# Tornado and its features prediction using ensembles

Aditya Mahale
*B00867619*

Guturu Rama Mohan Vishnu
B00871849

Smriti Mishra
B00904799

Benny Tharigopala
B00899629

Shiva Shankar Pandillapalli
B00880049

***Abstract*— Tornadoes are one of nature's devastating disasters that resembles a vertical funnel of rapidly spinning air at high speeds. Their winds may top 318 miles an hour with gust wind. The crucial point to notice here is the human lives and the resources that are being damaged. Hence, predicting and taking necessary actions ahead of time will relieve us from the huge loss. Predicting a natural phenomenon is not an easy task unless supported with sufficient historical data. In our case, we collected 1950-2015 data and tried to predict most of the features of the tornado including the starting and ending locations. We are not too far from accurate predictions, and we could achieve them with a little more research and data. The following paper describes our work on this topic where we built multiple machine learning models to predict each feature using the ensembles technique in machine learning.**

## I. Introduction

Tornadoes are swirling winds that spin at incredible speeds of up to 400 mph, making them one of the most devastating natural phenomena. They, like hurricanes, are created by extreme low-pressure zones that draw in strong winds. Because they have smaller bases, their destructive power is dramatically amplified. Tornadoes are unpredictable and frequently strike without warning.

Tornadoes can range from a few hundred yards to more than a mile in width. These tornadoes can be found throughout the United States, but the Midwest and South regions are particularly more common and vulnerable. When a tornado is approaching, you just have a few moments to make life-or-death decisions. To survive a tornado, you'll need to plan beforehand and react quickly. To deliver such advanced warnings, we aim to use historical data to estimate the possibility of a tornado in a specific region over time. Meteorologists are trying hard to forecast tornadoes with the intent to predict them soon. Hence, we chose to work on this.

## II. Data understanding and preprocessing

The data set we have chosen for our project is the Tornadoes data set from the year 1950 to 2015. The data set contains many attributes about the Tornadoes that occurred between these years in the United States of America, like location, date, magnitude of the tornado, injuries and losses etc. The data we have right now is around more than 60,000 entries, which we believe is big enough in order to do data analysis and to perform the prediction. The problem we have right now is a supervised learning problem and we have multiple target variables. We did not stick with just one target variable for the project, but we have chosen to put multiple target variables, one target variable for each individual; so that we might be in a good position to predict most of the tornado’s features. And we have chosen to proceed with the process of ensemble, so that we could get the best of all the models and accuracy. Our target variables are,

- Magnitude
- Length
- Crop Loss
- Property Loss
- Location (latitude and longitude)

Our dataset contains both continuous and categorical features. To understand those features, we have built the data quality report for both features separately. To observe for any specific patterns in these individual features, we have made some visualizations for all the features.

All the continuous features have been visualized through histograms as shown below.

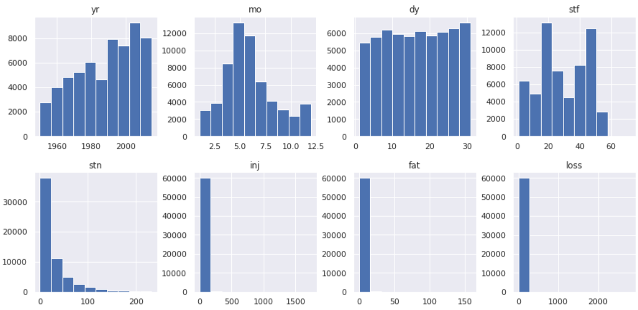


*Figure 1 Continuous Features Visualization (1)*

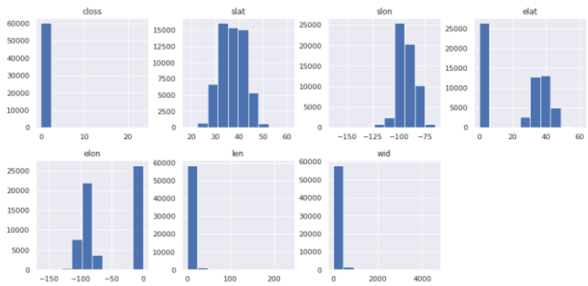


*Figure 2 Continuous Features Visualization (2)*

All the categorical features have been visualized through bar plots as shown below.

