Predicting Collision Severity in Seattle

IBM and Coursera Applied Data Science Capstone Project

Tom Walter, 04 September 2020

1.Introduction: Problem & Goal

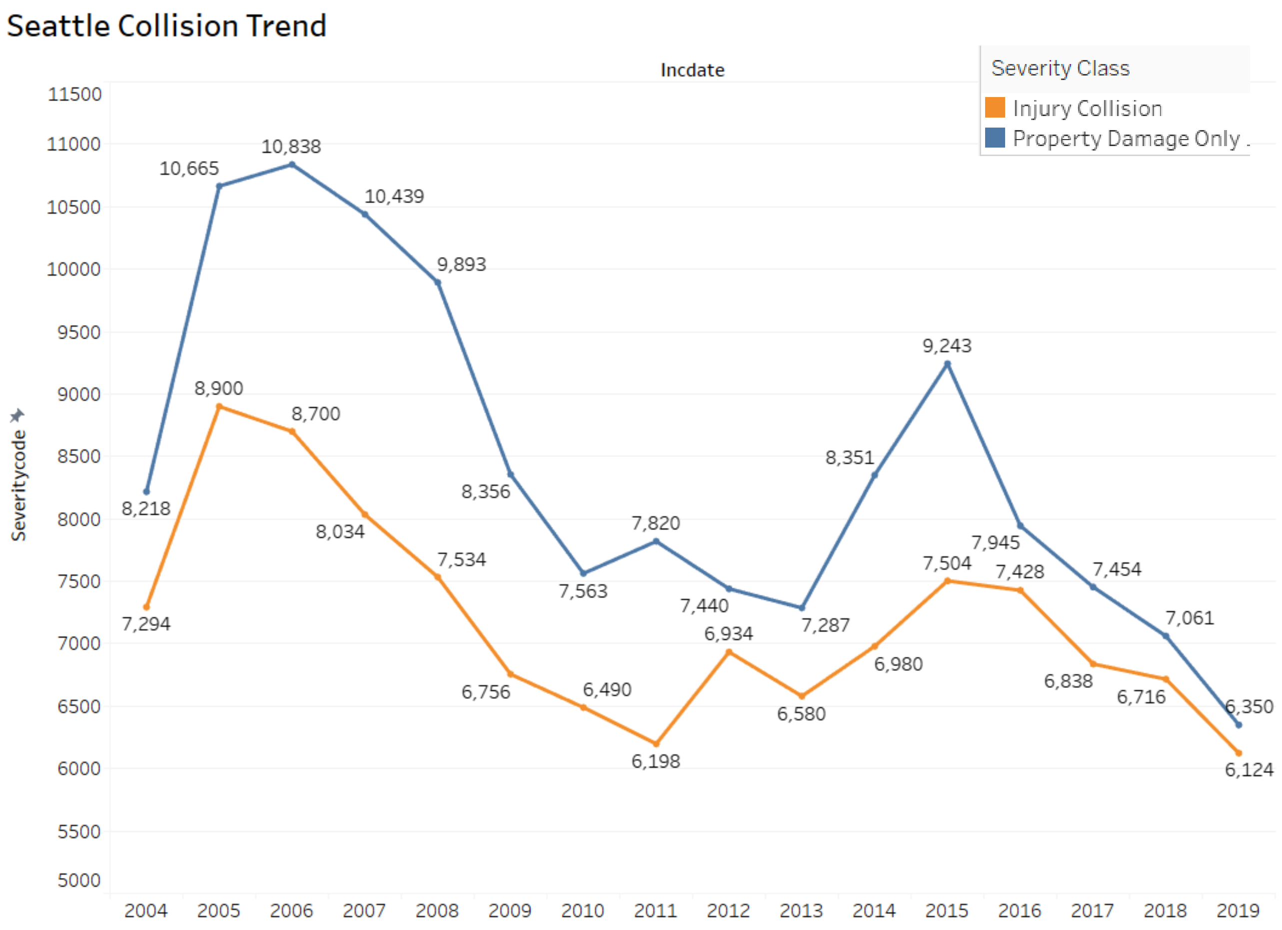
Problem:

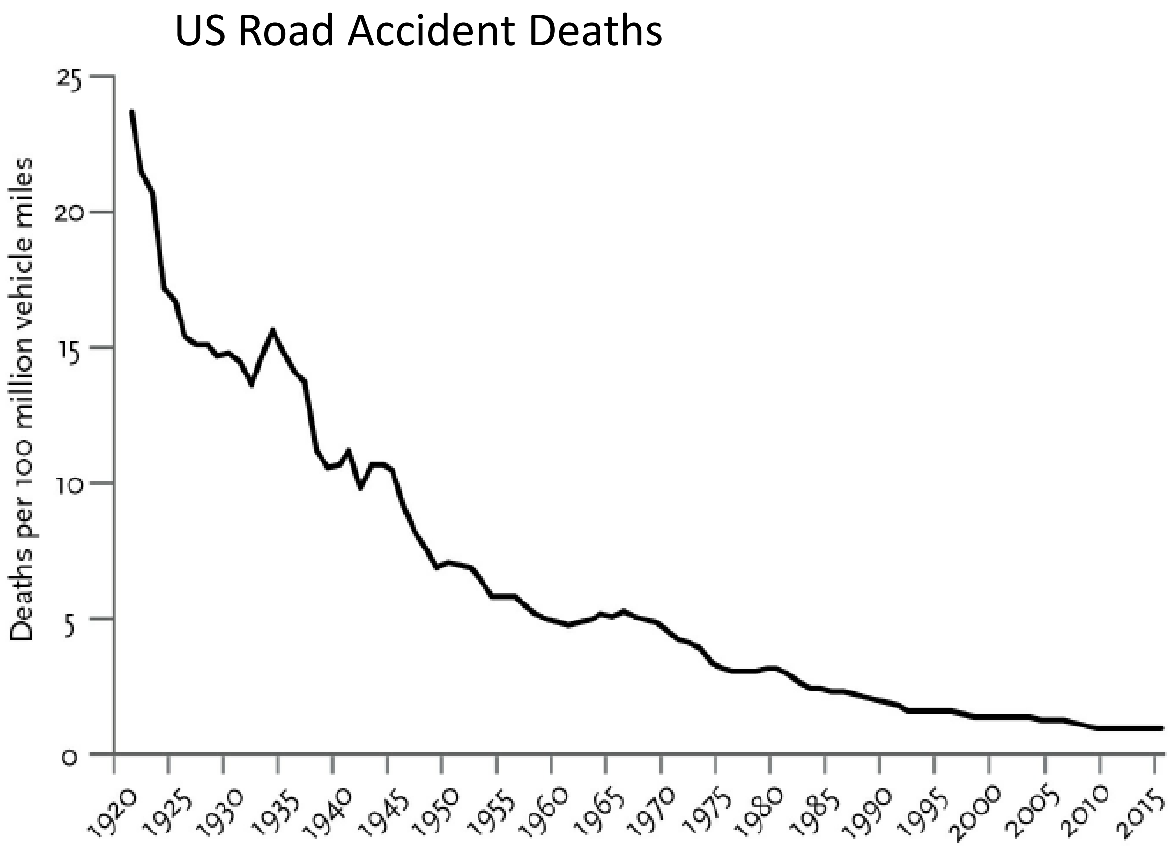
- accident collisions cause property damage, injury, or death
- since cars invented, many auto safety measures followed:
  - manufacturers: seatbelts & airbags
  - civil engineers: guard rails & traffic lights
  - government: traffic laws & enforcement (speeding, drunk driving, etc.)
- US road accident deaths at all time low
- BUT: every collision remains public health risk!

