Predicting Hurricane Survival: A Classification Analysis of WMO Data

Team 12
Connor Park, Data Science, Undergraduate (cdpark@wisc.edu)
Saarthak Aggarwal, Computer Science, Undergraduate (saggarwal29@wisc.edu)
Jon Starfeldt, Atmospheric and Oceanic Sciences, Undergraduate (jstarfeldt@wisc.edu)

Repository Link: Hurricane Survival
Dataset Link: https://www.kaggle.com/datasets/rahulsathyajit/wmo-hurricane-survival-dataset

A. Abstract

Hurricanes pose a substantial threat to human life, prompting our research to identify factors influencing survival during such disasters. This study investigates the attributes that can contribute to an individual's likelihood of survival during a hurricane. Using data from the World Meteorological Organization (WMO), our research uses machine learning models, XGBoost and a Multilayer Perceptron (MLP) neural network to identify these attributes. After fine-tuning our models to improve performance, results from the XGBoost classifier and MLP models reveal sub-par accuracies, indicating the need for a more complex model architecture.

B. Introduction

It is estimated that 10,000 people die each year worldwide due to hurricanes and tropical storms [1], with the majority of human deaths being caused by flooding. We want to find out what characteristics of a person make them more likely to survive a hurricane.

Hurricane Irma hit Florida as a Category 4 storm the morning of Sept. 10, 2017 with maximum wind speeds of 115 kts, ripping off roofs, flooding coastal cities, and knocking out power for many, totaling approximately fifty billion dollars in damage, making it the fifth-costliest hurricane to affect the United States to its day. A full report by the National Hurricane Center provides all the relevant information on the storm [2]. From this, we learn there were ninety-three deaths caused by Irma, ten direct (from the storm itself) and eighty-three indirect (from the effects of the storm). The increased occurrence of natural disasters such as this one have been of particular concern for the United Nations. The World Meteorological Organization (a specialized agency of the UN) has been collecting data about all the individuals that are living in and around Hurricane and Cyclone prone areas [3]. In the aftermath of Irma, the WMO wanted to find a pattern or a relation between the attributes of a person and whether an individual will survive a hurricane in the future or not.

Because this is a classification problem, we will be using a multilayer perceptron neural network. Multilayer perceptrons work like regression models in that input values are transformed via a series of multipliers (called weights) and nonlinear functions (called activations) that help

the input data best represent the desired output. Because the dataset we are using has many variables, we will be formulating a decision rule that chooses certain variables to input to our neural network and others to leave out.

For related works, we found another project that was done on Kaggle using the same dataset. The author focused on cleaning the data, encoding categorical variables, and using various machine learning models like logistic regression and a random forest classifier to predict survival. Their logistic regression showed a 50.6% accuracy and their random forest classifier achieved a 50.7% accuracy. For our project, we aim to refine the model selection and feature engineering process to surpass this accuracy.

C. Data

This data set, primarily focused on assessing the survivability of individuals during hurricanes and tropical storms, is a comprehensive collection of various personal attributes. It consists of 20 different variables, ranging from basic demographic details like Date of Birth (DOB), Marital Status (M_STATUS), and Gender (GENDER), to more specific preferences such as Favorite TV Show (FAV_TV), Preferred brand of car (PREF_CAR), and Favorite Cuisine (FAV_CUIS). Each data sample in the set represents an individual and includes their details across all these variables. The data set is extensive, covering a diverse range of factors that might influence survivability in extreme weather events. The dataset has samples for 5000 people that went through hurricane events, with 49% of the people being classified as not surviving and the 51% of the people labeled as surviving. A partial example of a sample is given below, with there being multiple more columns present in a full observation.

