Predicting Severity of Car Accidents

Applied Data Science IBM Capstone Project

Background

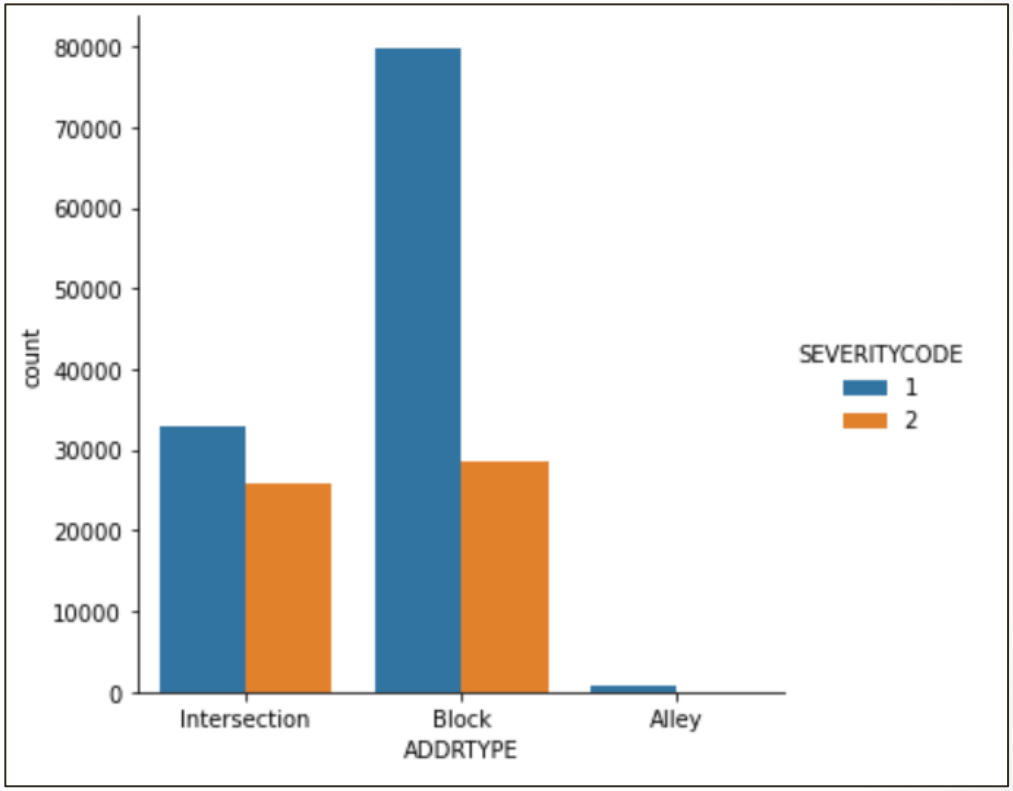
- Car accidents are a huge problem in the modern world.
- Many governments collect accident records to try to find what conditions cause car accidents.
- A predictive model could help address this issue and save lives.

Data acquisition and cleaning

- Car collisions records provided by the Seattle Police Department at Seattle.gov.
- Raw dataset contains 195,000 records and 38 features from the year 2004 to present, updated weekly.
- Duplicate, highly similar or highly correlated features were dropped.
- Cleaned data contains 167,000 records and 13 features.

