COVID-19 and Safety On Wheels in Chicago

Let crashes show you how to learn lessons

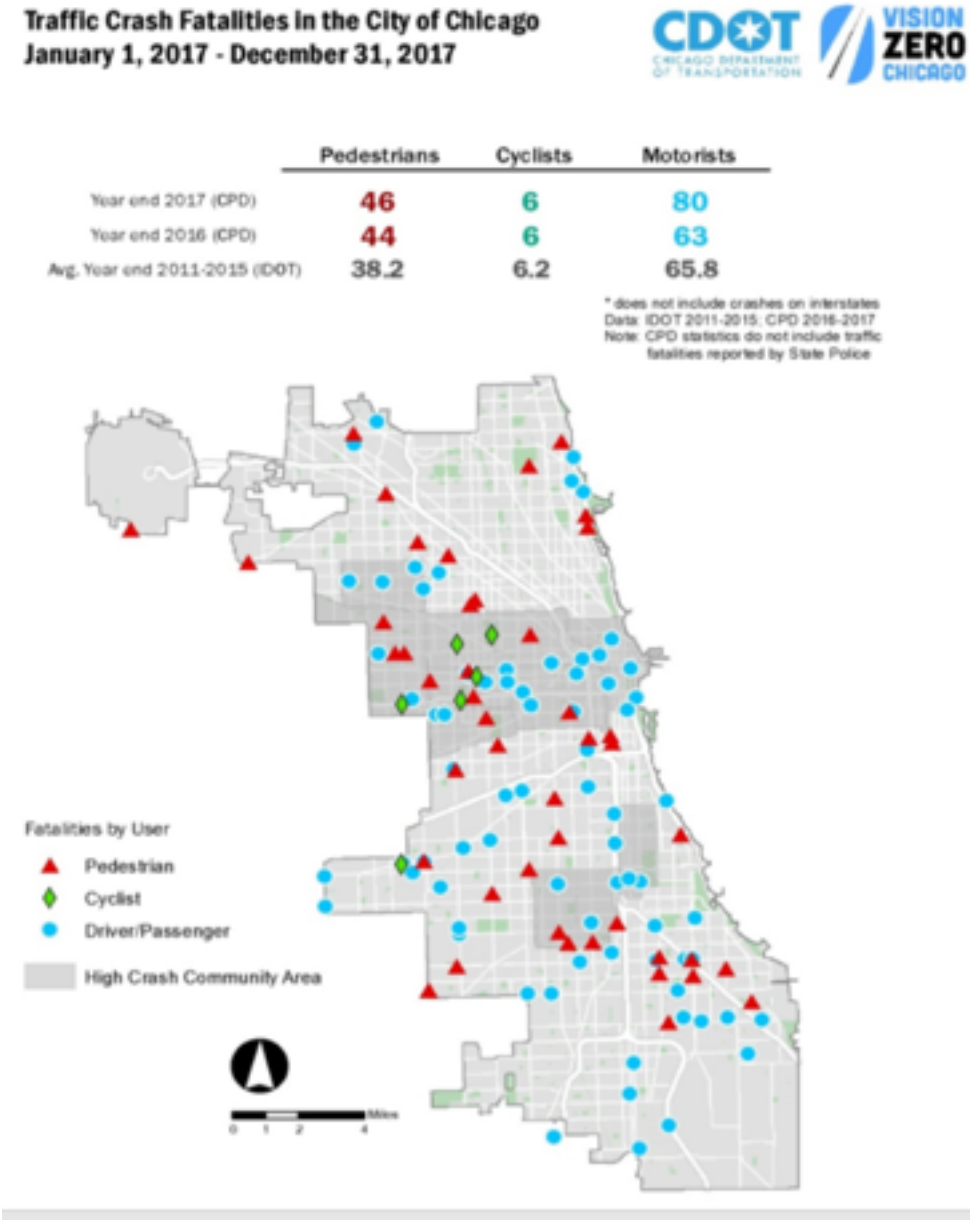
Chun-Min (Mindy) Jen

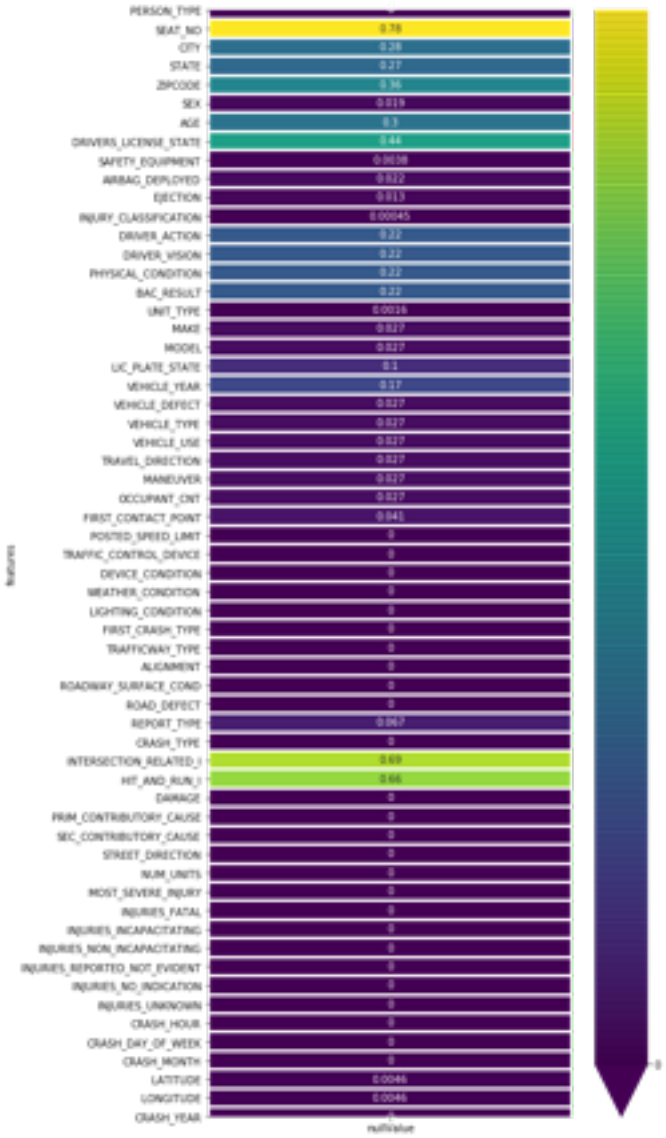
August 21st, 2020

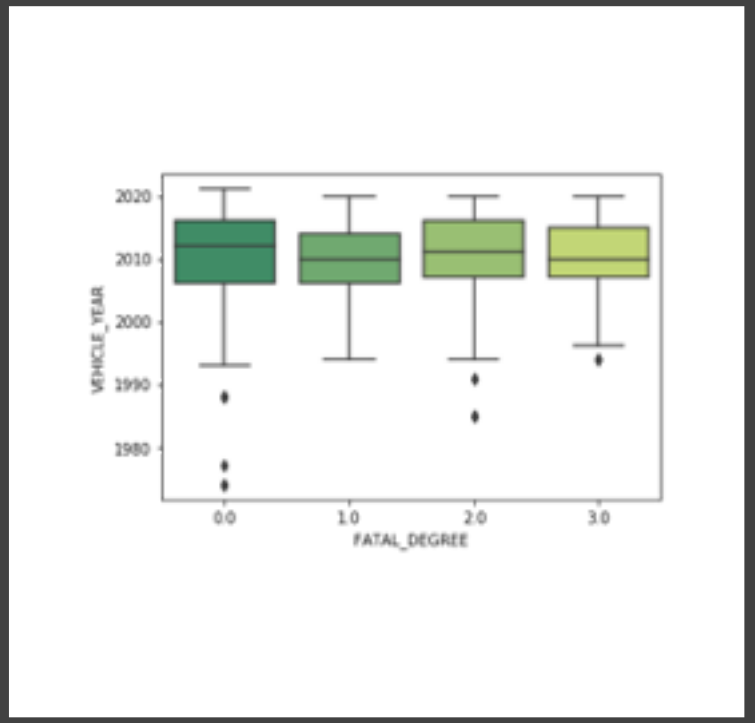
CHICAGO DATA PORTAL

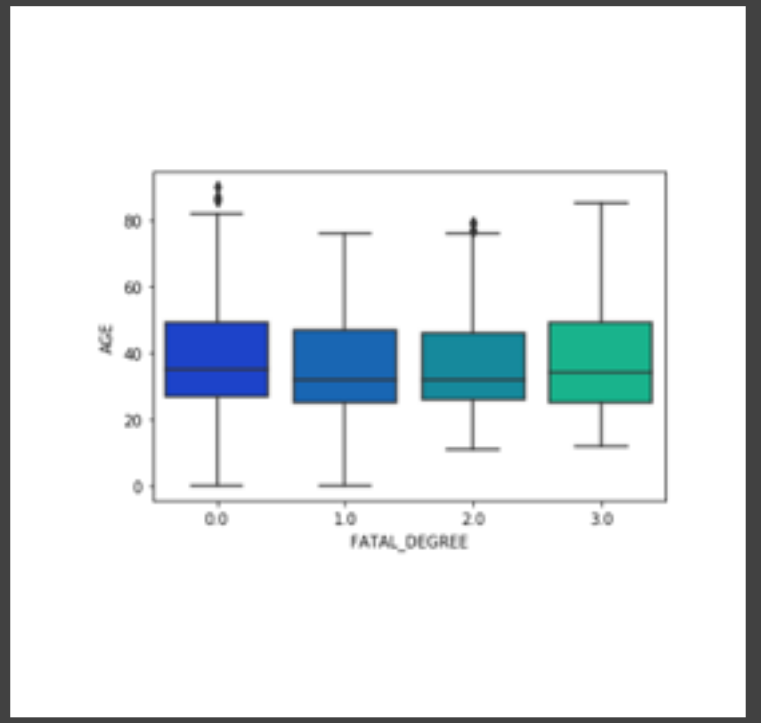
Outline

- **Data sources**
- **Data selections**
- **Data studies : EDA**
- **Data simplifications**









