Predicting Car Accidents' Severity

Lecture 1 Group K - Gary Ramos, Marlyn Tanuwandi, Thi Nguyen, Juhyeon Seo

Overview

INTRODUCTION

Context and Overview

METHODOLOGY

Data Cleaning and Modeling

RESULTS & DISCUSSION

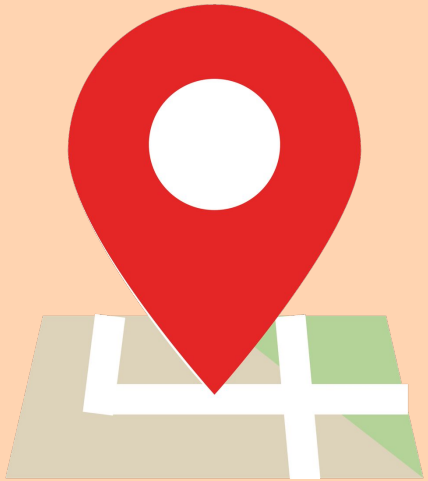
Final Constructed Model Analysis

LIMITATIONS & CONCLUSION

Setbacks, assumptions, and final words

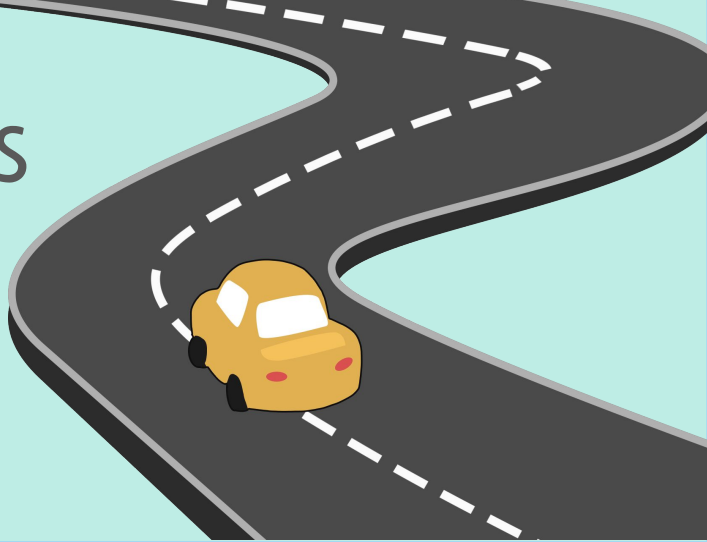
INTRODUCTION

01



LEADING CAUSE • UNITED STATES

CAR ACCIDENTS



6,000,000

