlation is significant

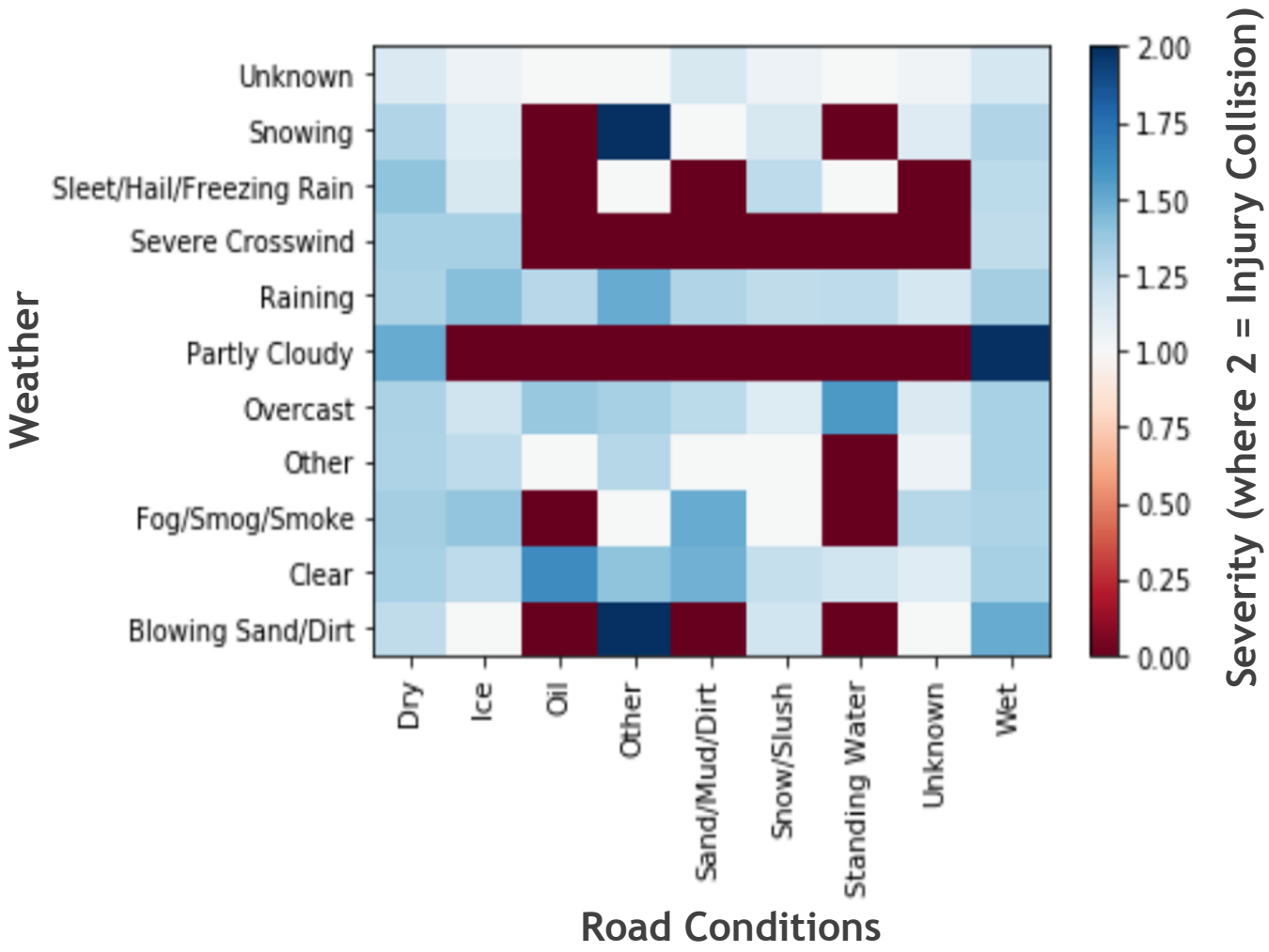
Methodology

To develop an accurate model to be used to predict the severity of accidents, the following steps were implemented:

- Categorical features were converted to numerical values
- Data was normalized to give data zero mean and unit variance
- Created a training and test set
- Used the training set to build an accurate model
- Tested the accuracy of the model using the following algorithms:
  - K Nearest Neighbor (KNN)
  - Decision Tree
  - Support Vector Machine
  - Logistics Regression

