# Agenda

1) Introduction
2) Deliverable 1 - Risk Score Model
3) Deliverable 2 - Accident Severity Prediction
4) Conclusions and Next Steps

# Deliverable 1

## CRSS Dataset - Pedestrian Risk Score Model

# CRSS - Crash Report Sampling System

“CRSS is **a sample of police-reported crashes involving all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that result in fatalities**.”

## Accident Dataset
**(27.6K Cases)**
**2017-2022**

### Non-Motorist Datasets

- Distracted
- Safety Equipment
- Crossing

### Pedestrian Safety **(Dependent Variable)**

- Minor/No Injury (21.4K)
- Serious Injury (4.8K)
- Fatal Injury (1.4K)

---

- Weather
- Work Zone
- Interstate Highway
- Light Condition
- Junction
- Intersection type etc.

### Vehicle Dataset

- Roadway alignment (straight, curved)
- Traffic Signal
- Speeding
- Number of Lanes
- Traffic Control Device
- Speed Limit etc.

https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813459

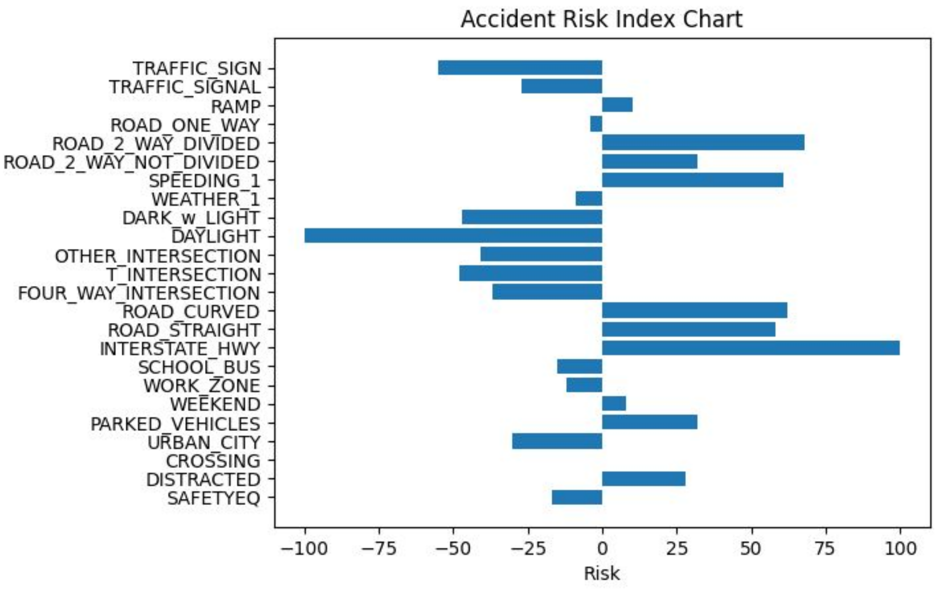
# CRSS - Approach 1

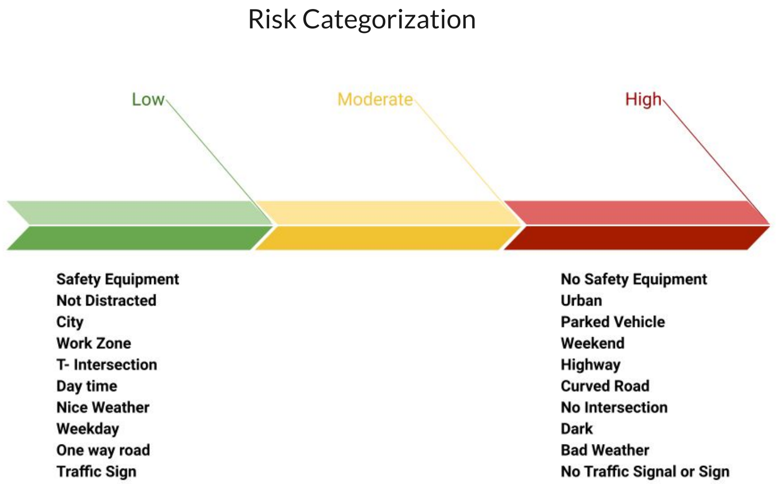
- **Binary Dependent Variable ->** 0 for minor/no injury and 1 for fatal/severe injury
- **Removed majority of the missing values :** Size dropped from 27.5K cases to 14.3K
- **One hot encoded all categorical variables**

