Utilizing Data Mining Techniques to Rethink Saffir-Simpson for Tropical System Landfalls

Neelesh Rastogi[1] and Sean Lukasiewicz[1]

***Abstract*— The Saffir-Simpson scale is the most well known system through which hurricanes are rated based on the intensity of sustained winds within the storm. It is a simple system placing the storms into six numbered categories and is often the first statement made when talking about the potential impact of a tropical storm. However, as more landfall events occur, we are seeing that the intensity of the wind is not always indicative of the potential impact a storm may have. Hurricane Michael in 2018 and Hurricane Sandy in 2012 both caused approximately $70 billion in damages to the continental US, but Michael was a Category 4 storm and Sandy was a Category 1 at their respective landfalls.**

I. METHODOLOGY

A. DATA COLLECTION

The bulk of the data used for this analysis came from open source R packages published by the National Oceanic and Atmospheric Administration (NOAA) and by Dr. Brooke Anderson of Colorado State Universitys (CSU) Department of Environmental and Radiological Health Sciences. Other attributes, such as moon phase and tidal data, came from various NOAA webpages and factors used in deriving and normalizing cost coming from the U.S Bureau of Labor Statistics website. Since most of the data is available through R packages, it was natural to use the language for the data collection and pre-processing.

The data was collected in a data frame utilizing the "storm_id" attribute from CSU's *hurricaneexposuredata* package as the index key. This attribute uses "name-year" format, and since all storms involved in this analysis occurred after the practice of naming storms started, this provides a unique key that is simple to use. We utilized the *separate* function in R to split the "DateTime" attribute in the *HURDAT* package into "Year", "Month" and "Day"; and then the *paste* function to concatenate the "Name" and "Year" attributes into the "storm_id." The *merge* function then allowed us to create our table for analysis. Tidal data and moon phase were entered manually from the NOAA Tides and Currents product webpage and the Almanac.com moon phase calendar respectively. Any data further missing from our data frame were manually entered from the official storm report PDFs available on the NOAA website. Microsoft Excel was used to manually input the data.

*This work was supported by St. John's University Data Innovation Lab.

[1]Under guidance of Dr. Christoforos Christoforou, who is with Faculty of Data Mining and Predictive Analytics, Computer Science, Mathematics and Science, St. John's University, 8000 Utopia Pkwy, Queens, NY - 11439 christoc@stjohns.edu

B. PRE-PROCESSING

To account for the natural variations within a storms track, we opted to derive an effective land speed and effective angle of impact to land. This accounts for any sudden acceleration or change in direction that may occur once a system starts to interact with land. Since the NOAA monitors a system in six-hour intervals, we looked at the twelve-hour window around the moment of landfall to create the displacement vector for this time frame. This displacement vector can be considered to be average velocity of the system from which we can take the effective land speed and effective direction of travel.

$$\text{v}=\frac{\sqrt{(62.17*\Delta\text{Latitude})^2+(69.17*\cos(Latitude_2)*\Delta\text{Longitude})^2}}{12} \quad (1)$$

$$Angle=\arctan\left(\frac{(69.17*\cos(Latitude_2)*\Delta Longitude)}{(69.17*\Delta Latitude)}\right) \quad (2)$$

The lunar cycle is approximately 28 days long; therefore, the moon phase can be simplified to a discrete scale of 0 to 27, where 0 is a full moon, 14 is a new moon, and 7 and 21 are half-moons. The effects on tidal ranges by moon phase are maximized during the full and new moon phases and minimized on the half-moon. This means that for every lunar cycle the tidal effect goes through two compete cycles. To handle this, we translated the data by subtracting 14 and taking the absolute value. We repeated this process again, this time subtracting by 7, to create a linear scale where 0 is the minimal tidal range and 7 is the maximum range.

