Traffic Collisions

Diego Fondevilla, Julianna Larios, Karina Bik

Project objective:

Help insurance companies set better premiums and improve emergency response by predicting crash severity, injury, and fault.

The Raw Data

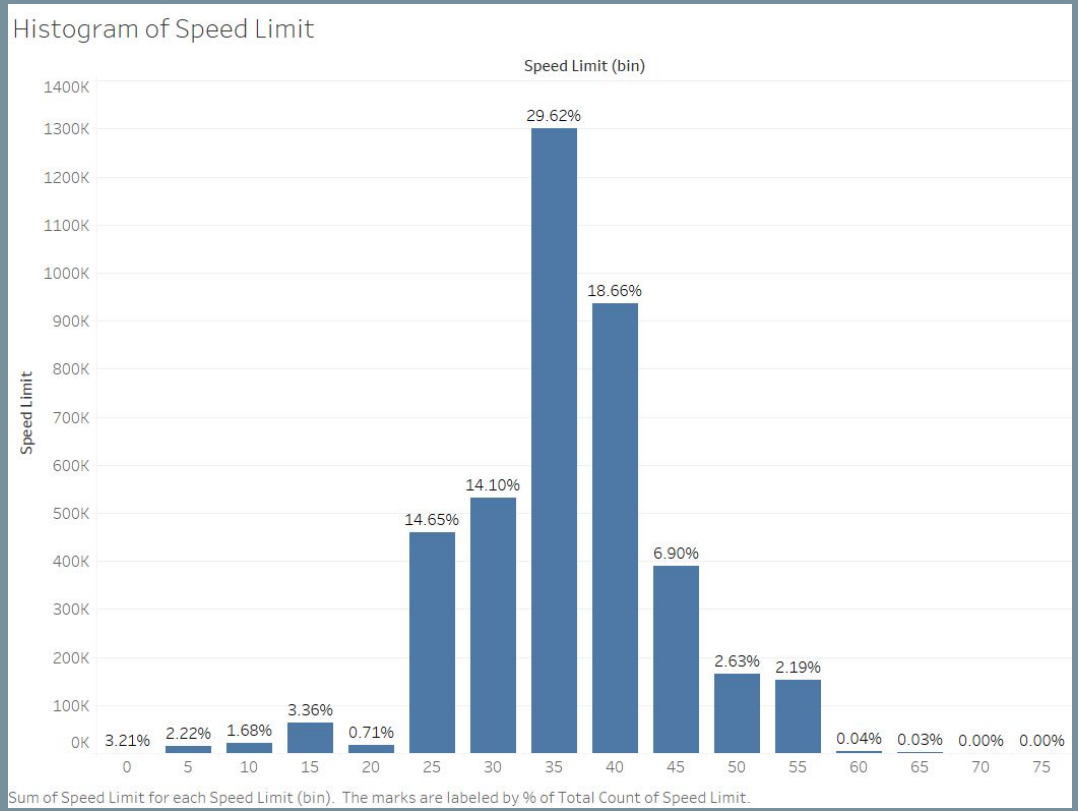
- Originated from data.gov
- Contains data on Motor Vehicles involved in traffic collisions within Montgomery County
- Collected via the automated Crash Reporting System (ACRS)
- Total number of variables:
  - 39
- 100,000+ rows (after cleaning)

