# Predicting Cost of Property Loss Incurred by Tornado activity in the USA

## ABSTRACT

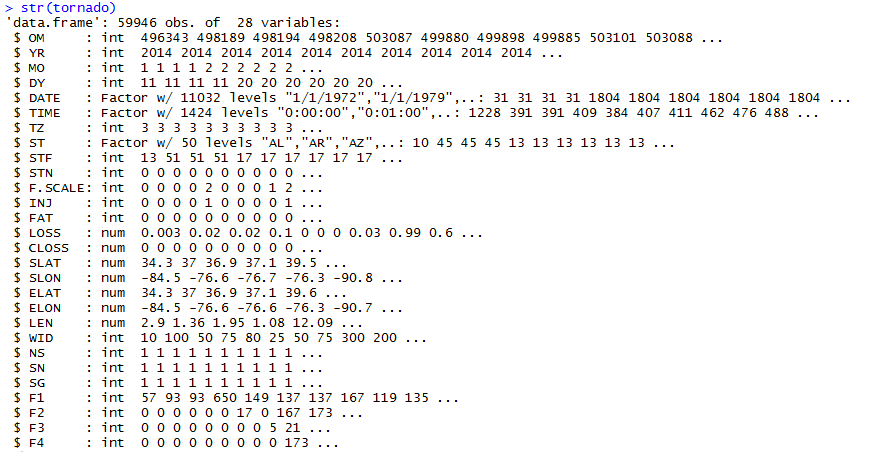
We illustrate a statistical model for predicting cost of property loss due to the tornado activity in the United States during 1950 to 2014. The model predicts the total loss of property given the tornado reports during the years 1950 to 2014 using the data from The National Oceanic and Atmospheric Administration’s (NOAA) Storm Prediction Center (SPC). The sustainability of positive amount of property loss in the dataset is estimated from a classification model using logistic regression on the training data. The model yields mean absolute error of ±$3.1 million of the total losses produced from the tornado which is classified based on it features such as length, width, intensity etc. These relationships are broadly consistent with understanding financial models and have a high commercial value for the insurance companies. The insurer can predict the damage which can occur due to the natural calamity and can estimate how much damage they will be responsible for, thus, it will help the company to calculate premium charges for the incoming customers in that area for an year.

## INTRODUCTION

The United States experiences more tornadoes than any country on Earth with an annual average during the 3 years (2009–11) exceeding 1350. According to the National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center, the annual average number of killer tornadoes since 2009 is 30 with total deaths exceeding 600. Annual statistics are only part of the story. The number of tornado reports varies widely from one year to the next. We are trying to extract information to generate a useful prediction model which estimates losses after a short duration of occurrence of the calamity. Most of the statistics is collected shortly after the tornado has occurred, however the property loss is a lagging statistic which requires more time before it can be properly assessed.

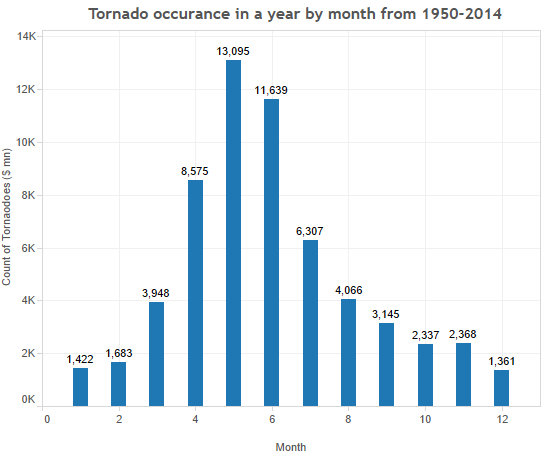
## DATA OVERVIEW

The Storm Prediction Center (SPC) maintains a data-set of all reported tornadoes in the United States from 1 January 1950 to the present. The SPC dataset is the most reliable archive available for tornado and related studies. (We download the dataset from <http://www.spc.noaa.gov/gis/svrgis/>.) The dataset has 28 features of which 11 variables are numerical fields namely, Tornado Number, Injuries, Fatalities, Loss, Crop-loss, Length, Width, Latitude, Longitude, E-Latitude, and E-Longitude. (Figure 1.)