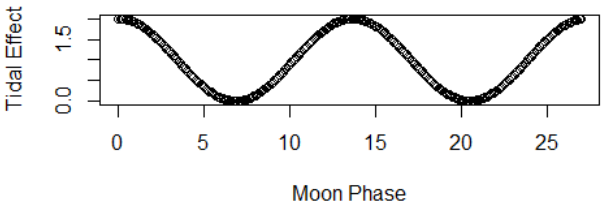

Fig. 1. Effect of moon phase on tidal range

The cost of the landfall events were normalized by using yearly consumer price index (CPI) values and state standard of living indices. The cost of each storm *j* was multiplied by the ratio of the 2018 CPI to the CPI of the year the storm occurred to account for inflation then divided by the average

cost of living index value of the states affected to account for location.

$$Cost_{Norm} = \frac{\frac{CPI_{2018}}{CPI} * Cost_j}{(\Sigma(Costof Living))/n} \quad (3)$$

Our scale structure came from a visual analysis of the plot of the normalized cost of each landfall event ordered from low to high. This revealed a curve that appeared exponential in nature, so we opted for a base-ten logarithmic scale similar to the Richter magnitude scale, where each point is ten times greater than the previous one. This also produced the most even distribution of events along the scale.

C. EXPLORATORY DATA ANALYSIS

Once the data was gathered, and pre-processed, a raw consolidated data file was generated and was further fed into a data frame (df) using pandas. With later df, implicit data values such as effective land speed, effective angle of impact, normalized cost and moon phase were derived and calculated. These calculations were carried out based on equations as mentioned in prior section. Our final Data frame consisted of the following values as shown in Table I. These values were then, utilized to find common patterns and observations.

TABLE I

ALL FEATURES OF OUR FINAL DATA FRAME.

Feature Name	Feature Type
storm_names	115 non-null object
eff_land_sp	115 non-null float64
direct	115 non-null int64
angled	115 non-null int64
cross	115 non-null int64
press_mbars	115 non-null int64
max_sust_winds_kts	115 non-null int64
storm_surge	115 non-null float64
storm_tide	115 non-null float64
moon_phase	115 non-null int64
low_neap	115 non-null int64
high_neap	115 non-null int64
high_ebb	115 non-null int64
high_tide_line	115 non-null float64
low_tide_line	115 non-null float64
norm_cost	115 non-null float64
cost_category	115 non-null category

Our data set is comprised of all landfall events along the North American Atlantic and Gulf Coast basins between the years of 1985 and 2017. The location of these events can be seen in Figure 2

On an initial analysis we found the following observations:

- Our current curated data-set currently consists of in total 115 storms, out of which we observed our data set having 48 storms which caused minimal damage and around 34 storms which caused severe to catastrophic level damage. Table II below shows all storm counts for all levels of damage caused during landfall.
- On grouping our data-set by year of storm events, we observed Most storms occurred in the year of 2004 followed by year 1998 and 2005. The frequency chart is shown below in Figure 3

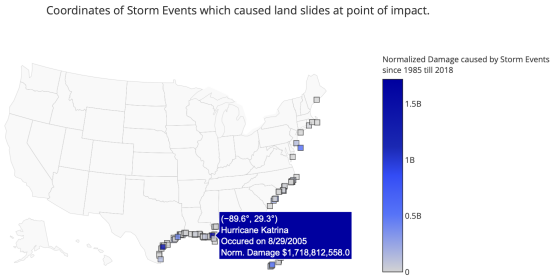

Fig. 2. All Storm events plotted via coordinates since 1985 - 2017.

TABLE II

COUNT OF ALL STORMS BASED ON THEIR DAMAGE.

minimal	48
low	6
moderate	8
high	19
severe	17
catastrophic	17

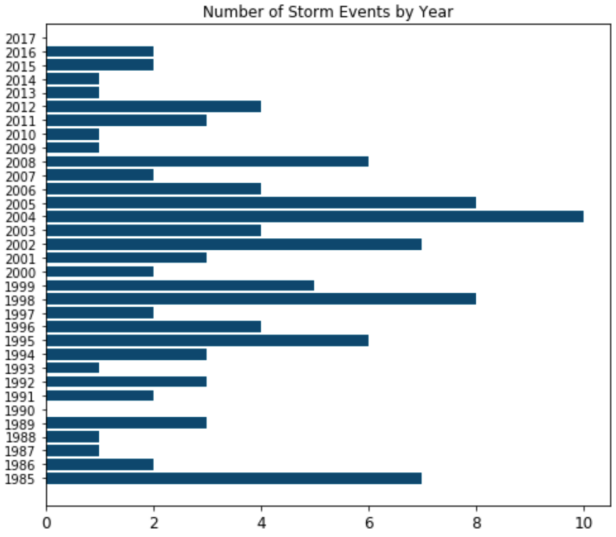

Fig. 3. Number of Storm Events by Year.

