UNIVERSITY of VIRGINIA | SCHOOL of DATA SCIENCE

Traffic Accident Severity in the Continental United States

Kaleigh O'Hara, Eileanor LaRocco, Layla Ranjbar

December 4, 2024

Table of Contents

Executive Summary

- Can we adequately classify the expected traffic resulting from accidents to assist Lyft and other rideshare drivers?
- **Objective:** want to minimize the wait time in traffic for the shortest ride
  - Want to maximize true positives

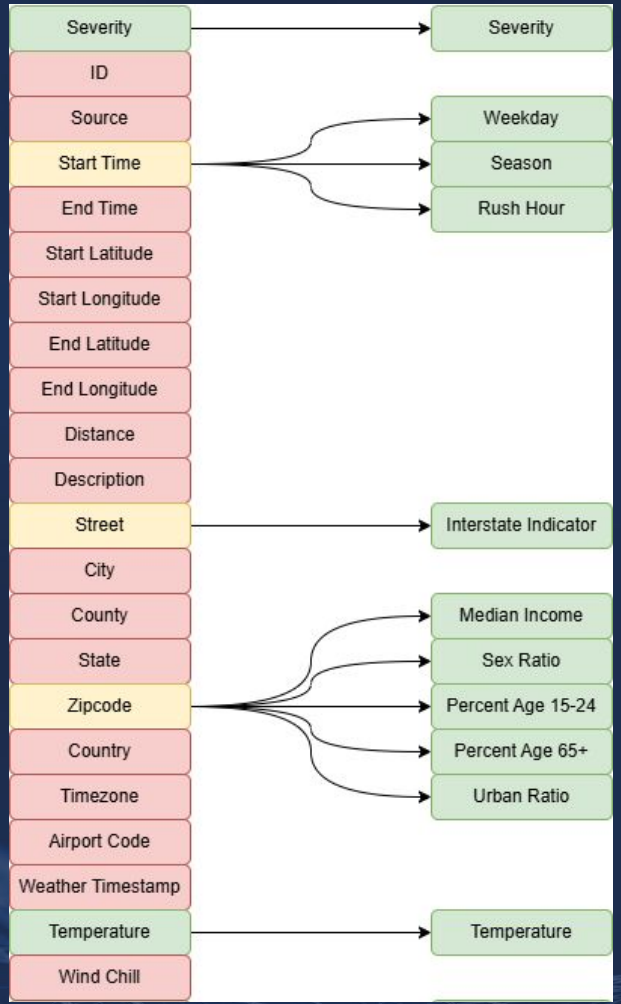
Project Significance

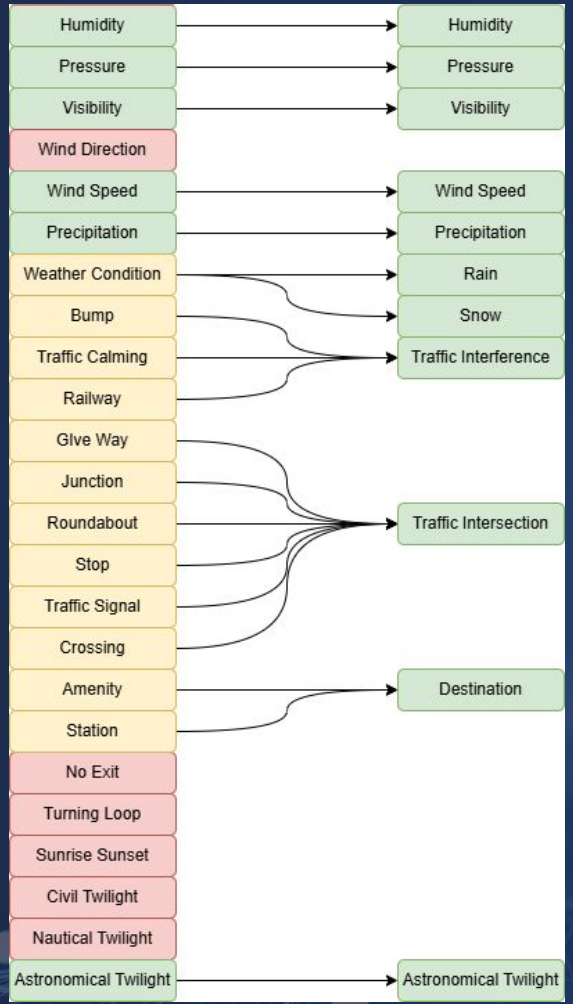
- Late arrivals results in customer dissatisfaction
- Decrease ride time to increase total number of rides provided

