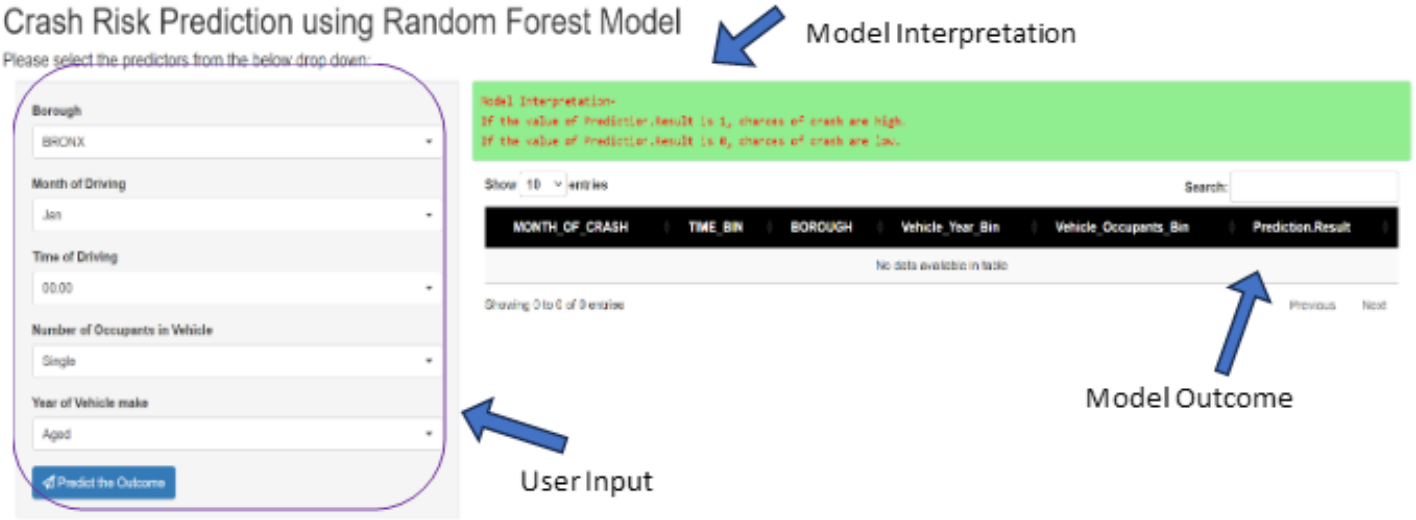
and on hitting the “Predict the Outcome” button it displays the predictive result of Random Forest model in a tabular form. Users can key in multiple entries and the data frame keeps a history of output for a session.



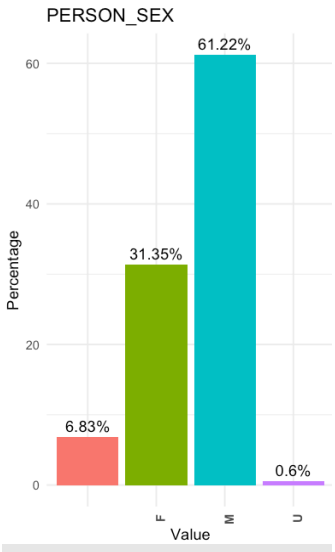
Person dataset

The initial hypothesis based on the Person dataset was:

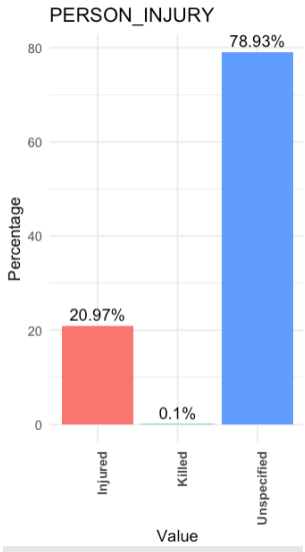
a. Wearing safety equipment could impact mortality in vehicle accidents.

b. Differences exist between person’s sex and age group involved in vehicle collisions respectively.

