Predicting Severity of Car Accidents

By: Merve Dumlu, Regina Thahir, Tori Wang, And Xingbo Zhao

Table of Contents

Introduction

Problem Statement

Car Accidents are a common problem, and one that almost all individuals will experience. Severe cases can cause significant damage and can even be fatal with 39,404 fatal car crashes in 2018.

In this presentation, we explore our attempts to predict severe car accidents.

There is a training dataset and a testing dataset.

The training data contains

- **35,000** observations and
- **44** variables.

Part 2
Methodology

Process

1

Data Cleaning

*Removed & imputed **missing** values
*Removed variables containing similar information of time: created new predictor “**Year**”
*Created new boolean predictors for **keywords** in “Description”: blocked, closed & caution.

2

Model Data

4 different models tried:

1. **Knn** classification
2. **Logistic** Regression
3. Classification **Tree**
4. **Random Forest**

3

Analyze and Compare

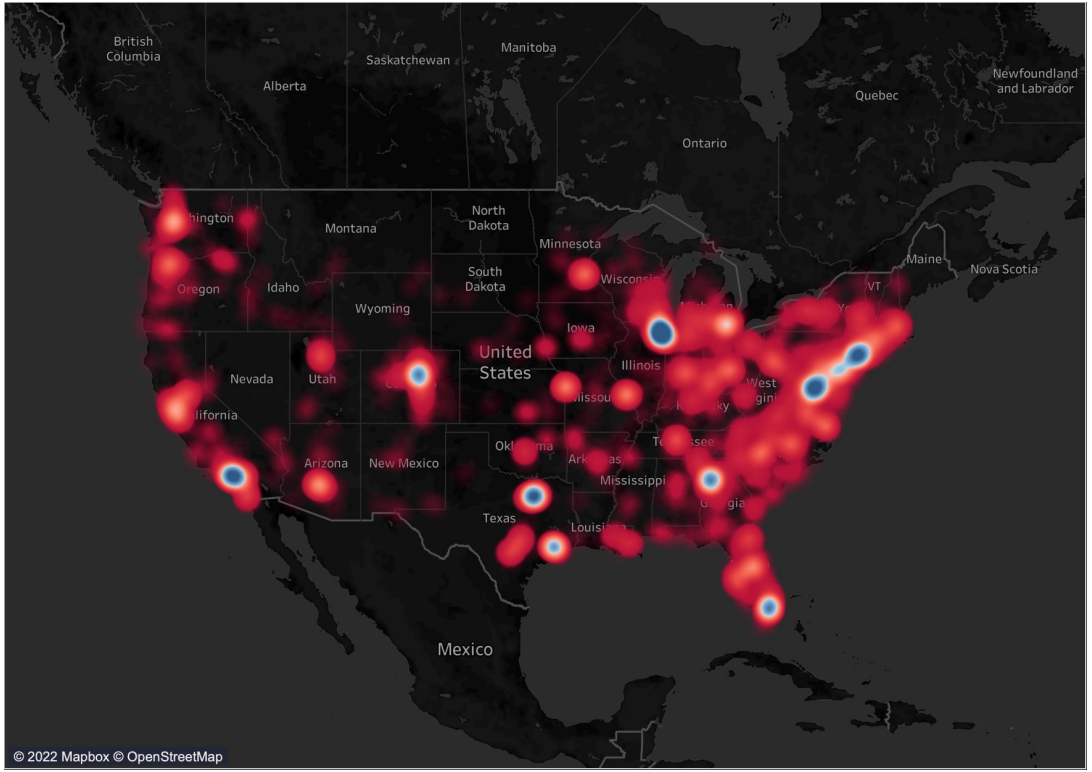
of Predictors involved
Confusion Matrix
Looked at accuracy rates from Kaggle

Data Cleaning: Removing NA values

- 14918 NA values in training dataset (0.008%)
  - Omitted all NAs since proportion < 5% of entire dataset

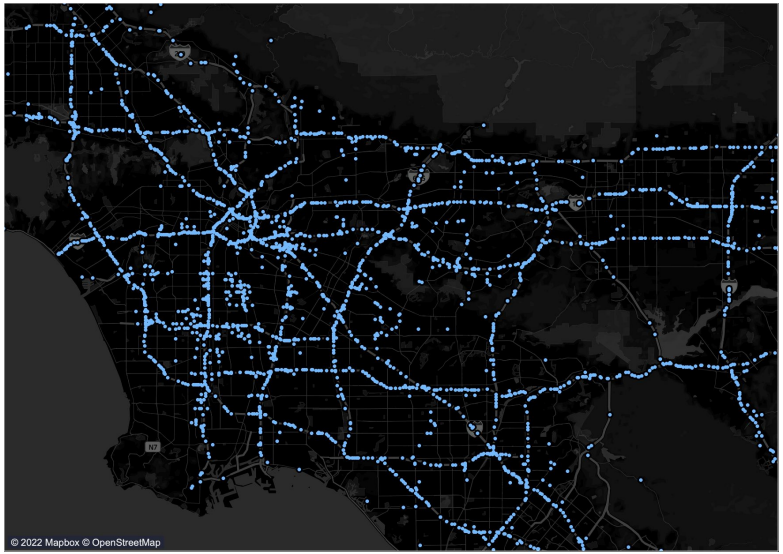
- Imputed missing values to avoid having NAs in our result:
  - For numerical variables, NAs => column mean.
  - For categorical variables,
    1) Converted into factors
    2) Variables with too many levels were not used in our model

