Name: Aaron Jay (ajj73), Carolyn Shi (cs947) **Course:** ORIE 4741

Storm Fatality Prediction

Introduction

In recent years, there has been a rise in occurrences of storm events in the United States, as well as weather-related injuries and fatalities. In 2018 alone, there have been nearly eight hundred deaths and over double that in injuries. We seek to understand the factors surrounding these fatalities, such as measurements of the storm itself or geographic factors of where fatalities have occurred. It is the goal of this project that the insights obtained from this work can be used by local and regional governments for forecasting damage from future storms. Having a reliable estimate of potential fatalities and a good understanding of factors that cause them will provide impetus for making infrastructural and economic preparations for the storm. Understanding the major factors that contribute to fatalities occur can also help EMS and local authorities identify areas of risk and better prepare for the types of injuries they need to treat.

Using the historical storm event data collected by the National Weather Service (NWS)[1], we hope to shed light on storm events in the future by predicting the damages that will be caused by these phenomenon. The measure we have chosen is the number of fatalities. This is motivated in part by the setting of the problem, as well as practical concerns from the data. The data contains information about storms since 1950, including location, storm measurements, and circumstances surrounding each fatality, and narratives (text summaries) of each event. Information about injuries are not presented in this dataset, which has forced us to limit our scope.

Data

Data Overview and Exploration

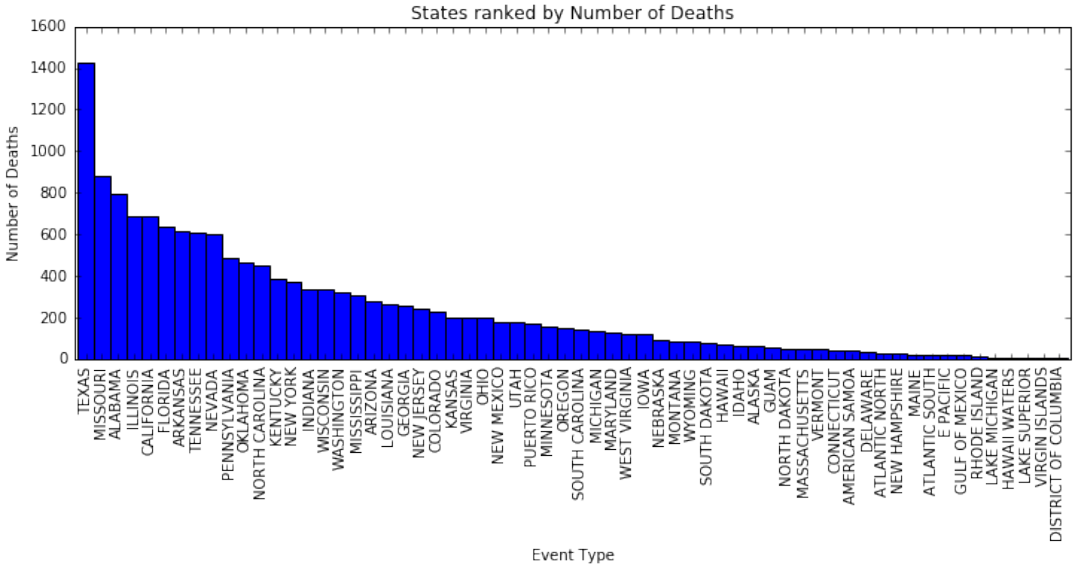
For each year of data provided National Weather Service (NWS), there are three files: Details, Locations, and Fatalities. The Detail files contain the bulk of the information about a storm, including its measurements and narratives. Location files provide geographic coordinates for where each storm event occurred. Fatality files contain additional information regarding a casualty, injury or death, including date/time, demographic information about the victim, and the location/manner in which they were harmed.
This data consisted only of tornado events until 1955, when information about thunderstorm, wind, and hail were added from publications containing accounts of those events. Then in 1996 the event types were expanded to what currently exists in the database: 48 different storm types. The due to the historical development of the data collected, a majority of the features are only relevant to the event types mentioned previously. Thus, we are choosing to focus our attention on storm events that have had more descriptive data collected. These are the four original categories: tornadoes, wind, hail, and thunderstorms, as well as other common events like floods and heat.

