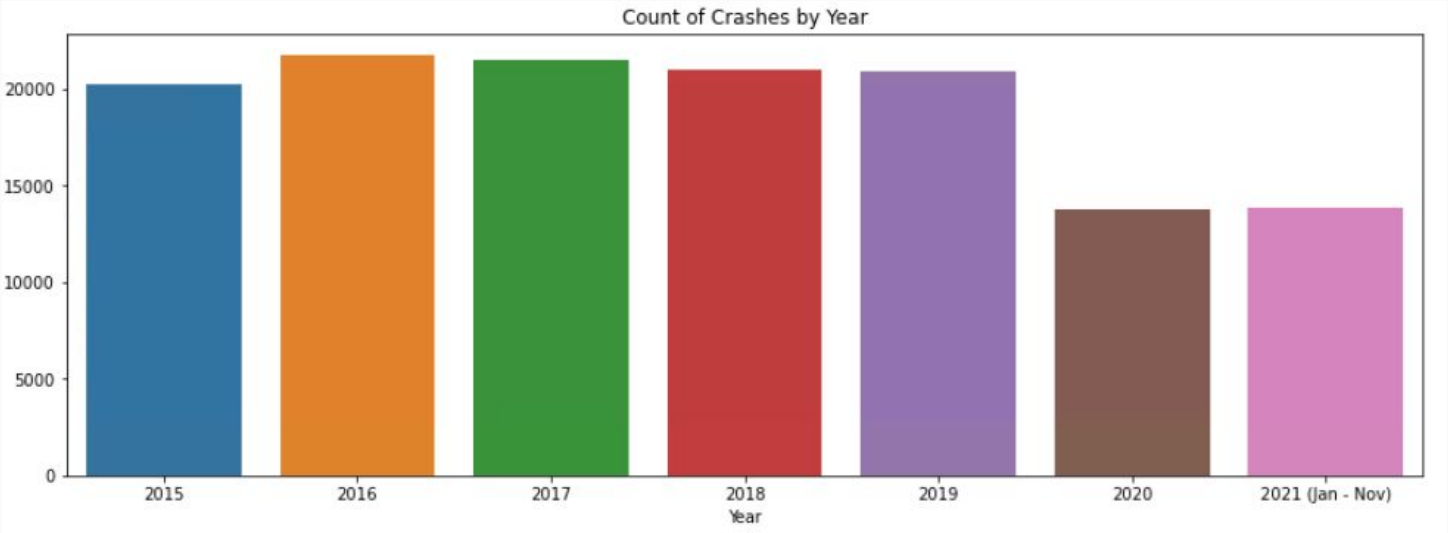
Analysis of Car Crash Data: Montgomery County, Maryland

Rufus Ayeni, Laura Minter, and Jalil Kabbaj

Background (*Presented by Rufus Ayeni*)

From 2015 to 2019, Montgomery Country averaged approximately 21100 automobile crashes per year. For the years 2020 and 2021 (January to November), the number of car crashes is 13800. The sudden drop in crashes is undoubtedly attributable to fewer drivers on the road, because of the pandemic which emerged circa March 2020.



Background (*Presented by Rufus Ayeni*)

Montgomery County officials suspect that the average number of crashes will return to pre-pandemic levels once the county (and country) returns to normal. In fact, despite there being fewer cars on the road, the county (and state) has noticed an uptick in risky driving such as speeding The NHTSA has also noticed a similar uptick despite fewer drivers, nationwide.

Montgomery County has asked us to perform a data analysis of their car crash data to determine if there are any factors that can be mitigated that will reduce the number of crashes and predict reckless drivers.

Problem Statement (*Presented by Rufus Ayeni*)

Can an analysis of Montgomery County crash data yield the following?

- Insights into which factors contribute most to crashes?

- Can features such as speed limit, time of day, and surface condition be used to predict whether a driver is at fault for an automobile crash?

Data Wrangling (*Presented by Rufus Ayeni*)

1. The data was collected from dataMontgomery
   https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632

2. Initial dataset of 134,000 observations from 2015 to 2021 (Jan - Nov)

3. Final dataset of 83,000 observations.

4. We eventually used two datasets for our analysis-- the larger (initial) dataset was used for EDA and the smaller (final) dataset was used for modeling.

