Author: Justine Huynh
Date: 2021/05/04

Death By Hurricane Irma

Executive Summary

This project focuses on the factors that determine whether a resident lives or dies from Hurricane Irma and whether it is possible to accurately classify whether someone's survival. Because Hurricane Irma wreaked so much havoc and killed many individuals, policymakers are trying to identify residents who are at high-risk of dying from the storm when it returns in the future.

I used Linear Discriminant Analysis, Decision Tree, Random Forest, and XGBoost to answer my research questions. After reading in hurricane data that was publicly available from Kaggle, I preprocessed the data: implemented ordering to any variables who had inherent ordering, created dummy variables, scaled some numeric data. Afterwards, I tested each model individually and recorded their performance metrics. Yet, since I wanted to determine which model was the best as a whole, I combined all the models and ran a seamless pipeline with specified hyperparameters for each model to smoothly determine which model was the best at its given hyperparameters. In the end, I discovered that all the models (except for the individual xgboost) poorly predicted whether someone died.

Background

Every year, hurricanes, cyclones, and typhoons rage through countries furiously. It is estimated that a total of 10000 people die each year from these monstrous storms. Tumultuous storms can cause power outages, destroy electric towers, flood the roads--damages that can take an enormous amount of money and time to rectify. One such deadly storm is Hurricane Irma. Hurricane Irma originated from a tropical wave on the West African coast. It slowly inched its way east, increasing wind speeds until it became a category 5 storm on September 5, with wind speeds above 157 mph. Hurricane Irma laid destruction on Antigua, Barbuda, Anguilla, the US and British Virgin Islands, and more neighboring islands until it finally laid waste on the US. On one deadly day, in the morning of September 10, 2017, Hurricane Irma raged through the East Coast of US, specifically in the states of Florida, Georgia, South Carolina, ripping away roofs, flooding the streets, and tearing trees apart (Klinger, 2018). After Hurricane Irma ravaged the East Coast, the World Meteorological Organization (WMO) wanted to find a pattern or a cycle in determining which individuals would survive or die in any hurricane or cyclone in the future given the individuals' characteristics (Sathyajit, 2017).

Because Hurricane Irma was extremely violent, there have been some studies done on this storm's impact. For instance, there was a similar study in 2018 studying the storm's environmental impacts. The research study focused on ecological factors such as sediment deposits, mudflat core tops, and vegetation reduction (Wingard, 2019). But this time, this data science project is focusing on the human side of the storm. Specifically, this project aims to discover which factors are the most important in determining whether an individual will survive

in the face of the storm and use those important variables to predict whether a certain individual will survive or die.

Determining the important factors of who will survive or die in a hurricane is important because it focuses on what aspects of people's lives have to be improved to increase their chances of survival. For instance, infrastructure could be innovated to better protect individuals against the storm, especially for those who live near the ocean since they are hit with the first wave of the storm. Or, if the majority of the people who die are poorer people, organizations can work to offer financial support to those who are struggling. Learning these factors are rather helpful for discovering any way to help the residents, especially the ones hardest hit, of hurricane-torn stricken regions in the future when another furious storm hits. Hurricanes may never disappear, but at least the country can learn how to better help and prepare its residents to survive.

Data

Datasets that have data about the people of a certain region who survived or died in the face of a storm are particularly helpful. Kaggle fortunately had one such dataset: WMO Hurricane Survival Dataset at this link (https://www.kaggle.com/rahulsathyajit/wmo-hurricane-survival-dataset). Each row corresponded to one citizen with characteristics such as how many miles away from the coast, gender, marital status, favorite song genre, education, and many more; each observation was also labelled whether the human flourished or perished after Hurricane Irma hit on September 10, 2017, notably in the column Class. This column will be the dependent variable I shall be predicting. The dataset has a total of 5021 rows and 24 columns, so fortunately the resulting training data and test data sizes will be sufficiently large enough to prevent underfitting.

The type of data that is useful for this project includes demographic variables such as: wealth, distance from the coast, whether they were employed, etc. Although the data did provide some pertinent demographic variables, most of the variables were seemingly irrelevant. Furthermore, interestingly, the data had two similar-sounding column names: Dist_Coast and DIST_FRM_COAST, 2 columns that actually did not have similar values. Since the website did not adequately distinguish between the two columns, I understood that Dist_Coast meant "Length of the beach/coast in kilometers" while DIST_FRM_COAST referred to how many kilometers away someone lived from the coast. Furthermore, although Dist_Coast was a numeric variable, DIST_FRM_COAST was a categorical variable with values in ranges such as: 0-100 km, 100-300 km, 300-500 km, etc. A numerical value measuring the actual distance from the coast would have been more helpful and simpler.

This dataset also lacks information on house characteristics and whether these houses collapsed or not. Some houses in Florida remained intact even after the deadly storm hit. Such sturdy houses were "concrete home[s] elevated on stilts" and "must be elevated above the flood plain to allow storm surge, which is the deadliest part of a hurricane, to pass underneath living

spaces" (Ovalle, 2017). Modelling future homes after these sturdy houses can be key to decrease future damage and save money on rebuilding houses and infrastructure.

Most of the data consists of categorical variables, with some numeric variables. Of these categorical variables, as stated before, it appears that there exists some irrelevant variables such as PREF_CAR (preferred car), GEN_MOVIES (favorite genre of movies), and FAV_SUPERHERO (favorite superhero), FAV_COLOR (favorite color), FAV_CUIS (favorite cuisine), personal preferences that have no influence on survival rate in Hurricane Irma's wrath. Other irrelevant variables include ID and DOB. However, I used DOB to calculate someone's age by comparing each person's birthdate to the date the study came out. This way, we are not completely eradicating the DOB variable--merely altering it to something more useful. At first glance, the age variable looked like it lacked data from people in their mid-30s or people in their mid-50s. However, each of the histogram's bins covered only about 2 years, an interval so small that it can be ignored. It is not a big deal if the data had few 35-year-olds but had a lot of 37-year-olds, as the age difference between them is not significant. Most of the data distributions are not too skewed.

For some of the categorical variables who do not have a natural ordering, I shall convert them to dummy variables. These variables, which encode residents' favorite things, include: favorite color, favorite cuisine, favorite alcohol, and others.

