First Responders

Jonathan Beltran, Niraj Saran, Marcus Salandy-Defour, and Kehinde Ajayi

Data A Team
Apr 16, 2022

Problem Statement

First Responders want to be better prepared by the time they reach an accident site

- ➢ Based on number of vehicles, people, time of day, weather conditions, how many fatalities and what level of injury severity to expect
- ➢ What are the worst days/times when they should be appropriately staffed
- ➢ What can be done to reduce the number of fatalities and severity of injuries

Data Sources and Information

National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS)/Crash Report Sampling System (CRSS)

- ➢ FARS contains data derived from a census of fatal motor vehicle traffic crashes in the 50 States, the District of Columbia, and Puerto Rico.
- ➢ To be included in FARS, a crash must involve a motor vehicle traveling on a trafficway customarily open to the public and must result in the death of at least one person (occupant of a vehicle or a non-motorist) within 30 days of the crash.
- ➢ FARS was conceived, designed, and developed by NHTSA's National Center for Statistics and Analysis (NCSA) in 1975 to provide an overall measure of highway safety, to help identify traffic safety problems, to suggest solutions, and to help provide an objective basis to evaluate the effectiveness of motor vehicle safety standards and highway safety programs.

<u>FARS data 'dictionary'</u>:

- ➢ Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2020:
  - ○ This multi-year analytical user's manual provides documentation on the historical coding practices of FARS from 1975 to 2020. In other words, this manual presents the evolution of FARS coding from inception through present. The manual includes the data elements that are contained in FARS and other useful information that will enable the users to become familiar with the data system.
- ➢ Fatalities and Coding and Validation manual
  - ○ Provides more detailed definitions for each data element and attribute for a given year.

US Census Bureau, Dept. of Labor, : for county population and location (latitude and longitude) data

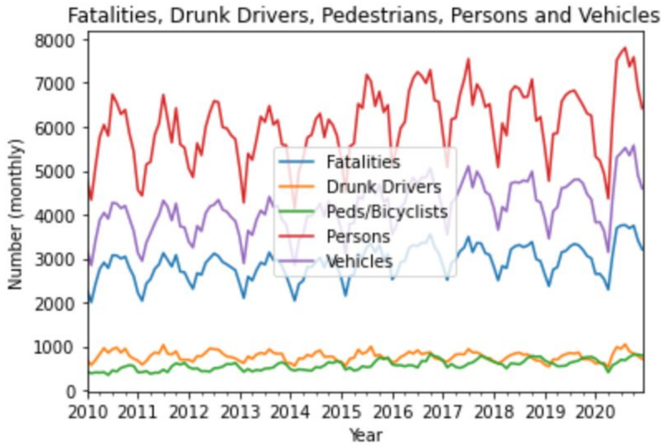
Data Gathering

- Downloaded FARS data from 2010-2020 (2.3 GB of data!)
- Each year comprised 23-36 .csv files
- >200 features across all files
- Types of data available:
  - Crash-level: Date/Time, GPS, Work Zone, EMS Arrival, Highway, etc.
  - Vehicle-level: Type, Make/Model, Traveling Speed, Registration, Rollover, etc.
  - Driver-level: Presence, License State/Zip/Status, Height, Speeding, etc.
  - Precrash-level: Speed Limit, Roadway Grade, Distracted, Vision Obscured, etc.
  - Person-level (MVO): Age, Sex, Seating Position, Air Bag, Ejection, etc.
  - Person-level (NaMVO): Location, Alcohol test, Safety Equipment, etc.
- Automated merging of .csv files based on desired features

Problem Statement

First Responders want to be better prepared by the time they reach an accident site

- ➢ Based on vehicle information, personal data, **time of day**, weather conditions, how many fatalities and what level of injury severity to expect?
- ➢ **What are the worst days/times when they should be appropriately staffed?**
- ➢ What can be done to **reduce the number of fatalities** and severity of injuries?



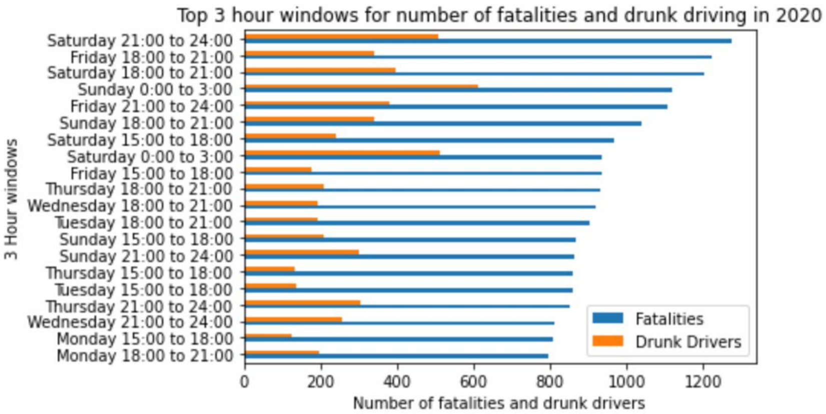
Number of fatalities, drunk drivers, pedestrians have been increasing in spite of latest tech, safety features and enforcement

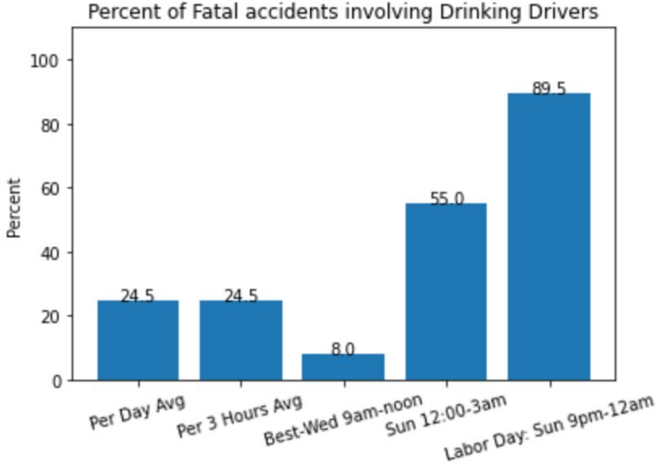
Methodology: Vector AutoRegression, Recurrent Neural Network

