Predicting Car Accident Severity

IBM Applied Data Science Capstone

Colin Olson - 1/22/21

Predicting Car Accident Severity

Problem Description

- Per the CDC, ~1.35 million people are killed on roadways around the world every year
- According to the CDC, crash injuries are the 8th leading cause of death globally
- In 2014, the National Highway Traffic Safety Administration reported the “economic and societal harm from motor vehicle crashes” cost $871 billion in one year
- Being able to accurately predict the severity of car accidents would help to reduce both the bodily and economic harm that the accidents create

Predicting Car Accident Severity is Valuable to Several Parties

- The Government:
  - Improve conditions with enhanced safety measures in at-risk areas
  - Station emergency personnel closer to areas with higher risk of severe accidents
- Private Companies:
  - Insurance Companies - adjust premiums based on severity of possible accidents
  - Car Manufacturers - develop new technology to improve driver safety
- You, The Driver:
  - Avoid areas with high risk of severe accidents (or be more cognizant in those areas)

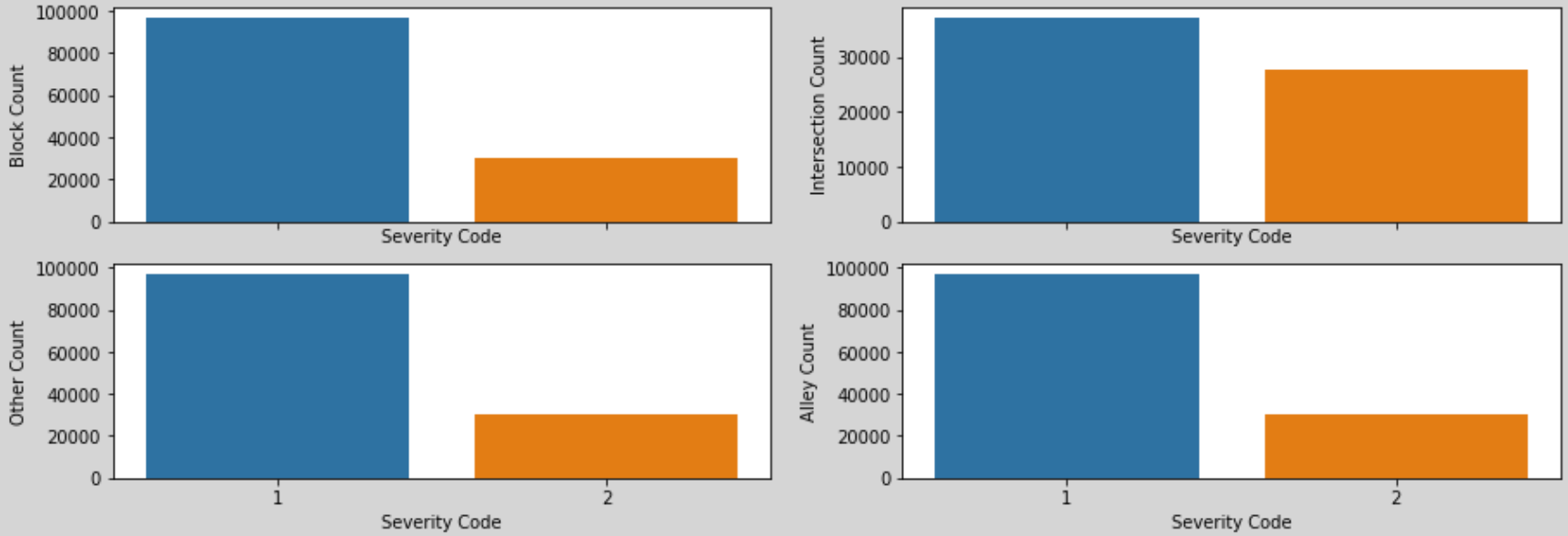
The Data

Acquisition and Cleaning

- Car Accident Data acquired from CSV link provided in Capstone Project description: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- In total, 194,673 rows and 38 features were imported in the raw data set
- Duplicate, similar, and unnecessary features were dropped
- The final, cleaned dataset contained 194,657 rows and 13 features

