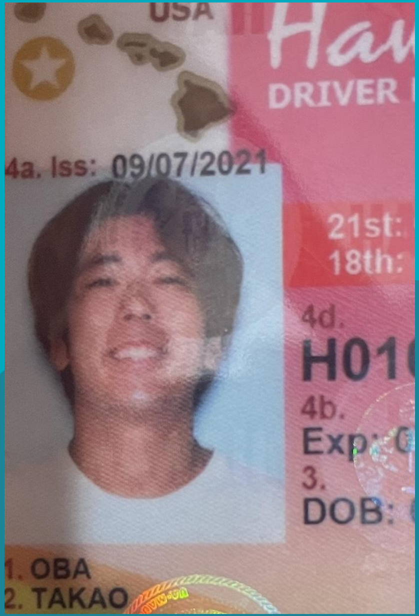
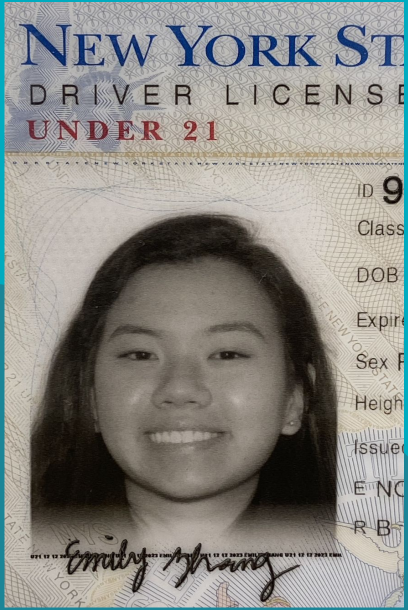
Predicting Severity of U.S. Car Accidents

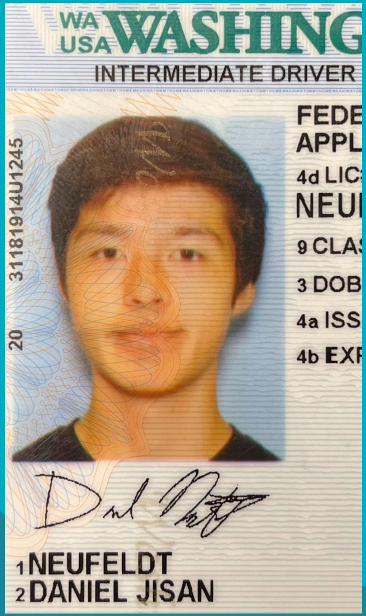
Team Emily



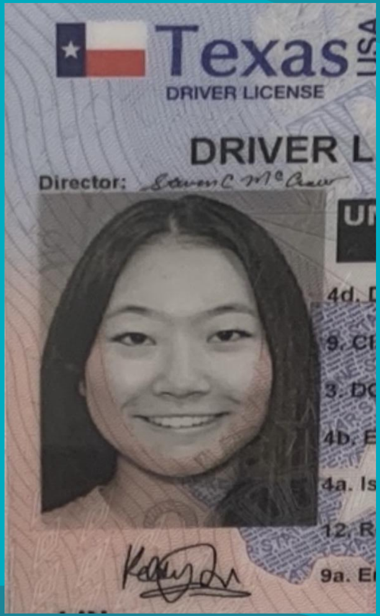
Takao Oba



Emily Zhang



Daniel Neufeldt



Kelsey Lin

Table of Contents

Introduction

Topic overview and data set context

U.S. Car Accidents

- Vehicle accidents are the most common cause of death of teenagers and young adults.
- Vehicle accidents can result in injury, disability, death, and property damage.
- Accidents result in financial costs to both society and the individuals involved.

By analyzing this dataset, we can gain a better understanding of what is happening on the roads and find factors that best predict the severity of an accident.