CAR ACCIDENTS

38,368

DEATHS

Car Accident Data Set

02

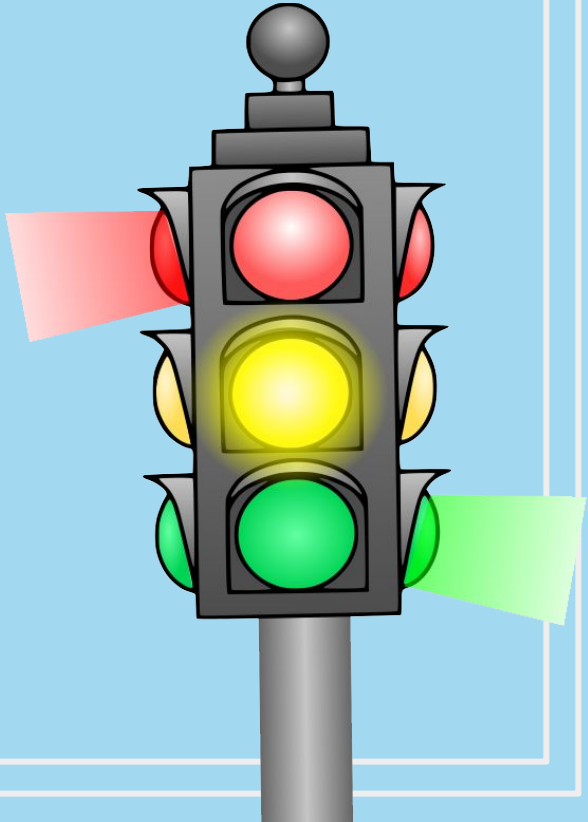
METHODOLOGY

01
Clean Data

02
Model Data

03
Analyze Models

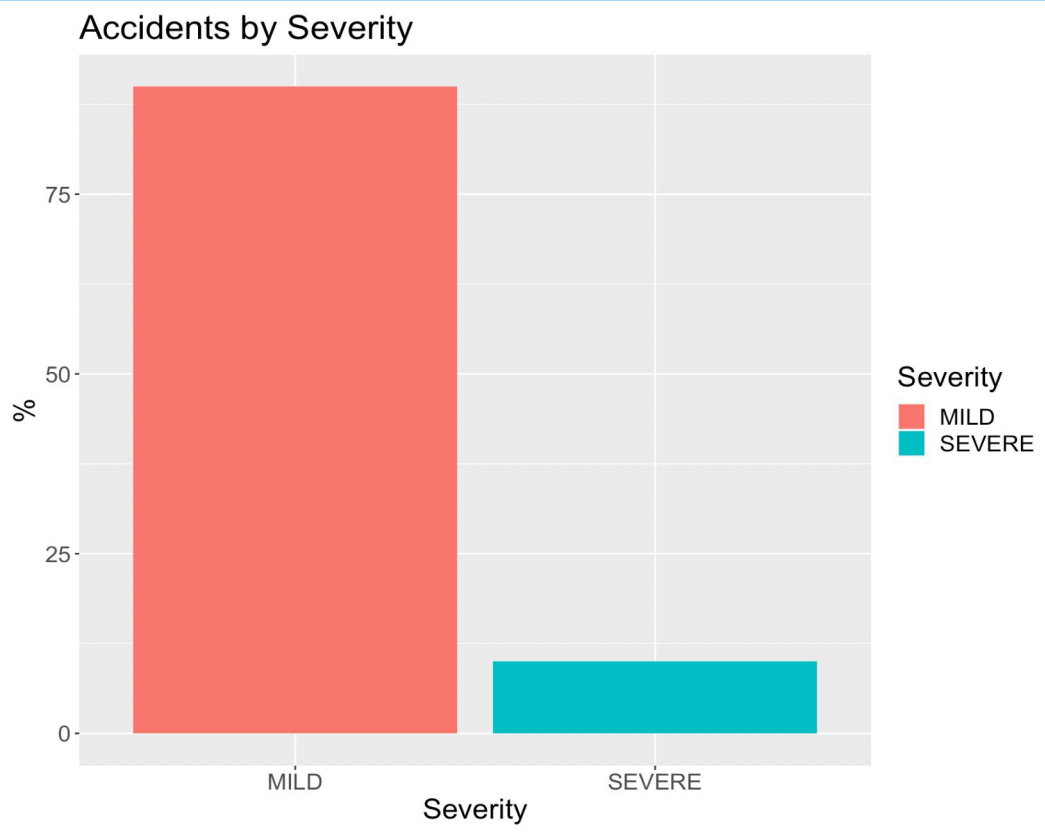
04
Compare Models



Exploratory Data Analysis (EDA)

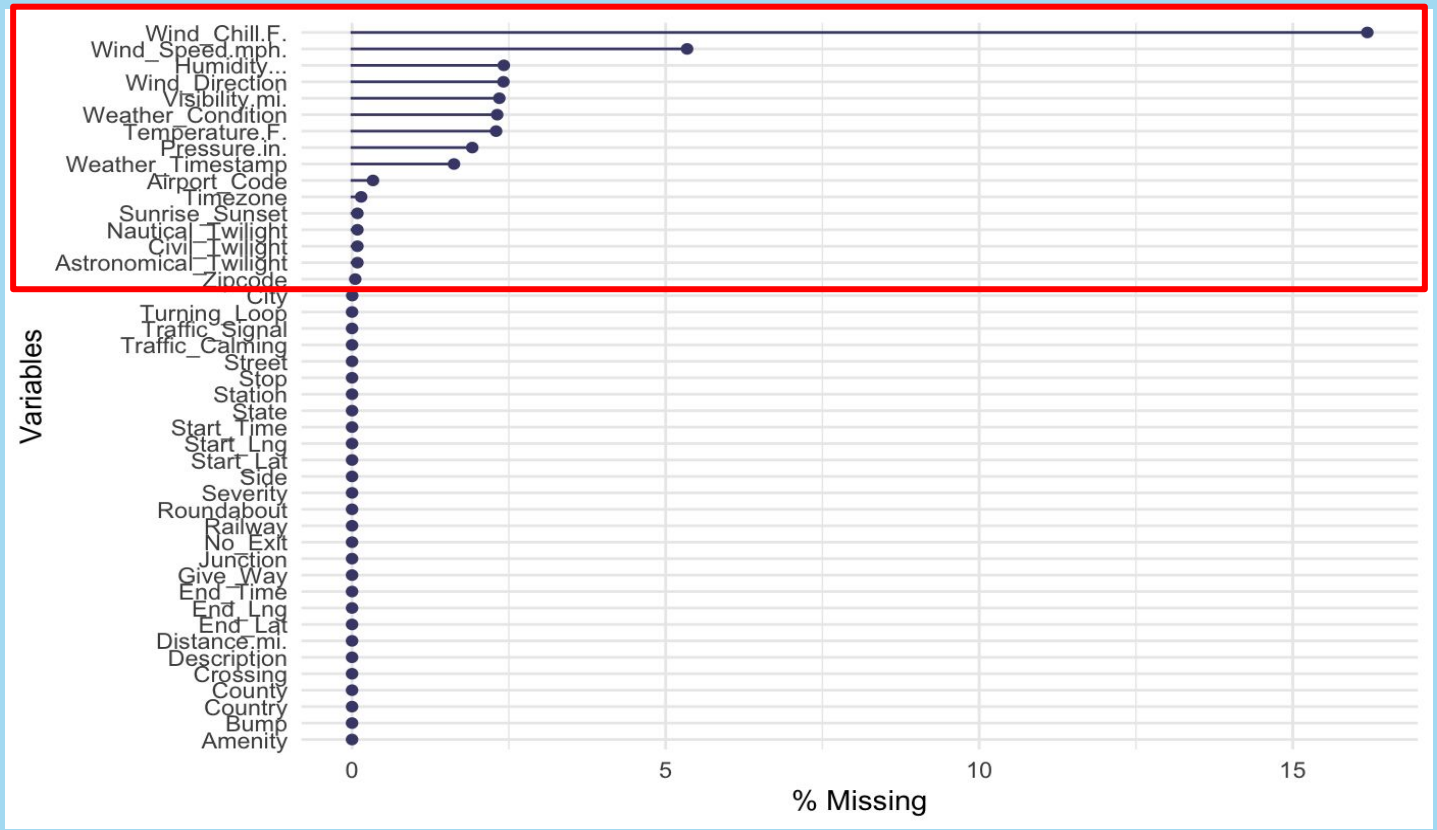
Car Accident Data Set

Exploratory Data Analysis (EDA)



	Total	Percentage
MILD	31482	90%
SEVERE	3518	10.1%

Exploratory Data Analysis (EDA)



Clean Data: Exploratory Data Analysis (training)

