

Predicting Severity of Car Accidents

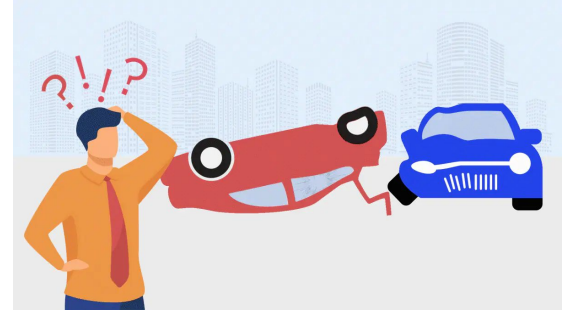




Table of Contents:

1. Introduction
2. Data Overview
3. Methodology
4. Conclusions

Severity of Car Accident



- Although the fatalities in car accidents has been significantly declined in the past decades, there were still a lot of fatalities caused by severe crashes.
- By being able to predict the severity of a vehicle accident, it will be helpful for the local traffic enforcement and fire department to rescue and prevent those fatalities.
- Our goal is to have the model that accurately estimate the severity based on the known predictors



Data Overview

Missing Values and New Variables



Car Accident Data Set

- 43 variables
- 35,000 observations for training and 15,000 observations for testing



Missing Data

- Temperature: 2.297%
- Wind_Chill: 16.189%
- Humidity: 2.42%
- Pressure: 1.917%
- Visibility: 2.349%
- Wind_Speed: 5.343%

Only Wind_Chill has missing values more than 10% and also Wind_Chill and Temperature are high correlated, so we decided that it will be beneficial if we drop the variable Wind_Chill. The rest of the missing values we impute with estimated values.

Subsetting Predictors

With so many different predictors and overlap, we wanted break up the predictors and subset to analyze which specific areas are classified as SEVERE and which areas are classified as MILD.

Three stand out predictors we found are: `Season`, `WeatherCon`, and `stateLevel`.

New Variables from Description and Street Name and Weather

- Highway = whether the accident occurred on highway
- roadBlocked, Incident = whether the word “block”, “incident” appeared in the description
- WeatherCon = whether the weather condition is clear or overcast or not

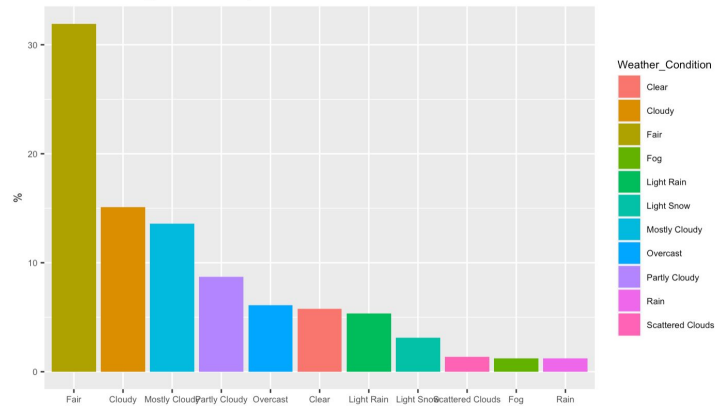
roadBlocked	Mild	Severe
False	29156	2962
True	2326	556

Incident	Mild	Severe
False	21122	3329
True	10360	189

WeatherCon	Mild	Severe
False	29356	2522
True	2126	996

New Variables - Weather Con

Severe: Percentage of accidents by weather condition

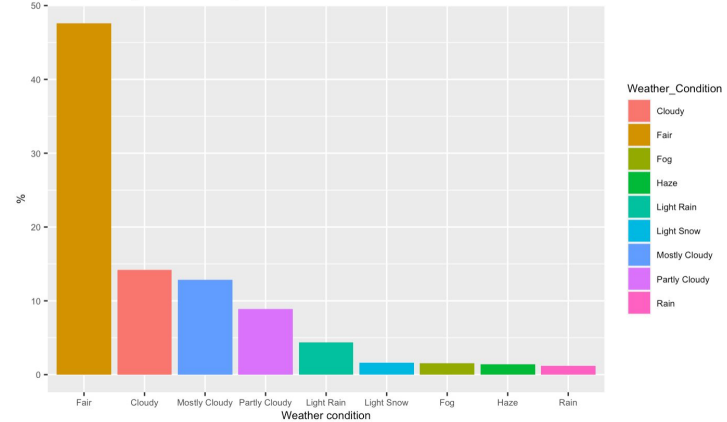


Weather_Condition
<chr>

percentage
<dbl>

Fair	31.932773
Cloudy	15.126050
Mostly Cloudy	13.602941
Partly Cloudy	8.718487
Overcast	6.092437
Clear	5.777311
Light Rain	5.357143
Light Snow	3.098739
Scattered Clouds	1.365546
Fog	1.207983