Data Engineering Flowchart

Dropped Outliers

- Wind Speed > 318 mph
- Pressure > 31.42 in
- Temperature > 130 F
- Max Visibility = 40





Auxiliary Data Summary

- Lyft data lacked information about the geographic infrastructure and the drivers involved in the accident
  - Searched for additional features in US Census and ACS to add demographic info

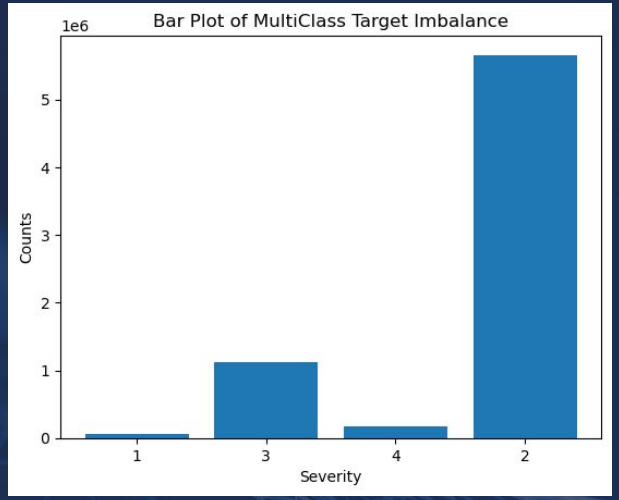
- 2020 US Census
  - Raw count or urban/rural population converted to Urban-Rural ratio column
- 2015-2019 American Community Survey Estimates
  - Gender ratio
  - Median Income
  - Percent of drivers 18-24
  - Percent of drivers 65+

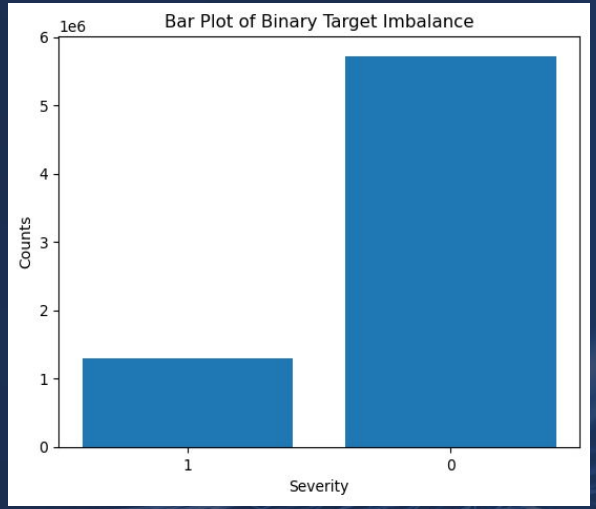
Final Feature Dataset

Weather	Infrastructure	Time	Population
• Temperature • Humidity • Pressure • Visibility • Wind Speed • Precipitation • Rain • Snow	• Interstate Indicator • Traffic Interference • Traffic Intersection • Destination	• Astronomical Twilight • Weekday • Season • Rush hour	• Gender Ratio • Percent Age 15-24 • Percent Age 65+ • Urban Ratio • Median Income

Data Summary

- Target Variable: Severity
  - Low severity (severity 1, 2)
  - High severity (severity 3, 4)