Variables	# of NAs
City	1
Zipcode	18
Timezone	51
Airport_Code	117
Weather_Timestamp	569
Temperature.F	804
Wind_Chill.F	5666
Humidity	847

Pressure.in	671
Visibility.mi	822
Wind_Direction	845
Wind_Speed.mph	1870
Weather_Condition	810
Sunrise_Sunset	30
Civil_Twilight	30
Nautical_Twilight	30
Astronomical_Twilight	30

13211

NAs total

Clean Data: Exploratory Data Analysis (testing)

Variables	# of NAs
Zipcode	4
Timezone	15
Airport_Code	36
Weather_Timestamp	264
Temperature.F	357
Wind_Chill.F	2485
Humidity	373
Pressure.in	312

Visibility.mi	358
Wind_Direction	376
Wind_Speed.mph	843
Weather_Condition	371
Sunrise_Sunset	12
Civil_Twilight	12
Nautical_Twilight	12
Astronomical_Twilight	12

5842

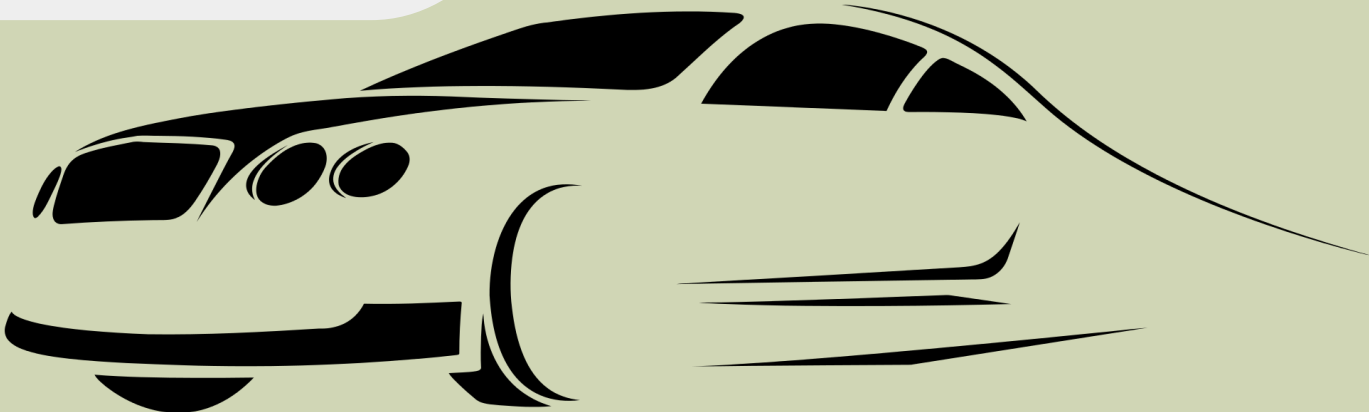
NAs total

Clean Data: Impute missing variables using mice

METHOD

- Use information from other variables in the dataset to predict and impute the missing values
- Run 5 imputations
- Create a dataset after the imputation

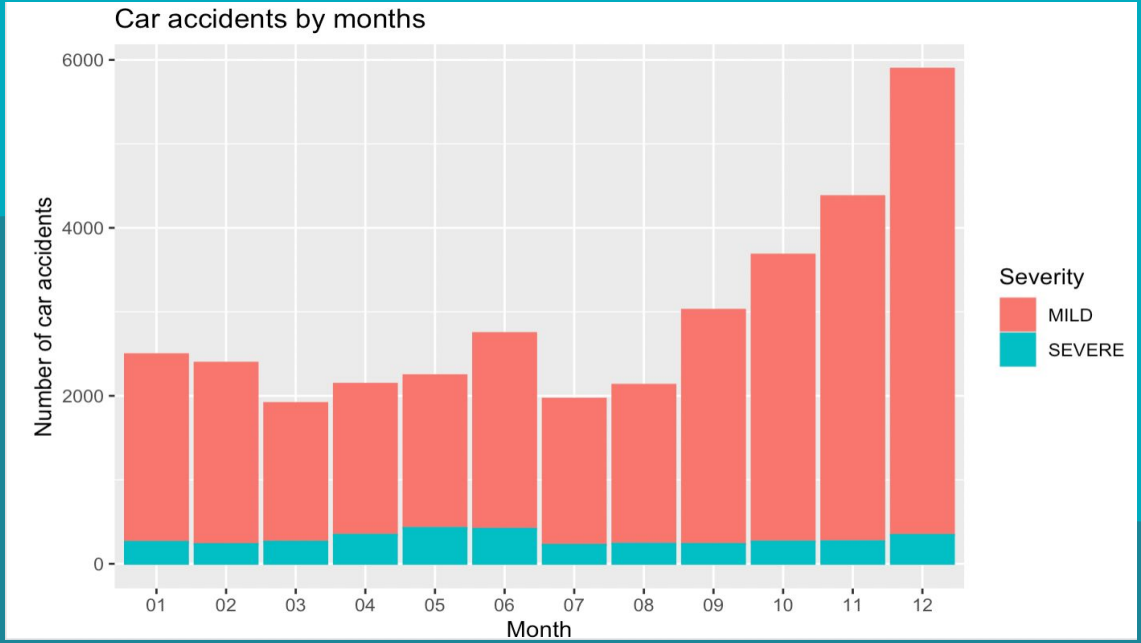
	Training Data	Testing Data
# of NAs	2531	1114



