business context

stakeholder:

Chicago Department of Transportation

goal:

reduce traffic crashes by determining their main causes

project requirements:

keep project methods *interpretable*

original dataset

- from the City of Chicago website

- detailed records dating **back to 2015**

- 3.8 million rows & 146 columns

- **after cleaning**: 885 thousand rows & 15 columns

- target **imbalance** addressed with over/under-sampling

Target Value Counts
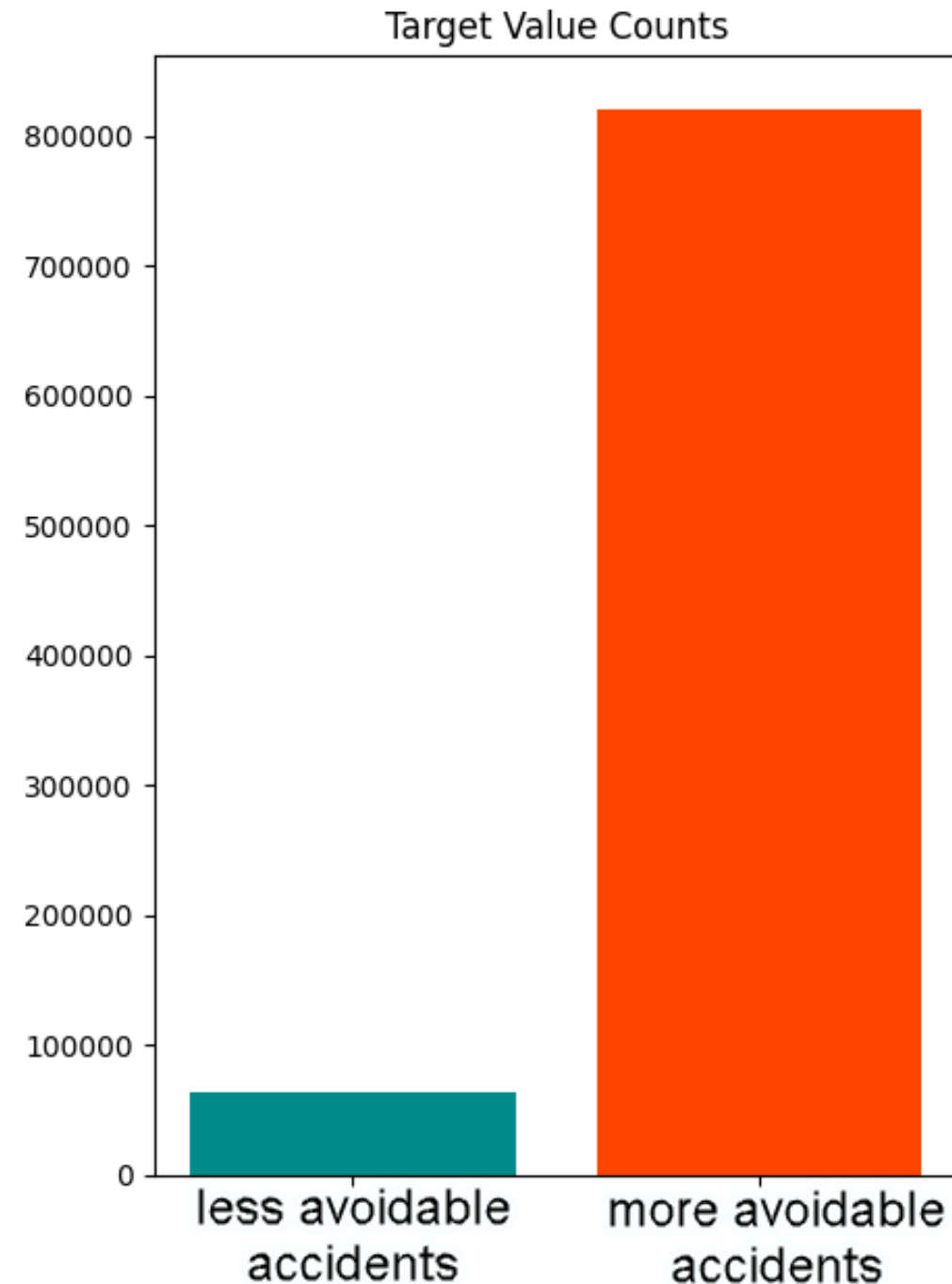
Methodology

narrow columns down

try different modeling techniques

optimize the best model

reducing columns

146 columns = too many for an **interpretable** model

unnecessary columns reduced via:

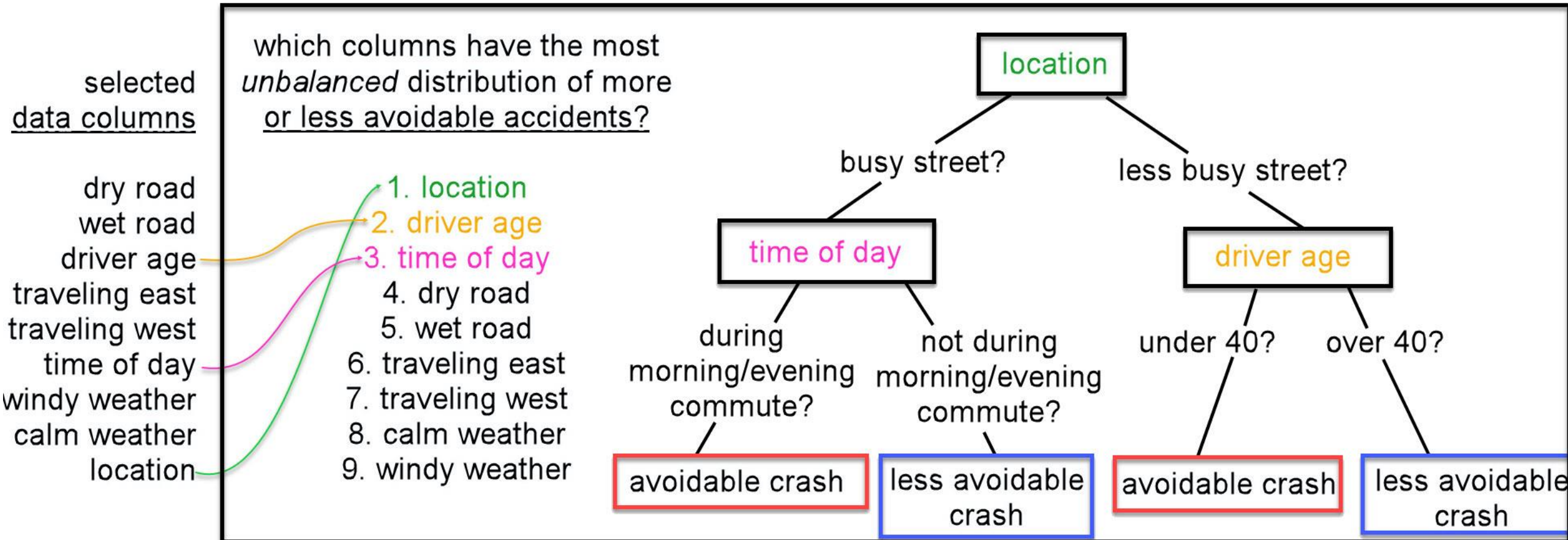- domain knowledge
- inherent redundancy

further columns reduced by:

- models that tell **how much data is explained** by each column
- **statistical** models

simplified decision tree example

decision tree model

selected data columns

which columns have the most *unbalanced* distribution of more or less avoidable accidents?

dry road
wet road
driver age
traveling east
traveling west
time of day
windy weather
calm weather
location

1. location
2. driver age
3. time of day
4. dry road
5. wet road
6. traveling east
7. traveling west
8. calm weather
9. windy weather

location

busy street?   less busy street?

time of day   driver age

during morning/evening commute?   not during morning/evening commute?   under 40?   over 40?