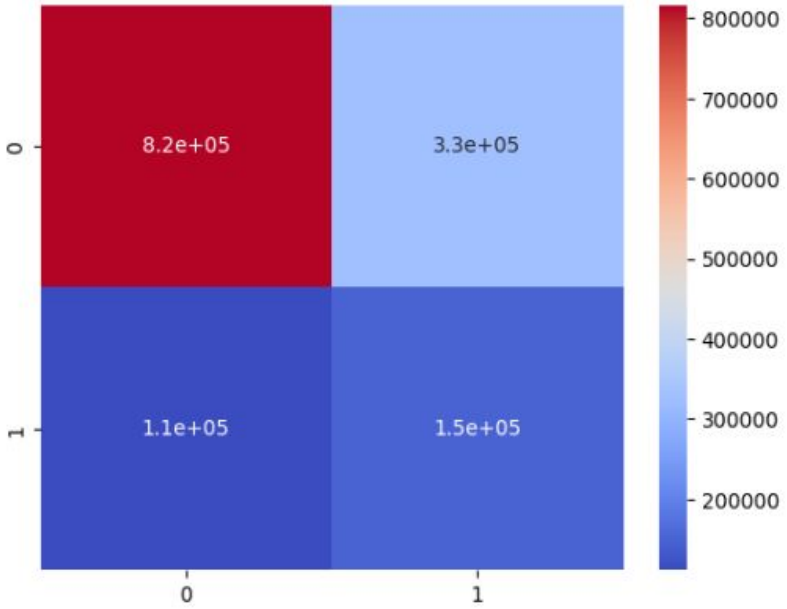




01 Logistic Regression

- Undersampling
- Best Model Parameters:
  - Maximum iteration: 10
  - Regularization parameter: 1
  - Elastic net parameter: 0
- Most Important Features:
  - Season
  - Traffic Intersection
  - Wind Speed
  - Temperature
  - Interstate Indicator