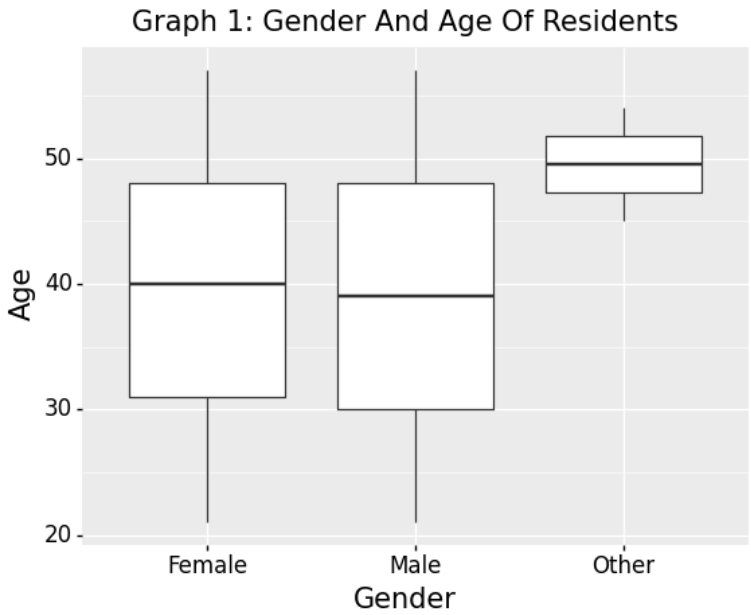
Table 1: Descriptive Statistics For Numeric Variables

Statistic	Coast Length (KM)	Age	Survived?
Count	4951.0	4951.0	4951.0
Mean	761.701	39.417	0.513
Std. Dev.	424.024	10.327	0.5
Min	40.0	21.0	0.0
25%	388.0	31.0	0.0
50%	756.0	39.0	1.0
75%	1126.0	48.0	1.0
Max	1500.0	57.0	1.0

Above is the descriptive summary statistics for numeric variables. Note that the mean of the survived column is 0.458, close to 0.5, which indicates a relatively equal distribution between the people who survived and those who died. Note that the age variable consists of values between 21 and 57. The dataset does not include residents who were seniors or minors.

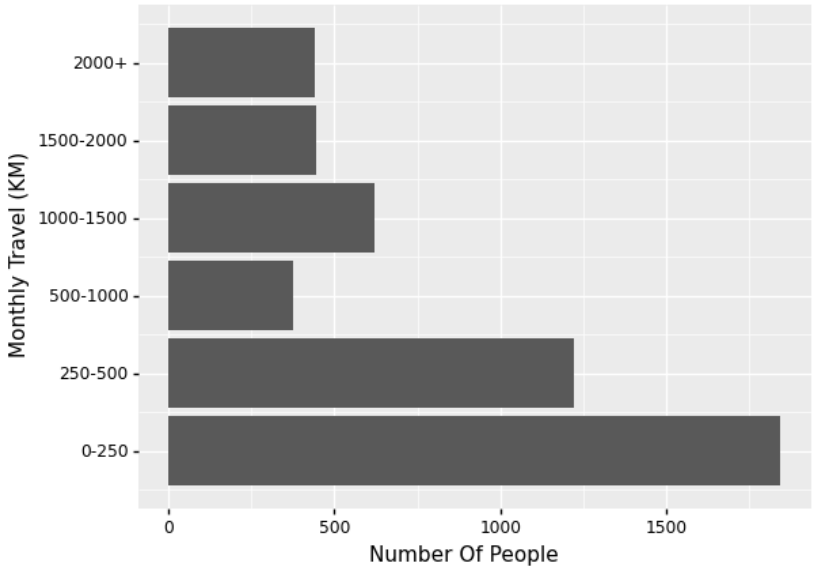
Furthermore, on average, the average length of the coast residents lived nearby was about 763 kilometers.

I then graphed the relationships between the variables.



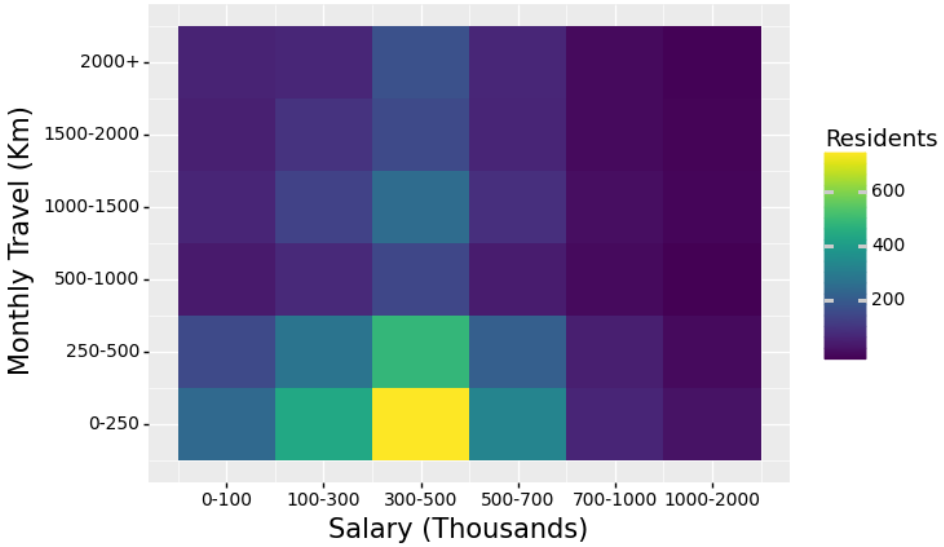
According to the boxplot, it appears that the distribution of age is uniform for females, males, and non-binary people. The mean and distribution of age appears to be equal between females and males. Note there are few non-binary people, and these people tend to be older adults (at least 45 years old). Also note that the age ranges from 21 to around 60 years old, thus neglecting the elders and youth.