avoidable crash   less avoidable crash   avoidable crash   less avoidable crash

Model results:
most important columns

feature importances in final model

Feature	
latitude	
longitude	
driver age	
hour of day	
dry road conditions	
wet road conditions	
day of week: Wednesday	
traveling north	
day of week: Sunday	
traveling east	
traveling west	
lighting condition: daylight	
maneuver: parked	
clear weather	
maneuver: skidding/control loss	
rainy weather	
maneuver: turning right	
snowy/slush road conditions	
maneuver: turning left	
maneuver: straight ahead	
cloudy weather	
maneuver: stopped in traffic	
maneuver: changing lanes	
maneuver: passing/overtaking	
icy road conditions	

importance ratio of feature (they add up to 1.0)
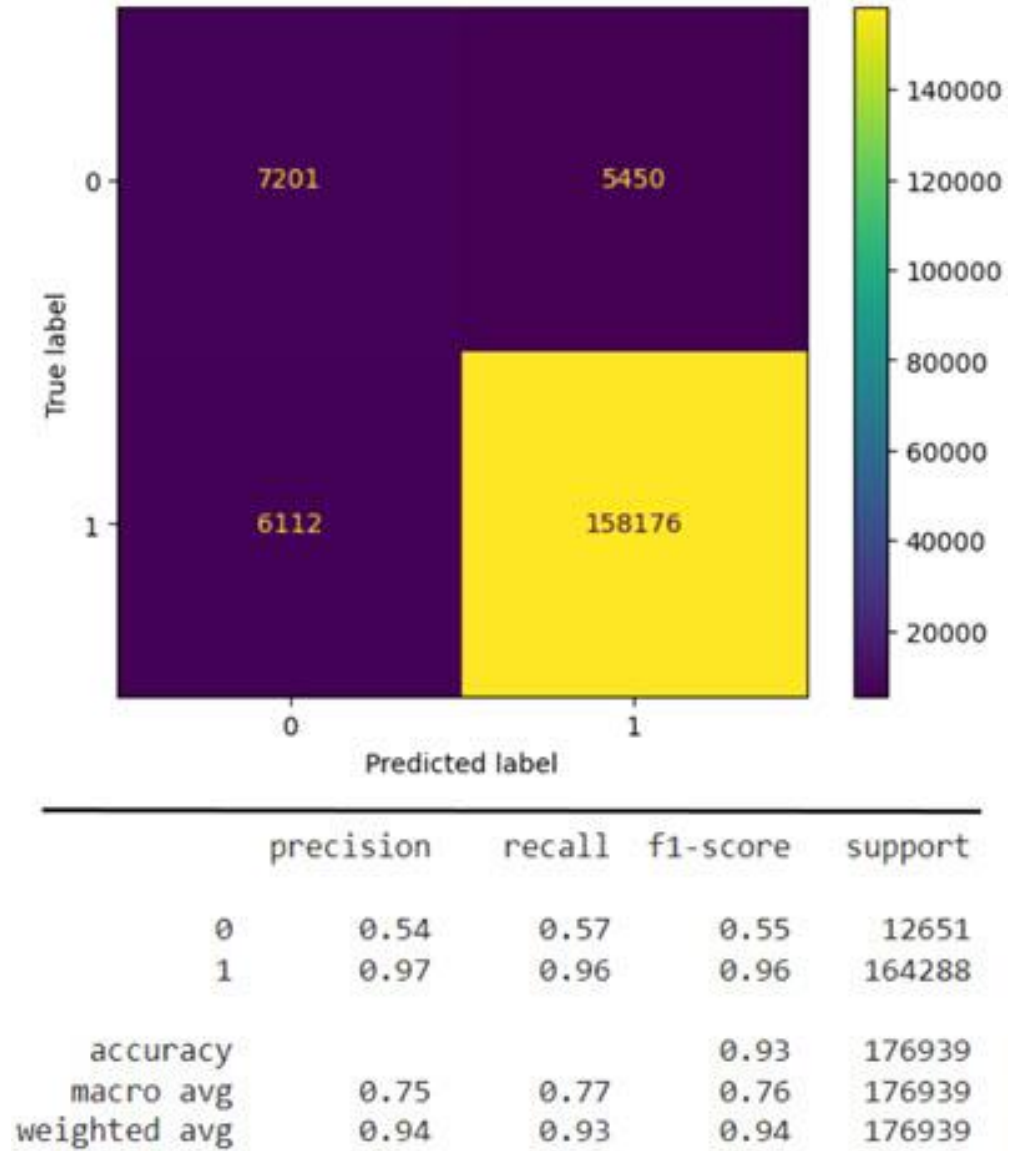
best performing model results

- 93% correct predictions **overall**

- **less avoidable accidents** ('0' in the diagram) only correctly predicted 54%

- considering class imbalance issues, this is not a great score, but it is **acceptable**