Goal:

- Machine Learning: Classification
  - predict severity class
  - identify causes of severe collisions
  - help reduce severity & total number of accidents
- Deployment Options:
  - electronic warning signs
  - road improvements
  - future: feed data to AI cars
- Make drivers & pedestrians saver!

US Decline in Accident Severity





2. Data, Hypothesis, & Feature Selection

Dataset:

- Collision Data by SDOT Traffic Management Division in Seattle, WA
- from 2004 to 2020
- 194,673 reported collisions
- 38 attributes about collisions
- unnecessary features dropped

Hypothesis:

- severity of collision is a function adverse driving conditions and negligent human behavior

y = severity class

y = f(x) = driving conditions + human behavior + timing

Table 1: Pre-Selected Features

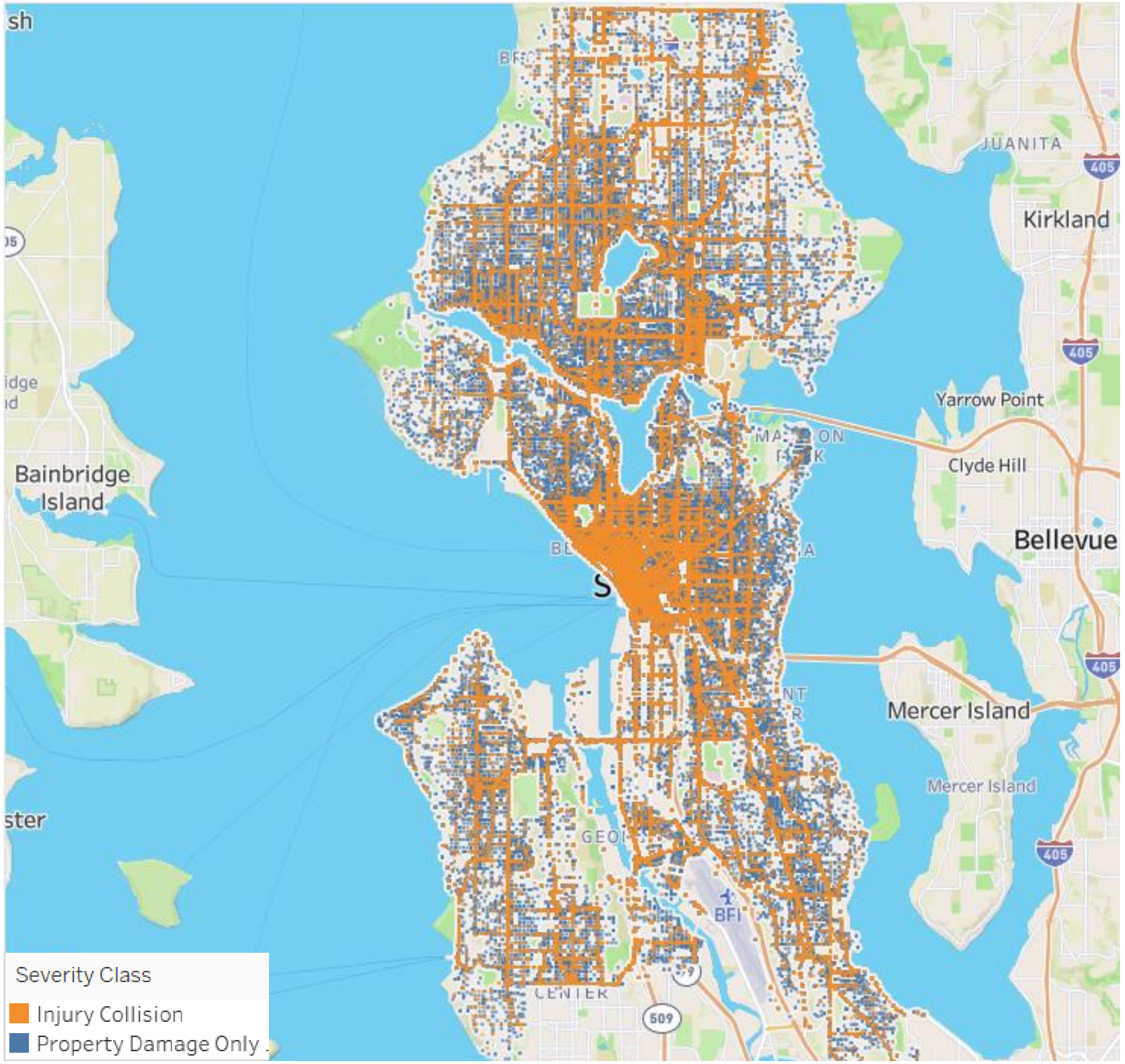
	Feature	Description
1	SEVERITYCODE	severity class of the collision: 1 = property damage 2 = injury collision
2	LONGITUDE	longitude
3	LATITUDE	latitude
4	JUNCTIONTYPE	category of junction
5	WEATHER	weather conditions
6	ROADCOND	road conditions
7	LIGHTCOND	light conditions
8	INCDATE	date of the incident
9	INDTTME	date & time of the incident
10	INATTENTIONIND	whether collision was due to inattention
11	UNDERINFL	whether driver was under the influence of drugs/ alcohol
12	SPEEDING	whether speeding was a factor in the collision

3. Exploratory Data Analysis

Driving Conditions - Location

- LONGITUDE and LATITUDE pinpoint collisions on map
- missing values are replaced by mean