Results

Weather and Road Conditions

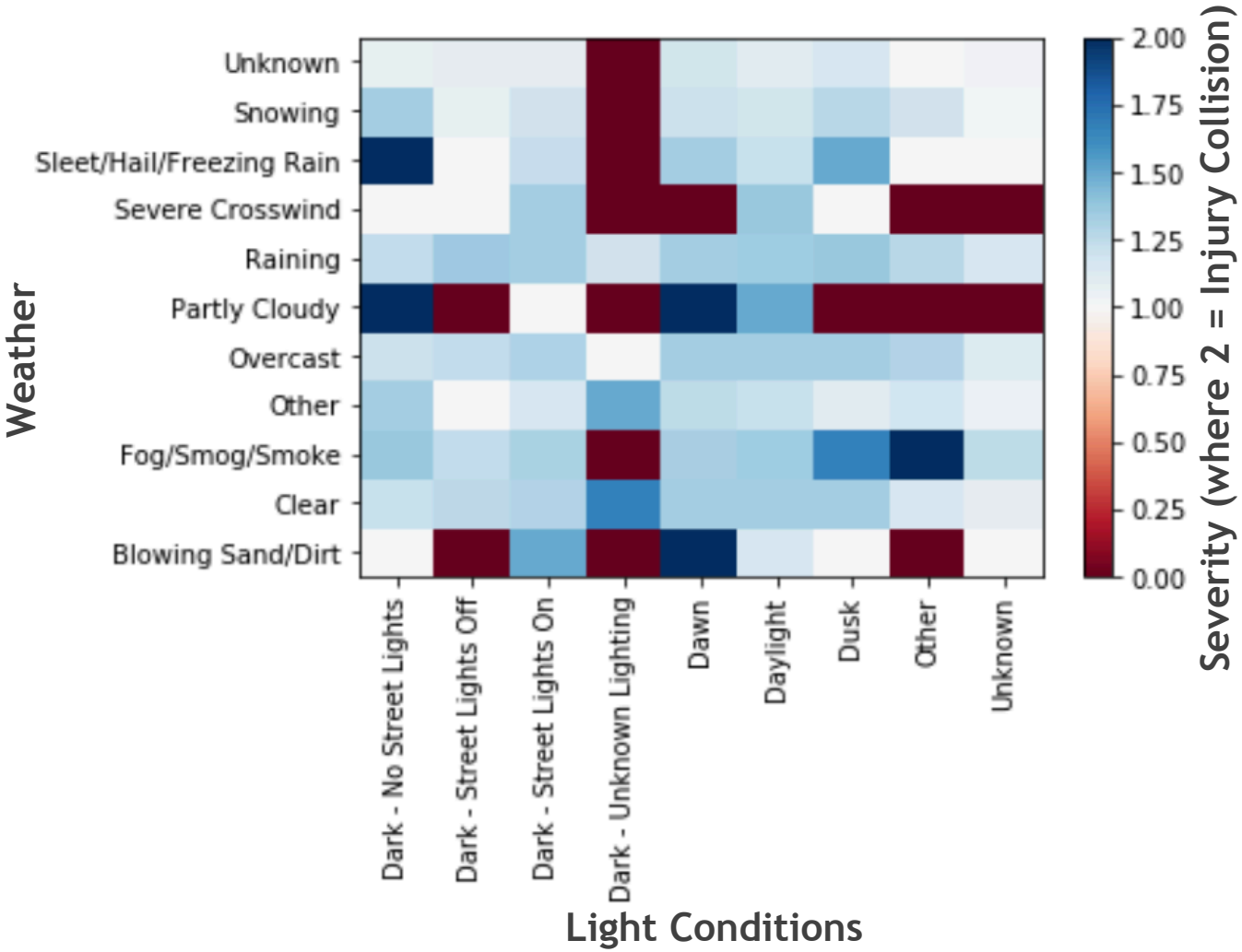


Combinations leading to severe accidents

- Partly Cloudy + Wet Roads
- Snowing + Other Roads
- Blowing Sand/ Wet + Other Roads

