Prediction models for estimating severity of collision in Seattle

Data Resource

Data set: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Data description file: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Introduction

In our daily life, there often happens collision in our cities. The bad weather, the bad road conditions, and so on factors could lead to traffic accidents.

From the brief report from the privies, the weather report, the conditions of the position, and so on, the information may help instructors make predictions.

Business Audience

Main audience should be the traffic managers in big cities like Seattle.

Data

Collision information since 2004 to present

194673 rows, 38 columns

Main attributes: location, severity, collision type, number of pedestrians, bicycles, and vehicles involved, injuries and fatalities, time, weather, road condition, light condition, and speed.

Methodology

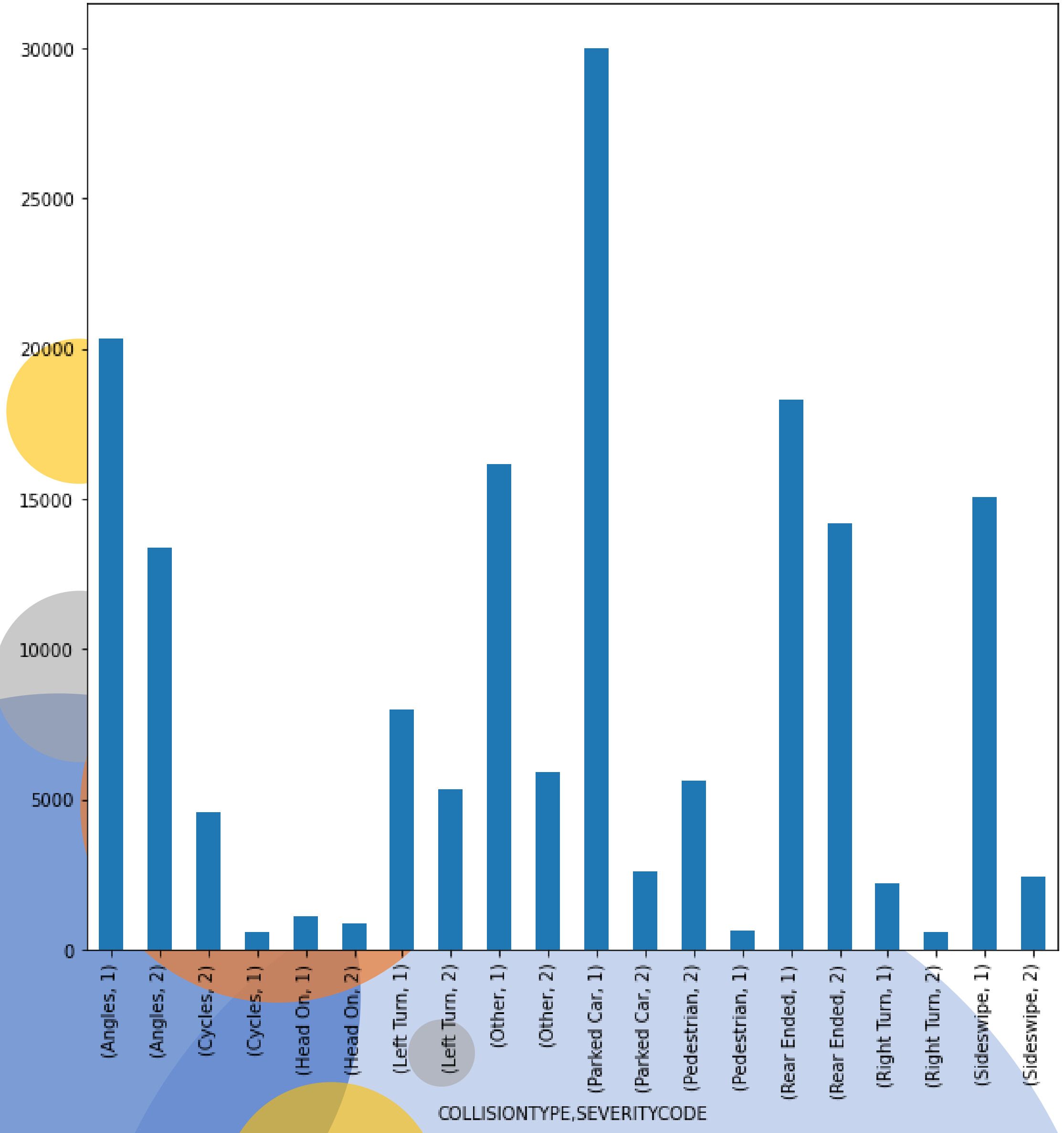
1. Select several variables

2. Wash data set

3. Visualize data

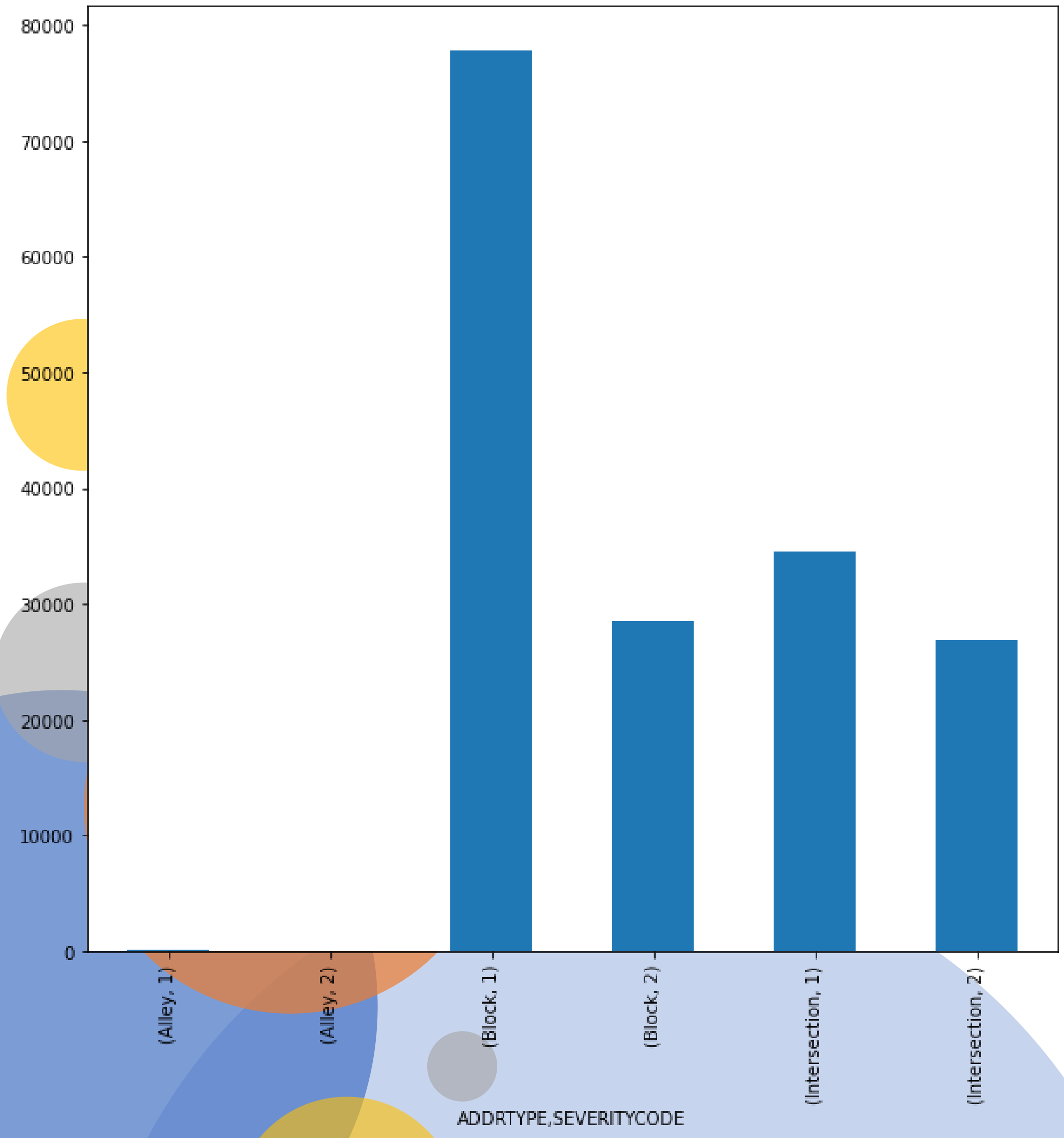
4. Build predictive models

Data Visualization



Bar plot with variables collision type and severity of collision shows that the condition ‘collision type = cycles’ is more likely to cause more severe collisions.

