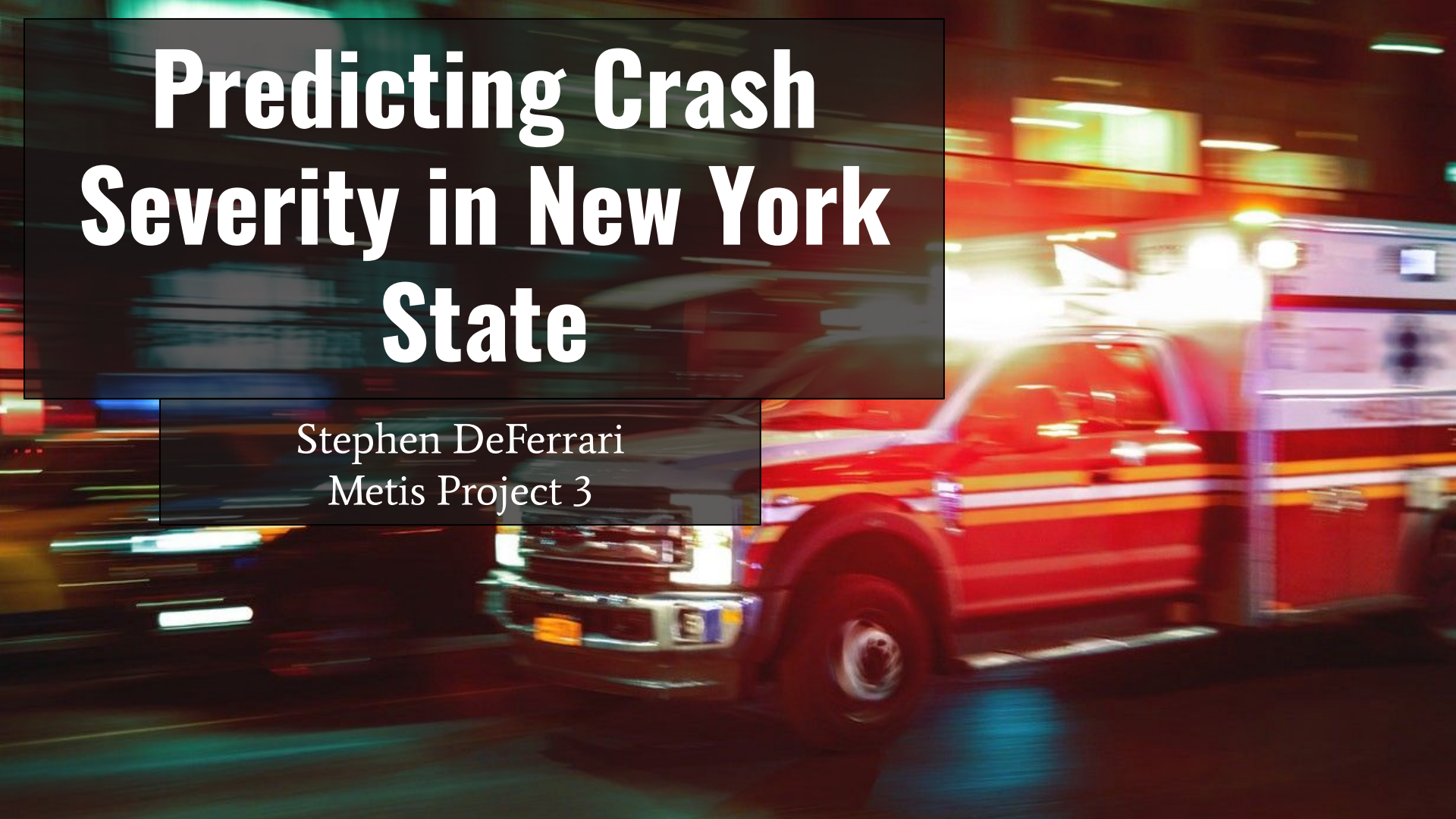


Predicting Crash Severity in New York State

Stephen DeFerrari
Metis Project 3

Motivation

Motor vehicle traffic crashes are the leading cause of injury related death for New York State residents.

During the period from 2012-2014 there was an annual average of 1,098 deaths, 12,093 hospitalizations and 136,913 emergency department visits due to traffic incidents

Objectives

Can we use a classifier to help EMS services predict the severity of an accident during a specific time of day/day of week with specific weather conditions to help departments better coordinate ambulance services?

- What steps can governments take to reduce accident severity?
- What messages should they be sending to people to increase safety?