- could be improved with a model that's **harder to interpret**

	precision	recall	f1-score	support
0	0.54	0.57	0.55	12651
1	0.97	0.96	0.96	164288
accuracy			0.93	176939
macro avg	0.75	0.77	0.76	176939
weighted avg	0.94	0.93	0.94	176939

mapping more avoidable accidents

- concentrated on **main roads**

- especially dense in **downtown**

▲ more avoidable crash
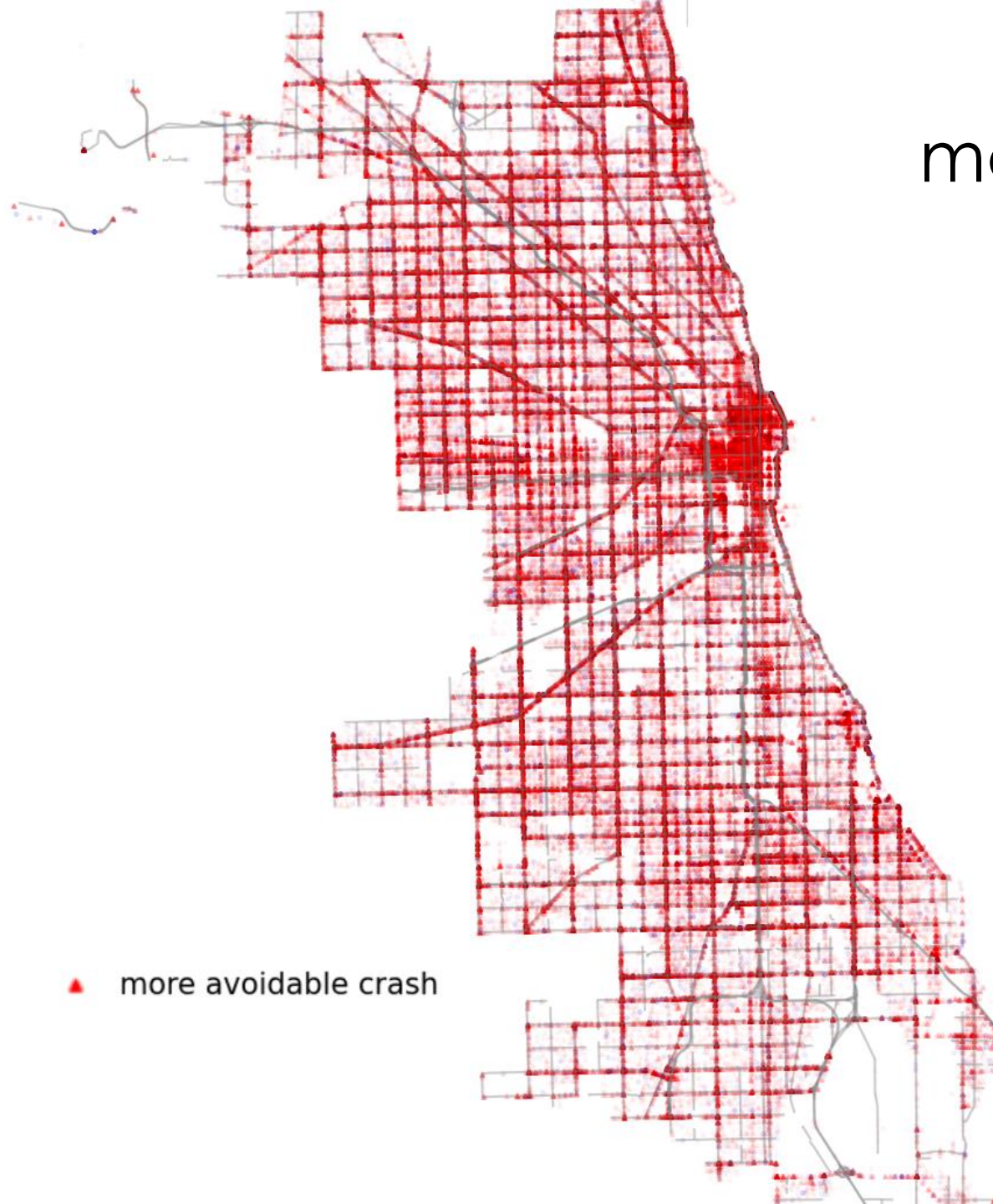
targeting drivers aged 23-40