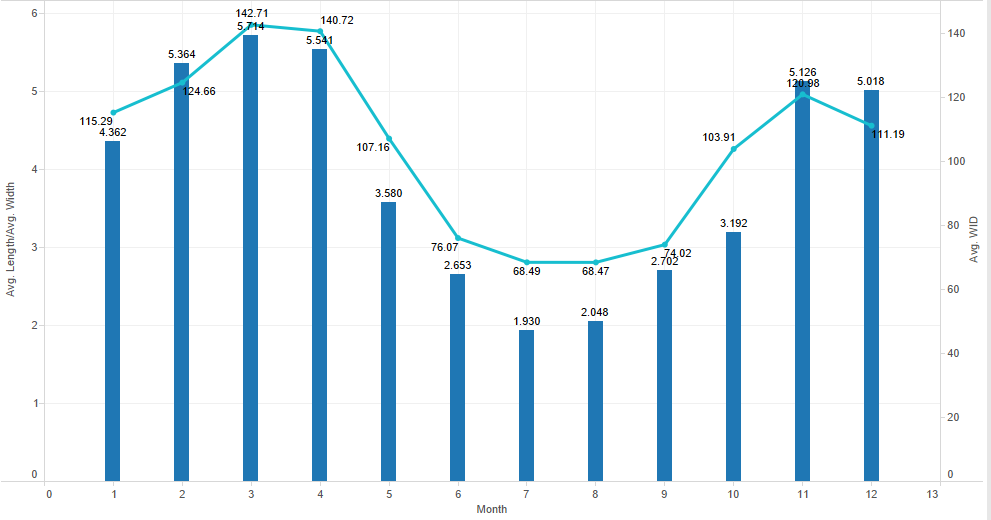
**Figure 1.**

## DATA EXPLORATION AND VISUALIZATION

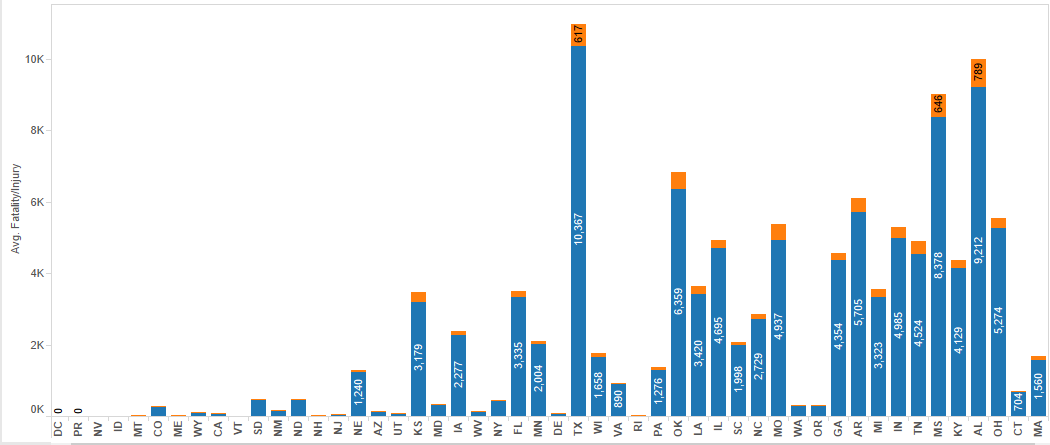
The analysis of the dataset has been performed in Weka, Azure, R, Tableau and MySQL. All graphs have been generated in R and Tableau.

The code and the output snippets have been shown for reference and reader’s clear understanding.

Figure 2 shows the distribution of all tornado reports by month. There is a marked peak in activity during May with the main season running from April through July. Here we focus on this month three- period. Of the 59,947 tornado reports in this region over the period 1950–2014, 66.08% of them occurred in April, May, June or July. We have classified these months as storm seasons and the other months as no-storm seasons.  **Figure 2.**

Figure 3 shows the distribution of all tornado intensity reports by month. There is a marked peak in the intensity for the seasons of fall and spring. There is a significant dip in the intensity of the tornado in summer season.

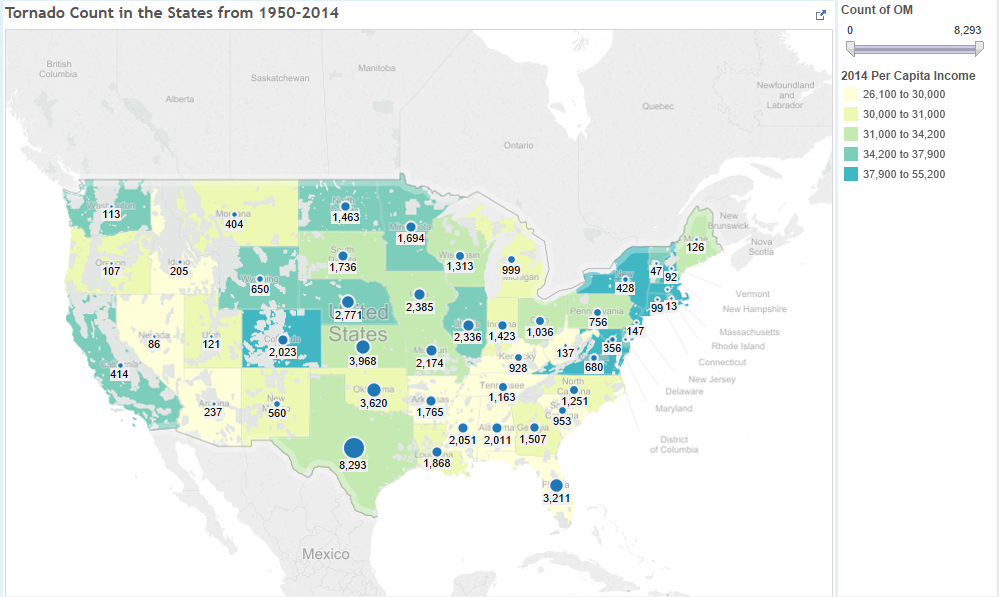
**Figure 3.**



**Figure 4.**

Figure 4 shows the distribution of fatalities and injuries in the given US states during 1950 to 2014. The worst hit state being MA with an average count of 10,367 injuries and 637 fatalities. For some states such as DC, Porto Rico, Idaho the data is not available.

