Predicting the Severity of Accidents in Seattle, Washington Using Machine Learning

By: Dolu Okeowo
Date: 09/23/2020

Outline

Data

Result/Discussion

Introduction

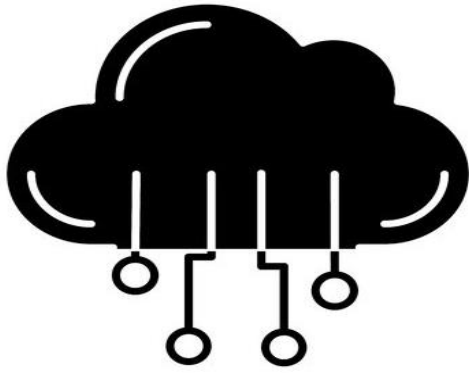
Methodology

Conclusion

Introduction

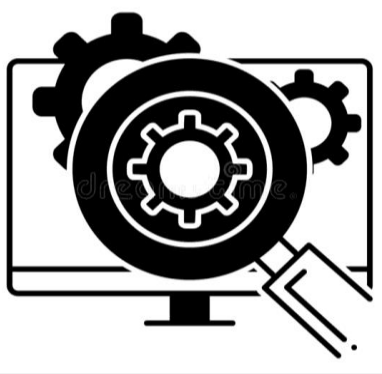
- Objective:
  - To predict the severity of accidents in Seattle into property damage (Class 1) and Injury (Class -2) using weather, road and light condition
- Predicting accidents saves lives and properties and increases overall life expectancy
- Help first responders deploy appropriate personnel to accident scenes

Data



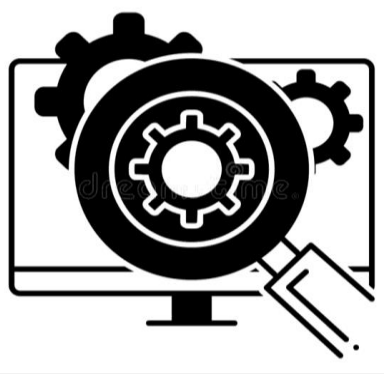
- Accident data for **Seattle, Washington**
- Consist of **37 features and 194,673 records**
- Some of the features include: Weather, RoadCond, LightCond, X, Y, Speeding etc.
- Target variable for classification:
  - **Severitycode**:
    - Property Damage – Class **1**
    - Injury – Class **2**

Methodology



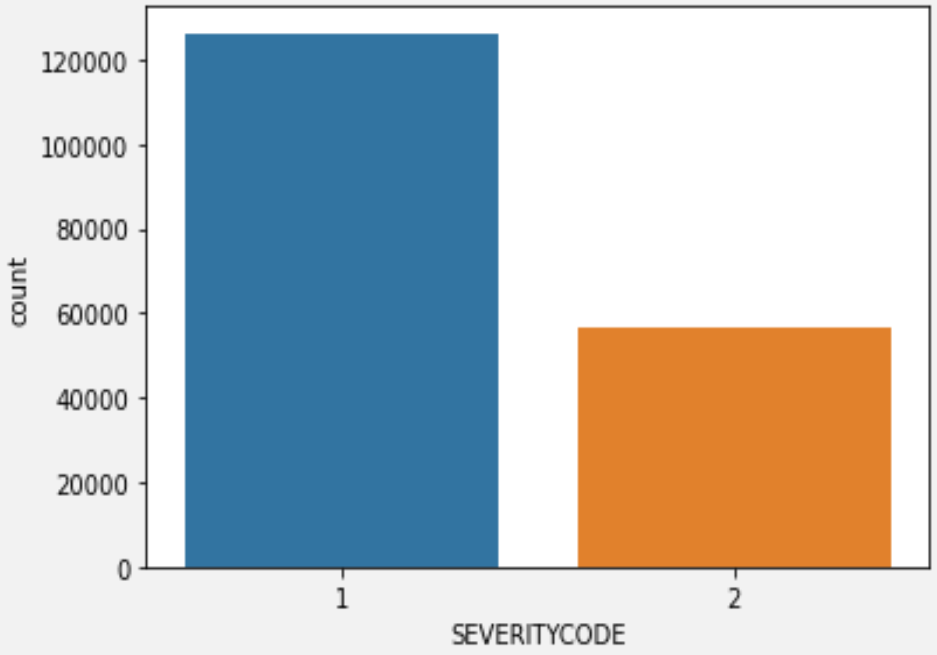
- Data Processing & Cleaning
  - Missing Data
    - Total records: 194,673
    - Missing records: 5,336 (5%)
    - Remaining: 189,337