Graph 2: Number Of People For Different Monthly Travel Distances



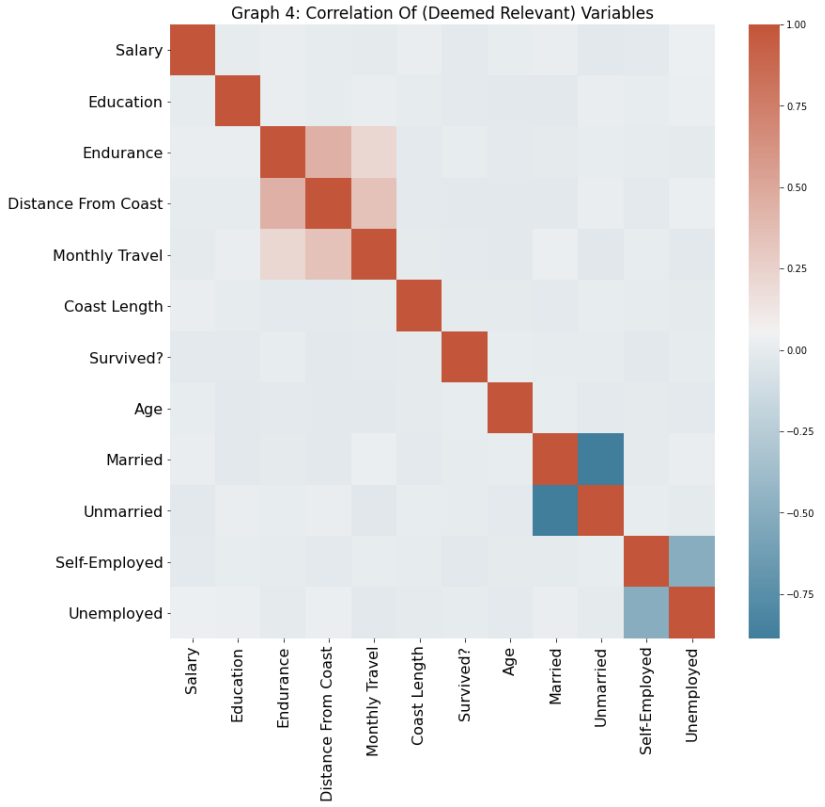
There appears to be a right skew for the monthly travel distribution (tail is on the bigger side of monthly travel). Most of the people travelled very little per month. The biggest group of people travelled less than 250 km, while the next biggest group of people travelled between 250-500 km. Relatively few people travelled more than 500 km per month.

Graph 3: Residents' Monthly Travel Based On Salary



Graph 3 charts the number of people for each monthly travel category for each salary bin. Note that there is a disproportionately huge number of people with salaries between $300K and $500K who do not go on monthly travel at all. This imbalance could bias the predictions. Overall, most of the residents had salaries between 300-500K and do not usually travel at all. Note that, interestingly, the dataset did not input numerical values for the salary or monthly

travel but rather put the numbers into categories, bins. As I later learned, this too happened to the variable distance from the coast.



I chose to look at the correlations of mainly the traditional demographic variables that I deemed important and ignoring the variables that indicated favorite food/color/car, as including all the other irrelevant variables would make the graph huge and difficult to read. There seems to be a small positive correlation among endurance level, distance from the coast, and monthly travel. There appears to be a deep negative correlation between unmarried and married, which makes sense as both these categories are mutually exclusive, with an additional reference category (Divorced). The same can be said for the deep negative correlation between self-employment and unemployment. Otherwise, most of the variables are uncorrelated with each other.

Unfortunately, this data does have its limits. The data did not have any specific location or region variables for any location-based fixed effects (although the data does have a variable that measures distance from the Atlantic Coast). The lack of these location factors could misleadingly bias Ordinary Least Squares coefficients, making some variables significant when in reality they are not. This can be particularly frustrating when we want to estimate the numerical effect each variable has on a resident's chance of survival. Furthermore, this dataset is concerned with only individuals in the aftermath of only one hurricane, Hurricane Irma specifically. Because of this, the data may not be as generalizable as a dataset documenting an individual's survival status for multiple hurricanes raging in different regions (Bailey, 2016).

Another limitation is that this dataset has some imbalances. This dataset has a slight gender-imbalance (with very few non-binary residents), a disproportionately small number of divorced residents, and a mediocre range of different ages. The age ranged from 21 to 57, thus neglecting youth, children, and elders, populations that could prove to be even more vulnerable to the deadly storm and the aftermath.

Another challenge from the dataset is that some of the variables that are thought to be inherently numeric (salary, monthly travel, etc) are actually inputted as bins. This loses the specificity of the data for these variables (and also forced me to apply inherent ordering).

Finally, another limitation of the dataset is that the data dictionary did not explicitly offer helpful definitions of the variables. Because of this, I had to apply my best judgement on the insight and definitions of some of these variables. For instance, the variable ENDU_DATA was defined as "endurance level." Unfortunately, that definition was not detailed or helpful enough to fully understand how the surveyors measured "endurance level," though I understood it loosely as, "how fierce is the motivation to survive" or how "hardy" someone was. On another instance, there were two seemingly similar variables (Dist_Coast and DIST_FRM_COAST) whose inputs were actually quite different. The data dictionary defines DIST_FRM_COAST as "distance from the coast" in kilometers but no explicit definition for Dist_Coast. I later defined Dist_Coast as the length of the coast, perhaps in a particular city.

Methodology

I will be completing a number of variable preprocessing. One of them is preserving inherent categorical ordering. A closer look at the ENDU_LEVELS (endurance level) reveals that this variable is categorical, with values such as: 1-3, 4-6, 10+, etc. I re-categorized the endurance variable to ensure inherent proper ordering. Just as the ENDU_LEVELS have natural ordering, so does the SALARY column, a variable that binned income values into String bins ("7K-1Million", "0-100K", "100K-300K", etc), so I plan to re-categorize the SALARY column into ordinal values. I also plan to re-categorize the EDU_DATA variable, which measures the highest level of education ("Graduate", "High School", "Post-Graduate", etc) and has intrinsic ordering.

I will be using the following variables as my features: age, distance from coast, education, employment, gender, monthly travel (distance someone travels per month), marital status, religion, and salary. The binary numeric class variable (whether someone survived or not) will be the target, the dependent variable.

Since most of the variables are categorical (and the dependent variable is categorical), I can use random forest or decision trees as a non-parametric modelling technique. Just as tree trunks split into branches and then into leaves, decision trees split data using a splitting criterion and stopping criterion to create homogeneous nodes. Although decision trees can be biased if some classes are dominant, the data fortunately has an equal balance of those who survived and those who died. Decision trees are intuitive, can handle numbers or categories, and can quickly visualize important features in predicting classes. Decision trees are prone to overfitting,

unfortunately. Random forests create ensembles of decorrelated, unpruned (fully-grown) decision trees by choosing the best split among a small set of random features. This makes random forests generally more robust against overfitting. This ensemble method, however, has 2 conditions: 1) base classifiers are independent and 2) base classifiers perform better than random guessing. Furthermore, random forests will not perform well if none of the features have predictive power (Gupta, 2020).