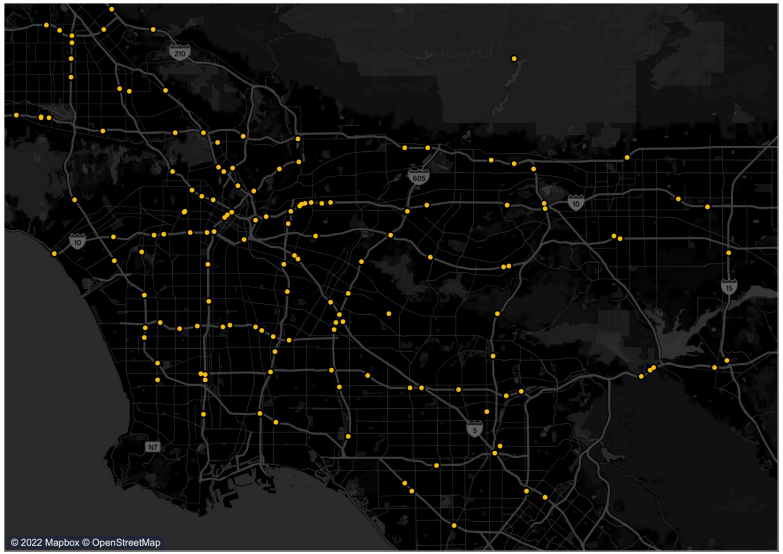
Exploring severity ~ locations



Severity Heat Map

Mild vs Severe Cases on LA Highways





Mild Cases

Severe Cases

Most Important Predictor - “Description”

Key word | severity rate	MILD	SEVERE
Closed	0.003676471	0.996323529
Blocked	0.8049303	0.1950697
Caution	0.9998574077	0.0001425923

Model Type 1: KNN Classifier

We create a K Nearest Neighbor Model and test different values of K.

- Lower K is less flexible
- We determine that K = 40 is optimal

K = 40 KNN model using:
Start_Lat, Start_Lng , Junction, Temperature.F. , Wind_Chill.F., Humidity…, Pressure.in., Visibility.mi., Civil_Twilight, Caution, Blocked, and Closed

Training misclassification rate: 0.06781874
Accuracy rate of 0.90035 on Kaggle

Confusion Matrix

		Predicted
Actual	MILD	SEVERE
MILD	27027	1903
SEVERE	71	106

Model Type 2: Logistic Regression

We construct a logistic regression model on predicting Severe accidents(0 for MILD, 1 for SEVERE) with variables that we initially find a correlation with the response.

Predictors:
Start_Lat, Start_Lng , Junction, Temperature.F. , Wind_Chill.F., Humidity…, Pressure.in., Visibility.mi., Civil_Twilight, Caution, Blocked, and Closed

Confusion Matrix

		Predicted
Actual	MILD	SEVERE
MILD	27096	1467
SEVERE	2	542

Training misclassification rate: 0.05046896
Accuracy rate of 0.915 on Kaggle

Model Type 3: Classification Tree

Using rpart to build classification tree

Important variables:
closed_boolean, State, Year, Distance.mi., End_Lng, Start_Lng, caution_boolean, Pressure.in.