Methodology

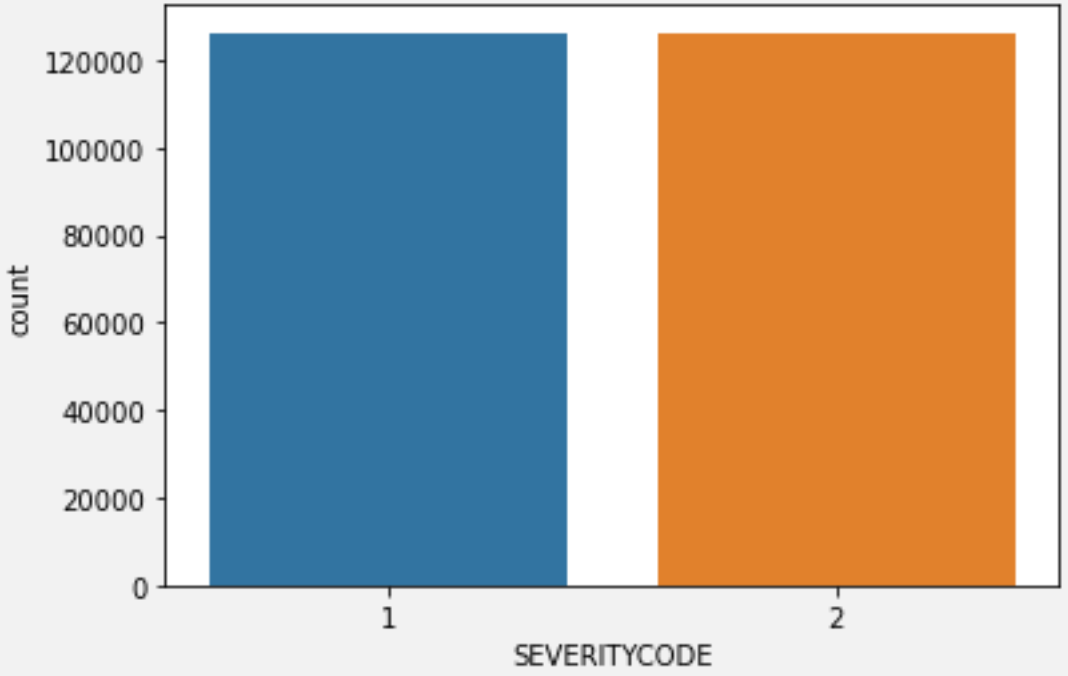


- Balancing data: Synthetic Minority Over-sampling **(**SMOTE) oversampling technique