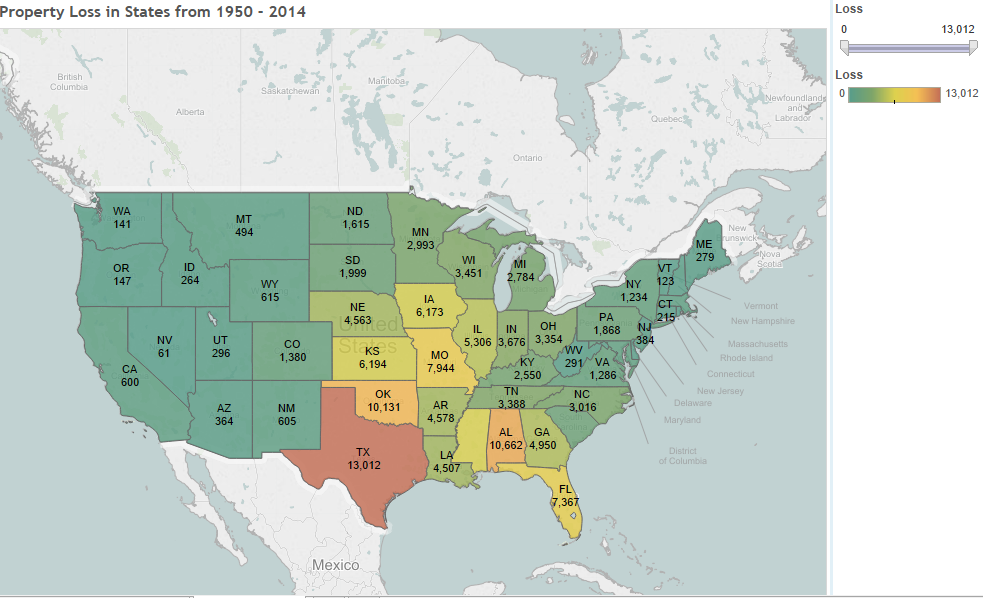
Figure 5 shows the distribution of tornado count in the given US states over the period of 50 years. TX is the state hit by maximum tornadoes in the last 50 years with ~ 8200 tornadoes. It can be observed that the states which are hit by tornadoes badly, the per capita income has not been raised as much as compared to the other states. The occurrence of natural calamities might also be contributing factor in lower rate of increase in the per capita.



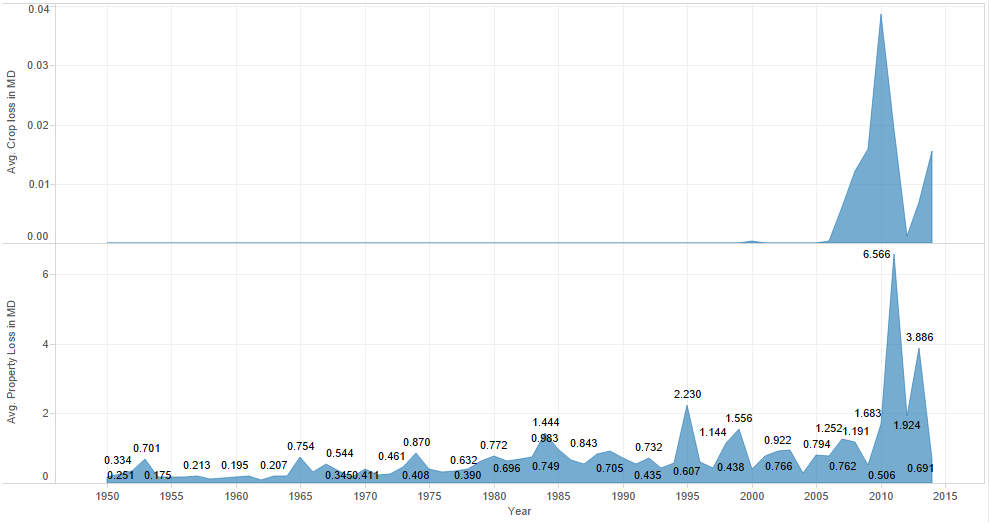
**Figure 5.**

Figure 6 shows the property loss incurred by various state in the United States of America in the last 50 years. It can be seen that the worst hit state has been the state of Texas and the maximum losses of ~ $13K have occurred in this state. There are state with very few losses as well such as WA, NV but it can be seen that even with very few tornado hits in these states the loss is comparatively higher than those states which are hit by tornadoes more often.

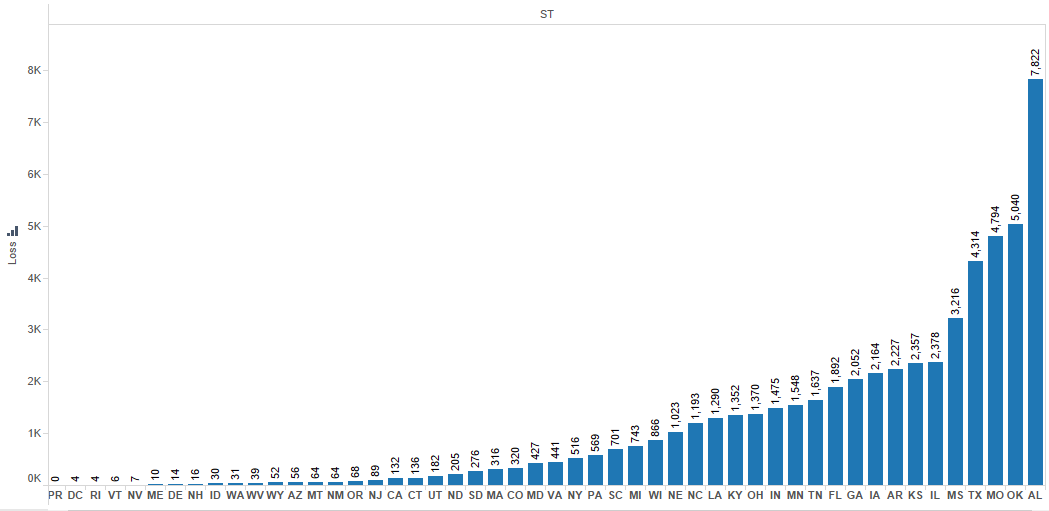
With the above observations we can infer that the most badly hit states are those in the central region thus making it prone to tornadoes.



**Figure 6.**

The figure 7 below describes the loss trend present in the dataset. It can be seen that the crop loss has a lot of null values until 2005. A peek in the property and crop loss can be described by a major calamity in 2010 occurring once in several years.

