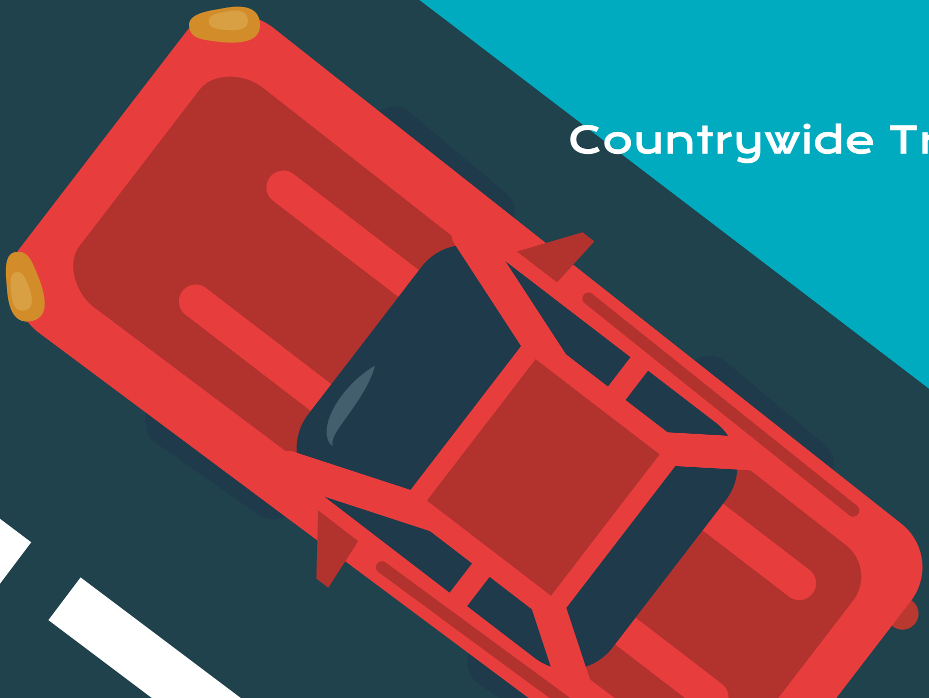
Predicting Car Accidents' Severity

Countrywide Traffic Accident Dataset (2016-2021)



By: Priyanka Iragavarapu, Charles Barnes, Ellen Wei
(Group M - Lec 1)

Table of Contents

01

INTRODUCTION

A background on US Accidents and
Traffic Accident Dataset overview

Car Accidents in the US

Countrywide

Covers 49 states from
Feb 2016 - Dec 2021

Youth fatalities

Most common cause of
death of young adults

By age

Majority of deaths
occur 25-64

Types of crashes

Usually fatal when
fixed objects involved

Countrywide Traffic Accident Dataset

50000

Observations

Training data: 35000
Testing data: 15000

Each observation represents a car accident incident

44

Variables

Detailed information recorded with each incident
Ex. Start time, Description, State