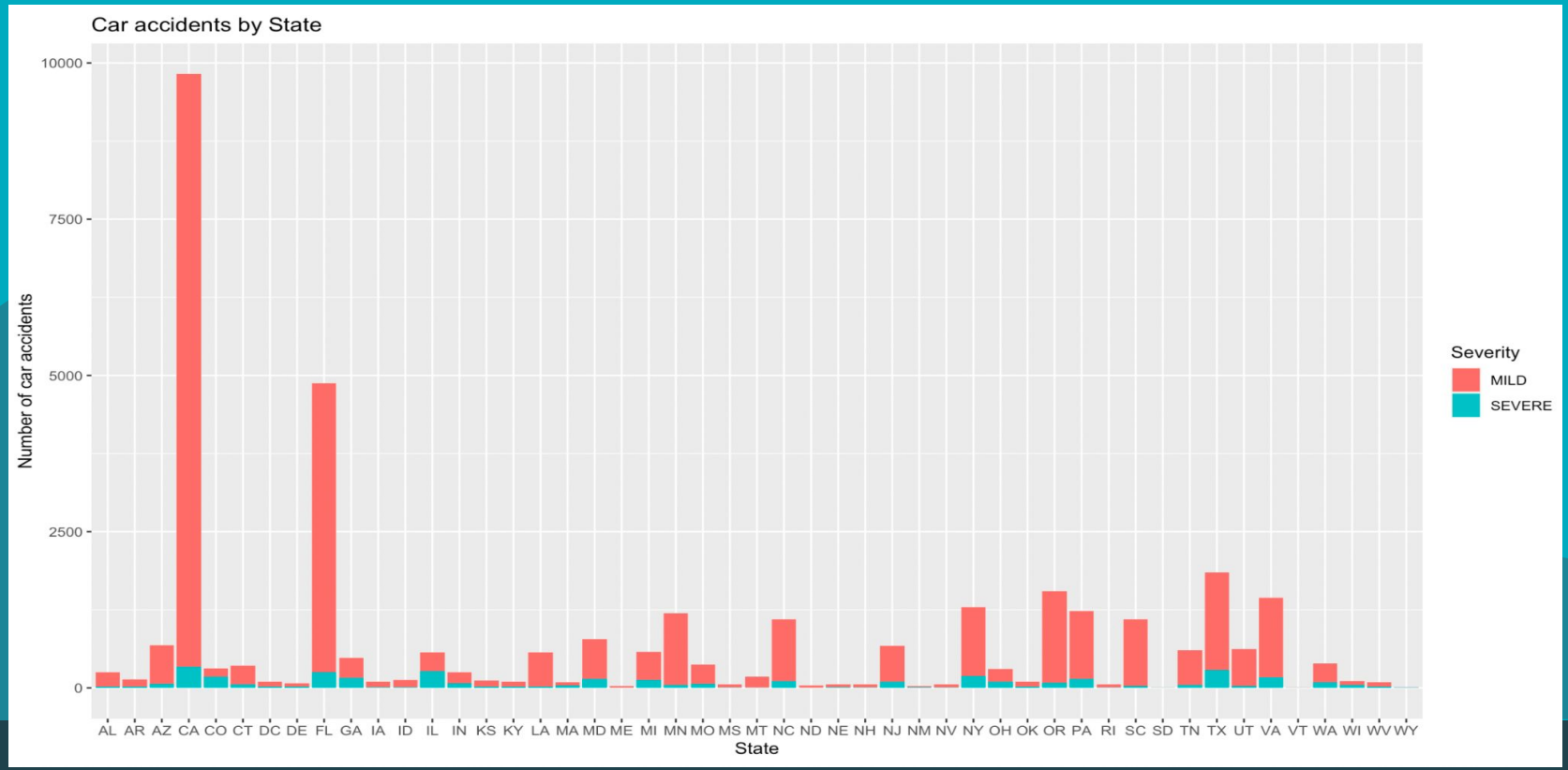
Clean Data

- Merge data based on County and State
- Features engineering
  - Split the time from the start_time and end_time of the dataset and create the time each accident took place
  - Creating new variables based on Description predictor in which using the common words that appears the most in the description variable

Data Visualization



Data Visualization



Clean Data: Remove insignificant variables and variables with missing values

Weather_Timestamp	Wind_Direction	index	Start_Time
Airport_Code	Wind_Speed.mph.	Transit_County	End_Time
Timezone	Visibility.mi.	Wind_Chill.F.	Country
Walk_County	Sunrise_Sunset	End_Lat	Street
City	Civil_Twilight	End_Lng	County
Description	Nautical_Twilight	Weather_Condition	new_date
Zipcode	Astronomical_Twilight	MedianHouseholdIncome_County	is.blocked

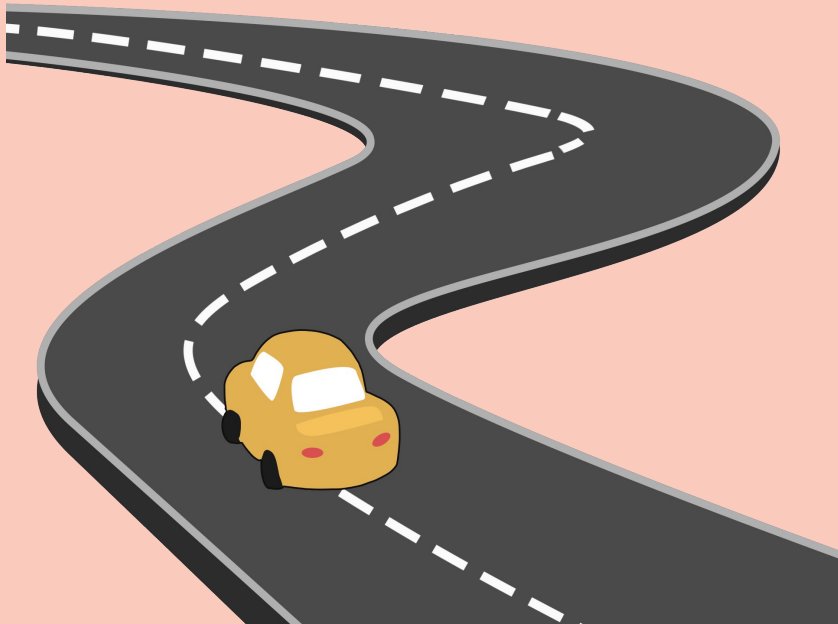
Clean Data: Updated Data

Clean Data: Add New Variables

- Merge fatal data
- Encode variables (convert Severity, Side, State, year to factors)
- Assign median to NAs (Population_County, Drive_County, Humidity..., Pressure.in.)