Data Wrangling - Feature Selection (*Presented by Rufus Ayeni*)

1. Initial dataset had 43 features.

2. Final dataset has 14 features.

3. Administrative features such as **`Report Number, Local Case Number, Agency Name`**, and **`ACRS Report Type`** were removed, because they don't contribute to either EDA or modeling.

4. Additional driver and accident features such as **`Person ID, Vehicle ID, Driverless Vehicle (0), Vehicle Year, Vehicle Make, Vehicle Model, Latitude, Longitude, Location, Injury Severity, Vehicle Damage Extent`** were deleted, from the modeling dataset, but were kept in the in the EDA dataset.

Data Wrangling - Feature Selection (*Presented by Rufus Ayeni*)

Feature	Null Count
Non-Motorist Substance Abuse	129925
Related Non-Motorist	129063
Municipality	118553
Off-Road Description	120888
Circumstance	108076

Features with high null counts (>100K) were deleted instead of deleting rows or imputing values, because we determined that they didn't contribute to modeling and we wanted to preserve observations as much as possible.

Data Wrangling - Value Imputing (*Presented by Rufus Ayeni*)

Feature	Null Count
Driver Substance Abuse	24277
Traffic Control	20692

With respect to `Driver Substance Abuse` and `Traffic Control`, we decided not to delete the rows. Instead, we imputed the missing values with the existings values of `NONE DETECTED` AND `NO CONTROLS`, respectively.

Reasoning: Substance abuse and traffic controls are fairly significant pieces of information to be omitted. If either value was missing, then we assumed neither existed.

EDA (*Presented by Jalil Kabbaj*)

Balance of Data for Driver At Fault, Driver Substance Abuse, and Driver Distracted:

Figure 1

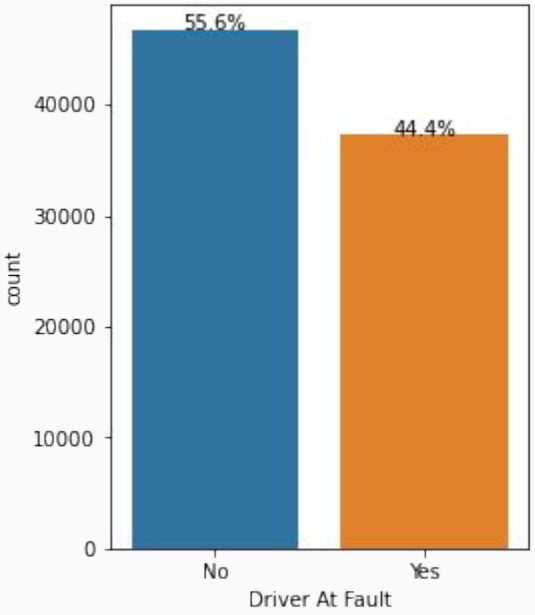


Figure 2

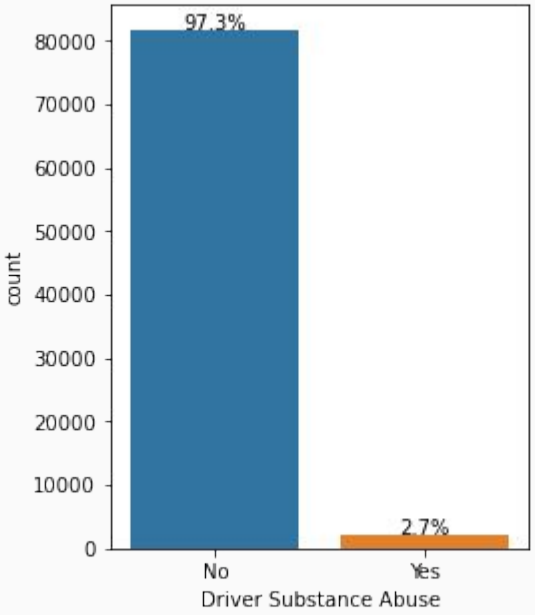
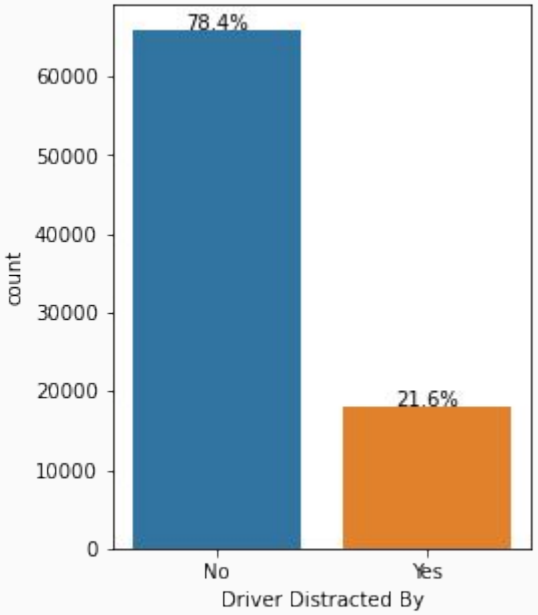