Exploratory Data Analysis

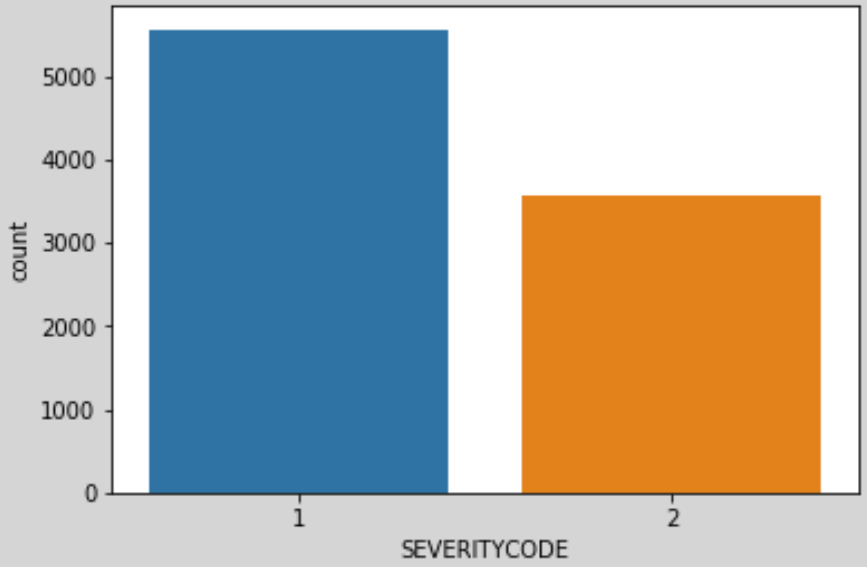
Different address types impacted the ratio of severe accidents



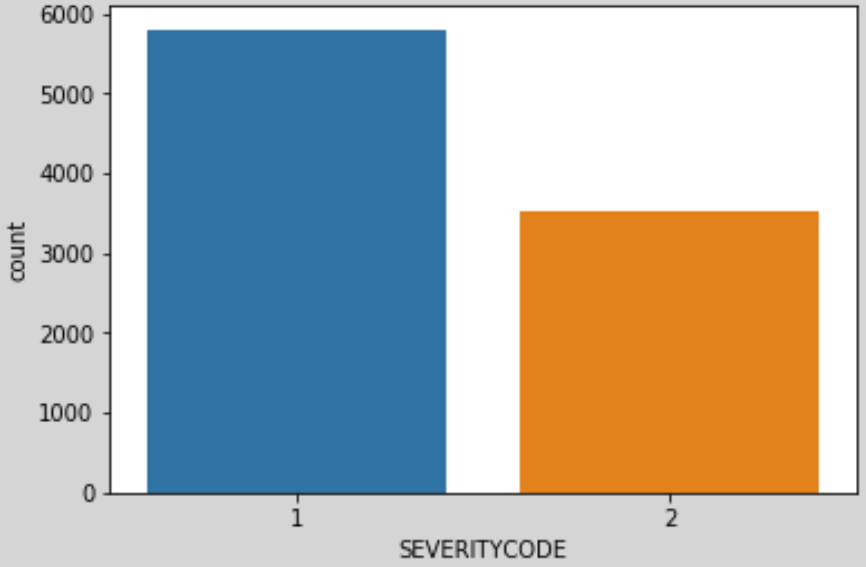
Exploratory Data Analysis

The ratio of severe accidents to non-severe accidents increased when:

- The driver was under the influence



- The driver was speeding



Exploratory Data Analysis

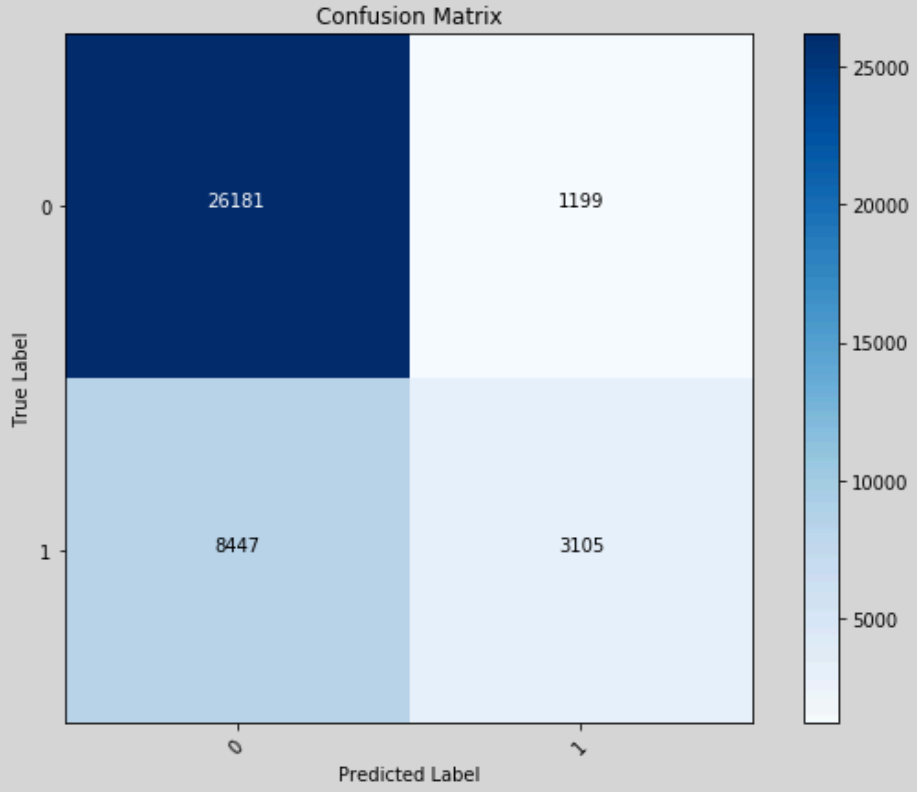
Other features examined include:

- Vehicle Count
- Person Count
- Pedestrian Count
- Cyclist Count
- Junction Type
- Severity where inattention occurred
- Weather conditions
- Road conditions
- Lighting conditions

Modeling

K-Nearest Neighbors

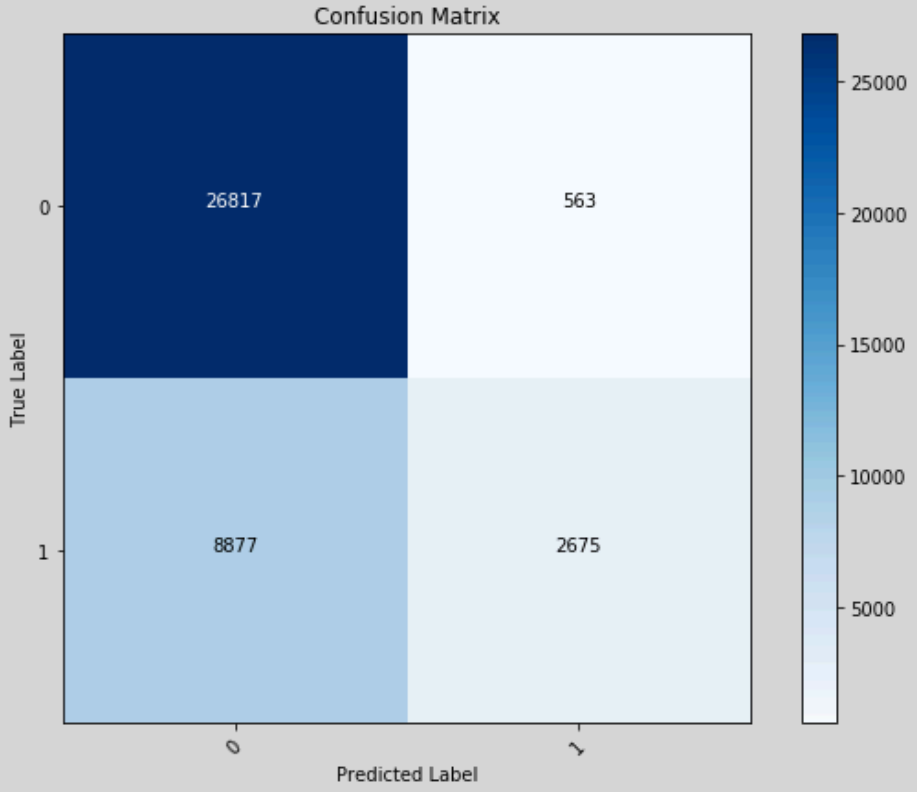
- Jaccard Score:
  - 0.752235
- F1 Score:
  - 0.844439
- Subset Accuracy Score:
  - 0.752235



