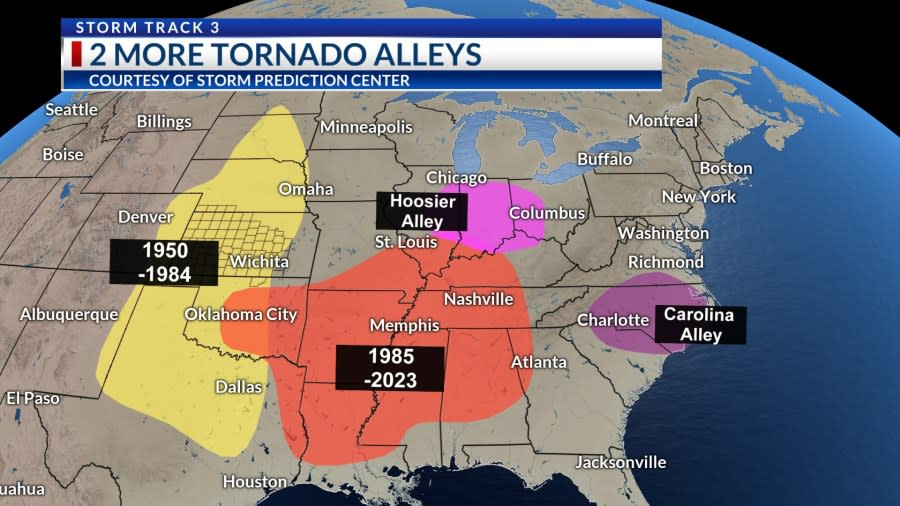
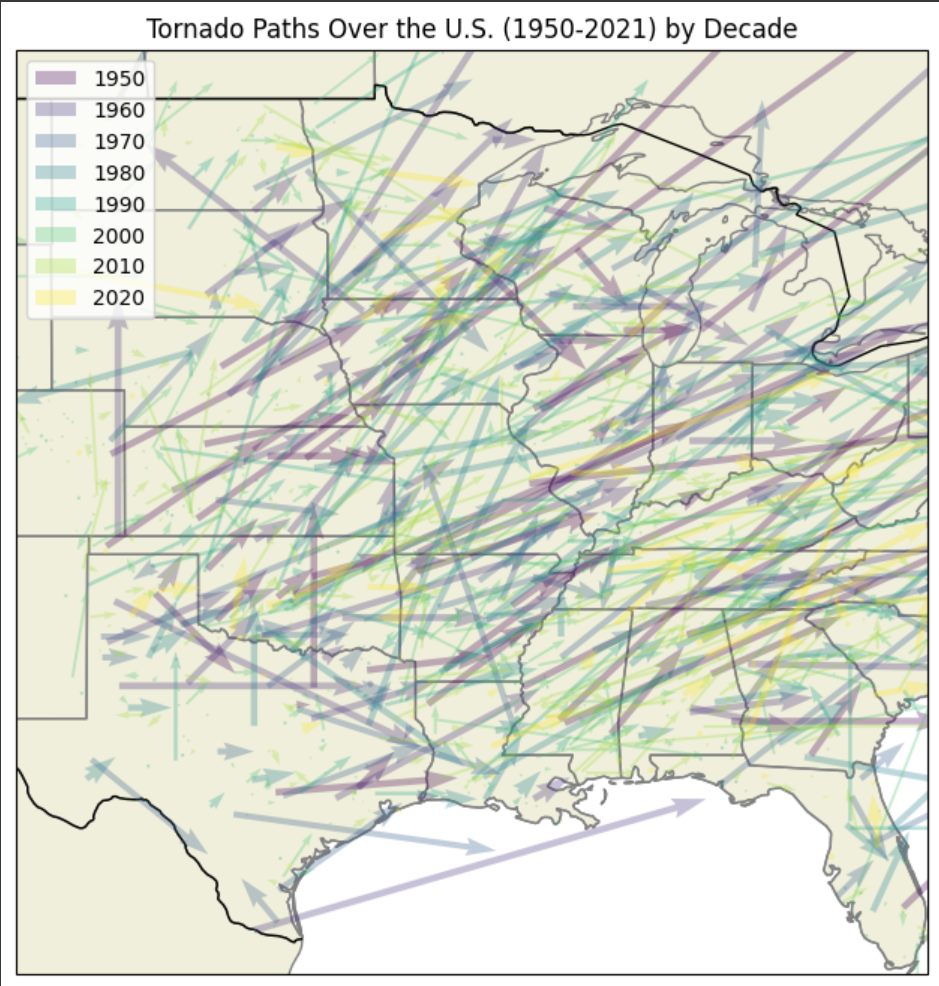
**Data Overview**