Data Cleaning Process

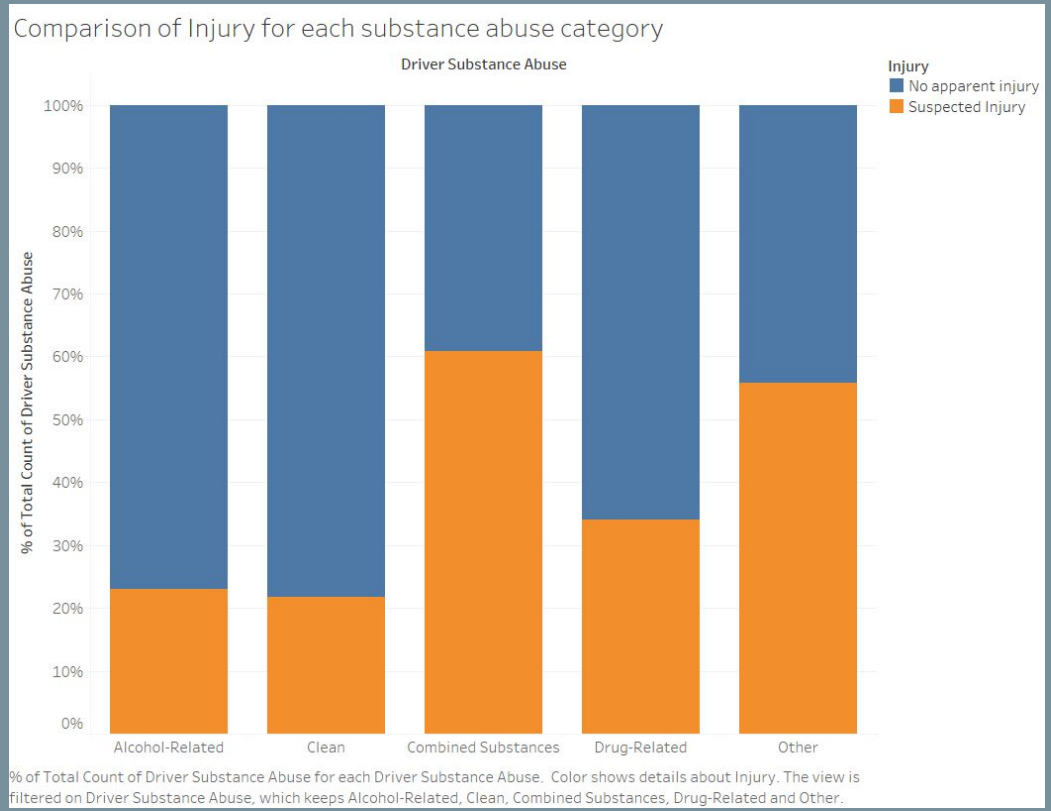
- Created a new dataset using only variables we need
- Removed rows with empty cells, duplicates and have a value of “N/A”
- Factored categorical variables
- Merged factor levels that have similar words

Summary Plots

Summary Plot 1



Summary Plot 2



Models

Logistic Regression

Outcome Variable: **Suspected Injury**

Predictor Variables: Weather, Light, Speed Limit, Collision Type, Driver Distracted by, Driver Substance Abuse

Decision Tree

Outcome Variable: **Crash Severity**

Predictor Variables: Weather, Surface Condition, Vehicle Body Type, Speed Limit, Driver Substance Abuse

Random Forest

Outcome Variable: **Driver At Fault**

Predictor Variables: Collision Type, Weather, Light, Traffic Control, Driver Substance Abuse, Driver Distracted By, Vehicle Damage Extent, Vehicle Body Type, Vehicle Movement, Speed Limit

Model 1: Logistic Model

Variables Explained

- Outcome: Injury
- Predictor Variables
  - Weather
  - Light
  - Speed Limit
  - Collision Type
  - Driver Distracted by
  - Driver Substance abuse

Additional Data Manipulation (Stratified Data Sampling)

```
Suspected Injury No apparent injury
           35257                90259
```

```
sampled_crash <- crash_clean %>%
  group_by(Injury) %>%
  sample_n(35000)
```

- Originally, the proportion between suspected injury vs no apparent injury was too big
- Use stratified sampling to make the proportion even
- Made sure testing and training set didn't have the same issue