Another non-parametric model I used later on was xgboost. Xgboost implements gradient-boosted decision trees and are efficient and flexible. Boosting involves calculating the errors from prior models and adding new models to reduce loss. Note that boosting compares the model outcomes with the previous model outcomes and assigns weights to whichever outcome was misclassified (Pathak, 2019). The mechanism is called gradient-boosting because Xgboost uses gradient descent. Xgboost can perform both regression and classification, is fast, and is resistant to overfitting. Unfortunately, xgboost can be computationally expensive, require specific hyperparameters to maximize generalization performance, and can be difficult to interpret (Brownlee, 2021; Gupta, 2020).

As for parametric models, I can use Linear Discriminant Analysis (LDA). As a parametric model, LDA tends to be simpler in form and not as flexible as non-parametric models. LDA can not only classify objects but also compute the posterior probability an observation belongs to a class, or the confidence in labelling each human has survived or died (since the binary dependent variable is numeric). LDA reduces dimensions (for instance, given 13 features, LDA will choose 3 features to classify) and creates a linear plane decision boundary by assuming the prior, likelihood, and marginal probabilities for each (an altered form of Naive Bayes). LDA can be negatively impacted by outliers--luckily, none of the variables appear to have any outliers. LDA also works well with NAs and when the classes are well-separated. Unfortunately, LDA does not work well with irrelevant features, so I plan to eradicate those irrelevant variables before modelling (Sande, 2020).

I ran each model individually and calculated their performance metrics such as: recall, validation score, train score, accuracy, etc. But since I wanted to see which model as a whole was the best at predicting someone's death or survival, I ran a streamlined pipeline with some hyperparameters applied to each model to find out which model performed best at whichever hyperparameter.

Decision Trees have been used in literature. GeeksForGeeks, for instance, created a website explaining the mechanics of decision trees to predict which person was male, though, a majority of the website discussed the procedures of the tree instead of the actual prediction. The website even explained how decision trees use information gain, Gini Index, and entropy when determining which variable to split on. The tree decides which feature to split if splitting the feature yields a high information gain, low Gini index, and low entropy (Sharma, 2021).

Random forests are also popular. For instance, a study in health care involved using random forests (and other reference models and classification models) to classify the types of clinical emergency department outcomes in comparison with a conventional approach.

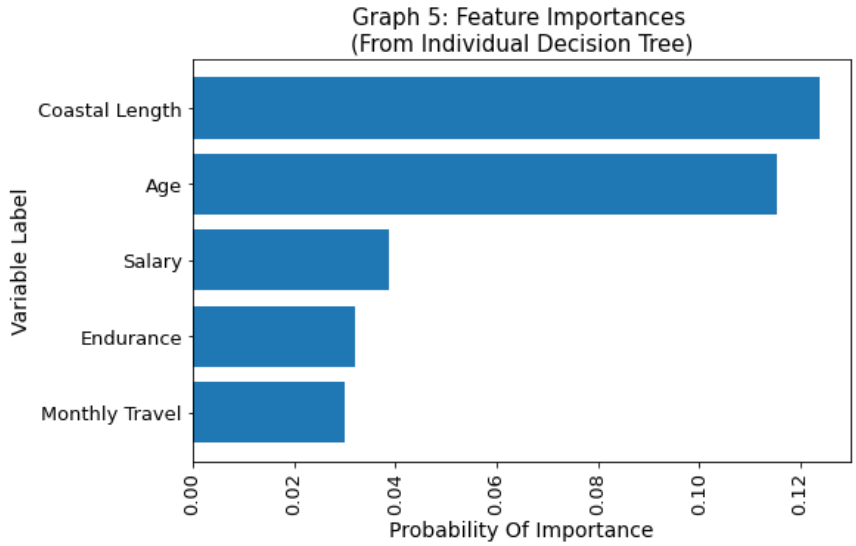
According to the paper, random forest's list of selected important features was consistent with the gradient boosted decision tree's selected important features. The random forest model had a higher ROC than the reference model. Limitations existed, unfortunately--the data excluded hospital visits without data from conventional classification and did not include all pertinent clinical variables (pre-treatment response, medication history, allergy history, etc). However, the study concluded that these Machine Learning models, including random forest, consistently had better predictive ability than the conventional approach (Goto et al, 2019).

Finally, LDAs have also been used in studies. Yang Xiaozhou wrote about using an LDA to recognize and identify images of different digits. The author describes the mechanics of LDA and even compares it with Quadratic Dimension Analysis (QDA). The author writes that although LDA is a robust classification and reduces dimensions, sometimes a linear boundary decision line is not enough to adequately classify. QDA assigns more flexible decision boundary lines but uses significantly more dimensions and parameters than LDA does. Yang claims that a regularized LDA (placing "certain restriction[s] on the estimated parameters") can compromise and get the best of both worlds, especially in a high-dimensional setting or when the number of features outnumber the sample size (Yang Xiaozhou, 2020).

Results

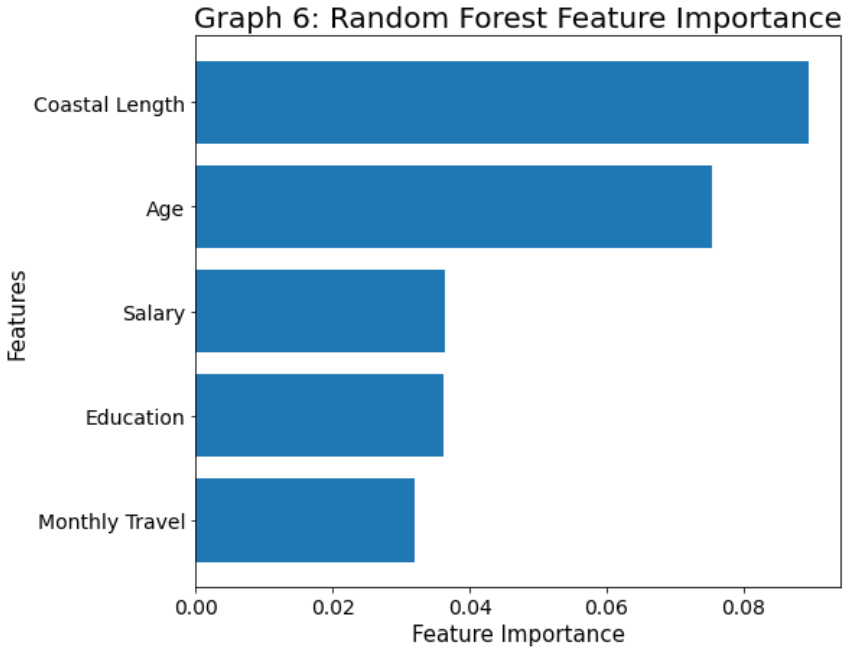
The streamlined pipeline determined that xgboost was the best model. It has a learning rate of 0.8, a max depth of 5, 10 estimators, and an alpha and lambda regularization constant of 15.

