Fatal Encounters with Police

Recommendations for a National Policy on Policing

Alexander Vizzi
Data Science Intensive Capstone Project, cohort April 2020

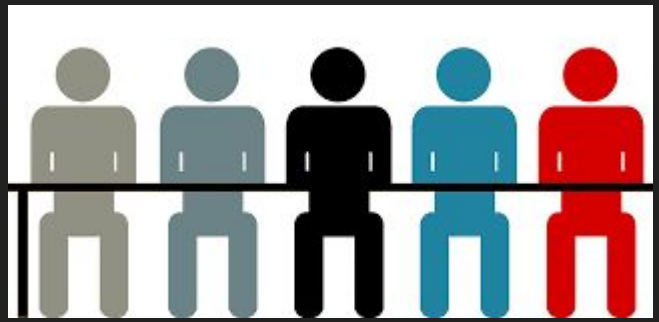
The Problem

- ~1000 people shot by police / year
- Black and Hispanic deaths disproportionate
- No national standards
- Public outcry
- What factors predict that a victim will be Black or Hispanic? Can policymakers target these factors to reduce disproportionate Black/Hispanic deaths?

Photo by Steve Apps, Wisconsin State Journal

Who Wants to Know?

President-elect Joe Biden



National Police Oversight Commission

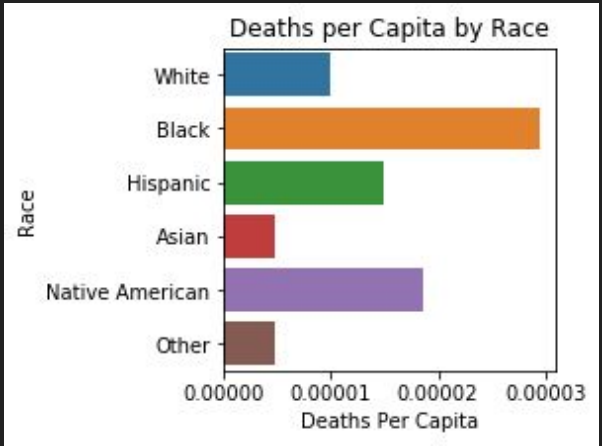
Advocacy groups

Police chiefs

The Data

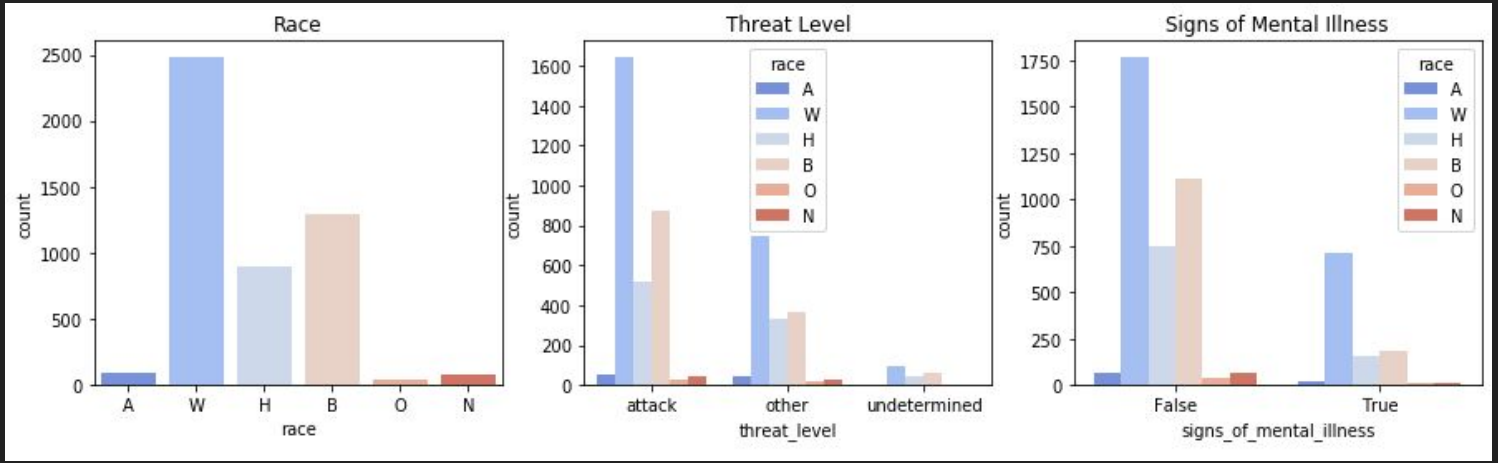
- Data collection not standardized among states/cities
- Hasn’t been rigorously collected before 2014
- I sourced from Washington Post Fatal Force Police Shootings Database

Black/Hispanic deaths not proportional to population.



