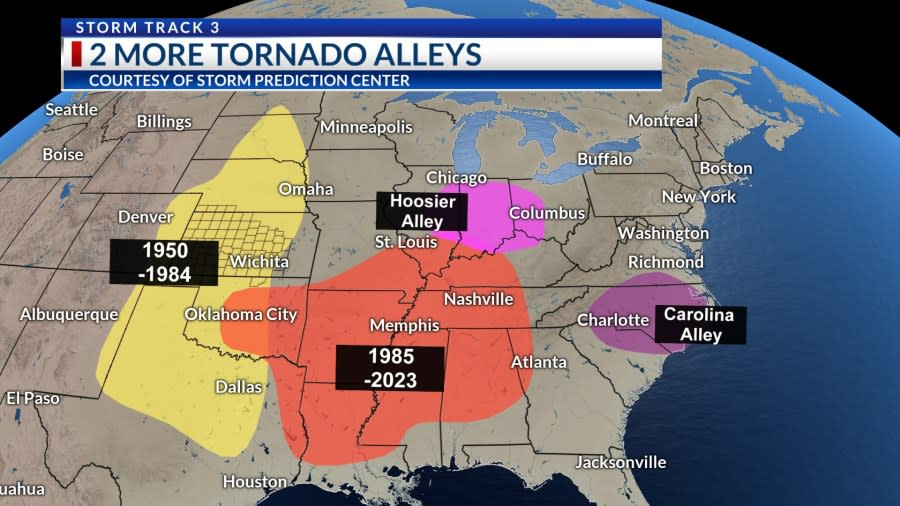
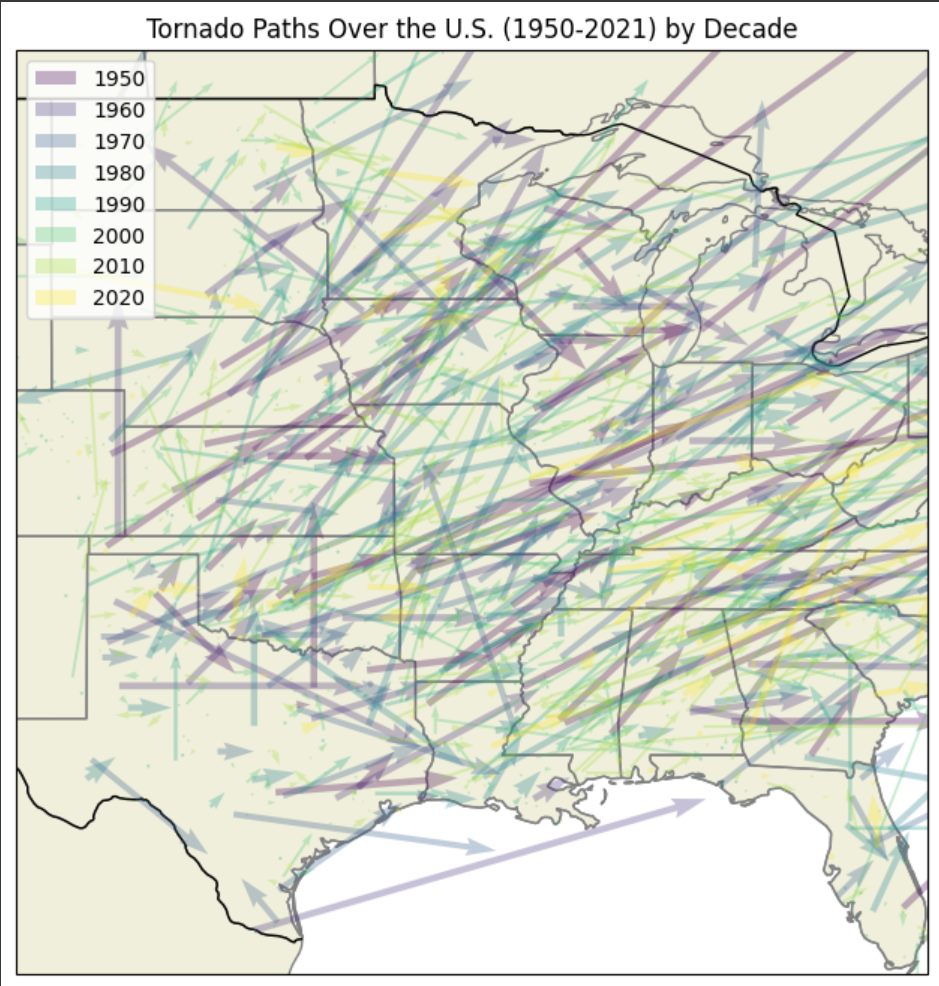
**Data Overview**