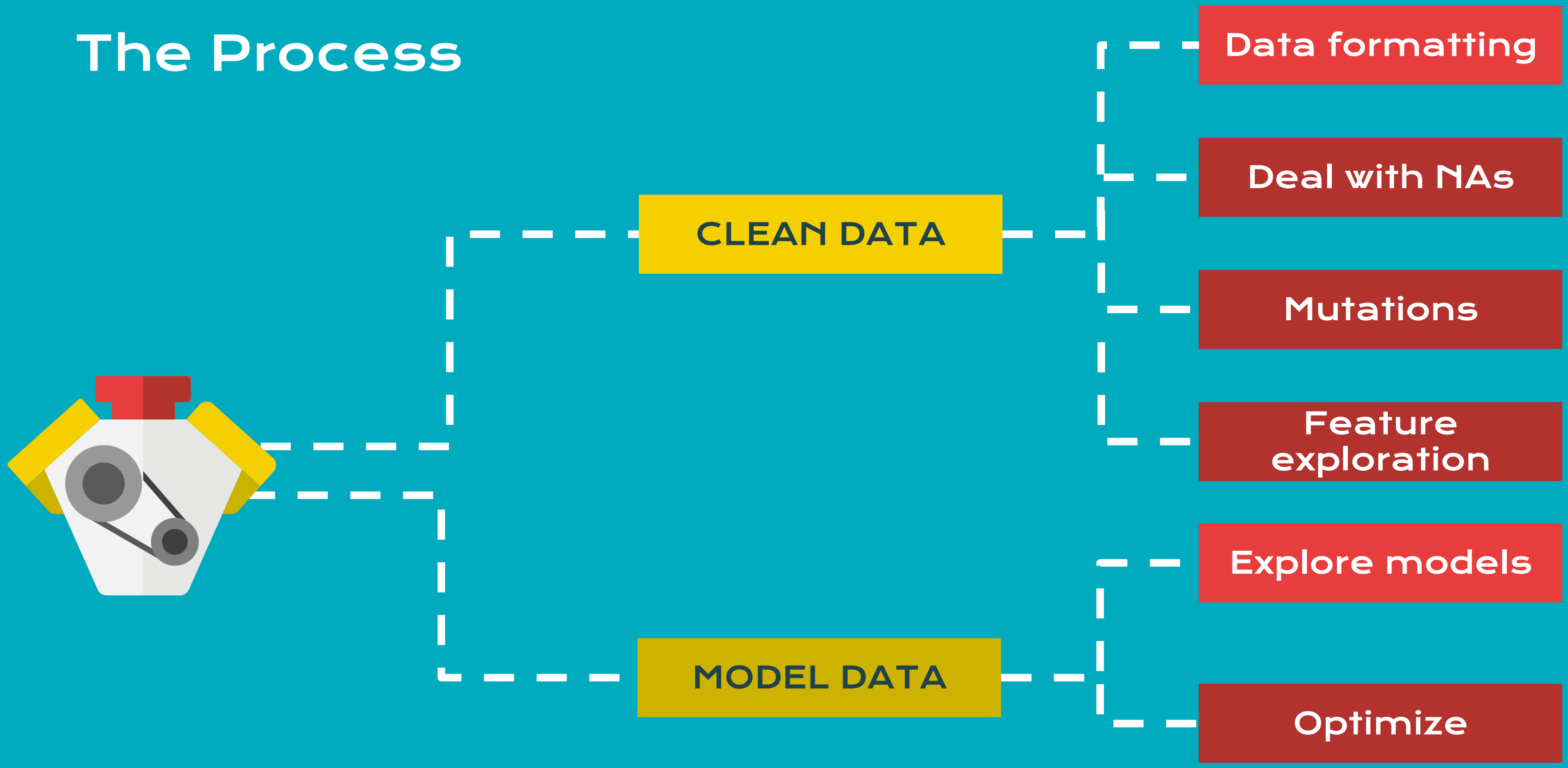
Response variable: SEVERITY (mild vs. severe)

02

METHODOLOGY

Data cleaning and model exploration

The Process


CLEAN DATA
MODEL DATA
Data formatting
Deal with NAs
Mutations
Feature exploration
Explore models
Optimize

Data Cleaning - NAs

Numerical

- First used medians
- Tried package 'mice'

Mutated variables

Description: Word count, Characters,
Binary variables for presence of words
Time: Month, Season, Year, Hour Start,
Time of Day, Rush Hour,
Weekend/Weekday
Location: Local road, Region

Categorical

- Tried to consolidate categories
- Most models can't handle 50+ levels -> removed
- Package 'Hmisc'

Data augmentation - mutations

Description analysis:

- Word cloud visualization
- Common words
- Categorical variables

New grouped variables:

- Start Time Category (ex. morning)
- Weekday
- Traffic object, building, sign

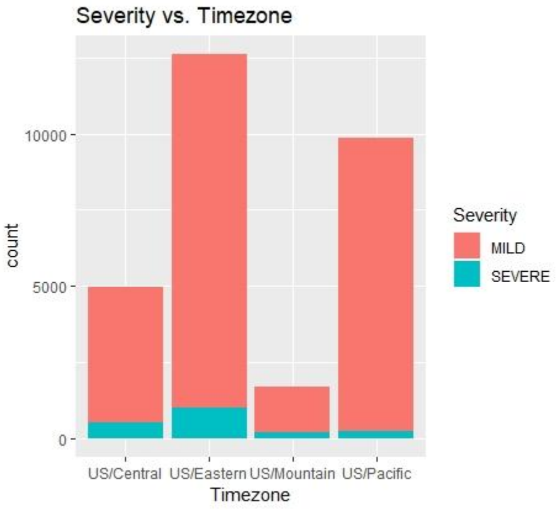
"Caution"