**Figure 7.**



**Figure 8.**

Figure 8 above shows the distribution of property losses per state in last 50 years. This has been distributed in the order of state severity where the severity is defined as the amount of losses that the state has incurred due to the tornado activity in the last 50 years of available data in the dataset. Thus we have classified states broadly in four state severity categories.

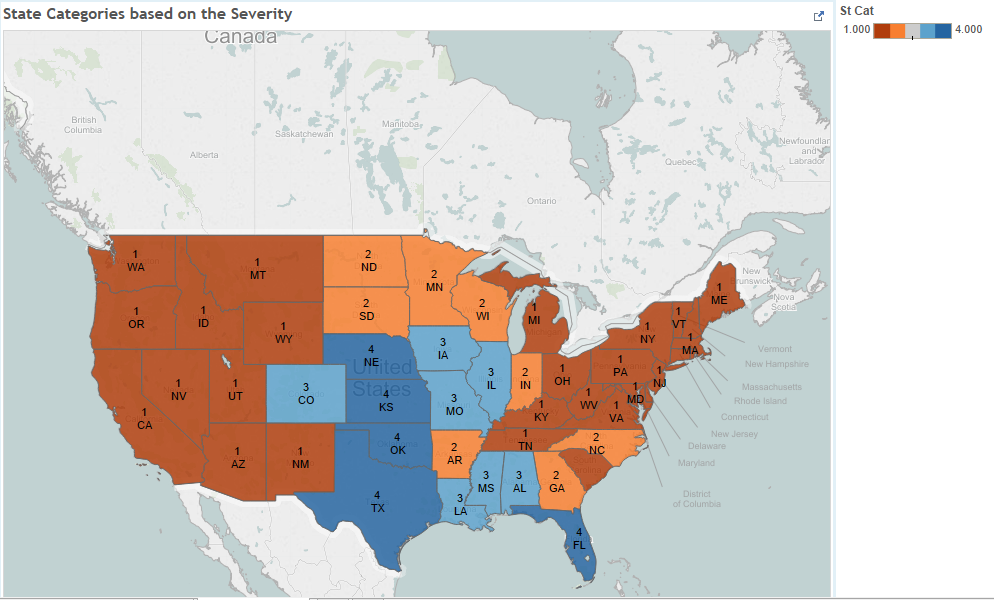
## VARIABLE SELECTION

After analyzing the complete dataset provided by NOAA, we determined seven variables that would be helpful to predict the property loss.

Table1: List of Variables

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Property Loss | Outcome Variable: Amount of loss (in Millions of Dollars) |
| Storm Seasons | Categorical Variable: Tornado Season:1, Off-Season: 0 |
| State Severity | Categorical Variable of Percentage of National Tornadoes in State: 1(<2%), 2(2-3%), 3(3-4%), 4(>4%) |
| F-Scale | Categorical variable of Fujita Scale that measures tornado severity. 0-5, 5 being most destructive |
| Injuries | Number of Injuries occurring due to the tornadoes |
| Fatalities | Number of Fatalities occurring due to the tornado |
| Length | Distance the tornado traveled (in miles) |
| Width | Diameter of the Tornado(yards) |

Looking at our data, there was a significantly higher frequency of tornados in April through July, and significantly less in other months. This was a consistent pattern, which was similarly present since 1950. Thus, Storm Season is a binary variable indicating whether the tornado occurred during tornado season or not. The State variable was consolidated into four categories based on the percentage of all tornados that hit the state. For example, Texas contains roughly 9% of all tornados that occurred in the United States for our dataset. Given this, we categorized the states accordingly into the State Severity variable.



**Figure 9.**

The State variable was consolidated into four categories based on the percentage of all tornados that hit the state (see figure 8). For example, Texas contains roughly 13.8% of all tornados that occurred in the United States for our dataset. Given this, we categorized the states accordingly into the State Severity variable. Figure 9 shows a color-coded map of the categorized states. Another categorical predictor is the Fujita scale (F-Scale). This is based on an aerial survey of the affected area that gives a categorical estimate of the severity of damage. The scale is measured in increasing severity from zero to five. Further details of all variables are included in Table 1.

## PROCEDURE OVERVIEW

In order to predict a dollar estimate for damages, two regression models were fit to the data. Our approach is to use the nineteen years of the data for training and validating our models. Our ultimate goal is to predict the approximate dollar amount of property loss from a given tornado. To estimate this using a linear regression, a logarithmic transformation of the independent variable Property Loss was necessary in order to ensure the assumptions of linear regression hold, specifically that the dependent variable exhibits a linear relation with each of the predictors. Due to the high frequency of zero-valued observations for property loss, taking the logarithm will cause computational problems. Thus, before fitting the linear regression we first classified the observations in the dataset by predicting whether or not there was a positive amount of property loss sustained. To perform this classification of the data we fit a logistic regression on the training and validation portion of the dataset. Independent of the logistic regression, we fit a linear regression model on the non-zero observations of property loss in 1996- 2014 dataset. A k-fold validation process was run and the model with the lowest mean absolute error was chosen. The two models were tested to determine how well they were able predicted future values of property loss. We could then predict whether or not a particular storm resulted in positive property loss. The observations from the validation data for which we had predicted positive property loss were input into the linear regression model to predict the actual dollar amount of damage sustained by that tornado. This process resulted in the predicted estimate of property damage. Statistics of the results were collected and then assessed for the level of commercial utility they present. The subsequent sections illustrate the result of this procedure and conclusions drawn from those outcomes.