Map of Collisions

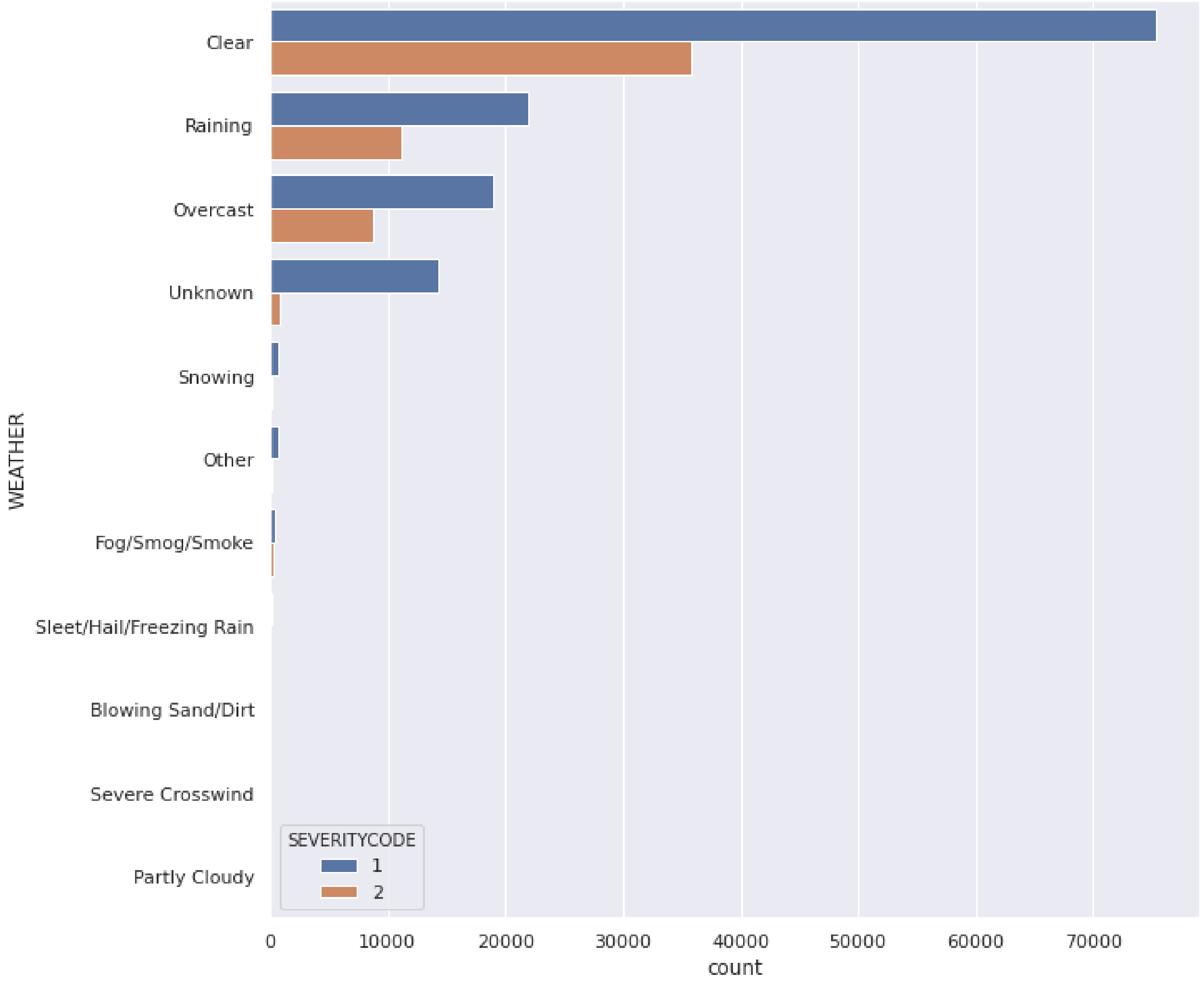


3. Exploratory Data Analysis

Driving Conditions - Weather

- mode = clear weather
- mode will replace 5,081 missing values

- worse weather ≠ worse collision
- *counter-intuitive conclusion*

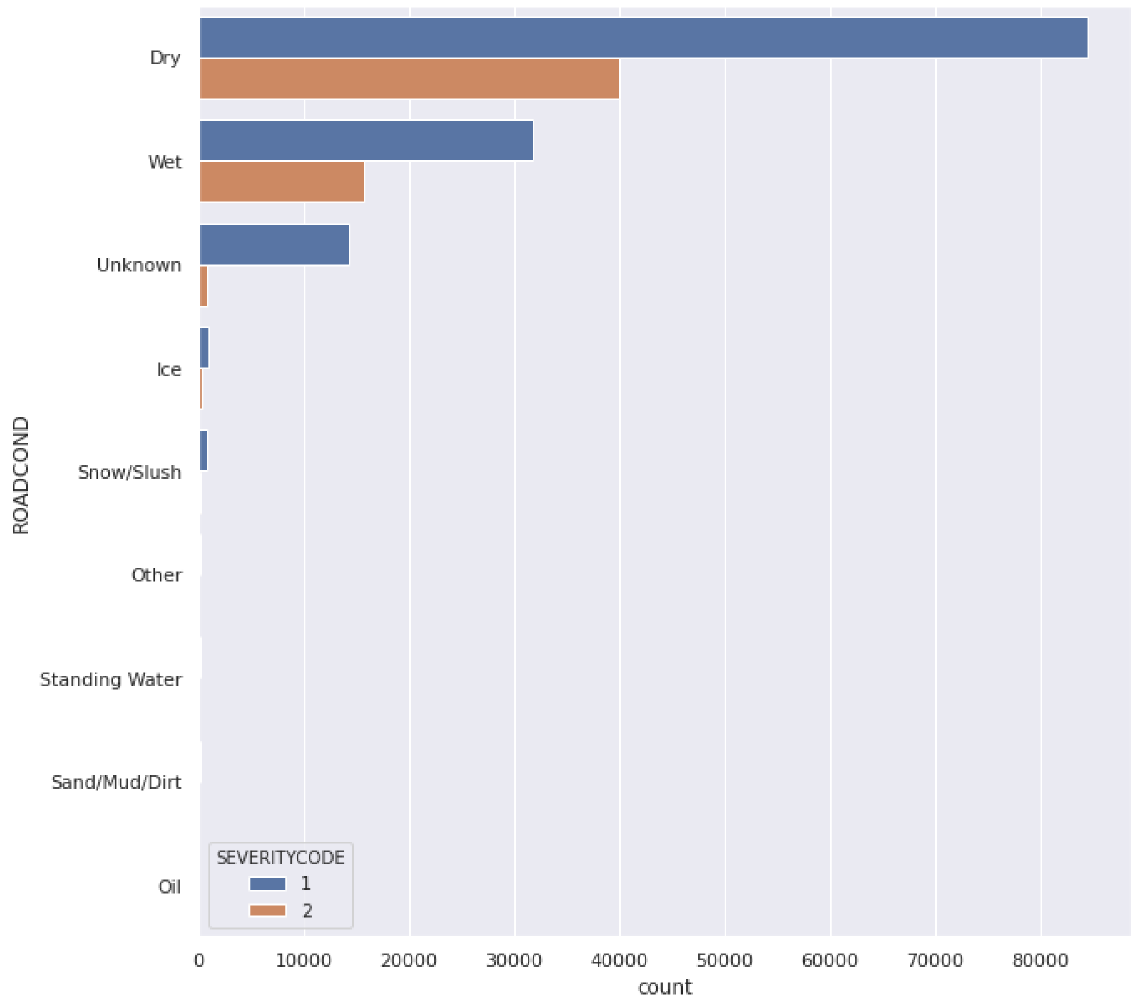


3. Exploratory Data Analysis

Driving Conditions - Road

- mode = dry roads
- mode will replace 5,012 missing values

- worse road ≠ worse collisions
- *counter-intuitive conclusion*