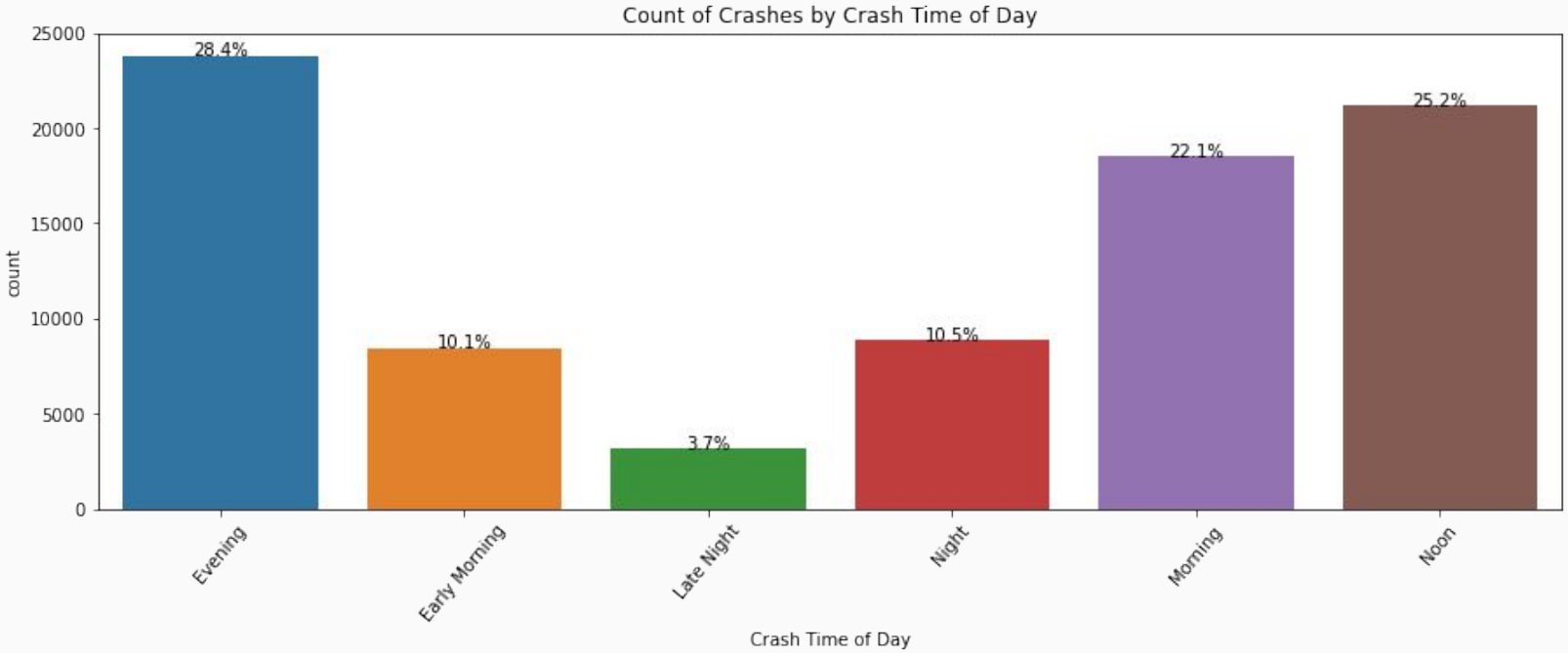
Figure 3



EDA (*Presented by Jalil Kabbaj*)

Distribution of Crashes Amongst 6 Different Times in a Day:

Figure 4



EDA (*Presented by Jalil Kabbaj*)

4 Most Common Vehicles and Severity of Injury:

Table 1

Severity of Injury	**Passenger Car**	**Sports Car**	**Motorcycle**	**Transit Bus**
Fatal Injury	0.10%	0.00%	6.30%	0%
Suspected Serious Injury	0.80%	1.00%	19.90%	0.10%
Suspected Minor Injury	8.50%	8.60%	33.90%	1.00%
Possible Injury	12.80%	13.40%	21.80%	2.20%
No Apparent Injury	77.80%	76.90%	18.10%	96.70%