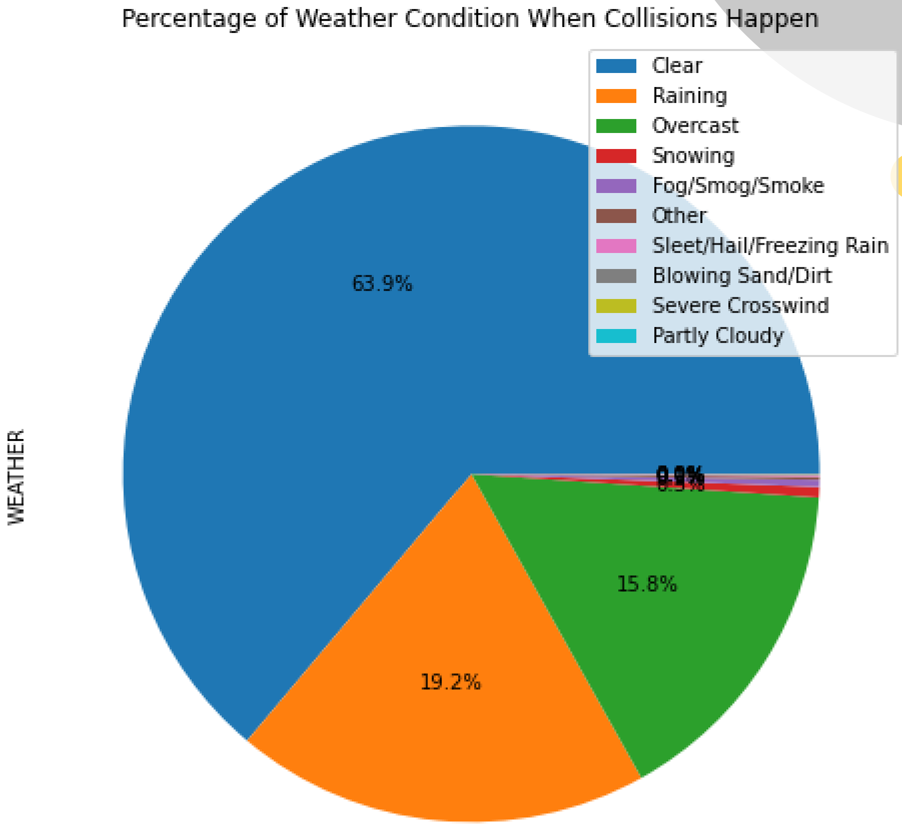
Data Visualization



Bar plot with variables address type and severity of collision shows that the condition ‘address type = intersection’ is more likely to cause more severe collisions.

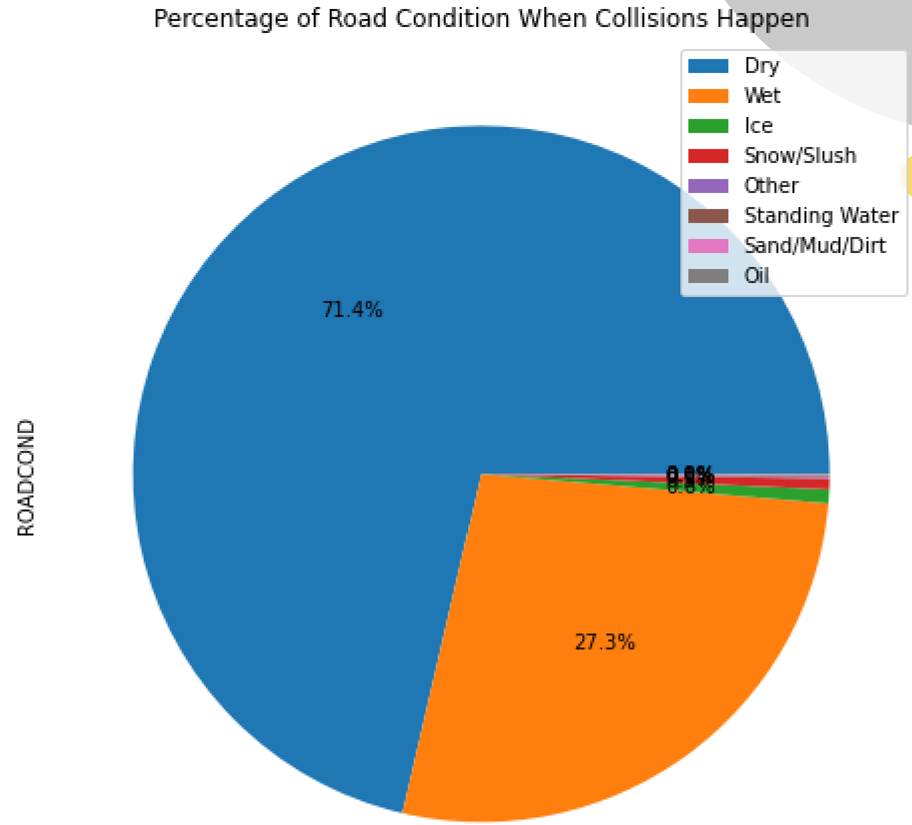
Data Visualization

Pie plot with variables weather and severity of collision shows that the condition ‘weather = raining and overcast’ is more likely to cause collisions.



Data Visualization

Pie plot with variables road condition and severity of collision shows that the condition ‘road condition = wet and ice’ is more likely to cause collisions.



Models

- Decision tree model
- Support vector machine
- Logistic regression model

Results

	Accuracy	F1-score	Log_loss
Decision tree	0.7225	0.6722	NA
SVM	0.7175	0.6499	NA
Logistic regression	0.7240	NA	0.5734

Discussion

From the results, the accuracies are very close. Therefore, the decision tree model and the logistic regression model are preferred to be applied. Between them, the decision tree model will be preferred when a simple and easy to explain model is needed. The logistic model is very useful to predict the probability of severity of collisions.

Further Directions

In this project, many other variables such as positions, the date and time, and detailed collision descriptions are not used in building the models. Further, since the alerting model should be faster, the number of inputs should be easier to gather and record. Simpler models are needed in real cases. In this way, it can shorten the time for traffic managers to collect necessary information and make a decision in a short while.