Methodology of Project

Clean and combine Data → Feature Engineering → Modeling → Tableau Dashboard

Data:

- NY DMV accident records from 2014 - 2016
- NY Census Information
- NY Impaired Driving Rates

Models:

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- XG Boosted Trees

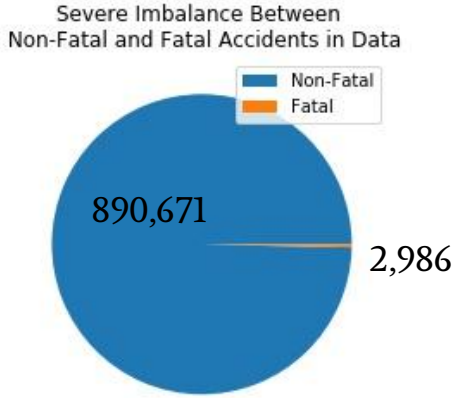
Data Cleaning and Engineering

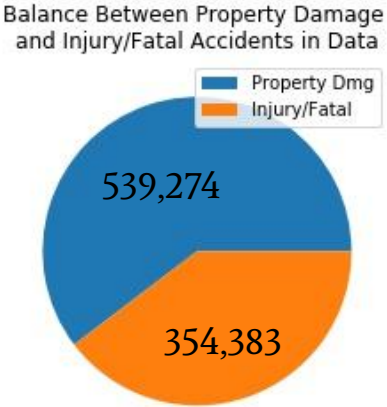
DMV Data contained records of close to 900,000 accidents

Final Features included:

- Crash severity/description
- Road/Weather conditions
- What was struck (car, tree, etc.)
- County Name
  - Added pop density, DUI rates

With Counties - 90 features
Without Counties - 27 features





* An “All Property Dmg” prediction results in 60% accuracy

Modeling

Interpretability vs Black Box

Interpretation Models	Logistic Reg P: 69 R: 31	Decision Tree P: 57 R: 51	Logistic Reg with .35 threshold P: 40 R: 90
Black Box Models	KNN P: 68 R: 42	Random Forest P: 75 R: 46	Random Forest with .35 threshold P: 58 R: 72

Modeling Takeaways:

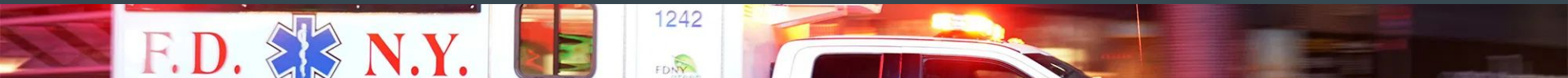
Recall chosen as key attribute

ROC curves were pretty similar across the board

- High diminishing returns

XG Boosting did little to help

.35 threshold applied to counter slight imbalance in data



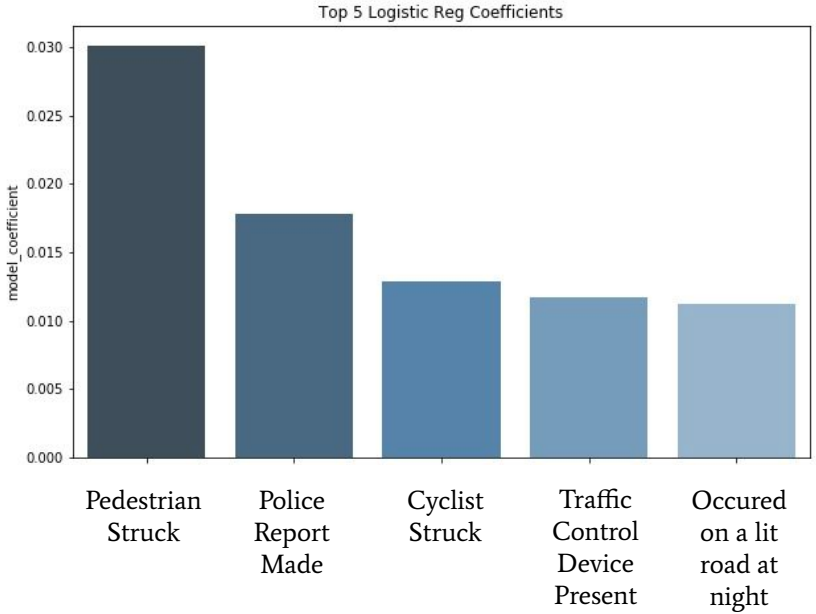
Logistic Regression Takeaways

Pedestrian and cyclist involved accidents are more severe

- Local governments need to work more on creating bike lanes/pedestrian safe areas