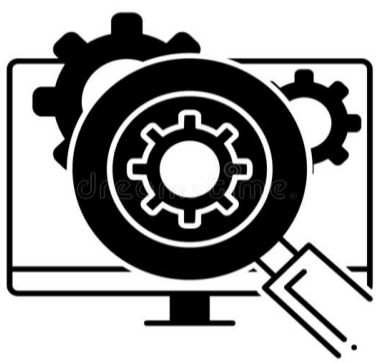
Unbalanced Data



Unbalanced Data

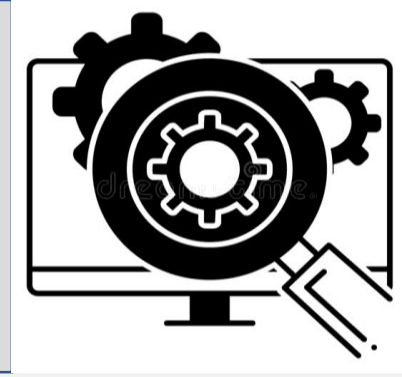


Methodology



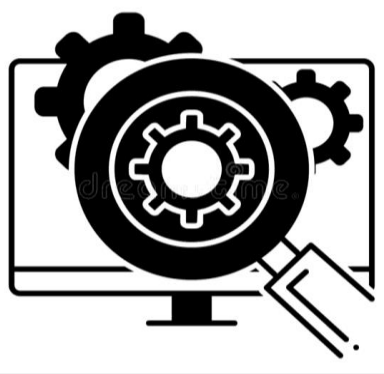
- **Model Selection**:
  - **Supervised Classification**
    - Decision tree
    - Logistic Regression
    - Naïve Bayes
    - Linear Discriminant Analysis
- **Train, Test Data:**
  - The data set was randomly split into **70% training and 30% test data**