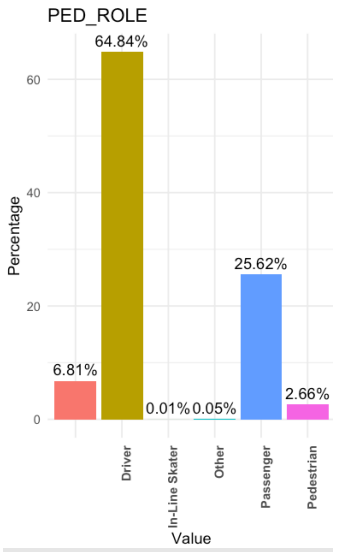
Some histograms and findings shown as below:



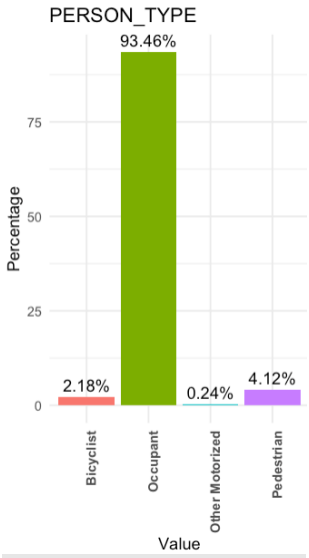
Male has almost twice the probability of involved in accidents as female.



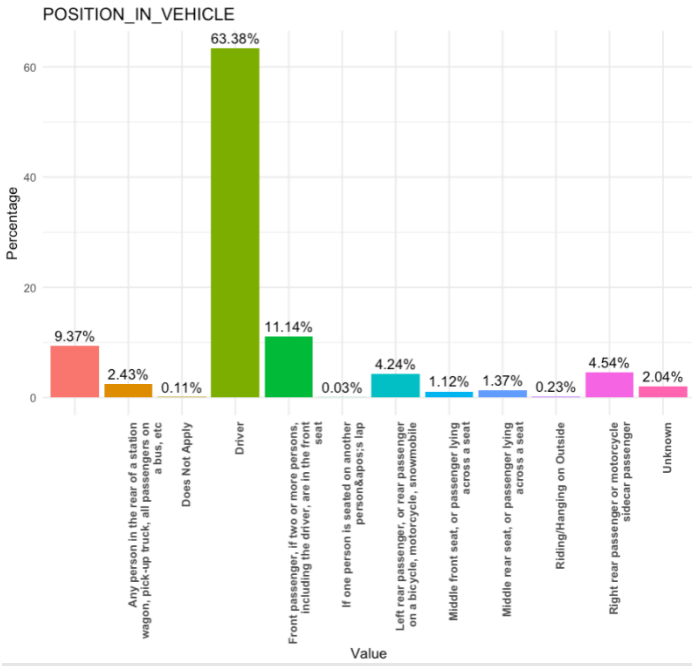
79% data are Unspecified, could be uninjured person.



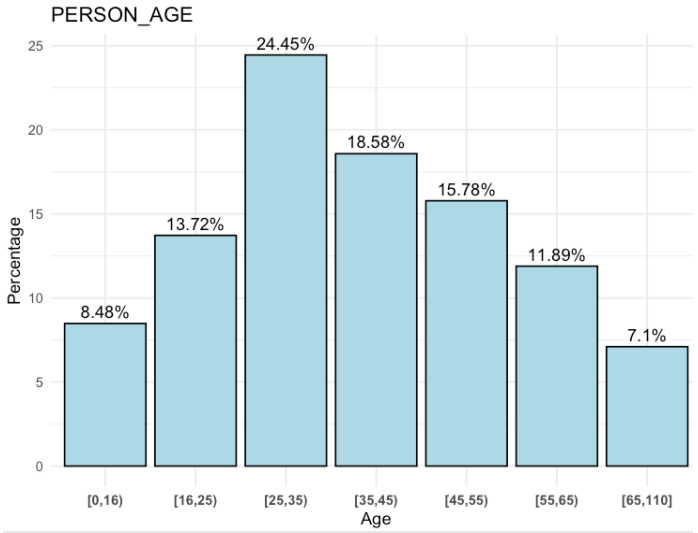
Obviously, drivers and passengers are more involved in accidents.



Occupant includes Driver, Passenger, Other, blank values, In-Line Skater.



Driver is much more involved in accidents; front passenger is more involved than left rear or right rear passengers.