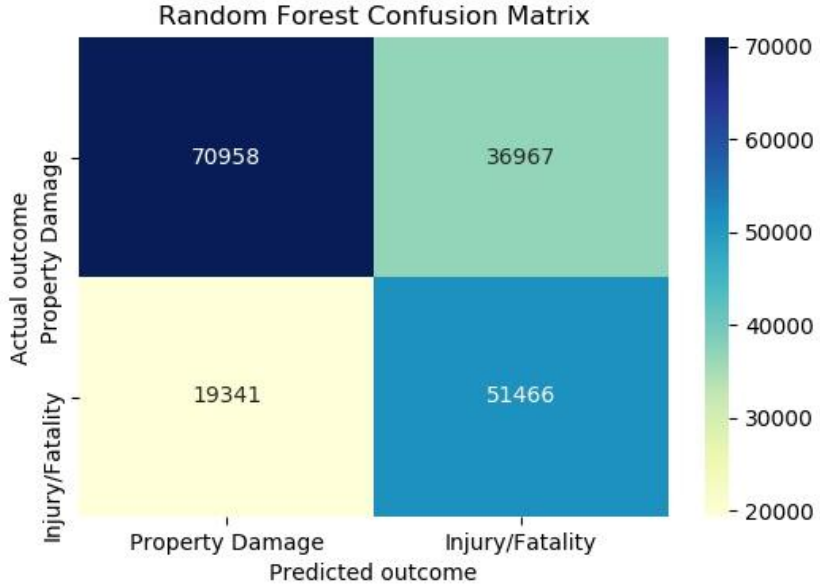
Lit roads are more dangerous than unlit roads at night

- Road lighting needs to improve



Random Forest with Threshold

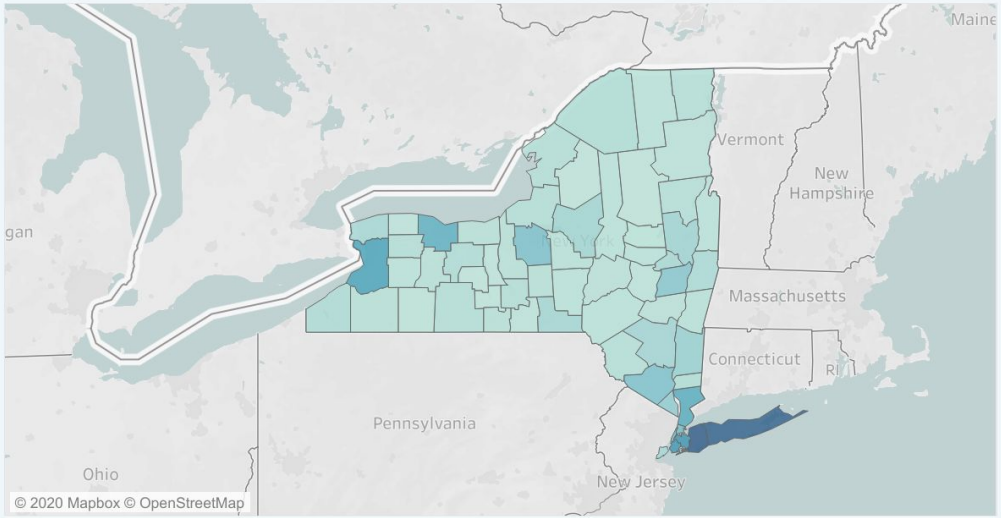
- Random Forest with .35 threshold applied resulted in best overall balance in scores.
- While recall was my target, I wanted to ensure that my precision was still somewhat balanced to avoid resource drain.



Accuracy: 68 **Recall**: 72
Precision: 58 **F1 Score:** 64

Tableau Dashboard

New York State Traffic Accidents (2014 - 2016)

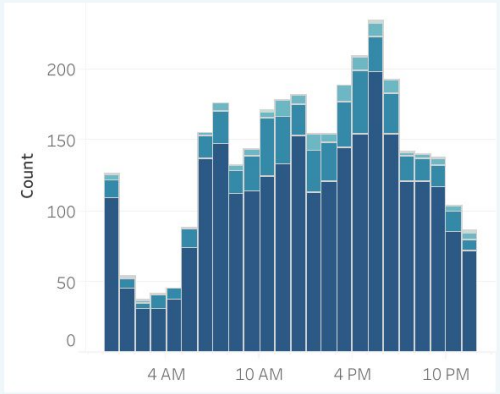


Count of motor-vehicle-crashes-case-inf..
630 98,983

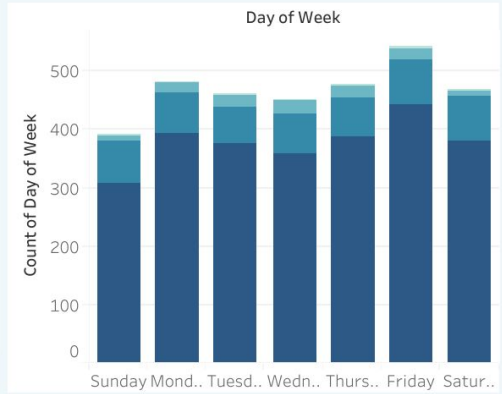
Accident Causes

Event Descriptor (group)	
Other Motor Vehicle, Colli..	1,186
Fixed Object Collision	976
Deer	876
Animal, Collision With	87
Other Object (Not Fixed)*,..	36
Pedestrian, Collision With	25
Overturned, Non-Collision	24
Other*, Non-Collision	19
Fire/Explosion, Non-Collis..	12