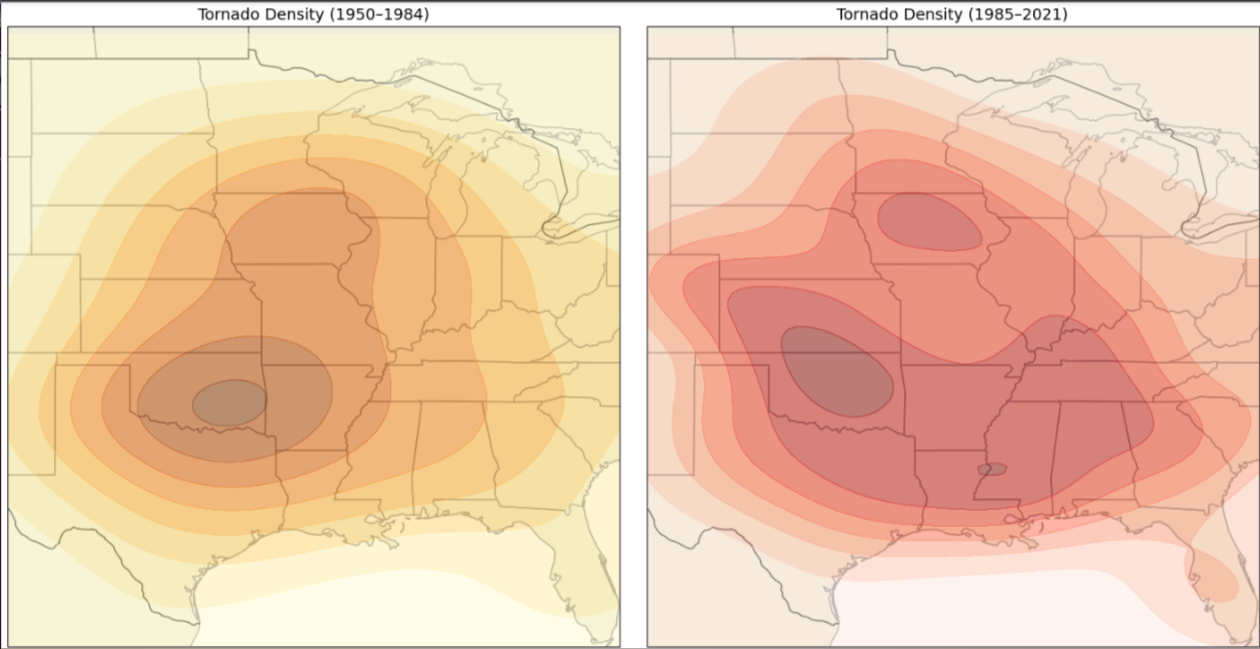
Our dataset contained a little over 68,000 rows ranging from 1950-2021, and contained temporal data, various geographic information (Starting and ending latitude and longitude of a tornado, state FIPS number, and total states affected by a tornado) In addition we also had tornado characteristics like 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**



