Exploring and Predicting Car Accident Severity in the U.S.

By Andrew M, Eric W, Gabriel W, Saurabh S

Problem Statement

- Car accidents are a leading cause of death in US
- Major financial burden

Based on US Accident data from 2016-2021, what recommendations and observations can be made towards reducing severity of automobile accidents?

- Benefits:
    - General public (i.e. pedestrians, bikers, drivers, etc.)
    - Policymakers and city planners

Data Acquisition

- Compiled US Accident data hosted on Kaggle (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)
- Data consists of car accident data across the US from 2016-2021
- Contains geospatial, temporal, weather related features.
- Approx. 3,000,000 samples

Data Cleaning

- Imputing Values:
  - Missing precipitation amount imputed with 0 where weather condition was not wet (majority)
  - Missing wind speeds imputed with 0 where wind direction was specified as “calm”
- Removing unnecessary features
  - Latitude, longitude, wind direction, and others not relevant to analysis
- Missing Data
  - Small remaining number of missing values (relative to full dataset) removed
  - After cleaning, 2,731,050 samples

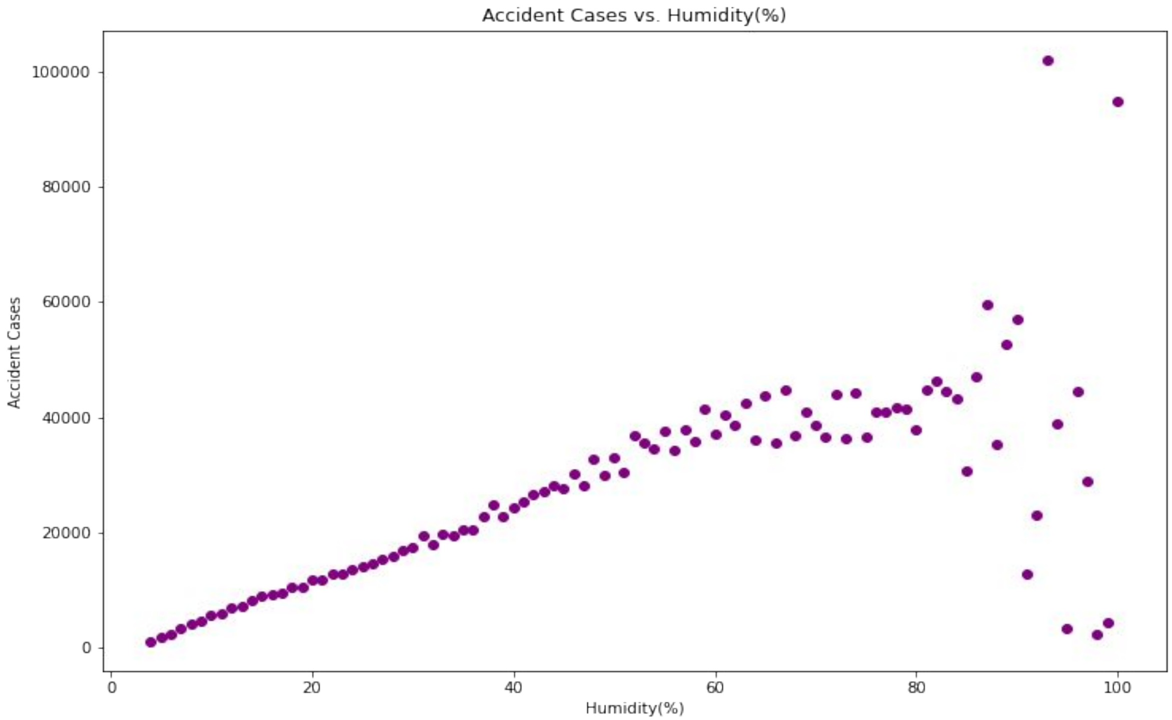
Features

- Accident Severity - the target feature. Provided on 1-4 scale but reduced to 0-1 (low, high) for purposes of analysis.
  - Target heavily imbalanced with 90% of dataset being low severity
- Weather Features
  - Precipitation amount, Humidity, Air pressure, Wind speed, Temperature, and others
  - Includes a column with weather categories (‘rainy’, ‘windy’, ‘snowy’, etc.)
- Road Features
  - Nearby Signs, Rotaries, Speed Bumps, and others
  - Street Type (Highway, Freeway, Interstate, etc.)
- Time Features
  - Year, Date, and time; time of day (sunrise, sunset)