- EDA: Pre-crash, weather, drinking drivers, dozens of features
- 3 hour windows: trends, weekends, holidays
- Model selection: most popular for Time Series
- VAR: bi-directional, multi-variate
  - Forecast future values: fatalities, drunk driver, pedestrians, persons and vehicles
  - Stationarity: ADF, ACF, PACF
  - Decompose: trend, seasonality and residuals
- RNN: Deep learning for sequential input
  - Same variables as VAR; predict Fatalities
  - Keras TimeseriesGenerator: sequence length, batchsize
  - GRU, LSTM: alleviate the vanishing gradient problem
  - Regularization: Dropout, EarlyStopping
- Metrics: RMSE and MAPE - intuitive interpretation

EDA Findings
Sobering Truth

- Number of fatalities have been increasing steadily every year. Percent involving drunk drivers is ~25%
- During the first Covid lockdowns of Mar-Apr 2020, the number of fatalities dropped significantly, but soon exceeded pre-pandemic levels by summer
- % Drunk Drivers in fatalities:
  - Safest 3 hours: Wed 9am-noon: 8%
  - Really dangerous: Sun 12:00-3am: 55%
  - Worst: Labor Day Sun 9pm-12am: 90%
- The range is from 8% to 90% - clearly calls for better sobriety checks and roving patrols during the hours from 9pm to 3am on Holidays and weekends (bar closing time!)

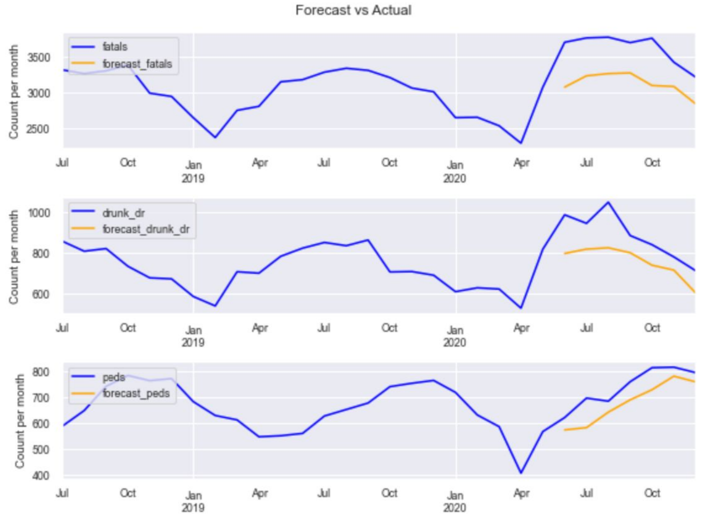




Vector AutoRegression findings

Baseline RMSE			Test RMSE			Baseline MAPE			Test MAPE		
fatals	drunk_dr	peds	fatals	drunk_dr	peds	fatals	drunk_dr	peds	fatals	drunk_dr	peds
762	190	190	503.7	139	67	0.20	0.14	0.23	0.14	0.14	0.08

- Monthly series performed better than weekly and daily, even though it has fewer (12*11=132) observations; test is 5% (7)
- Order of timeseries (number of lags) is13
- Diffs needed to make stationary: Monthly 1, Weekly and daily: 0
- Test RMSE substantially better than baseline RMSE, for all 3 variables
- For fatals, the test RMSE of 503 is 34% better than the baseline of 762
- The Mean Absolute Percent Error (MAPE) - percent error - forecast for fatals is off by 14% in an average month or 503
- Forecast for peds is most accurate: off by 8% or 67 peds per month, in an average month

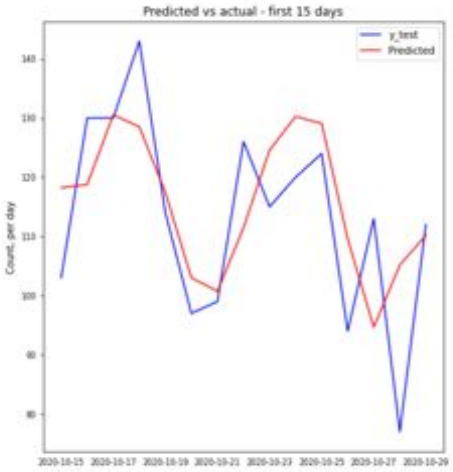


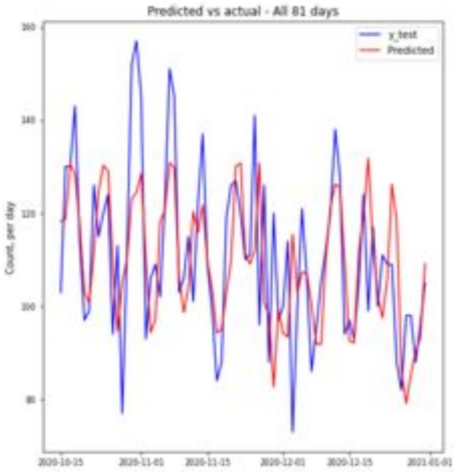
Forecast is pretty accurate, and follows the curve quite well, even though most recent data had covid blips

RNN Findings

Baseline RMSE	Training RMSE	Test RMSE	Baseline MAPE	Training MAPE	Test MAPE
23.25	14.69	15.32	0.15	0.12	0.10