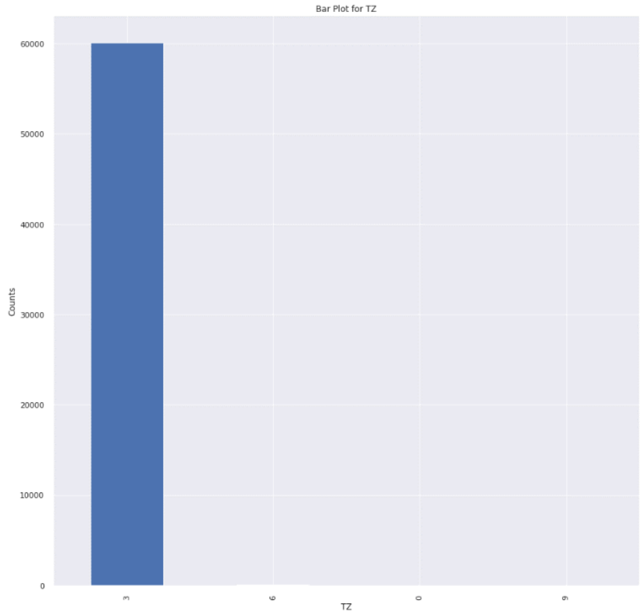


*Figure 3 Bar plot for Time Zone*

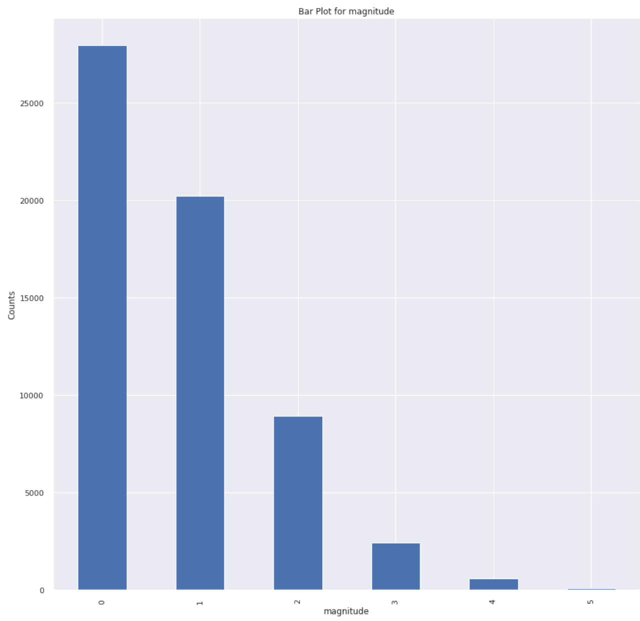


*Figure 4 Bar plot for Magnitude*

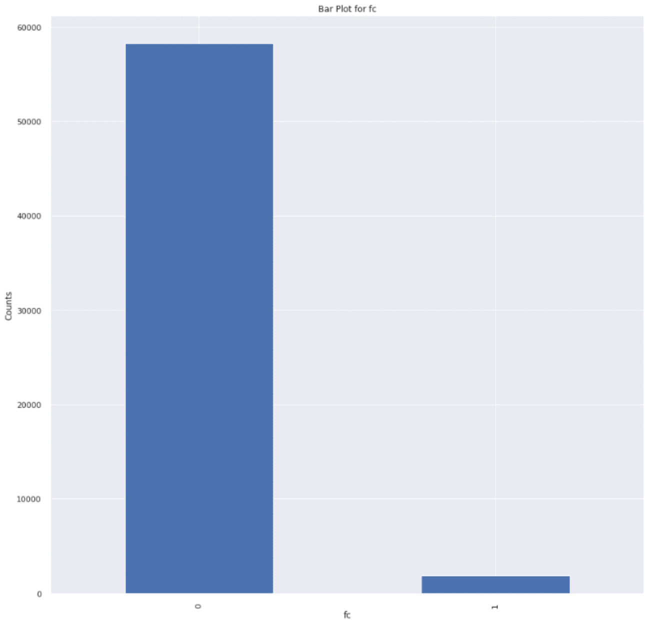


*Figure 5 Bar plot for F-Scale Reading*

Generally, data issues in the data sets are of three type: missing values, irregular cardinality and outliers. We have checked our data for any of these issues and if there are any, we have built a data quality plan in order to resolve that issue.

### A. *Missing Values*

As we can observe from the data quality report of the continuous features below, there are no missing values present in these features.