Findings

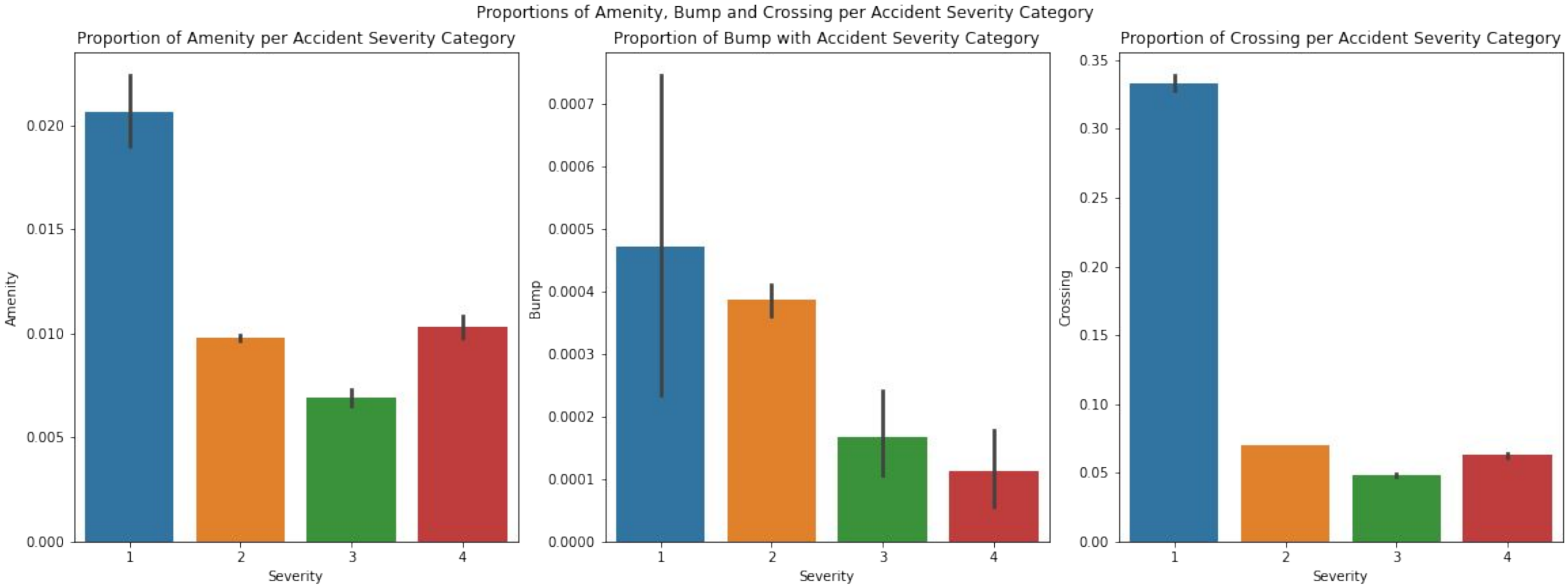
- There appears to be a positive, linear relationship between humidity and frequency of accidents

- Most accidents occur for humidity levels around 0-40%



Findings

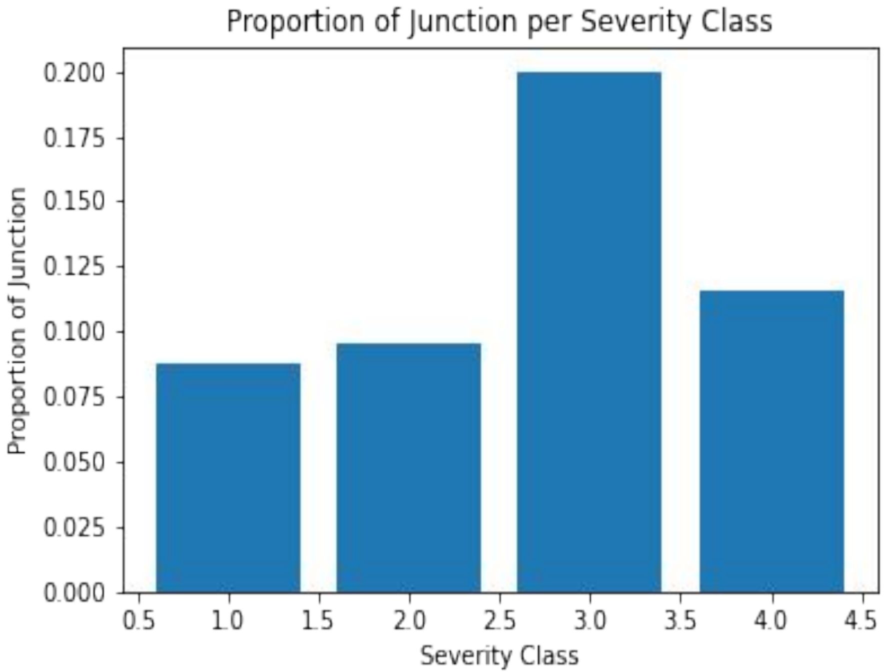
- Low severity accidents (class 1 and 2) have a higher proportion of bumps, crossing, and amenities
- The same can be said for the presence of roundabouts, railway stations, and traffic signals