- Daily timeseries spanning 11 years
- Total 4018 observations, validation data 2% ie 81
- Various models - varying sequence of length size
    - LSTM and GRU layers with regularization
    - Different number of hidden layers and nodes and Dropout.
- Surprisingly, the simplest model performed the best!
- The RMSE and MAPE confirm the graphs. Training RMSE substantially better than baseline
- Test RMSE only marginally higher than the Training RMSE





- Predictions are quite good, more so in first few days
- Over 81 days, predictions don't quite match peaks
- Towards the end overshoots the data considerably

Recommendations and Next Steps

- Model is good to go!
- Reduce the 90% drunk driver involved fatalities
  - Sobriety checkpoints and roving patrols between 9pm-3am on weekends and holidays
  - Alcohol Ignition locks

Best Model

- Both models did quite well, though RNN slightly better on fatalities
- VAR forecasts on more than one target; did really good forecasting pedestrian fatalities
- **RNN Best Params:**
  - 2 GRU hidden layers of 16 and 8 nodes
  - Dense hidden layer with 8 nodes, output layer (1 node, relu activation)

Next Steps

Additional areas to attempt:

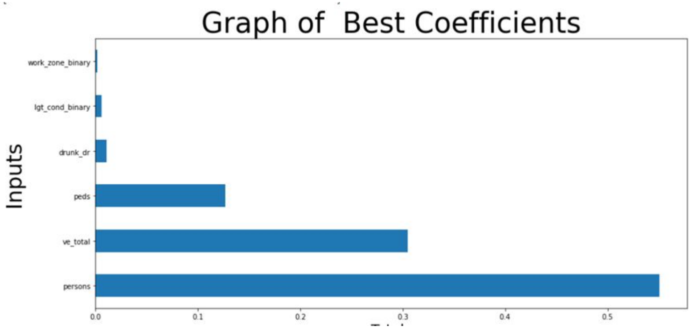
- Try Linear Regression with exogenous variables – weather, distracted, construction zone
- VAR: investigate how to improve weekly and daily forecasts
- RNN: Can try weekly and monthly resampling to compare with VAR. Could try further hyper-parameter tuning

Problem Statement

First Responders want to be better prepared by the time they reach an accident site

- Based on vehicle information, **personal data,** time of day, **environmental factors**, weather conditions, how many fatalities and what level of injury severity to expect?
- What are the worst days/times when they should be appropriately staffed?
- What can be done to reduce the **number of fatalities and severity of injuries?**

Best Features- Injury Severity

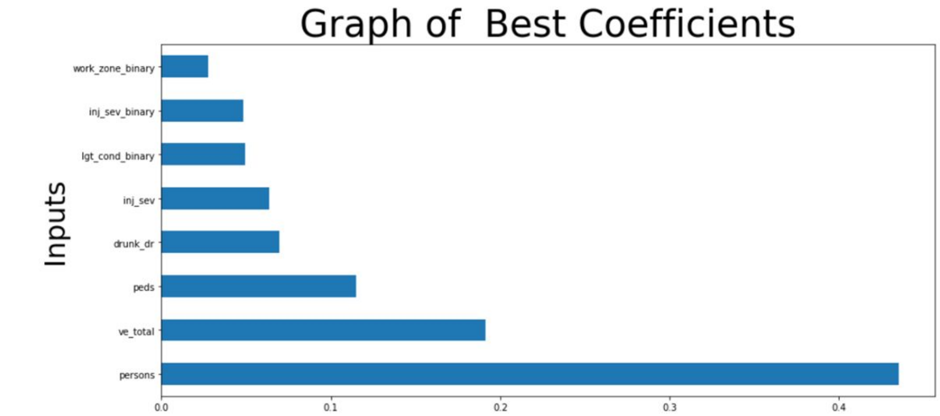


	Specs	Score
0	persons	76386.771434
2	ve_total	22767.693429
1	peds	3093.861243
3	drunk_dr	920.511683
4	lgt_cond_binary	426.749213
5	work_zone_binary	396.459942

- The above chart shows which features do best at predicting injury severity

Best Features- Fatalities

- Persons is the best predictor- to no surprise
- I was very interested to see that work zone and light condition had little to no effect



	Specs	Score
1	persons	26074.760656
0	inj_sev	2927.346604
3	ve_total	2886.277084
2	peds	2084.111283
5	inj_sev_binary	741.583297
4	drunk_dr	157.526996
7	work_zone_binary	14.987448
6	lgt_cond_binary	0.110774

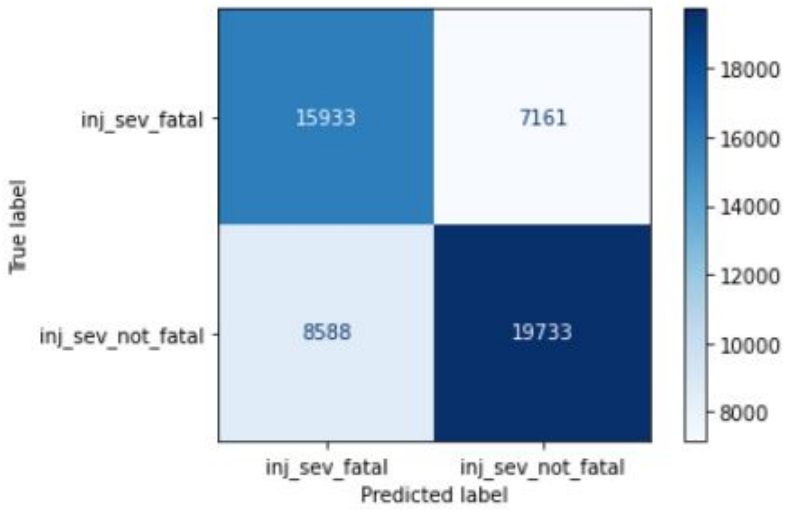
Confusion Matrix

Logistic Regression Score

```
0.6936886122726831
```

Baseline Prediction

```
     0.550826
0    0.449174
Name: inj_sev_binary, dtype: float64
```