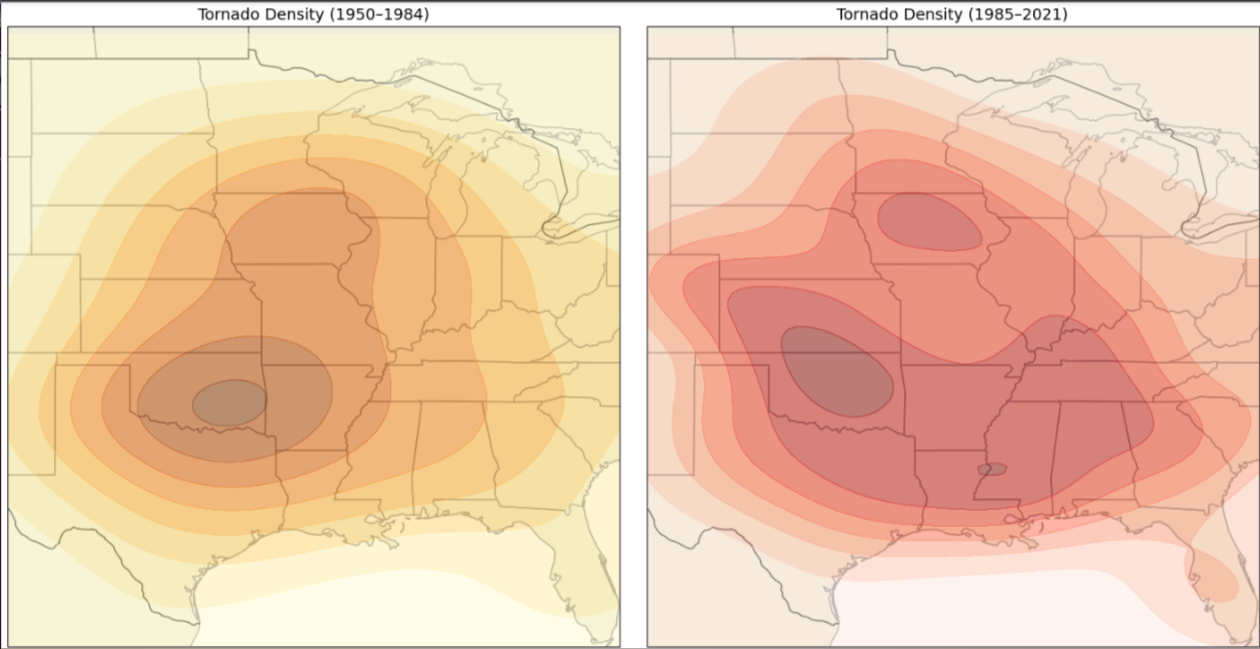
Our dataset contained a little over 68,000 rows ranging from 1950-2021, and contained temporal data, various geographic information (Starting and ending latitude and longitude of a tornado, state FIPS number, and total states affected by a tornado) In addition we also had tornado characteristics like EF scale, which measures the intensity of a tornado (which will be covered further in depth later on) length, width, injuries, fatalities, estimated crop damage, and finally estimated property damage.