In the raw data, there are 50,000-60,000 examples for each year of storm event details (though there are significantly fewer examples with fatalities). The features of the data can be grouped into the following categories:

Name	Description	Data Types
Date/time	Date and time when a storm event occurred.	Date/time
Geographic Details	Coordinates where the event occurred, as well as place names (state, county, city).	Real, Text
Event Type	NWS classification of weather events among 48 available types.	Nominal
Casualties and Damages	Counts of injuries, fatalities, and economic damage estimates.	Real
Event Characteristics	Measurements and classifications of a storm event (F-scale for tornadoes, wind speeds, hail size, etc.).	Real, Ordinal, Nominal
Narratives	A textual account of developments during a storm event. Includes measurements of weather conditions and accounts of fatalities.	Text
Fatality Details	Details about a storm event victim, and circumstances of death.	Real, Nominal

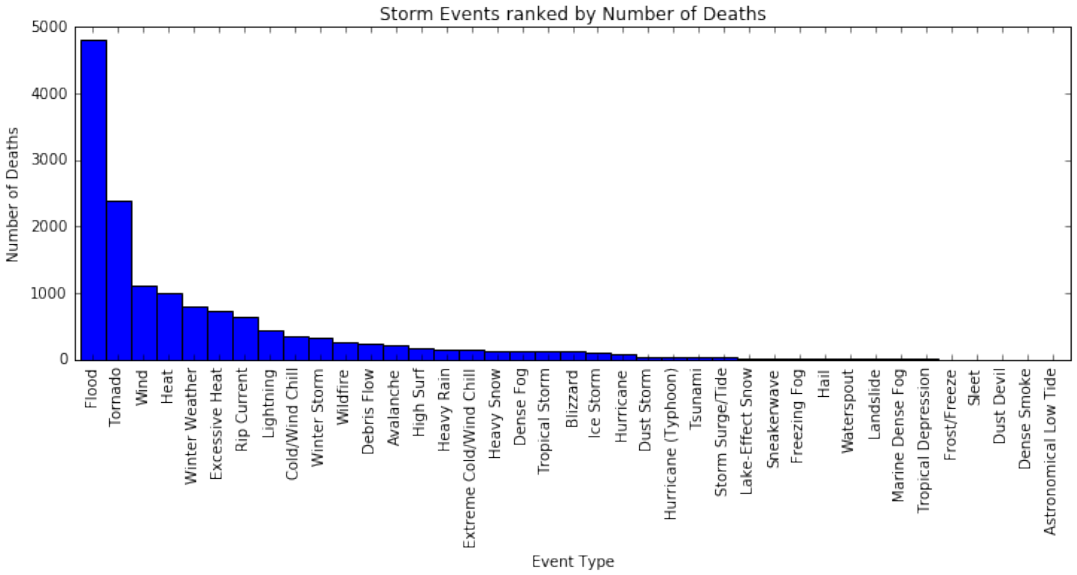
Due to the historical data collection process, we knew that there were more records for certain weather events. In order to avoid this bias in our models, we restricted our scope to data from 1996 onwards, when records of all 48 event types were collected.

We investigated two non-numeric factors for casualties in the dataset: location (state) and event type. This was done in order to quickly determine whether the data needed to be partitioned for more effective models, or whether we we would need to include additional data from other sources.



Since there is a skewed pattern towards certain states having casualties, we use this as a basis for including more geographic data in the future. As a result, we took two other categories of features into consideration: infrastructure and terrain.

For infrastructure, we hypothesized that the number of casualties was also related to the availability of urgent care providers and emergency services. Thus, we used a hospital dataset[2] to determine the following: distance to the nearest hospital, as well as the number of hospitals available locally (in the county) and regionally (in the state).



For terrain, we obtained data on US land classification[3] based on standards set by the International Geo-sphere–Biosphere Programme (IGBP). These classes indicate the type of vegetation present in a general area.

The main dataset is actively managed and updated by the National Weather Service, thus it is well-formatted and well-documented. We have not had or anticipate huge issues with corrupted data.

Data Preprocessing and Missing Value Handling

We took two approaches to adding features to the data: simple computations involving numeric columns, and feature encoding for non-numeric columns.

Within the NWS dataset, we initially combined the *Direct Deaths* and *Indirect Deaths* columns in order to have one feature to predict. Features we were able to extract were *Duration* (based on the recorded start and ending times), and *Narrative Length*. By adding on the hospital dataset, we were able to compute the distance from a fatality to the closest hospital. This was done through the use of a k-d tree, which allowed us to find the closest hospital in linear time. For the remaining nominal and ordinal features, (F5 scale for tornadoes, Flood causes, Fatality Locations, etc.), we used one-hot, ordinal, and multi-hot encoding. In the end, we ended up with a dataset of 59 features.