D. CLASSIFICATION

Once we identified our initial observations within exploratory data analysis, we then conducted supervised machine learning processes to identify significant traits within storm systems to identify top features for predicting potential storm damage and potential cost incurred after landfall.

We initially ran an ensemble of classification models like XGBoost, Support Vector Machines, Random Forest, Decision Tree, K-Nearest Neighbor and Logistic Regression on our curated data set, which was randomly split in 70-30 ratio with a parameter of `(random_state = 40)`, and was further looped for a total of 30 iterations, giving us 30 unique accuracies and mean squared errors for all our 6 models. The similar procedure was replicated over Saffir-Simpson scale which solely relied of maximum sustained wind speed as its training variable. Below table III shows

TABLE III

COMPARISON OF ALL ACCURACIES AND MSE'S MEANS AND STD. DEV. ACHIEVED.

MSE Comparison					**Accuracy Comparison**			
	Our Dataset		**Saffir Simpson Model**		**Our Dataset**		**Saffir Simpson Model**	
Models	Mean	Std dev	Mean	Std dev	Mean	Std dev	Mean	Std dev
Decision Trees	4.21E+00	5.4E-01	4.605263	1.81E-15	0.514431	3.52E-02	0.447368	1.69E-16
KNN	6.55E+00	9.0E-16	6.684211	9.03E-16	0.394737	1.13E-16	0.342105	1.69E-16
Logistic Regression	4.68E+00	9.0E-16	4.394737	1.81E-15	0.421053	2.82E-16	0.421053	2.82E-16
Random Forest	4.26E+00	6.8E-01	5.312281	1.12E+00	0.489813	6.18E-02	0.428947	3.73E-02
SVM	4.84E+00	1.8E-15	6.631579	1.81E-15	0.447368	1.69E-16	0.421053	2.82E-16
XGBoost	4.55E+00	9.0E-16	4.973684	1.81E-15	0.578947	1.13E-16	0.447368	1.69E-16

the mean and standard deviation of all accuracies and mean squared errors achieved through this process.

The above mentioned classification algorithms took 80 historic storm events as its input and was further tested on the other 35 tropical system landfalls. The results of all models were then compared to Saffir-Simpsons output with all of our data models showing improvement over Saffir-Simpson. Our best individual model was found to be Decision Tree Classifier with an mean accuracy of 51.44 percent% compared to Saffir Simpson Scale, which was 44.74% accuracy.

E. FEATURE OPTIMIZATION

Based on our findings we then ran a Recursive Feature Elimination process along with a 5 fold cross validation, over our selected model of Decision Tree Classifier.

Recursive Feature Elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the models coef_ or feature_importances_ attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and cl-linearity that may exist in the model.

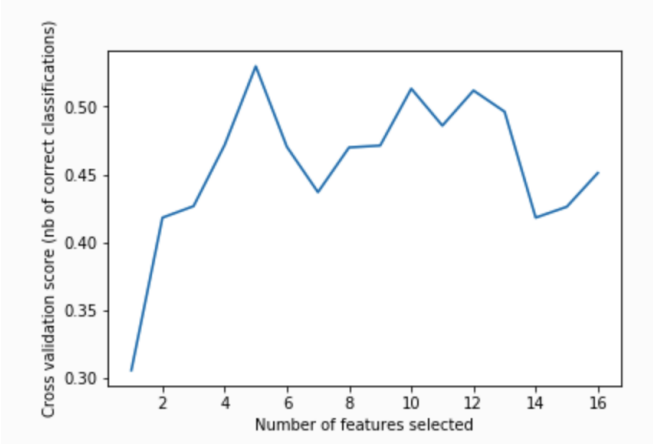

Fig. 4. Optimal Number of Features.