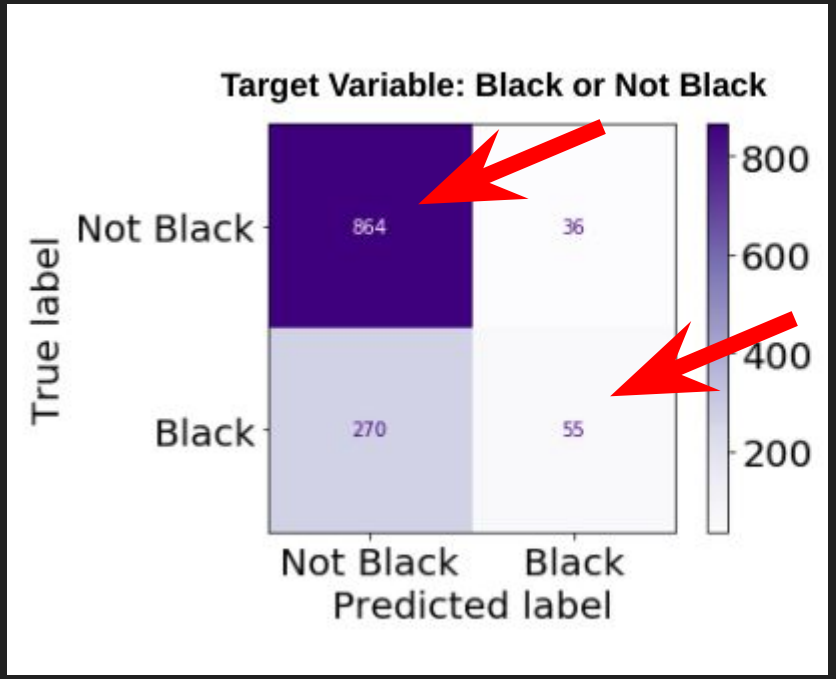
Data Exploration

Proportions of races don't seem to change when considering recorded details of the shooting.

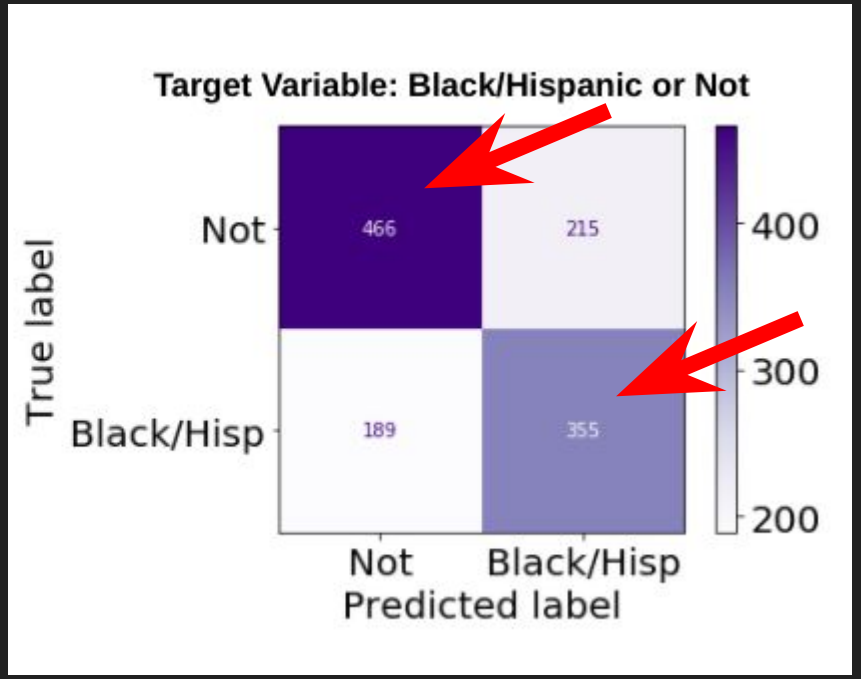


Which Target Variable to Choose?

Models were better at predicting Black & Hispanic deaths than Black deaths alone.



VS.