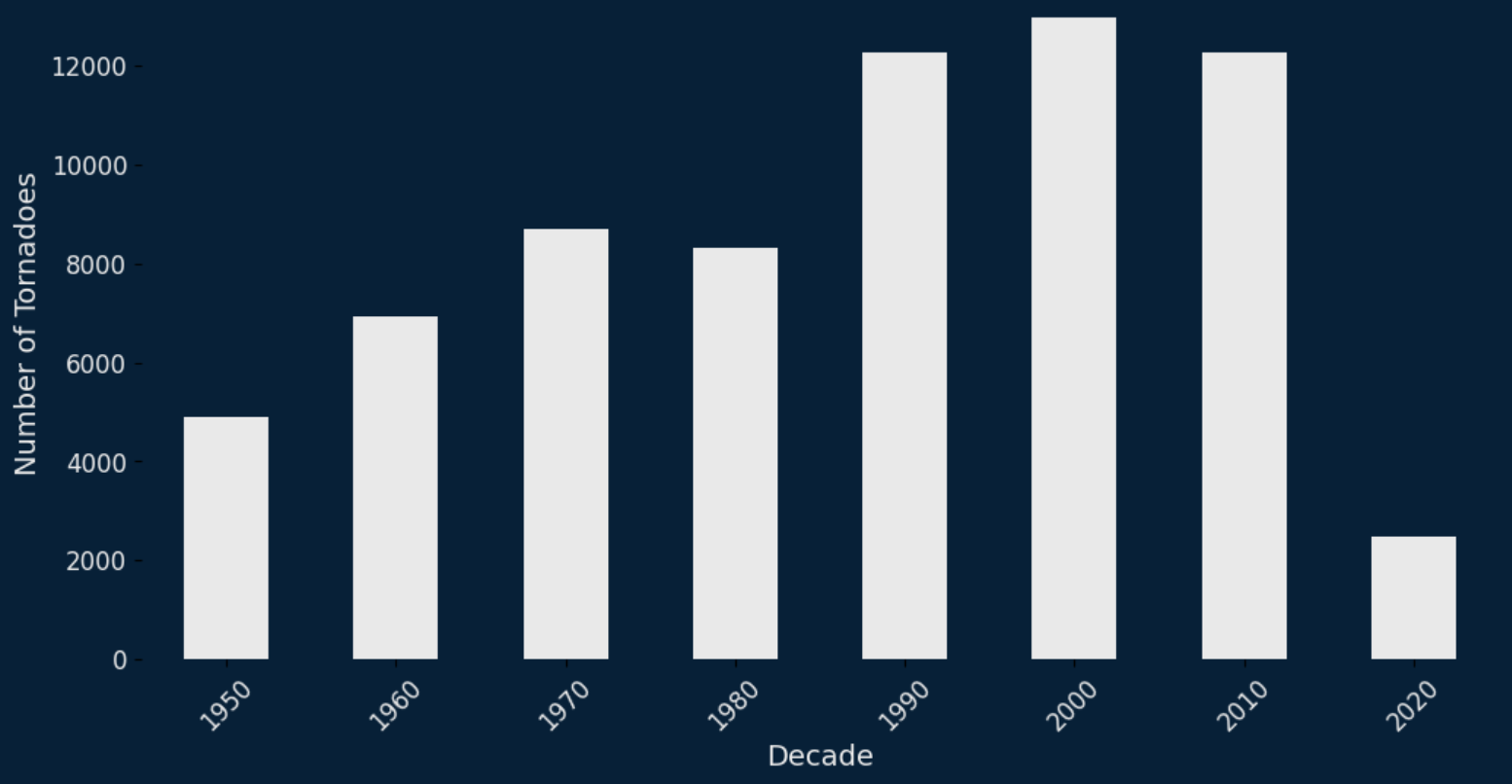
As you can see in the photo above the infamous Tornado Alley (yellow area) was where the majority of tornadoes occurred from 1950-1984, hover starting in 1985, we see a shift to more of the midwest to an area called Dixie Ally (Red Orange area) two smaller areas that are also known for tornadoes are 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 and 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?

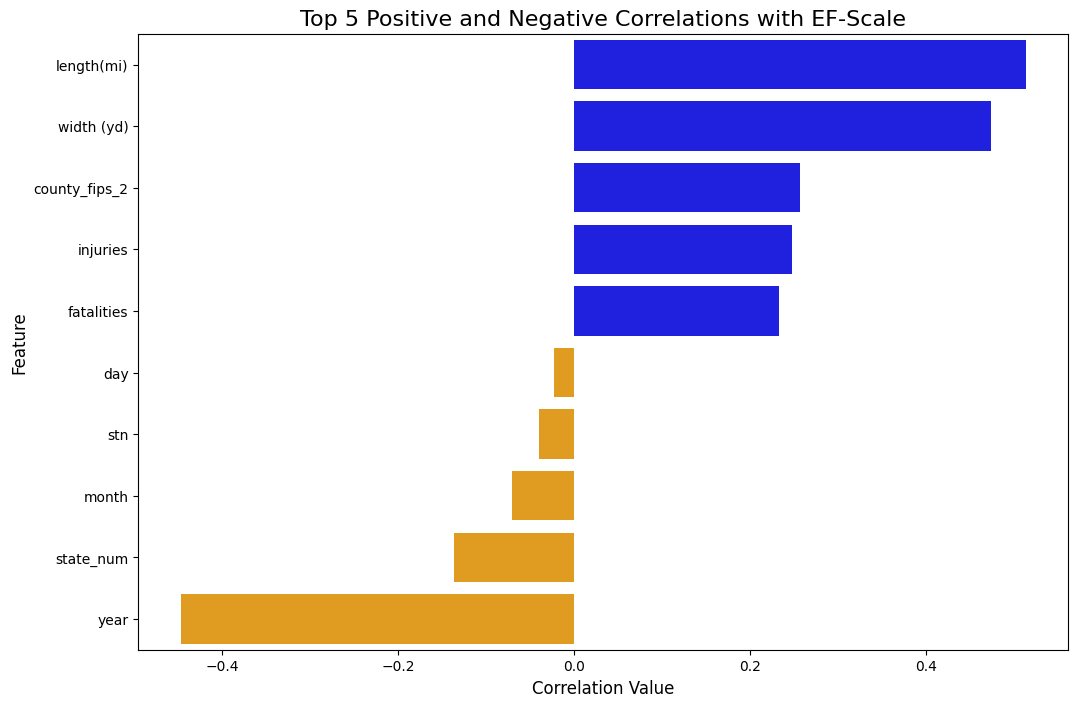
How do tornado characteristics differ between EF-ratings?