Nearly all of the features had missing entires in the resulting dataset, though most columns did not require imputation due to the context of the information (for example, a tornado should correctly have missing data in the *Flood Cause* column). For columns that did need to be filled in, we attempted to find substitute columns in the dataset that could help us calculate/infer a proper value. We used this idea frequently for geographic data, since those columns were the link to other information. One example of our method was replacing missing coordinates in the Location file with corresponding coordinates of an event in the Details file. Another example was imputing missing coordinates by simply finding the coordinates of the centroid of a county.

Models

The dataset for training models was separated into training, validation, and test sets (64%, 16% and 20%, respectively). We also chose to focus on the three categories with the most events which are: floods, tornadoes, and winds. We first train models on the tornadoes dataset since it is the event type that has the most events and then we will use our model to predict deaths for floods and winds. We will use Mean Absolute Error (MAE) and the coefficient of determination to evaluate the success of each model.

Linear Models

1. Linear Regression

First, we fit an ordinary least squares regression with squared loss to serve as a baseline predictor for deaths. We will see that our validation set error is actually lower than our training set error which may be a result of many factors that include, training-validation-test set split and a small dataset. However, we use cross validation to get a more accurate estimate of the error. We will see this pattern continue through the rest of our models.

	MAE	Coefficient of Determination
Training Set	2.713147725769235	0.7422334404169714
Validation Set	2.400473944482946	0.3709905056002396
10 fold Cross Validation	-2.9377890416372208	

2. Ridge Regression

Next, we try to improve the model by adding a regularizer. We start with Ridge regression to address any collinearity issues we may have between the features. Unfortunately, this does not improve the model as the MAE and R^2 values remain almost identical.

	MAE	Coefficient of Determination
Training Set	2.7681370134536283	0.7393472765426321
Validation Set	2.411621933412783	0.36694207551684077
10 fold Cross validation	-2.9577066757162074	

3. Lasso Regression

After Ridge Regression failed to improve our model, we turn to using a Lasso regularizer with alpha = 0.1. The MAE of the testing and validation sets resulted in lower values and higher R^2 value for the validation set, making it the best model thus far.

	MAE	Coefficient of Determination
Training Set	2.5006414400993053	0.7203888518301895
Validation Set	2.037223035723083	0.41676908476165936
10 fold Cross validation	-2.6194774842481463	

4. Elastic Net

We try an Elastic Net model that is a combination of the l1 and l2 regularizers in an attempt to extract the benefits of both models. However, as you can see from the results in the MAE and R^2, that was not the case.

	MAE	Coefficient of Determination
Training set	3.2846997819732455	0.09079347919467362
Validation set	2.7248610257742496	0.16251227895806752
10 fold cross validation	-3.1663515912905873	

5. Stochastic Gradient Descent with Squared Loss

We fit a model using the Stochastic Gradient Descent method to minimize squared loss. This model achieves the best R^2 value by at 0.05 margin but the MAE is still comparable to the other models.

	MAE	Coefficient of Determination
Training Set	2.963852439515893	0.7295720010091092
Validation Set	2.5664755376037545	0.3158095185452503
10 fold cross validation	-3.128446489014581	

6. Stochastic Gradient Descent with Huber Loss

We try Stochastic Gradient Descent again but with Huber Loss in order to control for any outliers that we may have. We checked the maximum residual error of a 10-fold cross-validation of SGD with squared loss and it was one magnitude larger than the maximum residual error of the other models. While the MAE improves from the SGD with squared loss, it is not any better than Lasso Regression and has the worst coefficient of determination of all the linear models.

	MAE	Coefficient of Determination
Training set	2.845655810793211	-0.03729603387299951
Validation set	2.046287889048312	-0.06555597004403935
10 fold validation set	-2.677254237619757	

None of the linear models performed great, although a couple had average results that might be acceptable. Most data scientists would agree that these models are not ready to be applied and utilized. We thought that since many of our variables are binary that linear models may not be the best to predict our output variable.

Trees

We try to fit some decision trees because they can capture non-linear relationships between the features and address any underfitting problems that our previous models may have.

