Predicting Severity of Car Crashes

Avani Kanungo, Kathy Nguyen-Ly, Jenna Schindele, Sarah Zhari
(Lecture 2, Group E)

Table of Contents

01

Introduction

Car crash context and data set overview

Goal: Predicting Severity of Traffic Delays after a Car Accident

Data: 35k training, 15k testing

Variables with NA Values

PREDICTORS WITH NAs

CATEGORICAL	LOGICAL	NUMERICAL
City	Sunrise_Sunset	Temperature.F.
Zipcode	Civil_Twilight	Wind_Chill.F.
Timezone	Nautical_Twilight	Humidity...
Airport_Code	Astronomical_Twilight	Pressure.in.
Weather_Timestamp		Visibility.mi.
Wind_Direction		Wind_Speed.mph
Weather_Condition		

Used the library mice to impute numerical missing values

02

Methodology

Data cleaning and modeling

Methodology Process

Data Cleaning

Create
new variables

Modeling

Create numerous
models using our
training data

Analyze Models

Calculate accuracy and
error rates

Model Selection

Choose the best model
based on accuracy and
simplicity

Creation of Numerical Variables

- Time, Latitude, and Longitude themselves may not be useful
- The following may indicate higher likelihood of SEVERE:
  - Large difference in time (more time to clear accident)
  - Large difference in Lat/Lng (larger area affected by accident)

- Components of Date can also be useful:
  - Year: certain years (COVID) may have had decreased driving rates
  - Month: holiday months often have increased driving rates
  - Hour: dusk to dawn hours have increased chance of accidents

Numerical Variables	
Change in Time	= End_Time – Start_Time
Change in Latitude	= End_Lat – Start_Lat
Change in Longitude	= End_Lng – Start_Lng
Year	= Year(Start_Time)
Month	= Month(Start_Time)
Hour	= Hour(Start_Time)

Creation of Variables from Description

- Cannot use “Description” predictor as is since each description is different from each other
- Generated word clouds using the “Description” predictor for both MILD and SEVERE accidents
- Based on word clouds, created logical predictors for whether the following words/phrases were in the description:
    - Accident
    - accident
    - Closed
    - closed
    - Traffic/traffic
    - Blocked/blocked
    - Caution
    - Incident/incident
    - Road closed due to accident

Creation of Variables from Weather_Condition

WEATHER CONDITION

"Partly Cloudy"
"Cloudy"
"Mostly Cloudy"

"Light Rain"
"Rain"

"Thunderstorm"
"T-Storm"

"Windy"
"Fair / Windy"
"Fair"
"Clear"

LOGICAL VARIABLES

Rain
Cloudy
Windy
Thunderstorm
Haze
Snow
Clear

Finalized Model Data Set

Observations

We kept all 35,000 observations from the original data set

Predictors

We reduced the original 43 predictors (numerical, categorical, and logical) to 25 predictors (categorical, logical)

Logistic Regression & Tree Modeling

Logistic Regression:

- We constructed a logistic regression model using all 25 predictors to predict the Severity of a car accident
- When applied to the testing data, this model produced a **93.46%** accuracy rate