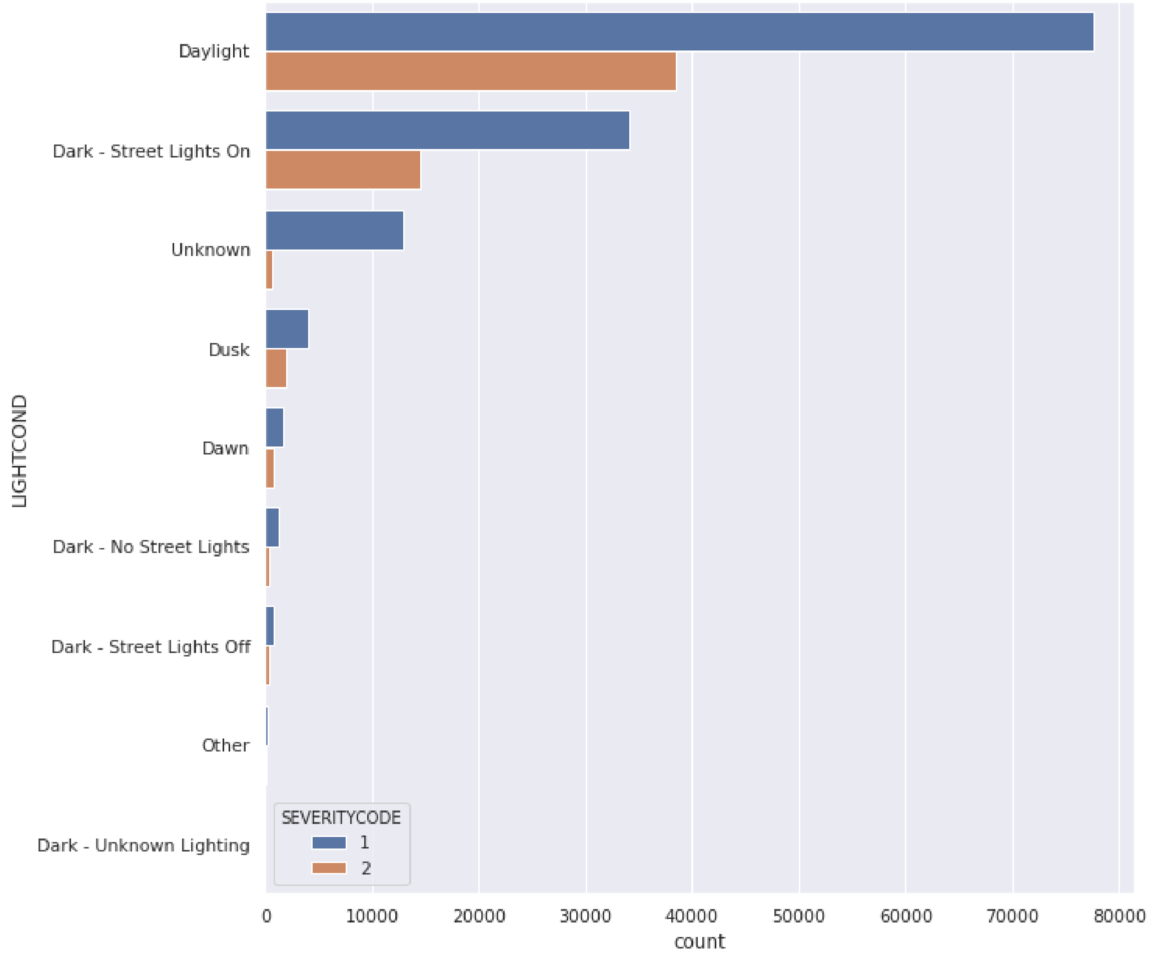


3. Exploratory Data Analysis

Driving Conditions - Light

- mode = daylight
- mode will replace 5,170 missing values

- worse lighting ≠ worse collisions
- *counter-intuitive conclusion*

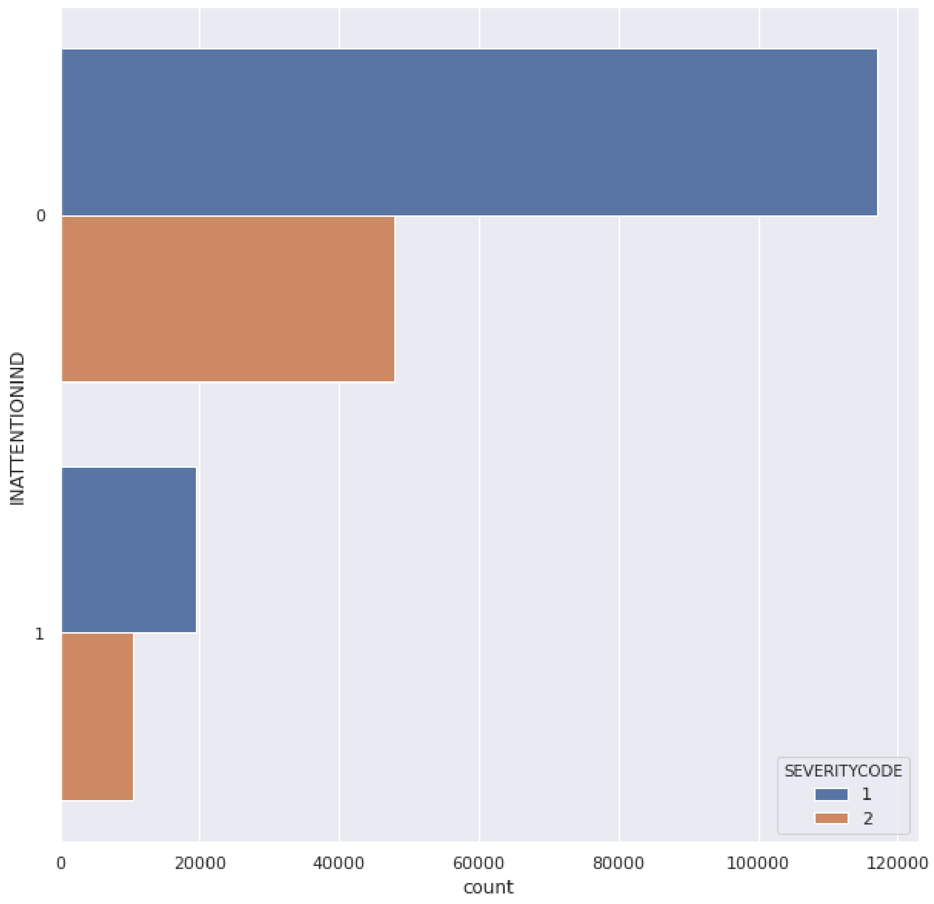


3. Exploratory Data Analysis

Human Behavior - Inattention

- mode = no
- mode will replace 164,868 missing values

- most accidents happen even when people pay attention
- *counter-intuitive conclusion*

