Group 5 Capstone
Juveriya Baig, Chris Haub, Angie Tran, Ryan Permenter

March 29, 2021

Topic: Accident Severity in Texas

TABLE OF CONTENTS

Inspiration and Overview

For our project, we chose to investigate how different weather conditions, as well as road types, affect traffic accident severity in Texas. We were excited to see how different parameters affected the severity of accidents. We thought it would be beneficial to know when the worst accidents happen so can be more careful when driving in those conditions.

We used a CSV data set from Kaggle that was then narrowed down to reflect accidents in Texas. The data points included severity, time of the accident, weather conditions, and location.

Machine learning was used to predict accident severity.

Data & Modeling Approach

The original dataset required some pairing down of data to focus on a specific area as the entire file was too large to work with efficiently. We chose to focus our efforts on Texas as that is where we live. Texas had about 350,000 data points. The severity score is on a scale of 1-4 while 98% of the data is either a 2 or a 3. In our model, we put emphasis on what factors make the severity a 1 or a 4.

Tableau

Tableau dashboards were created to visualize the data. The original CSV was too large so we used a sample. We chose a Texas-colored theme for the dashboard because we focused on one state. The dashboards visualized accident severity for 2016-2020, average accident severity for counties in Texas within that time, and weather conditions.

The first dashboard shows the number of accidents that happened in Texas in 2016-2020. You can see in the middle of 2020, with the shelter-in-place orders enforced, the number of accidents dropped to 3000 and increased again when the order was lifted after the month of August.

Secondly, the bubble chart demonstrates the number of accidents in the main cities of Texas. With a huge difference among numbers of accidents, big major cities' average severity level was not as high as small cities.

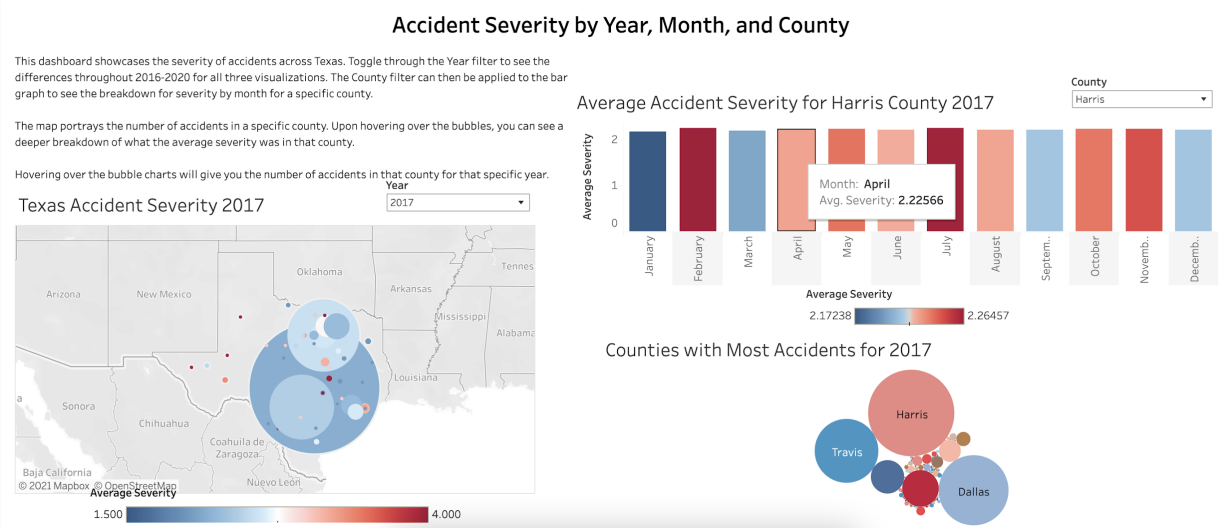
The bar graph indicates the relationship of how weather conditions affect the average severity of accidents in Texas.

The second dashboard showcases the severity of accidents across Texas. Toggling through the Year filter will show the differences throughout 2016-2020 for all three visualizations. The County filter can then be applied to the bar graph to see the breakdown for severity by month for a specific county.

The map portrays the number of accidents in a specific county. Upon hovering over the bubbles, one can see a deeper breakdown of what the average severity was in that county.