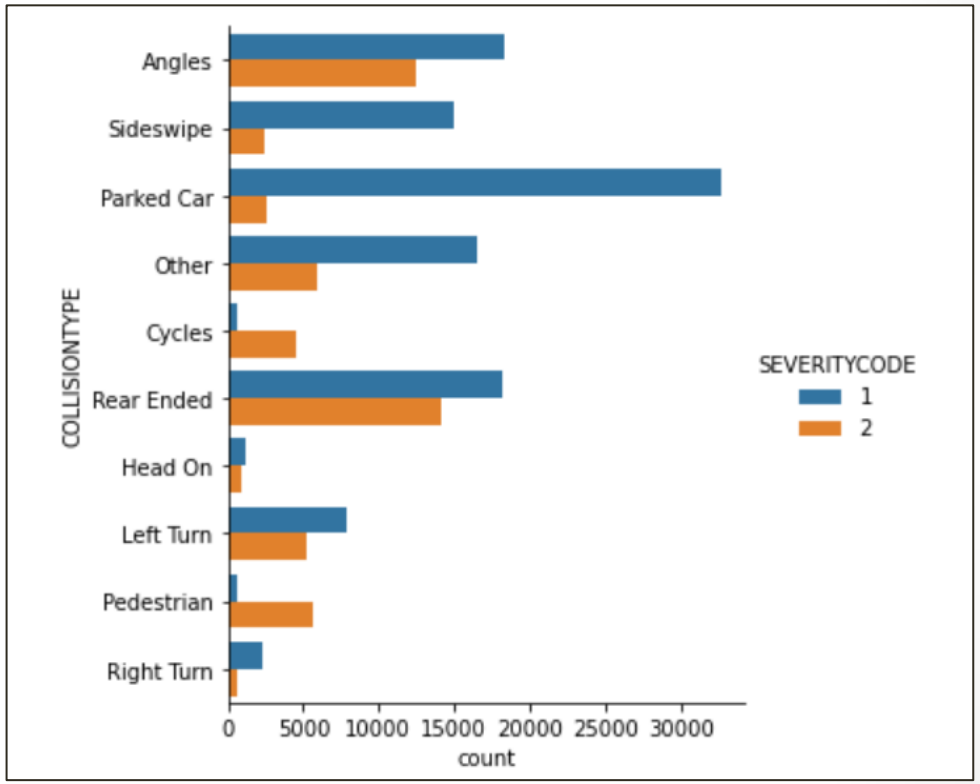
Exploratory Data Analysis

- Severity vs. Address type
- Chance of injury is much higher if the collision is in an intersection.
- Chance of property damage is much higher if the collision is in a block



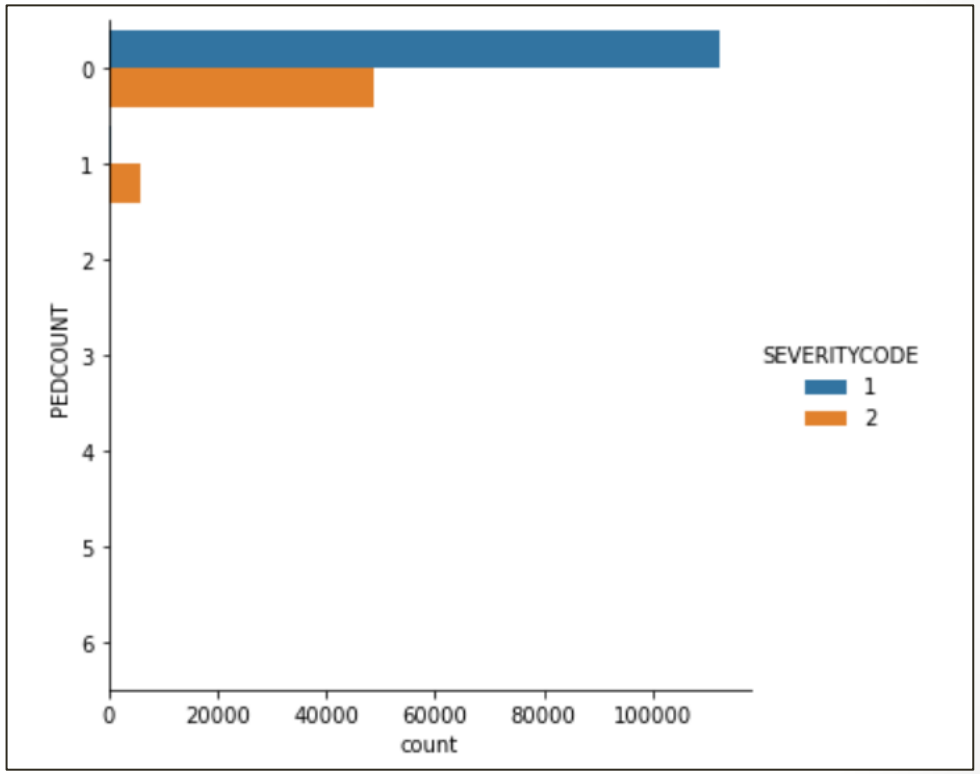
Exploratory Data Analysis

- Severity vs. Collision type
- Chance of injury is much higher if collision is from an angle or rear-ended.
- Chance of property damage is much higher if the collision is on a parked car.