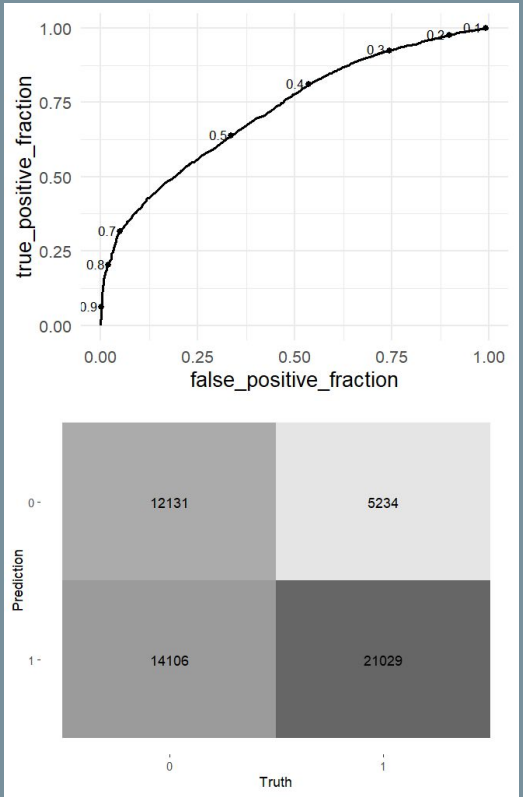
Model Coefficients

	Pr(>|z|)	
(Intercept)	0.000000946791679	***
WeatherCloudy	0.62107	
WeatherFog, Smog, Smoke	0.35179	
WeatherRain	0.64365	
WeatherSnow	0.01517	*
WeatherOther	0.000001049168015	***
LightTwilight	0.63405	
LightDAYLIGHT	< 2e-16	***
LightUNKNOWN	0.000001418888331	***
`Speed Limit`	< 2e-16	***
`Collision Type`HEAD ON	0.000000000219058	***
`Collision Type`Front to Rear	< 2e-16	***
`Collision Type`Sideswipe	< 2e-16	***
`Collision Type`SINGLE VEHICLE	0.09900	.
`Collision Type`Other	< 2e-16	***
`Driver Substance Abuse`Combined Substances	0.000000000000258	***
`Driver Substance Abuse`Drug-Related	0.06431	.
`Driver Substance Abuse`Clean	< 2e-16	***
`Driver Substance Abuse`Other	< 2e-16	***
`Driver Distracted By`External Distractions	0.27984	
`Driver Distracted By`Inattention to surroundings	0.00619	**
`Driver Distracted By`Not Distracted	0.000040301787559	***
`Driver Distracted By`Unknown	0.20895	
`Driver Distracted By`Other	0.38188	

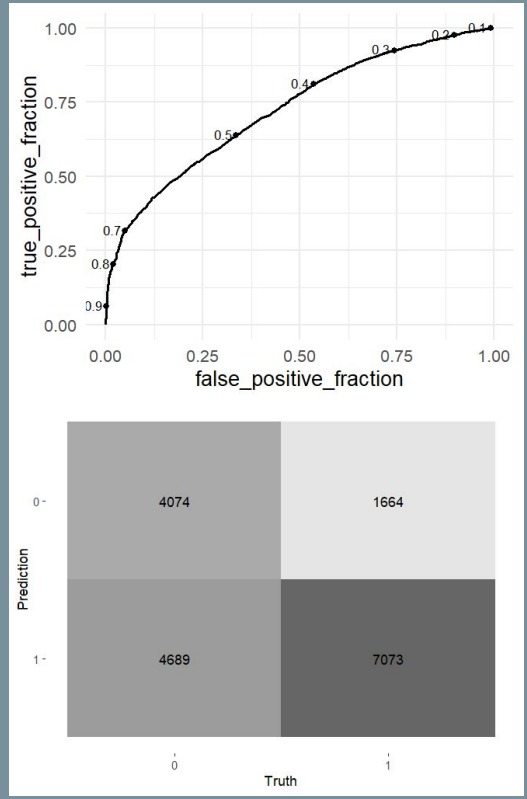
```
> exp(log_model$coefficients)
                                                  (Intercept)
                                                    0.4828998
                                        WeatherFog, Smog, Smoke
                                                    0.8726557
                                                  WeatherSnow
                                                    0.7819564
                                                LightTwilight
                                                    0.9756543
                                                 LightUNKNOWN
                                                    0.4709954
                                       `Collision Type`HEAD ON
                                                    1.2529471
                                      `Collision Type`Sideswipe
                                                    0.3381375
                                        `Collision Type`Other
                                                    0.3433432
                                `Driver Substance Abuse`Drug-Related
                                                    1.3090916
                                   `Driver Substance Abuse`Other
                                                    4.5340789
`Driver Distracted By`Inattention to surroundings
                                                    0.6829263
                                `Driver Distracted By`Unknown
                                                    0.8405348
                                                 WeatherCloudy
                                                    1.0154985
                                                   WeatherRain
                                                    0.9866284
                                                  WeatherOther
                                                    0.6119165
                                                 LightDAYLIGHT
                                                    1.2157336
                                                 `Speed Limit`
                                                    1.0233292
                                     `Collision Type`Front to Rear
                                                    0.6519697
                                    `Collision Type`SINGLE VEHICLE
                                                    0.9426153
                          `Driver Substance Abuse`Combined Substances
                                                    9.5925514
                                  `Driver Substance Abuse`Clean
                                                    0.6097200
                      `Driver Distracted By`External Distractions
                                                    0.8124966
                          `Driver Distracted By`Not Distracted
                                                    1.7548128
                                  `Driver Distracted By`Other
                                                    1.1359474
```