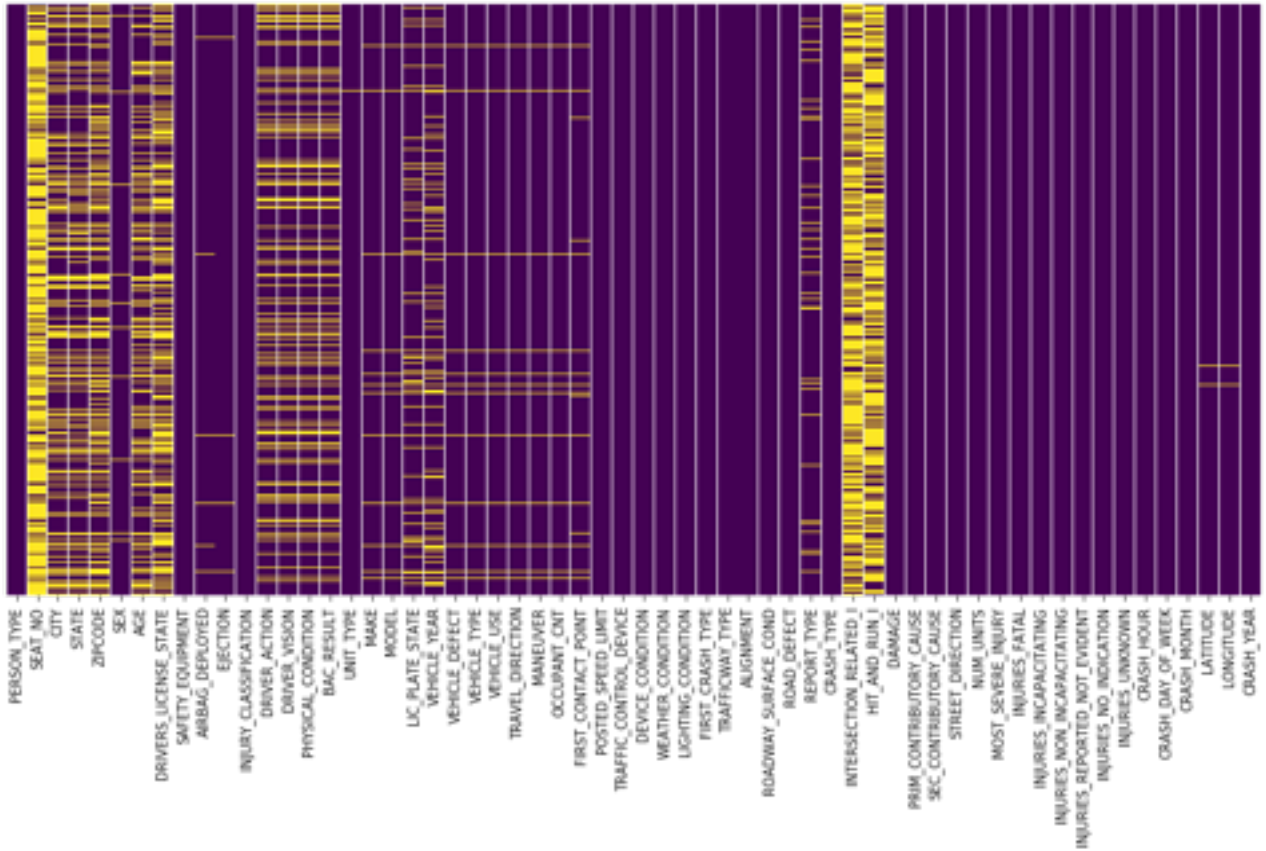
Data Selection - COVID-19 era

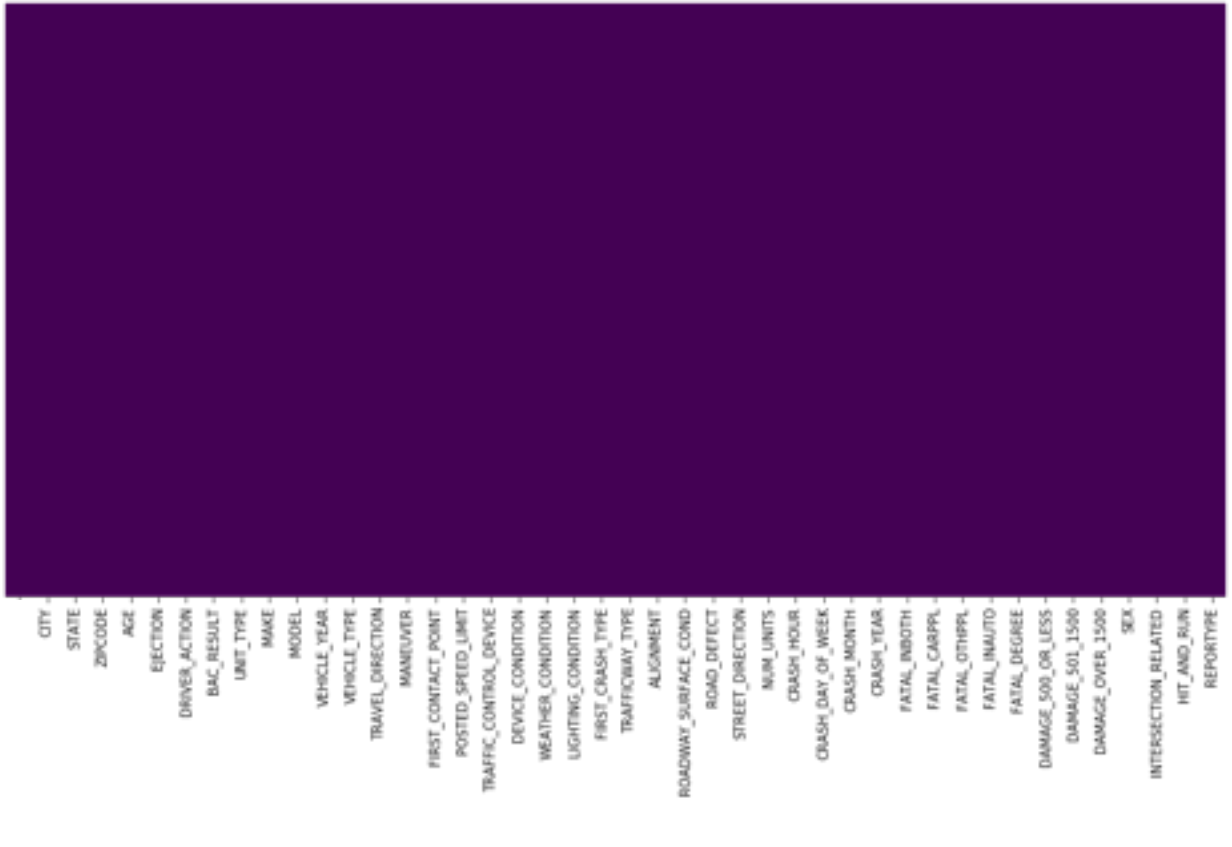
1. **Crashes**
2. **People**
3. **Vehicles**

Rows = 140'902 (April 1st 2020 – August 9th 2020)
columns = 73

Data Selections

73 features → **43 features**

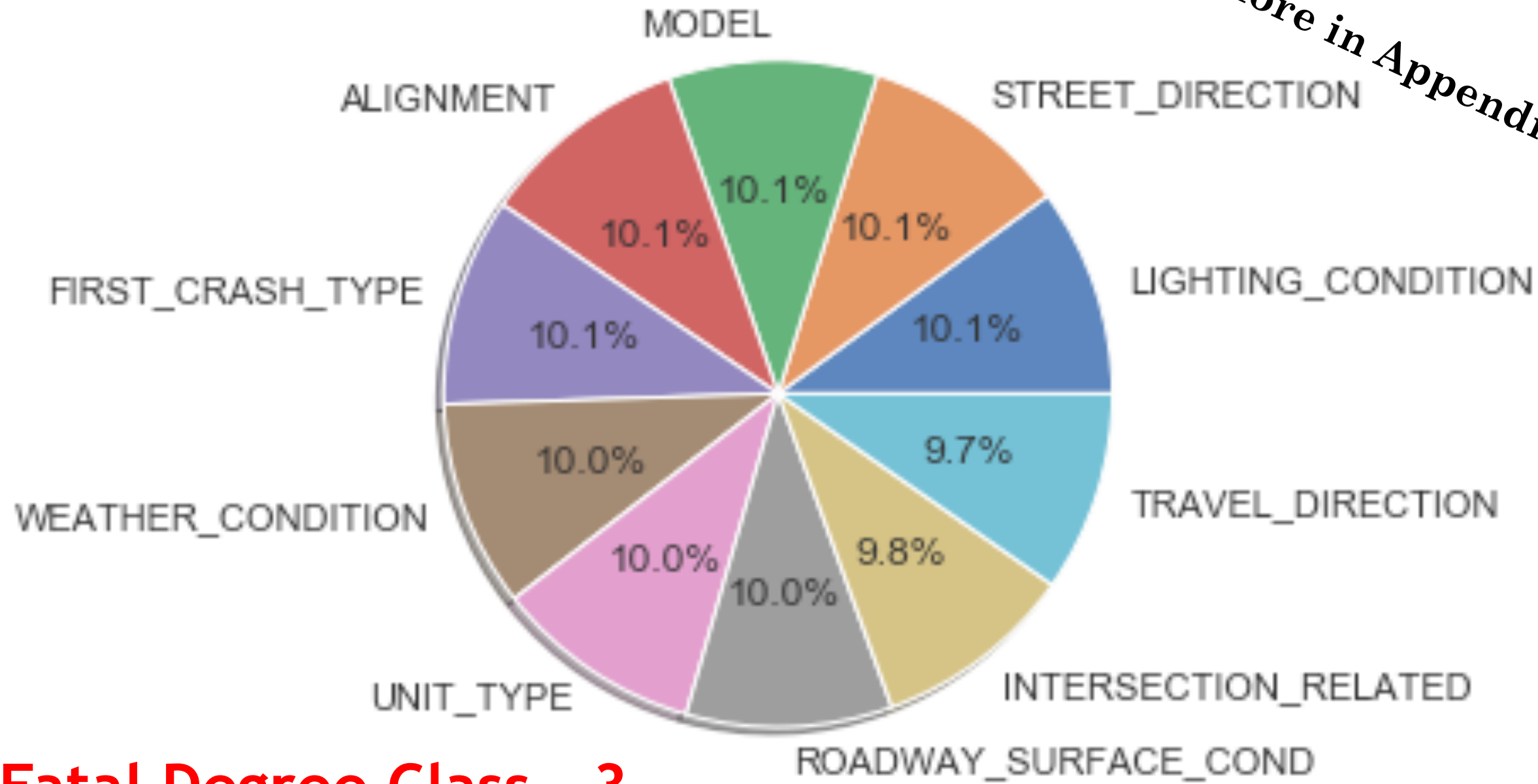




EDA:

Over COVID-19 outbreak, why did traffic accidents still take place, if Illinois Stay at Home order took effect on March 21st, 2020?