Interestingly, each individual model had different claims as to what features were considered important in classifying survival and death.

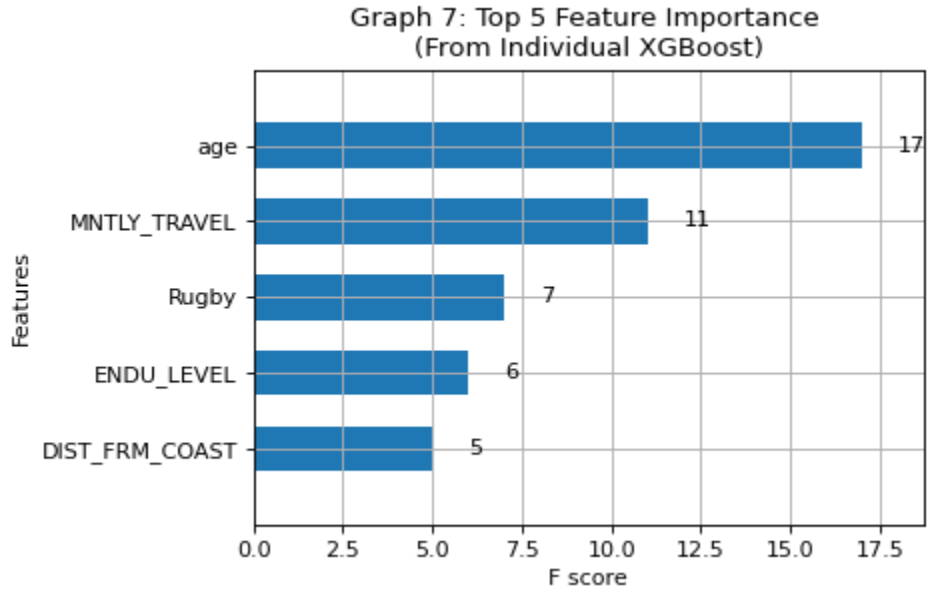


The individual decision tree determined that the most important variables we see are coast length, followed by age, then salary, endurance level, and monthly travel, and others. But

as a whole, distance of coast still does not have the best importance because its probability of accurately predicting whether someone survived or died is still no more than 15%.

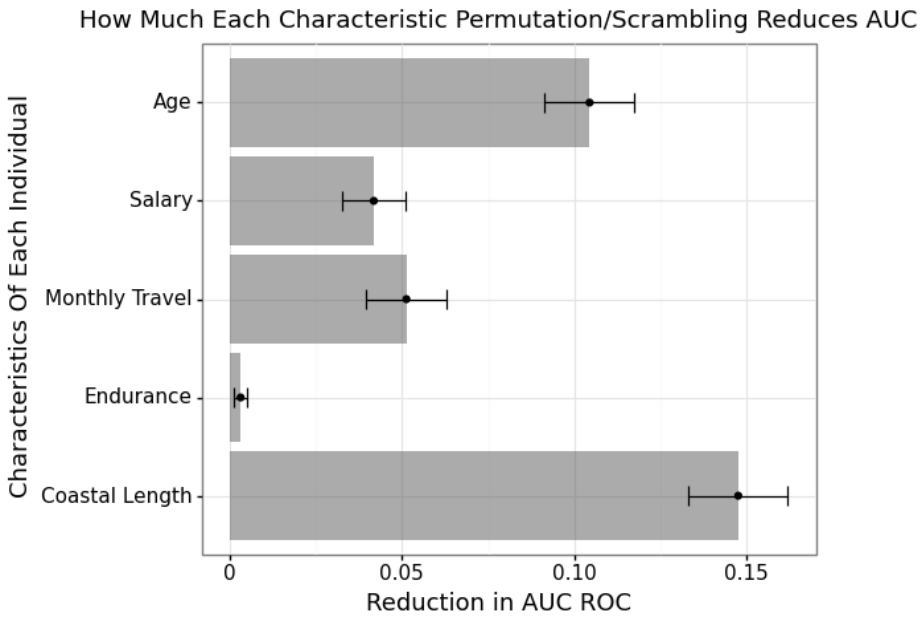


According to the random forest feature importance, coastal length is the most important feature, followed closely by age. The rest of the features have about half the importance or less. Note that the top 5 important features mirror the decision tree's chosen important features, and that both models scored coastal length as the most important, followed closely by age.



According to the feature importance plot, as determined by the individual xgboost model, age is the most important feature with the highest importance score, followed by monthly travel, then by having rugby as a favorite sport, then endurance level, then coast length.