**Model Results**

| Model - Test Dataset | F1 Score | AUC score |
|---|---|---|
| Logistic Regression - balanced | 0.52 | 0.73 |
| Decision Tree w/ Depth 10 | 0.3 | 0.69 |
| Bagging | 0.32 | 0.68 |
| Random Forest | 0.32 | 0.69 |
| Random Forest - balanced_subsamples | 0.47 | 0.68 |
| Random Forest - Upsampling | 0.49 | 0.68 |
| Random Forest - Downsampling | 0.48 | 0.67 |
| Gradientboost-Upsampled | 0.52 | 0.72 |
| Support Vector Machine-Upsampled | 0.48 | 0.71 |

# CRSS - Approach 1

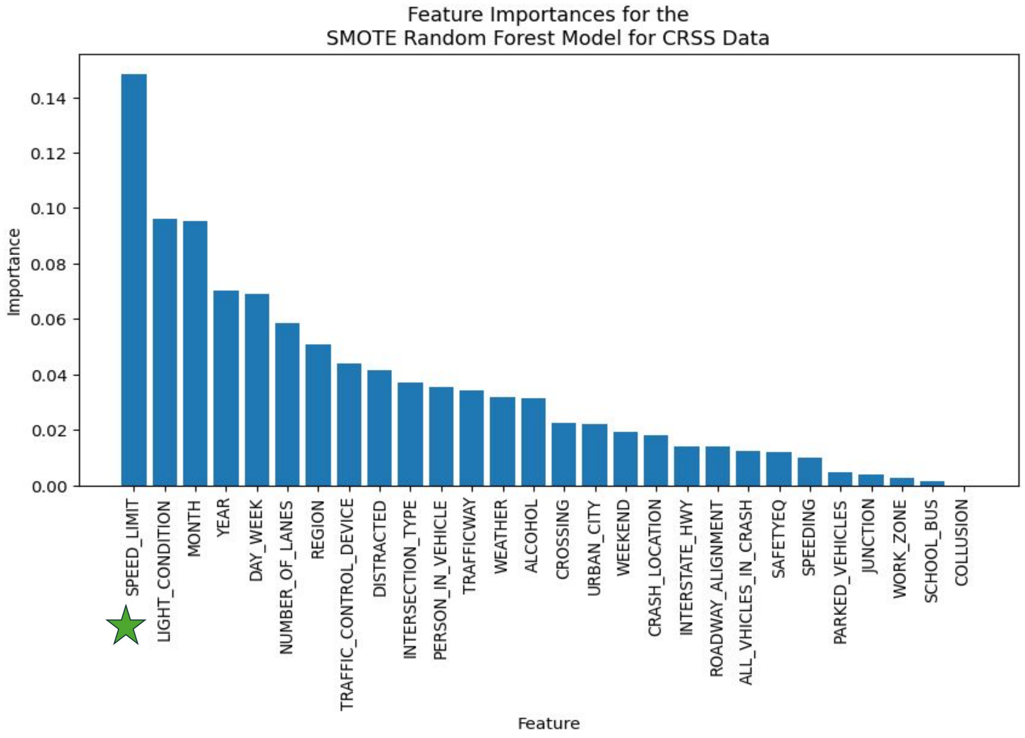


# CRSS - Approach 2

- **Ternary Dependent Variable:**
  - 0 for minor or no injury
  - 1 for severe injury
  - 2 for fatal
- Instead of removing missing values in approach 1, they were imputed

| **Model** | **Test Accuracy** |
|---|---|
| Random Forest Non-SMOTE | 77.63% ★ |
| Random Forest SMOTE | 75.42% |
| Neural Network Non-SMOTE | 77.18% |
| Neural Network SMOTE | 62.36% |