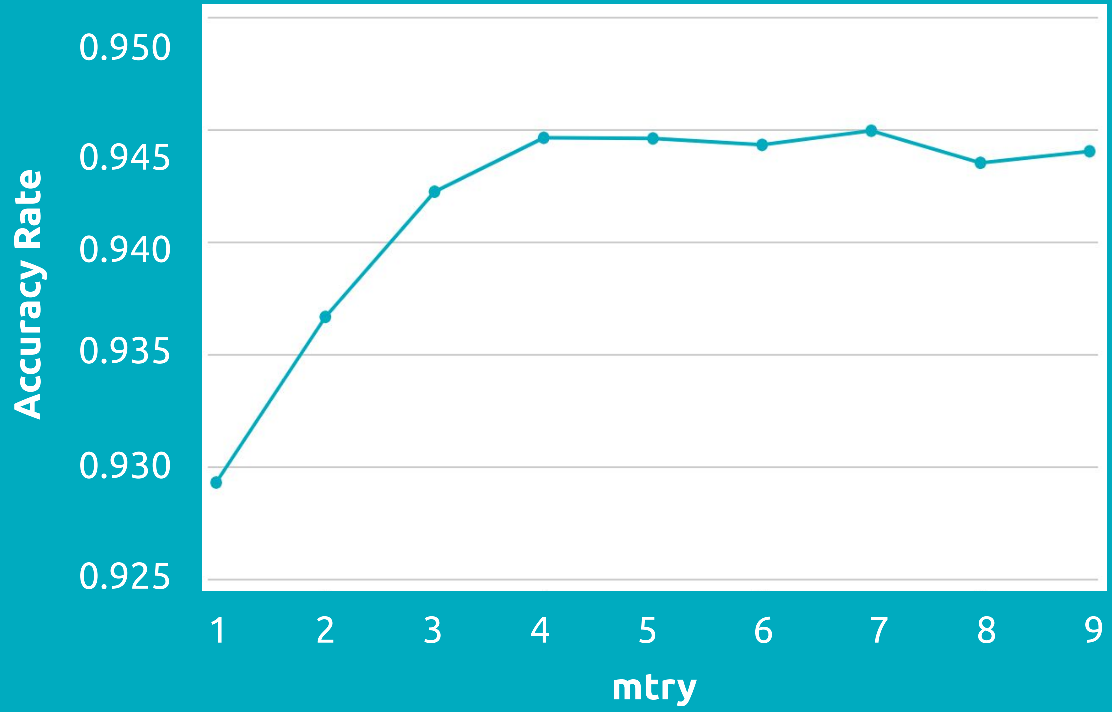
Tree Modeling

- We constructed a tree model using all 25 predictors to predict the Severity of a car accident
- When applied to the testing data, this model produced a **93.28%** accuracy rate

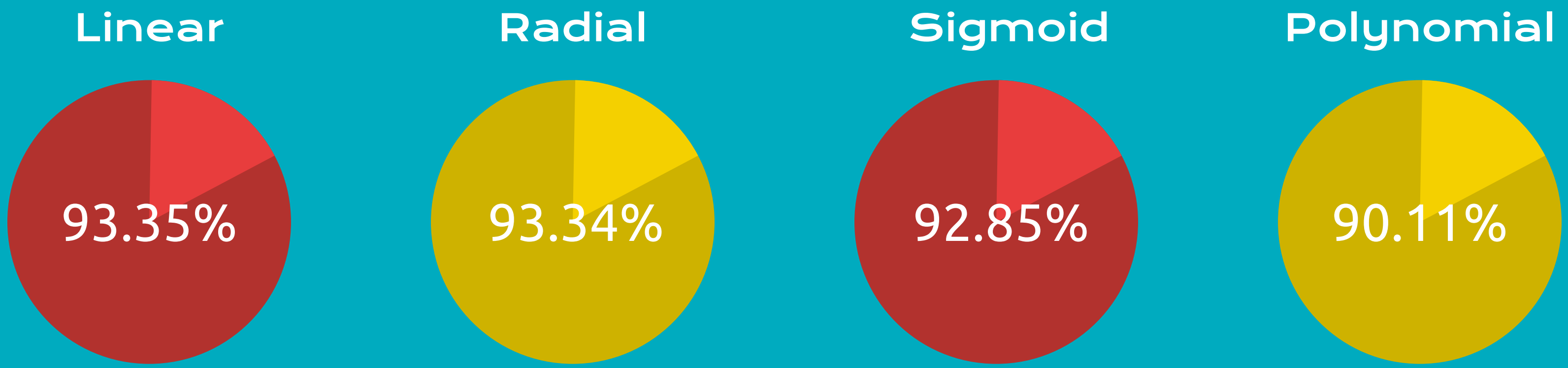
Year, Month, Hour, Side, State, Timezone, Change_Time, Change_Lat, Change_Lng, Closed, closed, Accident, accident, Traffic, Blocked, Caution, Incident, Rain, Cloudy, Windy, Thunderstorm, Haze, Snow, Clear, Road_Closed_Due_To_Accident