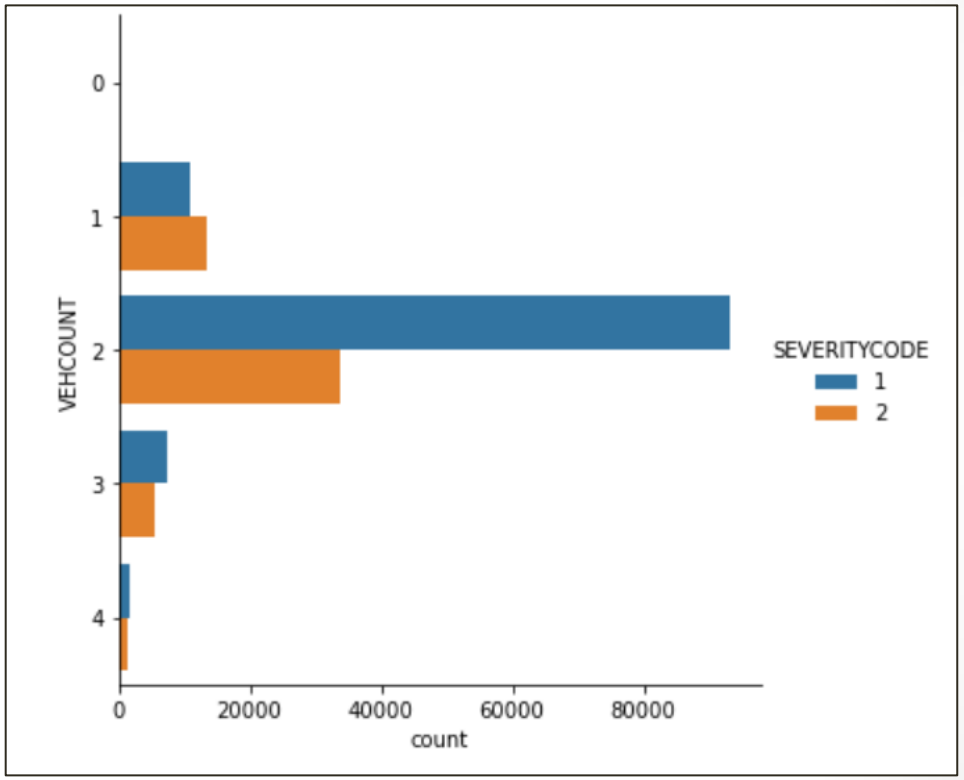
Exploratory Data Analysis

- Severity vs. Number of pedestrians involved
- Pedestrians are seldom involved in collisions.
- If they are, then the chance of injuries is high.



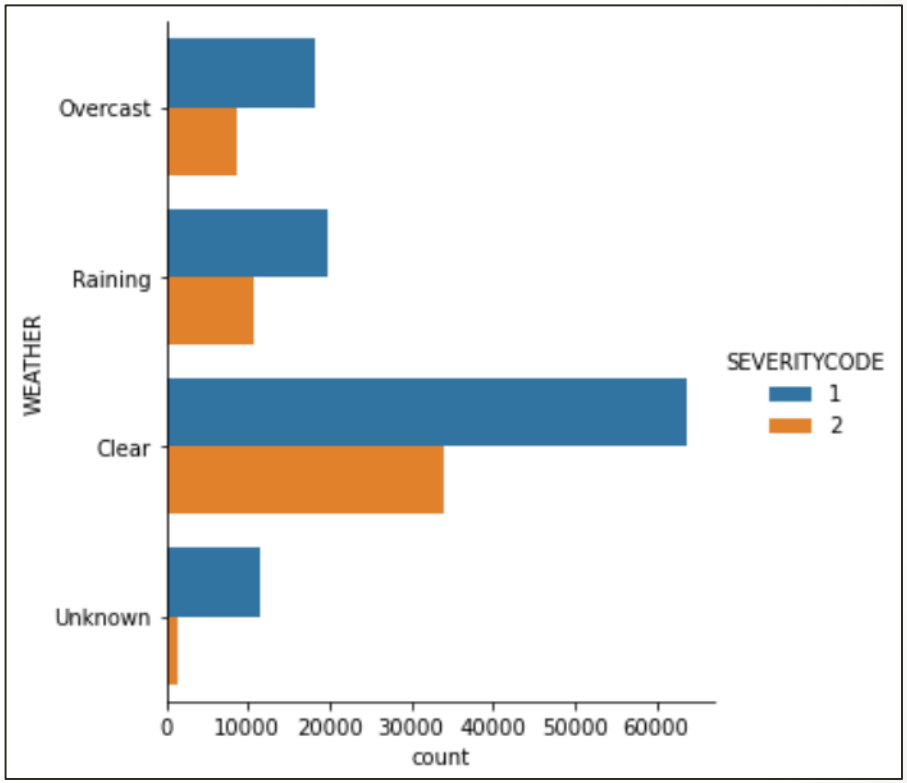
Exploratory Data Analysis

- Severity vs. Vehicle count
- Most collisions involve two vehicles.
- Chance of injury is lower in these cases relative to higher or lower vehicle count.