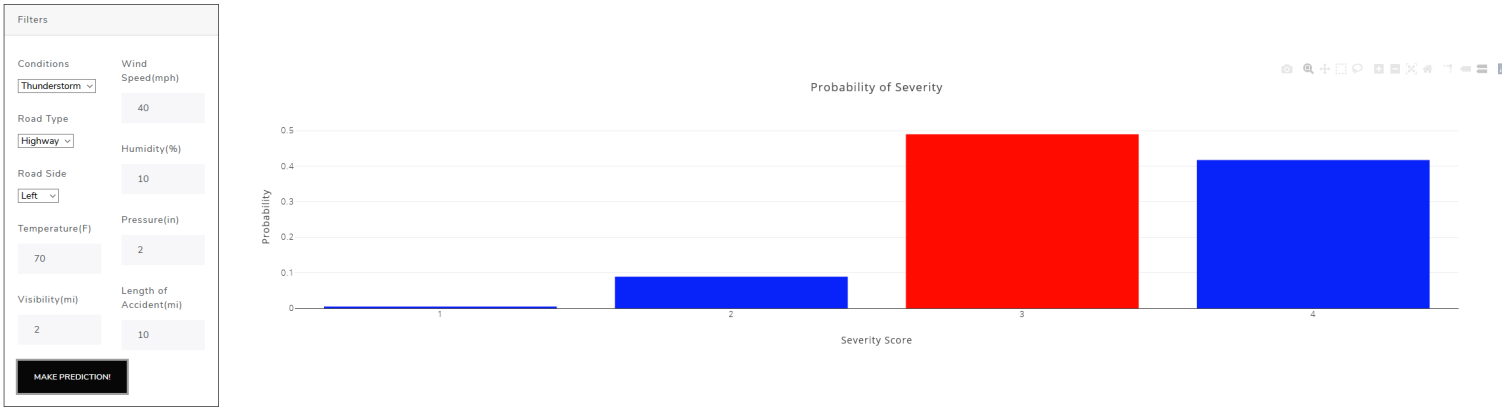
Hovering over the bubble charts will give you the number of accidents in that county for that specific year.

An interesting insight is comparing the years to 2020. Because of the pandemic, there wasn't much traffic, especially during the early months of lockdown. Some counties, such as Travis County, saw a decrease in the number of accidents. Others, such as Harris County, didn't see much of a difference when compared to previous years.



Machine Learning

The machine learning predictive classification model was designed to predict the severity of a car accident based on data surrounding the accident. This data includes weather condition, side of the road, precise location, distance length, time of day...etc. The severity was on a scale of 1-4. A severity score of 2 and 3 combined makes up 98 percent of all accidents. The model created tries to put emphasis on what conditions cause a 1 or a 4 severity. To do this we made three predictive models one to capture the less severe (1 or 2) and more severe (3 or 4). Then we input the data into a given less or more severe model to put emphasis on what differentiates a 1 from a 2 and a 4 from a 3. For the initial severity, the machine learning uses the XG boosted classifier, while both of the later models use the random forest classifier. The models were chosen based on the accuracy of their out-of-sample predictions or F1 score.



Given these factors this a chart of the probability of the severity of an accident

Conclusions

In 2020 as lockdowns went into effect the number of accidents decreased. While the number of accidents went down the severity did not. The severity of accidents actually increased in 2020. This is most likely to fewer vehicles being on the road resulting in speeds increasing due to less traffic.

Overall large counties had more accidents than smaller counties. Larger counties are more likely to have less severe accidents on average than smaller counties. Smaller counties produce more severe accidents, however they also have less accidents overall which skews the average. Harris county in Houston had the most number of accidents in the dataset.

It is hard to predict an accident with a severity of 4 and even more difficult to get a 1.

Limitations/Bias

We encountered file size limitations that made us focus our data set on Texas specifically. In the future, it would be most beneficial if we could complete an analysis for the entire United States. This would provide us with a more accurate macro picture of accident severity factors.

The main limitation is there is no data column that explicitly states how many cars were involved in the car accident. The dataset would benefit greatly from predictive modeling if that data were available.

Our dataset was lacking in uniformity in some areas as inputs were done at the scene of an accident and notes were not always input the same way every time.

The dataset dates back to 2016, most likely traffic conditions and patterns have changed greatly with Covid related adjustments to how people commute.

Weather itself can be difficult to bucket as conditions are fluid.

Future Work Recommendations

If we had more time we would include more data points to analyze accidents. This would provide us with a more accurate picture of what conditions are most likely to result in more severe accidents.

The data set used includes the timestamp of when an accident occurred. Would be beneficial to cross-reference the weather with the timestamp in the data set using the API from a weather source rather than relying on user input of conditions.

Research providing more detail about the severity of accidents. Our research only went so far as to bucket severity into 4 categories. Going forward would be beneficial if it could provide greater granularity of severity.

Works Cited

1. "How Severity the Accidents Is?" Kaggle, https://www.kaggle.com/deepakdeepu8978/how-severity-the-accidents-is

About Us

Angie Tran
Angie graduated B.S in Hospitality Management and has been working 6 years+ in customer service industries. She found interest in telling hidden stories within datasets. In her free time, she enjoys playing and walking with her corgi puppy.

Christopher Haub
Christopher has 10 years of experience in managing mortgage loan pools for government-sponsored entities and private equity. Christopher graduated from the University of Oklahoma prior to moving to Dallas. He enjoys learning new languages, most recently learning Mandarin. Christopher also enjoys fitness, including meditation, yoga, running, and HIIT workouts.

Juveriya Baig
Juveriya graduated college in 2020 from The University of Texas at Dallas with a Bachelor's in Marketing. She will be starting her Master's in Business Analytics in the fall of 2021 from SMU. In her free time, you can find Juveriya embroidering, binging shows, and hanging out with her family and friends.

Ryan Permenter

Ryan graduated from Texas Christian University in 2018 with a B.S in Criminal Justice. Ryan has experience in sales and marketing at Light Lab. In his spare time, he enjoys watching Dallas Stars games and lives within walking distance of the American Airlines Center. Ryan also enjoys playing video games.