More in Appendix

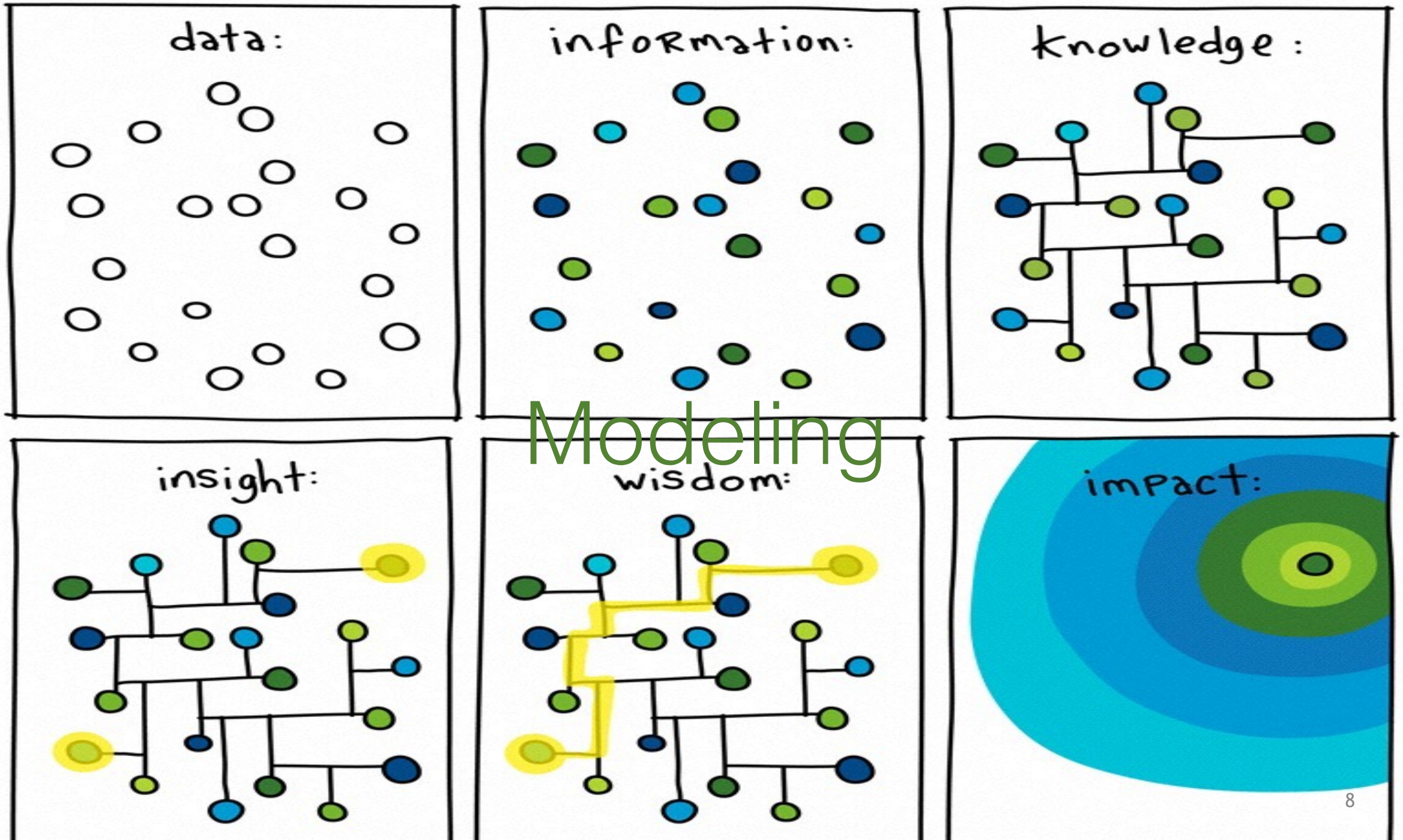


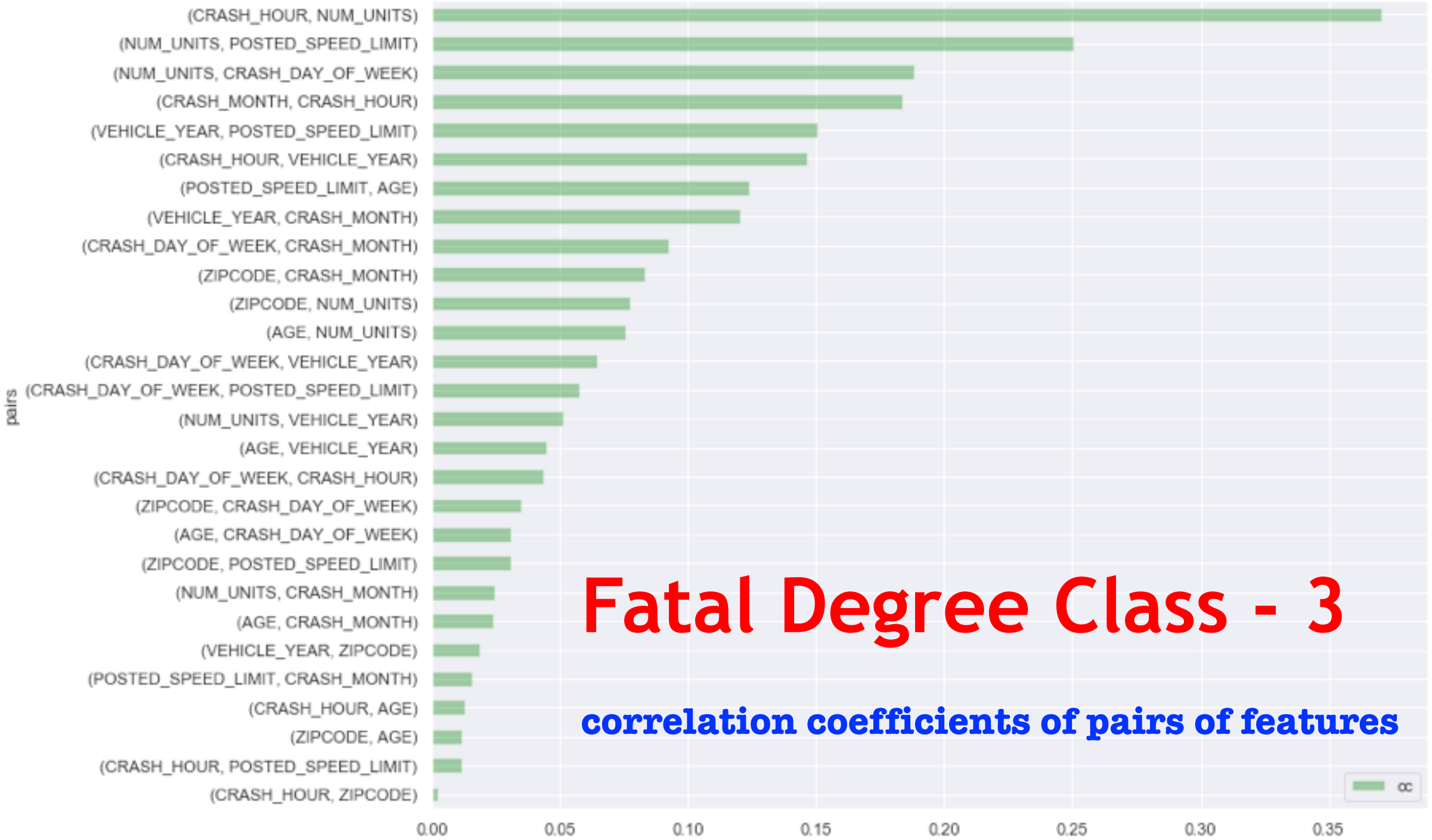
Fatal Degree Class - 3

Fatal Degree Class - 3

More in Appendix

1. **ANGLE**
2. **TURNING**
3. **REAR_END**
4. **SIDESWIPE_SAME_DIRECTION**
5. **PEDESTRIAN**
6. **HEAD_ON**
7. SIDESWIPE_OPPOSITE_DIRECTION
8. PARKED_MOTOR_VEHICLE
9. PEDALCYCLIST
10. FIXED_OBJECT
11. REAR_TO_FRONT
12. REAR_TO_SIDE
13. OTHER_OBJECT


data:
information:
knowledge:
insight:
wisdom:
impact:
Modeling


(CRASH_HOUR, NUM_UNITS)
(NUM_UNITS, POSTED_SPEED_LIMIT)
(NUM_UNITS, CRASH_DAY_OF_WEEK)
(CRASH_MONTH, CRASH_HOUR)
(VEHICLE_YEAR, POSTED_SPEED_LIMIT)
(CRASH_HOUR, VEHICLE_YEAR)
(POSTED_SPEED_LIMIT, AGE)
(VEHICLE_YEAR, CRASH_MONTH)
(CRASH_DAY_OF_WEEK, CRASH_MONTH)
(ZIPCODE, CRASH_MONTH)
(ZIPCODE, NUM_UNITS)
(AGE, NUM_UNITS)
(CRASH_DAY_OF_WEEK, VEHICLE_YEAR)
(CRASH_DAY_OF_WEEK, POSTED_SPEED_LIMIT)
(NUM_UNITS, VEHICLE_YEAR)
(AGE, VEHICLE_YEAR)
(CRASH_DAY_OF_WEEK, CRASH_HOUR)
(ZIPCODE, CRASH_DAY_OF_WEEK)
(AGE, CRASH_DAY_OF_WEEK)
(ZIPCODE, POSTED_SPEED_LIMIT)
(NUM_UNITS, CRASH_MONTH)
(AGE, CRASH_MONTH)
(VEHICLE_YEAR, ZIPCODE)
(POSTED_SPEED_LIMIT, CRASH_MONTH)
(CRASH_HOUR, AGE)
(ZIPCODE, AGE)
(CRASH_HOUR, POSTED_SPEED_LIMIT)
(CRASH_HOUR, ZIPCODE)
pairs
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Fatal Degree Class - 3
correlation coefficients of pairs of features

Summary of Data Interpretations

More in Appendix

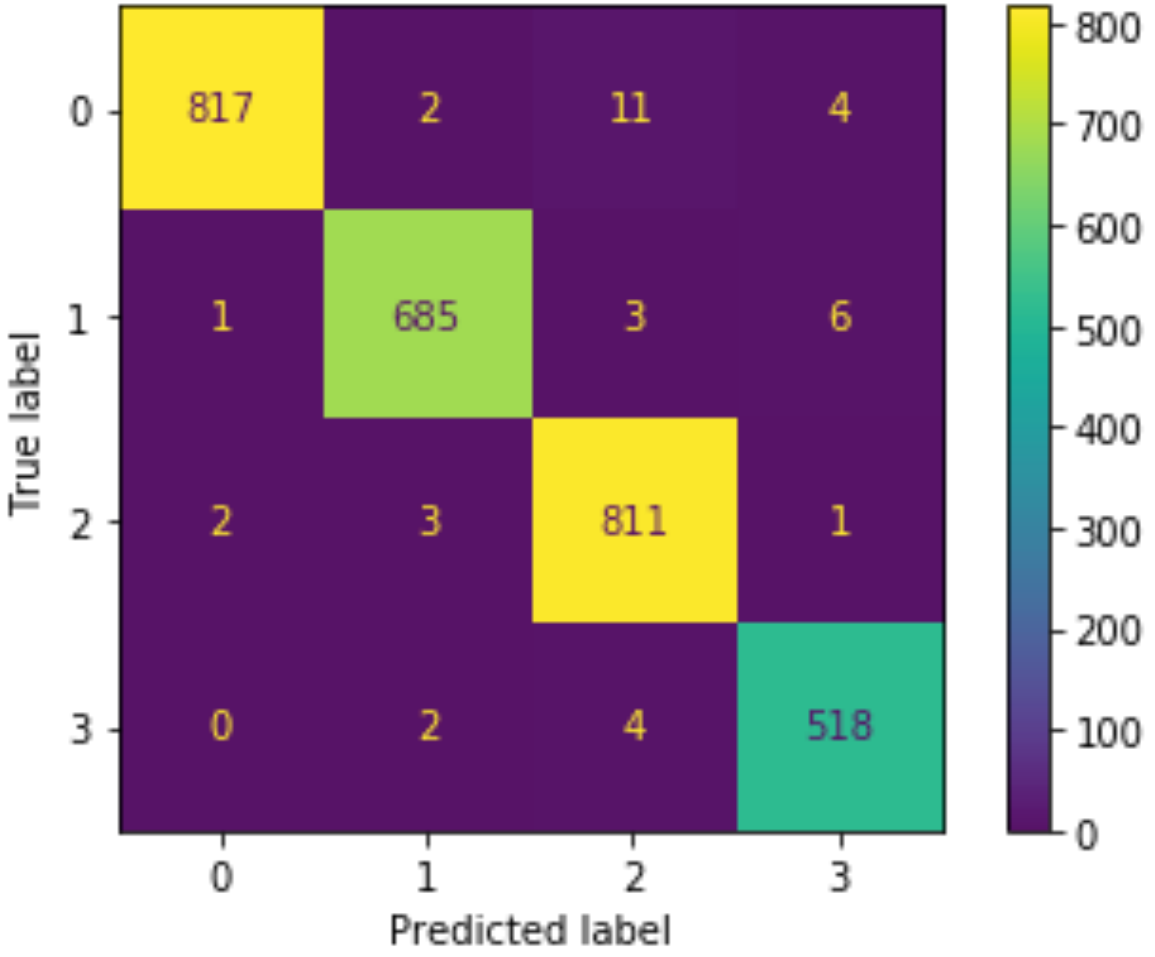
Tree-based models 27 features + 1 target	Accuracy in Test
Decision Tree (5)	**80.74%**
Bootstrap Aggregation (4)	**83.76%**
Random Forest (2)	**90.95%**
AdaBoost (6)	**67.75%**
Gradient Boost (3)	**84.22%**
XGBoost (1)	**95.59%**

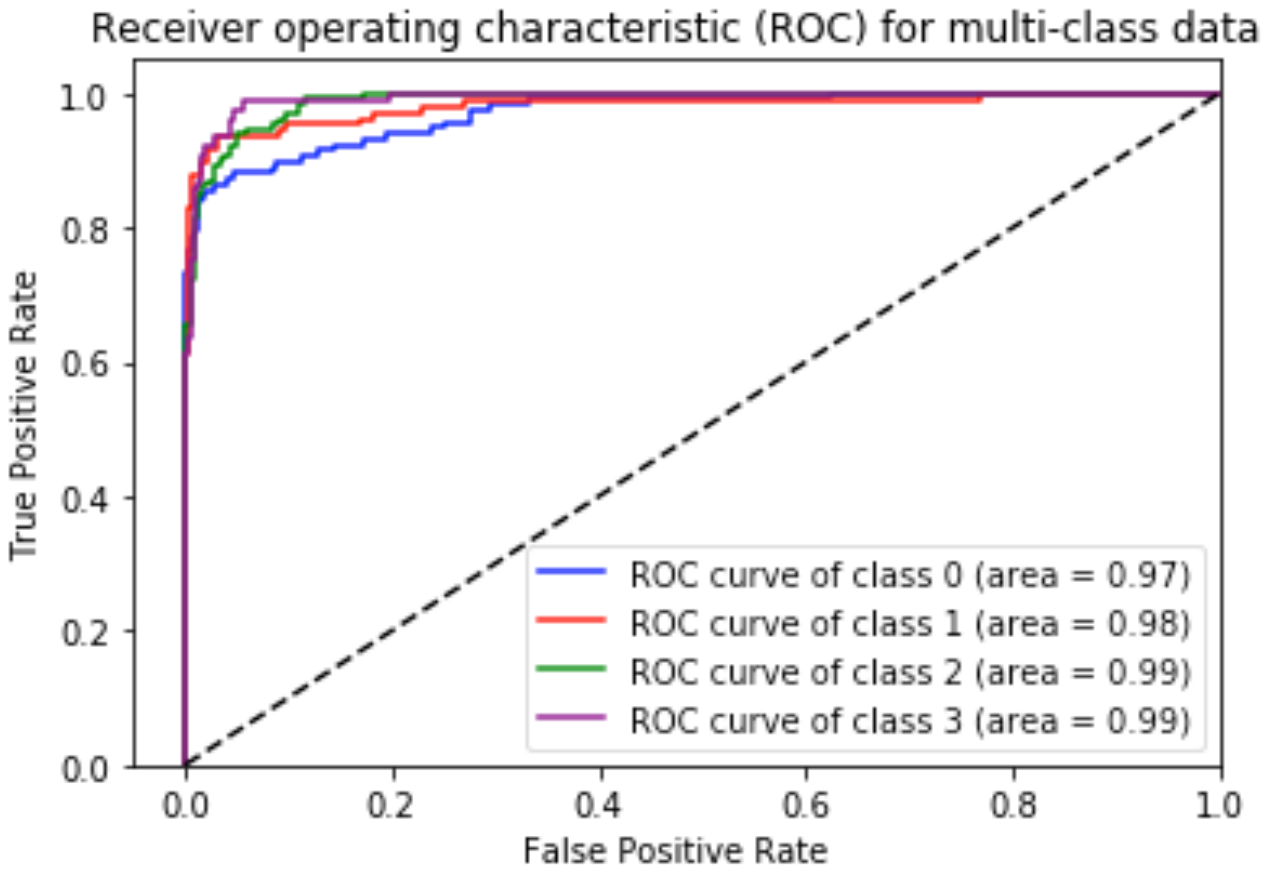
Grid Search found the following optimal parameters:

- learning_rate: 0.1
- max_depth: 6
- min_child_weight: 10
- n_estimators: 100
- subsample: 0.7
- Training Accuracy: 95.41%
- Validation accuracy: **95.59%**

multi-class	precision	recall	f1-score	support
0	0.990566	0.897436	0.941704	117.000000
1	0.981132	0.981132	0.981132	106.000000
2	0.948148	0.977099	0.962406	131.000000
3	0.892857	0.974026	0.931677	77.000000
accuracy	**0.955916**	**0.955916**	**0.955916**	**0.951276**
macro avg	0.953176	0.954230	0.954230	431.000000
weighted avg	0.957897	0.955910	0.955902	431.000000