Graph 8: Feature Importance



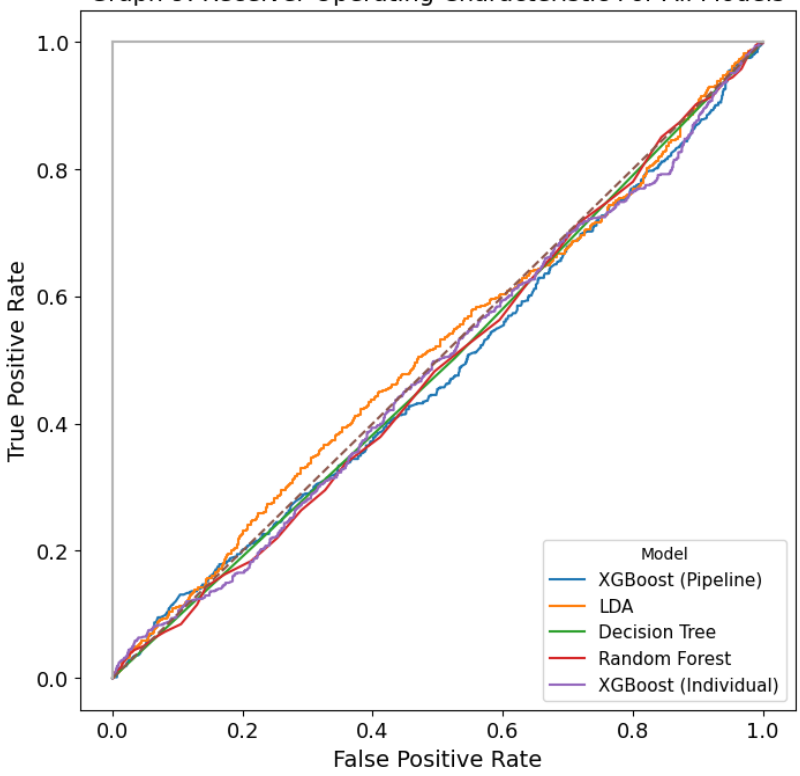
Above is the permutation method to achieve feature importance from the pipeline's xgboost model. Note that the most important feature from this permutation is coast length. The most important features are: coast length and age, then followed by monthly travel, salary, endurance level. Coastal length, the variable that can reduce the most AUC, at best reduces the AUC at around 15%. Permuting age, the next most significant feature, age, decreases the area under the ROC curve by about 10%. The rest of the features reduce the AUC only a little bit, so I did not consider them to be very important features.

Table 2: Feature Importance Summary

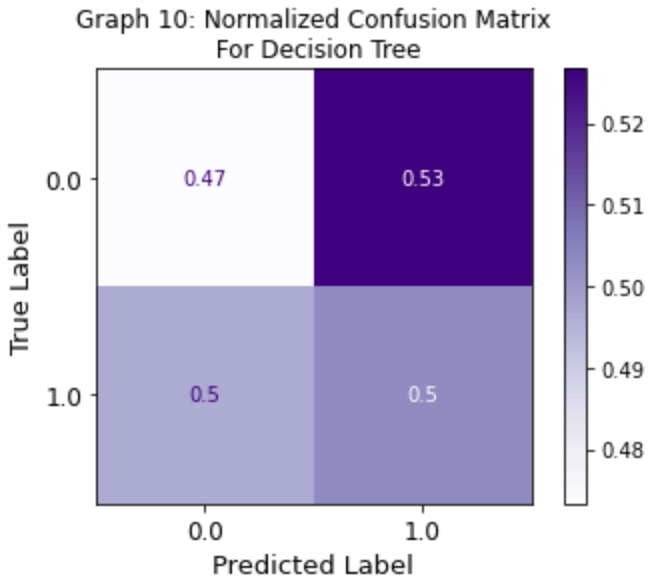
Model	Most Important Feature	Second Most Important Feature
Decision Tree	Coast Length	Age
Random Forest	Coast Length	Age
XGBoost (Individual)	Age	Monthly Travel
XGBoost (Pipeline)	Coast Length	Age

Above lies the summary of the top two important features of each model. It appears that coast length is the most important feature, followed by age and a little bit of monthly travel. It is interesting to point out that the pipeline xgboost and the individual xgboost model had slightly different most important features. Perhaps the coastal length was important because a longer length of the coast meant higher, more violent waves to destroy houses. Age also made sense, since older people tend to be more susceptible to pain and are not as hardy as young adults.

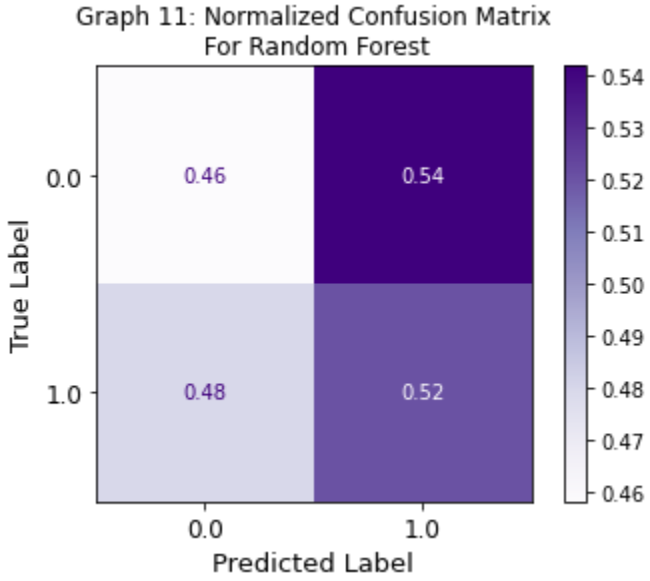
Graph 9: Receiver Operating Characteristic For All Models



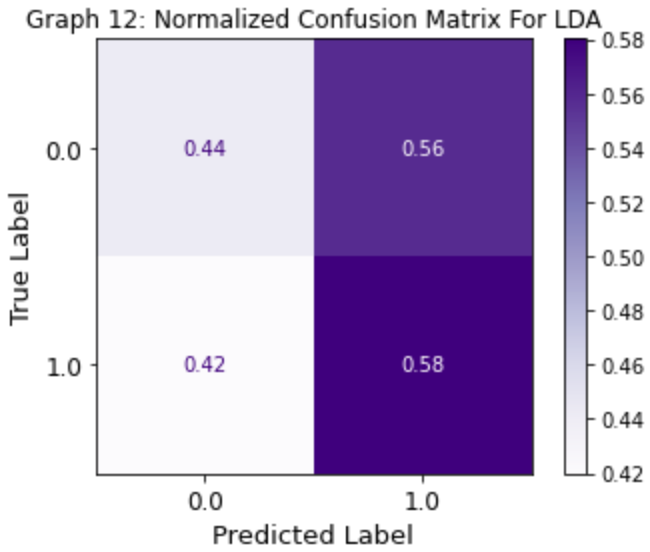
According to the ROC curve, the XGBoost model (chosen from the pipeline) correctly classified positive cases only about 50% of the time, barely better than a coin flip. In fact, all of these models were terrible in finding relationships and predicting positive cases and were marginally better than random guessing. There is no model that performed significantly well, unfortunately.



