PREDICTING CAR ACCIDENT SEVERITY

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE FINAL CAPSTONE PROJECT

AUTHOR: MURAD POPATTIA

Project OUTLINE:

- Introduction (Business Problem)
- Data
- Methodology
- Results
- Conclusion

Introduction

- **Goal:** to find potential location for a car accident given the weather and road conditions
- **Main target:** increasing public road safety in Seattle, Washington, United States.
- Unfortunate incidents often occur on the road. The unpredictability of these accidents make them so dangerous. It is usually the impatient nature of the human which leads to an accident. But wouldn't it be a blessing if the number of accidents that occur on a daily basis be reduced? **That is the problem which will be addressed in this project.**

Data

Data Set Summary

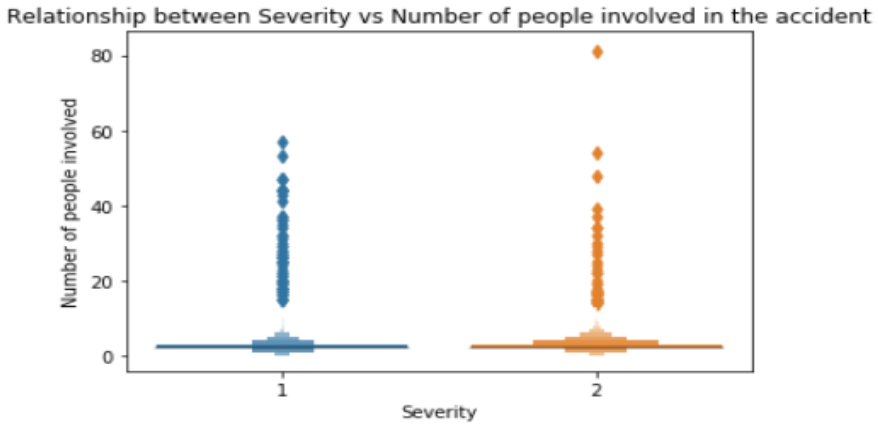
Data Set Basics	
Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Supplemental Information	
Update Frequency	Weekly
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Contact Information	
Contact Organization	SDOT Traffic Management Division, Traffic Records Group
Contact Person	SDOT GIS Analyst
Contact Email	DOT_IT_GIS@seattle.gov

Methodology

- Data Preprocessing:
  - Removing null values
  - Removing unwanted columns except for (Person Count, Vehicle Count, Attention ID, under influence verification, Weather, Road conditions, Lighting conditions)
  - One-hot encoding columns
  - Standardizing format type of values across columns

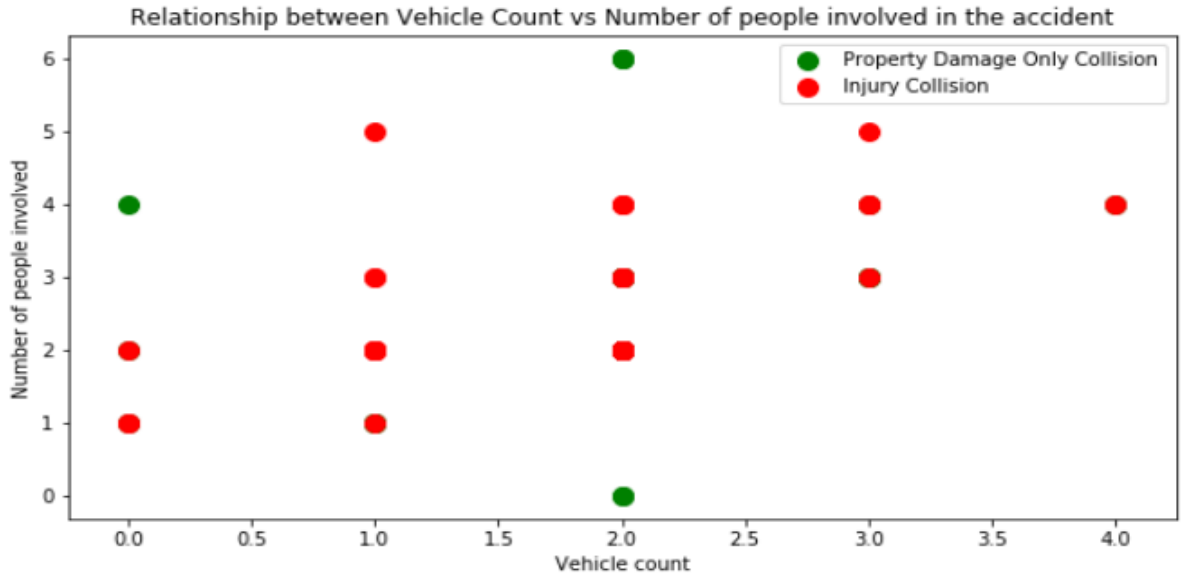
Methodology

- Exploratory Data Analysis:
  - Through this we could conclude that if there are higher people traveling, the possibility of severe collision is less since more people mean more traffic which leads to less possible speeding. The box plot below shows this analysis:



Methodology

- Seeing this plot, we could conclude that conclude that when the number of vehicles involved in the accident were high, there was more injury collision than the collisions which only damaged the property.
- The scatter plot below shows this analysis:



Methodology

- This data will then be used to create a train/test data set.
- We use the train and test split in the data to prevent overfitting and ensure that there is some out-of-sample data for testing the model.

Results

- A linear regression model was successfully created to predict the severity of the road accidents by evaluating various factors that occurred in the training set data. The best parameters were found for the model and the model accuracy was measured. The model accuracy was found to be 70.5%. Below are some performance metrics used for model evaluation.

	Jaccard Similarity Score	F1 Score	Log loss	Test Accuracy
0	70.58	0.608417	0.577406	70.57917

Conclusion

- Further additions to the project will involve improving the efficiency of the model by providing a more diverse dataset which would contain multiple degrees of severities.
- Moreover, we can also use a better algorithm with more hyperparameter tuning.
- The dataset as we see is imbalanced hence, we can use down-sampling as well in order to avoid biasness in results.
- However, as a benchmark, this has achieved respectable accuracy.