| | Count | Miss % | Card. | Min | 1st Qrt. | Mean | Median | 3rd Qrt | Max | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|
| om | 60114 | 0.00000 | 7422 | 1.00000 | 248.00000 | 41119.37575 | 509.00000 | 845.00000 | 613494.00000 | 130865.99363 |
| yr | 60114 | 0.00000 | 66 | 1950.00000 | 1974.00000 | 1987.97006 | 1991.00000 | 2003.00000 | 2015.00000 | 17.78788 |
| mo | 60114 | 0.00000 | 12 | 1.00000 | 4.00000 | 5.97244 | 6.00000 | 7.00000 | 12.00000 | 2.38881 |
| dy | 60114 | 0.00000 | 31 | 1.00000 | 8.00000 | 15.87637 | 16.00000 | 24.00000 | 31.00000 | 8.73719 |
| tz | 60114 | 0.00000 | 4 | 0.00000 | 3.00000 | 3.00110 | 3.00000 | 3.00000 | 9.00000 | 0.07738 |
| stf | 60114 | 0.00000 | 52 | 1.00000 | 18.00000 | 29.41992 | 29.00000 | 45.00000 | 72.00000 | 15.01885 |
| stn | 60114 | 0.00000 | 233 | 0.00000 | 4.00000 | 26.47623 | 15.00000 | 35.00000 | 232.00000 | 32.67265 |
| mag | 60114 | 0.00000 | 6 | 0.00000 | 0.00000 | 0.79615 | 1.00000 | 1.00000 | 5.00000 | 0.91234 |
| inj | 60114 | 0.00000 | 206 | 0.00000 | 0.00000 | 1.56130 | 0.00000 | 0.00000 | 1740.00000 | 19.17770 |
| fat | 60114 | 0.00000 | 48 | 0.00000 | 0.00000 | 0.09687 | 0.00000 | 0.00000 | 158.00000 | 1.54621 |
| loss | 60114 | 0.00000 | 472 | 0.00000 | 0.00000 | 2.15931 | 0.10000 | 4.00000 | 2800.10000 | 20.28468 |
| closs | 60114 | 0.00000 | 47 | 0.00000 | 0.00000 | 0.00213 | 0.00000 | 0.00000 | 23.52000 | 0.12218 |
| slat | 60114 | 0.00000 | 2319 | 18.13000 | 33.24000 | 37.15521 | 37.08500 | 40.97000 | 61.02000 | 5.12345 |
| slon | 60114 | 0.00000 | 4234 | -163.53000 | -98.60000 | -92.96113 | -93.95000 | -86.87000 | -64.90000 | 8.71912 |
| elat | 60114 | 0.00000 | 2291 | 0.00000 | 0.00000 | 20.95625 | 31.20000 | 38.15000 | 61.02000 | 18.89035 |
| elon | 60114 | 0.00000 | 3850 | -163.53000 | -94.25000 | -51.90289 | -81.76500 | 0.00000 | 0.00000 | 46.32270 |
| len | 60114 | 0.00000 | 2132 | 0.00000 | 0.10000 | 3.48072 | 0.60000 | 3.00000 | 234.70000 | 8.51933 |
| wid | 60114 | 0.00000 | 323 | 0.00000 | 13.00000 | 98.45460 | 40.00000 | 100.00000 | 4576.00000 | 195.72628 |
| fc | 60114 | 0.00000 | 2 | 0.00000 | 0.00000 | 0.03099 | 0.00000 | 0.00000 | 1.00000 | 0.17330 |

*Figure 6 Continuous Features Report*

Coming to the categorical features, we can observe that there are no missing values in these features too.

| | Count | Miss % | Card. | Mode | Mode Freq | Mode % | 2nd Mode | 2nd Mode Freq | 2nd Mode % |
|---|---|---|---|---|---|---|---|---|---|
| date | 60114 | 0.00000 | 11225 | 4/27/2011 | 207 | 0.34435 | 4/3/1974 | 128 | 0.21293 |
| time | 60114 | 0.00000 | 1429 | 16:00:00 | 955 | 1.58865 | 17:00:00 | 942 | 1.56702 |
| st | 60114 | 0.00000 | 52 | TX | 8484 | 14.11318 | KS | 4027 | 6.69894 |

*Figure 7 Categorical Features Report*

### B. *Irregular Cardinality*

From the above data quality report of the continuous features, we observed that the features **tz**, **mag** and **fc** have cardinality less than 10. So, we have decided to convert them into categorical features. And from the data quality report of the categorical features, we can conclude that **date** is neither a categorical feature nor a continuous feature and as well as **time**.

### C. *Outliers*

There are a few outliers in our data for the continuous features, but we have decided not to remove them since our dataset is only of 60k records and every record would be useful for the modelling and for the prediction.

To summarize the pre-processing of the data, we have performed the following steps to clean our data and transform it as required:

- Converted the features "**tz**", "**mag**" and "**fc**" from continuous to categorical.
- Dropping "**om**" column as it just denotes the tornado number. It does not add any information to the data.

- Breaking down “**date**” column into separate “**month**”, “**year**” and “**day**” columns.
- Dropping “**date**” column since we have broken down it to three “**month**”, “**year**” and “**day**” columns.
- Dropping the “**timezone(tz)**” column does not add any new information to the dataset since location information is already represented by other columns.
- Dropping both “**stf”** and “**stn”** columns because they are redundant, i.e., “**st”** column already represents the same data.
- Replacing the zero values in “**elat**” and “**elon**” columns with the starting values.

## III. Literature Review

Tornadoes are one of the major natural threats to human life. There are many places on earth in countries like North America, Australia, Europe, Africa, Asia etc., where we can see these tornadoes frequently.

Most tornadoes form from the thunderstorms as explained in an article [7] written by NOAA. Not all thunderstorms are responsible for creating tornadoes, there are a special kind of thunderstorms in the thunderstorms themselves, which are termed as supercell thunderstorms. Most of the time, these are responsible for the formation of the tornadoes. Since their wind speed is unimaginable as it reaches anywhere from 40 to 318 miles as per NOAA Fujita scale covering a massive amount of land. The length and width of these tornadoes are unpredictable most of the time. However, we will be trying to predict these based on the historical data using ensembles technique. The following figure shows the enhanced Fujita Scale from weather.gov [8].

| EF SCALE | |
|---|---|
| **EF Rating** | **3 Second Gust (mph)** |
| **0** | 65-85 |
| **1** | 86-110 |
| **2** | 111-135 |
| **3** | 136-165 |
| **4** | 166-200 |
| **5** | Over 200 |

*Figure 8 Enhanced F Scale*

Each year about 1200+ tornadoes hit the U.S. The NOAA of government of U.S has huge team to predict and detect these tornadoes which takes toll on human lives and damages a lot of resources. NOAA [9] works on predicting these tornadoes using various techniques like numerical weather prediction models, mobile Doppler radar to predict and position the tornadoes which includes the entire lifecycle scan. A few members at MIT [10] has recently started working on tornado and its prediction. A part of their work is to predict the path to be followed by the tornado.