Recommendations and Next Steps

- Be over prepared for situations involving multiple cars and many people
- Continue to educate people on the dangers involved in driving while distracted

Problem Statement

First Responders want to be better prepared by the time they reach an accident site

- ➢ Based on vehicle information, **personal data**, time of day, weather conditions, how many fatalities and **what level of injury severity to expect?**
- ➢ What are the worst days/times when they should be appropriately staffed?
- ➢ What can be done to reduce the number of fatalities and **severity of injuries?**

Dataset and Methods

- ➢ Person and driver-level data
- ➢ Motorists only
- ➢ 898,057 records (people) after cleaning
- ➢ Selected 42 features (out of ~70) with modest or high correlation coefficients, all categorical

Preprocessing steps:

- ➢ One-hot encoded all features
- ➢ 255 features went into models

Models evaluated:

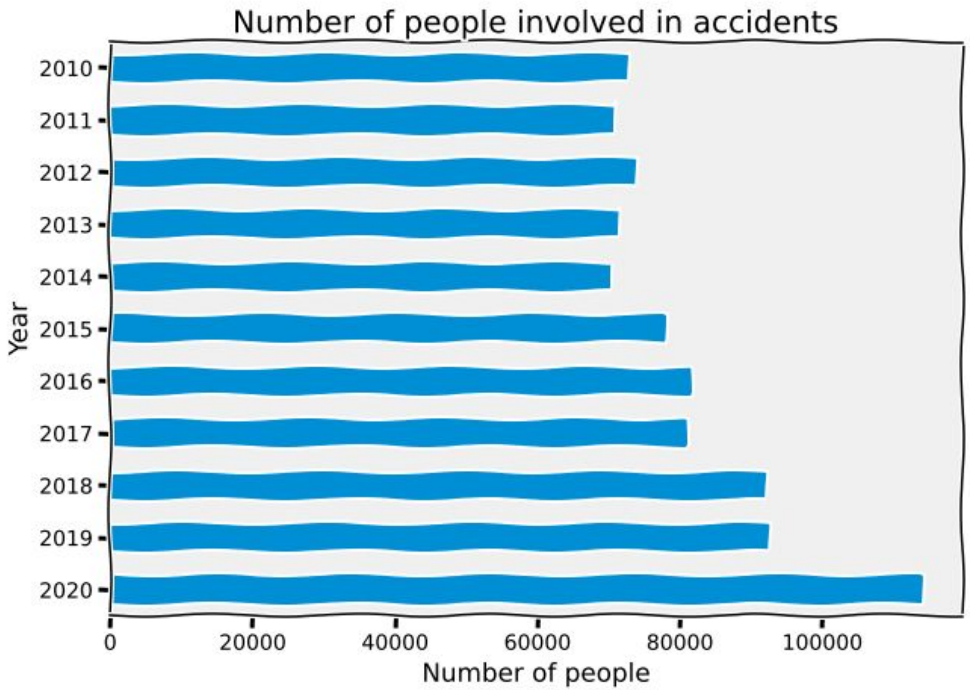
- ➢ Random Forest
- ➢ Logistic regression

Target: *Injury Severity*

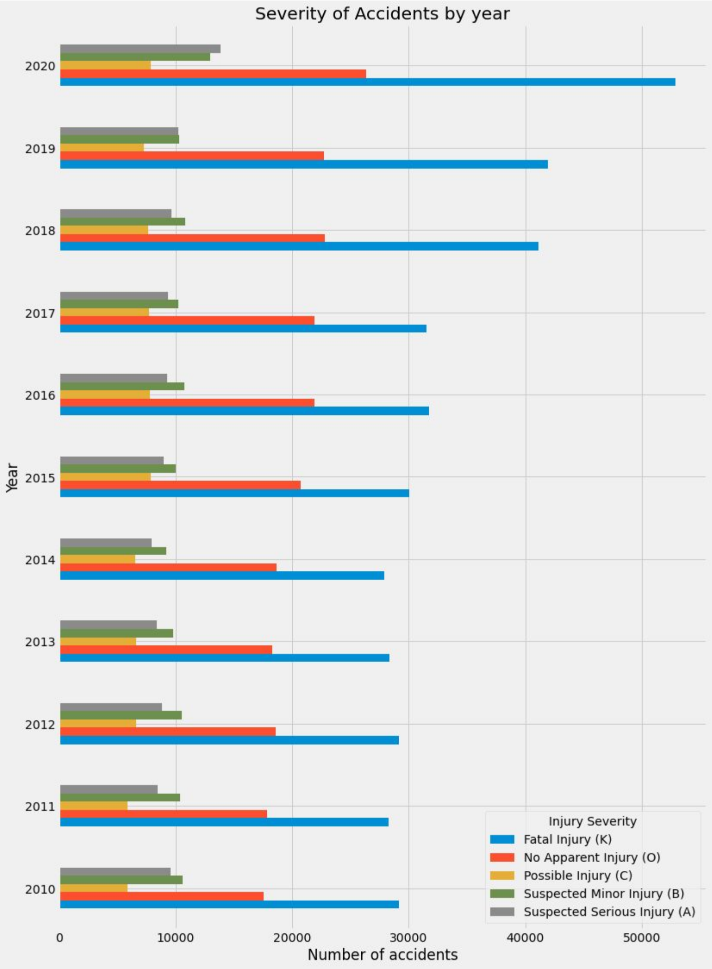
Class of Injury Severity	**% of dataset**
No Apparent Injury (O)	25.3
Possible Injury (C)	8.7
Suspected Minor Injury (B)	12.9
Suspected Serious Injury (A)	11.7
Fatal Injury (K)	***41.4***