Findings

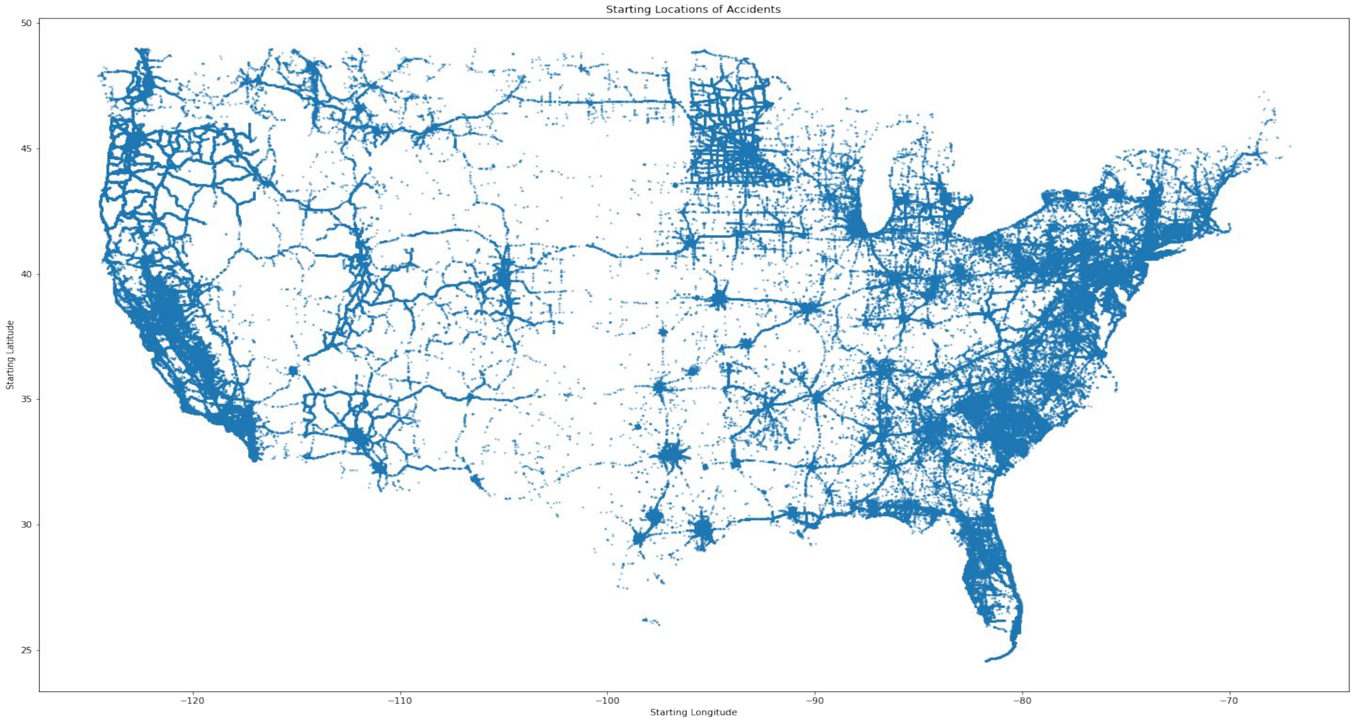
- High severity accidents (class 3 and 4) have a higher proportion of junctions/intersections