Data used for the Model

Accidents Data Set

Collects countrywide car accident from February 2016 to December 2021

Training

- **35000 rows** — Observations
- **44 columns** — Predictors plus the Y variable

Testing

- **15000 rows** — Smaller Observations
- **43 columns** — Predictors used to predict the severity

Methodology

Data cleansing and modeling

Our Methodology Process

Cleaning the Data

- Assessing the NAs
  - Imputation
- Examining Outliers

Determining Predictors

- Determine significance of predictors
- Creating new predictors

Model Creation

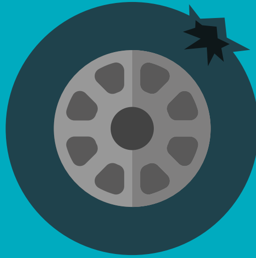
- KNN
- Logistic Regression
- Decision Tree
- Random Forest

Model Comparison

- Split Training Data
- Best Kaggle Score

Cleaning the Data

Removing the NAs



Categorical

Manual Imputation

- Extremely Small number of NAs for predictors we used
- For example, we manually imputed zip codes based on location

Numerical

missForest imputation vs Median

- missForest()
- Median Temperature using median(na.rm = T)
- We will get into this later in this presentation

Determining Correlation of Predictors

Numerical

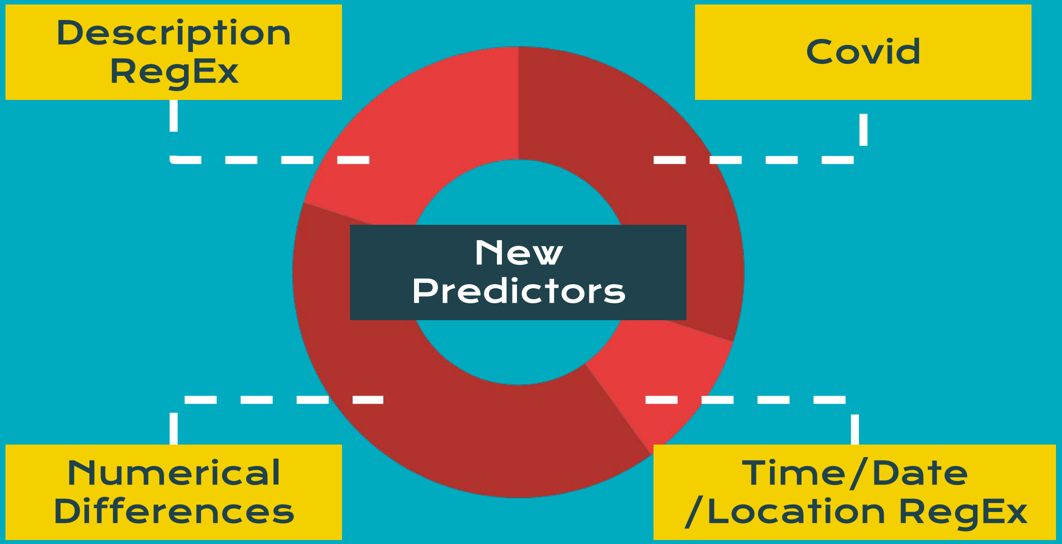
- Observed the data and determined which predictors were numerical
- Created density plots to see if the difference in the graph was significant enough

Categorical

- Created stacked bar plots to see correlation
- Removed variables that had too many categories (States, Zip, etc)
- Deduce if we should use these predictors or not


They Are All Bad!
BUT, we can change that...