Injuries of unknown severity comprised ~1.4% of the dataset, so they were dropped

EDA



Number of people involved in accidents
Year
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
0
20000
40000
60000
80000
100000
Number of people



Severity of Accidents by year
Year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
0
10000
20000
30000
40000
50000
Number of accidents
Injury Severity
Fatal Injury (K)
No Apparent Injury (O)
Possible Injury (C)
Suspected Minor Injury (B)
Suspected Serious Injury (A)

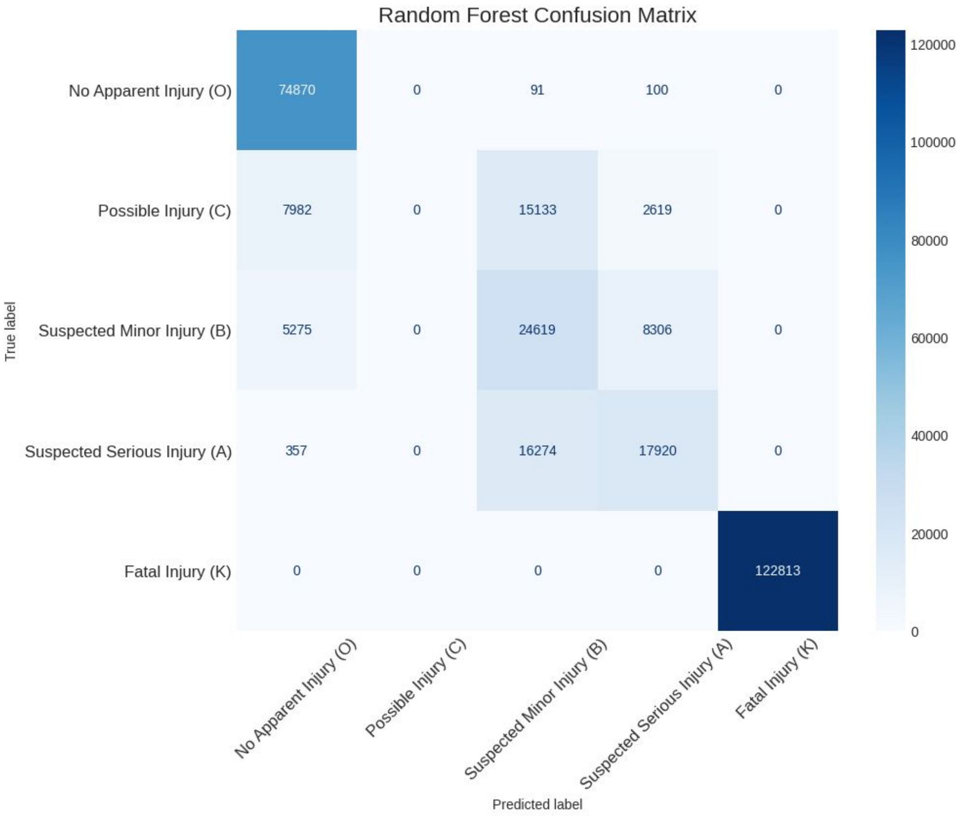
Random Forest model

Best Parameters:

- ➢ ‘ccp_alpha': 0.0005,
- ➢ ‘class_weight': None,
- ➢ ‘max_depth': 12,
- ➢ ‘min_samples_leaf': 7,
- ➢ ‘min_samples_split': 3,
- ➢ ‘n_estimators': 150

Random Forest Training Score: 0.810294
Random Forest Testing Score: 0.809947
Random Forest Categorical Crossentropy: **0.504**

	precision	recall	f1-score	support
No Apparent Injury (O)	0.846142	0.997455	0.915589	75061
Possible Injury (C)	0	0	0	25734
Suspected Minor Injury (B)	0.438708	0.644476	0.522048	38200
Suspected Serious Injury (A)	0.619105	0.518654	0.564445	34551
Fatal Injury (K)	1	1	1	122813
accuracy	0.810578	0.810578	0.810578	0.810578
macro avg	0.580791	0.632117	0.600416	296359
weighted avg	0.757441	0.810578	**0.779401**	296359