Results

Weather and Light Conditions

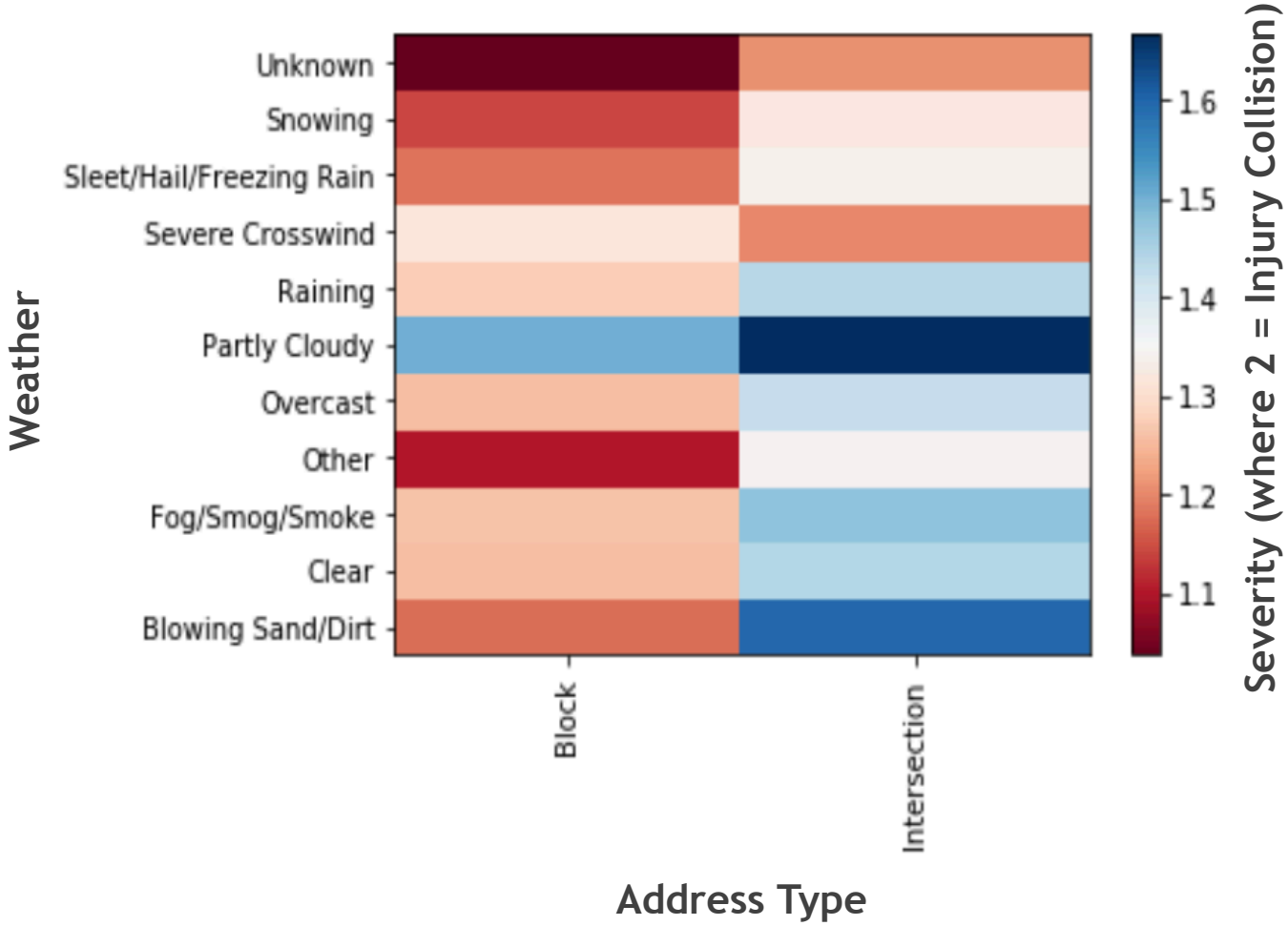


Combinations leading to severe accidents

- Sleet/Hail/ Freezing Rain + Dark – No Street Lights
- Partly Cloudy + Dark – No Street Lights
- Partly Cloudy + Dawn
- Blowing Sand/Dirt + Dawn
- Fog/Smog/Smoke + Other