Clean Data: Remove logical predictors

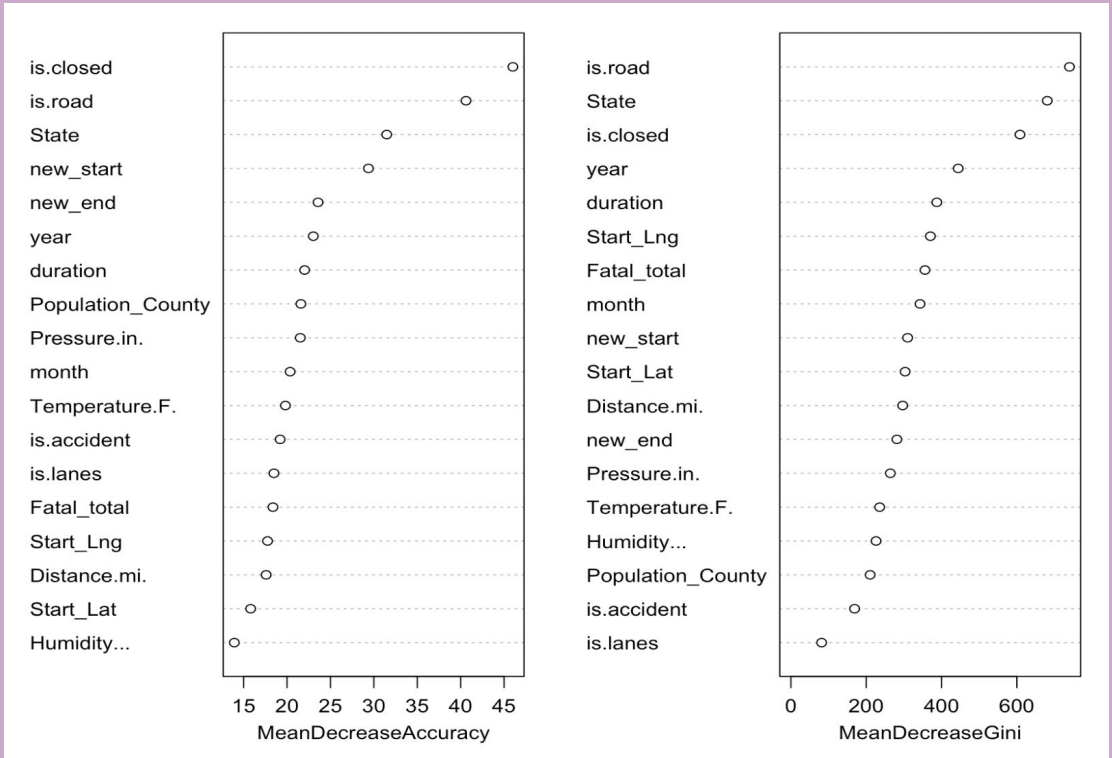
Bump	Station
Roundabout	Side
Amenity	Railway
Give_Way	Traffic_Signal
No_Exit	Crossing
Traffic_Calming	Junction
Turning_Loop	Drive_County
Stop	



Clean Data: New Data Set

Model Data: Random Forest Model (Full)

model.rf 18



Variable Importance Plot

mtry = 4
18 predictors

METHOD

Construct random forest model using all the predictors from training data.

Model Data: Random Forest Model (Full)