## CLASSIFICATION MODEL TO PREDICT A TORNADO AS LOSS OR NO-LOSS

Before we fit our linear regression model, we predicted whether or not property damage will result from a certain tornado. In our dataset, 37 percent of the tornados from 1950 to 2014 produced zero dollars of property loss. Our subsequent linear regression model will use logarithmic transformation of *Property Loss*. However, since the logarithm of zero is undefined we must first use logistic regression to predict the binary outcome of *Property Loss*. The linear regression will thus be used to predict the amount of property damage sustained, given we have predicted positive property loss from the logistic regression model. The linear regression will thus be fit to only a little more than half of the observations, since we will remove all the non-positive values of the outcome variable before fitting.

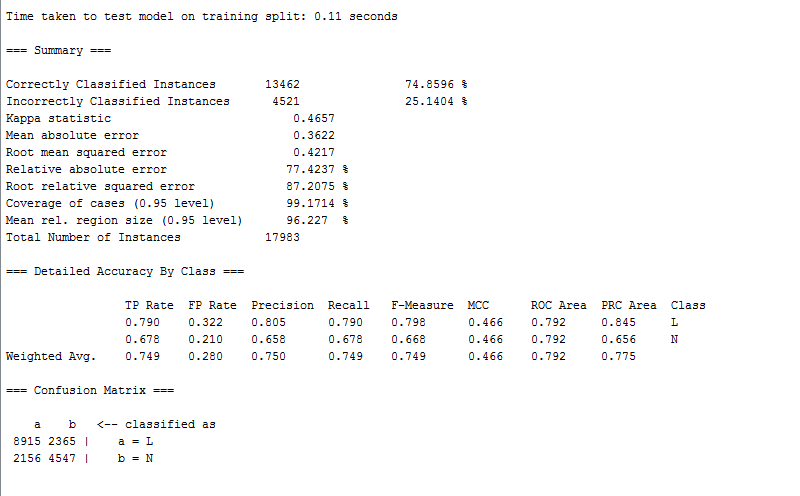
### Logistic Regression

The logistic regression is designed to classify the observations in the dataset into groups of whether or not the tornado had inflicted property damage. In our training dataset we converted our outcome variable of property damage, which is continuous, to being a binary outcome variable of either zero or one, where "1" indicates property loss and “0” indicates no property loss. We then fit this binary outcome variable to the following logistic regression model:

𝑃𝑟𝑜𝑝𝑒𝑟𝑡𝑦𝐿𝑜𝑠𝑠=1=𝑙𝑜𝑔𝑖𝑡 (𝛽0+𝛽1∗𝑆𝑡𝑜𝑟𝑚.𝑆𝑒𝑎𝑠𝑜𝑛+𝛽2∗𝐹𝑆𝑐𝑎𝑙𝑒+𝛽3∗𝐼𝑛𝑗𝑢𝑟𝑖𝑒𝑠+𝛽4∗𝐹𝑎𝑡𝑎𝑙𝑖𝑡𝑖𝑒𝑠+𝛽

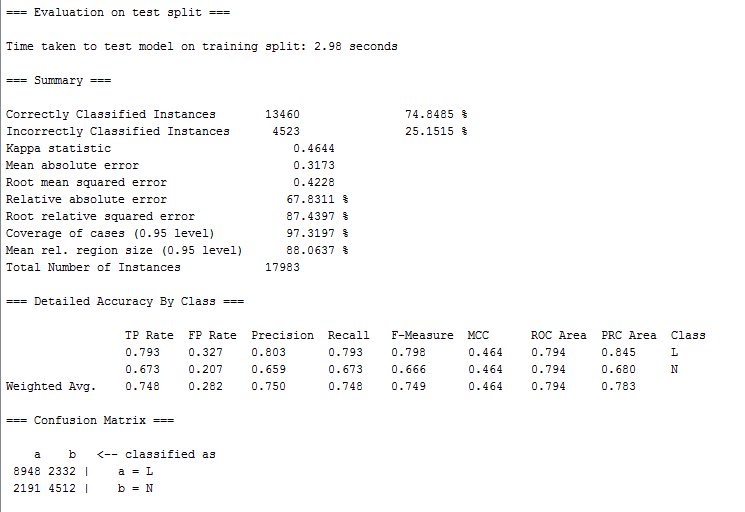
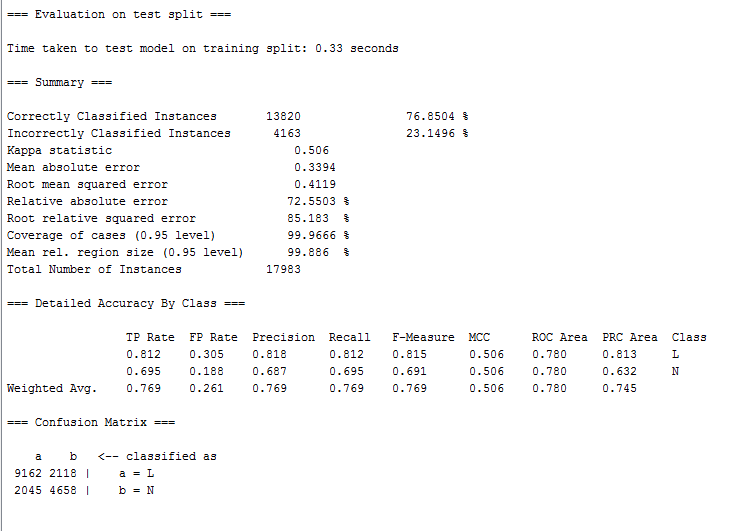
5∗𝑆𝑡𝑎𝑡𝑒.𝑆𝑒𝑣𝑒𝑟𝑖𝑡𝑦+𝛽6∗𝑊𝑖𝑑𝑡ℎ+𝛽7∗𝐿𝑒𝑛𝑔𝑡ℎ)