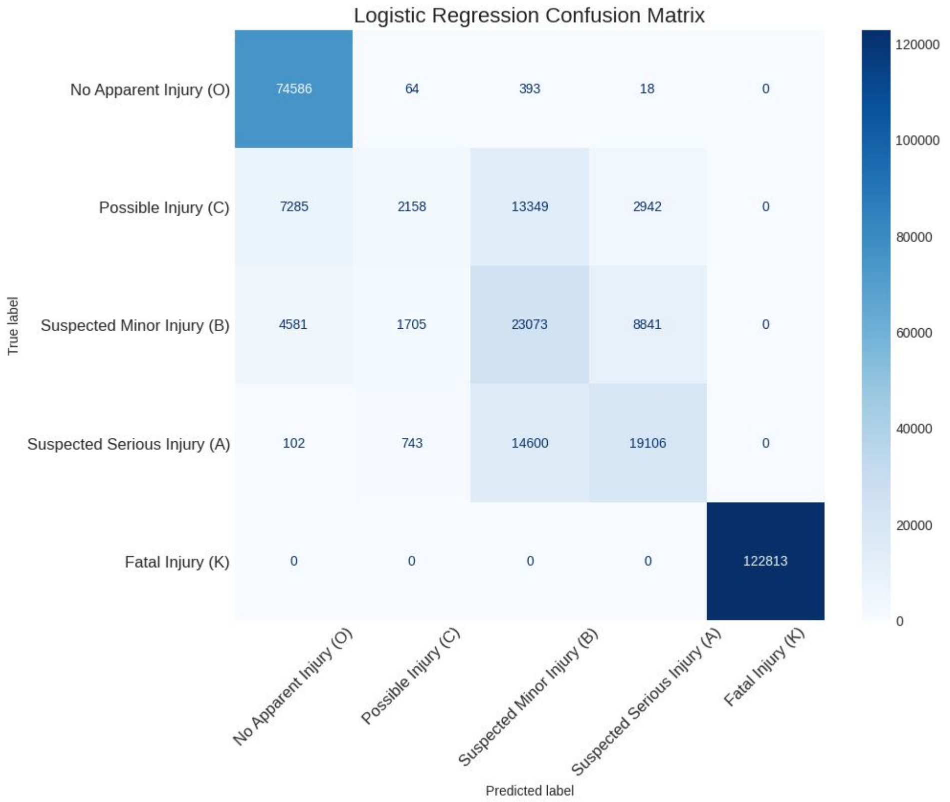


Logistic Regression model

Default parameters used

Logistic Regression Training Score: 0.817
Logistic Regression Testing Score: 0.816
Logistic Regression Categorical Crossentropy: **0.42**

	precision	**recall**	**f1-score**	**support**
No Apparent Injury (O)	0.861728	0.993672	0.923008	75061
Possible Injury (C)	0.462099	0.0838579	0.141955	25734
Suspected Minor Injury (B)	0.44876	0.604005	0.514936	38200
Suspected Serious Injury (A)	0.618177	0.55298	0.583764	34551
Fatal Injury (K)	1	1	1	122813
accuracy	0.815686	0.815686	0.815686	0.815686
macro avg	0.678153	0.646903	0.632733	296359
weighted avg	0.802702	0.815686	**0.794942**	296359



Strongest predictors of walking away unscathed

Feature	Meaning	No Apparent Injury (O)	Possible Injury (C)	Suspected Minor Injury (B)	Suspected Serious Injury (A)	Fatal Injury (K)
hospital_0.0	Not transported for treatment	149.313	0.67819	0.324509	0.0197441	1.54131
driverrf_22.0	Towing or Pushing Vehicle Improperly	4.08169	1.90653	0.431179	0.297971	1.0002
driverrf_86.0	**Skidding, Swerving, Sliding Due To: Pedestrian, Pedalcyclist, or Other Non-Motorist in Road**	**3.35802**	**1.53471**	**0.799223**	**0.244614**	**0.992524**
driverrf_21.0	Overloading or Improper Loading of Vehicle With Passenger or Cargo	2.66481	0.78911	0.729059	0.652211	1.00011
drugspec_98.0	Tested for drugs; drug unknown	2.3586	0.673608	1.02966	0.608543	1.00451
drimpair_7.0	**Blind/Visually Impaired**	**2.32689**	**0.27651**	**9.12449**	**0.170502**	**0.999019**
driverrf_94.0	**Emergency Medical Service Personnel**	**2.14167**	**0.724298**	**0.769555**	**0.838047**	**0.999593**
a_vroll_2.0	**No Rollover**	**2.10241**	**0.820107**	**0.607326**	**0.798639**	**1.19575**
ej_path_8.0	Other Ejection Path (e.g., Back of Pickup Truck)	2.05764	0.706143	1.06587	0.636276	1.01482
driverrf_32.0	Opening Vehicle Closure Into Moving Traffic or Vehicle Is in Motion	1.92235	0.690992	0.758967	0.992008	0.999902
air_bag_20.0	**Air Bag Available but Not Deployed for This Seat**	**1.89635**	**0.938932**	**0.757744**	**0.669538**	**1.107**
driverrf_13.0	Mentally Challenged	1.88872	0.40475	1.1272	1.15787	1.00227
rest_use_11.0	Child Restraint System – Rear Facing (Since 2008)	1.82324	1.39553	0.772464	0.505631	1.00625
driverrf_79.0	Slippery or Loose Surface	1.77508	1.00499	0.719332	0.775003	1.00551
driverrf_47.0	Making Right Turn From Left-Turn Lane or Making Left Turn From Right-Turn Lane	1.7511	0.581335	0.8757	1.12219	0.999633
rest_use_4.0	Child Restraint – Type Unknown	1.61098	1.19518	0.804716	0.627423	1.02867
driverrf_31.0	Starting or Backing Improperly	1.56456	0.844614	0.844712	0.892909	1.00331

Strongest predictors of fatality

Feature	Meaning	No Apparent Injury (O)	Possible Injury (C)	Suspected Minor Injury (B)	Suspected Serious Injury (A)	Fatal Injury (K)
a_hisp_1.0	Non-Hispanic	0.427391	0.606765	0.592438	0.562181	11.578
a_rcat_1.0	Race is White	0.534569	0.678298	0.659116	0.651507	6.42237
a_hisp_3.0	Unknown if Hispanic	0.600514	0.740183	0.746142	0.682295	4.4192
a_hisp_2.0	Hispanic	0.616186	0.78803	0.801751	0.780353	3.29166
a_rcat_8.0	Race unknown	0.716607	0.803955	0.804804	0.742527	2.90459
a_rcat_2.0	Race is Black	0.718879	0.825889	0.828798	0.806056	2.52121
extricat_1.0	**Extricated**	**0.445058**	**0.694154**	**1.14939**	**1.62861**	**1.7292**
a_eject_2.0	**Ejected**	**0.399741**	**0.740168**	**1.14617**	**1.78046**	**1.6562**
a_restuse_2.0	**Unrestrained**	**0.461789**	**1.0391**	**1.10467**	**1.14081**	**1.65368**
a_rcat_7.0	All Other Races	0.803669	0.90835	0.918948	0.90227	1.65212
a_mc_l_s_4.0	Motorcycle License - not applicable	1.39819	0.94826	0.834024	0.569663	1.58749
rest_mis_0.0	No Indication of Restraint Misuse	0.693329	0.98056	1.02081	0.910063	1.58332
hospital_0.0	Not Transported for Treatment	149.313	0.67819	0.324509	0.0197441	1.54131
dstatus_2.0	**Person was tested for drugs**	**0.751833**	**0.909962**	**0.839436**	**1.1765**	**1.48004**
alc_status_2.0	**Person was tested for alcohol**	**0.623199**	**0.913856**	**1.19066**	**1.00061**	**1.47382**