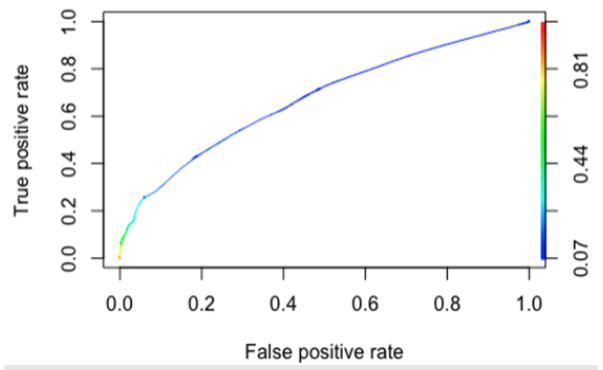
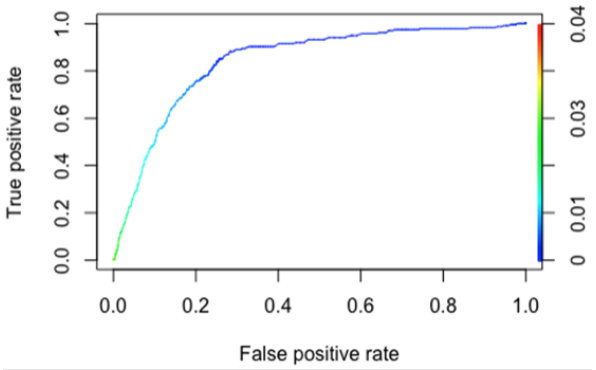


In person dataset, age ranges from -999 to 9999, had to filter to 0-110 (oldest driver in US history).
16-year-old is the legal age to get a permit in NYC.
Age group [25-35) ranks the highest percentage involved in vehicle accidents.

There are multiple areas of cleaning and improvement that the team could focus on. There are Missing/Unknown/ Unspecified/DoesNotApply values that are still existing in some columns, which could cause bias if removed without further analysis and consideration. There might also be a correlation between PED_ROLE, PERSON_TYPE and POSITION_IN_VEHICLE. SAFETY_EQUIPMENT (didn't show here) has several similar values, for example, "Lap Belt & Harness", "Lap Belt", "Harness", "Lap Belt/Harness", will be unified and integrated for future easier modeling.

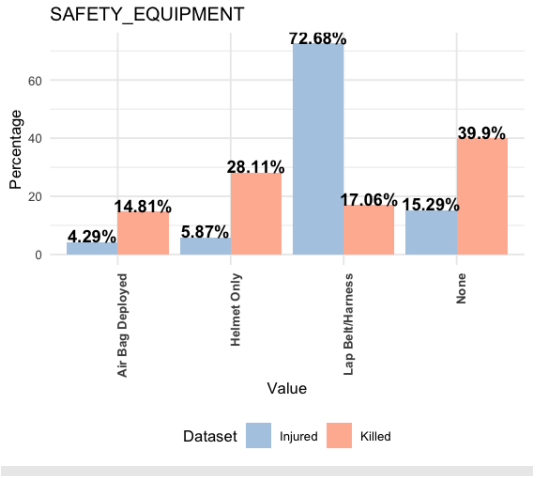
After some modification to values in certain columns, and data filtering, created two datasets for modeling, also split into 70% training set and 30% test set respectively. Create logistic model using training set, make prediction using test set, evaluate models using confusion matrix and ROC Curve. Results as below:

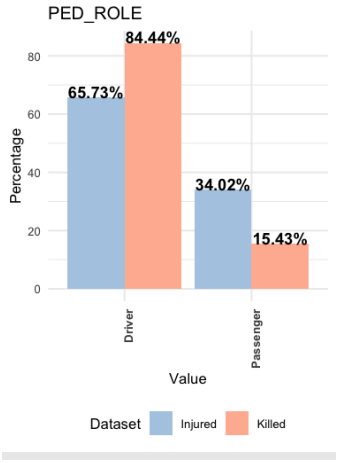
	Model 1		**Model 2**	
Data	all cleaned data		drop PERSON_INJURY == "Unspecified"	
Independent variables	PERSON_AGE, SAFETY_EQUIPMENT, PED_ROLE, PERSON_SEX (PERSON_TYPE & POSITION_IN_VEHICLE was removed due to vif result.)			
Dependent variables	PERSON_INJURY == "Injured" | "Killed" as 1, PERSON_INJURY == "Unspecified" as 0.		PERSON_INJURY == "Killed" as 1, PERSON_INJURY == "Injured" as 0.	
Model summary	All variables are statistically significant.		PED_ROLE isn't statistically significant anymore; Remove PED_ROLE and model again.	
Confusion Matrix	Reference Prediction 0 1 0 415066 43018 1 134534 41382	Accuracy: 0.7199 Sensitivity: 0.49031 Specificity: 0.75521 Balanced Acc: 0.62276	Reference Prediction 0 1 0 63426 41 1 20758 196	Accuracy: 0.7536 Sensitivity: 0.827004 Specificity: 0.753421 Balanced Acc: 0.790213

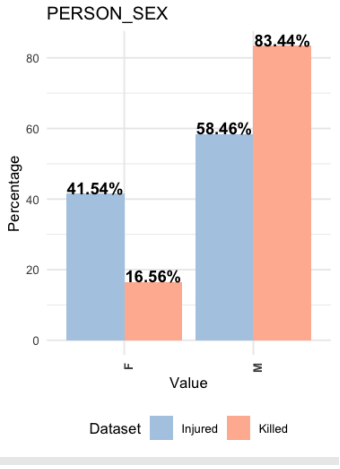
ROC Curve	 AUC value for logistic regression: 0.671097	 AUC value for logistic regression: 0.84247

Both models' accuracy at predicting PERSON_INJURY Is higher than 70%, but threshold did need to be chosen carefully.

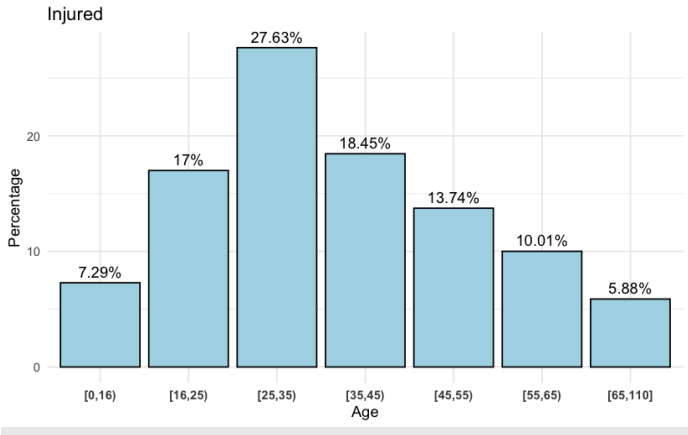
Combining with some histograms between Injured and Killed in model 2's data:

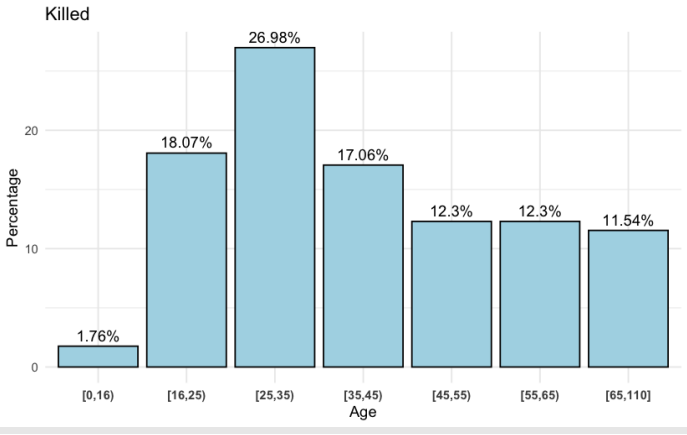






The percentage of Lap Belt/Harness dropped from 73% to 17%, none safety equipment jumped from 15% to 40%; the percentage of male driver in Killed is higher than in Injured.





Age group [25-35) ranks the highest percentage in both Injured and Killed in vehicle accidents.

Model2 shows the evidence that SAFETY_EQUIPMENT did play a big role between life and death in vehicle accidents. Age and Sex also influence the results, male driver is at higher risk of involved in vehicle accidents, also even higher percentage in Killed than in Injured in total accidents.