**A Bit of Background**



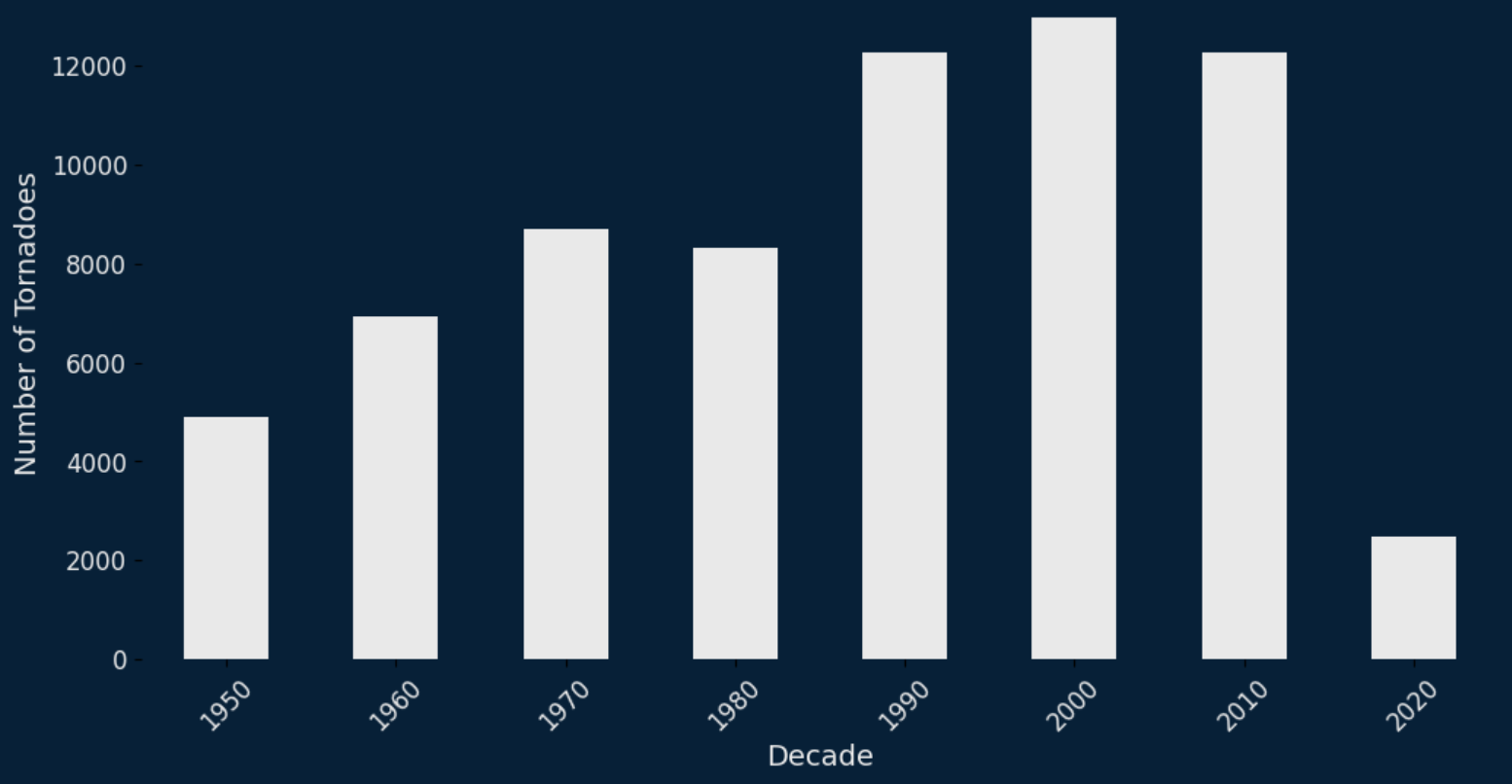
As you can see in the photo above the infamous Tornado Ally (yellow area) was where the majority of tornadoes occurred from 1950-1984, hover starting in 1985, we see a shift to more of the midwest to an area called Dixie Ally (Red Orange area) two smaller areas that are also known for tornadoes are Hosier ally (pink) and Carolina Alley (Purple). We wanted to see if our data lined up with the commonly known tornado trends, so we went ahead an tracked the start lat and long, and the end lat and long of tornadoes in our dataset, and drew a line between them. What we found was that generally our data also follows this trend. Something important to note is that due to visual clutter, only every 30th tornado is used for the visual, however when we change this interval we still see the trend hold true.





(Tornado occurrence before and after 1985)

When looking at tornadoes across decades it can seem like tornadoes have more than doubled since the 1950, however what is more likely is that as tornado tracking technology has gotten better, more tornadoes have been tracked, so it’s not the case that there are more tornadoes occuring, but rather that we just have the ability to track tornadoes more efficiently.



**How are tornadoes measured?**

Tornadoes are measured using the Enhanced Fujita Scale (EF Scale) which is based on estimated wind speeds and related damages of a tornado. When assessing damage people look at buildings, trees, and landscapes along the path of the tornado to determine the severity, so it’s not an exact science, but it gives a pretty good measurement of the potential windspeed of a tornado.

**Our Goals**

For this project we wanted to create a prediction model that could answer the following questions:

What characteristics can be used to predict tornado severity?

* What are the features of tornadoes that cause damage?
* What features make a tornado violent?

How do tornado characteristics differ between EF-ratings?