# Approach 2: Neural Network Models

- Configuration:
  - 2 dense layers
  - 1 batch normalization layer in between
- Non-SMOTE Model:
  - Predicted classes 1 & 2 poorly
- SMOTE Model:
  - Better at predicting classes 1 & 2
- Both versions predicted class 0 better than classes 1 & 2

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 32)                928

 batch_normalization (BatchN  (None, 32)               128
 ormalization)

 dense_1 (Dense)             (None, 3)                 99

=================================================================
Total params: 1,155
Trainable params: 1,091
Non-trainable params: 64
_________________________________________________________________
```

# Deliverable 2

Accident Severity Prediction:

- Part 1: FARS Dataset - Pedestrian Injury Prediction

# FARS - Fatality Analysis Reporting System

FARS Dataset: Provides yearly data on fatal injuries in vehicle-to-vehicle accidents

**Vehicles CSV Files**
1975 - 2022

- Hit/run accident?
- Drunk driver(s)?
- Car year manufactured

**Accidents CSV Files**
1975 - 2022

- State/city/county
- Weather/light conditions
- Work zone?
- **Dependent Variable:** Number of fatalities

Inner join on common field: “State Case Number”

Accidents + Vehicles Data
1975 - 2022

Time Period 1
1975 - 1981

Time Period 2
1982 - 1990

Time Period 3
1991 - 2000

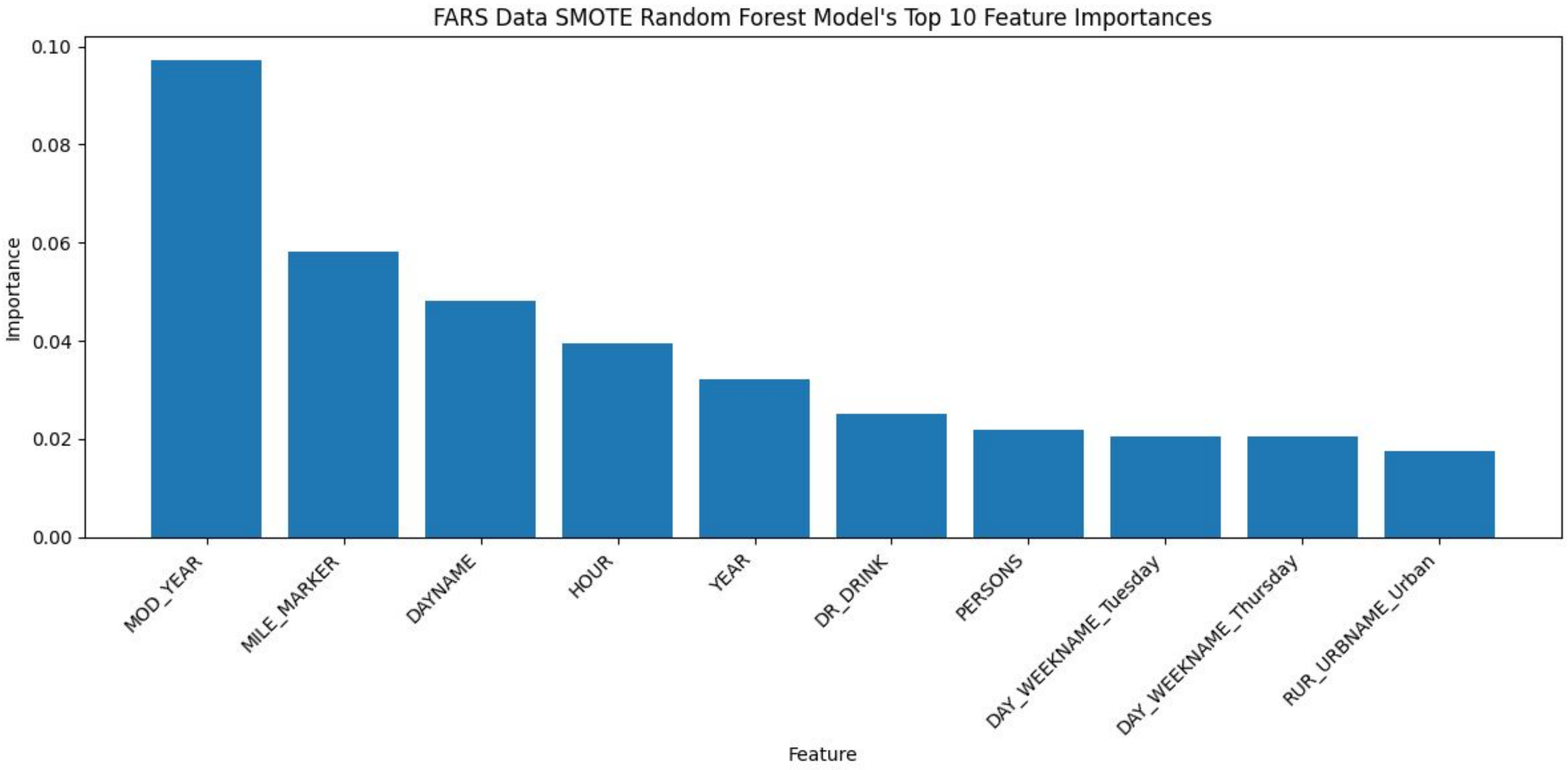
Time Period 4
2001 - 2007

Time Period 5
2008 - 2014

Time Period 6
2015 - 2022

https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/

# FARS Variable Importance Chart: 2015 - 2022



# Part 2 - Vehicle-to-Vehicle Traffic Severity Modeling

- **Dependent Variable "Severity" ->** classes 1 - 4