Modeling:

Classification algorithms used

Support Vector Classifier

Logistic Regression

Gradient Boosting Classifier

Random Forest Classifier

Model Tuning

Different techniques worked with each model when remedying imbalanced classes:

Synthetic Minority Oversampling Technique (using imblearn)

Class_weight= ‘balanced’ argument in Scikit-Learn

Used different techniques to regularize models:

L1 and L2 regularization using ‘penalty’ argument in Scikit-Learn

Manipulating ‘C’ parameter

Model Observations

Performances converged for all models, with SVC edging out the others.

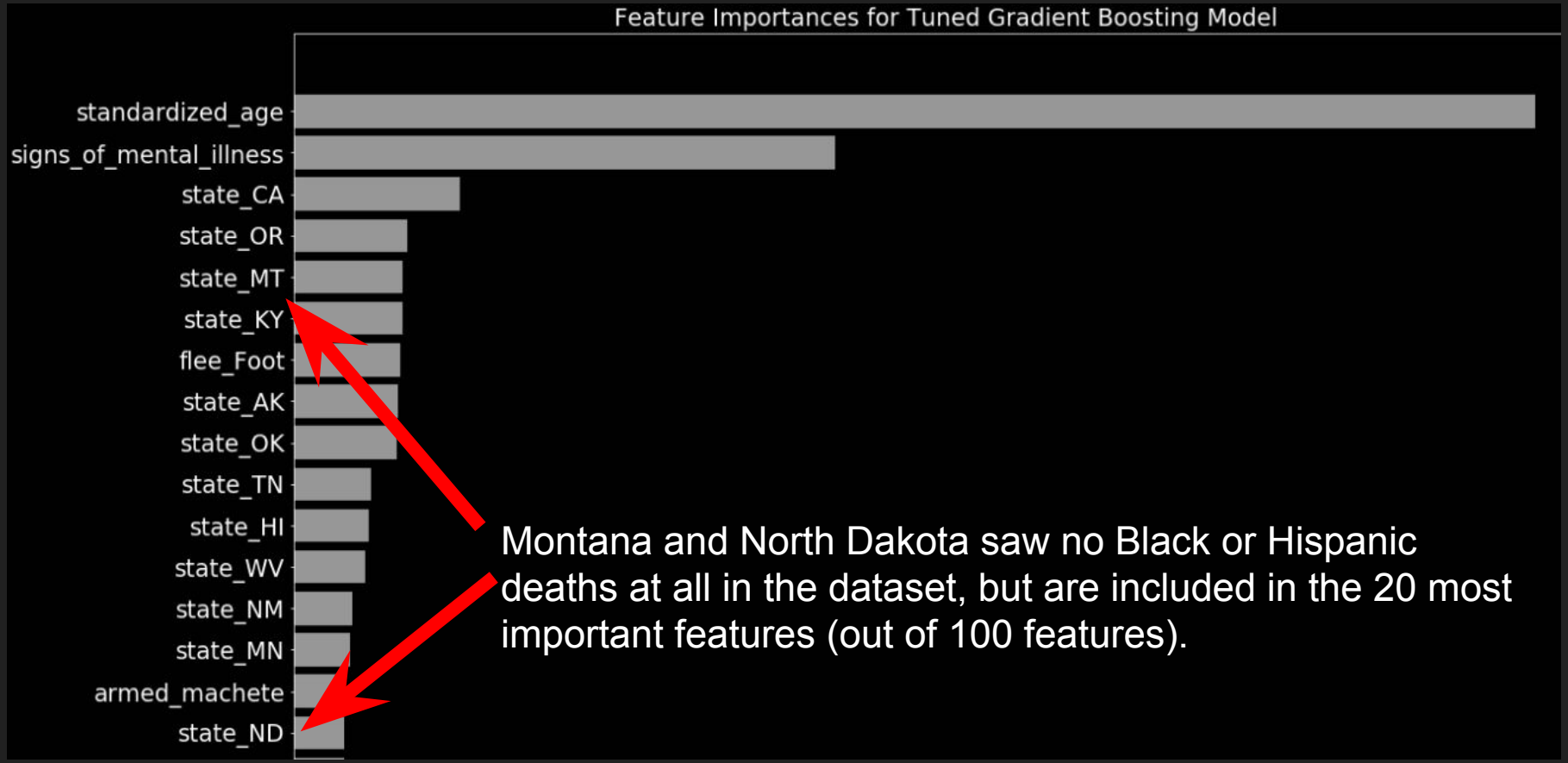
Model	Out of the Box F1	First Tuned F1	Regularized F1	Oversampled F1	Oversampled and Regularized F1
gbc	0.598901	0.610449	NaN	0.648026	NaN
lr	0.637523	0.634234	NaN	0.645161	0.623838
svc	0.637343	0.652927	0.650502	0.633452	0.640960
rf	0.568684	0.634310	NaN	0.633081	NaN

The difference between the best classifier's highest performance and the worst classifier's highest was 0.02.