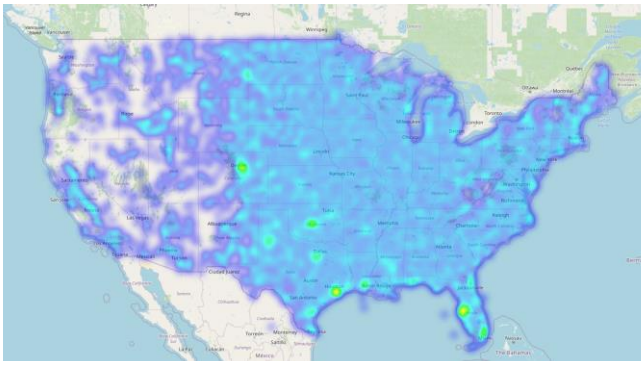

*Figure 9 Tornadoes starting locations from 1950-2015*

From the above map, we can see how frequent the tornadoes are in the United States of America in North America continent. Hence, we would like to focus on this part of the North American countries. Although we will be working on the entire country, we would like to focus more on the states which are affecting most. From the above figure 1, it is clear that the states towards the east and the middle of America ae affecting the most by these tornadoes.

Since a part of our objective is to predict the ending location of the tornado based on the starting geo location, predicting other features of the tornado based on the available information at any point of time. Members at Harvard [11] has done similar work to predict the start of the tornado, however they worked on the prediction based on the sounding and Reanalysis Data using linear regression. An article [12] written by Yves Jacquot on medium platform describes about the direction which the tornado might take, However the author has not done any work on predicting the ending location of a tornado based on the starting location.

Authors in the following [13] paper worked on predicting the tornadoes circulations using intelligent data analysis. However, they did not focus on predicting the final coordinates of the tornadoes and all the other features of the tornado. The source of the dataset [14] have a few solutions [15] predicting just the tornadoes magnitude based on multiple methods. However, we did not find solutions that were using ensembles particularly to predict almost all the features of the tornado. Especially the ending location of a tornado.

A few researchers have done their research in predicting the tornadoes an hour in Canada [16] before their start by using Multi-year Reanalysis of Remotely Sensed Storms (MYRORSS) and Gridded NEXRAD WSR-88D Radar dataset (GridRad) which was accurately predicting the tornadoes with an AUC above 0.9. Although we want to predict the tornadoes using the Neural Networks ensemble, we will not be using the Convolutional Neural Networks as we are not depending on the geographical climate image data to predict the features of the tornadoes.

Meteorologist of NOAA's Storm Prediction Center, Patrick Marsh in an interview with npr.org [17] has clearly stated that its is nearly impossible to predict the touch down of a tornado. They have mentioned that this is because of the lack of the data and the continuously changing climatic conditions and the weather in which the tornadoes are created.

From this [18] article, we can observe how the tornadoes are affecting the United States of America in terms of Economy from 1995-2020. Tornadoes not only have toll on human lives, but also devastate all kinds of properties like crops and other lands including the houses, cattle and other electronic and mechanical devices which are not moved out of the way in time. Because of the mix of both cold and hot airs in the tornadoes, they cause a lot of damage to the above mentioned increasing the cost of damage multiple folds as explained in the article.

In our work, we not only want to predict the ending location of a tornado based on the given location and a few other parameters, we also want to predict almost all the important features of the tornado based using ensembles methods using some weak learners. Since our work uses very less resources and weak models, we try ensembles to make our predictions better by combining a few weak models to get strong predictions.

## IV. DESCRIPTION OF THE SOLUTION

This section describes the development of machine learning models, to predict the magnitudes of tornadoes, tornado length and losses incurred for every occurrence of a Tornado. The methodologies implemented in this project have been adapted from the learning outcomes in CSCI 6409 – Process of Data Science (unless any other literature review has taken place). Rating the intensity of tornadoes is a very complex task, due mainly to the fact that they usually are not predictable, and when they are, it is only a few minutes before they occur. Measuring their wind speed, requires observers and doppler radars on site, is therefore not possible in most cases. Considering this, we have analyzed historic data relevant to Tornado occurrences obtained from United States' National Oceanic and Atmospheric Administration (NOAA) storm prediction center for the years between 1950 and 2015. Subsequently we developed Machine Learning models to predict the magnitude of a Tornado, length of a Tornado, crop loss and loss owing to an occurrence of a Tornado. The implementation of our solution is described in the following section.