- concentrated around drivers **aged 23-40**

- even younger drivers may not be driving as much or as far on main roads

Distribution of Avoidable Accidents by Age
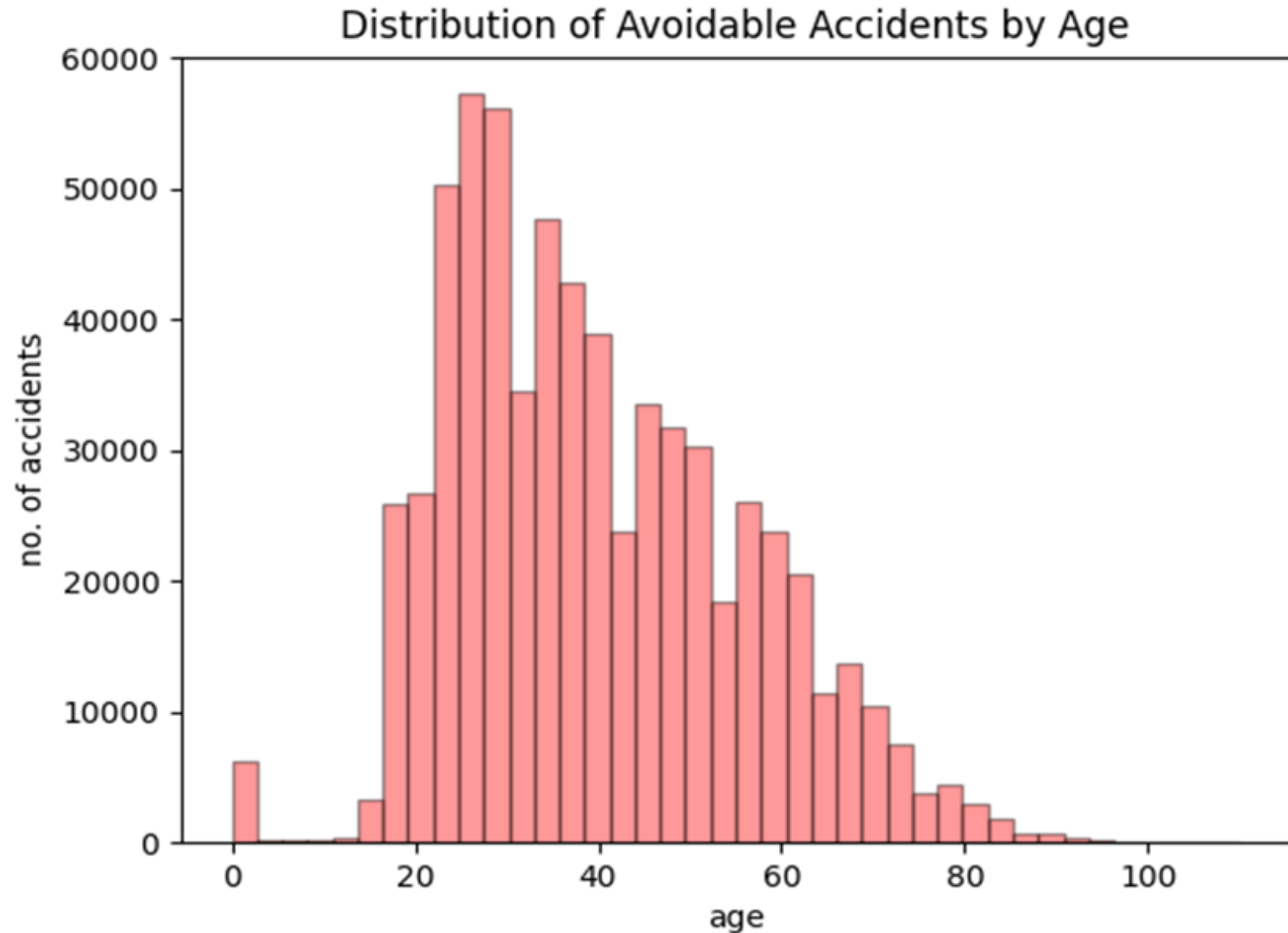
commutes see the most crashes