Random Forest Modeling

- Using the library randomForest, we constructed Random Forest models
- Modified the parameter, mtry, to values ranging from 1 to 9 to assess variance in accuracy rates
- Although mtry = 7 had the highest training accuracy, mtry = 6 resulted in the best testing accuracy



SVM Modeling



When predicting the testing data, SVM using a linear kernel had an accuracy rate of **93.33%**, while SVM using a radial kernel had an accuracy rate of **93.29%**. Therefore, we used the <u>linear</u> kernel for our SVM model.

Model Testing Accuracies

Logistic Regression	Tree	Random Forest	SVM
93.46%	93.28%	94.43%	93.33%

Evaluating our testing accuracies, we conclude that our Random Forest model is the best at predicting our testing data.

Proposed Models Summary

	MODEL PROPERTIES			
	Interpretable	Efficient	Accurate	Flexible
Logistic Regression	✓	✓	✗	✗
Tree	✓	✓	✗	✗
Random Forest	✗	✗	✓	✓
SVM	✗	✗	✗	✓

Final Model

Choosing the Model:

- Ideally, we would want to maximize accuracy while minimizing complexity
  - Increased accuracy rates generate higher predicting power
  - Simpler models are often easier to interpret
  - In reality, achieving both is difficult, so we must prioritize one over another
- We chose our first priority to be maximizing the accuracy rate
  - Therefore, our final model is: **<u>Random Forest (mtry = 6) using all 25 predictors</u>**

Year, Month, Hour, Side, State, Timezone, Change_Time, Change_Lat, Change_Lng, Closed, closed, Accident, accident, Traffic, Blocked, Caution, Incident, Rain, Cloudy, Windy, Thunderstorm, Haze, Snow, Clear, Road_Closed_Due_To_Accident