* What features are different between non violent 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 most times, and EF-5 ratings are rare in modern times because with improved technology what would have been an EF-5 is now categorized as an EF-4.

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)

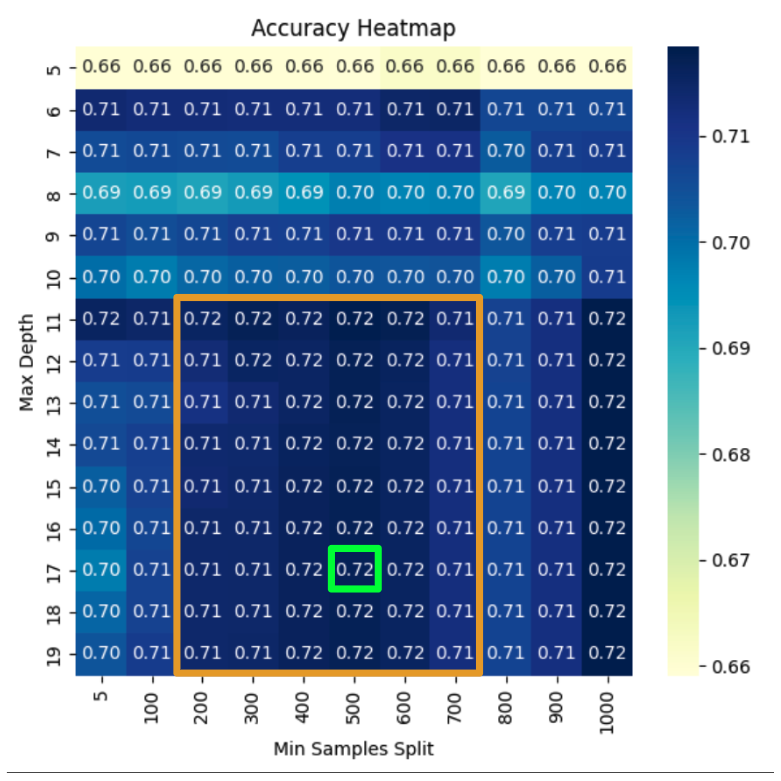


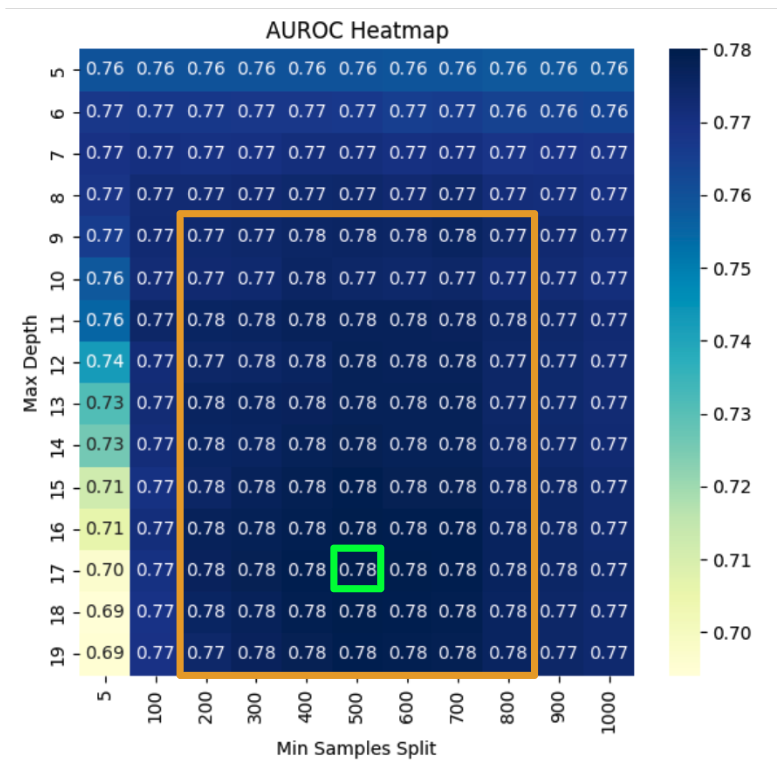
We first wanted to see which columns mattered the most so we made a correlation bar chart and found the top ten correlations. We see that length and width are some of the biggest predictors, while year length, width, and the state it occurred in are our highest mitigating factors.