ID	DOB	M_STATUS	SALARY	EDU_DATA	EMP_DATA	REL_ORIEN	FAV_TV	PREF_CAR	GENDER
3357	4/5/1994	Married	300k-500k	Post-Graduate	Unemployed	Agnostic	Friends	Ford	Female

C.1 Data Preprocessing

We began by looking through the list of variables in the dataset and throwing out variables that seemingly have no effect on a person's likelihood to survive a hurricane. We did this because many of the variables in the dataset are categorical and would require one-hot encoding, which would make the dataset far too messy due to the large number of variables. We also threw out variables that were not well documented in the Kaggle database, namely "Endurance Level", which came with no units. The full list of variables that were initially kept and removed from the dataset are listed in Table 1. Categorical variables that were kept were one-hot encoded and all the remaining data was collected in a Pandas dataframe.

We then split the dataset into training/testing/validation datasets with a 90/10 split, shuffling observations randomly. The partitioned data was used in the implementation of our models.

Date of Birth	Marital Status	Annual Salary (in	Education

(MM/DD/YYYY)	(Married/Unmarried /Divorced)	ranges of USD)	(Uneducated/High-School/ Graduate/Post-Graduate)
Employment Details (Employed/Self-Employed/ Unemployed)	Religious Orientation (Agnostic/Atheist/ Believer)	Favorite TV Show	Preferred brand of car
Gender (Male/Female/Other)	Favorite Cuisine	Favorite Genre of Music	Endurance Level
Favorite Sport	Favorite Color	Main News Source	Distance from Coast (km)
Monthly Travel	Favorite Genre of Movie	Favorite Subject in School	Preferred Alcohol
Favorite Superhero	Label: x (will survive) or y (will not survive)		

Table 1: Variables Included in the WMO Hurricane Survival Dataset. Variables in green are initially kept in the dataset while variables in red are initially cut from the dataset before further preprocessing.

C.2 Feature Selection

For data feature selection, we used vector classification with sklearn's LinearSVC function, penalizing with L1 norm to make the coefficient of insignificant variables zero. The regularization parameter was set to 0.02 to ensure we were only using the variables that were most important in predicting hurricane survival. Because we were dealing with many categorical variables, independent one-hot encoded vectors had to be taken as separate from their parent variable for feature selection purposes. To deal with this, variables that had any of their categories left after the execution of the LinearSVC were kept in the dataset and variables that did not have half of their categories left were dropped. The variables that were ultimately kept are listed below.

- Date of Birth
- Marital Status
- Education
- Gender
- Employment Details
- Main News Source
- Distance from Coast

D. Methods

D.1 XGBoost

In building the XGBoost model, a hyperparameter tuning process is employed to optimize the performance of an XGBoost classifier model. The primary focus is on fine-tuning two key hyperparameters: the learning rate and the number of estimators. The learning rate, which controls the step size in the gradient descent optimization process, is incrementally adjusted from a starting point of 0.003, increasing in steps of 0.002 up to a threshold of just below 0.1. This gradual adjustment allows for a thorough exploration of how different learning rates affect the model's performance. Simultaneously, within each iteration of the learning rate adjustment, a nested loop tests a range of values for the number of estimators, which dictates the number of boosting stages the model undergoes. This range spans from 25 to 150 in increments of 5. Each combination of learning rate and number of estimators is used to train the XGBoost model, and its accuracy is assessed against a test set. The approach is methodical and exhaustive, systematically iterating through a grid of possible hyperparameter combinations.

D.2 Multilayer Perceptron

Our multilayer perceptron (MLP) model has three fully connected layers. The input layer has eight neurons, the hidden layer has four and the output layer has two. The first two layers have a ReLU activation function and the last layer has a softmax activation function to output the probability of the specified person to survive or not survive a hurricane. The model is trained with categorical cross entropy loss and stochastic gradient descent optimization. In an attempt to further improve our accuracy, we trained the model with batch sizes between 5 and 25, fine-tuning this hyperparameter to get the best results. We also implemented adaptive learning rates, which improves the model by quickly progressing to an optima at first before more carefully taking steps towards optimization closer to convergence time.

E. Results and Discussion

The key to our classification analysis lies in the use of two machine learning models: XGBoost and a Multilayer Perceptron (MLP) neural network. Below, we present a detailed comparison of their performance metrics.