Results

Weather and Address Type

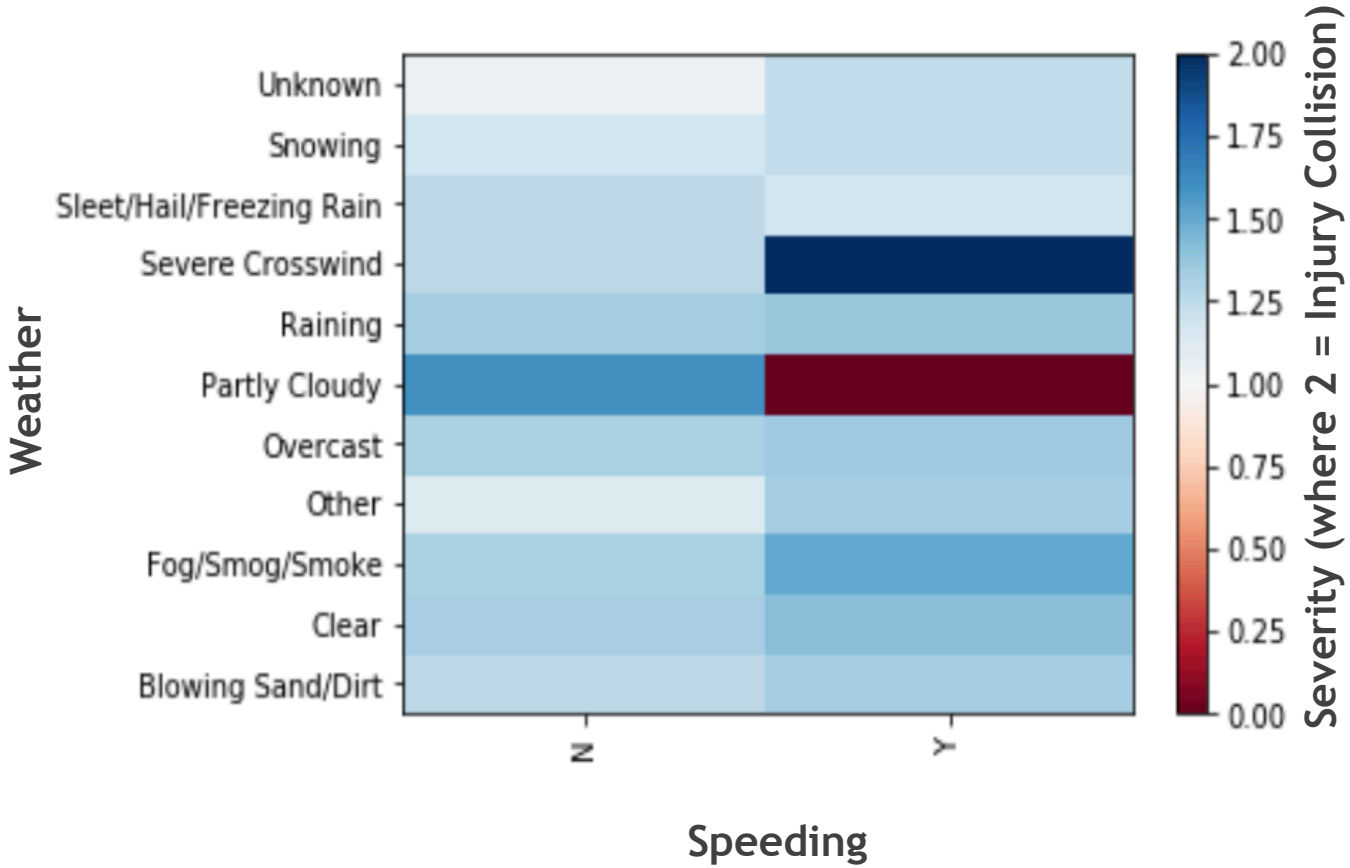


Combinations leading to more frequent severe accidents

- Partly Cloudy + Intersections
- Blowing Sand/Dirt + Intersections

Results

Weather and Speeding

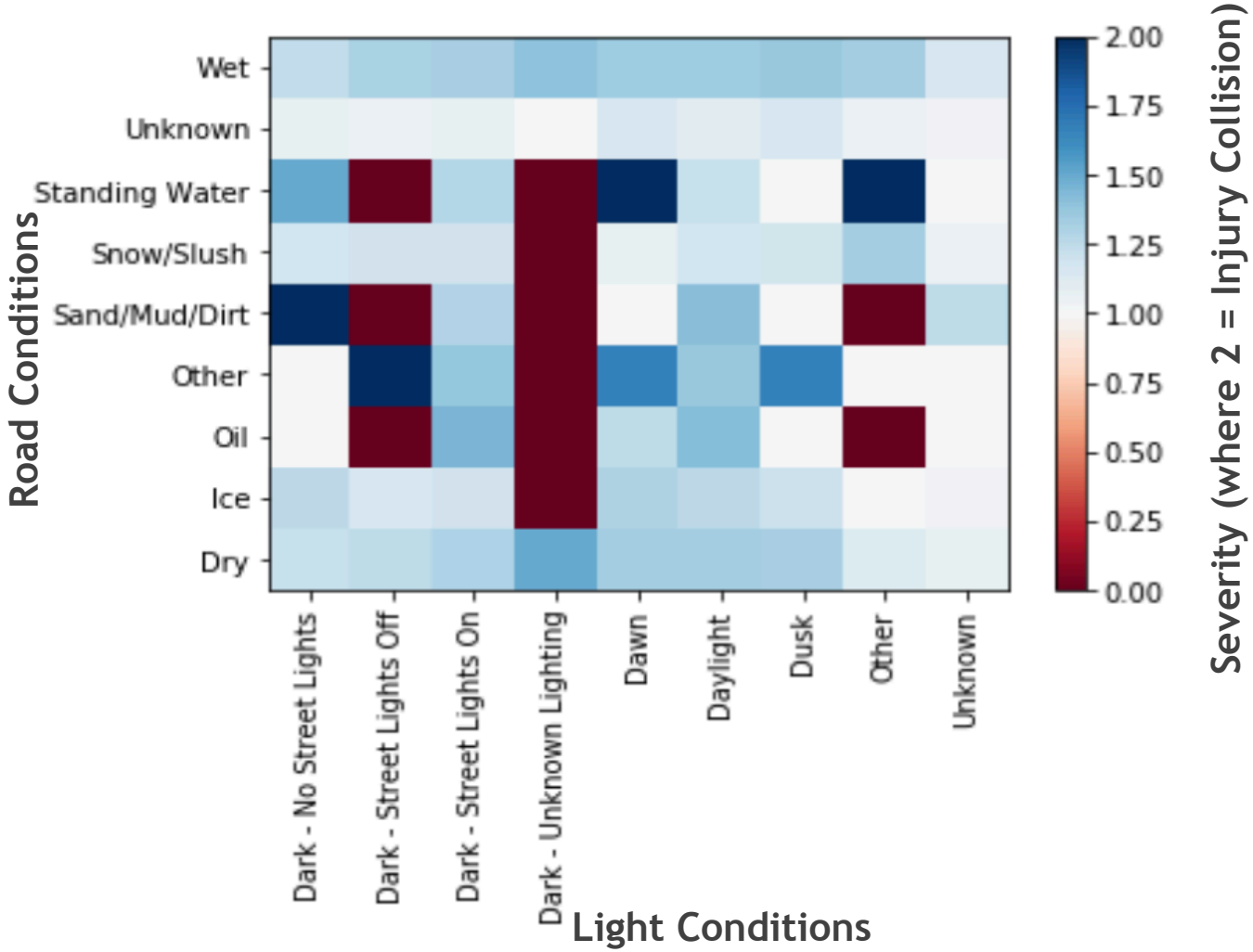


Combinations leading to severe accidents