**How was the data prepared for modeling?**

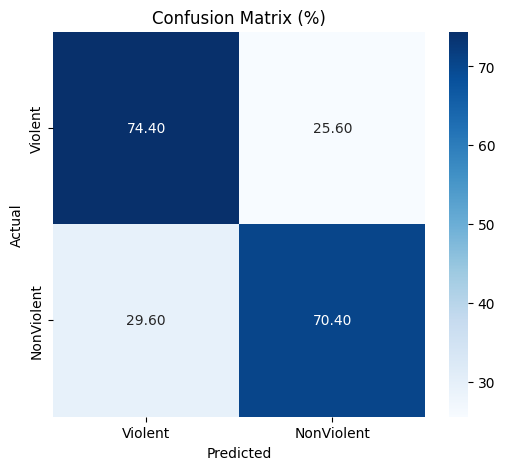
For the models, the data was balanced for each of the sections we are focusing on using SMOTE. Which works by generating synthetic samples for the minority class rather than simply duplicating existing ones.

For the first model, we wanted to start by focusing on non violent tornadoes. Firstly we made heatmaps to determine what depth and minimum sample split size yielded a reliable model for a decision tree of the non-violent data. We found that the best depth was 17 and the best min samples split was 500 when looking at accuracy (0.72) and auroc (0.78).

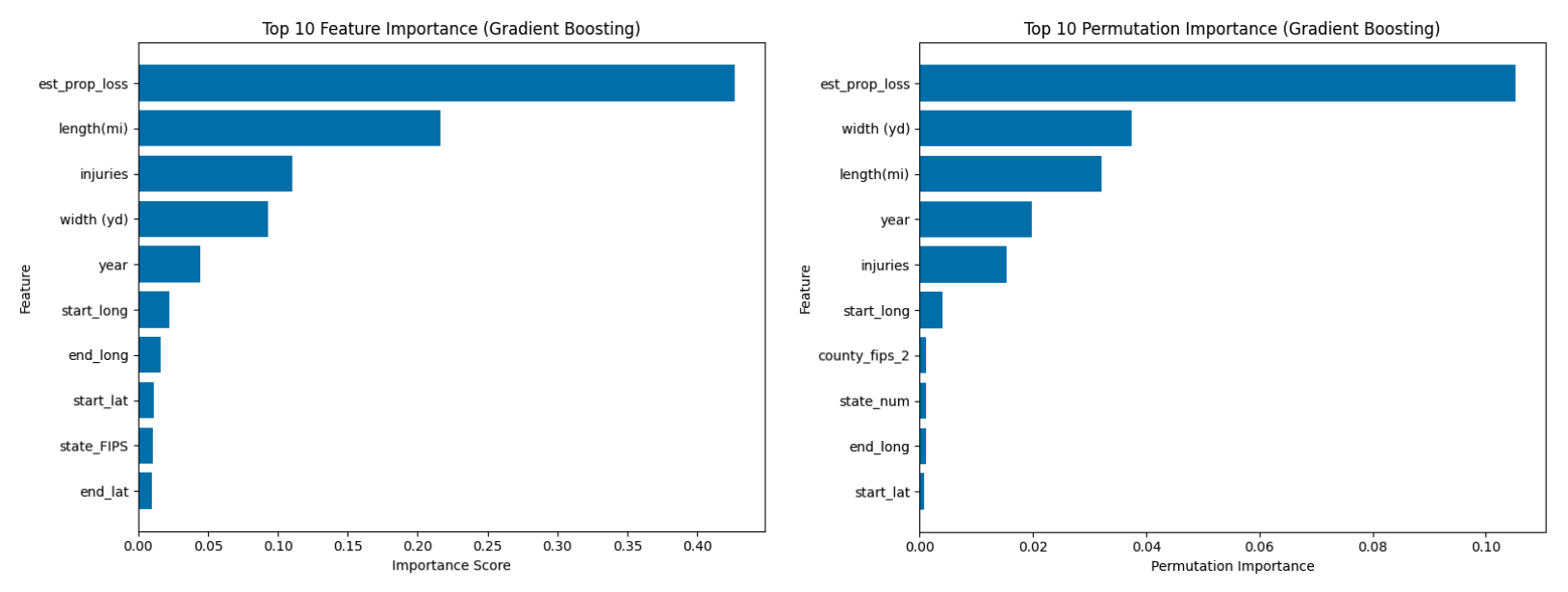




We were then able to plug these hyper parameters into a gradient boosting model to boost the performance a little more which led our accuracy to increase to 0.73 and our AUROC to 0.80.