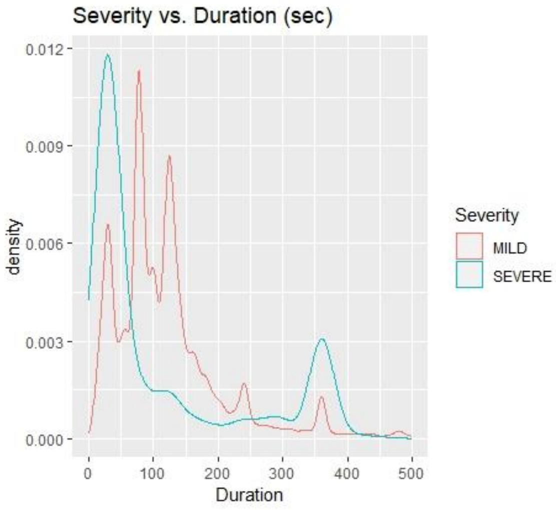
	Yes	No
Mild	7337	24145
Severe	1	3517

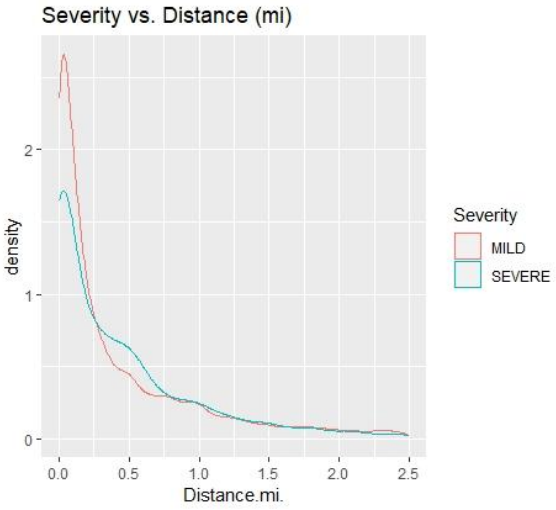
"Slow"

	Yes	No
Mild	4506	26976
Severe	3	3515

Variable Selection - Feature exploration







Cleaned Countrywide Traffic Accident Dataset

50000

Observations

Training data: # 35,000
Testing data: # 15,000

51

Variables

Removed categorical variables with large amount of NAs

Median for NAs in numerical variables

Mutated data to include additional variables

Initial Model Comparisons

Model	Pros	Cons
Logistic Regression	Good for classification	Low accuracy, hard to implement with the 90/10 split
LDA/QDA	Good for classification - simple boundaries	Dataset is extremely complex
KNN	Simple	Only numerical
k-means	Easy to understand	Only numerical
Multiple Linear Regression	Extremely basic	Low accuracy

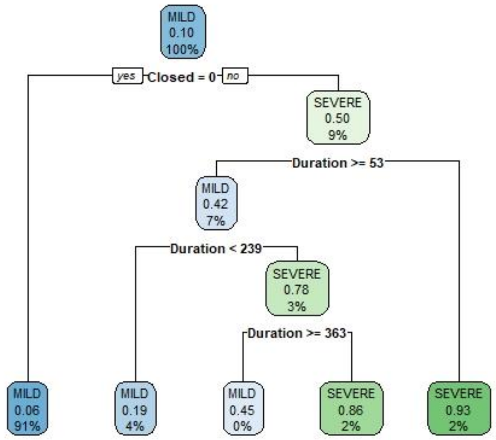
Models: Tree-based

- Could immediately decide whether an observation was Mild or Severe
- Reduces the dataset that is "difficult" to classify
- Closed, Duration, CautionOrSlow, Timezone were most important