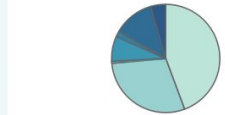
Time



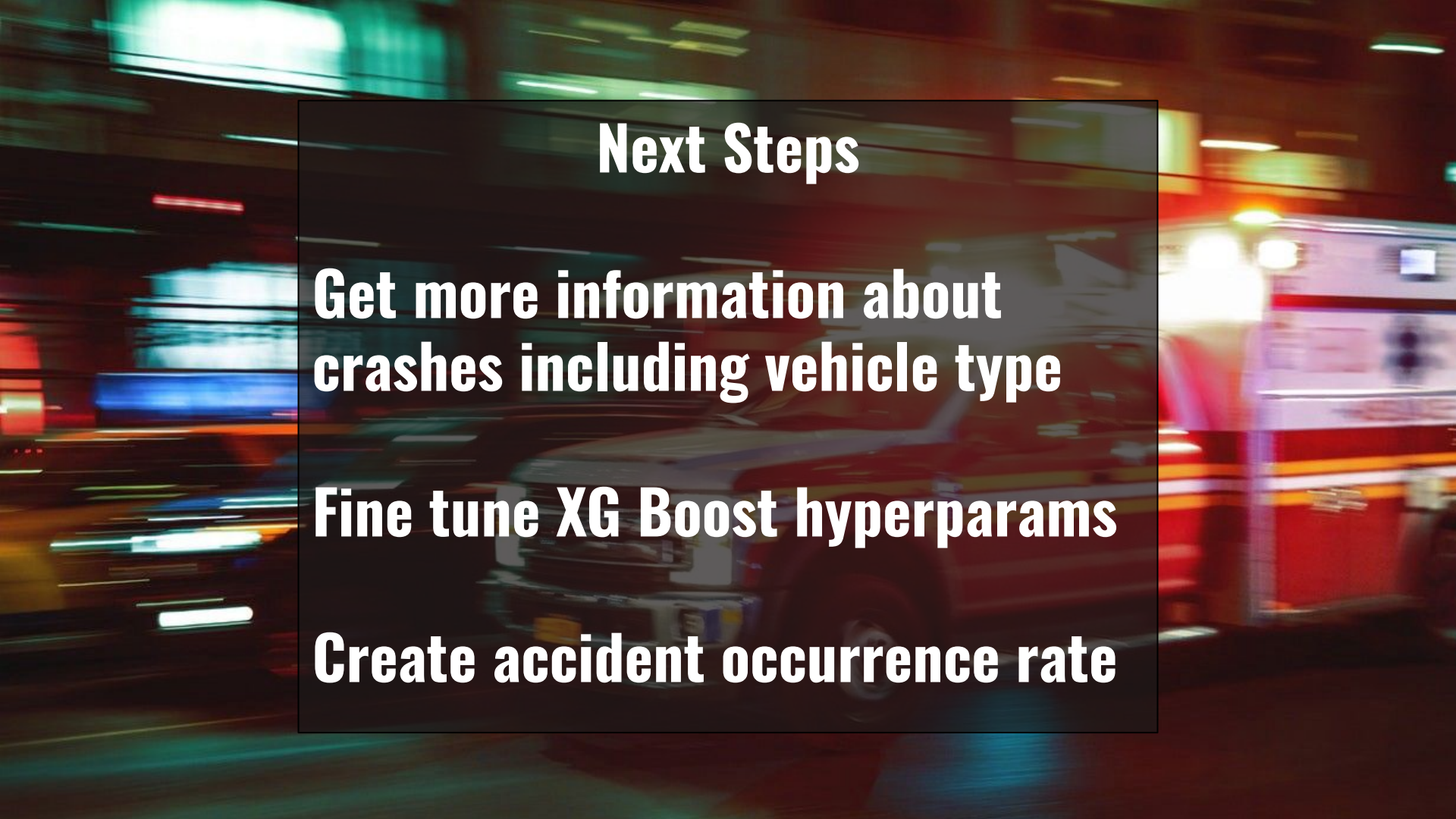
Day of Week



Weather



Road Type



Next Steps

Get more information about crashes including vehicle type

Fine tune XG Boost hyperparams

Create accident occurrence rate


Thank you,
Stay Safe
And
Watch for Deer