Summary of Data Interpretations

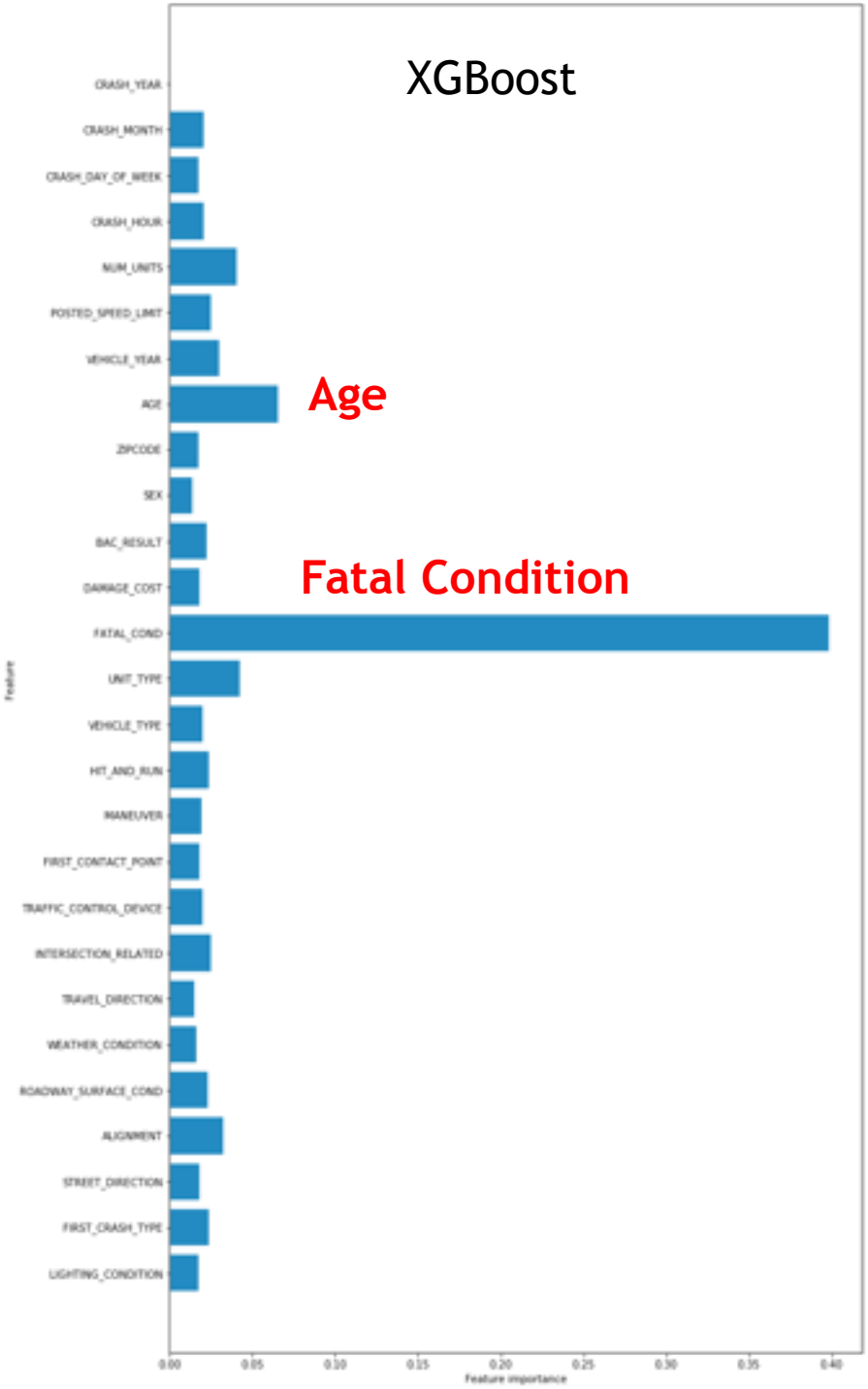




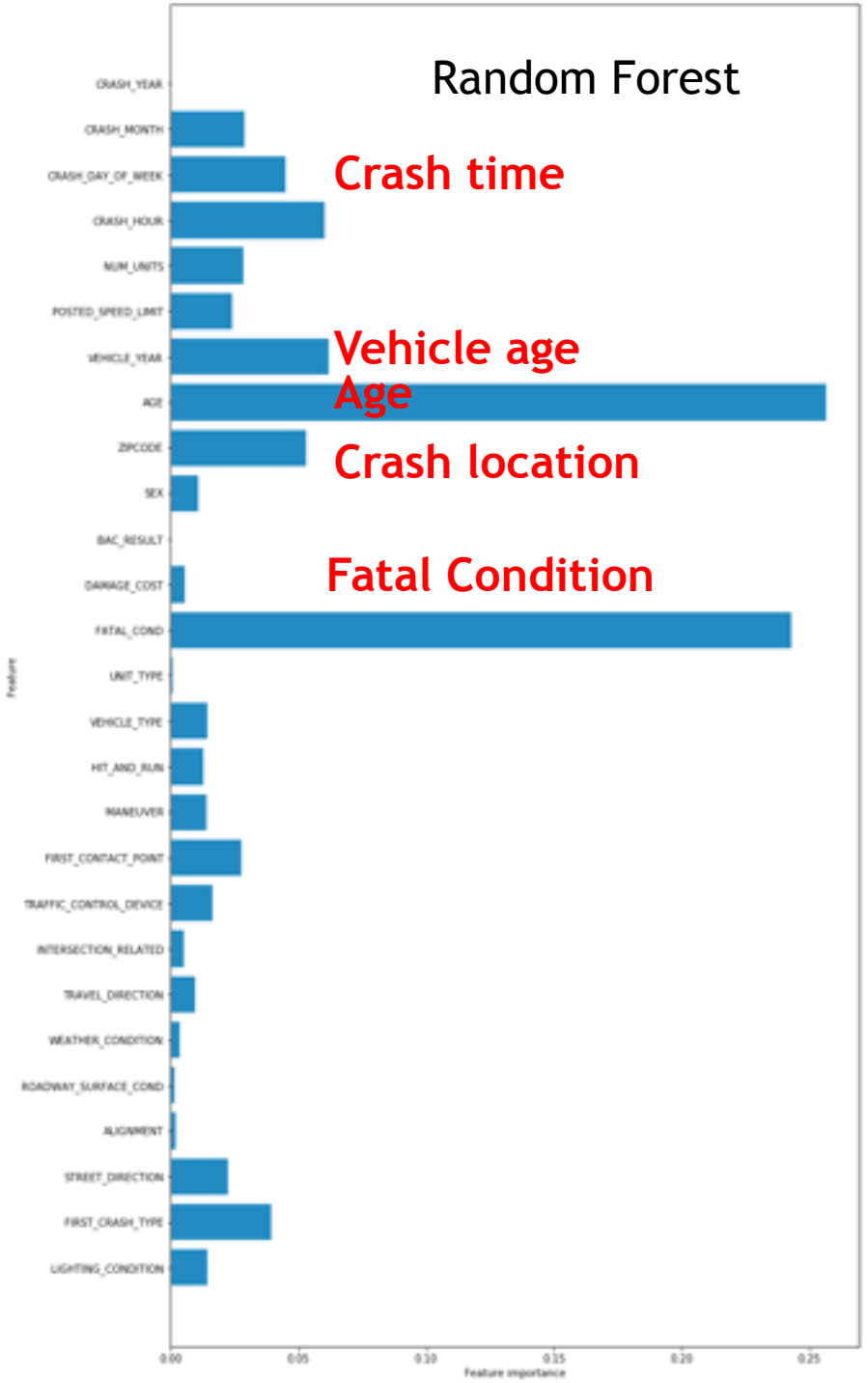
Recommendations - over COVID-19

Fatal conditions:

- 1 : pedestrian(s) + bicyclist(s) + other people surrounding only involved in crash
- 2 : vehicle only (neither 1 nor 3 included) involved in crash
- 3 : driver + passenger(s) (+ 1 possibly) involved in crash
- 4 : vehicle, driver and passenger(s) (+ 1 possibly) all involved in crash



Future Work - what're missing?



Thank-you

renjmindy@gmail.com

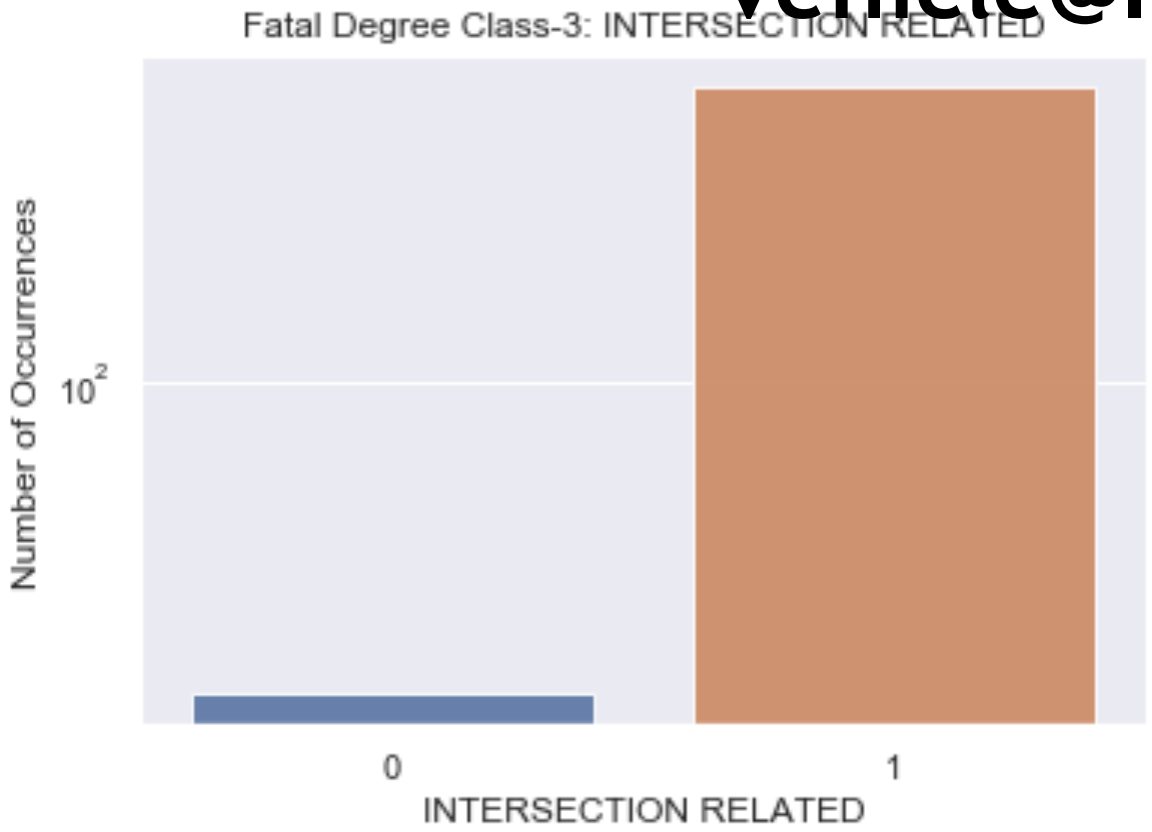
Appendix ---

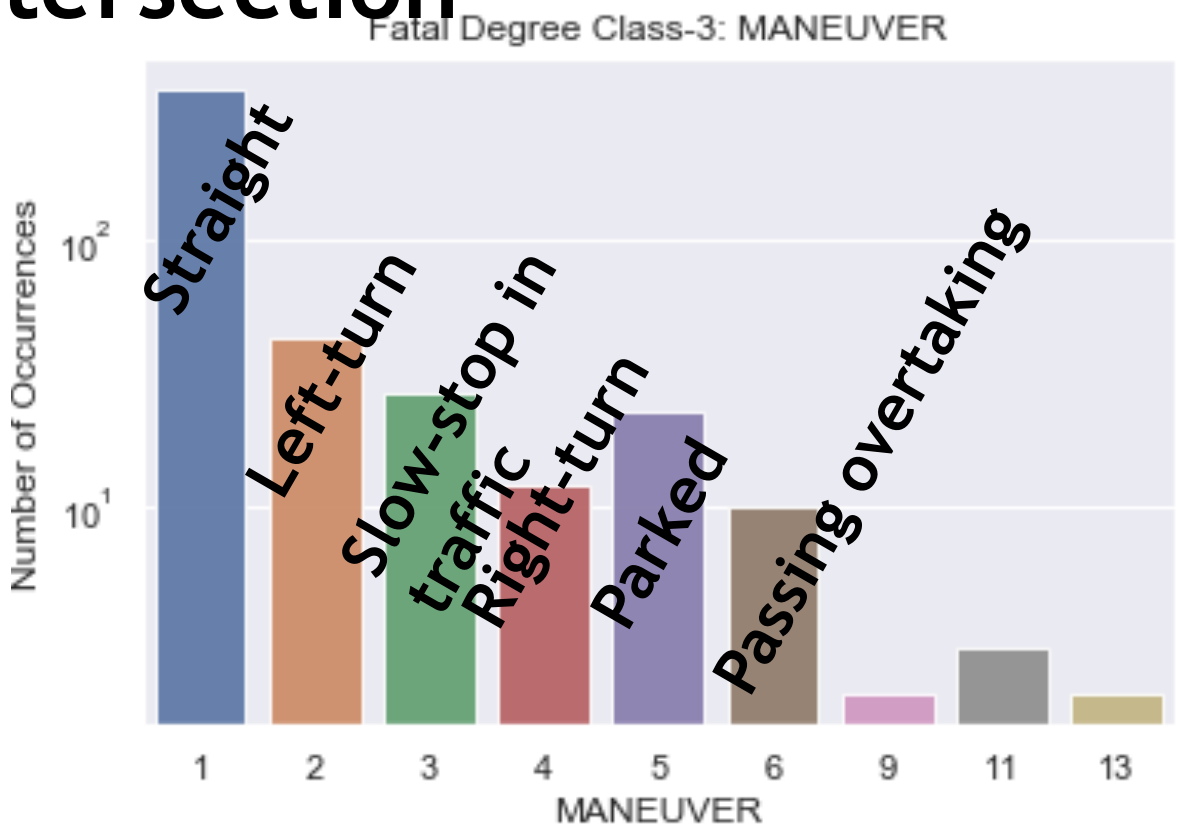
EDA-Q1:

What are Top 10 features yielding most information regarding crash causes and fatal degree levels?