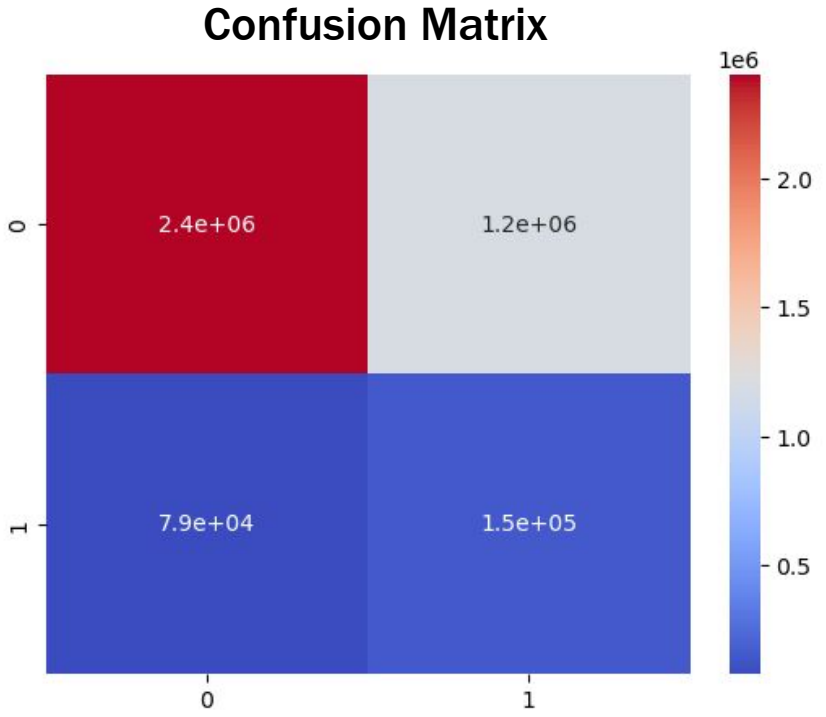
Confusion Matrix



Recall	Precision	F1	Accuracy
0.57	0.77	0.72	0.69

02 Random Forest

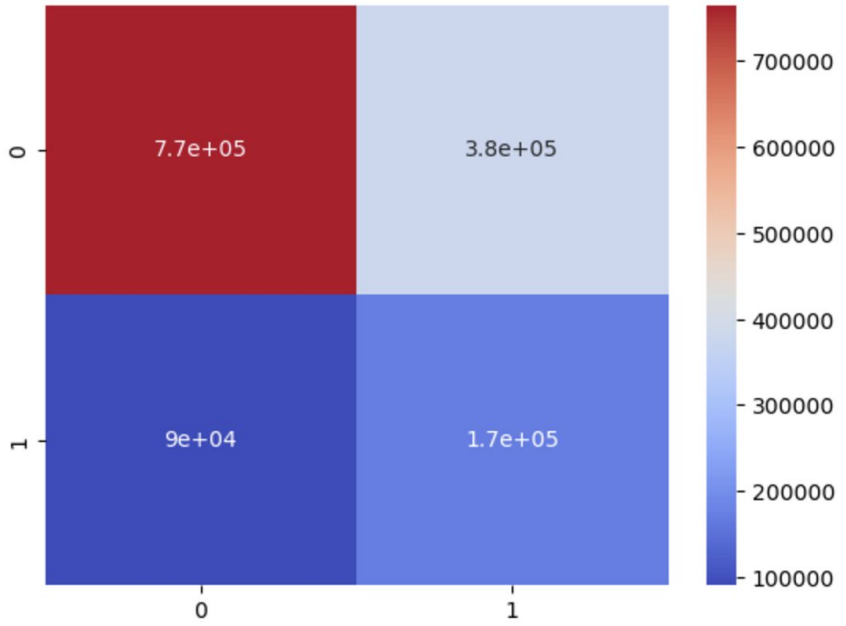
- Undersampling
- Best Model Parameters:
  - Maximum Depth: 6
  - Number of Trees: 10
- Most Important Features:
  - % population aged 65+
  - % population aged 15-24
  - Wind Speed
  - Traffic Intersection



Recall	Precision	F1	Accuracy
0.66	0.92	0.75	0.67

03 Gradient Boosted Trees

- Best Model Parameters:
  - Max Depth - 7
  - Max Iter - 3
- Most Important Features:
  - % population aged 15-24
  - Temperature
  - Wind Speed
  - Pressure
  - Traffic Intersection



Recall	Precision	F1	Accuracy
0.65	0.78	0.71	0.68

Champion Model

Random Forest

- Slightly better recall than gradient boosting method
- Likely less time to train than gradient boosting
  - If to be implemented in real time for ride share is important
- Top important features were similar to other models
- Models were all trained with different train/test splits and with different sampling methods to deal with class imbalance
- Training models on the same data would yield more confidence in results to pick the best performing model

Sensitivity Analysis (Random Forest)

- Sensitivity Analysis by changing parameters max depth and number of trees shows some sensitivity to parameter tuning
- Sensitivity expected for RF
  - changes in a single tree can be highly impactful
- Tested
  - max depth 3-6
  - number of trees 3-6
- If we needed a model to be continuously retrained on real-time data, this sensitivity might cause issues

maxDepth	numTrees	recall	f1	accuracy
3.0	3.0	0.690175	0.713544	0.614608
3.0	4.0	0.579387	0.780308	0.703519
3.0	5.0	0.608439	0.773524	0.693921
3.0	6.0	0.559532	0.805949	0.740353
4.0	3.0	0.768092	0.633417	0.520430
4.0	4.0	0.623520	0.762931	0.679304
4.0	5.0	0.465865	0.846424	0.802454
4.0	6.0	0.558597	0.805738	0.740055
5.0	3.0	0.612189	0.741588	0.650556
5.0	4.0	0.601329	0.780634	0.703869
5.0	5.0	0.485562	0.840872	0.793534
5.0	6.0	0.602058	0.788128	0.714422
6.0	3.0	0.529558	0.822159	0.764580
6.0	4.0	0.532845	0.821236	0.763161
6.0	5.0	0.628967	0.767320	0.685305
6.0	6.0	0.567323	0.808788	0.744403

Limitations & Assumptions

- Limited data documentation
- Spark does not have a native function for XGBoost
- Did not have access to demographics on the accident-level - used zipcode-level as proxy
- Our target variable is a proxy for time which may be an incorrect assumption
- Season varies by month depending on where in the US
- Models were not trained on the same train/test split or with the same method for dealing with data imbalance

Conclusions

- Recommendations for Lyft to create a model to best approximate traffic impact of accidents:
  - Reconsider the method of severity classification on the basis of the severe class imbalance
  - Evaluate model performance based on false positives to ensure that the drivers are using the best route for the rides
  - Further evaluation could be done with A/B testing to see if new models increase total ride counts to increase revenue for Lyft and drivers
  - Evaluation of model runtime as the recalculation of route info should be calculated quickly