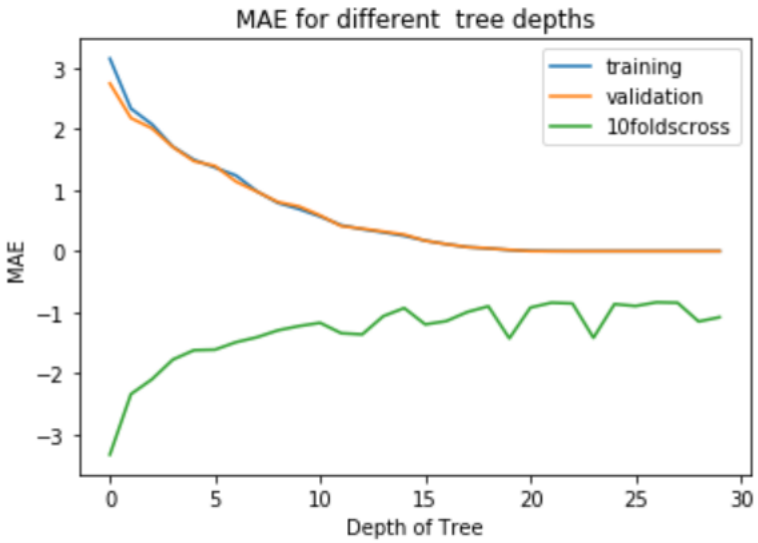
1. Decision Trees

We first fit a decision tree with depth = 2, which has an MAE that is on average better than the linear models and has an R^2 value that is significantly better on both the training and validation sets.

	MAE	Coefficient of Determination
depth = 2, training set	2.33312844849767	0.880827372560543
depth = 2, validation set	2.1778899115170036	0.6975003993299013
depth = 2, 10 fold cross validation	-2.33807590627926	

However, since this model underfits, we try different depths from 0 - 30 to find the best decision tree, which are shown in the plot below. We care more about the performance on the validation set than the training set since we want our models to be able to perform well on data it has not seen.

We can see that the training and validation MAE level off around depth = 20 and that the 10 folds cross validation error goes through a cyclical pattern. Thus, we choose depth = 21 as our best decision tree model with a training set MAE of 0.001880 and 10 folds cross validation of -0.868051.



2. Bagging

Our dataset is of an average size and in order to try more models, we want to try Bootstrap Aggregation or Bagging to address these issues. We try a bagging model but it results in an even higher MAE with lower accuracy.

	MAE	Coefficient of Determination
Training set	2.5412049417758054	0.7313259075882517
Validation set	2.1781516939068566	0.6803245466672825
10 fold cross validation	-2.7912628792407714	

3. Random Forest

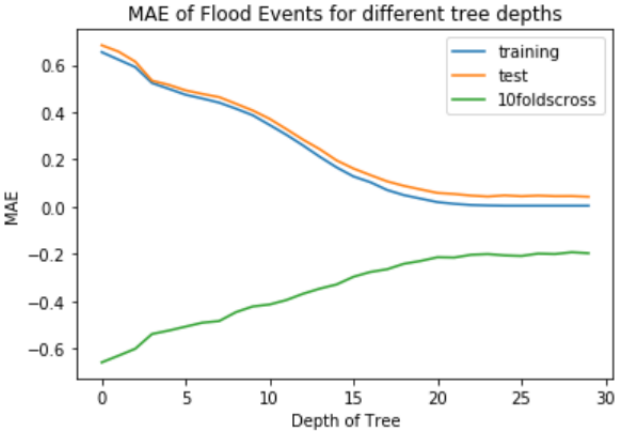
Lastly, we fit a Random Forest which splits at each node as a result of a subset of features instead of all the features like in the bagging model. Random forests can improve accuracy and address overfitting and we hoped that this would be an improvement on the decision tree with depth = 5 since it overfit but had a low training MAE.

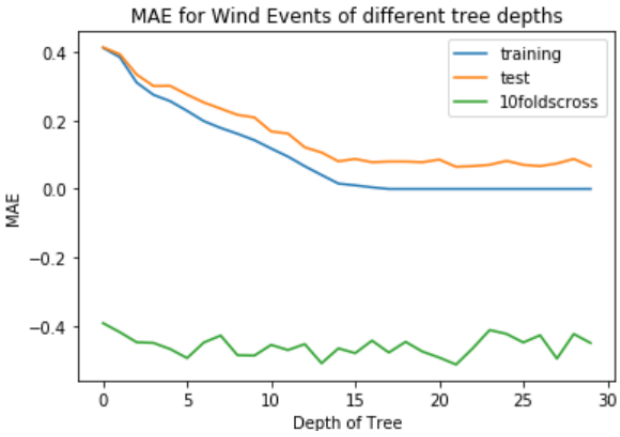
	MAE	Coefficient of Determination
Training set	2.4430864522558786	0.955396259762049
Validation set	2.3225600172824405	0.8169841130038518
10 fold cross validation	-2.7912628792407714	

Other Storm Event Types

After using Tornado Storm Events to select the best model to predict deaths, we try evaluating our chosen model on other types of storm events, specifically floods and winds.