03

Results & Discussion

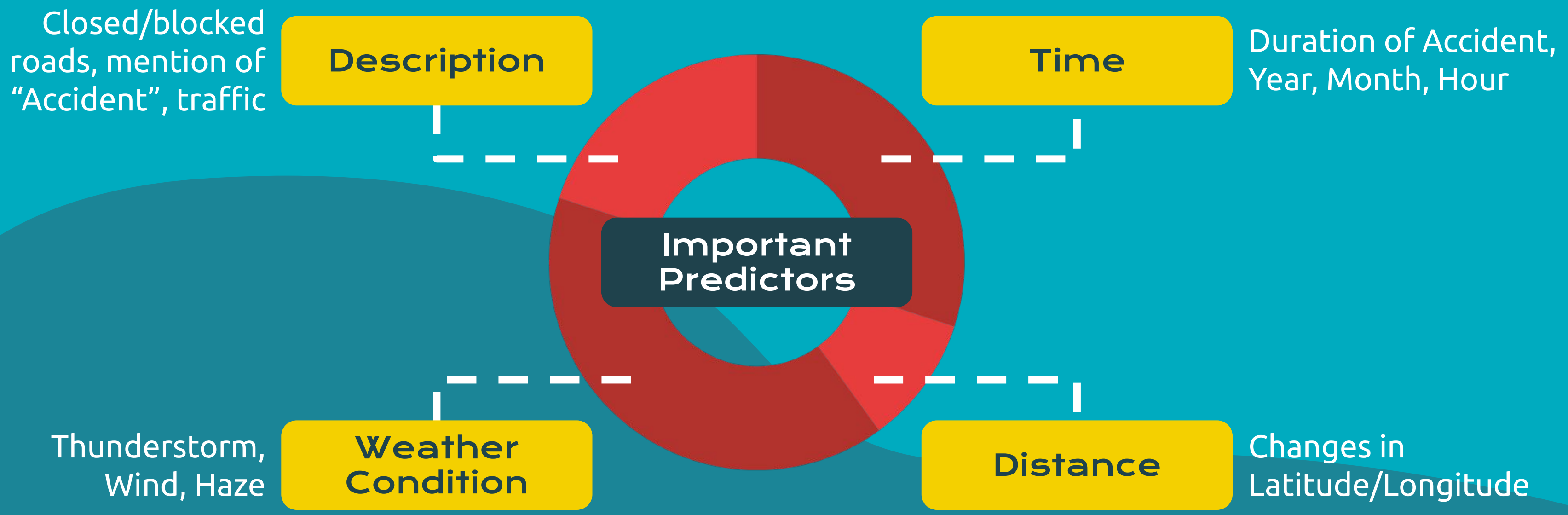
Final model analysis

Model Analysis

FINAL RANKING	FINAL SCORE	FINAL MODEL
3rd	94.55%	Random Forest, mtry = 6

# PREDICTORS	OBSERVATIONS	ACCURACY RATE
25	35,000	94.43%

Important Predictors



MOST Important Predictors

Most Important

Description

Roads closed were a strong sign of severe accidents – caused more traffic

Weather

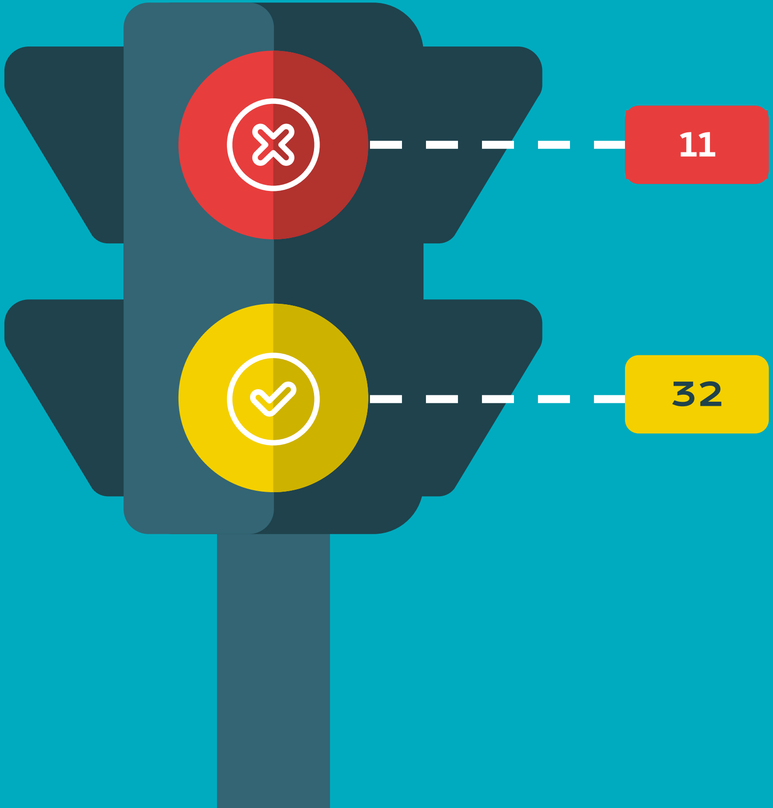
Weather can impact visibility and cause slippery roads – leading to severe accidents

04

Limitations & Conclusions

Setbacks, assumptions, and final words

Most Useful Variables Extracted from “Description”



Numerical Predictors

Most predictors weren’t helpful as is

Categorical Predictors

Extracting key phrases was key to our model

Limitations with Missing Values

ATTENTION!

Mice

Doesn't help with filling in missing categorical values

Predictors

Did not end up using the numerical predictors that had NAs; Created logicals for others based on if a word/phrase was present

Timezone

Used the mode of the state's timezone to fill in when missing

Additional Data That Could Be Useful

Car Specs

Size, condition, etc.

Type of Area

Urban vs. Local

Driver Condition

Drunk driver,
distracted driving

Responders

Police, EMS, Fire Dept

REMINDERS!

01 DON'T DRINK AND DRIVE

02 DON'T TEXT AND DRIVE

03 HAVE A GREAT BREAK!

Thank you for listening!
Questions?