- Severe Crosswind + Speeding

Results

Road Conditions and Light Conditions

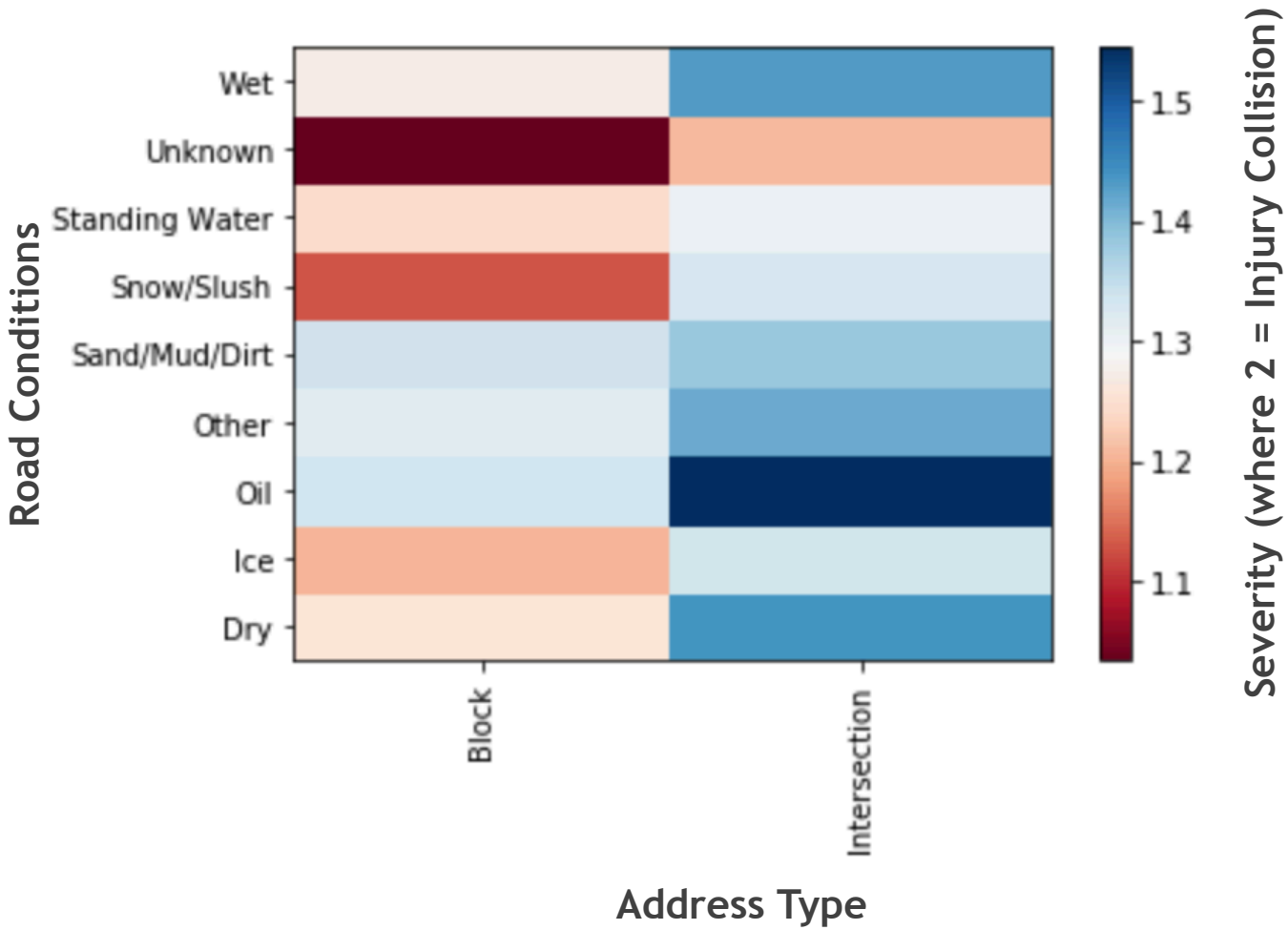


Combinations leading to severe accidents

- Sand/Mud/Dirt Roads + Dark – No Street Lights
- Other Roads + Dark – Street Lights Off
- Standing Water Roads + Dawn
- Standing Water Roads + Other