CONFUSION MATRIX

	MILD	SEVERE
MILD	31079	403
SEVERE	1530	1988

MISCLASSIFICATION ERROR RATE: 5.52%

Variable Importance Plot

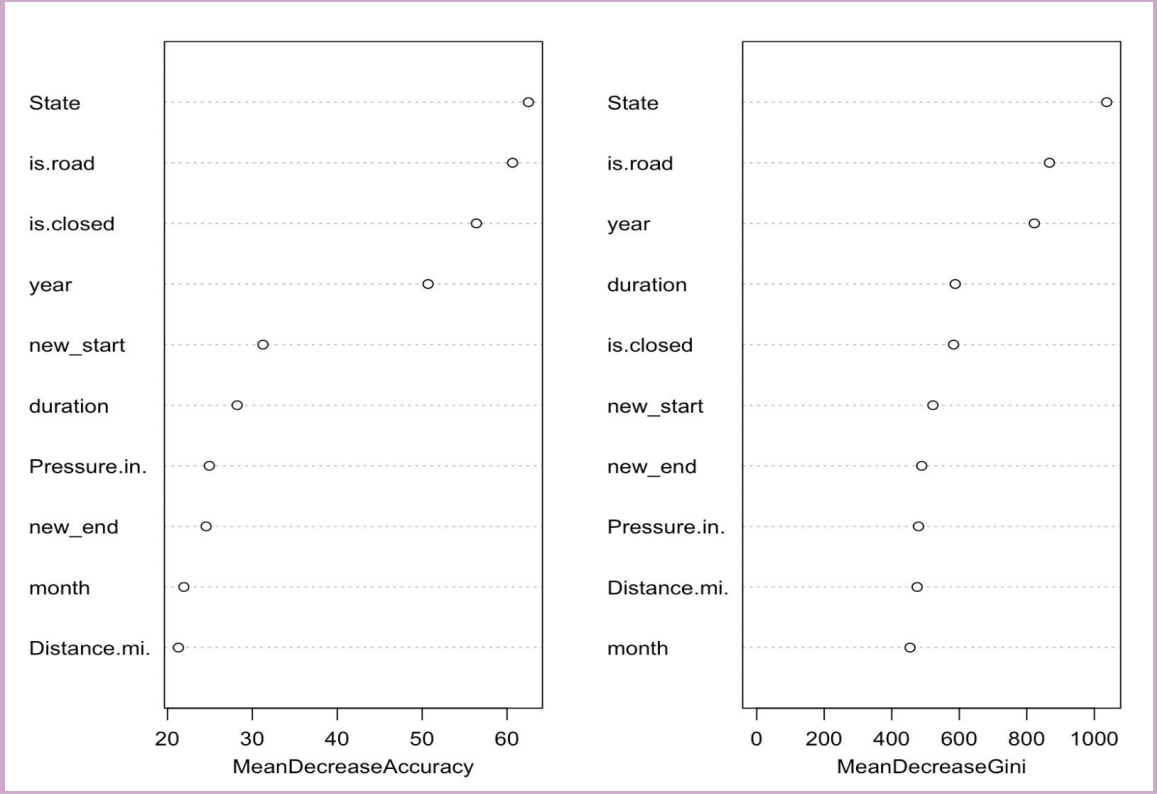
mtry = 4
18 predictors

METHOD

Construct random forest model using all the predictors from training data.

Model Data: Random Forest Model (10 predictors)

model.rf 10



Variable Importance Plot

mtry = 4

10 predictors

METHOD

Construct random forest model using 10 predictors from training data.

Model Data: Random Forest Model (10 predictors)

CONFUSION MATRIX

	MILD	SEVERE
MILD	30917	565
SEVERE	1543	1975

MISCLASSIFICATION ERROR RATE: 6.02%

Variable Importance Plot

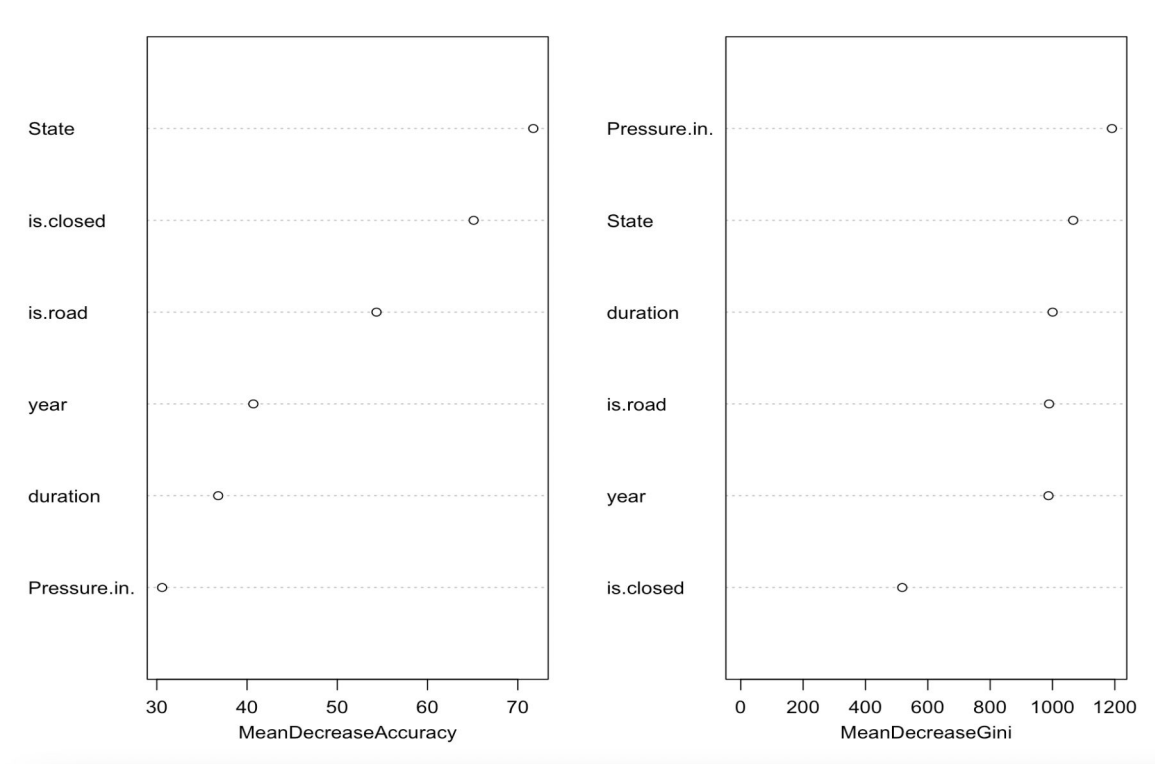
mtry = 4
10 predictors

METHOD

Construct random forest model using 10 predictors from training data.

Model Data: Random Forest Model (6 predictors)

model.rf 6



Variable Importance Plot

mtry = 4
6 predictors

METHOD

Construct random forest model using 6 predictors from training data.