EDA (*Presented by Jalil Kabbaj*)

Looking at Severity of Injury and Detection of Toxicity:

Table 2

Severity of Injury	**Intoxicated Driver**	**Non-Intoxicated Driver**	**Likelihood of Injury due to Intoxication**
Fatal Injury	1.10%	0.03%	36
Suspected Serious Injury	1.90%	0.82%	2.32
Suspected Minor Injury	7.00%	7.92%	0.88
Possible Injury	7.90%	12.02%	0.65
No Apparent Injury	82.20%	79.20%	1.03

EDA (*Presented by Jalil Kabbaj*)

Accidents by Route Types and Accidents by Surface Condition:

Figure 6

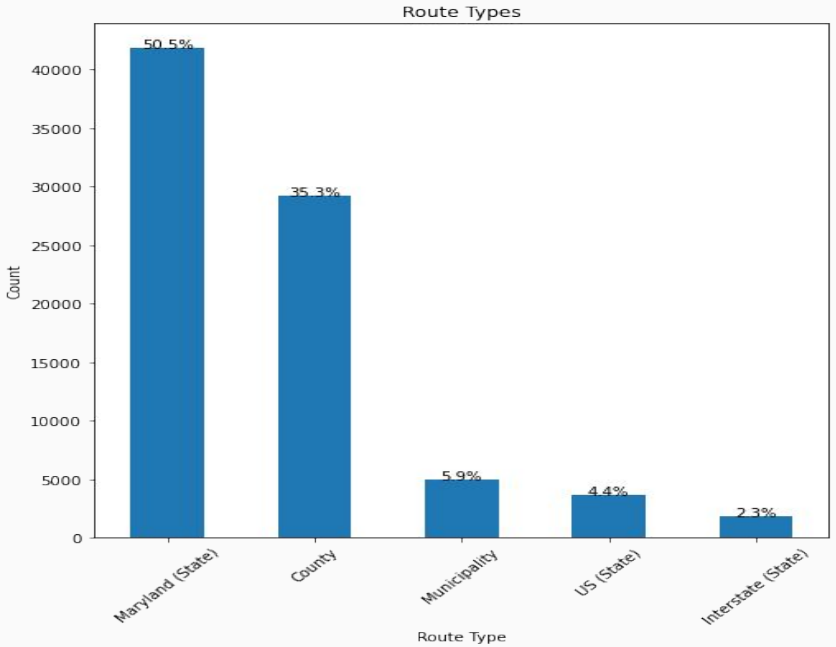
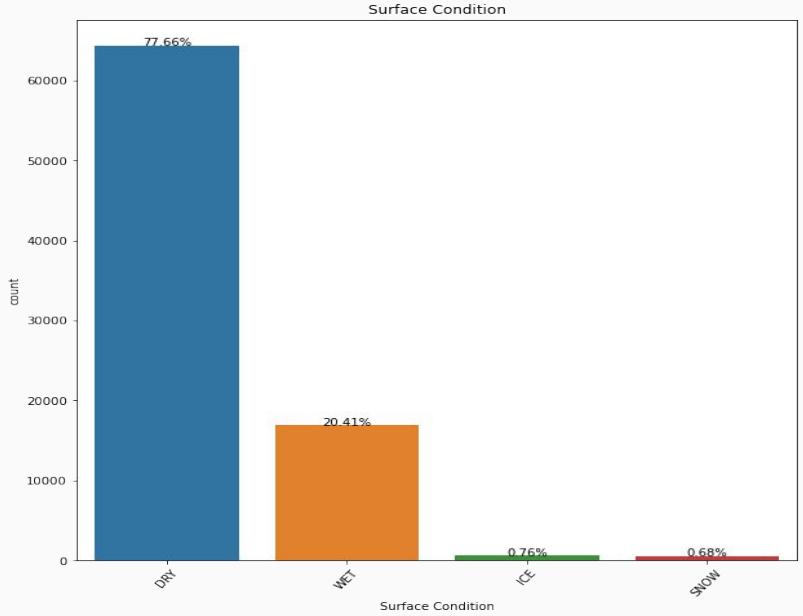