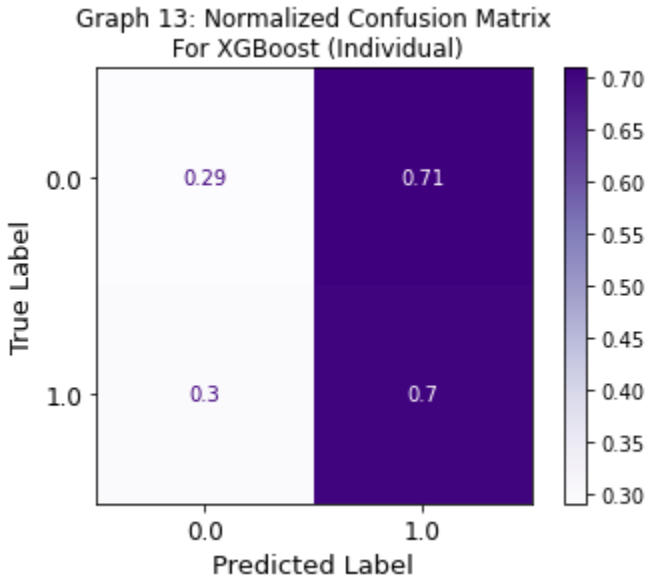
The decision tree had poor predictive accuracy overall. It appears that, given how high both the true positive and false positive rate compared to the false negative and true negative rate, the decision tree tended to label more cases as positive instead of negative. Its ability to differentiate between positive and negative cases is suboptimal.



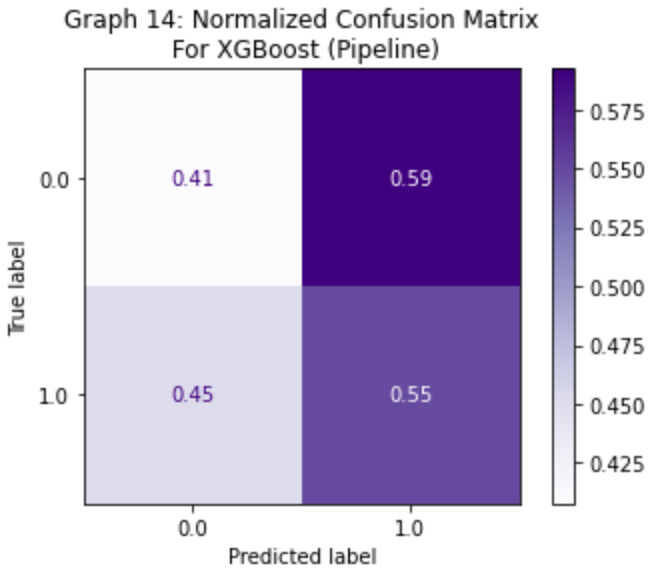
Random forest had poor predictive accuracy. For each case, random forest was more likely to predict it was a positive case (resident died), which explains the higher percentage for true positives and false positives. Random forest was better at predicting positive cases than negative cases. However, overall, random forest's ability to differentiate between positive and negative classes is suboptimal.



LDA also had poor prediction accuracy, though, LDA was better at predicting positive cases than negative cases. Since the false positive rate is also almost as high as the true positive rate, it appears that LDA could have automatically labelled each case as positive.



Xgboost still had poor predictive accuracy when it came to detecting negative cases but mediocre predictive accuracy for detecting positive cases. The individual xgboost had the highest true positive rate, yet also the highest false positive rate. It appears that the individual xgboost model tended to assign more positive labels than negative labels to cases, since both its false positive and true positive rates are high. Given that a case was positive, the individual xgboost could correctly identify its class fairly well, yet given a negative case, it identified the class poorly.



The false positive rate is significantly higher than the true negative rate, and the true positive rate is higher than the false negative rate. It appears that the pipeline xgboost model tended to give out more positive labels to cases than negative labels, since both its true positive rate and false positive rate were equal. If the case was positive, the pipeline xgboost could accurately identify it somewhat well; for a negative case, the model could identify it poorly.

Table 3: Model Metric Summary

Model	MSE	RMSE	MAE	AUC	Accuracy	Recall	Train Score	Validation
Decision Tree	0.511	0.715	0.511	0.488	0.489	0.503	1.0	0.487
Random Forest	0.51	0.714	0.51	0.489	0.49	0.521	1.0	0.517
LDA	0.488	0.698	0.488	0.511	0.512	0.58	0.576	0.514
XGBoost (Individual)	0.5	0.707	0.5	0.497	0.5	0.702	0.596	0.511
XGBoost (Pipeline)	0.52	0.721	0.52	0.479	0.48	0.55	0.657	0.509

The random forest had the highest MSE and almost tied with Decision Tree in RMSE and MAE, though most models performed relatively similarly for those MSE, RMSE, and MAE. Note that LDA had the lowest MSE, RMSE, and MAE, probably because LDA is a parametric model and tends to be simpler than non-parametric models. LDA had the highest AUC and accuracy, though that does not mean LDA is amazing at predicting, given that LDA's AUC was only marginally better than the lowest AUC. The decision tree and random forest had very high training scores but terrible validation scores, indicating that they drastically overfit the data. The pipeline xgboost somewhat overfit the data too, given that the difference between the train score and validation score was about 15%. LDA had the smallest difference between training score and validation score, and these two scores were also low, indicating that LDA underfit the data. The individual xgboost had a surprisingly high recall compared to other models, indicating this accurately classifies positive cases relatively well. What's even more surprising is that the chosen

xgboost from the pipeline actually had a noticeably lower recall score, unfortunately. However, both the individual xgboost and pipeline xgboost had relatively higher recall than the other models' recall scores. LDA had the highest validation of 0.514 while decision tree had the lowest validation of 0.495, so LDA performed marginally better than all the other models on validation scores.

Conclusion

My models generally classified poorly, notably in predicting positive cases. For each metric, none of the models did distinctly well except for the recall score for xgboost. Xgboost's recall scores, for both the individual and the pipeline version, were surprisingly relatively higher than all the other models' recall scores, probably because xgboost is quite a flexible, efficient model. However, the models all did poorly in correctly classifying negative cases, making it seem as if they tended to over-predict those who died. The streamlined pipeline declared that the most important, significant feature was the length of the coast. This was in line with my individual decision tree's feature importance chart labelling age and length of the coast as significant attributes. However, an individual xgboost declared that age and monthly travel were significant features. Although none of the models universally agreed as to what the most important features were and in what order, the most common features they praised were first coast length and then, age, and then monthly travel. However, since only one model, the individual xgboost, declared monthly travel as the top 2 most relevant features, it is still unclear how much of a predictive power monthly travel has on survival.