Results

Road Conditions and Address Type

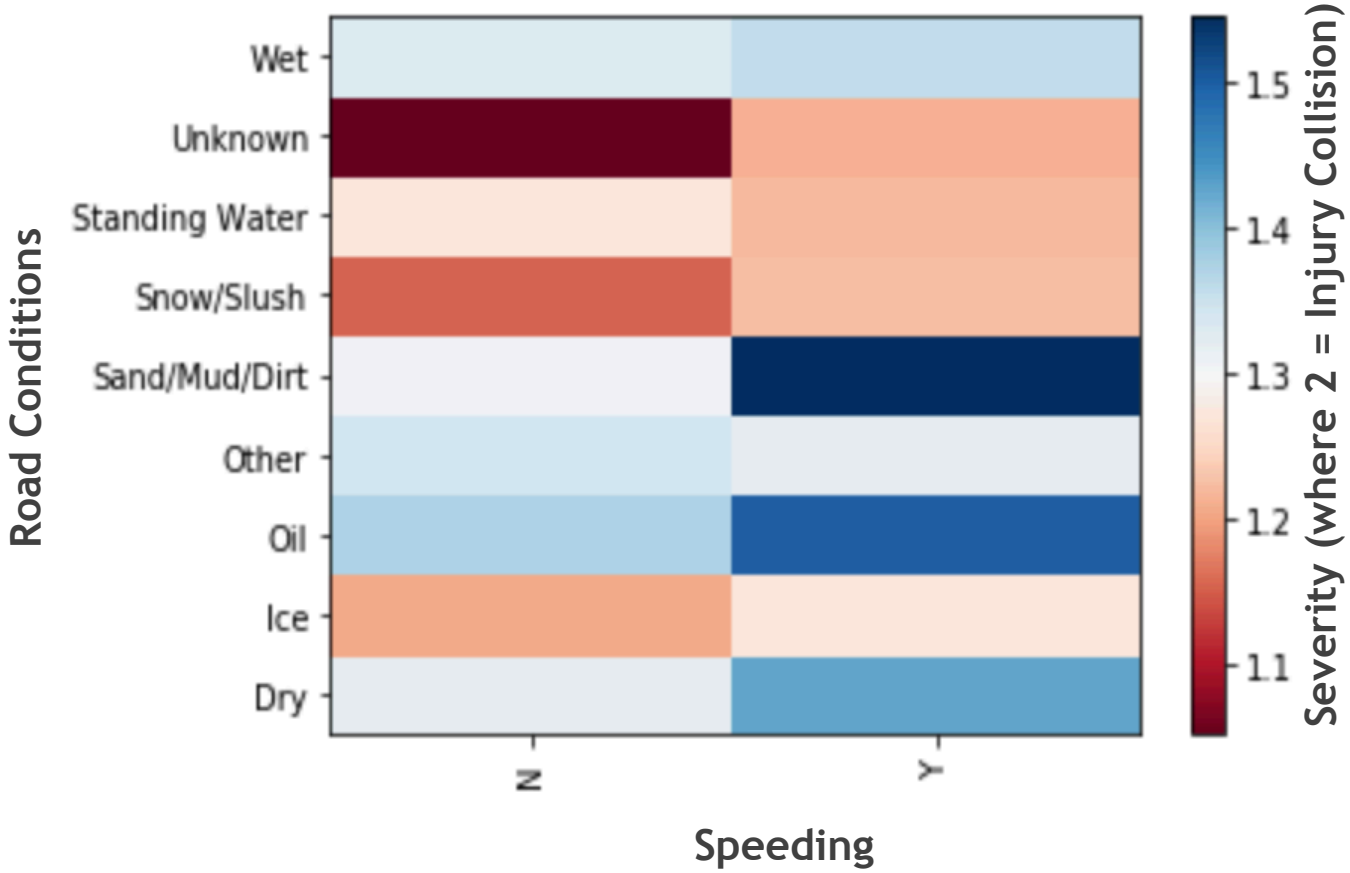


Combinations leading to more severe accidents

- Oil Roads + Intersection

Results

Road Conditions and Speeding

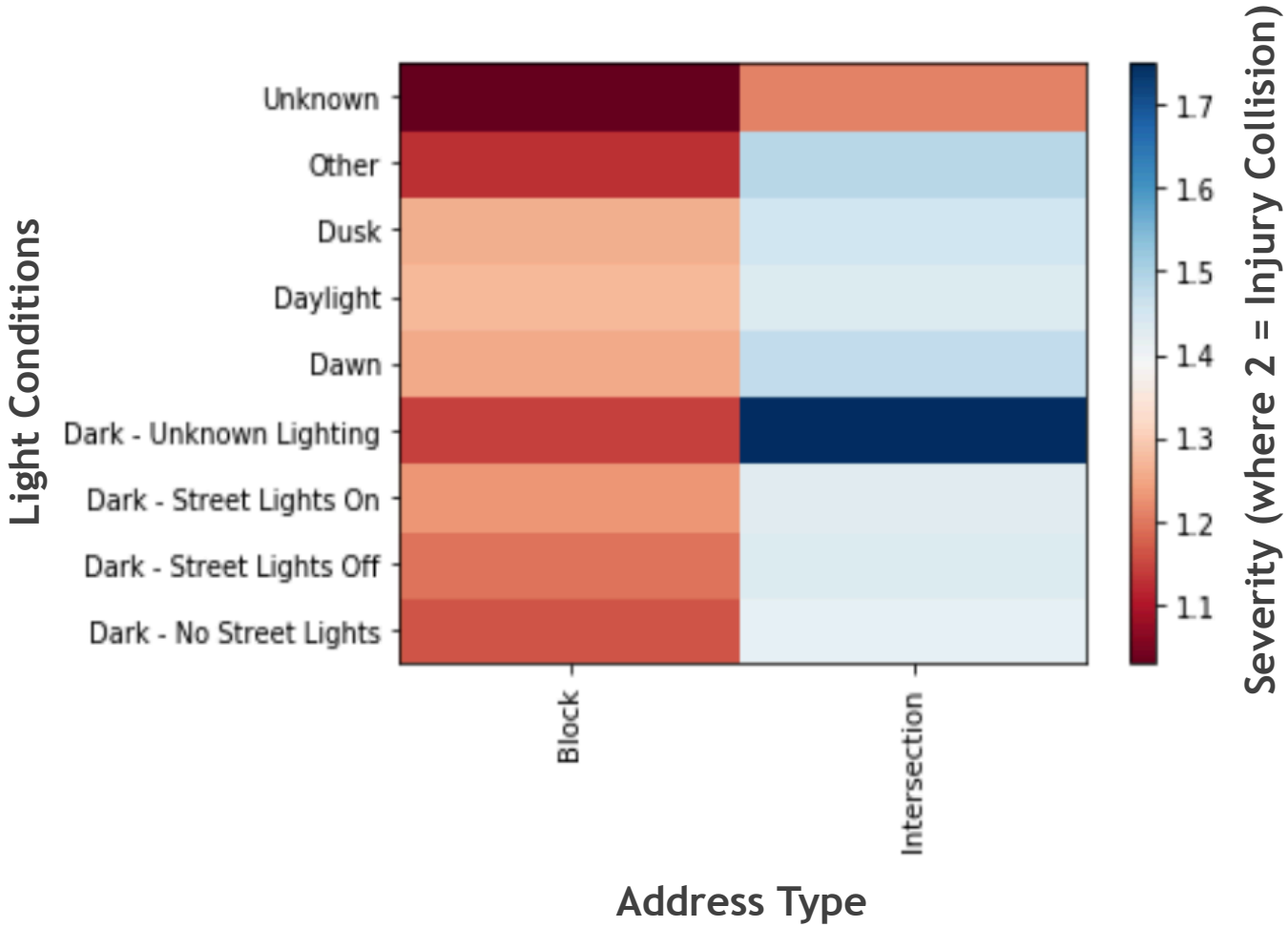


Combinations leading to more severe accidents

- Sand/Mud/Dirt Roads + Speeding

Results

Light Conditions and Address Type

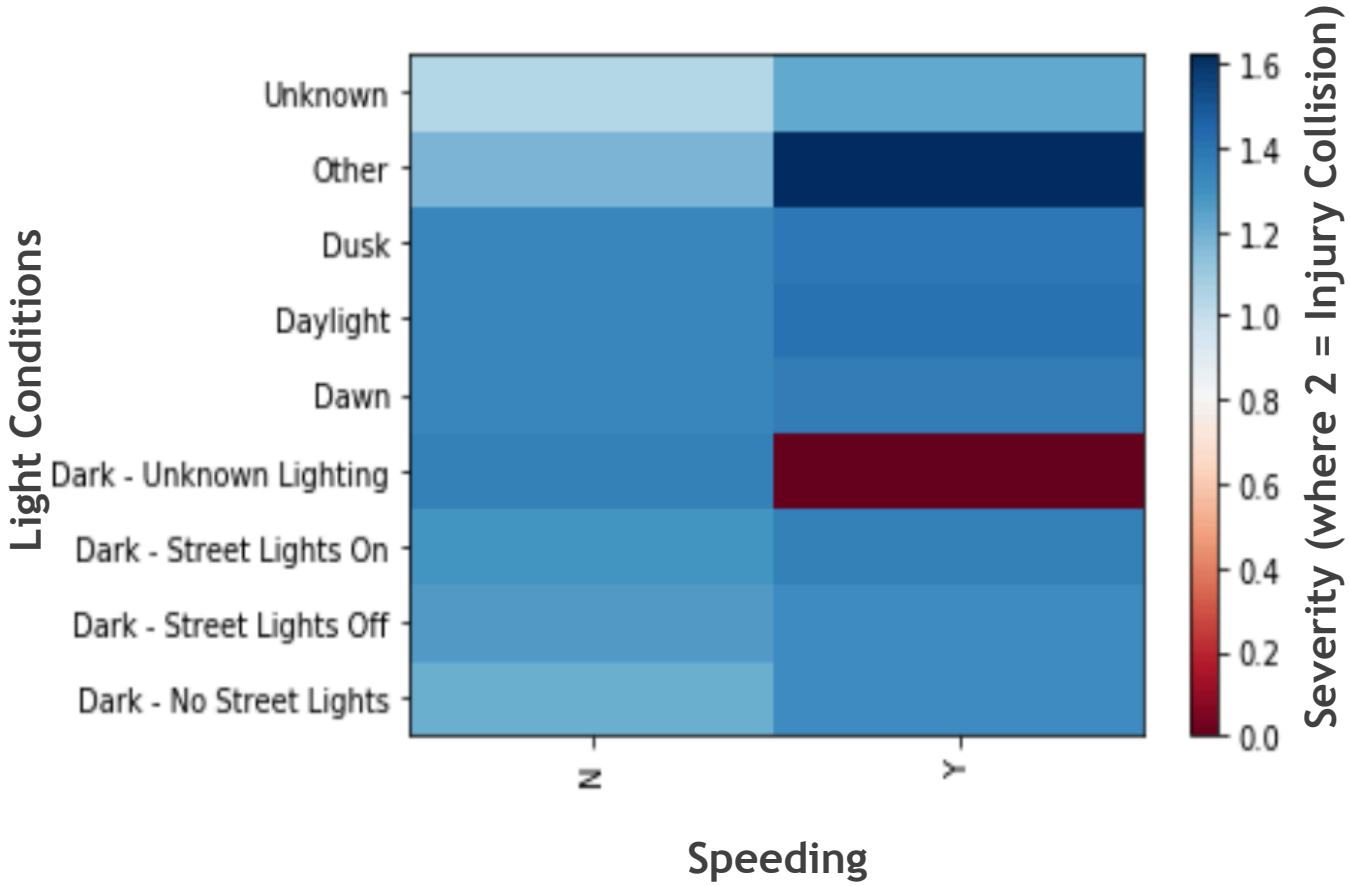


Combinations leading to more severe accidents

- Dark – Unknown Lighting + Intersections

Results

Light Conditions and Speeding

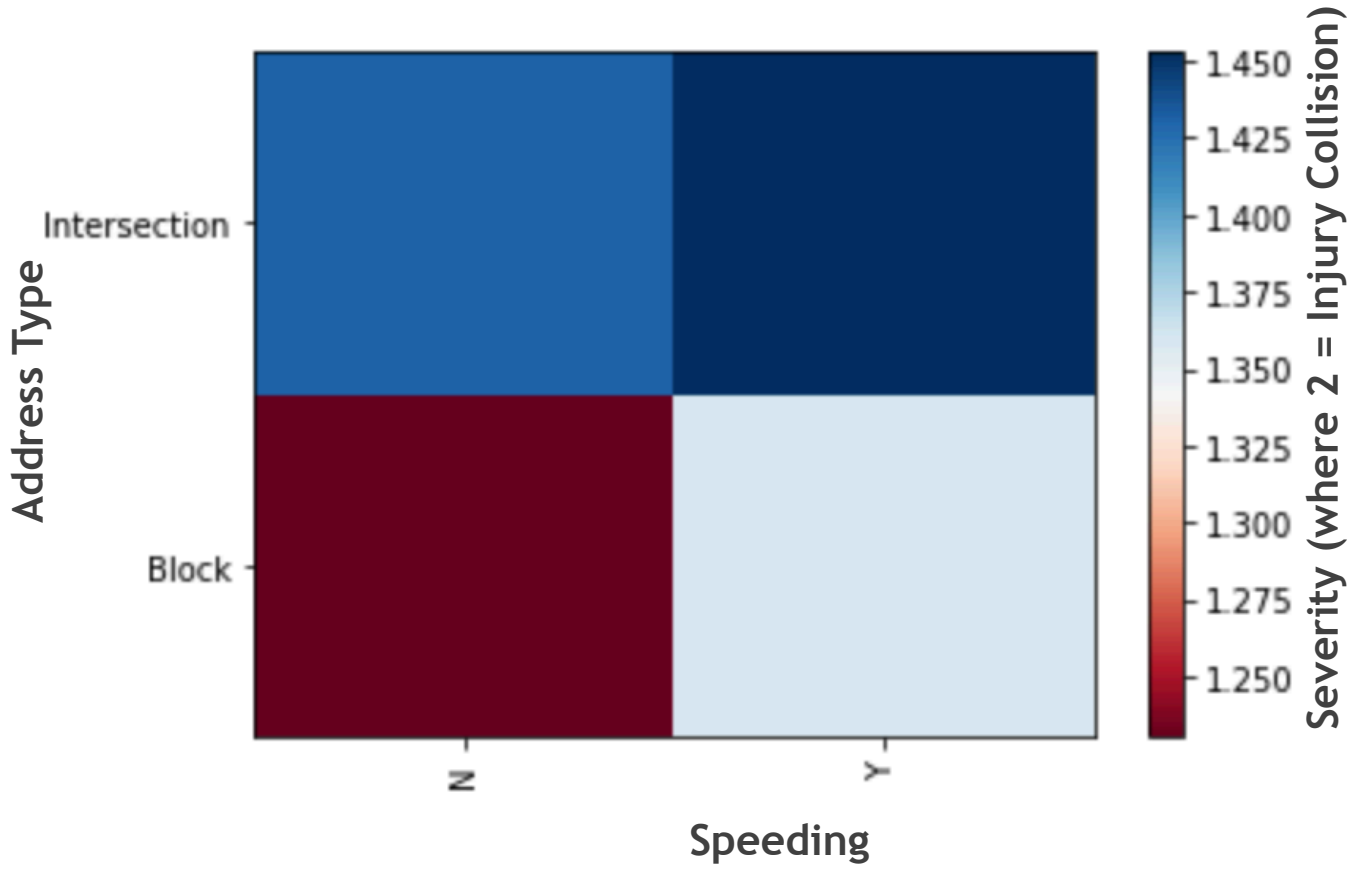


Combinations leading to more severe accidents

- Other Lights + Speeding

Results

Address Type and Speeding



Combinations leading to more severe accidents

- Intersections + Speeding

Results

Variable Significance Relative to Accident Severity

Variable	Pearson Coefficient	P-value	Impact
Weather	-0.10599	0	The correlation has strong evidence of statistically significance, and the linear relationship is negative.
Road Conditions	-0.04601	6.743e-87	The correlation has strong evidence of statistically significance, and the linear relationship is negative.