Figure 7



Model selection: driver fault in crashes

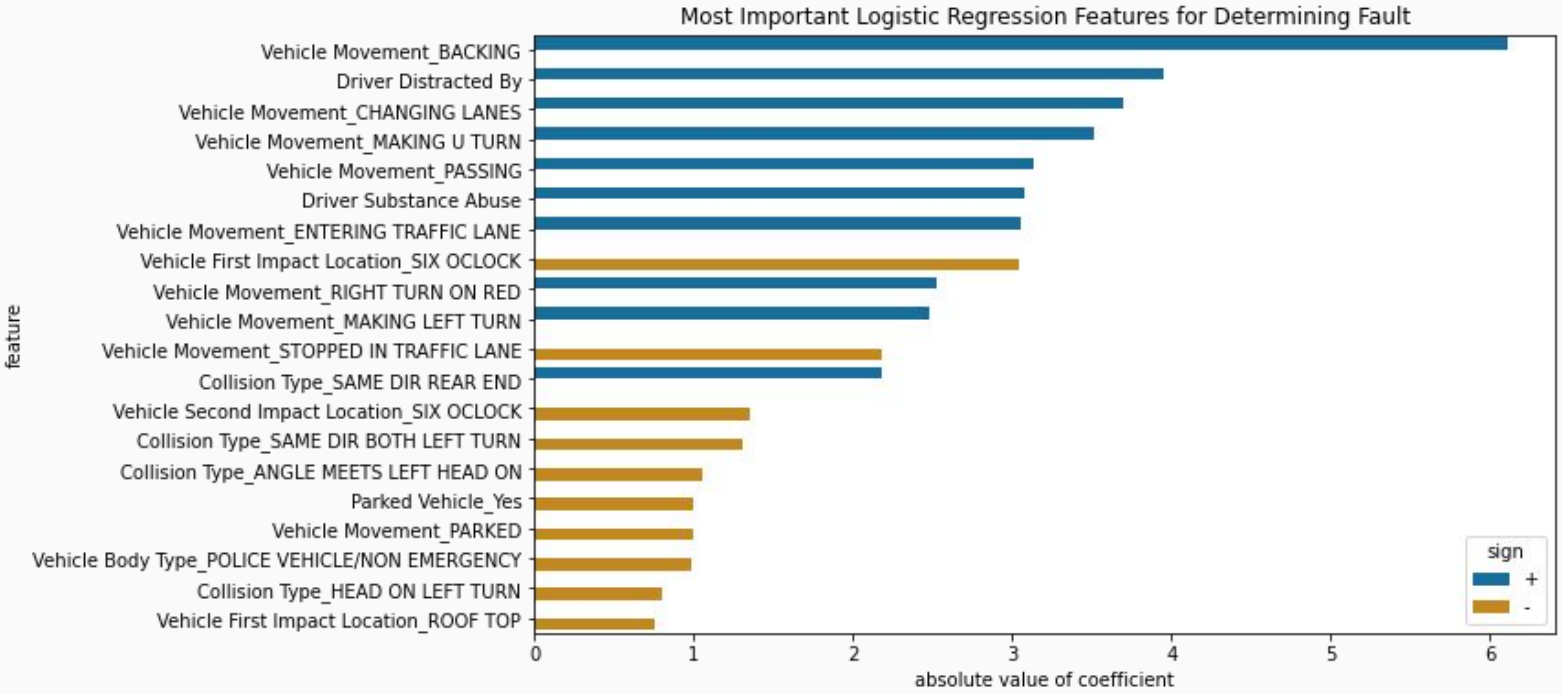
(*Presented by Laura Minter*)

- Model requirements
  - Binary classifier
  - Strong learner
  - Interpretable feature strengths
- Models of interest
  - Logistic Regression
  - Random Forest
  - Null

Model metrics (*Presented by Laura Minter*)