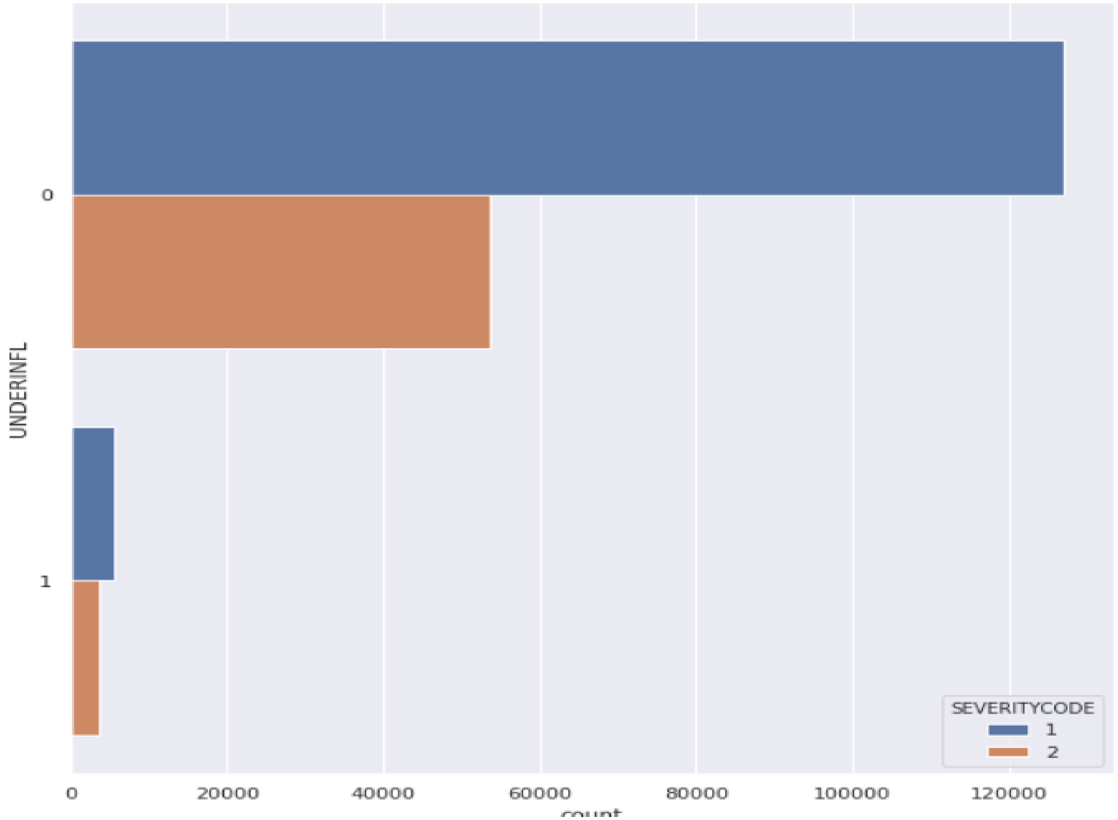


3. Exploratory Data Analysis

Human Behavior – Alcohol/Drugs

- mode = no
- mode will replace 4,884 missing values

- most accidents happen when people are sober
- *counter-intuitive conclusion*

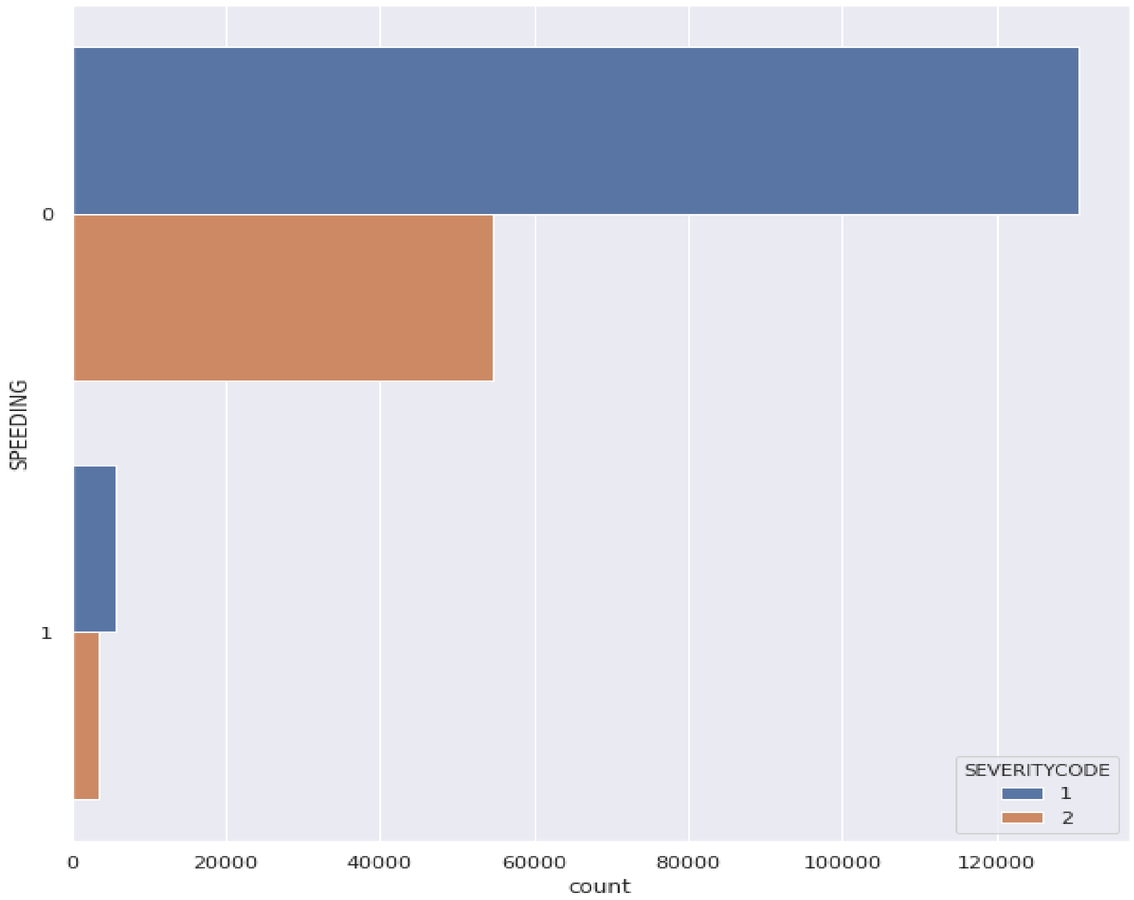


3. Exploratory Data Analysis

Human Behavior – Speeding

- mode = no
- mode will replace 185,340 missing values

- most accidents happen when people obey speed limit
- *counter-intuitive conclusion*