Light Conditions	-0.05663	1.127e-130	The correlation has strong evidence of statistically significance, and the linear relationship is negative.
Speeding	-0.03736	6.760e-58	The correlation has strong evidence of statistically significance, and the linear relationship is negative.
Address Type	0.20064	0	The correlation has strong evidence of statistically significance, and the linear relationship is positive.

Results

Evaluating Model Accuracy

The accuracy of the model using each test can be seen below:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.686	0.600	NA
Decision Tree	0.695	0.575	NA
SVM	0.697	0.575	NA
Logistic Regression	0.696	0.585	0.585

The model accuracy was between 58-70%. It is not very accurate and requires future modification.

Discussion

Key Observations

- From the data visualizations (heat maps), it was identified that certain combinations of road conditions + light conditions, weather + speeding, weather + light conditions, and weather + road conditions, lead to severe accidents every time (as identified on the corresponding slide).
  - 6 of the key observations within the above categories included “Other or unknown” conditions. This indicates important data that should be accurately documented.
- The variable statistic significance relative to accident severity indicated that there is strong evidence for a significant relationship between all variables and accident severity, as well as a linear correlation between the variables (both positive and negative). It is important to note that these were the only tested variables and thus may not be the only statistically significant variables.
- The accuracy of the model from each analysis was between 58-70% accurate. This indicates that the model has room for improvement.

Conclusion and Recommendations

- Data visualization and data modelling techniques were used to analyze the impact of 5 key variables on the severity of accidents.
- The severity of an accident typically involves a combination of key variables. No single variable can be used to predict the outcome of an accident.
- Accuracy of the models have room for improvement.
- Recommended next steps include:
  - Analyzing the data based on driver's age or experience,
  - Reviewing the accident severity based on location within the city, and
  - Develop models to include multiple variables.