Where Storm Season and State Severity are factor variables. Before fitting the model we selected a random sample containing 30 percent of our dataset to be used for model validation. Fitting the full model, we produced the logistic regression output with the associated confusion matrix:

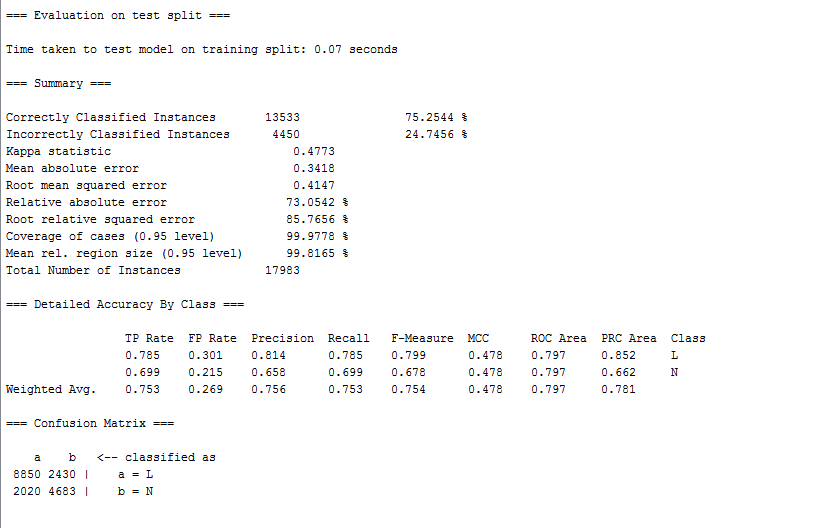
** Figure 10.**

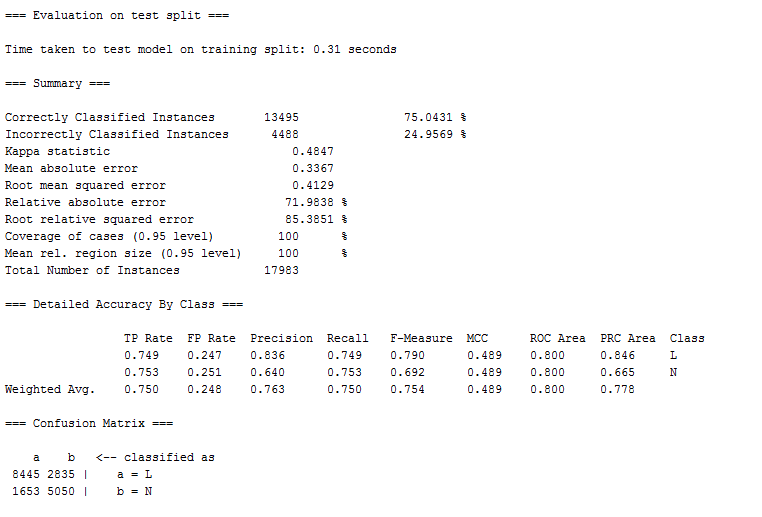
### Comparison with CART, Random Forest, Neural Networks and Adaboost

We compared the result from logistic regression with other algorithms and concluded that CART algorithm yields an error percentage of 23.1% which is a best fit for the classification model we are trying to achieve. Figures 11, 12 and 13 show the misclassification tables produced from CART, Random Forests, Neural Networks and Adaboost respectively.

** Figure 11.**

**Figure 12.**



**** **Figure 13.**

**Figure 14.**

**Figure 15.**

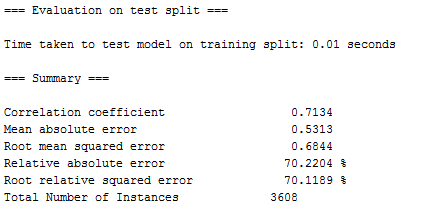
### Multiple Linear Regression

After classifying the observations in the dataset using logistic regression - predicting if there was property damage sustained - we fit our multiple linear regression model. The purpose of this regression is to estimate the dollar amount of property damage endured, given we have already predicted positive property damage in the logistic regression. To fit the model we must begin by removing the observations in our data set that have zero property damage. The relationship of *Property Loss* with most of the independent variables appears non-linear we took the logarithm of *Property Loss*. Since the logarithm of zero is undefined, the observations with zero property damage must be removed. However, since our linear regression model is attempting to estimate the amount of *Property Loss*, given positive amount of damage, the predictive significance of the model has not changed. *State Severity* and *Storm Season* are categorical non-ordinal variables, and *F-scale* is ordinal. The logarithmic transformation appears to correct much of the non-linearity between the predictors and the response. However the variables *Injuries* and *Fatalities* continued to appear to violate the linearity assumption. By taking the square-root of both these predictors the non-linearity is corrected. We begin by fitting the full model and then performing backwards elimination on the variables to find improvements in the error terms. The full model thus uses: *FScale, sqrtInjuies, sqrtFatalities, State Severity, Storm Season, Length,* and *Width.* Recall that *State Severity* is a factor variable with categories equal *to State.Severity1-4*, where *State.Severity1* is the base category. S*torm Season* is a binary variable indicating if the tornado took place during storm season. The full model when fit to the whole training data set gives the following output:

Model:

log*𝑃𝑟𝑜𝑝𝑒𝑟𝑡𝑦𝐿𝑜𝑠𝑠= 𝛽0+𝛽1*∗*𝑆𝑡𝑜𝑟𝑚.𝑆𝑒𝑎𝑠𝑜𝑛+𝛽2*∗*𝑆𝑡𝑎𝑡𝑒.𝑆𝑒𝑣𝑒𝑟𝑖𝑡𝑦+𝛽3*∗*𝐹𝑆𝑐𝑎𝑙𝑒+𝛽4*∗*𝑠𝑞𝑟𝑡𝐼𝑛𝑗𝑢𝑟𝑖𝑒𝑠+𝛽5*∗*𝑠𝑞𝑟𝑡𝐹𝑎𝑡𝑎𝑙𝑖𝑡𝑖𝑒𝑠*

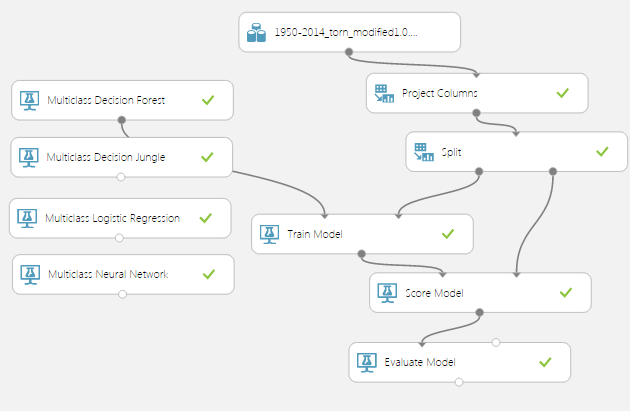
*+𝛽6*∗*𝑊𝑖𝑑𝑡*ℎ*+𝛽7*∗ *𝐿𝑒𝑛𝑔𝑡*ℎ

Figure 17 shows the output from the multiple linear regression. We took the square root to estimate the overall RMSE of the model. Doing this we produce an RMSE of 3.1 (million dollars). Looking carefully we can see that the high RMSE is due to a few outliers producing high error residuals in the predictions of some of the calculated MSE's, where few number of error terms are severely dragging the average upward while randomly sampling.

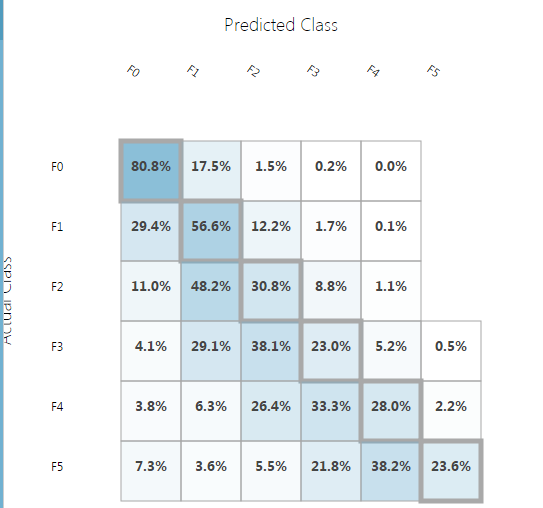
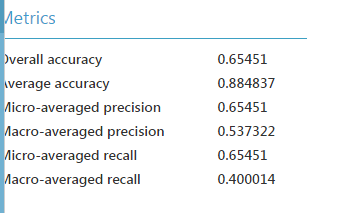
**Figure 16.**

## MULTICLASS CLASIFICATION BASED ON THE F-SCALE (TORNADO SEVERITY)

We have performed multiclass classification on the dataset to classify the records based on the intensity of tornadoes. We have used Microsoft Azure to perform this classification.

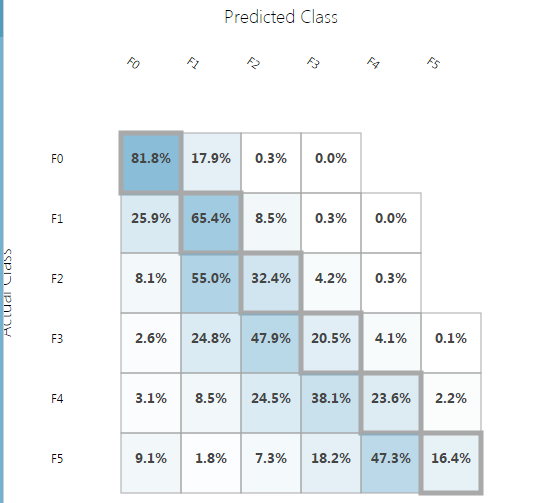


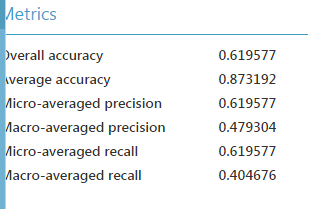
**Figure 17(a).**

Firstly we used Multiclass Decision Jungle to classify the data. Given below are the confusion matrix and accuracy metrics for the same:

**Figure 17(b).**

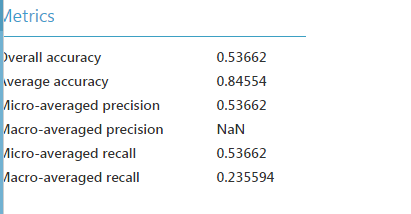
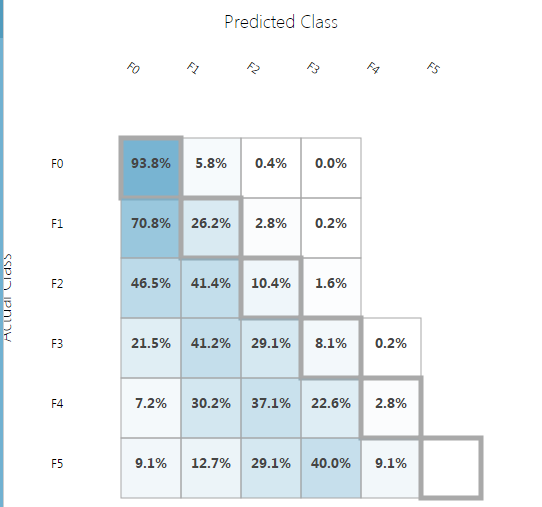
We then performed Multiclass Decision Forest on the dataset. Given below is the result of the analysis