Confusion Matrix + ROC Plot

Training Data:



Testing Data:



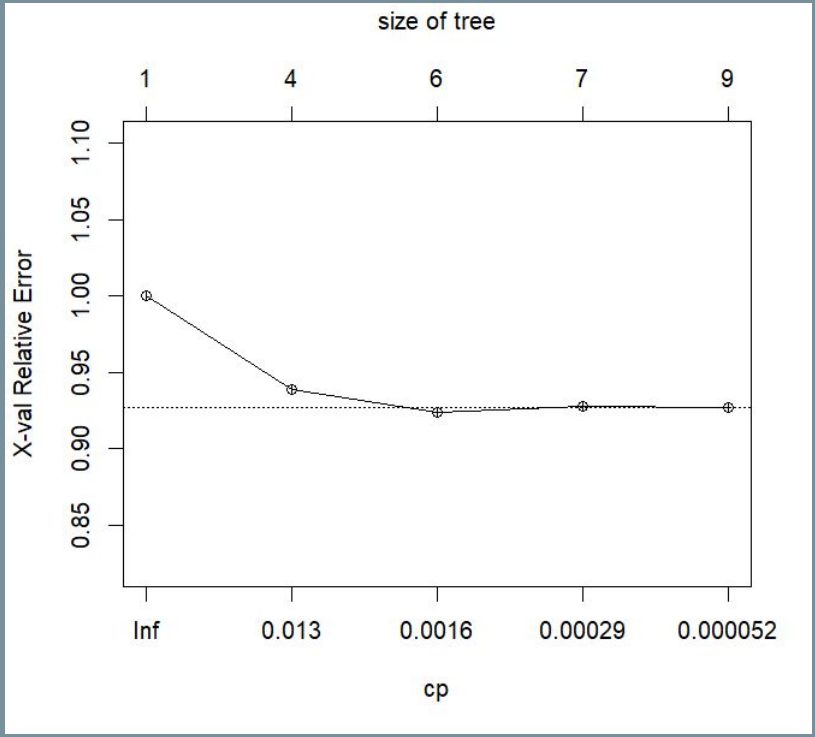
Cutoff: 0.4

- Ensures more true positives in exchange for more false positives

Model 2: Decision Tree

Variables Explained

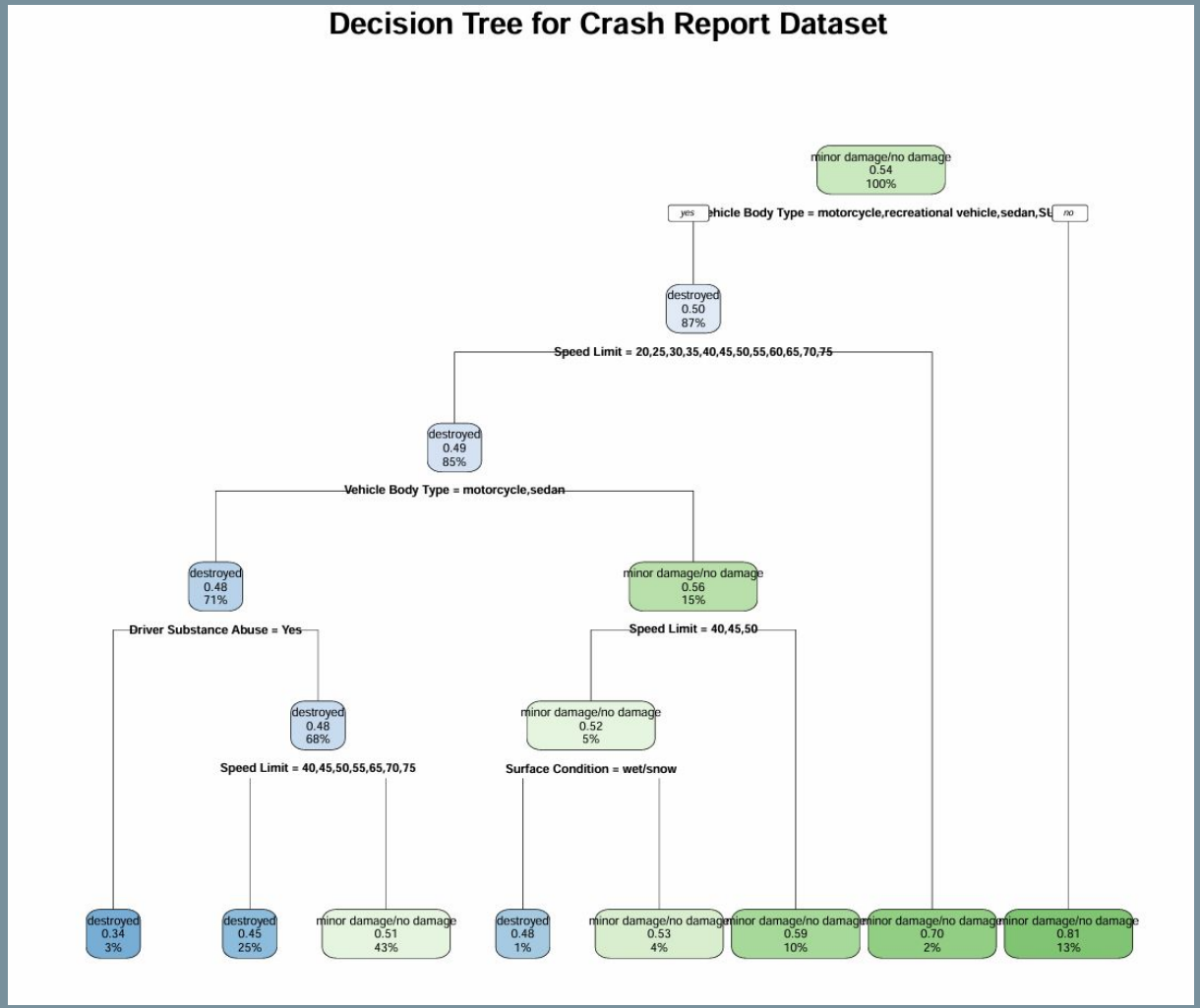
- Outcome: Crash Severity
- Predictor Variables
  - Weather
  - Surface Condition
  - Vehicle Body Type
  - Speed Limit
  - Driver Substance Abuse



	CP	nsplit	rel error	xerror	xstd
1	0.0203146199	0	1.0000000	1.0000000	0.003144901
2	0.0077554912	3	0.9390561	0.9390561	0.003125843
3	0.0003116864	5	0.9235452	0.9238018	0.003119482
4	0.0002658502	6	0.9232335	0.9277804	0.003121203
5	0.0000100000	8	0.9227018	0.9268270	0.003120795

- minsplit/minbucket:
  - since the dataset is so large, I choose a minsplit of 100 and a minbucket of 1000
- cp:
  - xerror starts to stabilize around nsplit = 5, so I used a cp of 0.00001
- maxdepth:
  - to avoid a very complex, deep tree I used a maxdepth of 5

Decision Tree for Crash Report Dataset



Confusion Matrix for the Test Set:

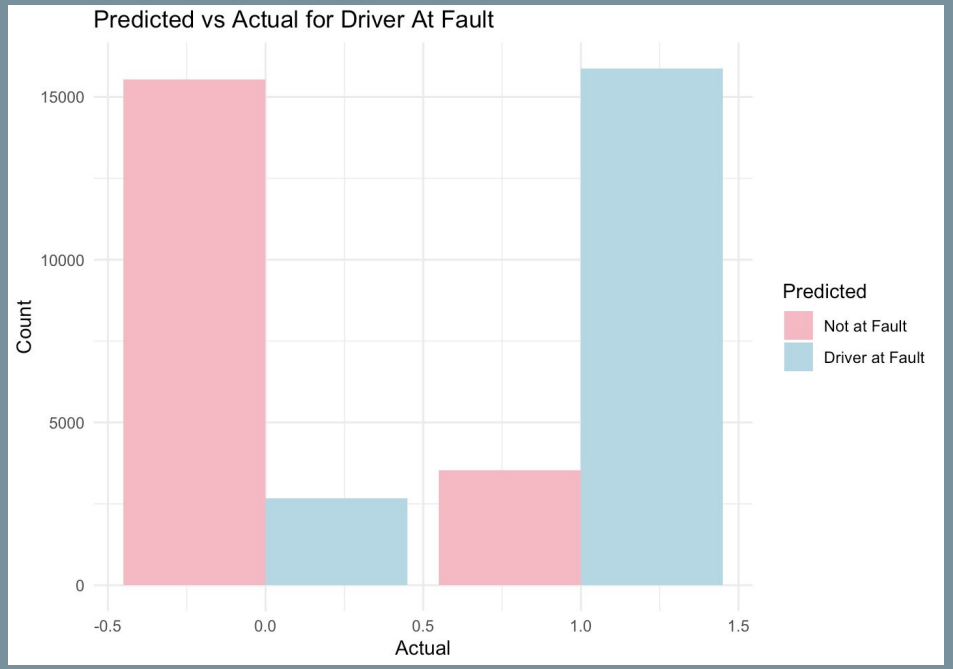
	Reference	
Prediction	destroyed	minor damage/no damage
destroyed	4655	3602
minor damage/no damage	8743	11755

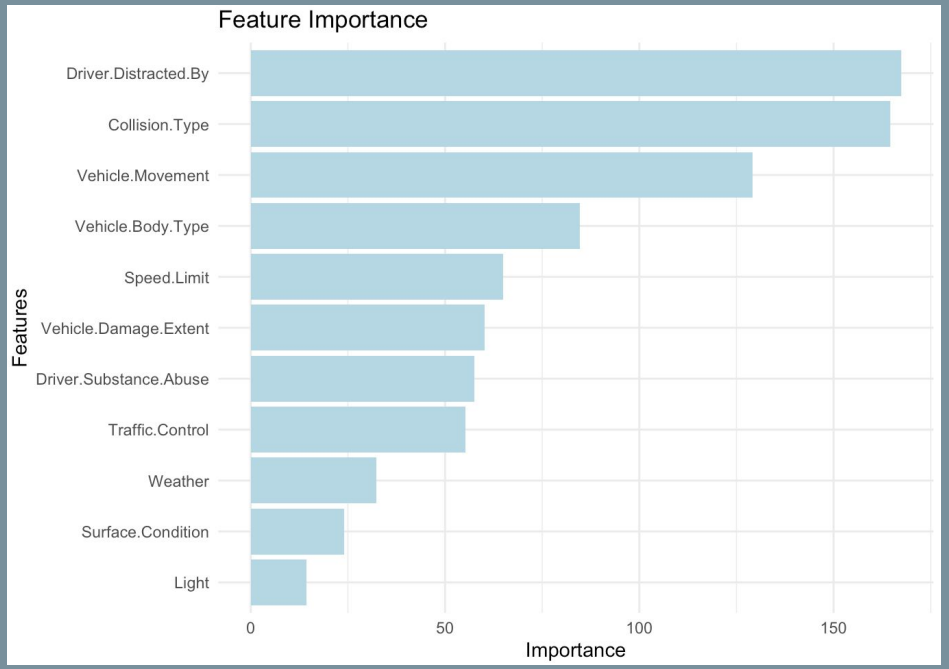
Model 3: Random Forest

Variables Explained

- Outcome: Driver At Fault
- Predictor Variables
  - Collision Type
  - Weather
  - Light
  - Traffic Control
  - Driver Substance Abuse
  - Driver Distracted By
  - Vehicle Damage Extent
  - Vehicle Body Type
  - Vehicle Movement
  - Speed Limit