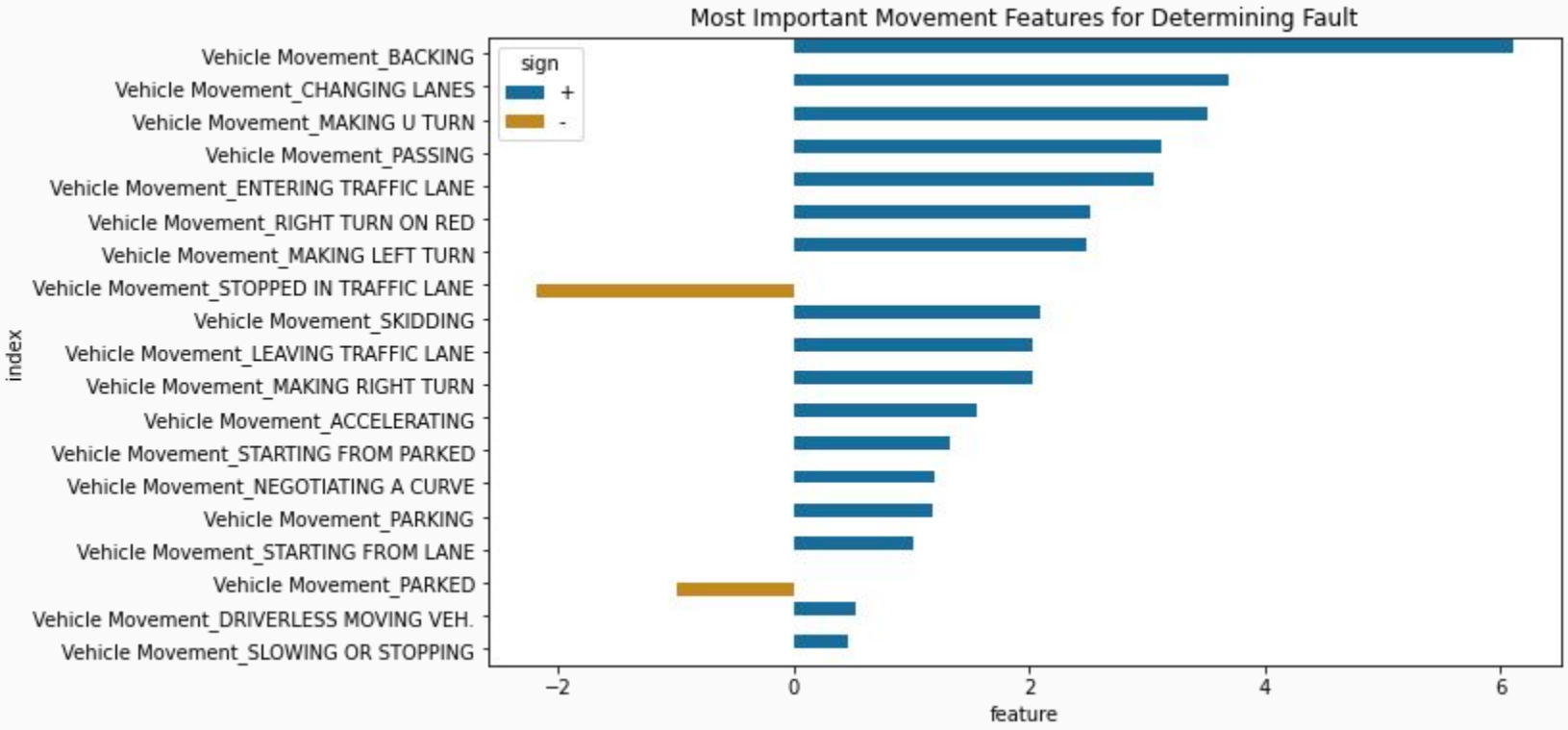
Model	**Parameters**	**Accuracy**		**F1 score**	
		Training	**Testing**	**Training**	**Testing**
Logistic Regression	no penalty max_iterations 1000	0.891	0.894	0.876	0.879
Random Forest Classifier	max_depth 14	0.906	0.900	0.891	0.884
Null	driver never at fault	0.556	0.556	UNDEF	UNDEF

Overall important features (*Presented by Laura Minter*)