Mild: Percentage of accidents by weather condition



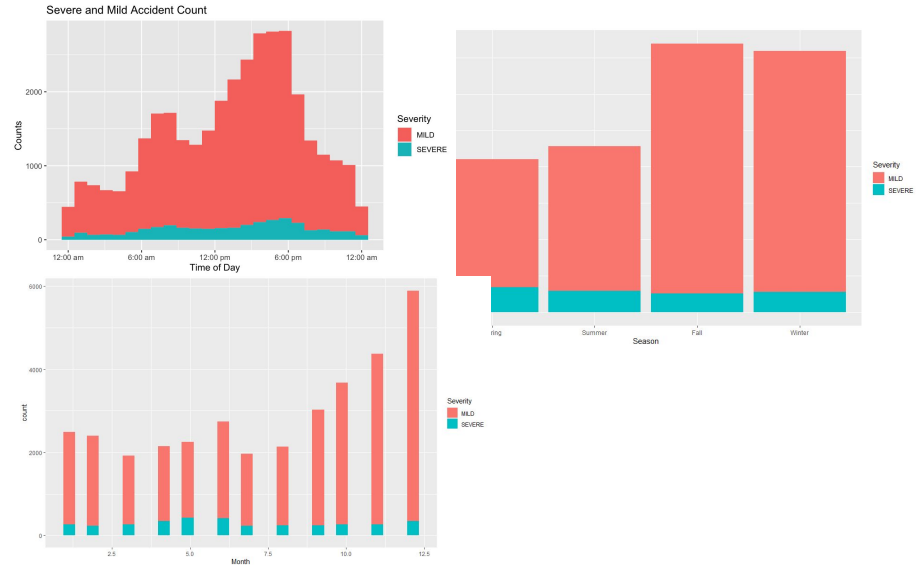
Weather_Condition
<chr>

percentage
<dbl>

Fair	47.624866
Cloudy	14.220254
Mostly Cloudy	12.857034
Partly Cloudy	8.874294
Light Rain	4.379869
Light Snow	1.634336
Fog	1.550328
Haze	1.386131
Rain	1.210478

New Variables from Timestamps

- Year, Month, Hour = the year, month, hour of the starting time of the accident
- Season = spring(3, 4, 5), summer(6, 7, 8), fall(9, 10, 11), winter(12, 1, 2)
- Morning_rush = whether happened in morning rush hour (7-9AM)
- Evening_rush = whether happened during evening rush hour (4-6PM)



New Variables from States

- stateLevel = whether the state is in top two states with highest amount of accident cases (CA, FL), the following seven states (TX, OR, VA, MN, NY, PA, SC), or the rest of states
- stateLevel2 = whether the state is in the top seven states with highest proportion of severe accident and having more than 100 cases (IL, SD, WI, CO, GA, MA, OH), or the rest of the states
- StateParty = whether the state is blue or red states

stateLevel	Mild	Severe
Tier 1	14108	593
Tier 2	8698	954
Tier 3	8676	1971


stateLevel2	Mild	Severe
Tier 1	1079	801
Tier 2	30403	2717

StateParty	Mild	Severe
Blue	24039	2468
Red	7443	1050



Methodology

Data Cleaning and Modeling



Logistic regression

We tried to build a logistic regression model with the best predictors selected from stepwise method with lowest BIC. The best predictors are:

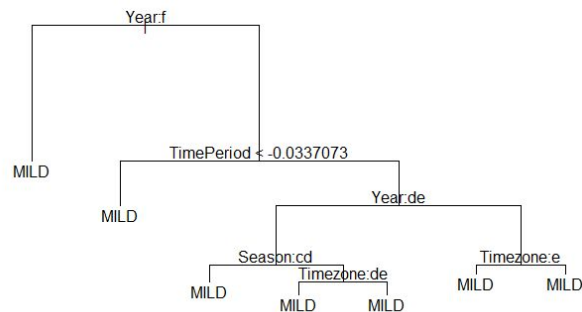
Side, Timezone, Nautical_Twilight, stateLevel, Year, Season, WeatherCon, Distance.mi.