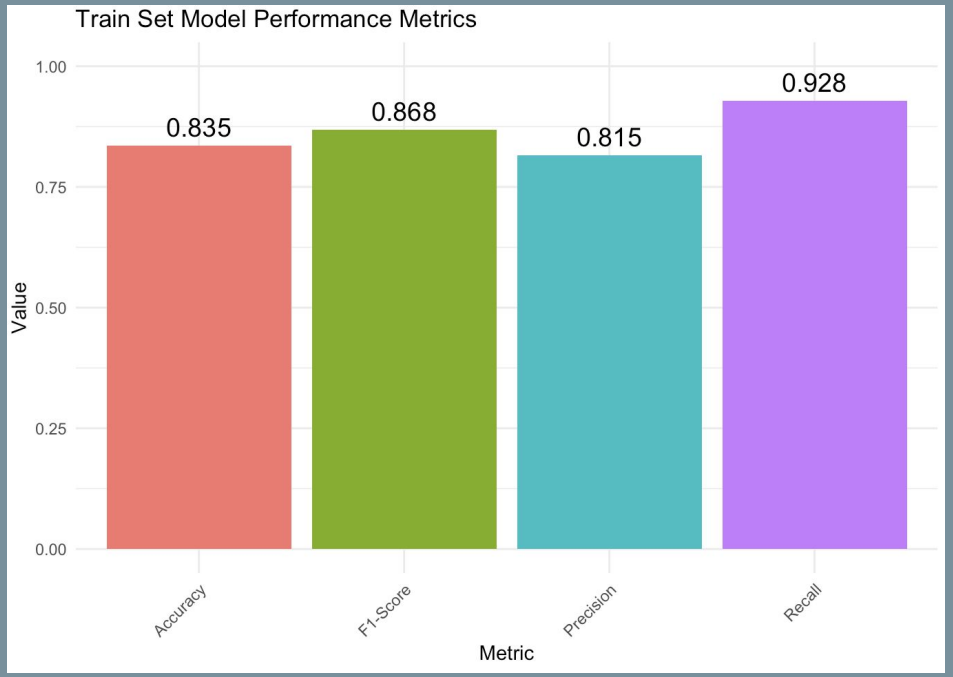
Predicted vs Actual Plot + Feature Importance