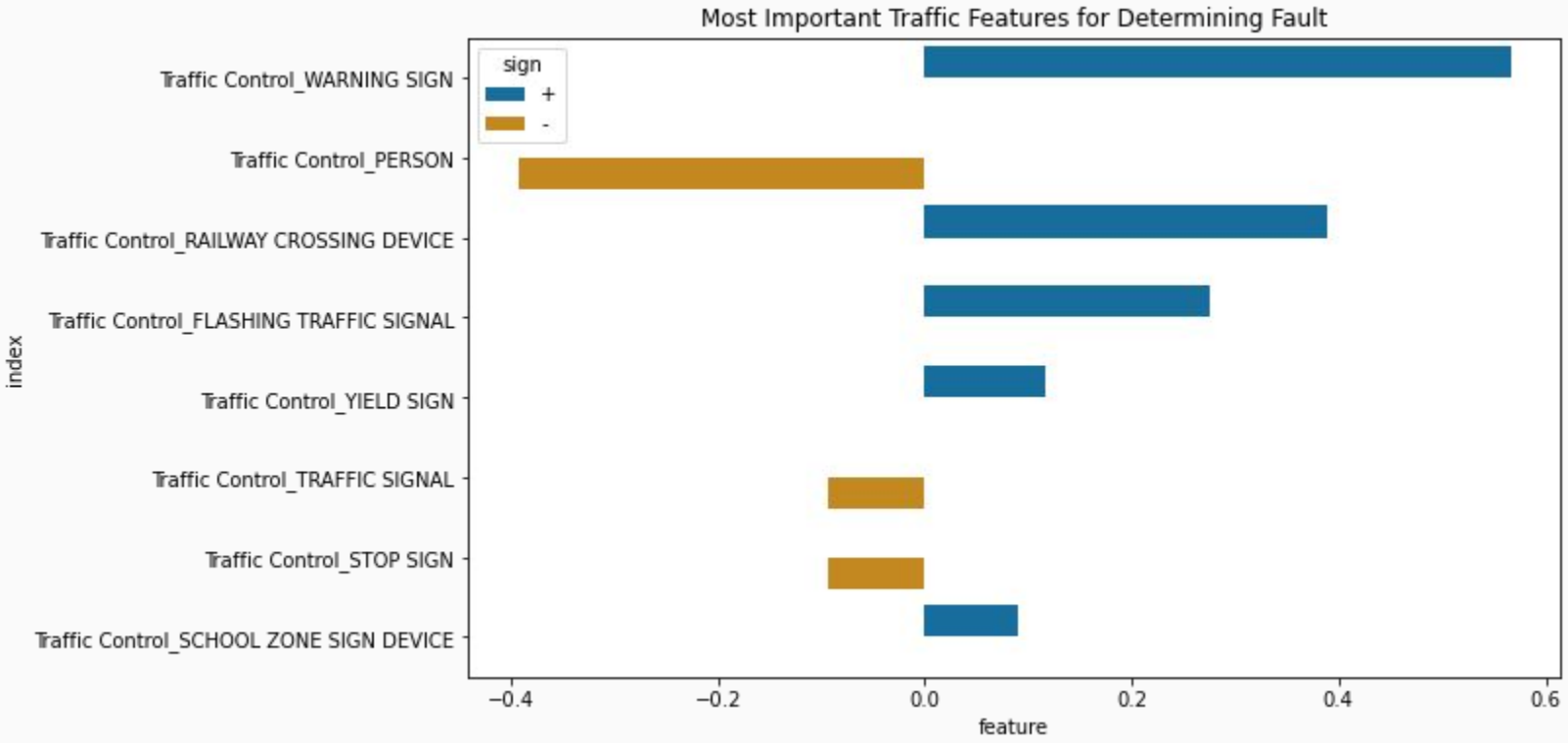
Baseline assumes crash occurred around noon in daylight and dry conditions with no traffic controls present with a car moving at constant speed.

Categorical Highlight 1: Movement (*Presented by Laura Minter*)



Baseline: moving at a constant speed

Categorical Highlight 2: Traffic Control (*Presented by Laura Minter*)



Baseline: no traffic controls

Summary of results (*Presented by Rufus Ayeni*)

- Able to build a model to predict driver fault with 90% accuracy
- Largest contributors to crashes overall
  - Vehicle movement is the largest contributor to crashes (Backing, changing lanes, u-turn)
  - Driver distraction
  - Substance abuse
- Other interesting insights
  - Many crashes happen in broad daylight (noon)
  - Most crashes happen in dry conditions
  - Most crashes occur on state highways (50.5%)

Recommendations (*Presented by Rufus Ayeni*)

To reduce fatalities and serious injuries:

- continue campaigns to address drunk driving and distracted driving
- consider supplementing alternative forms of transportation (e.g., cab rides)

To reduce overall accidents:

- encourage other forms of transportation during daylight (bus, public bikes, scooters)
- increase controls on state highways (‘your speed’ signs, speed traps, traffic controls)
- defensive driving PSAs highlighting the dangers of everyday maneuvers
  (Backing up, u-turns, lane changes, passing and entering traffic)
- increase awareness around traffic controls and consider enforcement mechanisms