Recommendations and Next Steps

- An LR or RF model can predict whether a person will die in a crash with 100% accuracy, but it can predict the severity of injury that a person would sustain with only ~82% accuracy
- Factors that increase the likelihood of walking away unscathed include: being visually impaired, not rolling over, slippery roads (all likely proxies for *reduced speed*)
- Factors that increase the likelihood of dying include: being ejected from car, being unrestrained, being tested for drugs or alcohol, being extricated (likely proxy for speed)
- Based on these factors, we recommend:
    - Enforcing seat belt laws
    - Enforcing speed limits
    - Implementing checkpoints for sobriety/drugs

Next steps:

- Address high degree of multicollinearity in dataset (great use case for PCA)
- Incorporate more geographic data
- Try XGBoost
- Try neural network

	Random Forest (optimized)	Logistic Regression (defaults)
Categorical Crossentropy	0.50	**0.42**
F1-Score (weighted average)	0.779401	**0.794942**

Problem Statement

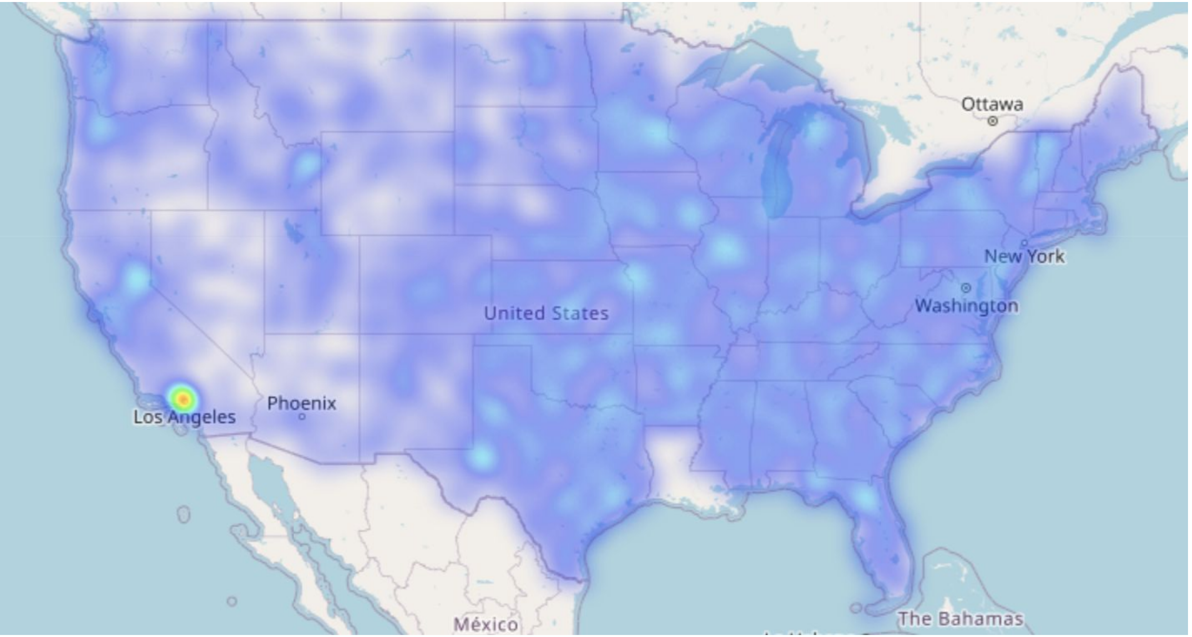
First Responders want to be better prepared by the time they reach an accident site

- ➢ Based on **vehicle information**, personal data, time of day, weather conditions, how many fatalities and **what level of injury severity to expect?**
- ➢ What are the worst days/times when they should be appropriately staffed?
- ➢ What can be done to **reduce** the number of fatalities and **severity of injuries?**

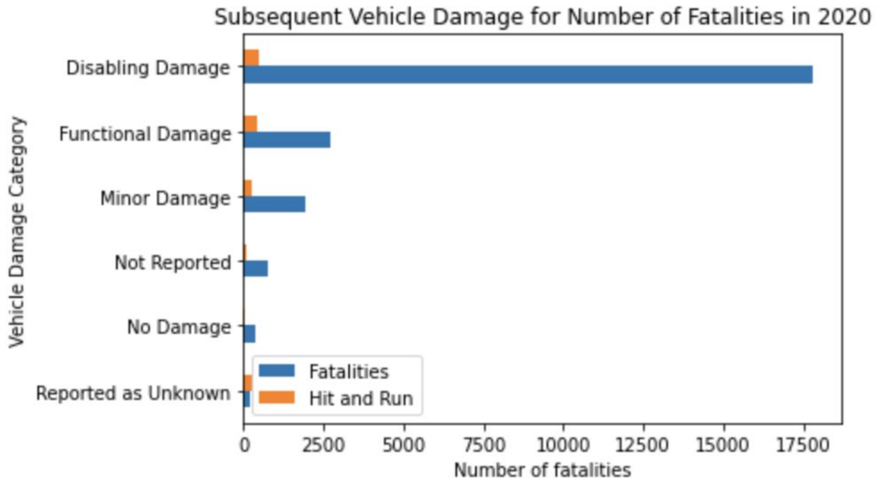
Methodology

- EDA: Vehicle data, top makes/models involved , vehicle damage, dozens of features
- Feature and Preprocessing - binarize target feature, mapping accident data, choosing features
- Model selection: Logistic Regression, Random Forest Classifier
- Logistic Regression
  - How much do each feature lead to fatalities
- Random Forest Classifier
  - Same variables as LR; predict Fatalities
  - Hyper tuning using GridSearch
- Metrics: Accuracy