As, RFE requires a specified number of features to keep, cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

Based on the findings, as shown in Figure 4f 5 features were found to be the optimal number.

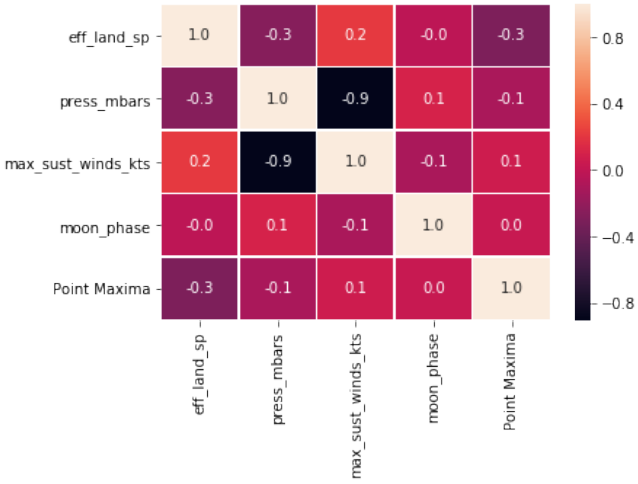

Fig. 5. Correlation Values between each Feature pairs.

These 5 optimal features were:, Point_Maxima, moon_phase, max_sust_winds_kts, press_mbars and eff_land_sp. A pair grid analysis as shown in Figure 6 and 5, shows us a linear relationship between moon phase and other features. Also most of these feature points seemed to be clustered closely as well, when plotted in pairs.

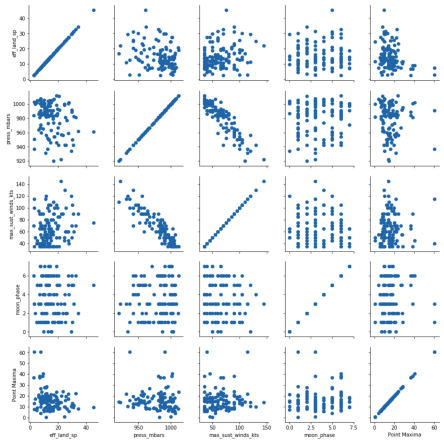

Fig. 6. Pair Grid between each Feature pairs.

F. UNDERSTANDING DECISION TREE MODEL

We used Decision Trees (DTs) as our optimal predictive model as they are a non-parametric supervised learning

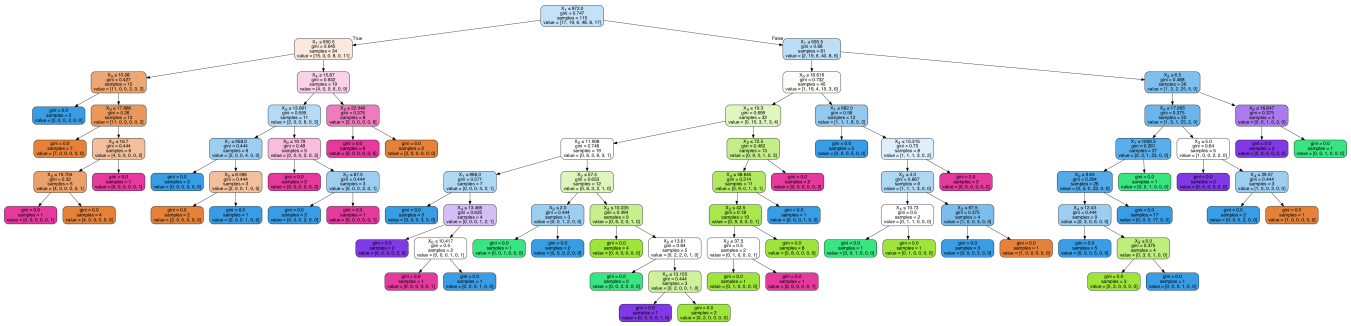

Fig. 7. Decision Tree Visualization using Pydotplus package.

method used for classification. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model gets. For our model our optimal parameters for Decision tree was maximum_depth = 10.