**Implementation Approach**

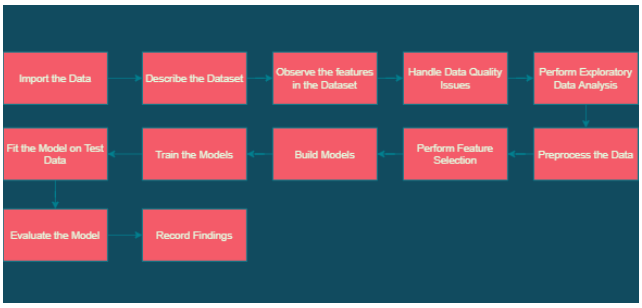


*Figure 10: Generic Block Diagram for Prediction of Tornado Attributes*

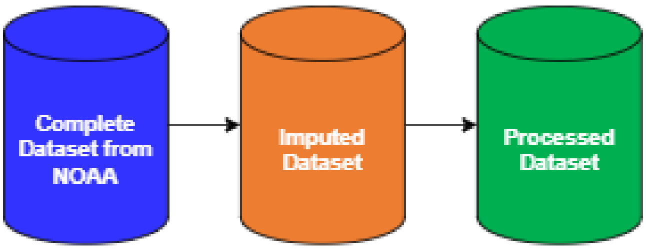


*Figure 11: Data Flow and Transformation*

**Data Analysis and Transformation:**

This component deals with obtaining the data from the source, analyzing, wrangling and transforming it according to our requirements. The dataset was first analyzed to comprehend the various features that were relevant to the prediction of our desired attributes and identify any data quality issues. Subsequently, data imputation was performed to replace values that seemed to be inappropriate. For instance, certain instances in the Dataset contained values of 0 for the feature "len" or length. These values were then replaced according to the respective magnitude of the Tornado by the mean length corresponding to each magnitude. In addition, features which are of little importance were dropped.

**Data Exploration, Visualization and Preprocessing:**

After handling data quality issues, we observed the distribution of data and visualized the relationships between various features in the dataset. Further the correlation between each target features and input features were examined by rendering visualizations such as Box plots, bar plots and other charts. This was followed by processing the data by performing operations like data normalization and label encoding.

**Model Development and Validation:**

After the dataset was prepared, feature selection was performed using score functions such as "chi2" for classification tasks and "mutual_info_regression" for regression problems. After the features of major importance were obtained, the models were trained on a portion of the dataset and subsequently were fitted on test datasets to evaluate the performance of the model. The performance of the models was evaluated using metrics such as "Mean Squared Error" and "R2-Score".

In the case of certain target features, the Hyperparameters of relevant models were tuned to improve the performance of the model.

The solution when combined with real-time data from storm prediction centers, can help predict the magnitude, other characteristics, and effects of tornadoes.

| | st | inj | fat | slat | slon | elat | elon |
|---|---|---|---|---|---|---|---|
| 0 | MO | 3 | 0 | 38.77000 | -90.22000 | 38.83000 | -90.03000 |
| 1 | IL | 3 | 0 | 39.10000 | -89.30000 | 39.12000 | -89.23000 |
| 2 | OH | 1 | 0 | 40.88000 | -84.58000 | 0.00000 | 0.00000 |
| 3 | AR | 1 | 1 | 34.40000 | -94.37000 | 0.00000 | 0.00000 |
| 4 | MO | 5 | 0 | 37.60000 | -90.68000 | 37.63000 | -90.65000 |
| 5 | IL | 0 | 0 | 41.17000 | -87.33000 | 0.00000 | 0.00000 |
| 6 | TX | 2 | 0 | 26.88000 | -98.12000 | 26.88000 | -98.05000 |
| 7 | TX | 0 | 0 | 29.42000 | -95.25000 | 29.52000 | -95.13000 |
| 8 | TX | 12 | 1 | 29.67000 | -95.05000 | 29.83000 | -95.00000 |
| 9 | TX | 5 | 0 | 32.35000 | -95.20000 | 32.42000 | -95.20000 |
| 10 | TX | 6 | 0 | 32.98000 | -94.63000 | 33.00000 | -94.70000 |
| 11 | TX | 8 | 1 | 33.33000 | -94.42000 | 33.45000 | -94.42000 |
| 12 | TX | 0 | 0 | 32.08000 | -98.35000 | 32.10000 | -98.33000 |
| 13 | TX | 0 | 0 | 31.52000 | -96.55000 | 31.57000 | -96.55000 |
| 14 | TX | 32 | 0 | 31.80000 | -94.20000 | 31.88000 | -94.12000 |

*Figure 12: Sample Input to Predict Magnitude of Tornadoes*

Let us suppose we have data in the above format. With the help of the models that we have developed, Meteorologists can make accurate assumptions about the magnitude of an impending Tornado in a particular State of the United States of America. This information will be vital for the general population to prepare for the calamity and take necessary precautions or evacuation protocols.

| | st | mag | inj | fat | slat | slon | elat | elon |
|---|---|---|---|---|---|---|---|---|
| 60104 | AR | 1 | 0 | 0 | 33.58000 | -91.96000 | 33.67000 | -91.94000 |
| 60105 | LA | 1 | 0 | 0 | 32.60000 | -94.02000 | 32.60000 | -94.02000 |
| 60106 | LA | 1 | 0 | 0 | 32.61000 | -93.88000 | 32.64000 | -93.88000 |
| 60107 | LA | 1 | 0 | 0 | 32.76000 | -93.13000 | 32.77000 | -93.12000 |
| 60108 | LA | 1 | 0 | 0 | 30.50000 | -90.83000 | 30.52000 | -90.80000 |
| 60109 | LA | 1 | 0 | 0 | 30.08000 | -90.54000 | 30.08000 | -90.53000 |
| 60110 | AR | 2 | 0 | 0 | 34.70000 | -90.92000 | 34.78000 | -90.90000 |
| 60111 | MS | 1 | 0 | 0 | 31.54000 | -89.53000 | 31.61000 | -89.49000 |
| 60112 | FL | 1 | 0 | 0 | 30.76000 | -87.24000 | 30.77000 | -87.23000 |
| 60113 | NC | 0 | 0 | 0 | 34.88000 | -80.53000 | 34.89000 | -80.52000 |