- **Datetime Breakdown ->** Datetime becomes year, month, day, hour, minute columns
- **Encode categorical variables**
- **Removed redundant columns**
  - **Windchill, when we already have temperature**
  - **End Latitude**
- **Scaled numerical values**

| Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| LightGBM | 0.75 | 0.75 | 0.75 | 0.75 |
| XGB | 0.76 | 0.76 | 0.76 | 0.76 |
| MLP | 0.72 | 0.7152 | 0.7204 | 0.716 |
| KNN | 0.48 | 0.47 | 0.48 | 0.47 |
| LSTM | 0.62 | 0.62 | 0.63 | 0.62 |
| GRU | 0.63 | 0.64 | 0.64 | 0.64 |

Model Results for Downsampled Data

# Conclusions and Next Steps

# Conclusions

1. **Overall Impact - Deliverable 1:**

- CRSS models identified key factors that impact accident severity using variable importance plots
- Models can predict accident risks, urging organizations to improve road safety
- **Recommendations:**
    - Decrease speed limit for high risk roads
    - Increase lighting on rural highways. The lighting condition was ranked 2nd most important

# Conclusions

1. **Overall Impact - Deliverable 2A:**
   a. Like CRSS models, FARS models' variable importance plots are crucial to predicting the number of fatalities, so organizations can prioritize certain factors
   b. **Recommendations:** Keep US fleet up to date, year of manufacture and mile marker (specific section) on the road were important factors.

2. **Overall Impact - Deliverable 2B:**
   a. Advanced models like XGBoost revealed critical predictive road features such as 'Amenities', Traffic Signal', and 'Station'
   b. **Recommendations**: Improve traffic signals and infrastructure near high-traffic amenities to reduce severe accidents.