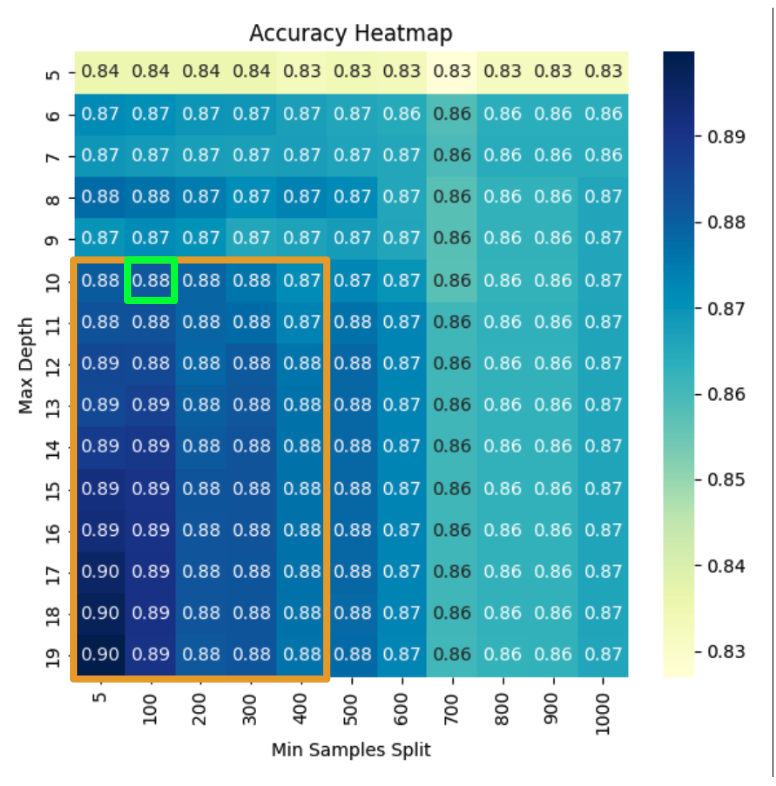
Using the gradient boosting model, we then looked at the top features and permutations. Which were Estimated Property Loss, Width (yd), Length (mi), Injuries, and Year being the top 5 features and permutations.

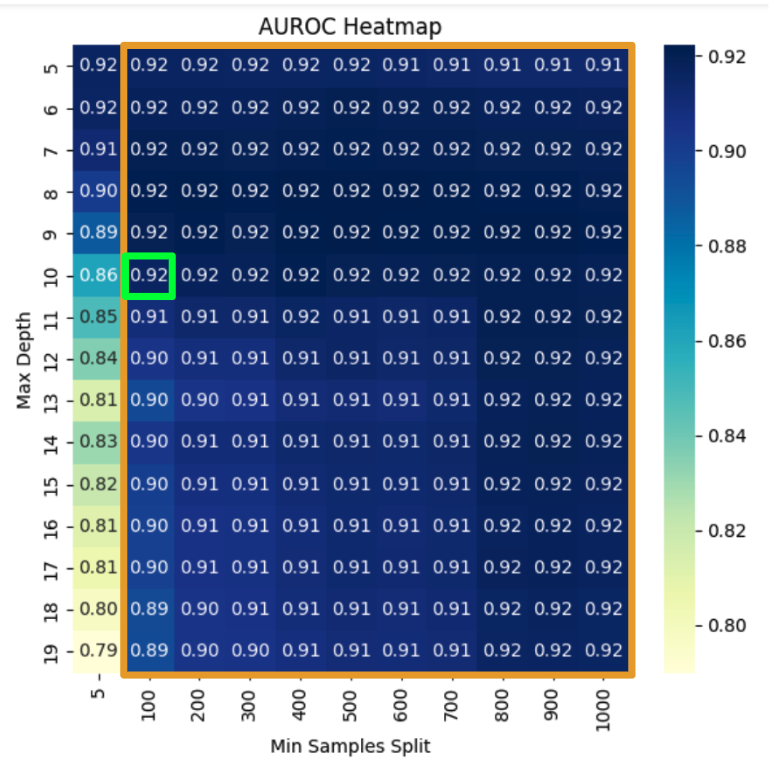


We used permutation importance to double check whether the features are actually important. This works by randomly shuffling the features, and measuring them again, so if they are actually important it will affect the models performance, and as we can see below the graphs before and after permutation look similar which shows us that the features are pretty stable. Now let's look at violent tornadoes.