*Figure 13: Sample Input to Predict Length, Width, Property Loss and Crop Loss*

With approximation of the magnitude of an impending Tornado, the length, width, crop loss and property loss can also be predicted. In addition, farmers can set up wind barriers and prune off branches in time to minimize crop loss.

## V. DATA ANALYSIS, RESULTS AND EVALUATION

### *A. Check if the data preprocessing was suitable for the dataset and models used.*

The dataset includes information related to space and time associated with the tornado-affected regions in the USA. For example, features such as date, time zone, and state give information about the location or spatial occurrence of a tornado. The features such as time and time zone are related to the time occurrence of the tornado. There are features such as magnitude, length, width, and loss which also tell about the intensity/devastation caused by the tornado.

Before passing information to the model, the data needed preprocessing. For example, features such as date needed to be transformed into different features like a month, year, and day in a numerical format.

Categorical features such as state had to be converted into a continuous feature using one hot encoding. Starting and ending latitude/longitude needed proper representation. In the dataset, the ending location of the tornado was mostly zero-valued. We changed the ending location of such instances to the starting location. The data was ready to be passed to the models with the described transformations.

*B. Check if the target feature is balanced or imbalanced.*

In our project, we have built models to deal with problems associated with a tornado. Among all the problems, our primary goal was to predict starting and ending location of a tornado to create a kind of warning system. The target variables of the location of the tornado were balanced and did not need any transformation except updating the ending latitude and longitude of a tornado with zero values. We also built models for dealing with problems such as predicting loss, crop loss, and the length of the tornado. We dealt with the outliers in those target variables by performing imputations.

*C. Check if the evaluation metric is suitable for the data and models.*

We decided to use three main evaluation metrics for the regression model we have planned to build.

1. Mean Squared Error
2. Root Mean Squared Error
3. Mean Absolute Error

We discuss in brief below about each of the metric mentioned above.

**1. Mean Squared Error**

The mean or average of the squared differences between predicted and expected target values in a dataset is used to calculate the MSE. Mean Squared Error (MSE) is a popular error metric for regression problems for several reasons, one of which is the squaring concept, which has the effect of inflating large errors and thus has the effect of "punishing" models more for larger errors.

**2. Root Mean Squared Error**

The root mean square error is the residuals' standard deviation, where the residuals are the measure of how far the data points are from the regression line. In other words, it indicates how concentrated the data is around the line of the best fit. This is another popular error metric for regression as generally regression prediction models are frequently trained using MSE loss, and their performance is assessed and reported using RMSE.

**3. Mean Absolute Error**

The mean absolute error is the average difference between the observations (true values) and model output (predictions). The changes in MAE, in contrast to the RMSE, are linear i.e the MAE does not give distinct sorts of errors more or less weight; instead, the scores rise linearly as the amount of error increases. We decided to use these three regression measures indicated above for the assignment due to its aforementioned behavior.

**Data Analysis**

We are analyzing spatiotemporal data associated with tornadoes. In this project, we are analyzing multiple aspects of a tornado in the USA. A tornado is a phenomenon whose occurrence in nature depends on several factors related to atmospheric weather changes and geographical location. In our dataset, we are only having information limited to the location and intensity of a tornado. Hence, we are predicting the attributes of a tornado with this limited dataset without adding any weather-related features.

The length of a tornado gives a good idea about its intensity of a tornado. Figure 14 shows the distribution of the length of a tornado in a specific year.

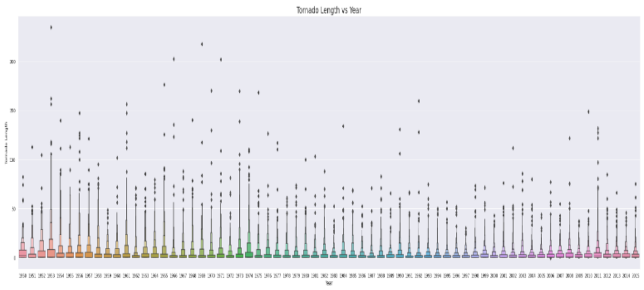

*Figure 14 Tornado Length Box Plot - Year 1950 – 2015*

Since we are dealing with spatiotemporal data, it would be interesting to visualize the mean length of the tornado in a state. Figure 15 shows that some states are more affected by tornadoes than others.