On plotting the decision tree as shown in Figure 7 We understand on how our model was able to classify values for storm events to be catastrophic or minimal.

The value row in each node tells us how many of the observations that were sorted into that node fall into each of our 5 categories. We can see that our feature X4, which is the moon-phase, was able to completely distinguish one storm event type of catastrophic from the rest.

The biggest drawback to decision trees is that the split it makes at each node gets optimized for the data -set it is fit to. This splitting process will rarely generalize well to other data. However, in future work we can generate huge numbers of these decision trees, tuned in slightly different ways, and combine their predictions to create some of our best models for predicting storm damages today.

II. CONCLUSION

The initial goal of our study was to create a new scale and algorithm that could effectively predict the impact of a tropical weather system making landfall in the continental United States. We were able to show improvement over the Saffir-Simpson scale but were unable to reach our intended goal of at least 75% accuracy. Further analysis will focus on tweaking some of the data parameters and finding ways to make the input data more comprehensive.

III. REFERENCES

[1]B. Anderson, Hurricaneexposuredata. https://github.com/geanders/hurricaneexposure

[2]D. Gershgorn, Artificial intelligence is great at predicting the size of hurricanes, but humans still need to figure out their impact, Quartz. [Online]. Available: https://qz.com/1072215/artificial-intelligence-is-great-at-predicting-the-size-of-hurricanes-but-humans-still-need-to-figure-out-their-impact/. [Accessed: 28-Feb-2019].

[3]S. Giffard-Roisin, M. Yang, G. Charpiat, B. Kgl, and C. Monteleoni, Fused Deep Learning for Hurricane Track Forecast from Reanalysis Data, in Climate Informatics Workshop Proceedings 2018, Boulder, United States, 2018.

[4]T. Loridan, R. P. Crompton, and E. Dubossarsky, A Machine Learning Approach to Modeling Tropical Cyclone Wind Field Uncertainty, Mon. Wea. Rev., vol. 145, no. 8, pp. 32033221, May 2017.

[5]A. Sujithkumar, A. W. King, M. Kovilakam, and D. Graves, Predicting the trajectories and intensities of hurricanes by applying machine learning techniques, AGU Fall Meeting Abstracts, vol. 31, Dec. 2017.

[6]Predicting Hurricane Damage with Machine Learning, Datanami, 22-May-2017. [Online]. Available: https://www.datanami.com/2017/05/22/predicting-hurricane-damage-machine-learning/. [Accessed: 28-Feb-2019].

[7]Detecting Storm Intensity from Satellite Imagery Using Machine Learning, GIS Lounge, 11-Sep-2018. [Online]. Available: https://www.gislounge.com/detecting-storm-intensity-satellite-imagery-using-machine-learning/. [Accessed: 28-Feb-2019].

[8]Consumer Price Index, US Bureau of Labor Statistics.

[9]Current Hurricane Data Sets. [Online]. Available: http://www.aoml.noaa.gov/hrd/data_sub/hurr.html. [Accessed: 28-Feb-2019].

[10]GOES-East - Latest Full Disk Images - NOAA / NESDIS / STAR. [Online]. Available: https://www.star.nesdis.noaa.gov/GOES/fulldisk.php?sat=G16. [Accessed: 28-Feb-2019].

[11]Hurricane Point Rainfall Maxima.

[12]Methodology, Deep Learning-based Hurricane Intensity Estimator. [Online]. Available: www.domain.org. [Accessed: 28-Feb-2019].

[13]Past Atlantic Storm Tracks - Data.gov. [Online]. Available: https://catalog.data.gov/dataset/past-atlantic-storm-trackstopic=disasters-legacy_navigation. [Accessed: 28-Feb-2019].

[14]Storm Surge — U.S. Climate Resilience Toolkit. [Online]. Available: https://toolkit.climate.gov/topics/coastal/storm-surge. [Accessed: 28-Feb-2019].