- Suggests that intersections increase accident severity



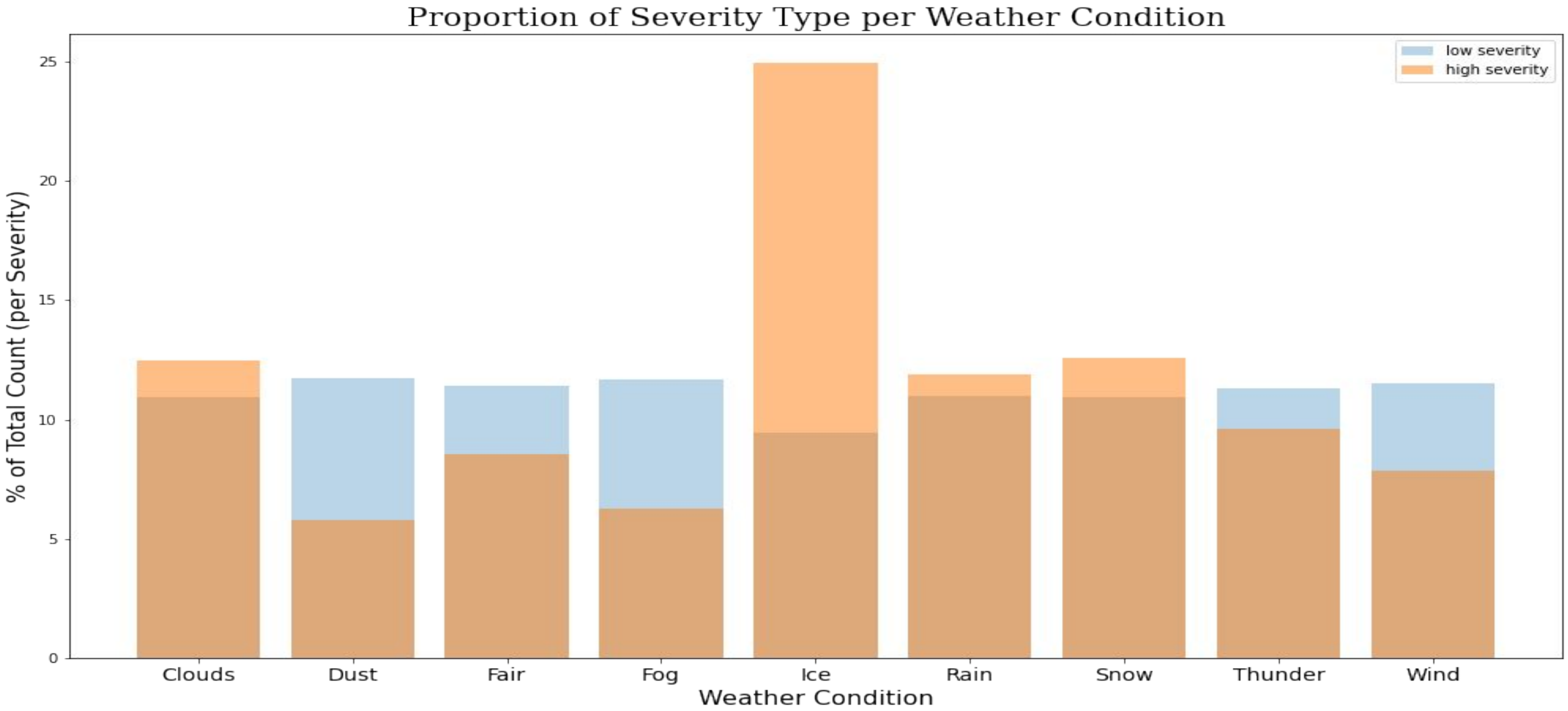
Findings

- Frequency of car accidents are largely concentrated around coastal areas
- Higher accident severities in California and many of its cities/counties as well as the PST



Findings

- Low severity accidents are even across the board, suggesting they occur at approximately the same rate independent of weather conditions
- In icy conditions, there is a higher proportion of high-severity accidents than low-severity accidents