The misclassification rate is 11.55% and we scored 0.89724 on the testing dataset

Actual Predicted	Mild	Severe
Mild	30959	3004
SEVERE	523	514

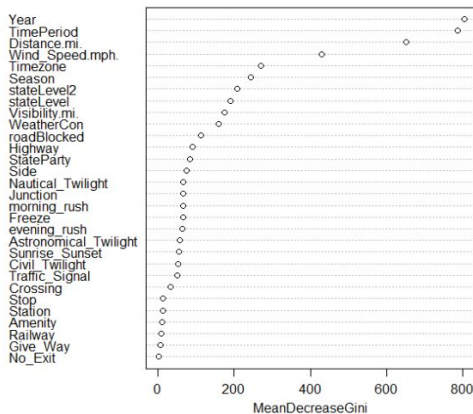
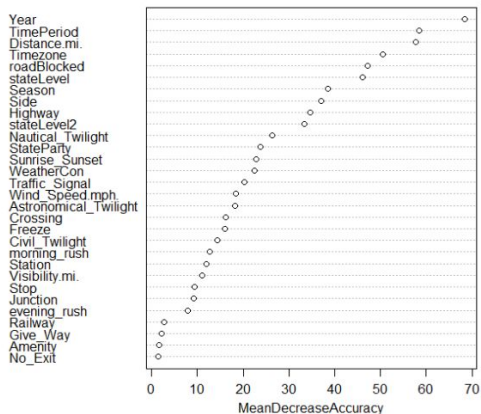
Tree Model

Tree model results underfitting as it labeled every case as “MILD” accidents. It has misclassification rate for training dataset 10.05% and scored 0.89848 on testing dataset.



Random Forest (Full)

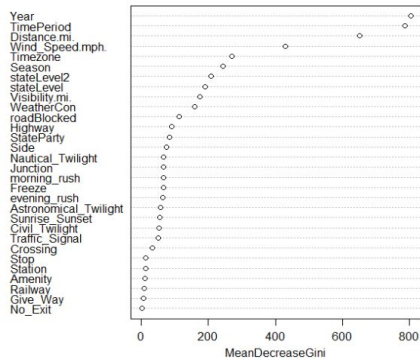
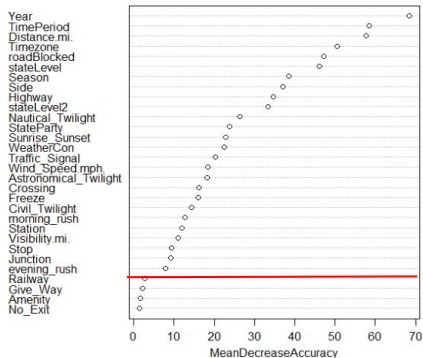
Using Random Forest with full model we got misclassification rate for training dataset 0.0258 and scored 0.90026 on the testing dataset.



Predicted \ Actual	Mild	Severe
Mild	31475	897
SEVERE	7	2621

Random Forest

We found the best mtry is 17, so we picked those variables that have high importance score on both Mean Decrease Accuracy and Mean Decrease Gini and fit with Random Forest as mtry = 17. The resulting two models both scored 0.9192 on testing dataset



Predicted \ Actual	Mild	Severe
Mild	31469	91
SEVERE	13	3427

Predicted \ Actual	Mild	Severe
Mild	31481	18
SEVERE	1	3500

Boosting

With the importance variables, we constructed boosting models and altered the parameters such as number of tree and the maximum depths. By using Boosting and tuning the parameters, we are able to get our best model with three level of depth and fitting 150 trees. The training misclassification rate in around 8.22% and scored 0.91697 on testing dataset.

Predicted \ Actual	Mild	Severe
	Mild	SEVERE
Mild	30995	2390
SEVERE	487	1128

Best Model

	Logistic	Tree	Random Forest (Full)	Random Forest (mtry=17)	Random Forest (mtry=17, Gini)	Boost
Training MCR	11.55%	10.05%	2.58%	2.97%	0.05%	8.22%
Testing Score	0.89724	0.89848	0.90026	0.9192	0.9192	0.91697

The models with highest testing score are the random forest models with mtry = 17 and the most importance variables from Mean Decrease Accuracy and Mean Decrease Gini. The model is not easy to interpret and take time to perform; however, it is more flexible and accurate than the others.



Conclusion

Limitations, Challenges, and Improvements



Limitations and Challenges



- The COVID-19 pandemic introduces a new dimension of variability.
- Accidents are unpredictable, we are only classifying them not predicting when they occur.
- Missing values.
- Variable creation.
- Some states have small amounts of observations.
- Small number of severe accidents.
- In our training data, trends for severe and mild accidents were very similar.

Conclusion

Year, distance, time period, and location are strong indicators of whether an accident is severe or not. Having more data would be helpful in refining the model.



References

US Car Accident Data from Kaggle - <https://www.kaggle.com/competitions/predicting-car-accidents-severity>

Driver Demographics - <https://driving-tests.org/driving-statistics/>

Images from Google - <https://images.google.com/>

Thank you and drive safe :)