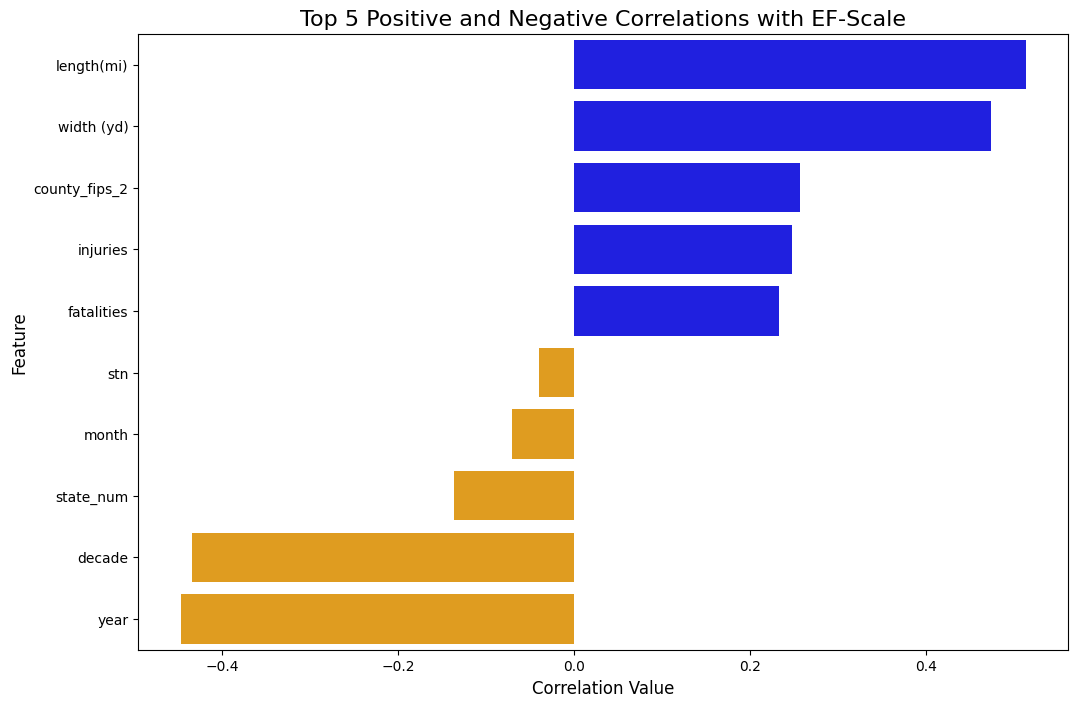
* What features are different between the majority of all damaging tornadoes and violent tornadoes?

**Onto The Model**

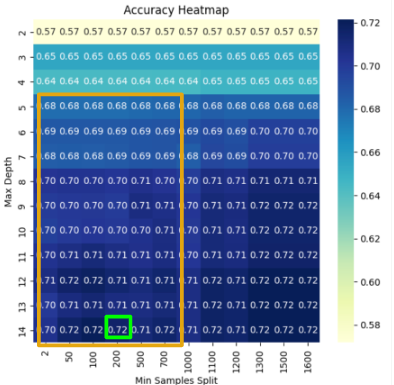
Before we get onto our prediction model, we need to explain the data we are using. We decided to exclude EF-0 and EF-5 tornadoes as the EF-0’s are difficult to track since they often don’t even touch down, and EF-5 tornadoes are extremely rare in modern times because with improved technology what would have been an EF-5 is now categorized as an EF-4.

We first wanted to see which columns mattered the most so we made a correlation bar chart and found the top ten correlations. We see that length and width are some of the biggest predictors, while year and decade are our highest mitigating factors.

Since we are trying to figure out what differences exist between our analysis we split up the tornadoes into two categories,non violent (EF1 and EF2) and violent (EF3 and EF4)



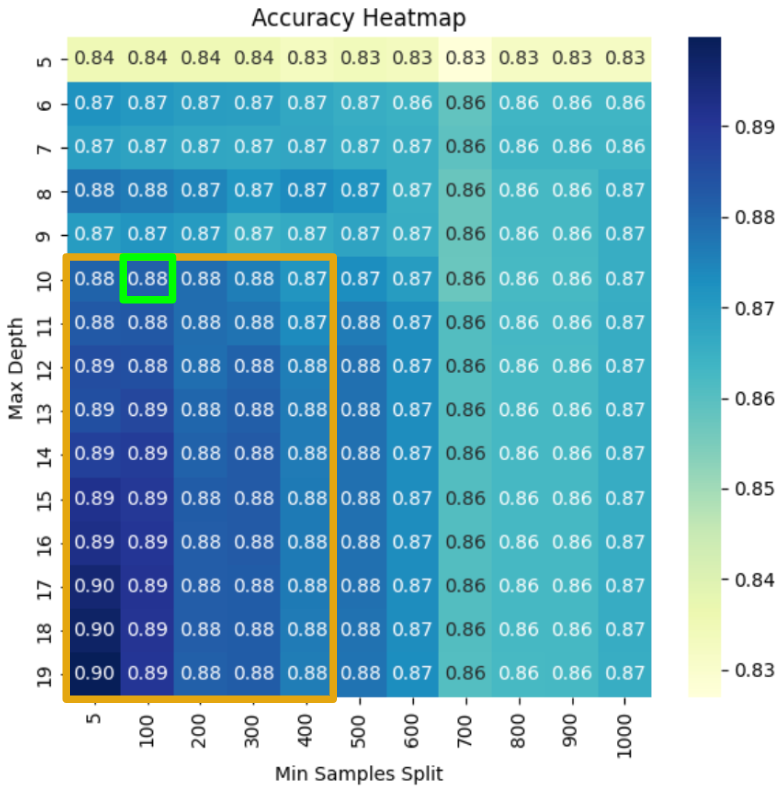
Next we made a heatmap to determine what depth and minimum sample split size, yielded the highest accuracy score for a decision tree of the non-violent data. We found that the best depth was 14 and the best min. split was 200.