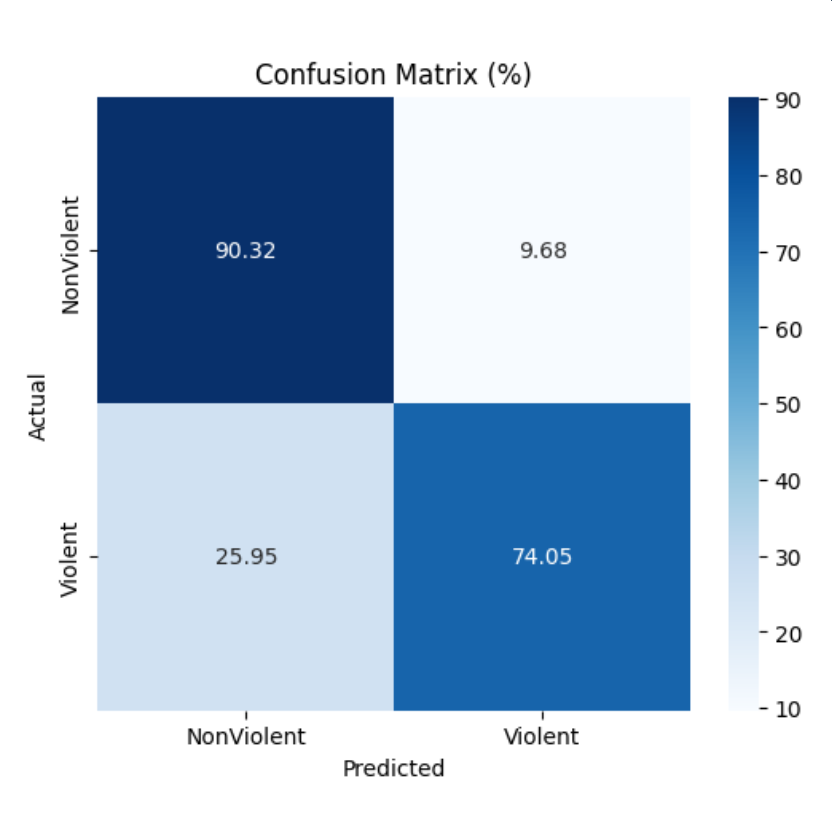
**Predicting Violent Tornadoes**

Following the previous steps we created a heatmap to find a reliable model for violent tornadoes. We found that the best hyper parameters to be a max depth of 10 and a min sample split of 100. This led to an accuracy of 0.88 and an AUROC of 0.92.