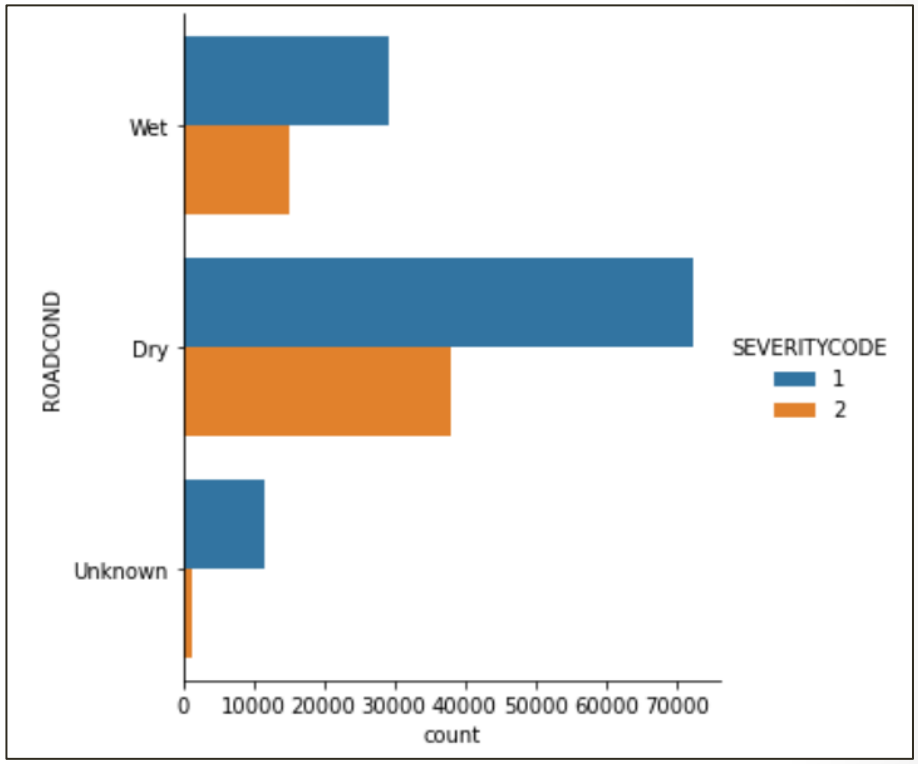
Exploratory Data Analysis

- Severity vs. Weather conditions
- Counter-intuitively most collisions occurred in clear weather.
- Varying ratio of injuries to property damage was noted.



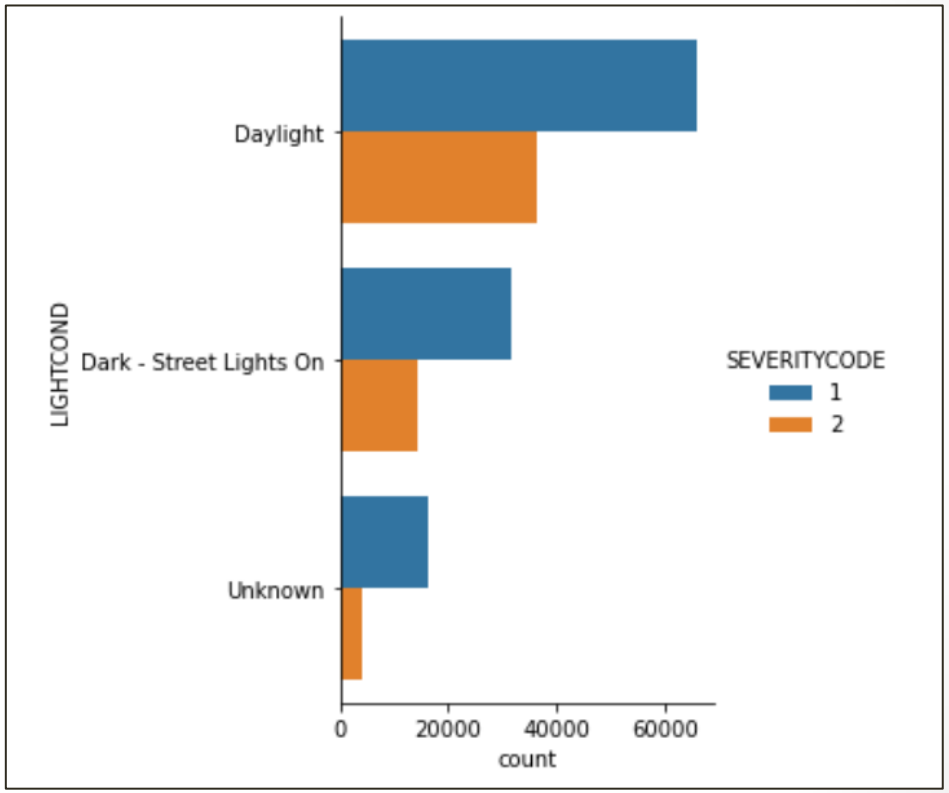
Exploratory Data Analysis

- Severity vs. Road conditions
- Counter-intuitively most collisions occurred in dry road conditions.
- Varying ratio of injuries to property damage was noted.