Model Data: Random Forest Model (6 predictors)

CONFUSION MATRIX

	MILD	SEVERE
MILD	30681	801
SEVERE	1624	1894

MISCLASSIFICATION ERROR RATE: 6.93%

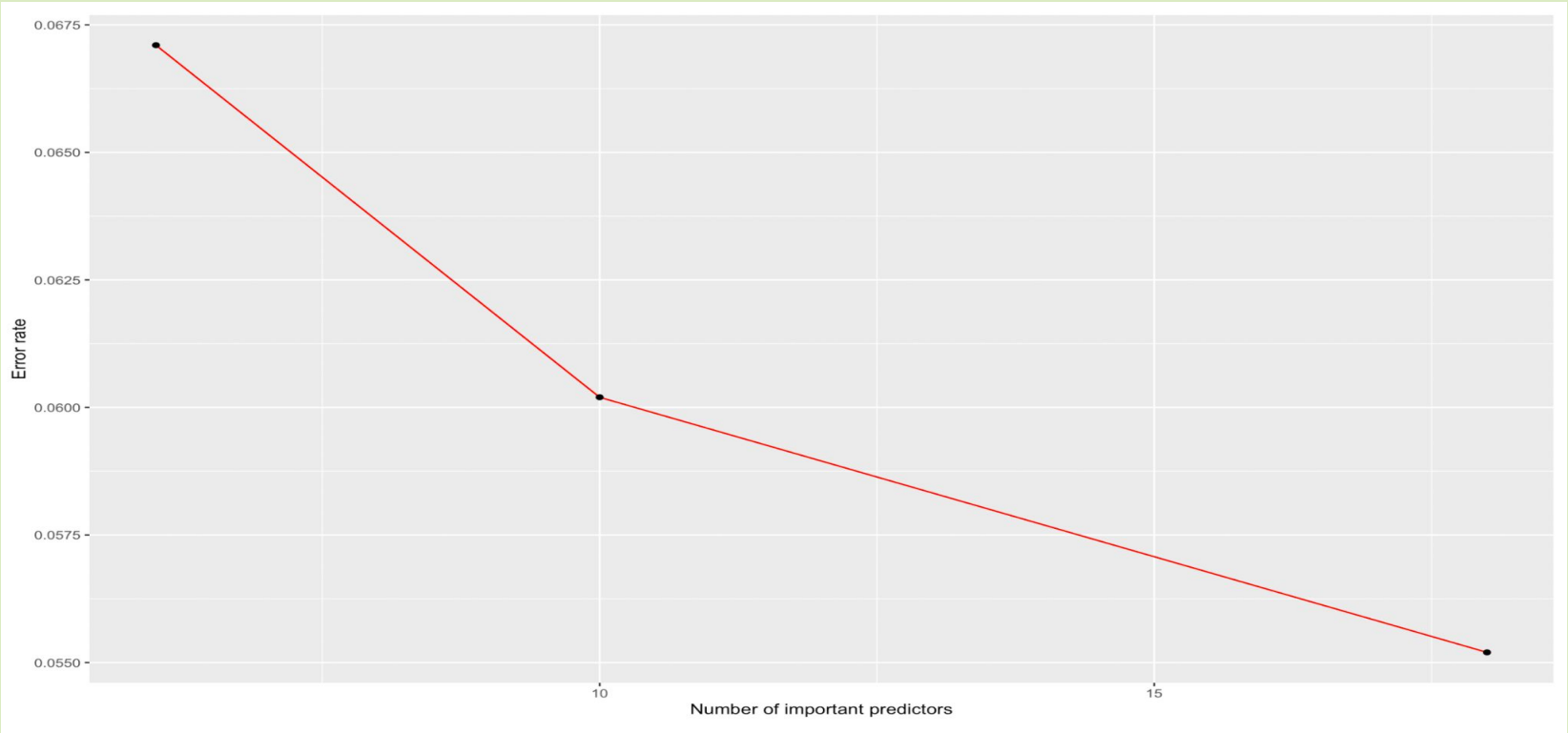
Variable Importance Plot

mtry = 4
6 predictors

METHOD

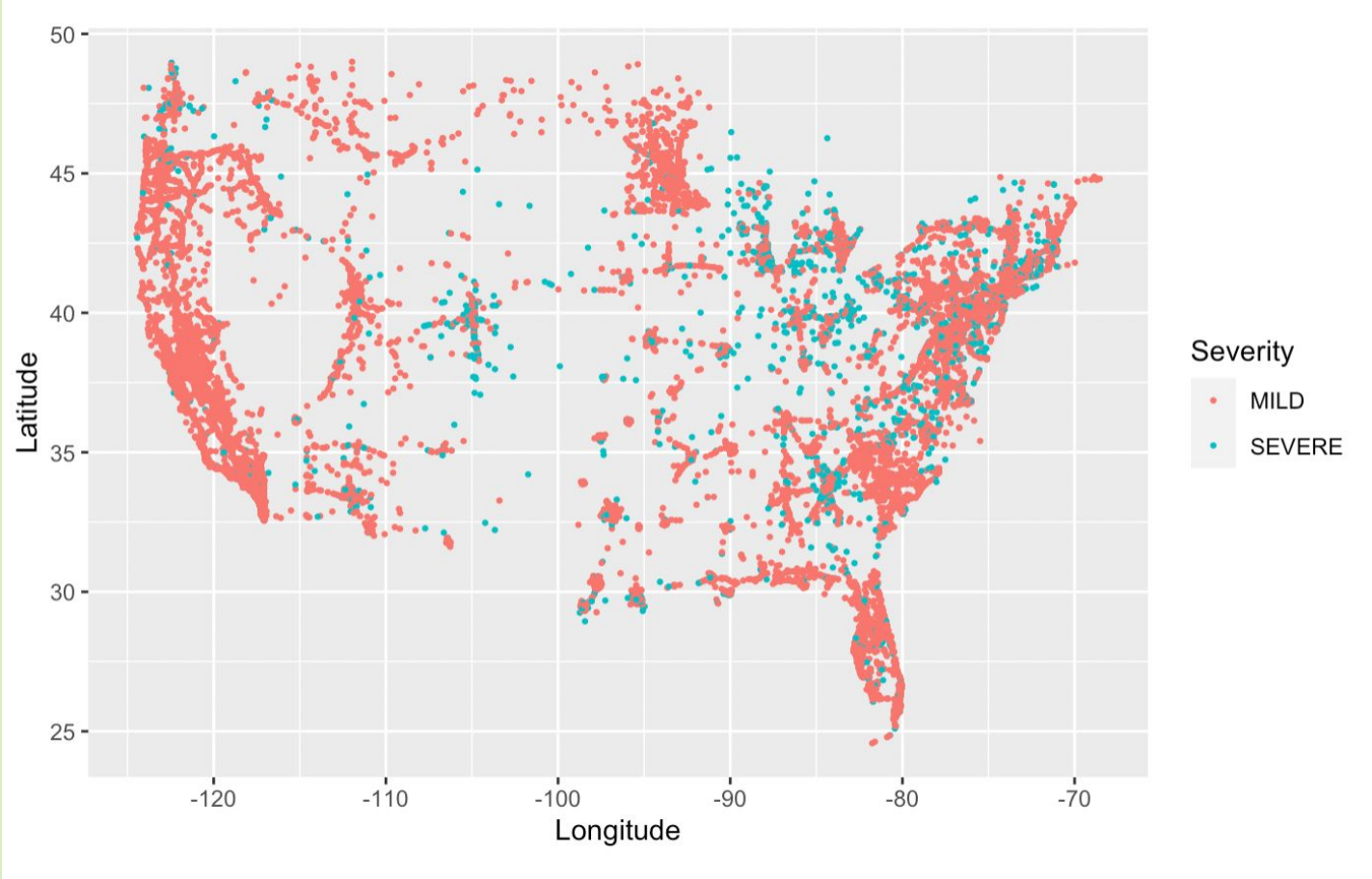
Construct random forest model using 6 predictors from training data.