Confusion Matrix

		Predicted	
Actual		MILD	SEVERE
	MILD	26922	176
	SEVERE	1170	839

Training misclassification rate: 0.04624317
Accuracy rate of 0.933 on Kaggle

Model Type 4: Random Forest

Full Model Attempt

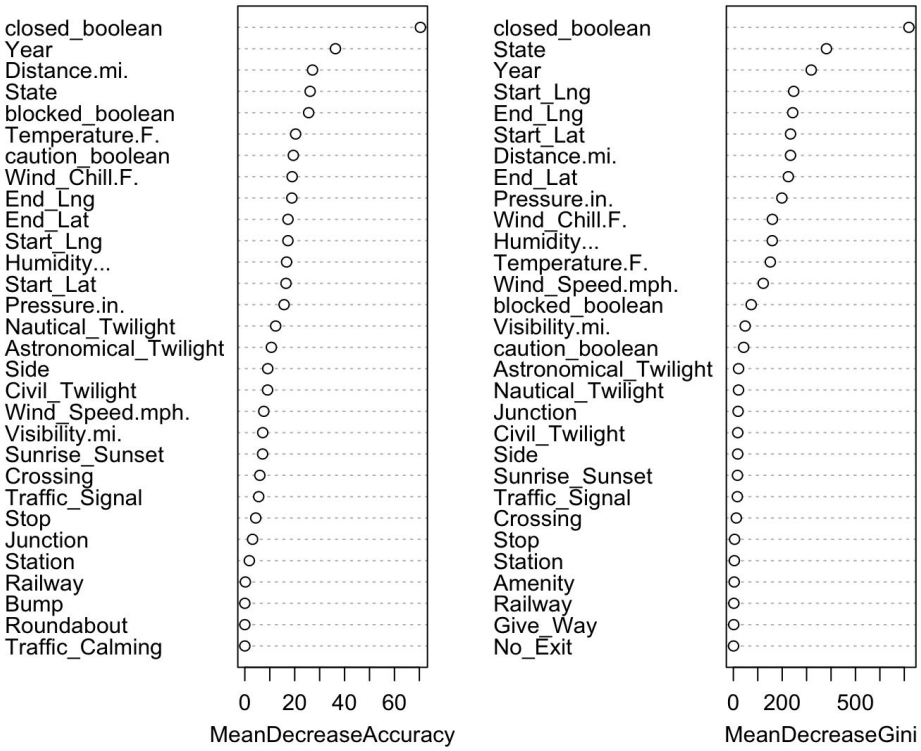
Number of trees: 128
No. of variables tried at each split: 7
OOB estimate of error rate: 4.26%
Accuracy: 95.74%

Confusion Matrix

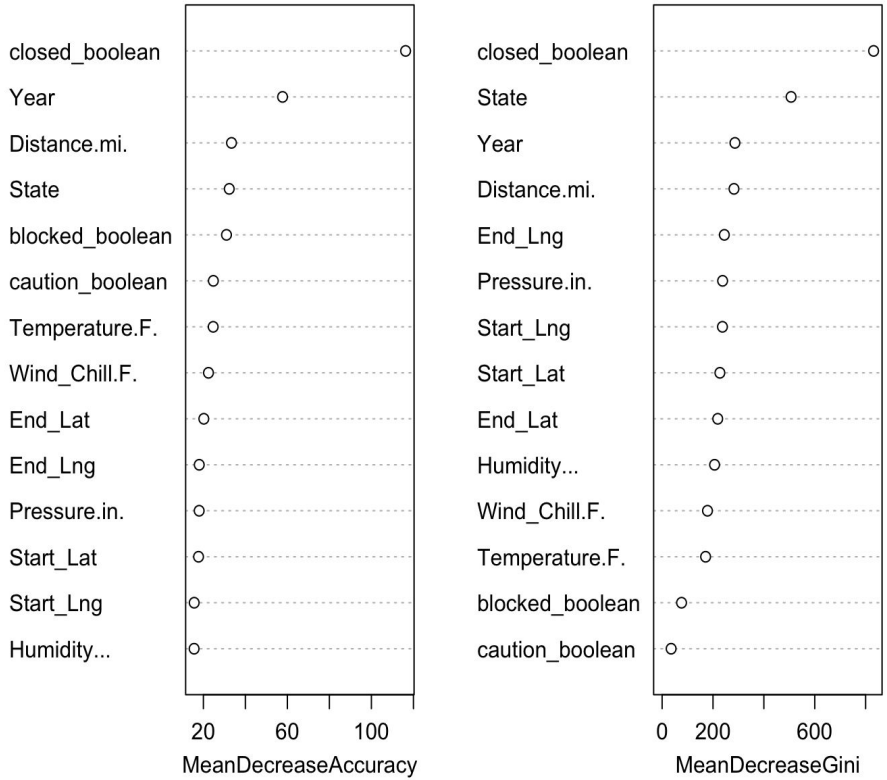
Actual	Predicted: MILD	Predicted: SEVERE
MILD	27098	0
SEVERE	3	2006

Training misclassification rate: 0.000103068
Accuracy rate of 0.93857 on Kaggle

Random Forest Model



Final Random Forest Model



Model Type 4: Random Forest

Final Model Attempt
Subset of 15 most important predictors

Number of trees: 128
No. of variables tried at each split: 3
OOB estimate of error rate: 4.34%
Accuracy: 95.66%

Confusion Matrix

Actual	Predicted: MILD	Predicted: SEVERE
MILD	27098	0
SEVERE	5	2006

Training misclassification rate: 0.00017178
Accuracy rate of 0.93857 on Kaggle

Part 3 Results

Model Used: Random Forest Model

Kaggle Ranking/Score: Rank 10 (0.93857)

#Predictors: 15

Important Predictors

Dates

Year, Month

Location

State, County

Weather

Year, Month

Description Key words

Blocked, closed, caution

Conclusion

Limitations and Setbacks

Data Cleaning- Missing Values

- Replacing numerical predictor's NAs values with the mean may have skewed some variables

Model

- Some variables were not very influential for our GLM and KNN
- Complexity of Random Forest model

Final Thoughts

Our final score of 0.93857 passes the 0.9 threshold of cases that are SEVERE by nature. We are able to gain valuable insight into the predictors of classifying cases including interesting finds within Description and Timestamp variables.

Further work can be done to analyze variables such as county, street, zipcode, and city, as these may provide further insights.