Methodology

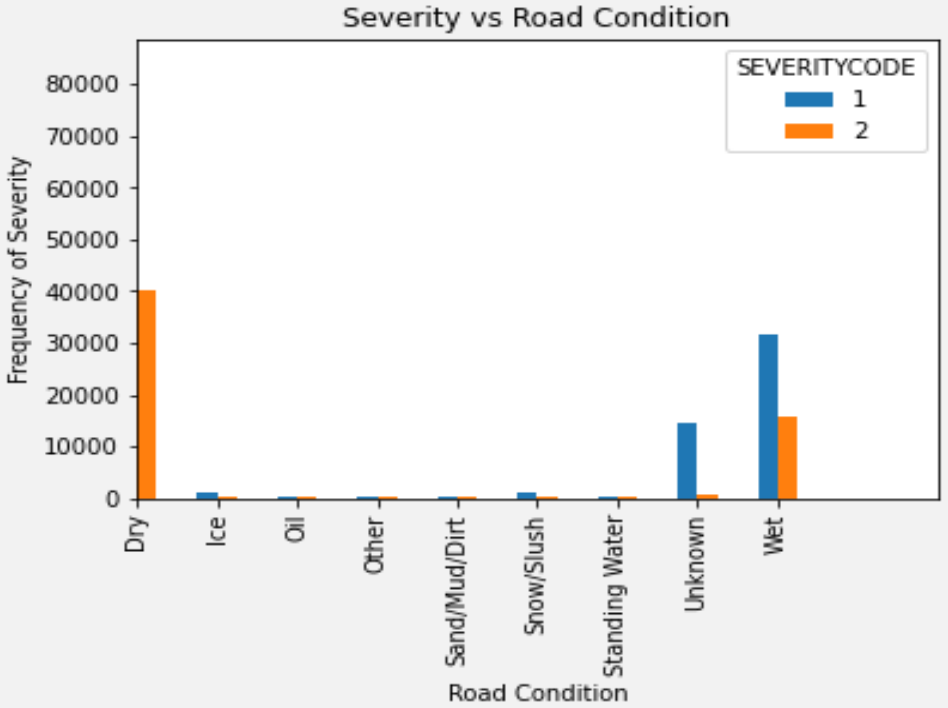


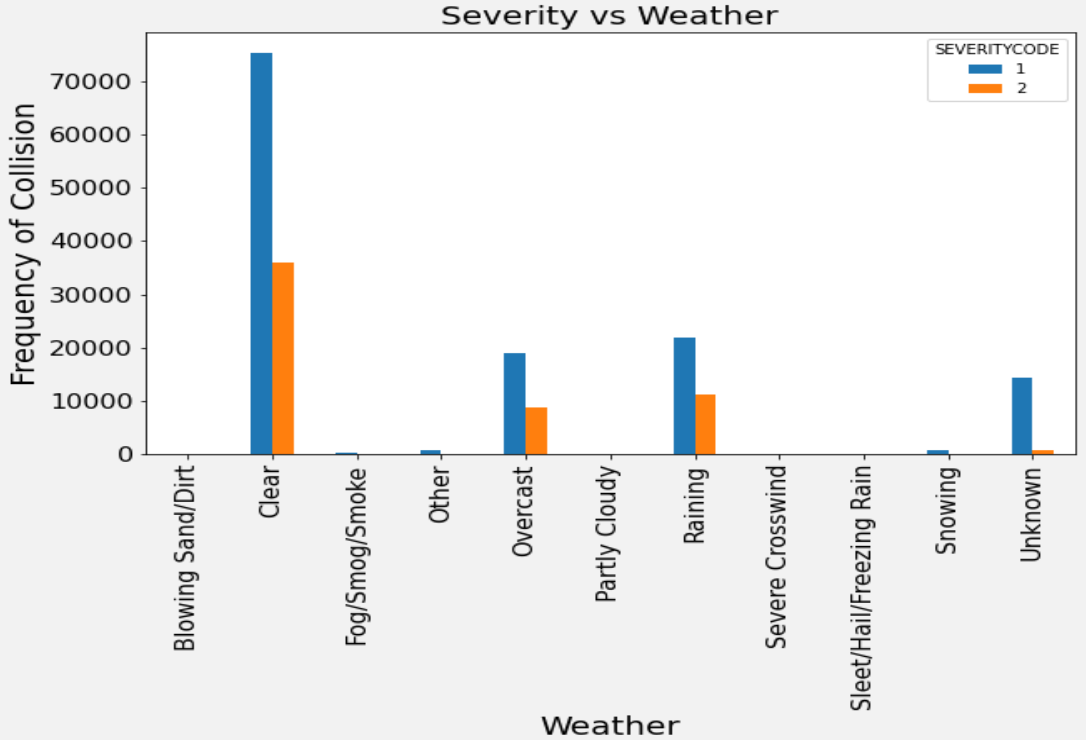
- **Feature selection:**
  - Weather condition
  - Road condition
  - Light condition