Model Data: Number of important predictors vs error rate



Model Data:

Graph of geographic dispersion of accidents by severity levels



Model Data: Generalized Linear Model (GLM)

Confusion Matrix

	MILD	SEVERE
MILD	31009	1906
SEVERE	473	1612

Misclassification error rate: 6.78%

Model Data: Final Model

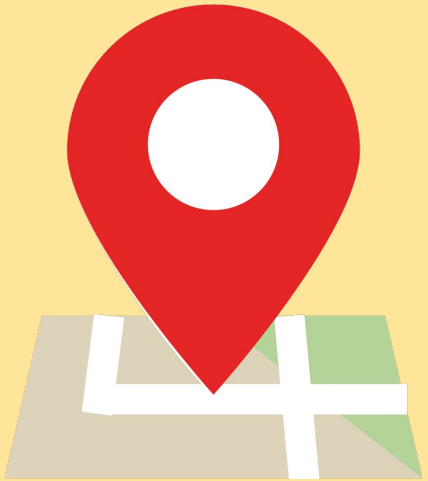
	RF 18	RF 11	RF 6	GLM
MCR	5.52%	6.02%	6.93%	6.78%

CHOOSING THE MODEL

When choosing the model, we observe the misclassification rates. Comparing our models' MCRs, we decided to choose the random forest model using 18 predictors.

RESULTS & DISCUSSION

03



Final Model Key Points

Model

Random Forest
(mtry = 4)

Observations

35000 Accidents

Predictors

18 Predictors

MCR

5.52%

Rank

2nd Place

Final Kaggle Score

0.9459

Most Important Predictors

is.road

Logical variable:
Did the word road appear in description variable

Year

Categorical Variable:
Year the accident occured

is.closed

Logical variable:
Did the word closed appear in description variable

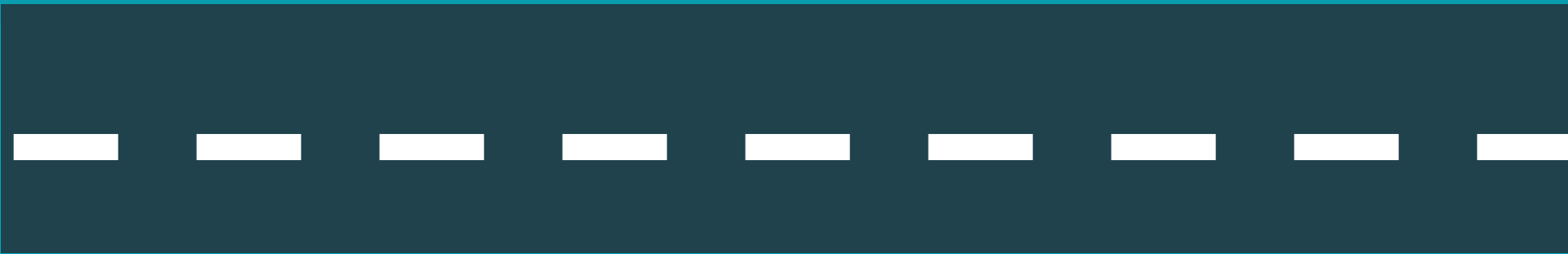
04

LIMITATIONS & CONCLUSION

Limitations

Increasing accuracy required more trees

Computationally expensive



Conclusion

Low misclassification rate

Predictive model

Detailed relationship between key factors and accident severity can be further studied.

References

- Almohalwas Stats 101C Lectures and Discussions
- Almohalwas, Akram. “Predicting Car Accidents’ Severity.” https://www.kaggle.com/competitions/predicting-car-accidents-severiy
- Insurance Institute for Highway Safety (IIHS) – https://www.iihs.org/about-us

THANK YOU