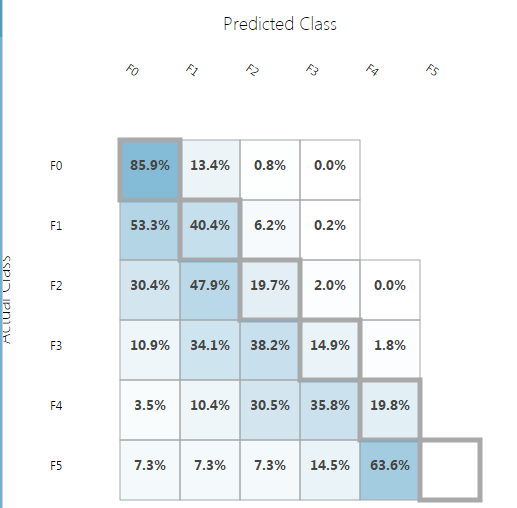
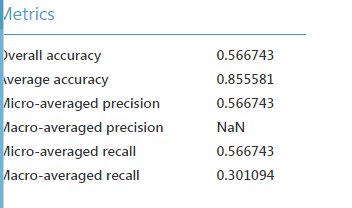
**Figure 17(c).**

We then performed Multiclass Logistic Regression, given below is the analysis of the same



**Figure 17(d).**

We also performed analysis using Multiclass Neural Networks



**Figure 17(e).**

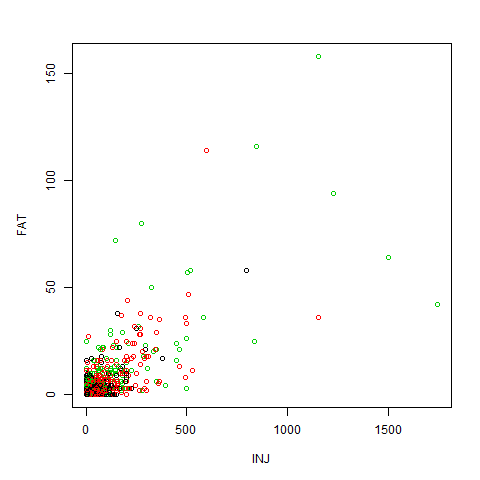
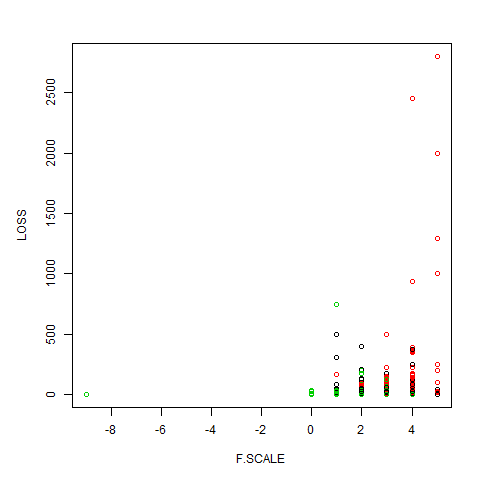
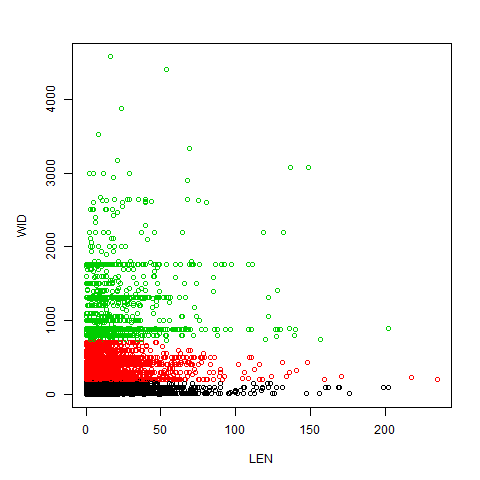
### Conclusion:

Looking at the Accuracy Metrics of the above performed analysis, we can conclude that we get the best accuracy with Multiclass Decision Jungle. The overall accuracy of Multiclass Decision Jungle is **0.6545** and the Average Accuracy is **0.8847**.

## CLUSTERING ALGORITHM

We studied the different features of tornado dataset using the clustering algorithm using K-means clustering. We clustered the dataset in 3 clusters and plot scatterplots to study the properties of tornado dataset. Figure 18 shows the scatter plots formed by clustering the dataset before the logarithmic transformation.

We can see that clear clusters are formed when comparing the length and width as two features of the dataset. Other properties are compared as well such as injuries and fatalities in which the clusters are more overlapping but still distinct.



**Figure 18.**

## CONCLUSION

We demonstrate a strategy for prediction of property losses due to tornado activity using NOAA tornado dataset. The modeling is done using a multiple linear regression model formulation where the loss sustainability is classified using classification model. The principal findings of the study include the following:

The CART model proved useful for initial classification of the observations, producing a classification rate of about 76% percent on the validation data. This proved successful in the ability to predict positive values of property damage.

Similarly, the predictors were also able to estimate the dollar amount of property loss with an error of $3.1 million while used in a multiple linear regression. We hope this predictive process provides commercial utility for approximating damage inflicted by tornados.