Creating New Predictors



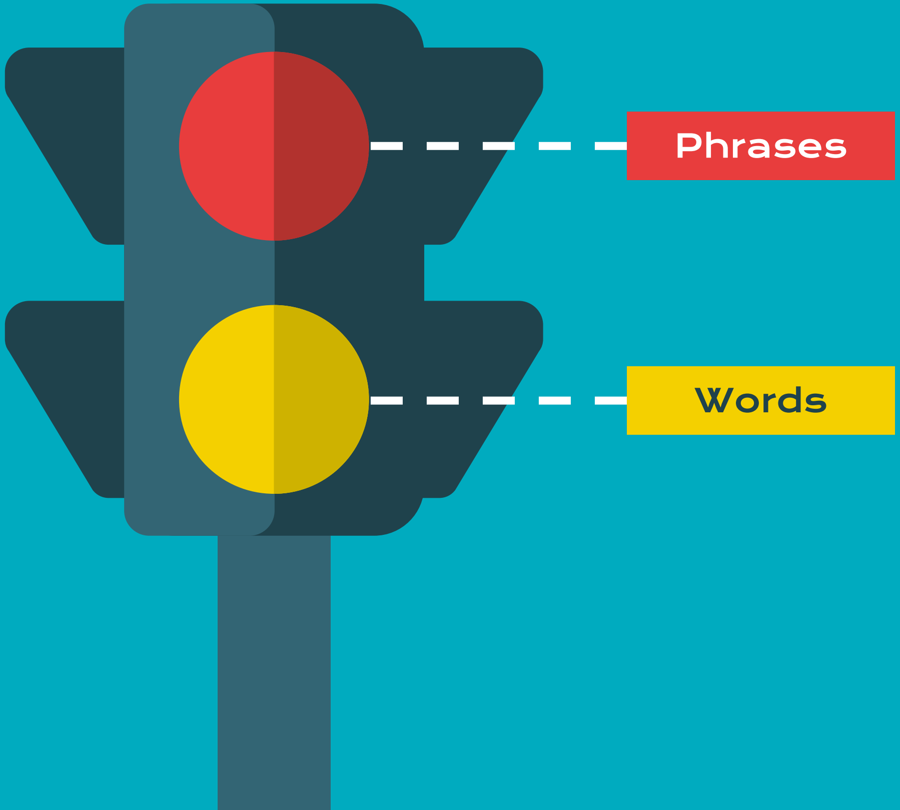
Scan Description for key words/phrases

Precovid, Covid, Postcovid

Differences between predictors, rather than start/end times

Holidays, rush hour, and State/Zip Code, etc.

Description Regular Expression



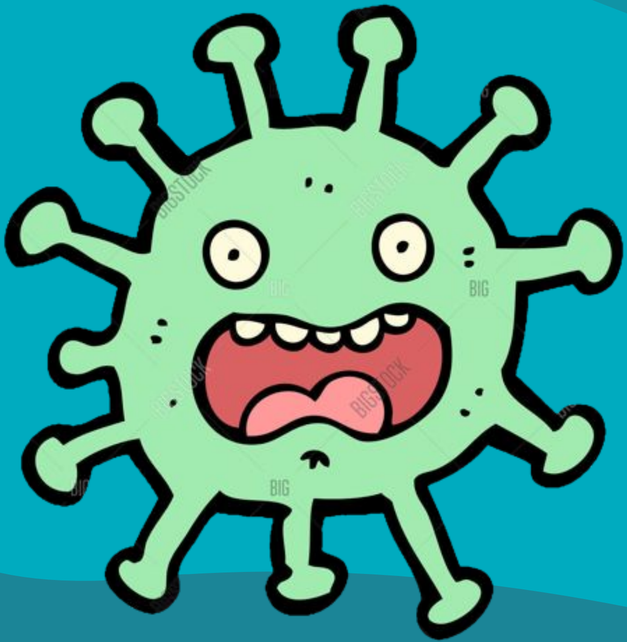
- “Road Closed”
- “Closed Due”

- “Road”
- “Closed”
- “Accident”
- “Blocked”
- “Incident”
- “Caution”
- “Alert”

Covid Predictor

Split 'Date' column into three new variables

- Precovid (before April 1st, 2020)
- Covid (between April 1st, 2020 and before Jan 1st, 2021)
- Postcovid (after Jan 1st, 2021)



Numerical Differences

Time

Latitude

Longitude

Procedure

- Utilized str_split() and difftime() functions
- 4.26 hour difference

Procedure

- Subtracted Start_Lat from End_Lat
- Found absolute value to get difference and then scaled

Procedure

- Obtained scaled difference in longitude in car crash.
- Subtracted Start_Lat from End_Lat

Time, Date, & Location - Regular Expression

Time

Rush hour:
- Calculate difference in time (timediff)
- Look into specific time frame

Date

Holiday Season:
- Notable national holidays in US

Most Car Accident Months:
- More people are on the road during summer months

Location

States, Zipcode, Insurance, Teen Drivers, Car Crashes Fatalities Teen:
- Use external sources to extract notable locations

Model Creation

Types of Models That We Created		
01	KNN	Uses only numerical predictors
02	Logistic Regression	Not as accurate as decision trees
03	Decision Tree	Random Forest performed better
04	Random Forest	Our Final Model!