More model observations

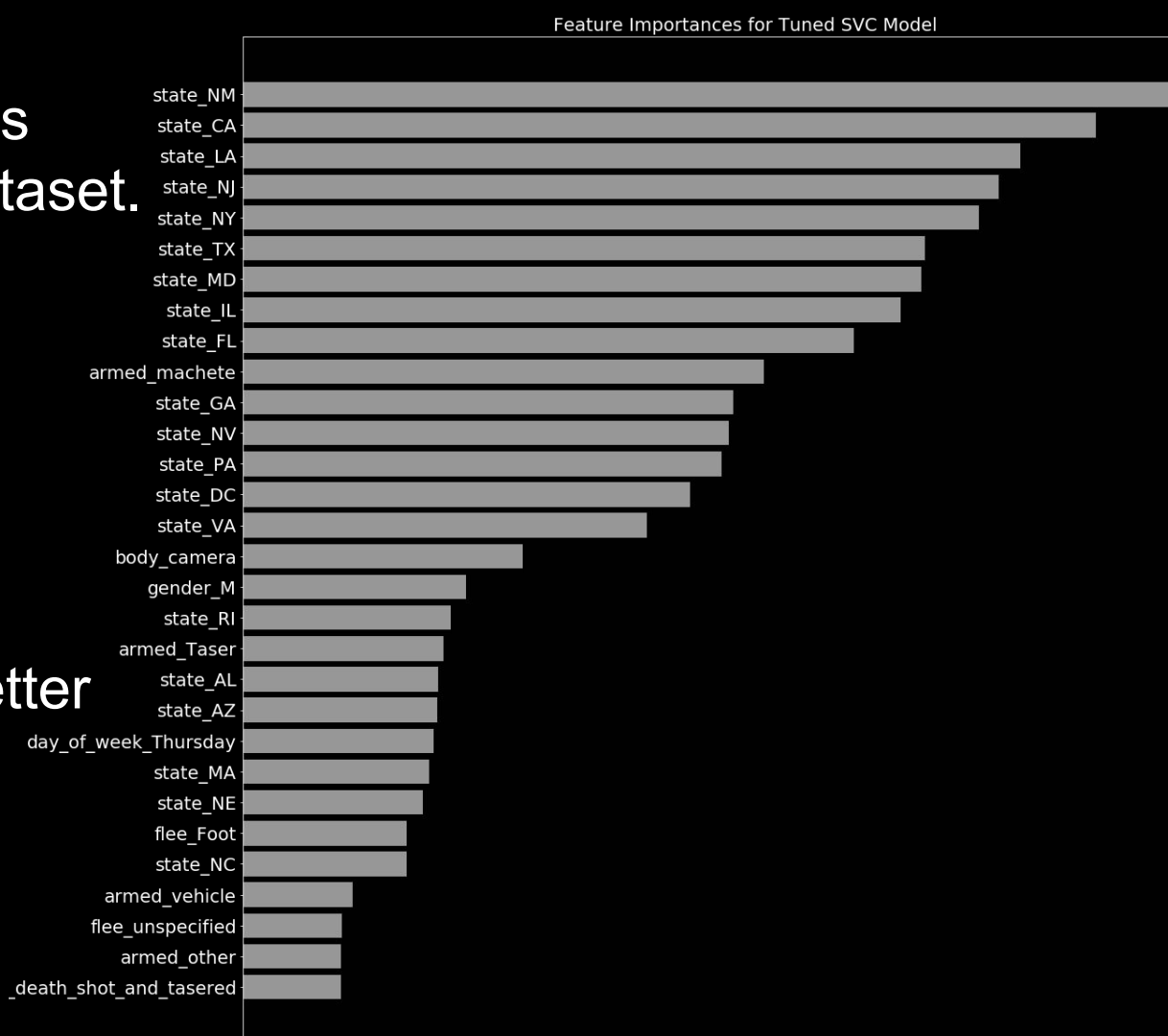
Despite similar performances,
tree based models emphasized certain features
while linear models emphasized others.

But some features of the tree based models were
misleading and illogical.


Feature Importances for Tuned Gradient Boosting Model
standardized_age
signs_of_mental_illness
state_CA
state_OR
state_MT
state_KY
flee_Foot
state_AK
state_OK
state_TN
state_HI
state_WV
state_NM
state_MN
armed_machete
state_ND
Montana and North Dakota saw no Black or Hispanic deaths at all in the dataset, but are included in the 20 most important features (out of 100 features).