Modeling

Decision Tree

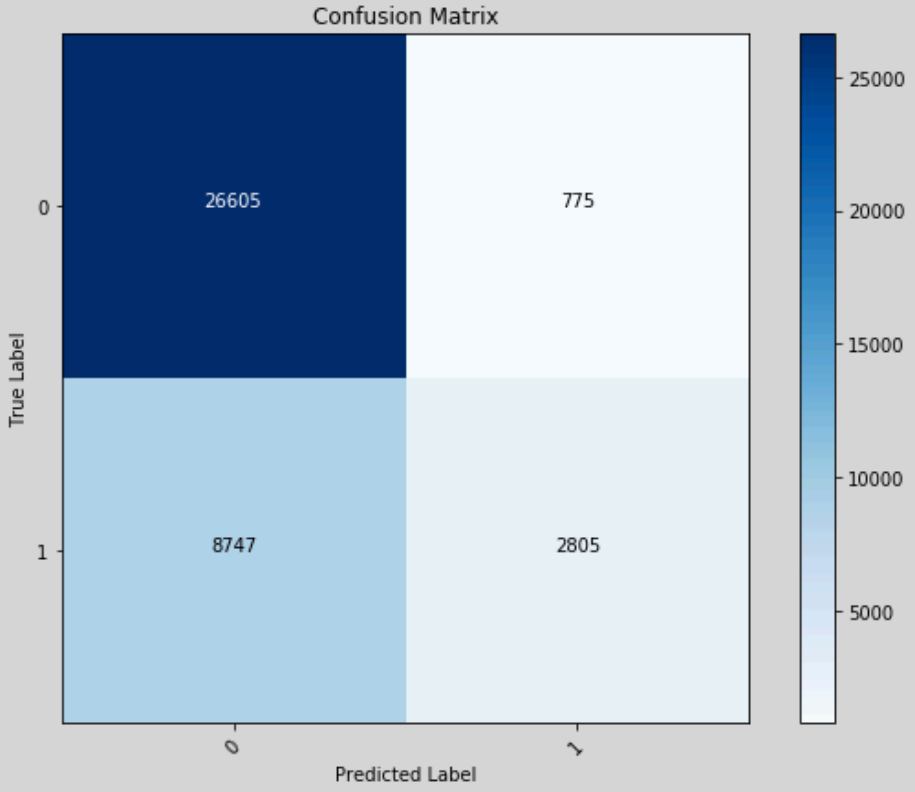
- Jaccard Score:
    - 0.757526
- F1 Score:
    - 0.850335
- Subset Accuracy Score:
    - 0.757526



Modeling

Linear Regression

- Jaccard Score:
    - 0.755420
- F1 Score:
    - 0.848211
- Subset Accuracy Score:
    - 0.755420
- Log Loss:
    - 0.514902



Conclusion

Future Directions

- Three useful models were built to predict the severity of a car accident given certain features
- The most accurate model, the Decision Tree model, should be implemented
- While these models were fairly accurate, the accuracy scores can definitely be improved
    - A more robust dataset, including additional features as well as filling in the data marked ‘Unknown’ and the missing data, would help improve these scores