One of the hardships in dealing with classification problems is that sometimes, models do not perform well. Their ability to recognize positive cases could be barely better than random guessing. However, just as embracing models who classified well is important, so too is embracing models who did not perform well. In this paper, as a whole, none of the models performed well. This, however, could give insight that the hurricane truly did randomly kill any resident who was in its path, ignoring political or socio-economic boundaries.

Some limitations of the project stem from the dataset. The dataset did not contain data for elders or youth, so these findings are not generalizable for the whole population. Furthermore, the dataset failed to record location data such as cities. This project could have failed to account for different survival baselines for location-specific effects.

Because my models performed poorly, more research and data is needed to further expand on this project. These models should never be used for policy-making unless more accurate, more robust models are developed. However, policy-makers should make sure to not rely too heavily on the model predictions. If the model inappropriately targets a senior who is very physically fit and hardy, the money and resources could be disproportionately spent to make the senior less vulnerable to the storm while these funds could have better helped a poor immigrant family with many children survive. The dataset also had a lot of irrelevant variables such as favorite movie genre, favorite cuisine, etc. and lacked some key demographic variables such as race, location, etc, which could have led to poor predictive accuracy. Policy-makers

should make sure to collect more key demographic variables if they want to help people survive the storm. What we do know is that a few of the biggest predictors of whether someone survived was the length of the coast, age, and how far they travelled each month. When the region is preparing for a storm, the policymakers could enact policies that target aid towards people living along the coast or the elder population, as old age increases recovery time and the chances of developing chronic illnesses and complications. Elders are particularly vulnerable to natural disasters because of "their impaired physical mobility, diminished sensory awareness, chronic health conditions, and social and economic limitations" (Benson, n.d.).

Implementation Appendix

The dataset is available on the kaggle website. No scraping the web is necessary to obtain the data. Unfortunately, the dataset was not clean for this project.

Since the dataset was not exactly clean, I had to pre-process the data. The dataset had a lot of categorical variables. Some of these variables have categories that have inherent ordering, such as education: no education, high school, undergraduate, post-graduate. I made sure my data models understood that these categorical variables had a natural ordering.

Surprisingly, some variables that are traditionally considered numeric (salary, monthly travel, etc) actually did not have numeric inputs but rather categories, or bins. Therefore, I too had to apply inherent ordering, though, I still lost the specificity of some of those variables had they been numeric.

Some of these categorical variables, on the other hand, had categories that did not have any natural ordering, such as the variable Gender: female, male, and other. I subsequently decided to construct dummy variables for these variables so I could have numerical values for easier modelling. I made sure to also create a reference category to prevent multicollinearity. For instance, gender's reference category was the "other" label.

Besides dummy variables and inherent ordering, I also scaled some numerical variables to prevent some variables with huge ranges from garnering an unfair weight advantage in modelling analysis and creating bias. For instance, the variables age and distance of the coast have drastically different ranges. Age ranged from 21 to approximately 57 (which means the data neglected the youth and elders). Coast length ranged from around 50 to about 1500. Those were the only variables I scaled, as most of the other variables had small, single-digit values after I applied inherent ordering or dummy variables.

One surprising insight was that the individual xgboost scored a significantly higher recall score than all the other models. What's also surprising is that the pipeline's chosen model, although also xgboost, scored a lower recall score than the individual's recall, even though I offered the same hyperparameters.

Another surprising insight was that although the pipeline xgboost declared coastal length as the most important variable, coastal length was not even considered to be the top 6 most important features for individual xgboost. All the other models listed age as the 2nd most

important feature and monthly travel as a very mildly important feature, so it was very surprising when individual xgboost listed monthly travel as high as the second most important feature.

Bibliography

Bailey, M. (2016). *Real Stats: Using Econometrics for Political Science and Public Policy.* Oxford University Press.

Benson, William F. (n.d.). *CDC's Disaster Planning Goal: Protect Vulnerable Older Adults.* Center for Disease Control and Prevention. https://www.cdc.gov/aging/pdf/disaster_planning_goal.pdf

Brownlee, Jason. (2021 Feb 17). *A Gentle Introduction to XGBoost for Applied Machine Learning.* Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning–Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. *JAMA Netw Open*. 2019;2(1):e186937. doi:10.1001/jamanetworkopen.2018.6937

Gupta, Shailaja. (2020 Feb 28). *Pros and cons of various Machine Learning algorithms.* Towards Data Science. https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6

Klinger, H., & O'Hara, K. J. (2018, Aug. 1). *2017 Hurricane Irma: Facts, FAQS, and how to help.* World Vision. https://www.worldvision.org/disaster-relief-news-stories/2017-hurricane-irma-facts

Ovalle, D. (2017, Sept. 15). *Keys homes, battered but standing, may be a model for reducing damage in Florida.* Miami Herald. https://www.miamiherald.com/news/weather/hurricane/article173408496.html

Pathak, Manish. (2019 Nov 8). *Using XGBoost in Python.* Datacamp. https://www.datacamp.com/community/tutorials/xgboost-in-python

Sande, Sumaiya. (2020, Nov. 4). *Pros and Cons of popular Supervised Learning Algorithms.* Analytics Vidhya. https://medium.com/analytics-vidhya/pros-and-cons-of-popular-supervised-learning-algorithms-d5b3b75d9218

Sathyajit, R. (2017). *WMO Hurricane Survival Dataset*. Kaggle. https://www.kaggle.com/rahulsathyajit/wmo-hurricane-survival-dataset

Sharma, Abhishek. (2021 Mar. 17). *Decision Tree Introduction with Example.* GeeksforGeeks. https://www.geeksforgeeks.org/decision-tree-introduction-example/

Wingard, G.L., Bergstresser, S.E., Stackhouse, B.L. et al (2019, Nov. 22). *Impacts of Hurricane Irma on Florida Bay Islands, Everglades National Park, USA.* Estuaries and Coasts 43, 1070–1089 (2020). https://doi.org/10.1007/s12237-019-00638-7

Yang, Xiaozhou. (2020 May 9). *Linear Discriminant Analysis, Explained.* Towards Data Science. https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b