- the **morning commute** shows a more isolated spike

- **evening commute** stretches from 3-7pm

Distribution of Avoidable Accidents by Hour-of-Day
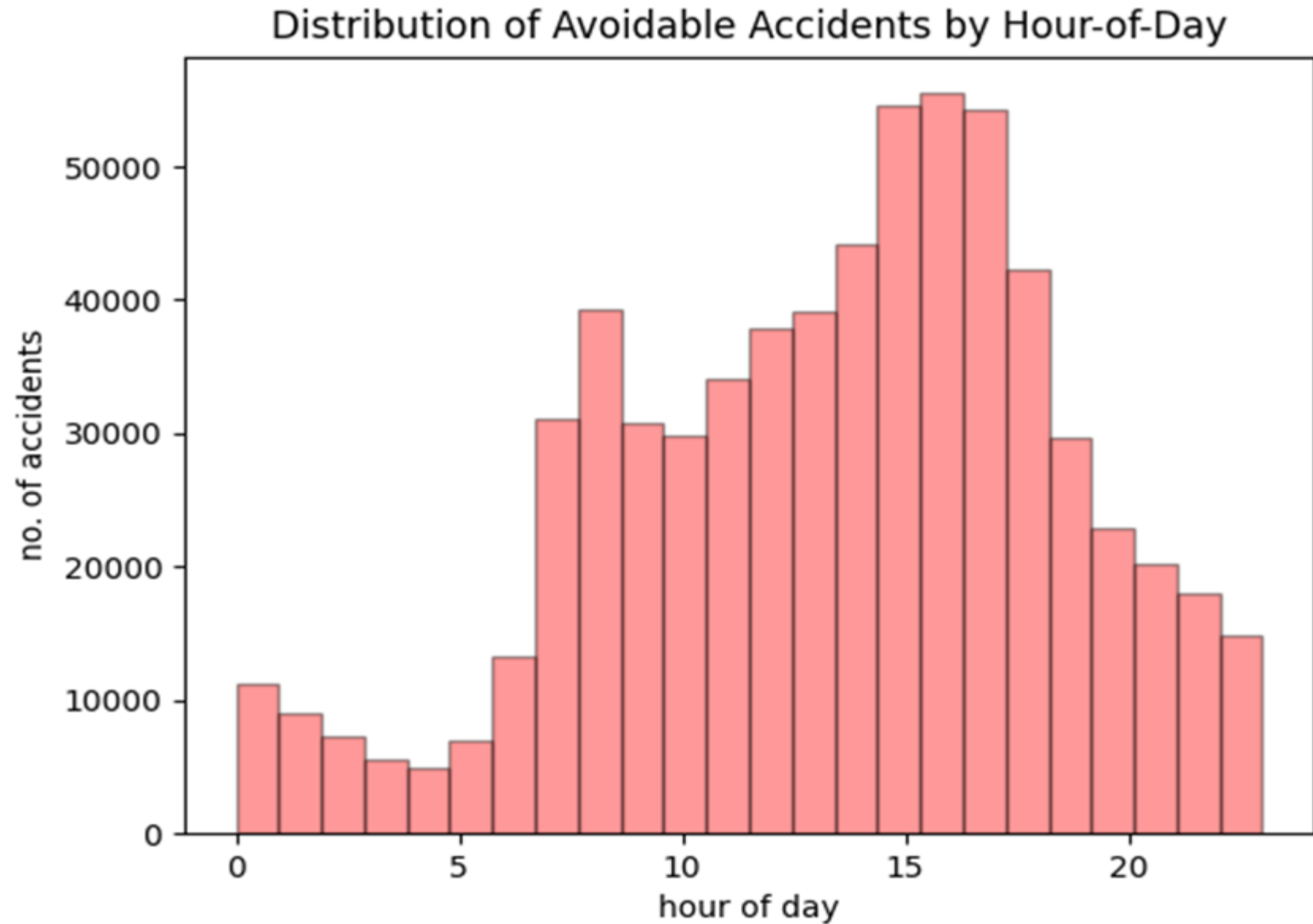
recommendations

Ads focused on safety in **heavy traffic** for **drivers between 23-40**

Road sign/traffic signal studies in Chicago's **middle/downtown**

Safety PSAs over the radio **during commute times**

Thank You

joel.mott8@gmail.com

Flatiron School