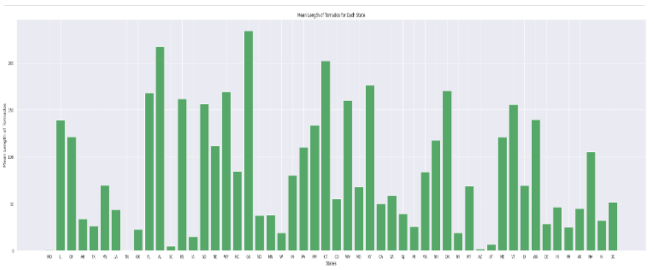

*Figure 15 Tornado Length by State*

The magnitude and length of a tornado are highly correlated. Figure 16 shows that magnitude of a tornado increases with the increase in length.

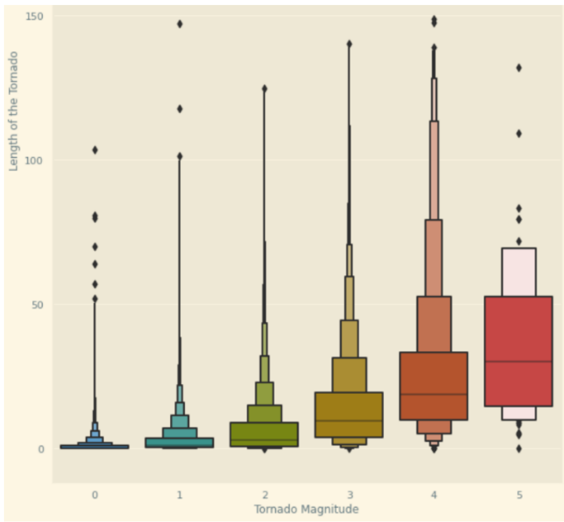


*Figure 16 Magnitude vs Length*

**Model Analysis**

We are predicting/classifying following target variables:

Predict – Starting and Ending Location (Latitude, Longitude) of a tornado

Classify – Magnitude of the tornado

Predict – Loss and Crop Loss associated with a tornado

Predict – Length of a tornado

**Predicting Starting and Ending Location (Latitude, Longitude) of a tornado**

Predicting starting and ending locations of a tornado without weather information is an arduous task. We have used neural network regressors for achieving high accuracy with little information we have got. We have trained five different versions of the neural network model using PyTorch. The loss drops significantly in the initial epochs of training the neural network. We are achieving an accuracy of 97%-99% for predicting the location of the tornado.

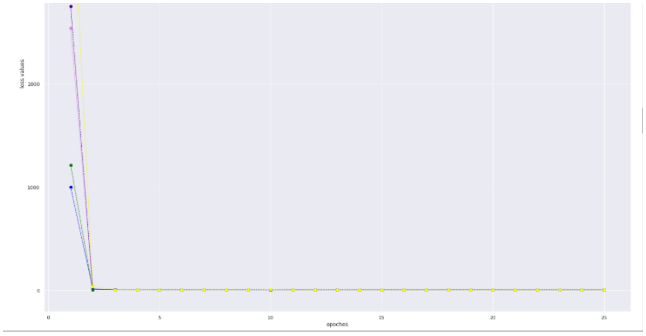

*Figure 17 Neural Network Cost of 3 models*

**Classifying Magnitude of a tornado**

The magnitude has 6 different levels from 0 to 5. Each level shows the intensity of the tornado. The higher value represents a larger magnitude. We have used a random forest classifier for classifying the magnitude of the tornado. We got an accuracy of around 71%. This model gave us better accuracy results as compared to other models such as Logistic regression.

**Predicting Loss and Crop Loss of Tornado**

For crop loss, we used a random forest regressor. We found that the ensemble model produces pretty good results because of using multiple models. Any spurious results produced by one of the models get canceled out due to other models.

For predicting loss, we have used two models: A linear regressor and the XGBoost model. The line regression model produced an R2 value of about -0.23 and a Mean Absolute Error of 2.608. We tried to predict the same feature(loss) using ensemble model XGBoost. The XGBoost model gave us much better results with an R2 value of 0.127 and a Mean Absolute error of 1.52. The XGBoost model produced better results because it uses an ensemble of models. The instances which are incorrectly classified get higher weight and get passed to the next model.

**Predicting Length of Tornado**

For the prediction of the length of the tornado, we used a linear regression model as our baseline model and Random Forest Regressor as the final candidate model. The linear regression model gave us an R2 score of 0.298 and a Mean Absolute Error of 3.52.


*Figure 18 The Random Forest Regressor gave us a much better result - an R2 score of 2.49.*

## VI. CONCLUSIONS

Tornado forecasting remains a near-impossible task, as explained in section 2. However, we did our part in predicting the tornadoes features like latitude and longitude of the starting and the ending location of a tornado, length and the width of the tornado including the damage cost that is caused by the tornado formation which is followed by the magnitude of the tornado.

We built a neural network that was uncovering the relationship between all the features of the tornado and the location of the tornado which we did not anticipate in the beginning. We know from the beginning that predicting the starting and ending location of a tornado is not that easy, however, we were able to figure out the relationship between the parameters of the tornado and the starting and ending location. From the work that is clearly explained in section 3, it is clear that we were able to predict the location based on previous learnings. However, we are certain that this could be enhanced further by the weather conditions as the tornadoes are formed because of the supercell thunderstorms.