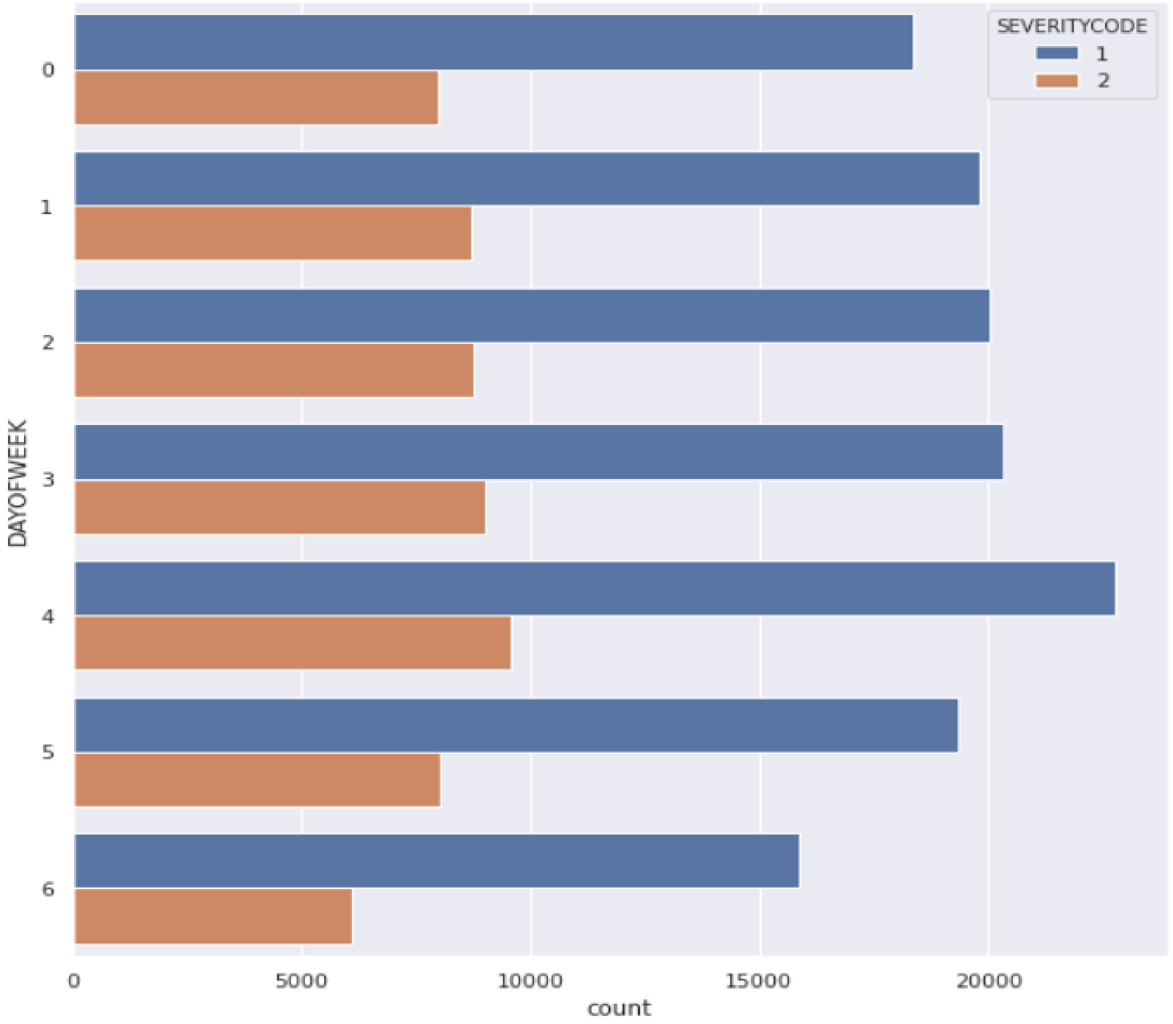


3. Exploratory Data Analysis

Timing – Day of the Week

- mode = Friday

- more accidents happen during workdays than on weekends
- *intuitive conclusion*



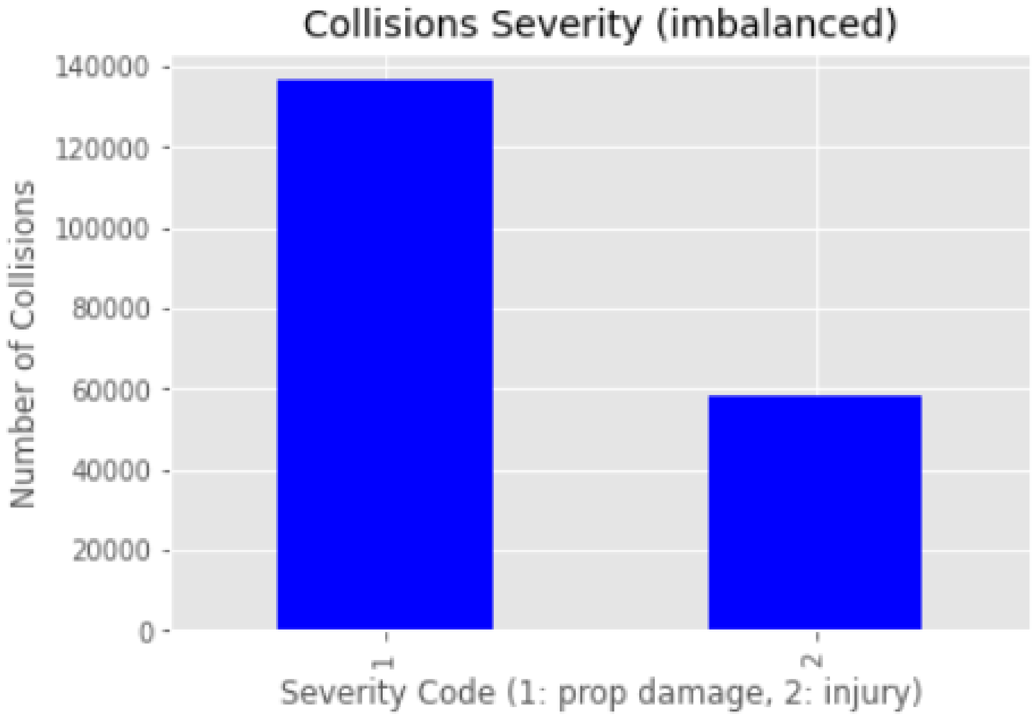
(0 = Monday, 1 = Tuesday, 2 = Wednesday, 3 = Thursday, 4 = Friday, 5 = Saturday, 6 = Sunday)

4. Machine Learning

Balancing the Dataset

- total collisions = 194,673
- class 1 = 136,485
- class 2 = 58,188

- imbalance will bias machine learning to majority class
- Random Under Sampling is used for balancing labels
- 58,188 (class 1) + 58,188 (class 2) = 116,376 cases for ML