Data Studies

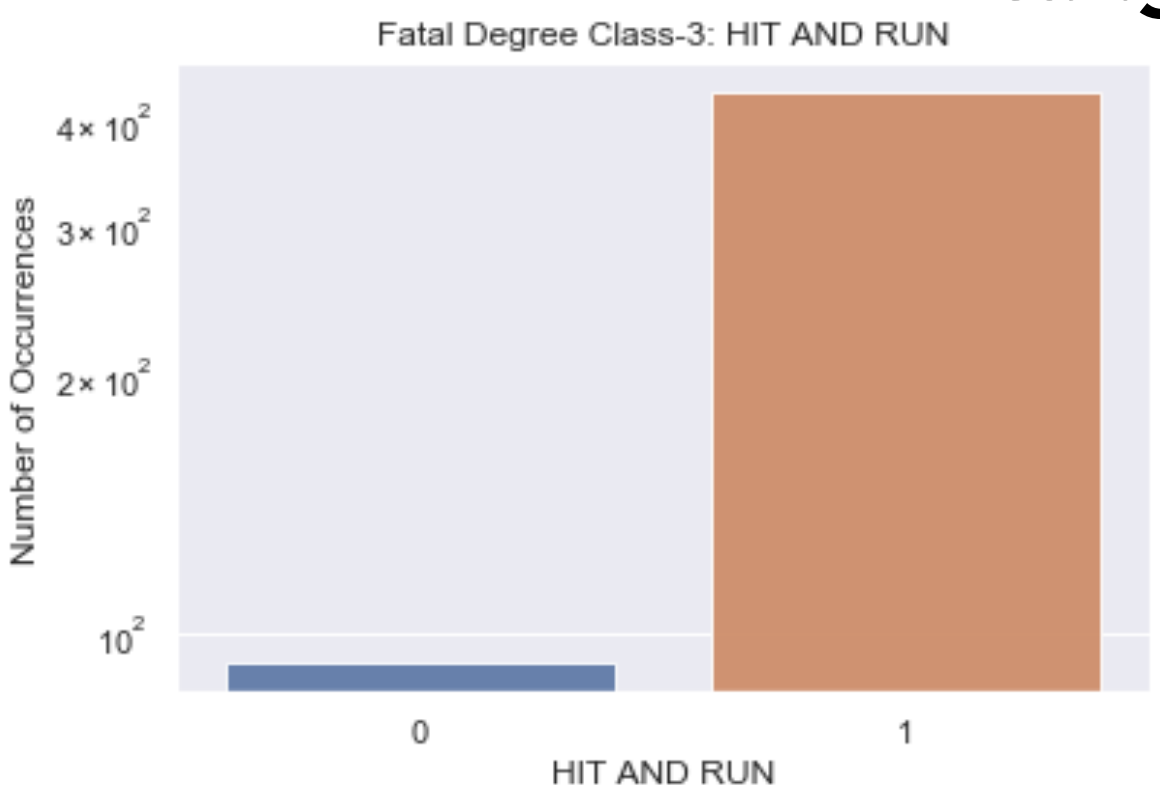
Turning vehicle@Intersection

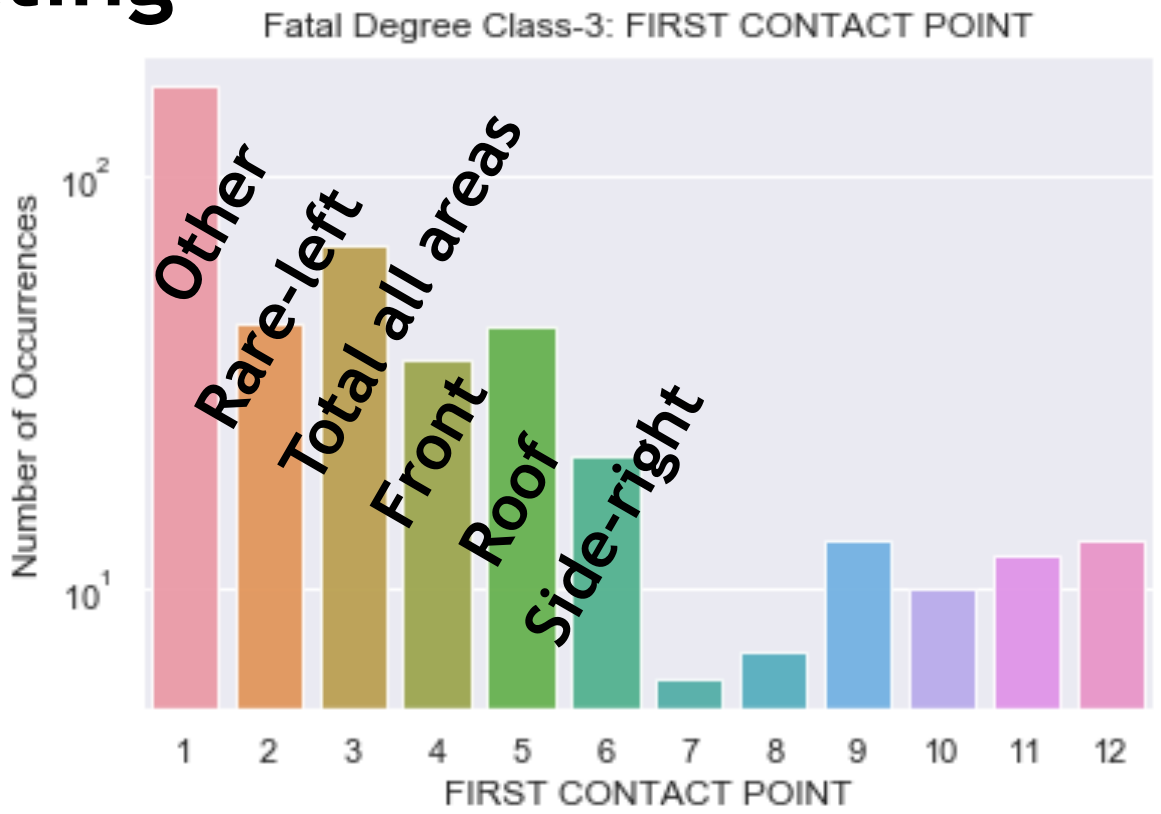




Data Studies

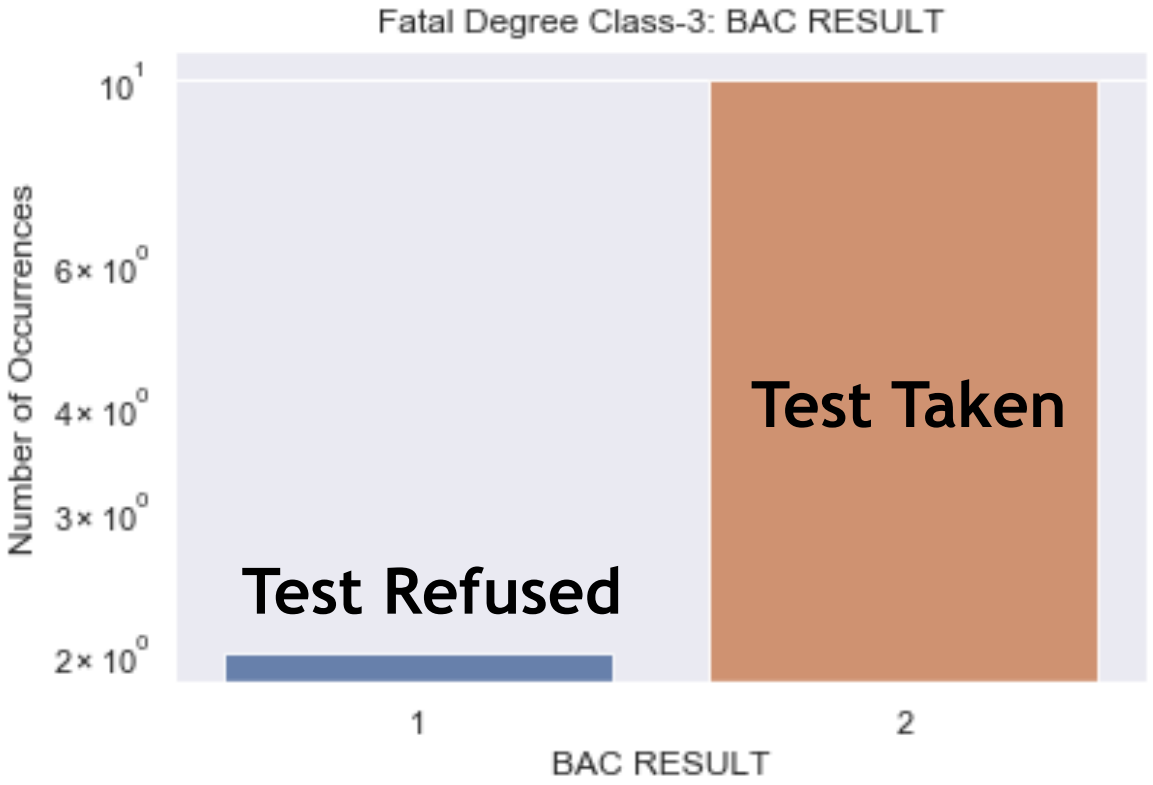
Rare-end collisions from tailgating

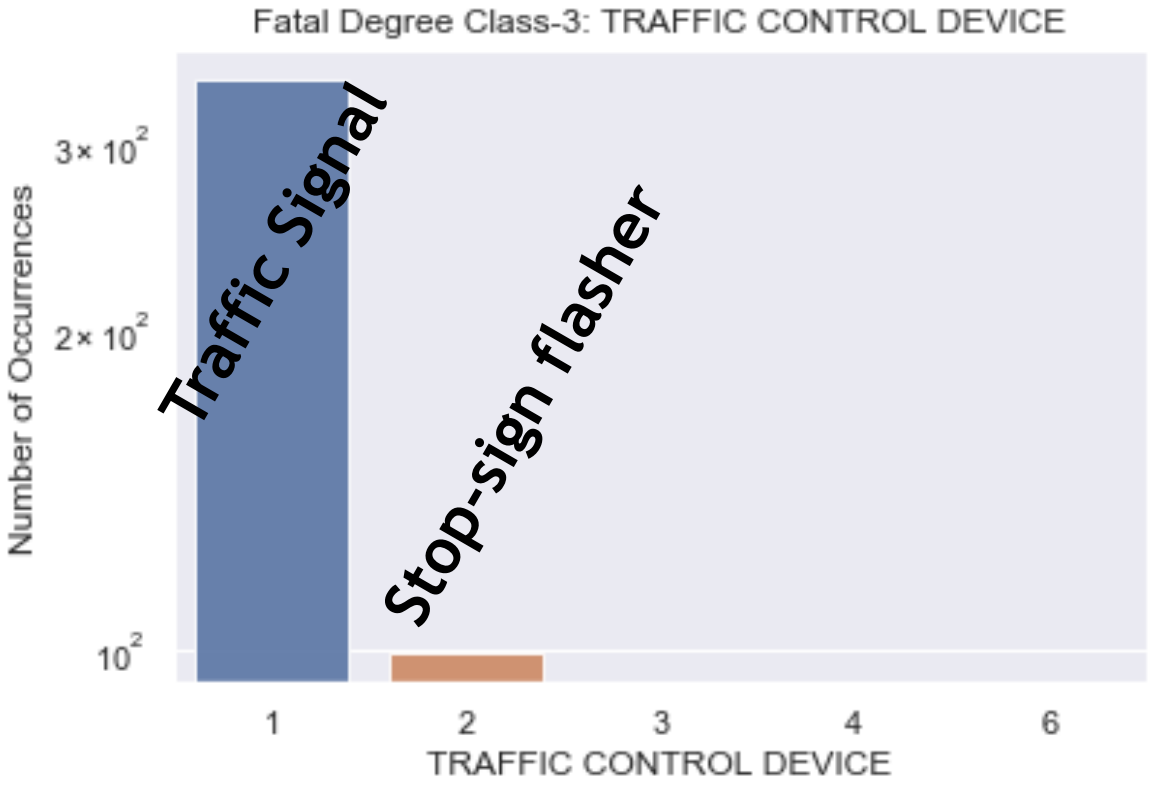




Data Studies

Driving under influence by factors





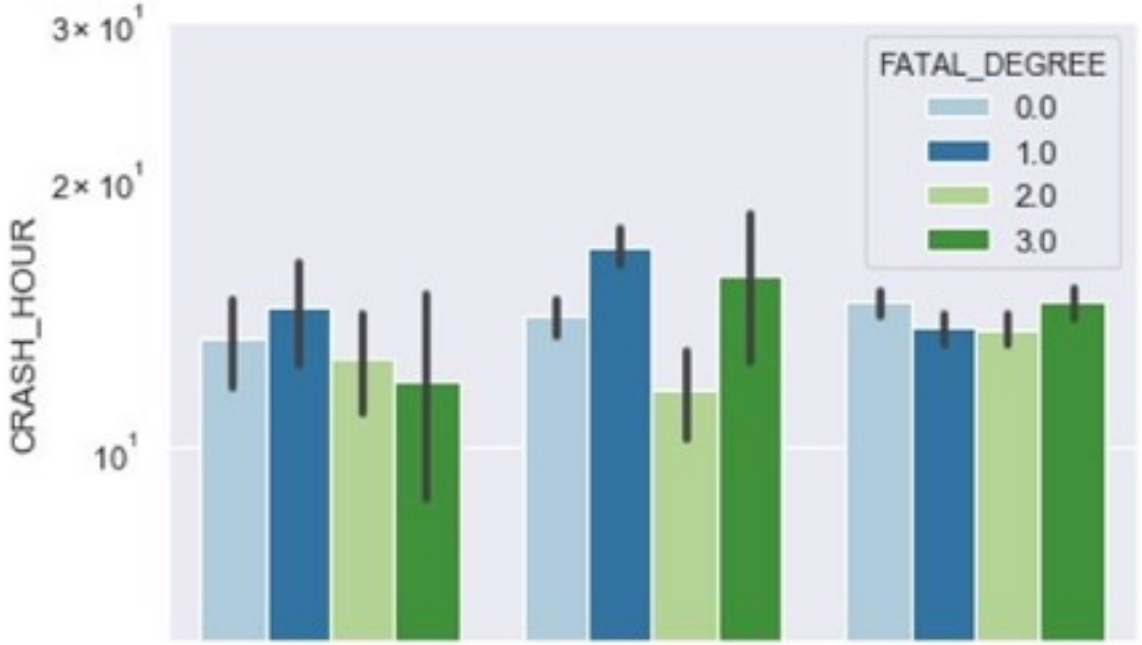
EDA-Q3:

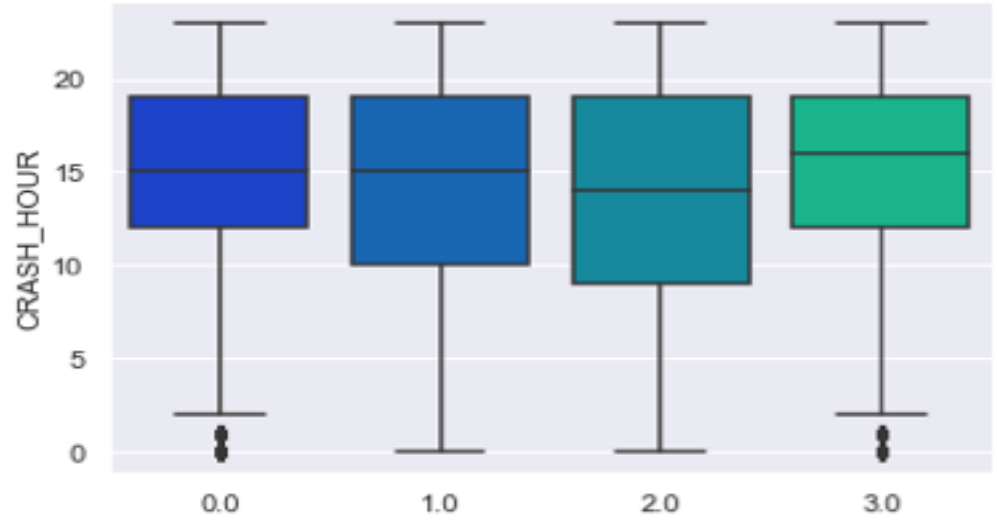
Following Q1, what're time-dependent? What're not? Any other factors possibly involved? What are they?

Data Studies

More in Appendix

Time Dependence





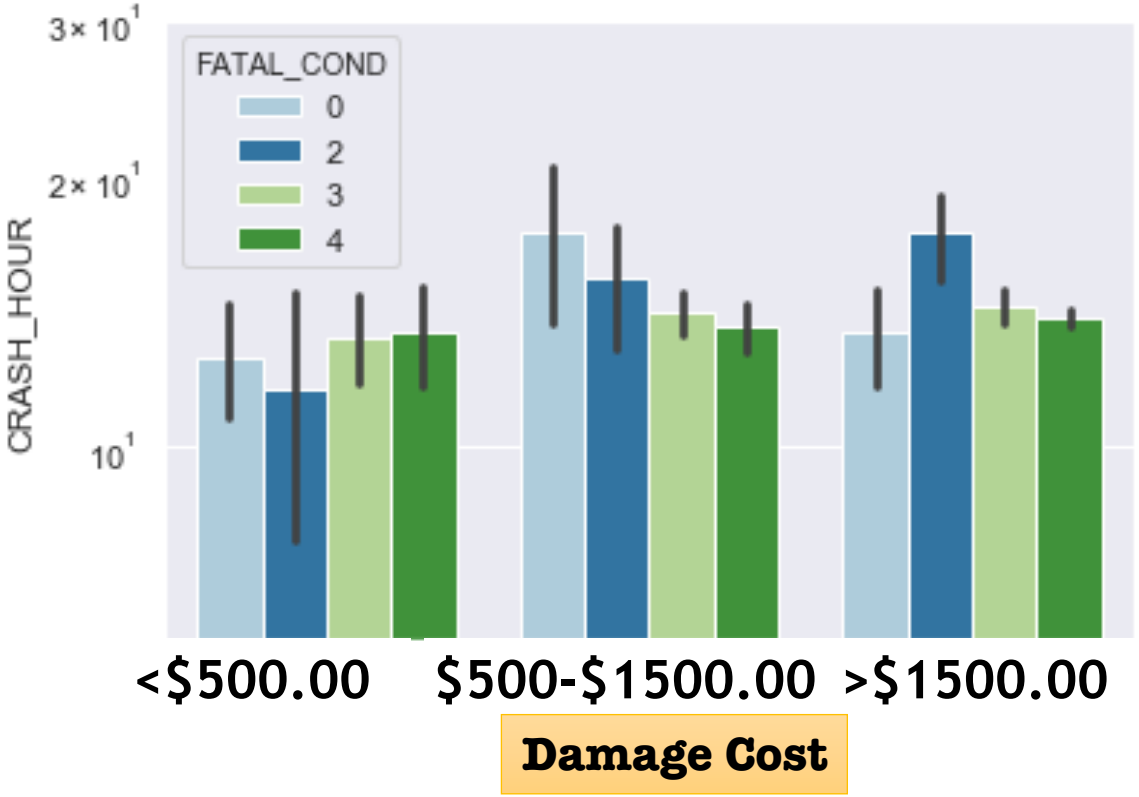
Fatal degrees:

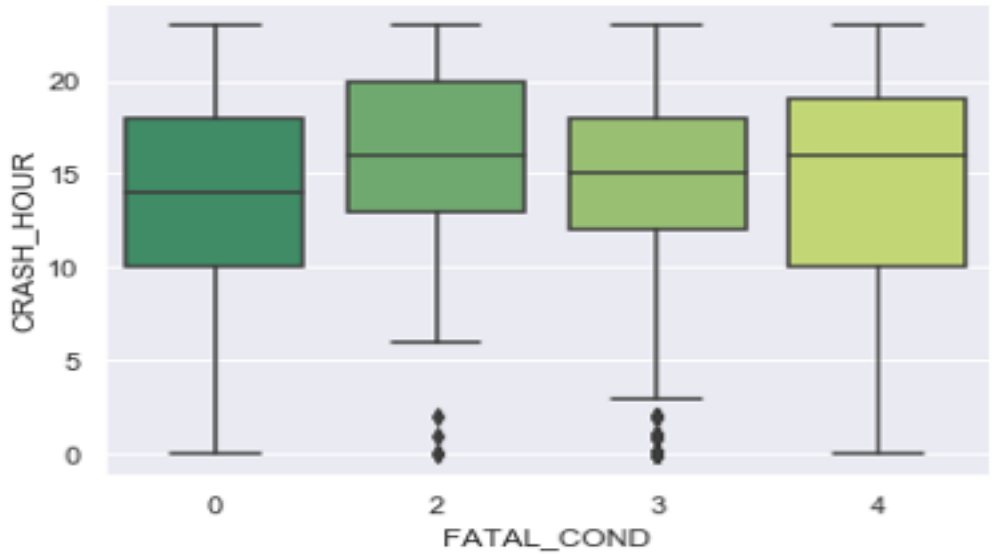
0 : no indication (-1 , 0.05)
1 : not evident (0.05 , 21.05)
2 : non-incapacitating - incapacitating (21.05 , 51.05)
3 : incapacitating - death (> 51.05)

Data Studies

More in Appendix

Time Dependence



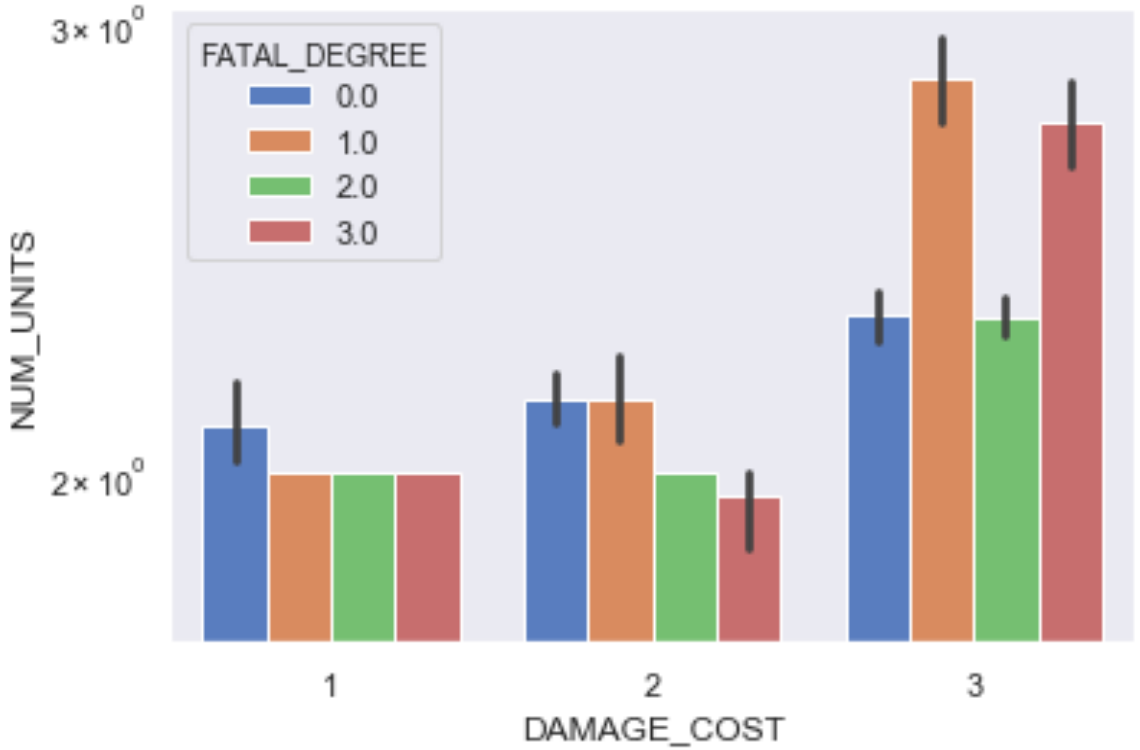


Fatal conditions:

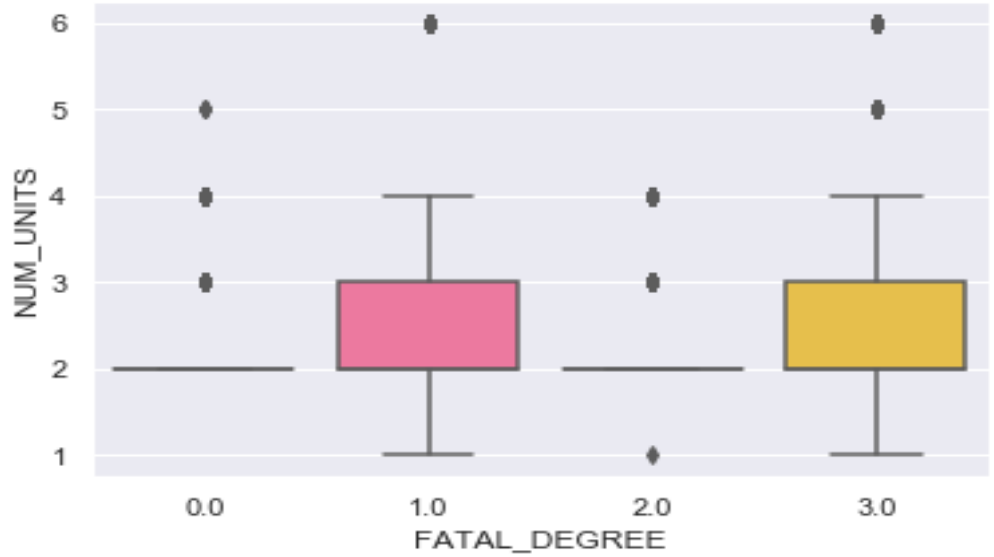
- 1 : pedestrian(s) + bicyclist(s) + other people surrounding only involved in crash
- 2 : vehicle only (neither 1 nor 3 included) involved in crash
- 3 : driver + passenger(s) (+ 1 possibly) involved in crash
- 4 : vehicle, driver and passenger(s) (+ 1 possibly) all involved in crash