As you can see in the below graphics, both test errors plateau around a tree depth of 20 or couple notches before for wind events. For flood events, the 10 fold cross validation error also converges around depth = 20. For wind events, the 10 fold cross validation error does not seem to improve very much by increasing the tree depth but it does not get worse either so picking depth = 20 is also reasonable for wind events.





Results and Implications

Model Assessment and Insights

We ran two types of models: linear regressions and decision trees. For linear regression, we chose to start off with baseline of squared loss, then moved to add a l1 (Lasso), l2 (Ridge), and combination l1/l2 regularizer (Elastic Net). The ordinary least squares, lasso, and ridge all performed similarly in terms of accuracy with Lasso Regression having the smallest error and best accuracy out of the three. SGD with Huber Loss had about the same MAE for the 10 folds cross validation as Lasso regression but since the coefficient of determination for SGD with Huber loss was close to 0, indicating that the model was not a better predictor than just a constant, we choose Lasso regression for the best linear model.

For trees, we started off with a basic decision tree with depth = 2. This resulted in a much better R^2 value than any of the linear models. It has a better training set MAE than all of the other models thus far, however Lasso regression still has a better validation MAE (but not 10 fold cross validation). Following this, we experimented with different tree depths from 0 to 30 and found that the training error levels off around depth = 20 and the 10 folds cross validation error follows a cyclical pattern. We choose depth = 21 as the best decision tree model to minimize depth since adding more layers complicates the interpretability of the model with not much model improvement. We also fit bagging and random forest trees to our data but they do not improve upon the 10 folds cross validation set MAE.

Taking a look at the weights each model produced provided some interesting results. Table 1 in the Appendix shows the weights out of 100 of each of the tree models. The features of closest hospital distance and number of hospitals in the county and state have an interesting evolution from the model on the left to the right which could be argued go from simple to more complex. In the most simple decision tree of depth = 2, none of the factors matter, but as you go to depth = 5 and using the bagging model, the different factors rise in weight. In the RandomForest model, one could make the very loose conclusion that the number of hospitals

affects deaths from storm events and thus local government can take action accordingly. Also, every model gave the highest weight to Fatality Location of Long Span Roof. A long span roof is a roof that exceeds 12m in span [4]. The fact that this factor is so heavily weighted indicates that more deaths occur around long span roofs and may suggest that homeowners should not have long span roofs or shorter roofs to have a lower likelihood of accidents during a storm event. These iare the kind of insight that we hope to glean in order to help governments and individuals avoid fatalities during storm events in the future.

Fairness

Since the predictions are made from features based on weather events and geographic location, there are no features that are explicitly tied to protected groups or are regulated by federal law. Though the original dataset does have personal details regarding each victim (gender, age, etc.), these columns were not used since we only considered the number of fatalities that occurred. The predictions made by the model are meant to inform existing government preparations for storm events. The only bias that existed in the original data was the number of records for each type of storm event, but that too was addressed by restricting how much data we used for our analysis.

Moving Forward

Though we were able to successfully find a model that performed well, future work in this topic should be concerned with the improvement of data quality.

Through historical developments, we note that there is a plethora of data for a small number of storm events. As such, there are additional columns that provide more details about such events that contributed to the success of our models. For future work, we recommend extracting keywords and measurements from the episode and event narratives. This could allow the same models to be applied to other storm events such as heat or winter weather.

Our question was originally concerned with both the number of fatalities, as well as the number of injuries. Unfortunately, the Nation Weather Service provides no data regarding injuries, which limits the scope of the models. This in turn adversely impacts the usefulness of these models when planning for storm events.

Conclusion

Overall, we are pleased to find that existing storm events data is enough to give good predictions for storm fatalities. We also observe our methods can easily be tailored to different kinds of events. Thus, we encourage local governments to consider using such machine learning models to inform their decisions when planning for a storm.