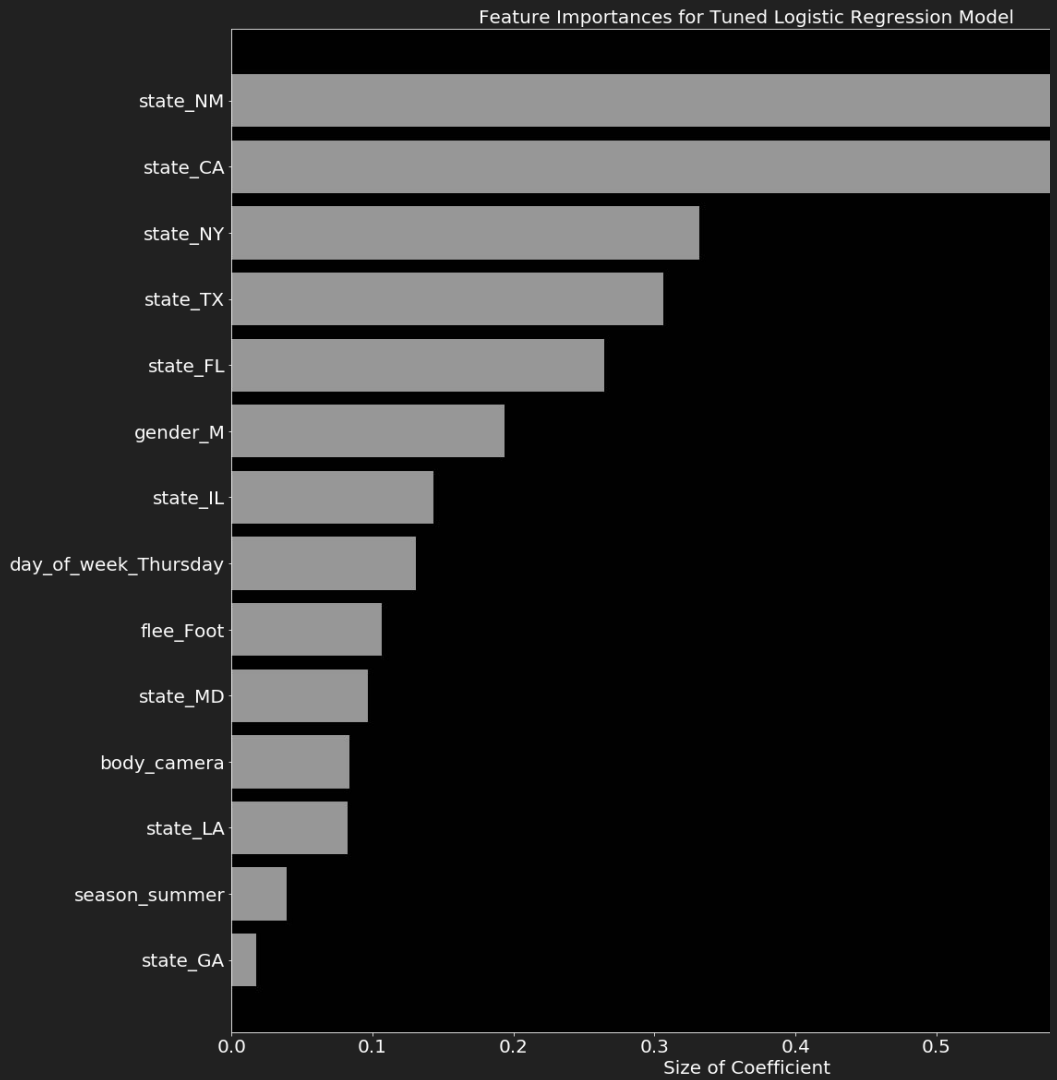
SVC's important features made sense with the dataset.

Despite inconsistent performances, the logistic regression models highlighted similar features to the better performing SVC models



Takeaways from important feature plots:

- Circumstances of shooting not as influential
- State differences are key



Policy Recommendations

- Demographics matter: diverse states appear often
- Other factors matter: some less diverse states appear
- Mental health training, policies that teach tactics for unarmed/fleeing civilians may not matter
- Body cameras may not matter (for this goal): these predicted Black/Hispanic deaths
- Should standardize data collection, include regional specifics

Further Research

- Include demographic features
- Include state-specific and county-specific features
- Include differences in state restrictions on use of force, racial diversity of police force, changes to policing that have been attempted already