Methodology

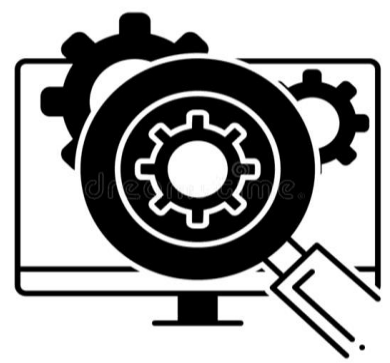


- Data Exploration
  - Road condition
  - Weather condition

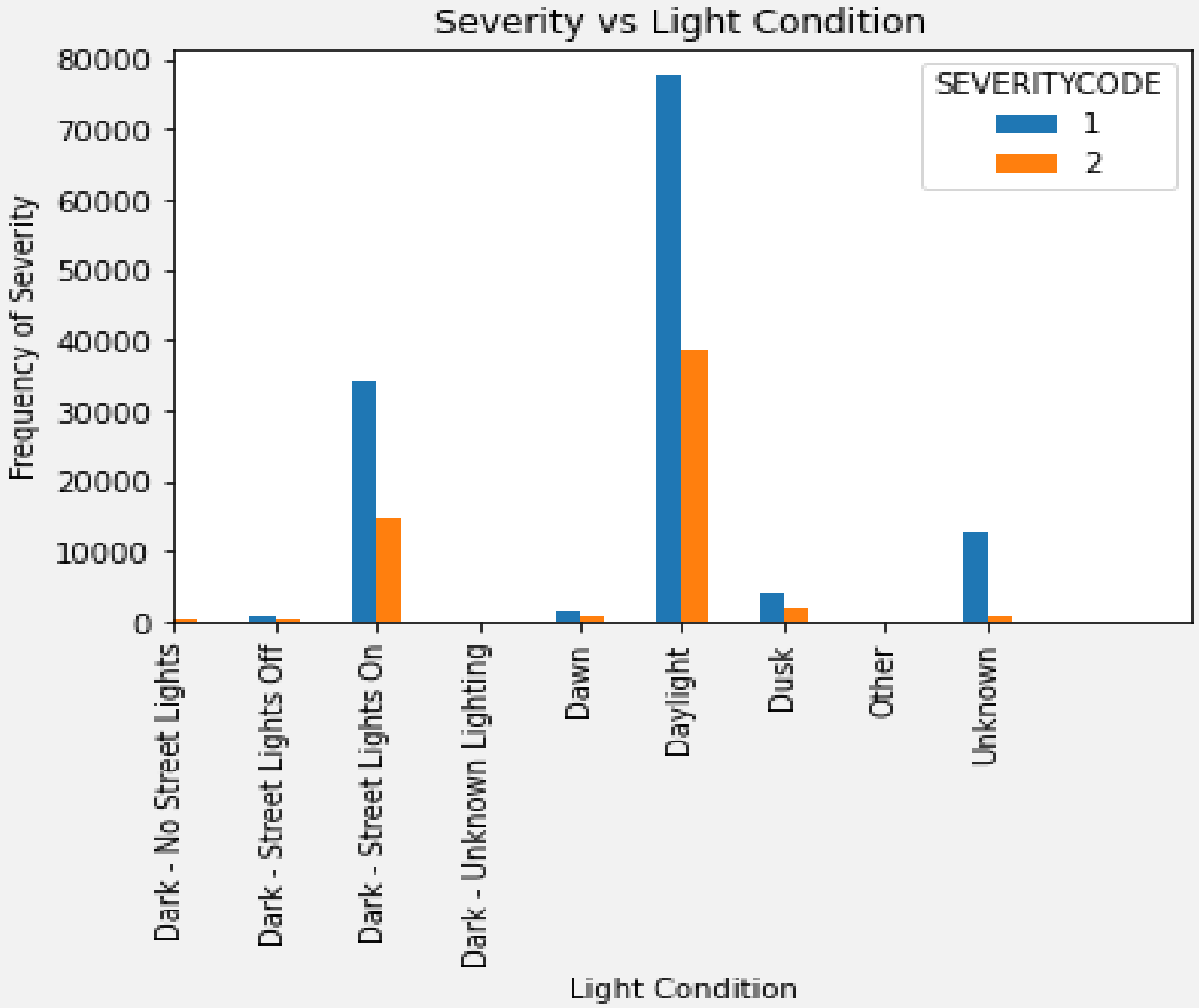




Methodology



- **Light condition**: We observe that most of the accidents occurred during “daylight.” for both property damage and body damage.



Results

- The 4 models used had an accuracy of about 0.5

Models	F1 Score	Precision	Recall	Accuracy
Decision Tree	0.42	0.63	0.31	0.56
Logistic Regression	0.41	0.63	0.3	0.56
Naïve Bayes	0.66	0.5	0.98	0.5
Linear Discriminant Analysis	0.44	0.6	0.35	0.55

Discussion

- The models had very low accuracy
- More features are required to improve the model accuracy
- Implementing hyper-parameter tuning could also help improve the model
- Ensemble learning by combining the 4 different models could also result into a better accuracy of the model
- The models can further be improved by using cross-validation

Conclusion

- We can conclude from the results obtained from the metrics that predicting the severity of an accidents will require more features to make better prediction.
- Using different algorithms further confirms that the low accuracy obtained was not due to the choice of the algorithms.
- However, the model obtained could still be used to predict the severity of accidents. Motorist as well as first responders would be beneficiary to the accident model prediction tool when deployed