Modeling Approach

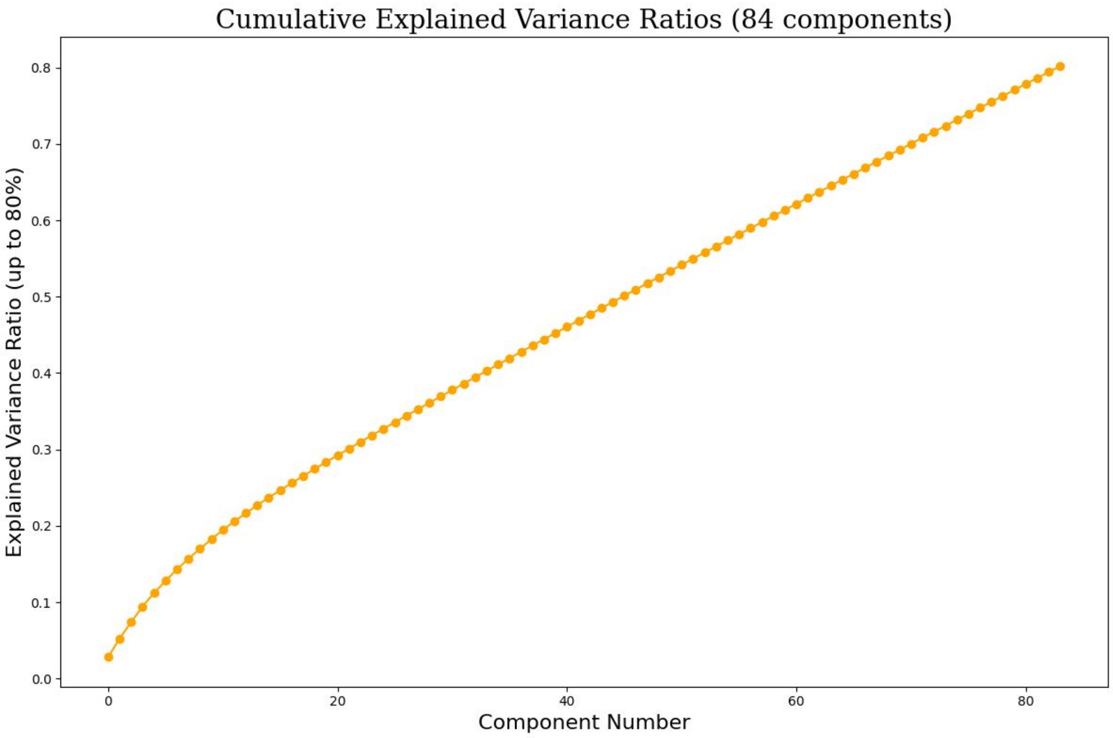
- Explore clustering and PCA analyses to improve subsequent classification models via transfer learning
- Goal:
    - Optimize for recall/sensitivity, balancing with accuracy
        - Reduce false negatives (incorrectly predicting a high-severity accident to be low-severity)
        - Maximize amount of correctly labeled accident severities
    - Interpretability to support policy/behavioral decision-making

Clustering

- Large dataset, so hoped to find subcategories of data that could be used to assist in analysis and modeling
- KMeans and KPrototypes algorithms - both ran really slowly!
- Required very small subsets of data to experiment with
- No clear clustering behavior, trying on larger subsets of data was time prohibitive

PCA

- Attempt to explore clustering from another angle - are there components that explain a large amount of the data?
- Answer: No! 84 components explain 80% of variance.



Classification

- Random sampling with 1 million sample dataset was used for modeling.
- MinMaxScaler used to transform features by scaling them to 0 and 1 values
- Performed synthetic minority oversampling technique (SMOTE) to deal with imbalanced classes.
    - Majority class had around 90% observation whereas the minority class had 10% observation.
    - This technique generated synthetic data for the minority class
    - After implementing SMOTE, we had a 50/50 split across both classes