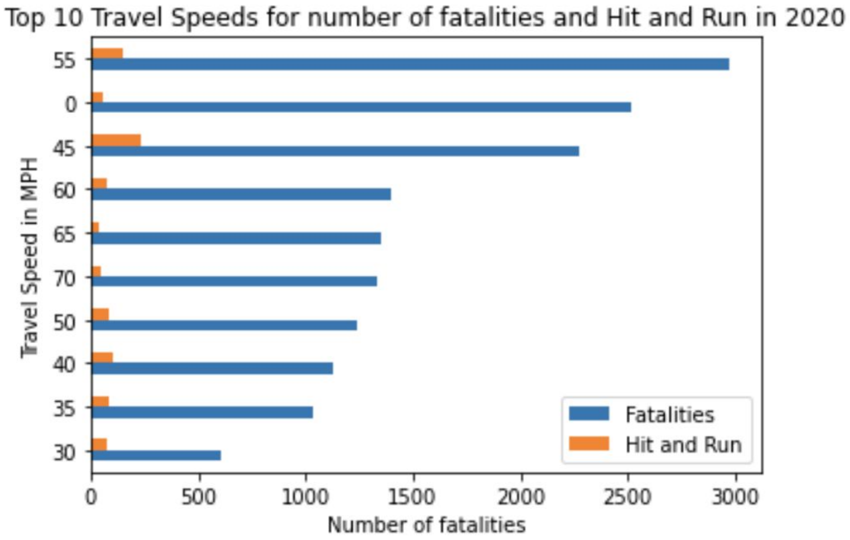
EDA - Heat Map of all Accidents 2020



EDA - Vehicle



Subsequent Vehicle Damage for Number of Fatalities in 2020
Vehicle Damage Category
Disabling Damage
Functional Damage
Minor Damage
Not Reported
No Damage
Reported as Unknown
Fatalities
Hit and Run
0
2500
5000
7500
10000
12500
15000
17500
Number of fatalities



Top 10 Travel Speeds for number of fatalities and Hit and Run in 2020
Travel Speed in MPH
55
0
45
60
65
70
50
40
35
30
Fatalities
Hit and Run
0
500
1000
1500
2000
2500
3000
Number of fatalities

Models

Baseline Score	Log Reg Score (test)	Random Forest(test)
60.70%	66%	70%

Feature Name	Coefs
man_collname_The First Harmful Event was Not a...	0.029669
harm_evname_Tree (Standing Only)	0.139073
rolinloc	0.111207
impact1name_Non-Collision	0.082512
harm_evname_Rollover/Overturn	-0.015460
m_harm	0.082018
harm_evname_Ditch	0.080731
rollovername_Rollover, Tripped by Object/Vehicle	0.010125
vtrafwayname_Two-Way, Not Divided	0.058037
rollover	0.012953
body_typ	0.019498
vpicmodel	0.008724
body_typname_Two Wheel Motorcycle (excluding m...	0.041122
harm_evname_Embankment	0.046947
vpicbodyclass	0.021848
harm_evname_Curb	0.047565

harm_evname_Embankment	0.046947
vpicbodyclass	0.021848
harm_evname_Curb	0.047565
deformed	0.069540
deformedname_Functional Damage	-0.033970
numoccsname_06	-0.147411
deformedname_Minor Damage	-0.041209
man_collname_Sideswipe - Opposite Direction	-0.065887
numoccsname_02	-0.194723
numoccsname_05	-0.183693
man_collname_Sideswipe - Same Direction	-0.078494
vtrafwayname_Two-Way, Divided, Positive Medi...	-0.073186
numoccsname_04	-0.217450
numoccsname_03	-0.235137
impact1name_6 Clock Point	0.005136
man_collname_Front-to-Rear	-0.109931
harm_evname_Motor Vehicle In-Transport	-0.221605

Recommendations and Next Steps

Based on the results, the Random Forest performed better than the lr model although the performance is on par with the baseline score.

- With this, we do know that coefficients like the speed of the vehicles before the accident, make, model, etc. can help us infer the severity of a person's injury
- To help avoid accidents in the future and make the jobs of first responders easier, more enforcements of speed limits with higher patrolling should be in effect in highways that are higher prone to accidents

Next Steps:

- Continue to create and select better features from other sources
- Try to combat the unbalanced class of the target variable by getting accident data where no deaths were involved. This could help us predict non fatal injuries more accurately

Conclusions and Recommendations

- ➢ Better sobriety checkpoints and roving patrols during the hours from 9pm to 3am on Holidays and weekends
- ➢ Police departments to enforce speed limits and increase highway patrol
- ➢ Enforce seat belt laws and implement checkpoints for drugs/alcohol
- ➢ Be over prepared for situations involving multiple vehicles and people

References

- FARS data:
- Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2020: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813254
- Fatalities and Coding and Validation manual: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813251
- Census data: https://www.census.gov/data/datasets/time-series/demo/popest/2020s-counties-total.html