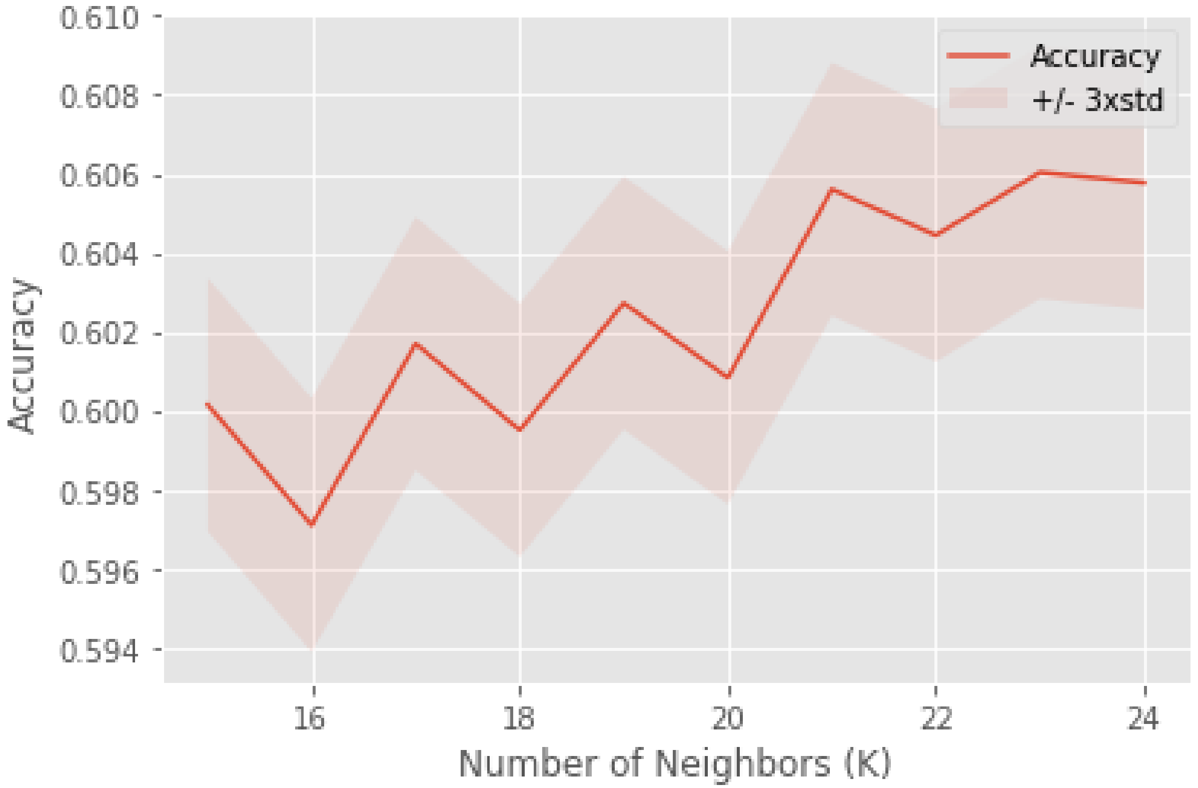
4. Machine Learning

- assumption = combination of independent features in dataset will have recurring patterns connected to severity class
- ML will find the patterns & combination of features that 'predict' severity class
- other application of ML classification
  - spam, fraud, & churn prediction
  - handwriting & face-recognition
  - extreme events
  - medical diagnosis
- common classification algorithms:
  - K-Nearest Neighbors
  - Decision Tree
  - Random Forest
  - Logistic Regression
  - Artificial Neural Networks

4. Machine Learning

K-Nearest Neighbors (KNN)

- stores all cases and classifies a new case based on its similarity to its 'nearest neighbors'
- e.g. an unknown case is compared to 5 neighbor cases
  - 3/5 neighbors are class 2
  - 2/5 neighbors are class 1
  - unknown case classified as class 2
- find best K=number of neighbors