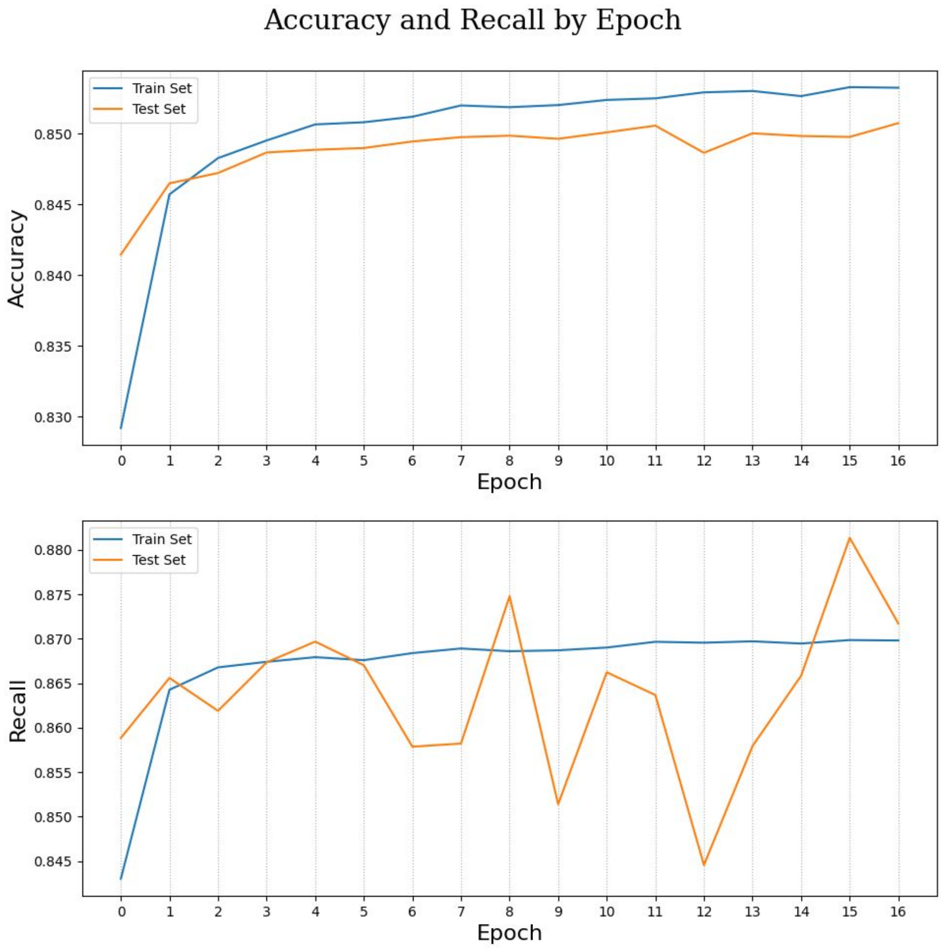
Classification

- We split the dataset into two subsets – Train (70%) and Test (30%)
- We took 2 passes at the Machine learning models, one with initial data and other with balanced data after performing SMOTE technique
- SMOTE Sampling methods provided much better results
- Random Forests has similar performance to Log. Reg but is significantly more computationally intensive
- XGBoost Model gave the highest accuracy of 95%