We already know that we are trying to predict these tornadoes to save lives and protect resources. We also put some effort into predicting the losses that could occur based on the parameters of the tornado that has already formed. We build an ensemble that could help us in figuring out the crop loss and the land loss which indirectly have an effect on the injuries and the fatalities because of the tornado. We achieved good accuracy in predicting these features which are explained in section 3.

We know that the length and the width of the tornado have a huge effect on the landscape and the loss concerning the economy, we tried to predict these costs in our models which we built using ensembles. Although the length and width of the tornado directly depend on the magnitude or vice versa, we wanted to predict these features as they give us the idea of how much of the area should be evacuated.

Magnitude is one of the most important features that need to be predicted as it is the deciding factor for most of the actions needed to be taken, hence we used the random forest to build an ensemble to predict the magnitude with good accuracy. Since the data has been changing because of the changing weather conditions, we were able to predict the magnitude with approximately 71% accuracy. However, we are confident this could be enhanced with the addition of the weather data which we could not do in this experiment.

Overall, our models in predicting the features of the tornado work even better if we could enhance the current dataset by combing it with the weather dataset that has the wind gust, wind speed, and current climatic conditions of the time. This could be future work as per our knowledge.

## References


[1] "Storm Prediction Center Severe Weather GIS (SVRGIS) Page," *www.spc.noaa.gov*. http://www.spc.noaa.gov/gis/svrgis/ (accessed Aug. 03, 2022).

[2] "sklearn.ensemble.BaggingRegressor — scikit-learn 0.23.2 documentation," *scikit-learn.org*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingRegressor.html (accessed Aug. 03, 2022).

[3] "sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation," *Scikit-learn.org*, Feb. 01, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (accessed Aug. 03, 2022).

[4] "3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.22.1 documentation," *scikit-learn.org*. https://scikit-learn.org/stable/modules/model_evaluation.html (accessed Aug. 03, 2022).

[5] "seaborn.boxenplot — seaborn 0.11.2 documentation," *seaborn.pydata.org*. https://seaborn.pydata.org/generated/seaborn.boxenplot.html (accessed Aug. 03, 2022).

[6] developer, "EF0 Tornado," *Facts Just for Parents, Teachers and Students*, Jul. 19, 2021. https://www.factsjustforkids.com/weather-facts/tornado-facts-for-kids/enhanced-fujita-scale/ef0-tornado/ (accessed Aug. 03, 2022).

[7] NOAA, 2022. Tornadoes. [online] Noaa.gov. Available at: https://www.noaa.gov/education/resource-collections/weather-atmosphere/tornadoes [Accessed 3 August 2022].

[8] Government, USA., 2022. The Enhanced Fujita Scale (EF Scale). [online] Weather.gov. Available at: https://www.weather.gov/oun/efscale [Accessed 3 August 2022].

[9] NOAA, 2022. Tornado Detection. [online] NOAA National Severe Storms Laboratory. Available at: https://www.nssl.noaa.gov/education/svrwx101/tornadoes/detection/ [Accessed 3 August 2022].

[10] MIT, 2022. Intelligent Tornado Prediction Engine | MIT Lincoln Laboratory. [online] Ll.mit.edu. Available at: https://www.ll.mit.edu/r-d/projects/intelligent-tornado-prediction-engine [Accessed 3 August 2022].

[11] Geogdzhayeva, M. and Repina, I., 2022. Using Machine Learning to Predict Tornadoes Based on Sounding and Reanalysis Data. [online] NASA/ADS. Available at: https://ui.adsabs.harvard.edu/abs/2019AGUFM.A31O2785G/abstract [Accessed 3 August 2022].

[12] Yves, J., 2022. Predicting tornado magnitude with Machine Learning. [online] Medium. Available at: https://medium.com/@yves.jacquot/predicting-tornado-magnitude-with-machine-learning-c76df84d7872 [Accessed 3 August 2022].

[13] Fuad, A., Boris, M. and Vladislav, P., 2022. [online] Arxiv.org. Available at: https://arxiv.org/ftp/arxiv/papers/1806/1806.05857.pdf [Accessed 3 August 2022].

[14] TENNIS, J., 2022. Storm Prediction Center. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/jtennis/spctornado [Accessed 3 August 2022].

[15] Kaggle, 2022. Storm Prediction Center. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/jtennis/spctornado/code [Accessed 3 August 2022].

[16] Lagerquist, R., Homeyer, C., Gagne, D. and Mcgovern, A., 2022. Deep Learning on Three-Dimensional Multiscale Data for Next-Hour Tornado Prediction. [online] researchgate.net. Available at: https://www.researchgate.net/publication/341339734_Deep_Learning_on_Three-Dimensional_Multiscale_Data_for_Next-Hour_Tornado_Prediction [Accessed 3 August 2022].

[17] Shapiro, A., 2022. NPR Cookie Consent and Choices. [online] Npr.org. Available at: https://www.npr.org/2019/05/30/728387095/scientists-know-how-tornadoes-form-but-they-are-nearly-impossible-to-predict [Accessed 3 August 2022].

[18] Fernández, L., 2022. Economic damage caused by tornadoes in the U.S. 2020 | Statista. [online] Statista. Available at: https://www.statista.com/statistics/237409/economic-damage-caused-by-tornadoes-in-us/ [Accessed 3 August 2022].