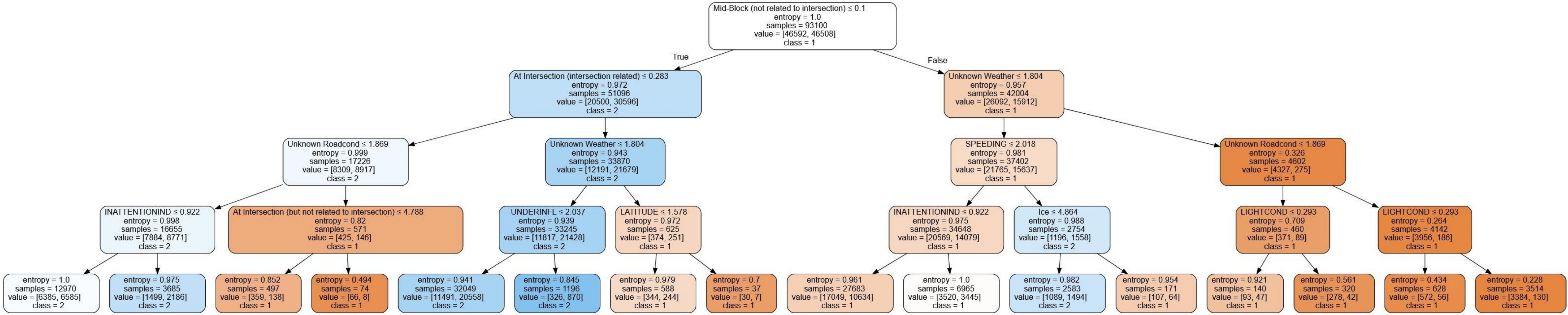


best general accuracy of 0.6060 with k=23

4. Machine Learning

Decision Tree

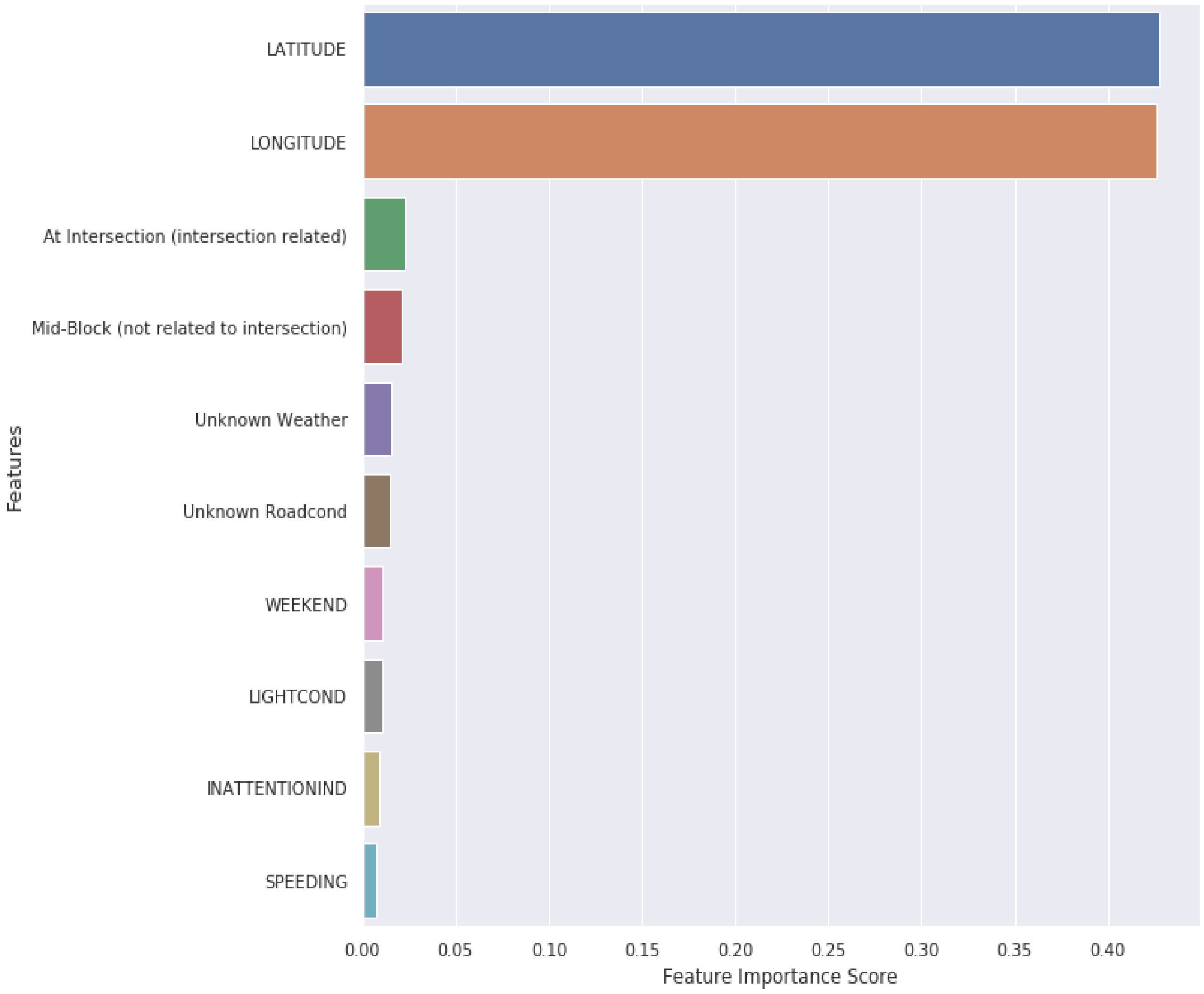
- are called tree
    - leaves = class labels
    - branches = conjunctions of features
    - leaves are pure when completely homogenous (no more entropy)
- trees mimic human decisions
- general accuracy = 0.5719
- example tree below (max depth=4)



4. Machine Learning

Random Forest

- are called forest:
  - multiple decision trees
  - random sub-samples of data
  - also reducing “entropy”
- general accuracy = 0.5870
- Feature Importance Scoring:
  - which features most important for outcome



4. Machine Learning

Logistic Regression

- common stat. method for binary classification
- can also estimate probability of a case falling into a class
- provide several solvers:
  - 'liblinear'
  - 'SAG'
  - 'SAGA'

Scores:

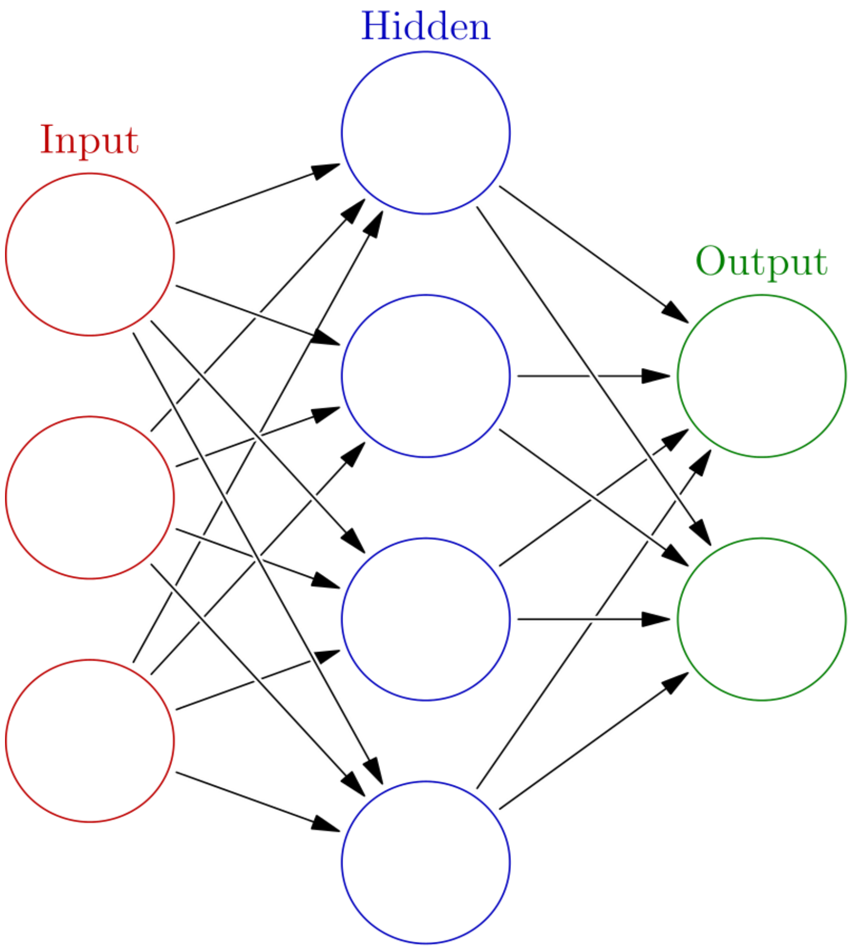
- liblinear gen. accuracy = 0.6199
- SAG gen. accuracy = 0.6199
- SAGA gen. accuracy = 0.6199

- liblinear is recommended for large-scale and high-dimension dataset

4. Machine Learning

Artificial Neural Networks

- loosely mirror neurons in a biological brain
- neurons have connections & layers
- learning:
    - layer of neurons receive input data
    - transform the data
    - send data to next layer of neurons
    - iteration until converge on functions with minimal error
- gen. accuracy = 0.6243



5. Evaluation

General Accuracy

- % of how many predictions were correct of all prediction
- higher = better, range 0-1

Jaccard-Score

- % overlap between predicted and actual class sets
- higher = better, range 0-1

F1-Score

- balance between true positives and false positives
- higher = better, range 0-1

Log-Loss

- only for models with probability estimation
- uncertainty of predicted probability
- lower = better, range 0-1

Table 2: Formal Evaluation Metrics

	Gen. Accuracy	Jaccard-Score	F1-score	Log-Loss
K-Nearest Neighbor	0.606032	0.414656	0.605196	NaN
Decision Tree	0.571920	0.411111	0.571595	NaN
Random Forest	0.587085	0.410946	0.587052	NaN
Logistic Regression	0.619995	0.440933	0.619862	0.643860
Neural Network	0.624377	0.447345	0.624300	0.640896

Table 3: Variation in Accuracy Scores

	Gen. Accuracy	Jaccard-Score	F1-score	Log-Loss
mean	0.602	0.425	0.602	0.642
std	0.022	0.018	0.022	0.002

5. Evaluation

Top 3 ML Classification Models:

1. Artificial Neural Networks
2. Logistic Regression (liblinear)
3. K-Nearest Neighbors

Link to the Full Report:
https://github.com/tom-walter/Coursera_Capstone/blob/master/Tom%20Walter%2C%20Full%20Report%20Capstone.pdf