KNN Model (n = 5)

- Only uses numerical predictors
- We used:
  - Time Difference
  - Latitude Difference
  - Longitude Difference
  - Temperature
- Scaled all variables
- The expected misclassification rate was a little high, so we should explore other modeling methods

Actual Sample Testing Severity

KNN	MILD	SEVERE
MILD	4289	399
SEVERE	200	112

This model resulted in an expected misclassification rate of 11.9%

Logistic Regression

- Able to incorporate our categorical predictors
- Because we are trying to find a binary result (either mild or severe accidents), this is an appropriate model.
- We chose to not use this model because the misclassification rate was higher than using a random forest

Actual Sample Testing Severity

Logistic Regression

	MILD	SEVERE
MILD	4438	325
SEVERE	51	186

The expected misclassification rate for this model was 7.52%

Decision Tree

- We constructed a decision tree with multiple branches but only one branch resulted in “SEVERE” (multiple predictor’s leaf nodes is “MILD”)
- In an attempt to solve this, we pruned the tree, however the result was the same.
- Thus, we moved to look at an alternative model

Actual Sample Testing Severity

Tree

	MILD	SEVERE
MILD	4470	508
SEVERE	19	3

This model resulted in an expected misclassification rate of 10.54%

Random Forest missForest Imputation

- Decision Trees showed a fairly low misclassification rate, however a small change in the data can largely charge the structure of the tree
- To overcome this disadvantage, we constructed the model using random forest.
- Further, for the numerical predictors missForest imputation was implemented in an attempt to better capture the frame of the missing values.

Actual Sample Testing Severity

Random Forest

	MILD	SEVERE
MILD	4462	298
SEVERE	33	207

The expected misclassification rate for this model was 6.62%

Random Forest Median Imputation

- Still able to utilize all of the relevant predictors
- Imputed the missing data with the median of respective column
- We chose this model over the missForest imputation random forest model because this model produced a better misclassification rate

Actual Sample Testing Severity

Random Forest

	MILD	SEVERE
MILD	4539	289
SEVERE	4	168

The expected misclassification rate for this model was 5.86%

Results & Discussion

Final constructed model analysis

Analysis: Key Parts

Final Model:
Random Forest
(mtry = 6)
Median Imputation

Observations: 50,000 accident records (35,000 from training, 15,000 from testing)

Predictors: 23 predictors (4 numerical, 19 categorical)

Accuracy Rating:
Kaggle Public Score: 0.93688
Kaggle Private Score: 0.93093

Discussion: Important Predictors

- Most relevant predictors were the newly created variables: Description RegEx, Numerical Differences, Covid, and Time/Date/Location RegEx

- The single predictor that contributed the most was Description RegEx. There are certain keywords and phrases in accident descriptions that are stronger indicators for the classification. In other words, classifying accidents as SEVERE is greatly attributed to the presence of words such as “road closed” for example.

Limitations & Conclusions

Setbacks, assumptions, and final words

Limitations

Lack of Predictors

Age - young drivers are more likely to get into car accidents
Gender - male drivers are at higher risk of fatal car accidents

Time Expenses

Due to using Random Forests, actually creating/testing the models were incredibly time consuming

Multicollinearity

Because we text mined many words/phrases, there could be a lot of overlap in variables

Predictor Accuracy

Certain data was drawn from external sources. There could be inaccuracies or biases in the used findings

Conclusions

01

Random Forest with Median Imputation had the lowest misclassification rate

02

The most significant predictors are description, covid dates, time difference, zip code, and state-death-rate

03

In order to improve our model, we would need to conduct further research

Citations

- https://injuryfacts.nsc.org/motor-vehicle/overview/crashes-by-month/
- https://www.perecman.com/blog/8-of-the-deadliest-holidays-for-driving/#:~:text=Memorial%20Day.,likely%20on%20Memorial%20Day%20weekend.
- https://www.carinsurance.com/Articles/car-insurance-rate-comparison.aspx