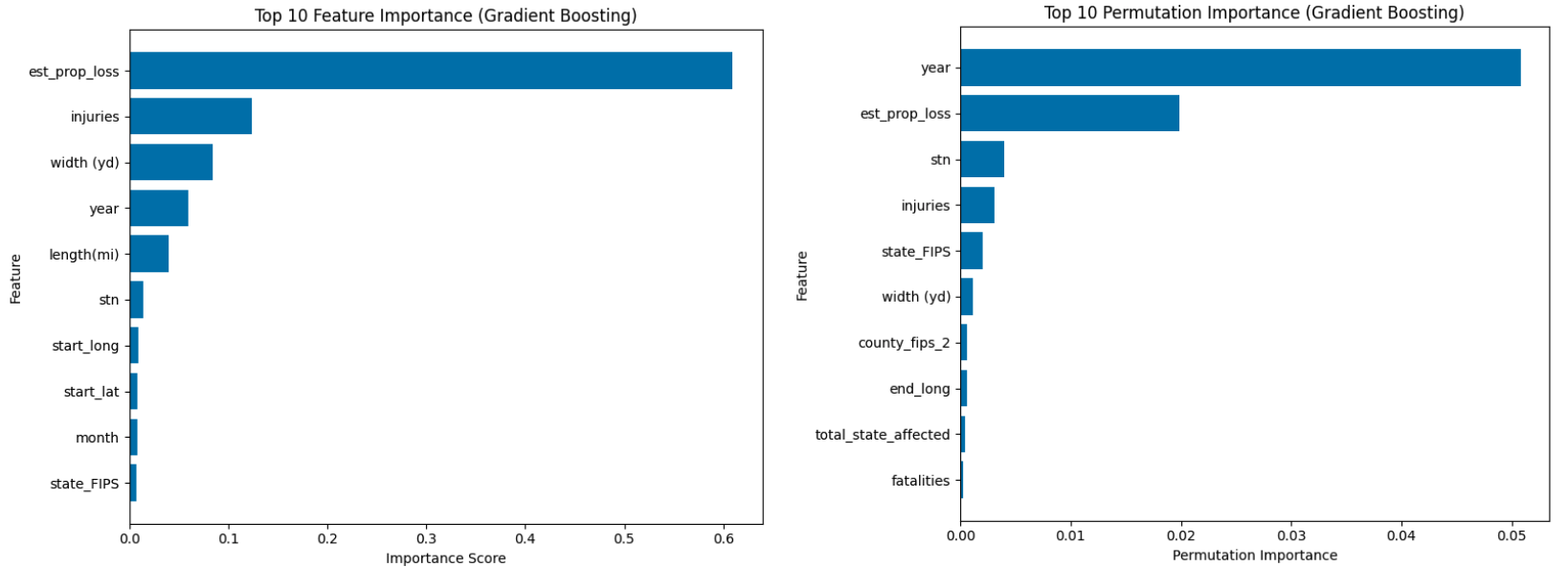




As before, we used the best performing hyper parameters and plugged them into a gradient boosting model to boost the performance which led to an accuracy of 0.89 and AUROC of 0.93.



We noticed that this model is doing a really good job at predicting non violent tornadoes, and retains similar performance for violent tornadoes as with predicting non violent tornadoes, with ~26% of violent tornados being classified as non violent. As before, we can expand on this by looking at the features of this model.



This is where things get interesting, as mentioned before for figure 2.4, we use permutation importance to double check importance. What stands out is that the top features of each graph are different. Not only that, a majority of the features in both graphs are different.

What does this mean? Well in short, we can see the model uses estimated property loss as its most important feature to rank violent tornadoes. As seen again in figure 2.4, estimated property loss is also the top ranked feature while retaining permutation importance. This is likely why this model over performed when predicting non violent tornadoes (figure 3.3).

But in this model, the top permutation is the year. Suggesting a few things, firstly, its likely that violent tornadoes are more complicated to predict and require more climate data to get more stable features. If that is the case, then why is the year showing up at all? Well, the year is heavily tied to long range climate trends, which heavily affects the occurrence of violent tornadoes. This is most evident with the outbreak in 2011, something that is not usually seen every year, and it is something we might be seeing in 2025 so far with larger scale tornadic outbreaks occurring more frequently than usual.

With that out of the way, from our data, we can still point out some of the features its targeting to predict violent tornadoes. We can somewhat pinpoint the top 5 features of this model to be Year, Estimated Property Loss, Injuries, The State is Occurred in, and the Tornados Width.

So, what can we take away from this?

**Conclusion**

Non Violent Tornadoes:

Gradient Boosting Classifier (figure 2.3): Accuracy (0.73) / AUROC (0.80)

The modeling for the non violent tornadoes (EF1-EF2) was pretty straight forward with the top features being Estimated Property Loss, the tornadoes Width (yd), Length (mi), Injuries, and Year. These are features that are expected when predicting tornadoes in general. But, since we know that year is really important for violent tornadoes, what does that tell us about non violent tornadoes? Well, essentially, non violent tornadoes are more common, suggesting rare weather patterns are not as important, its not surprising year is not as important for predicting violent tornadoes. It's important to note that year does not just stand for long term climate patterns or shifts, there are other features that are likely mixed in with that we don't have access to in this data set.

Violent Tornadoes:

Gradient Boosting Classifier (figure 3.3): Accuracy (0.89) / AUROC (0.93)

When modeling violent tornadoes (EF3-EF4), things get a little more complicated which is seen in the performance of the model and the heat maps (figure 3.1 & figure 3.2). We concluded that the top performing features of this model are, Year, Estimated Property Loss, Injuries, The Sate is Occurred in, and the Tornados Width (yd). With the year being the stand out feature of this model, that tells us more, likely climate data is needed for more stable features, especially long range climate data.

The Conclusion:

After looking through the models for each, we arrive at a conclusion for this work. We can see that most of the features of both models, permutation and feature importance both align around the coefficients seen in figure 1. Which as mentioned before is how tornadoes are measured in the first place, not surprising. But something we can confidently say is that more climate data is needed. We can say that non violent tornadoes are not dependent on long range climate trends, white violent tornadoes which are more rare tend to be more reliant on that factor. This work can and will most likely be expanded on in the future with more climate data to narrow down what leads to a violent tornado.