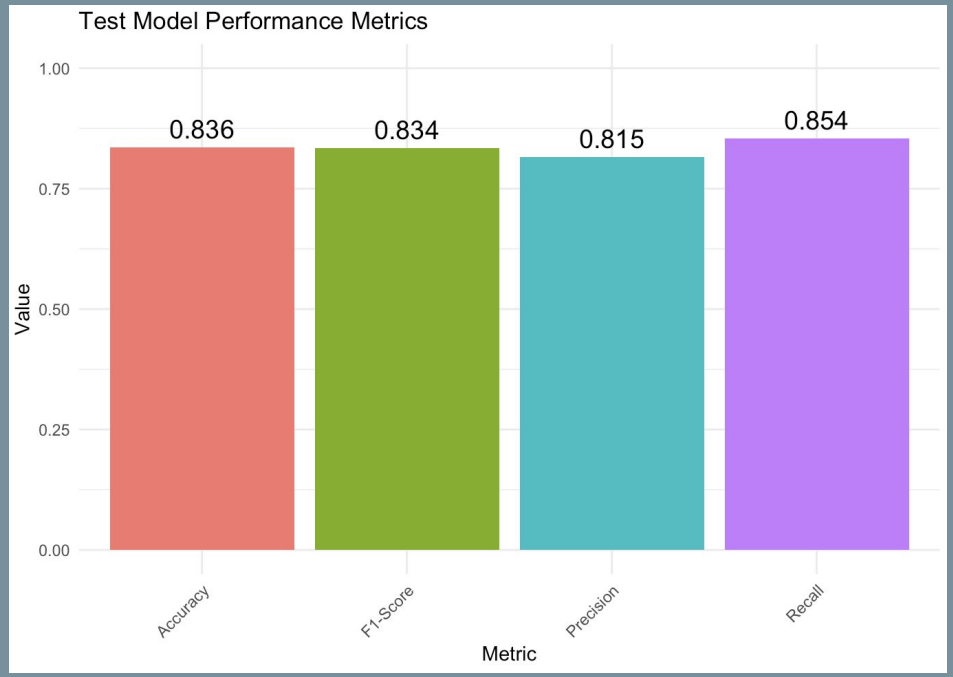




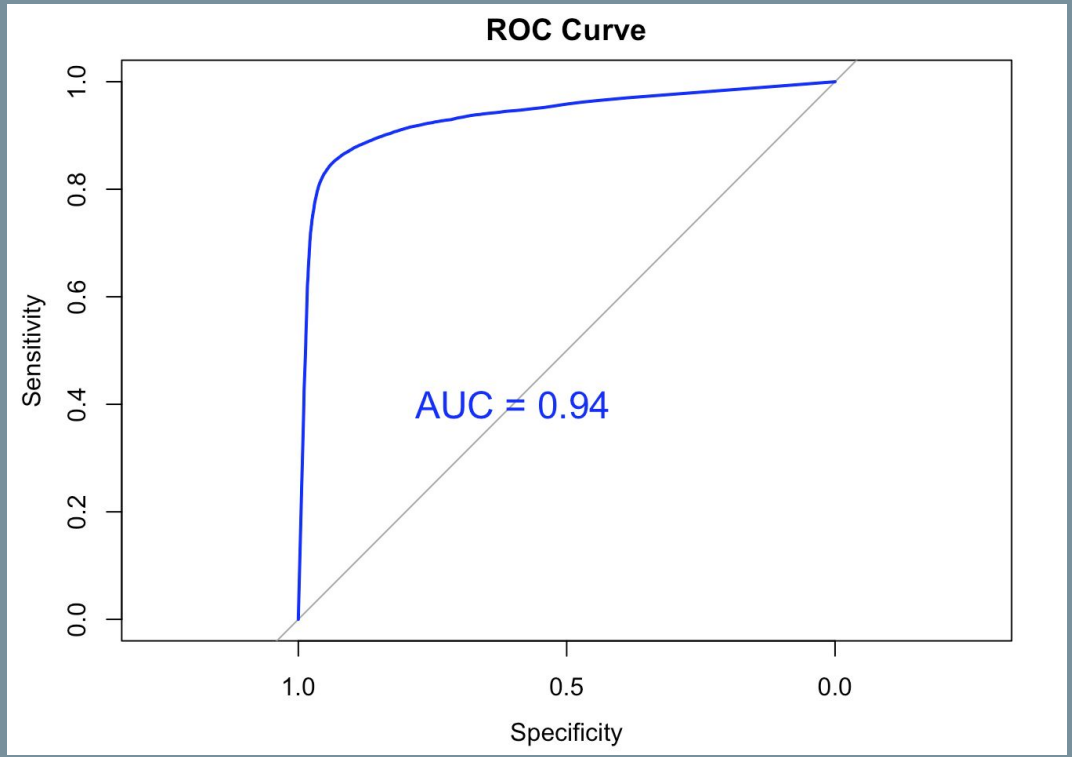
Trees used: 300, mtry: 3

Accuracy, F1-Score, Precision, and Recall of Train and Test set:





ROC Curve + AUC:



Model Performances

Comparison of Each Model's Performance

Logistic Regression

- Both the training set and testing set have an AUC of around 72%
- Both the training set and testing set has an accuracy of around 63%
- Overall, decent performance of the model but could be better

Decision Tree

- Accuracy of about 57% for both the train and test set.
- So out of all the predictions the model made on crash severity, only 57% were correct.
- Suggests that the model is underfitting.
- Overall, this means it is not a great model.

Random Forest

- Accuracy of about 83.5% for both the train and test set.
- The model's predictions of whether the Driver is at Fault is accurate 83% of the time.
- Metrics suggest the model performs better on the train set, worthwhile to look into potential overfitting.
- Good model, but may contain bias.

Thank you!