Base model: 92.89% training accuracy → Eventually got a 92.7% Public Score

Training Data Severity

Predicted Severity		Mild	Severe
	Mild	31330	2338
	Severe	152	1180



Models: Random Forest

<u>Why try RF?</u>

- 1 tree, variables can dominate
- Randomly choose a few predictors each split

<u>RF 5</u>

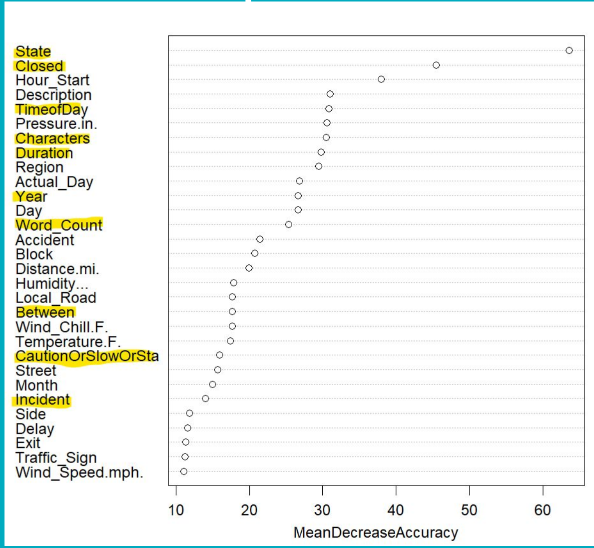
- 500 trees, default mtry
- Numerical: Distance
- Time: Duration, TimeofDay, Season
- Description: Characters, Word Count, Closed, CautionOrSlow
- Location: State

99.7% training, 93.76% testing

Predictions: 13,917 Mild, 1,083 Severe

	Training Data	
	Mild	Severe
Mild	31,478	108
Severe	4	3,410

Variable Importance Plot

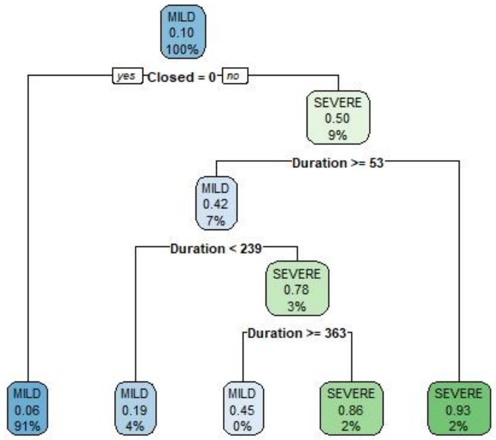


	Rev 12	
	Mild	Severe
Mild	31,475	454
Severe	7	3,064

31,478	108
4	3,410

Training Data Severity

		Mild	Severe
Predicted Severity	Mild	31,475	454
	Severe	7	3,064



Models: Random Forest

Back to the drawing board . . .

<u>RF 12</u>

- Numerical: Distance
- Time: Duration, TimeofDay, Season, **Year**
- Description: Characters, Word Count, Closed, **CautionOrSlowOrStationary, Incident, Between, Exit**
- Location: State

98.7% training, 94.25% testing

- 350+ severe cases incorrectly classified as Mild in training predictions compared to Rev 5

Predictions: 14,009 Mild, 991 Severe

Training Data

	Mild	Severe
Mild	31,475	454
Severe	7	3,064

Rev 5 Testing (columns) / *Rev 12 Testing* (rows)

	Mild	Severe
Mild	13,821	186
Severe	96	897

Optimizations

County Census Data

OVERFITTING

- Log total population
- Proportion 15-24 y/o
- Proportion 65+ y/o
- Log aggregate commute
- Average vehicle/household

Bagging

Mtry =13, 93.4% testing

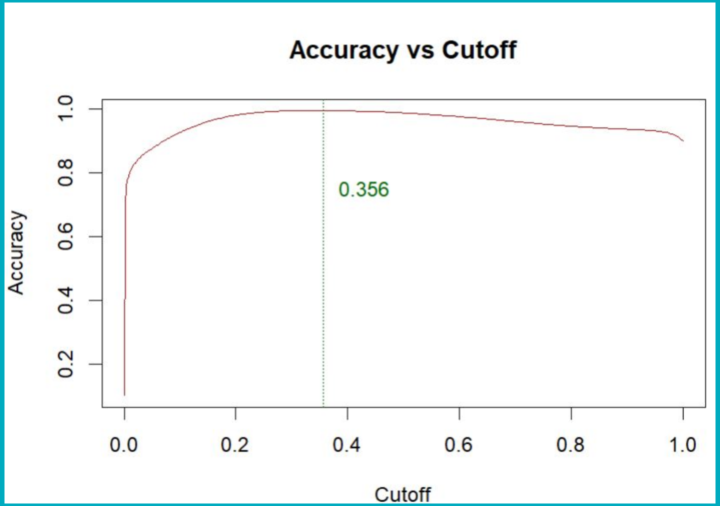
Boosting

OVERFITTING: high training accuracy, low testing accuracy

Importance and Probability Cutoff

Removing variables using variable importance
Changing probability cutoff using ROC