References

[1] National Climatic Data Center, NESDIS, NOAA, U.S. Department of Commerce, NCDC Storm Events Database, 2019. [Online] Available: `https://www.ncdc.noaa.gov/stormevents/`

[2] Oak Ridge National Laboratory, Hospitals, 2019. [Online]. Available: `https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals`

[3] NASA Earth Observations, Land Cover Classification (1 year), 2019. [Online]. Available: `http://neo.sci.gsfc.nasa.gov/view.php?datasetId=MCD12C1_T1`

[4] "Long span roof," Long span roof - Designing Buildings Wiki, 01-Dec-2019. [Online]. Available: `https://www.designingbuildings.co.uk/wiki/Long_span_roof`.

Feature	DecisionTree2	DecisionTree5	BaggingTreeAvg	RandomForest
TOR_LENGTH	46.015614	9.274940	29.7242	15.860476
TOR_WIDTH	0.000000	34.596589	2.6219	3.138983
DURATION	0.000000	0.000000	0.000000	0.000000
CLOSEST_HOSPITAL_DIST	0.000000	0.374783	0.000000	3.414959
NUM_COUNTY_HOSPITALS	0.000000	0.000000	0.000000	4.100639
NUM_STATE_HOSPITALS	0.000000	0.000000	3.9692	1.608653
EPISODE_LENGTH	0.000000	0.000000	2.8397	2.723155
LAND_COVER_CLASS_1.0	0.000000	0.000000	0.000000	0.002055
LAND_COVER_CLASS_2.0	0.000000	0.000000	0.000000	0.000000
LAND_COVER_CLASS_3.0	0.000000	0.000000	0.000000	0.222153
LAND_COVER_CLASS_5.0	0.000000	0.000000	0.000000	0.118332
LAND_COVER_CLASS_6.0	0.000000	0.000000	0.000000	0.229984
LAND_COVER_CLASS_7.0	0.000000	0.000000	0.000000	0.000000
LAND_COVER_CLASS_8.0	0.000000	0.000000	0.000000	0.000000
LAND_COVER_CLASS_9.0	0.000000	0.000000	0.000000	0.083261
LAND_COVER_CLASS_10.0	0.000000	0.000000	0.000000	0.000000
LAND_COVER_CLASS_11.0	0.000000	0.010151	0.000000	0.379373
LAND_COVER_CLASS_12.0	0.000000	0.000000	0.000000	0.000390
LAND_COVER_CLASS_13.0	0.000000	0.000000	0.000000	0.119783
LAND_COVER_CLASS_14.0	0.000000	0.000000	2.9762	0.661651
LAND_COVER_CLASS_15.0	0.000000	0.000000	0.000000	0.312615
LAND_COVER_CLASS_17.0	0.000000	0.000000	0.000000	0.000000
TOR_F_SCALE_EF0	0.000000	0.000000	0.000000	0.000014
TOR_F_SCALE_EF1	0.000000	0.000000	0.000000	0.019497
TOR_F_SCALE_EF2	0.000000	0.000000	0.000000	0.019057
TOR_F_SCALE_EF3	0.000000	0.000000	0.000000	0.152554
TOR_F_SCALE_EF4	0.000000	1.361635	0.000000	1.357006
TOR_F_SCALE_EF5	0.000000	2.966812	2.2898	3.971877
TOR_F_SCALE_F1	0.000000	0.000000	0.000000	0.000446
TOR_F_SCALE_F2	0.000000	0.000000	0.000000	0.000830
TOR_F_SCALE_F3	0.000000	0.000000	0.000000	0.189117
Ball Field	0.000000	0.000000	0.000000	0.000000
Boating	0.000000	0.000000	0.000000	0.000000
Business	0.000000	0.000000	4.9446	1.653103
Camping	0.000000	0.000000	0.000000	0.000000
Church	0.000000	0.485516	3.9036	7.648528
Golfing	0.000000	0.000000	0.000000	0.000000
Heavy Equipment/Construction	0.000000	0.000000	0.000000	0.000000
In Water	0.000000	0.000000	0.000000	0.002532
Long Span Roof	53.984386	50.105052	46.7308	44.818427
Mobile/Trailer Home	0.000000	0.000000	0.000000	0.761346
Other/Unknown	0.000000	0.000000	0.000000	0.116548
Outside/Open Areas	0.000000	0.000000	0.000000	0.066124
Permanent Home	0.000000	0.000000	0.000000	1.353415
Permanent Structure	0.000000	0.824522	0.000000	3.517999
School	0.000000	0.000000	0.000000	0.222119
Telephone	0.000000	0.000000	0.000000	0.000000
Under Tree	0.000000	0.000000	0.000000	0.000170
Vehicle/Towed Trailer	0.000000	0.000000	0.000000	1.152826

Table 1: Weights of the Features from the Tree models for Tornado Events