Data Studies

Crash-unit Dependence



< $500.00 **$500 - $1500.00** **> $1500.00**

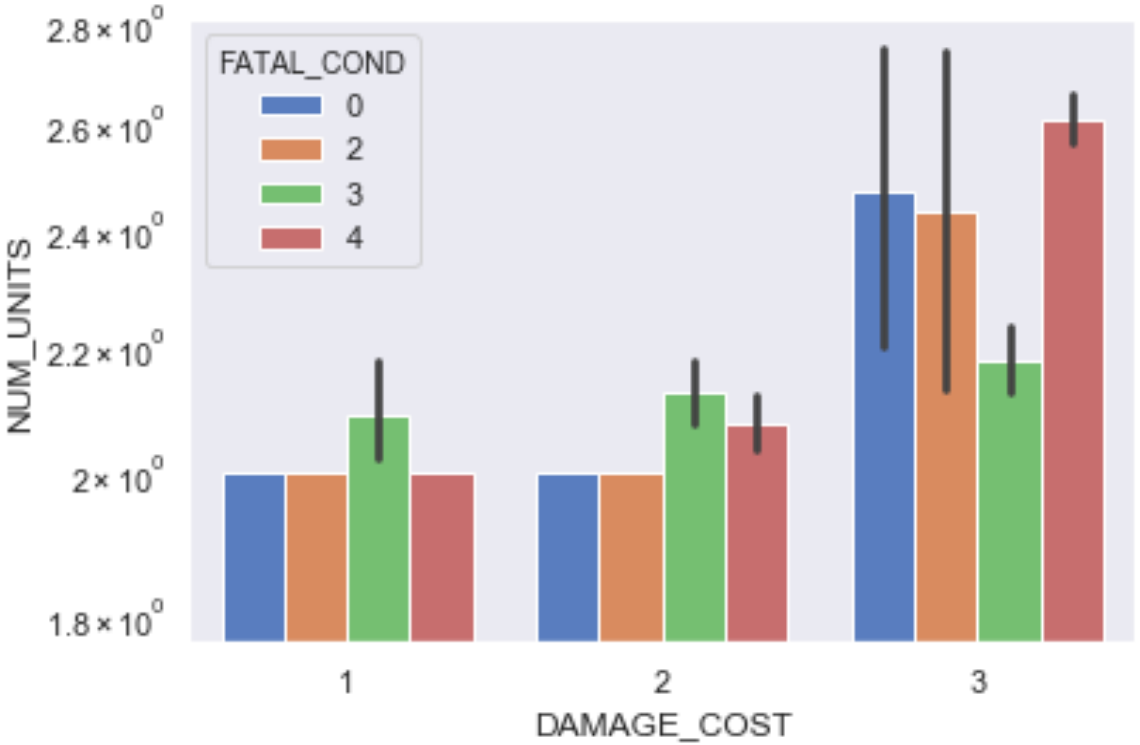


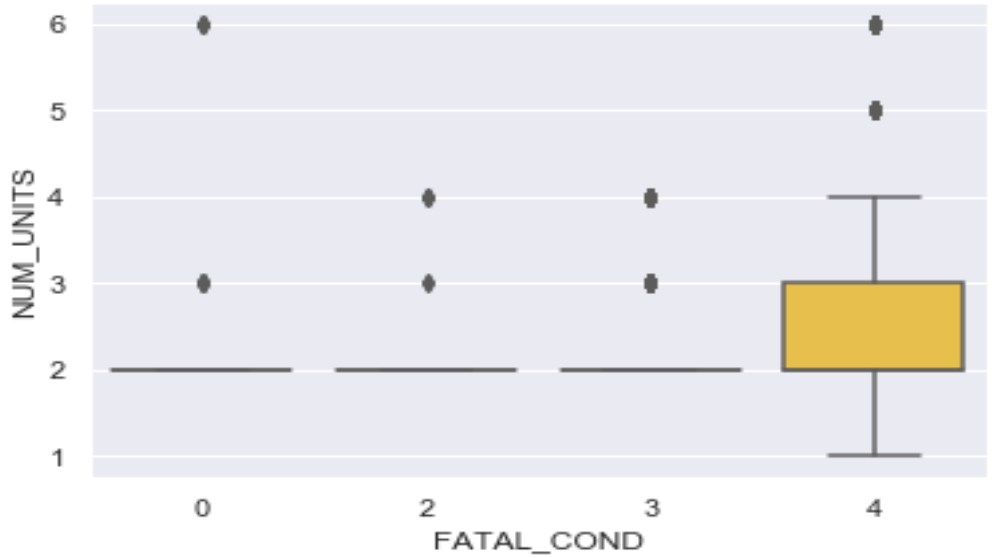
Fatal degrees:

- 0 : no indication (-1 , 0.05)
- 1 : not evident (0.05 , 21.05)
- 2 : non-incapacitating - incapacitating (21.05 , 51.05)
- 3 : incapacitating - death (> 51.05)

Data Studies

Crash-unit Dependence





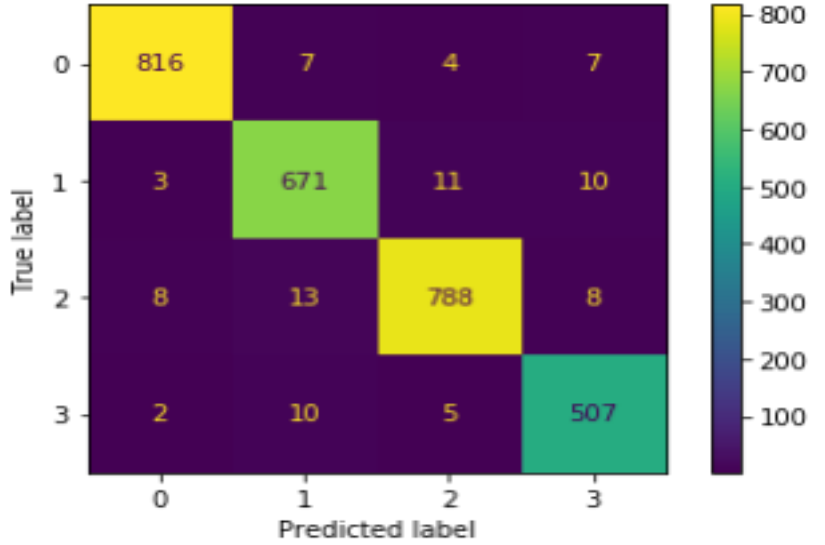
< $500.00 **$500 - $1500.00** **> $1500.00**

More in Modeling

Data Simplifications – Decision Tree

Accuracy:
79.58 %

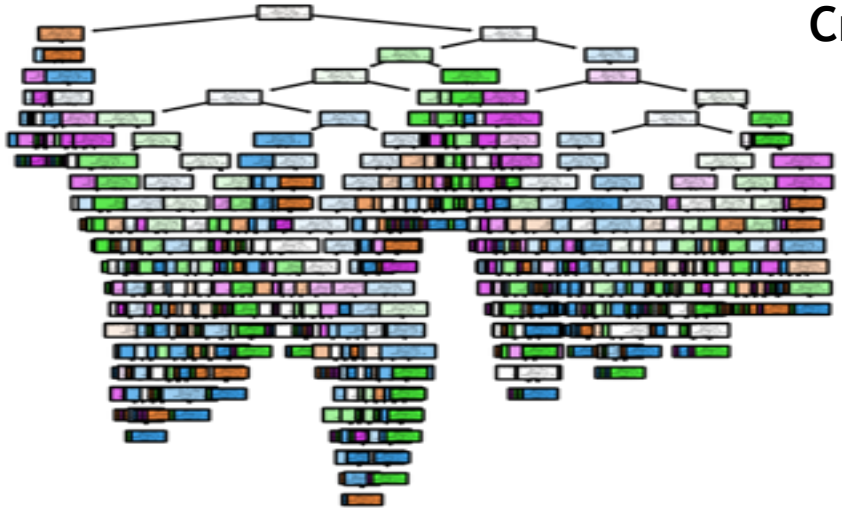
Criterion = **entropy**



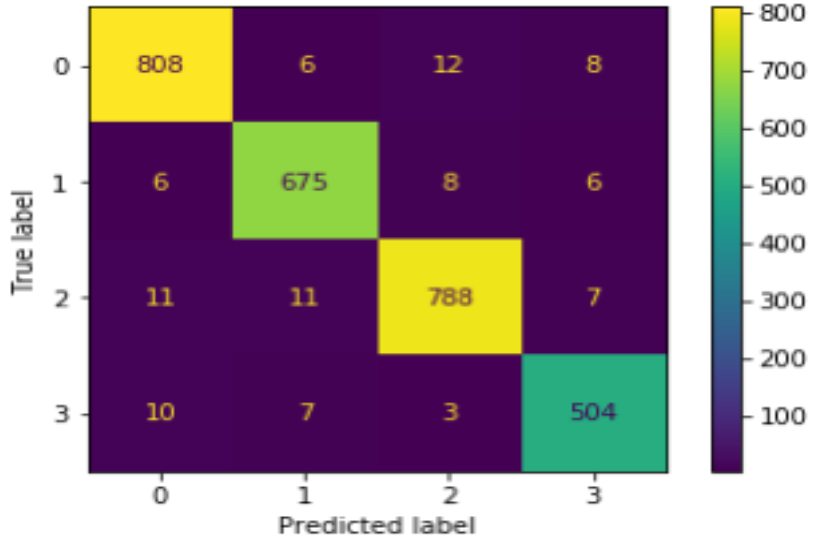
multi-class	precision	recall	f1-score	support
0	0.883929	0.846154	0.864629	117.000000
1	0.732143	0.773585	0.752294	106.000000
2	0.836066	0.778626	0.806324	131.000000
3	0.705882	0.779221	0.740741	77.000000
accuracy	**0.795824**	**0.795824**	**0.795824**	**0.795824**
macro avg	0.789505	0.794396	0.790997	431.000000
weighted avg	0.800242	0.795824	0.797147	431.000000

Data Simplifications – Decision Tree

Accuracy:
77.96 %



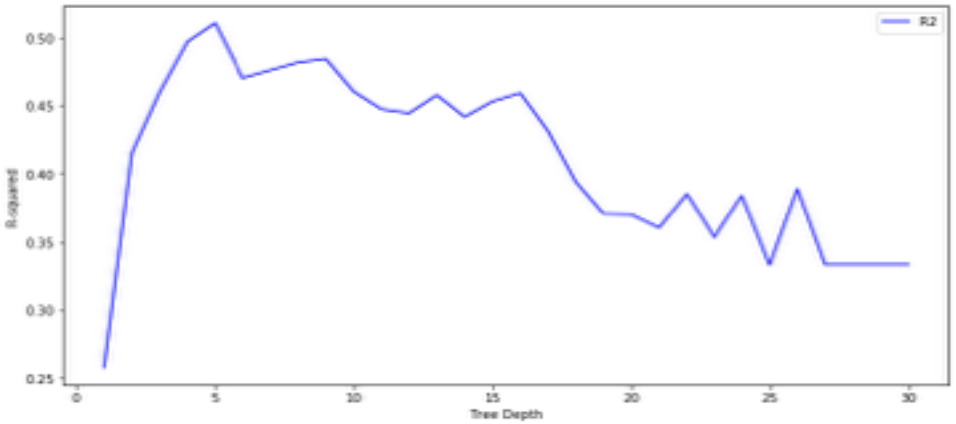
Criterion = **gini**

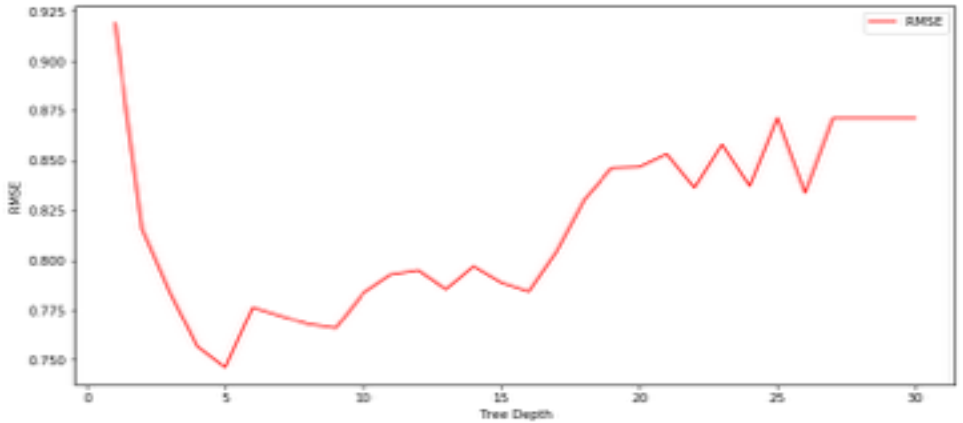


multi-class	precision	recall	f1-score	support
0	0.771186	0.777778	0.774468	117.000000
1	0.781818	0.811321	0.796296	106.000000
2	0.816000	0.778626	0.796875	131.000000
3	0.730769	0.740260	0.735484	77.000000
accuracy	**0.779582**	**0.779582**	**0.779582**	**0.779582**
macro avg	0.774943	0.776996	0.775781	431.000000
weighted avg	0.780201	0.779582	0.779682	431.000000