- 0.356 instead of 0.5 as the threshold for classifying as Severe
- 99.4% training, 93.21% testing



03

RESULTS & DISCUSSION

Final constructed model analysis

Analysis: Final Model and Key Ideas

Model
Random Forest

Observations
50000 incidents

Predictors
51 Traffic Predictors

Simplicity
Simple

Kaggle Score
Public: 0.94266
Private: 0.94080

Rank
Top 15

Competition: 13 / 36
Lecture: 8 / 17

Discussion: The Important Predictors

Caution+slow+stationary

Description words

Closed

Description word

Time of Day

Categorical

Duration

Numerical

State
Year
Incident
Between
Exit
Distance
Characters (description)
Word count
Season

04

LIMITATIONS & CONCLUSIONS

Setbacks, assumptions, and final words

Limitations

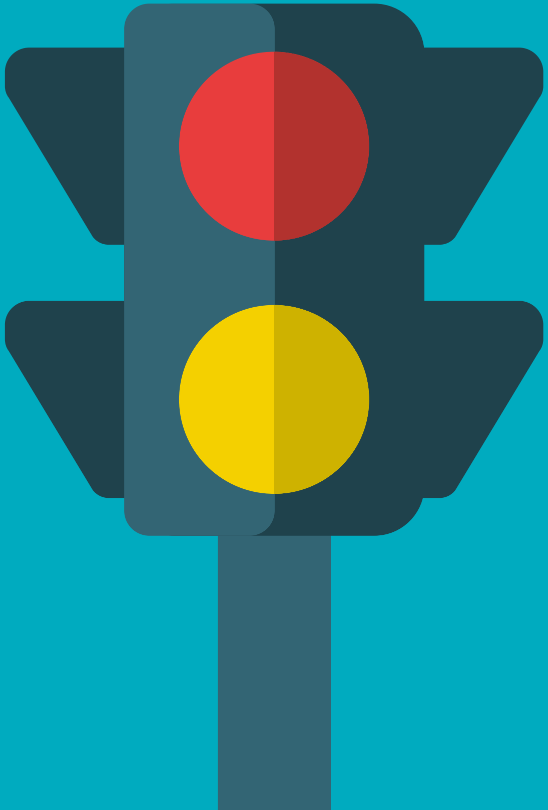
Data Cleaning

- Used medians for NAs at first
- NA imputation too computationally intensive
- Threw out categorical variables with large amount of categories (ex. city, county, zip code)

Modeling

- Random forest limitations
- Computationally intensive
  - Lack of visualization
- Overfitting occurred when using numerical predictors
- Can't use for inference - black box

Conclusion

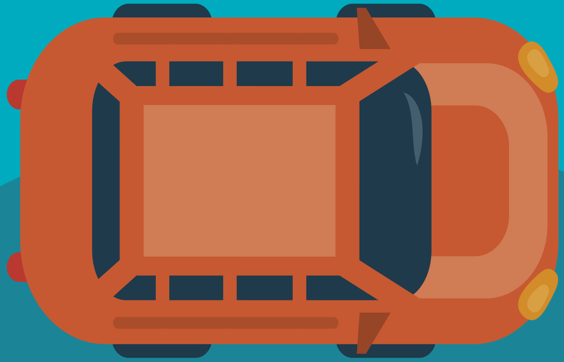


We get good predictions using our model and the random forest model is interpretable when we take a look at specific parameters.

Most importantly, our model is **simple**! It uses **only 5** of the original predictors, and the remaining are mutated predictors that we added.

We even attempted adding census data but found those predictors were not as significant as our 5 original predictors.

Thank you and happy holidays!



Special thanks to Professor Almohalwas for the quarter. We enjoyed the friendly Kaggle competition.