Model	Accuracy
Logistic Regression	82.70%
Random Forests	82.20%
XGBoost	95.08%

Neural Networks

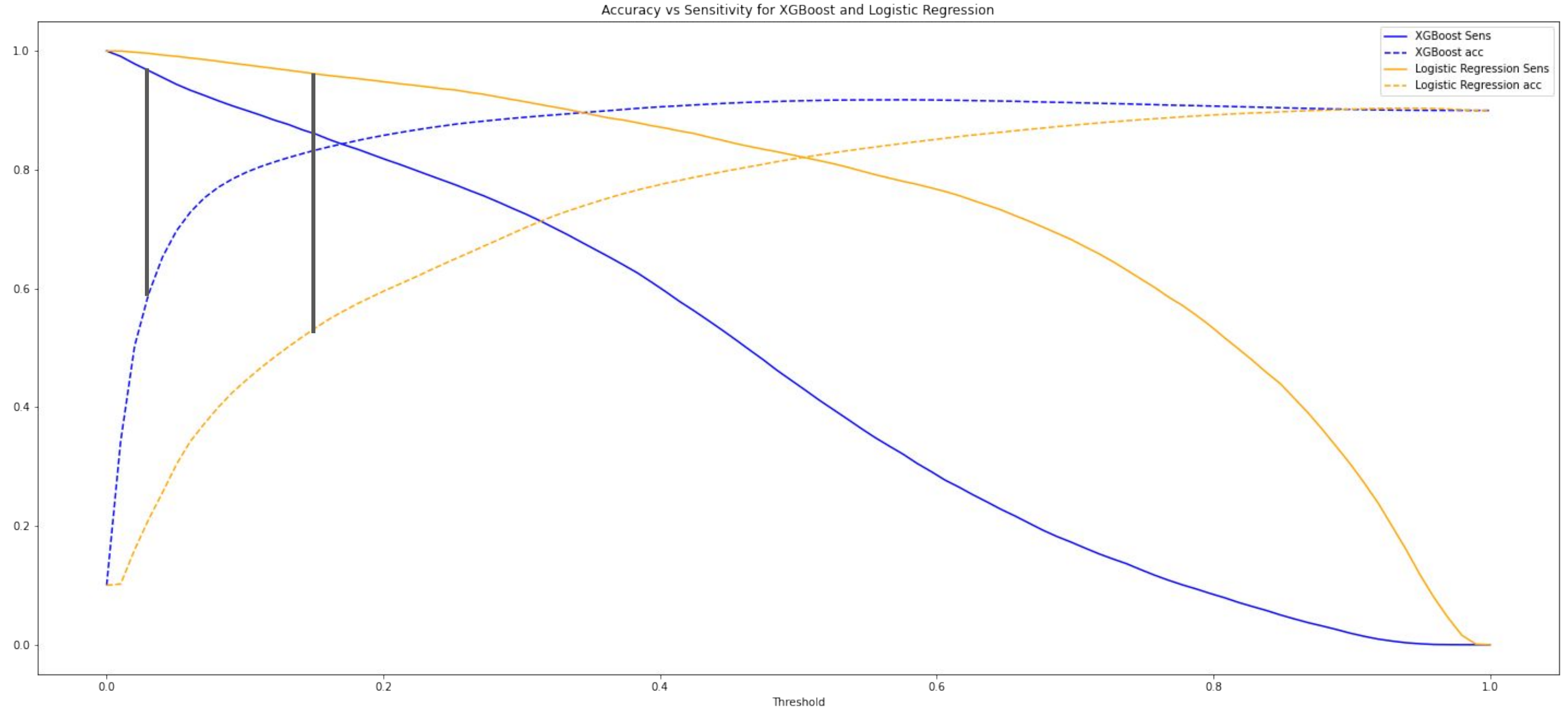
- Used to validate finding and/or see if additional accuracy could be achieved.
- Tried to optimize for recall, but not steady (though within a high range)



Neural Networks

- Simple 1 layer Neural Network
  - Cross-validated mean: Accuracy 0.8; Recall 0.9
  - Similar score to Logistic Regression; slightly better recall
- More complex 3 hidden layer Neural Network
  - Little difference in performance from 1-layer Neural Network
- Supports other models, but not interpretable

Model Comparison



Final Model and Demonstration

- XGBoost: 95% Sensitivity, 69.5% Accuracy
    - Baseline: 0% Sensitivity, 90% Accuracy

DEMO TIME!

Summary of Findings

Increasing Accident Severity

- Months: Winter Months showed strong model impact (icy conditions are more likely to make any accident highly severe)
- Time zone and State/County/City (PST and particularly California)
- Time: 6AM and 5pm (aligns with heavy traffic times)
- Presence of a junction/intersection

Increasing Accident Frequency

- Humidity (%)

Decreasing Accident Severity

- Presence of:
  - Bumps, Crossing, Roundabouts, Railway Stations, Traffic Signals

Limitations

- Missing key features
  - Features such as speed at time of accident, and other subjective measures are missing from data
  - All of our models have similar accuracy
- Computational power
  - Clustering on larger samples of dataset was time-prohibitive
  - Other models required substantial resources
- Imbalanced data
  - Needed to under-/over-sample data to reduce imbalance during model fitting
  - Despite this, models performed reasonably after adjustment sampling

Recommendations

- Speed bumps and signs work!
- Intersections and junctions are dangerous. Can extra measures be put in place in these locations?
- Salt/sand icy roads - any accident on ice is much more likely to be high-severity
- Look into traffic remedies to alleviate rush hour congestion or encourage conservative driving

Next Steps

- Collect more subjective data concerning accidents: speed of cars, possible distractions in area at the time
  - Models suggest that this data is not capturing the full reason for these accidents - there is likely something else going on
  - Collecting this data is much harder
- Utilize AWS / other cloud computing techniques for increased computational power to improve model tuning
- Explore our findings by incorporating further domain knowledge (i.e. states / time zone relations with accident severity)

Thank you!