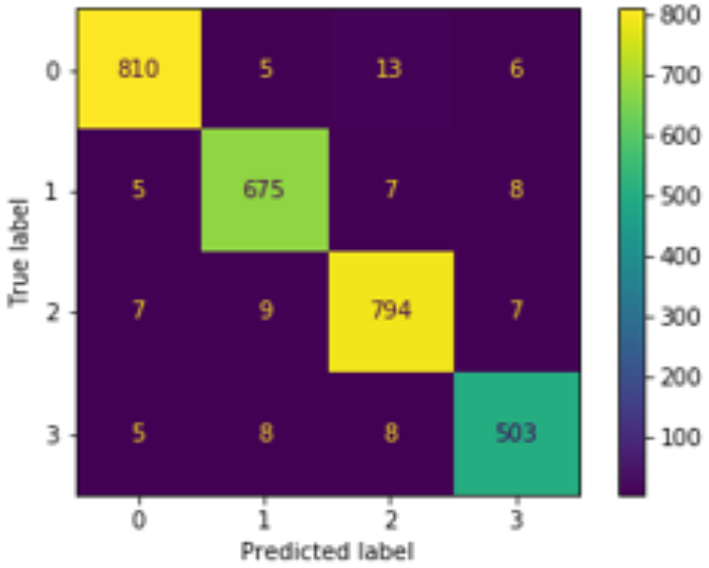
Data Simplifications – Regression CART tree

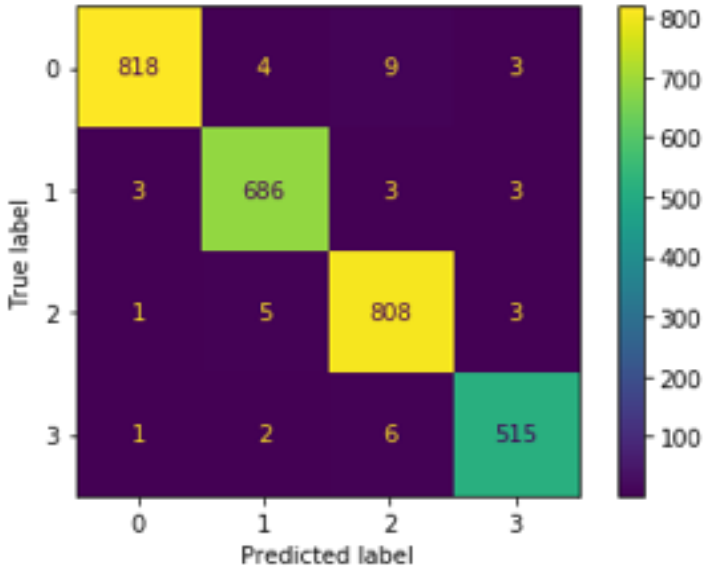
statistics	MAE	MSE	RMSE	R^2
Before tuning	0.37	0.73	0.85	0.36
After tuning	0.47	0.56	0.75	0.51





Data Simplifications – Bagged (top) / Random Forest (bottom)





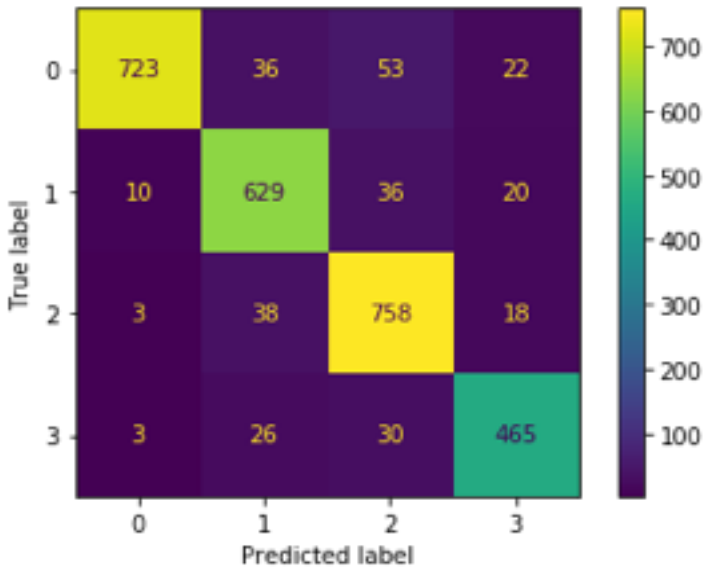
multi-class	precision	recall	f1-score	support
0	0.863636	0.811966	0.837004	117.000000
1	0.838095	0.830189	0.834123	106.000000
2	0.822222	0.847328	0.834586	131.000000
3	0.753086	0.792208	0.772152	77.000000
accuracy	0.823666	0.823666	0.823666	0.823666
macro avg	0.819260	0.820423	0.819466	431.000000
weighted avg	0.825017	0.823666	0.823975	431.000000

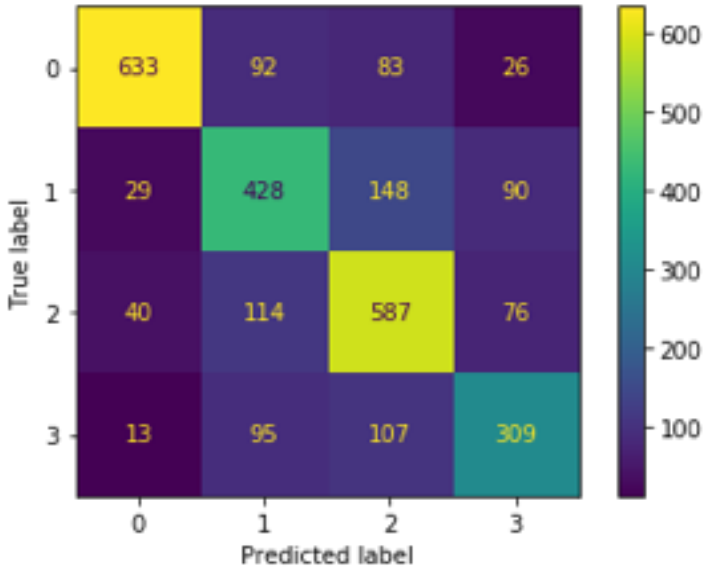
multi-class	precision	recall	f1-score	support
0	0.952830	0.863248	0.905830	117.000000
1	0.898148	0.915094	0.906542	106.000000
2	0.871429	0.931298	0.900369	131.000000
3	0.883117	0.883117	0.883117	77.000000
accuracy	0.900232	0.900232	0.900232	0.900232
macro avg	0.901381	0.898189	0.898964	431.000000
weighted avg	0.902186	0.900232	0.900287	431.000000

GridSearch CV

	Criterion		Max_depth	Min_samples_split	Min_samples_leaf	Training Score	Testing Score
DT	Entropy, Gini		11, 13, 15, 17, 19, 21, 23, 25	2, 5, 10	1, 2, 3	91.24%	**78.89%**
							'criterion': 'entropy', 'max_depth': 21, 'min_samples_leaf': 1, 'min_samples_split': 2
		N_estimators					
RF	Entropy, Gini	10, 30, 50, 70, 90, 110	2, 6, 10, 14, 18, 22	5, 10, 15	3, 6, 9	80.61%	**83.99%**
							'criterion': 'entropy', 'max_depth': 18, 'min_samples_leaf': 3, 'min_samples_split': 5, 'n_estimators': 70

Data Simplifications – Gradient (top) /AdaBoost (bottom)





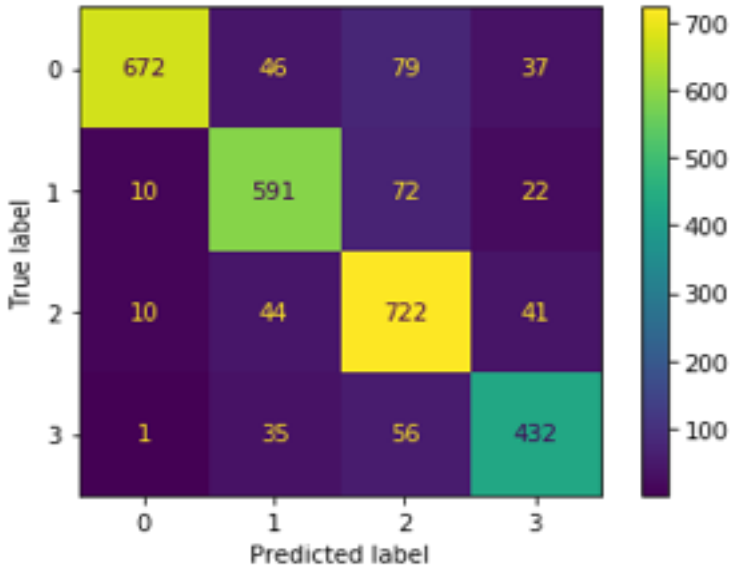
multi-class	precision	recall	f1-score	support
0	0.968750	0.794872	0.873239	117.000000
1	0.798246	0.858491	0.827273	106.000000
2	0.809160	0.809160	0.809160	131.000000
3	0.744444	0.870130	0.802395	77.000000
accuracy	**0.828306**	**0.828306**	**0.828306**	**0.828306**
macro avg	0.830150	0.833163	0.828017	431.000000
weighted avg	0.838237	0.828306	0.829801	431.000000

multi-class	precision	recall	f1-score	support
0	0.864078	0.760684	0.809091	117.000000
1	0.566372	0.603774	0.584475	106.000000
2	0.613139	0.641221	0.626866	131.000000
3	0.576923	0.584416	0.580645	77.000000
accuracy	**0.654292**	**0.654292**	**0.654292**	**0.654292**
macro avg	0.655128	0.647524	0.650269	431.000000
weighted avg	0.663287	0.654292	0.657650	431.000000

Tuning Regression Tree

	Criterion		Max_depth	Min_samples_split	Min_samples_leaf	Training Score	Testing Score
DT	Entropy, Gini		11, 13, 15, 17, 19, 21, 23, 25	2, 5, 10	1, 2, 3	91.24%	**78.89%**
							'criterion': 'entropy', 'max_depth': 21, 'min_samples_leaf': 1, 'min_samples_split': 2
		N_estimators					
RF	Entropy, Gini	10, 30, 50, 70, 90, 110	2, 6, 10, 14, 18, 22	5, 10, 15	3, 6, 9	80.61%	**83.99%**
							'criterion': 'entropy', 'max_depth': 18, 'min_samples_leaf': 3, 'min_samples_split': 5, 'n_estimators': 70

Data Simplifications - XGBoost



Grid Search found the following optimal parameters:

- learning_rate: 0.1
- max_depth: 6
- min_child_weight: 10
- n_estimators: 100
- subsample: 0.7
- Training Accuracy: 94.79%
- Validation accuracy: **95.13%**

multi-class	precision	recall	f1-score	support
0	0.990741	0.914530	0.951111	117.000000
1	0.945946	0.990566	0.967742	106.000000
2	0.946970	0.954198	0.950570	131.000000
3	0.912500	0.948052	0.929936	77.000000
accuracy	**0.951276**	**0.951276**	**0.951276**	**0.951276**
macro avg	0.949039	0.951837	0.949840	431.000000
weighted avg	0.952442	0.951276	0.951254	431.000000

Recommendations - over COVID-19

- **Hit-and-run**: Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid
- **Unit-type**: The type of unit - (1) driver; (2) parked; (3) driverless
- **First-contact-point**: (1) other; (2) rear-left; (3) total-all-areas; (4) front; (5) roof; (6) side-right; (7) side-left; (8) rear; (9) front-left; (10) rear-right; (11) front-right; (12) under-carriage
- **Traffic-control-device**: Traffic control device present at crash location
- **Vehicle-type**: The type of vehicle - (1) passenger; (2) sport-utility-vehicle-SUV; (3) van-mini-van; (4) pick-up; (5) truck-single-unit; (6) other
- **Roadway-surface-condition**: Road surface condition
- **Age**: Age of person involved in crash
- **Vehicle-year**: The model year of the vehicle
- **Num-units**: Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.

Future Work - what're missing?

- **Street-direction**: street address (N, E, S, W) of crash location
- **Alignment**: street alignment at crash location
- **First-crash-type**: type of first collision in crash

  ➡ **Travel direction**: the direction in which the unit was traveling prior to the crash

More feature engineered work could be considered by accounting for three more features as listed above! I am confident that model accuracy could approach to some level as close to 99% as possible!