XGBoost Classifier:

- Parameters:
    - Learning Rate: 0.005
    - Number of Decision Trees: 150
- Performance:
    - Accuracy: 59.27%
    - The XGBoost model, with its fine-tuned parameters, achieved a moderate level of accuracy, indicating a potential for further improvement. It's worth noting that the learning rate was deliberately kept low to mitigate overfitting risks, which could

have affected the model's ability to generalize. A snippet of the decision tree for the XGBoost model is pictured below. Each branch splits people into two categories based on a certain variable. For example, the top node is “AGE<44”, splitting people based on if they are younger than 44 or not. This also shows that a person’s age has the most predictive power on their hurricane survival. Each leaf has a corresponding score, positive showing that their likelihood of surviving a hurricane is greater with those characteristics, and negative showing their likelihood of survival to be lesser.

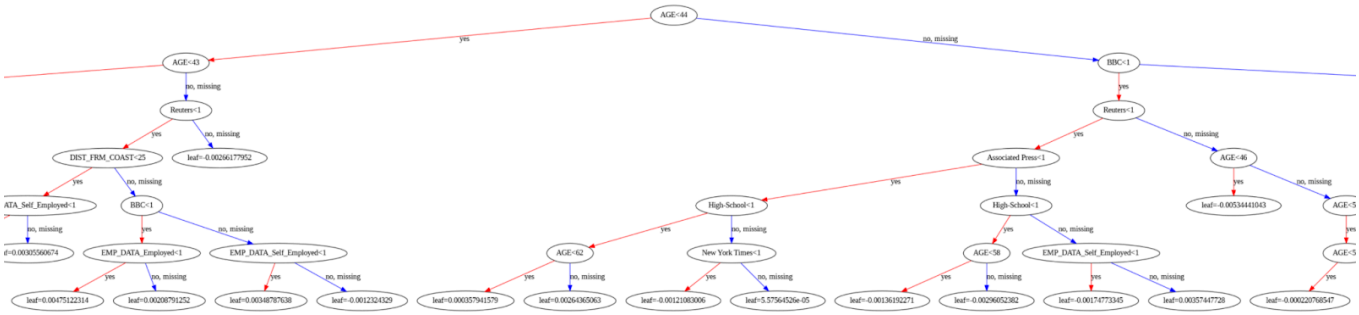

Figure 1: Visualization of the XGBoost Decision Tree

Multilayer Perceptron Neural Network:

- Parameters:
  - Layers: 3 (Fully Connected)
  - Shape: 8-4-2
  - Activation Functions: ReLU (Hidden Layers), Softmax (Output Layer)
  - Loss Function: Categorical Cross-Entropy
  - Optimizer: Stochastic Gradient Descent (SGD)
  - Epochs: 50
  - Batch Size: 5
- Performance:
  - Accuracy: 50.81%
  - The MLP did not outperform the XGBoost model. Given the basic architecture of the MLP, there is potential for improvement through a more complex network architecture. The ROC curve in Figure 3 shows the MLP to not have much dynamic predictive power, as its slope is invariant for much of the graph. A larger dataset or deeper neural network should be able to improve upon this issue in the future.

```
Model: "sequential_10"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_30 (Dense)            (None, 8)                 96

 dense_31 (Dense)            (None, 4)                 36

 dense_32 (Dense)            (None, 2)                 10

=================================================================
Total params: 142 (568.00 Byte)
Trainable params: 142 (568.00 Byte)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
None
```

Figure 2: MLP Neural Network Model Architecture

Quality of Results:

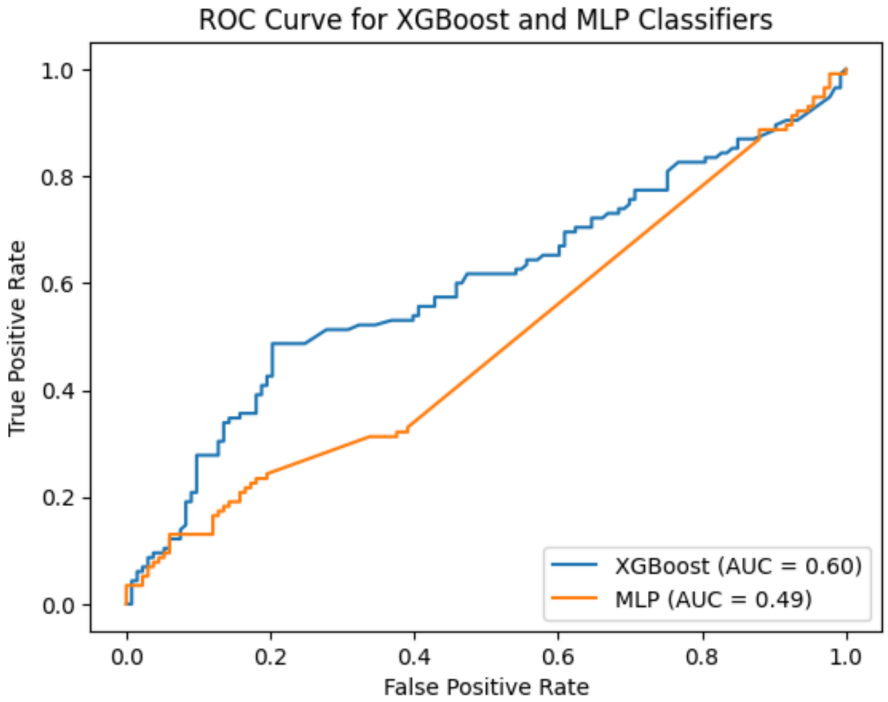

Figure 3: ROC Curve for XGBoost and MLP Classifiers

- The Receiver Operating Characteristic (ROC) curve for both classifiers was plotted to assess their diagnostic ability. The Area Under the Curve (AUC) scores for XGBoost and MLP were 0.60 and 0.49, respectively. This suggests that while the models can predict survival, there is significant room for improvement.

- The current model performances are somewhat below expectations. The quality of the results, gauged by the accuracy and AUC, suggest that neither model is currently extracting the full predictive signal from the feature set.

The results obtained from the analysis of the XGBoost classifier and the Multilayer Perceptron Neural Network reveal several insights into the prediction of hurricane survival and areas for further research.

The XGBoost model demonstrated a moderate accuracy of 59.27%. This is closer to our goal of improving upon the accuracy of the former work done on this dataset, though still not close to what we would hope for an operational model. On the other hand, the MLP yielded an accuracy of 50.81%. Since the accuracy level was akin to chance, it demonstrates the room for improvement. The overall quality of the results highlights the need for further refinement in model performance. This can be done through more sophisticated techniques such as ensemble methods and more complex neural network structures. It would also be beneficial to individually see each feature's impact on survival. If we can quantify the impact, we can further distill the explanatory variables in the models such that a better accuracy can be achieved.

F. References:

[1] Bergman, J. (2016). *Hurricane damage*. Hurricane Damage - Windows to the Universe. https://www.windows2universe.org/earth/Atmosphere/hurricane/damage.html

[2] Cangiolosi, J. P., Latto, A. S., & Berg, R. (2018, March 9). *Hurricane Irma*. National Hurricane Center Tropical Cyclone Report. https://www.nhc.noaa.gov/data/tcr/AL112017_Irma.pdf

[3] Sathyajit, R. (2017, December 19). *WMO hurricane survival dataset*. Kaggle. https://www.kaggle.com/datasets/rahulsathyajit/wmo-hurricane-survival-dataset/data

[4] Umansky, E. (2021, April 30). Hurricane Survival Prediction. Kaggle. https://www.kaggle.com/code/evgenyumansky/hurricane-survival-prediction#Modeling-and-Evaluation

[5] Tanner, G. (2021, October 20). Metrics Explained. ML Explained. https://ml-explained.com/blog/metrics-explained