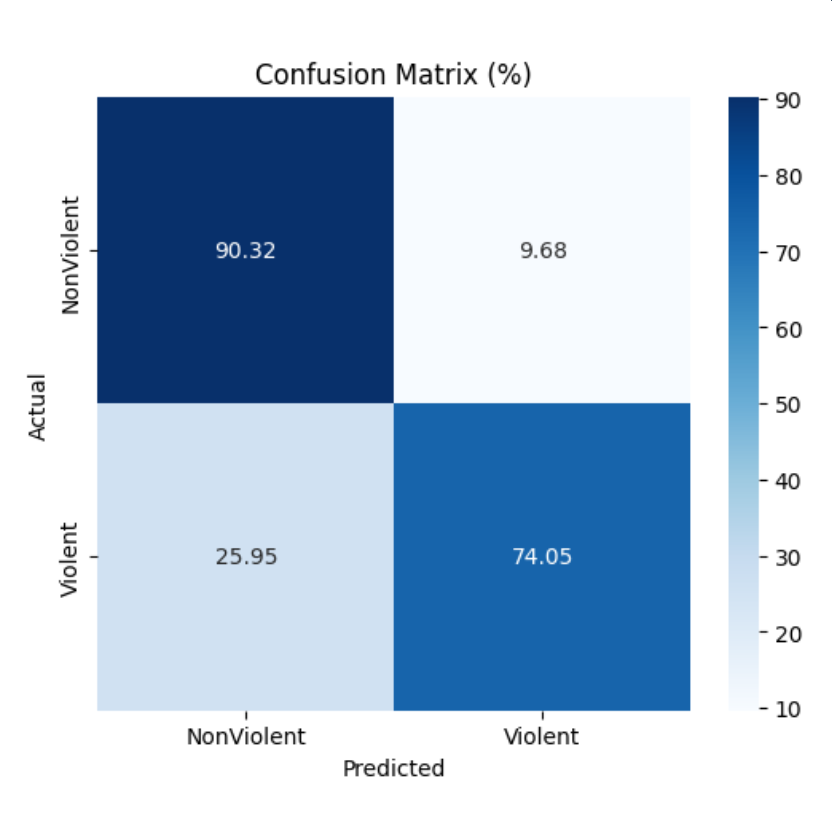


After using gradient boosting we were able to get the accuracy score up to 0.73 and the AUROC to 0.8 with Estimate Property Loss, Width (yd), Length (mi), and year being the top 4 coefficients. Afterwards we used a permutation importance graph to determine how factors affect the model. Permutation gets features, randomly shuffles them, and measures them again, so if they are actually important it will affect the models importance, and as we can see below the graphs before and after permutation look similar which shows us that the features are pretty stable.

**Predicting Violent Tornadoes**

Following the previous steps we created a heatmap to find the most accurate model, and that was at a max depth of 10 and a min sample split of 100. Our findings are far more accurate however, with an accuracy score of around 88%.





In the graph above we can see the breakdown, and see that our model does a worse job of predicting violent tornadoes compared to nonviolent one’s. When we take a look at what columns had the biggest impact on the prediction after gradient boosting (Which increased the accuracy score up to 89%), we see that Estimate Property Loss, injuries, Width (yd), Length (mi) are the top 4 features. However, after permutations, year, estimated property loss, and State FIPS were some of the top features, which suggests that there are more moving parts within the prediction of violent tornadoes, and are more challenging than prediction non-violent tornadoes.

**Wrapping up**

Through correlation analysis, decision trees, and gradient boosting, we identified key predictors such as estimated property loss, tornado width, length, and year. While our model performed reasonably well in predicting non-violent tornadoes, it demonstrated significantly higher accuracy in predicting violent tornadoes, achieving an accuracy score of 89% after gradient boosting.