Predicting the Severity of an Accident

The Problem

- 56,000 accidents last year in Seattle*
- Large amounts of data is being collected on accidents including its location, the weather, the number of people in the car, and many more
    - However it could be used more effectively
- Drivers are concerned for their safety when driving
- This information could be used to reduce the number of accidents

*via Kornfeld Law

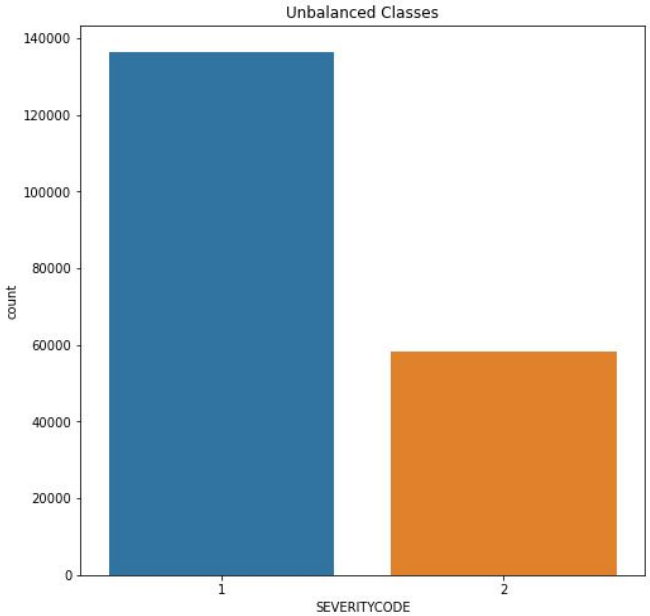
Methodology

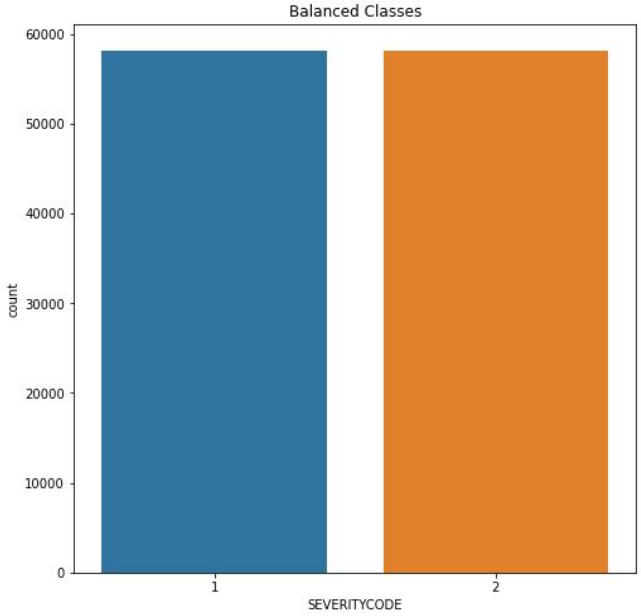
- Build a machine learning model using the nearly 200,000 accidents to be able to predict the severity of an accident given certain conditions
- The data contains many attributes, however only three were selected to avoid overfitting the model
    - Weather (overcast, raining, clear etc.)
    - Road conditions (wet, dry etc.)
    - Lighting conditions (daytime, nighttime, with or without street lamps etc.)

Methodology Cont.

- Balancing data
  - Before balancing, there were over 2x as many instances of severity code 1 compared to severity code 2
  - If the model were built without balancing, it could lead to a biased model
  - Shuffled the data and created a new dataframe with the same instances of both severity code 1 and 2

Before and After Balancing





Methodology Cont.

- Many machine learning techniques require numerical values for features
  - All three independent variables - weather, road conditions, lighting conditions - are categorical
  - Used a technique called label encoding to transform the variables into numerical values
  - Each unique value in a column is converted to a number - for example wet gets attached a 0, and dry gets attached a 1

Dataset after label encoding

TYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
2	Raining	Wet	Daylight	6	8	5
2	Overcast	Dry	Dark - Street Lights On	4	0	2
2	Clear	Dry	Daylight	1	0	5
2	Clear	Dry	Dark - Street Lights On	1	0	2
2	Clear	Dry	Daylight	1	0	5

Building the model

- Finally the model was built
  - Used logistic regression as it is useful for predicting the probability of a value belonging to a class
  - Split the data into X, or the independent variable, and Y, the dependent variable
  - Data was normalized to change the values to a common scale
  - Model was used to predict whether an accident would be a severity code 1 or 2

Performance

- It’s very important to evaluate the models performance
  - This was done using the Jaccard index, Logarithmic loss, and F1 score to get a balanced understanding of the models accuracy
  - Their scores were 0.527, 0.684, and 0.51 respectively
  - Overall the model was able to predict the severity code of an accident with moderate accuracy

Recommendations

- While the model was able to predict the severity of an accident, it was only able to do so with moderate accuracy
- To improve the model
  - One recommendation would be to add more independent variables to improve the accuracy of the model, but also its usefulness to users
  - Secondly, perform more feature engineering on the dataset to get the most out of the features chosen
    - This could include dealing with missing values, using an objective method for selecting features, and using a better method for transforming categorical variables into numerical ones