Exploratory Data Analysis

- Severity vs. Light conditions
- Counter-intuitively most collisions occurred in day light.
- Varying ratio of injuries to property damage was noted.

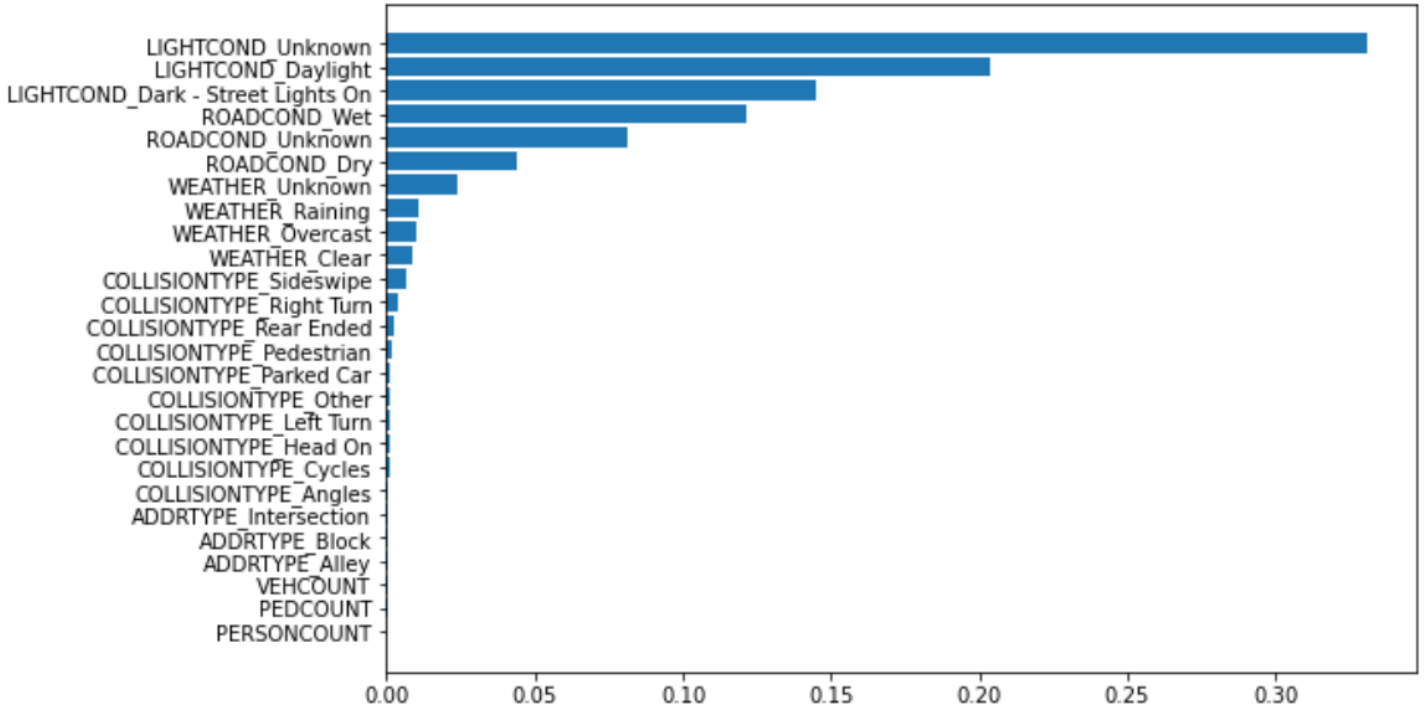


Classification model

- This is a classic supervised learning classification problem.
- four popular and well known classification models were examined:
  - Naïve Bayes
  - Logistic Regression
  - Random Forest
  - XGBoost
- performance metric: ROC_AUC
- XGBoost performed best with a score of 0.78

Feature Importance

- The most influential features are light conditions, road conditions and weather conditions.



Conclusion

- Results show that it is difficult to distinguish between Property Damage Only accidents and accidents with injuries.
- It is possible to ascertain what attributes are important and can indicate the possibility of an occurrence of a more fatal accident.
- Accuracy of the models has room for improvement.