Beyond the Titanic: Exploring the Oceanographic Factors Behind Maritime Disasters Along the Eastern Coast of the United States

Zev Burton, MS Candidate

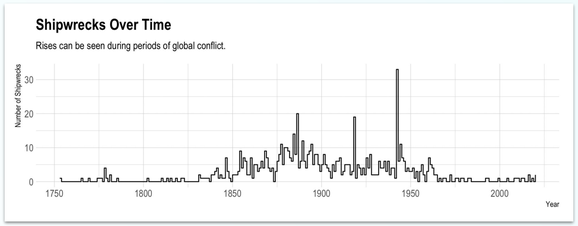
Department of Data Science and Analytics, Georgetown University

Abstract

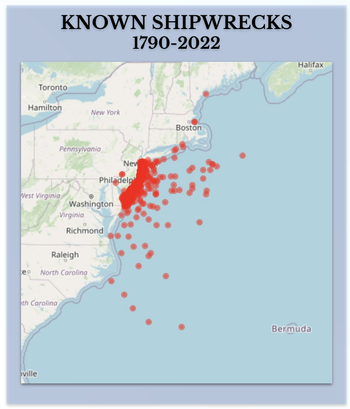
This study investigates the role of oceanographic features in contributing to shipwrecks along the eastern US coast. Using a database of shipwrecks and ocean pattern data from the National Oceanic and Atmospheric Administration, the study employs random forest and neural network models to identify crucial predictor variables. The results indicate that the strength of currents is the most significant factor in predicting shipwrecks, followed by water density and wind speed. The study emphasizes the need to incorporate current-based predictor variables in shipwreck prediction models and consider multicollinearity in predictive models. Further research could explore other factors involved in shipwrecks in the region and their relative significance. Overall, this study provides valuable insights to improve predictive models and mitigate the risks associated with maritime navigation along the eastern US coast.

Introduction

Shipwrecks pose a significant maritime issue due to the potential loss of life, property, and environmental damage they can cause. When a ship sinks, the crew and passengers aboard are put at risk of drowning, hypothermia, and other life-threatening injuries. In addition, shipwrecks can result in the loss of valuable cargo, which can have a significant economic impact. This is, of course, only glossing over the environmental and network impacts that a sinking can have. In this study, various machine learning algorithms were used to identify the key oceanographic features that contribute to shipwrecks on the Eastern Coast. Our study focuses on oceanographic features because they are often overlooked in studies of shipwrecks despite their importance in favor of vessel characteristics, human error, and, more often than not in history, myth. Our analysis aims to fill this gap in the literature by providing insights into the relationships between oceanographic features and shipwrecks. While it is understood that we cannot change the ocean, we hope that our findings will influence both the creation of shipping routes to avoid particularly hazardous and threatening waters and the increased monitoring of such oceanographic characteristics to inform real-time decision making.



Materials

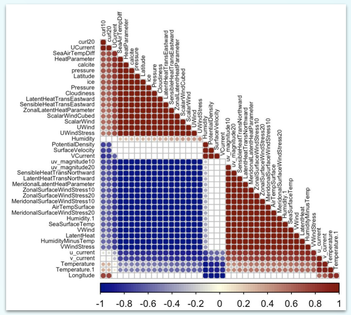


The New Jersey Maritime Museum has created a database of shipwrecks along the east coast. Key information, such as location, flag, and whether the ship was found or not is also recorded.

Additional data on the ocean patterns in this region was gathered from the National Oceanic and Atmospheric Administration, a Washington, D.C.-based scientific and regulatory agency within the United States Department of Commerce.

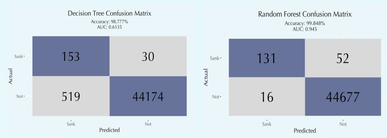
Model Selection

While natural, it is unfortunate for our case that ocean measurements are not independent of each other. The temperature on the surface of the water is likely to have some impact on the temperature 10 feet below. This correlation plot is shown below. The labels are not nearly as important as seeing the wide swaths of correlation between variables.

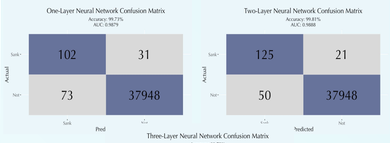


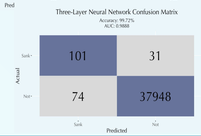
Instantly, we are prohibited from using SVM and logistic regression, since they require a lack of multicollinearity. While an attempt was made to remove multicollinearity, it left us with virtually no data to go off of. Just in case, we ran the logistic and SVM models, to little benefit other than satisfying our curiosity..

Fortunately, we had the computing power available to test several models, including three neural networks. Below are the confusion matrices for the Decision Tree and Random Forest Models:



To determine optimal hidden layer sizes and node amounts, we trained neural networks on various sizes and layers on a smaller test set and validated with a validation set. We then applied this to the entire dataset. Below are the confusion matrices for the optimal one-layer (30), two-layer (50, 50), and three-layer (60, 60, 60) neural networks.



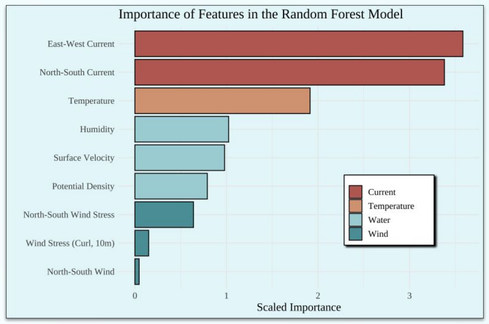


There is an inherent difficulty in comparing traditional classification models with more advanced neural networks. The solution is to look at both accuracy and AUC, the former of which is much less important when the vast majority of the data is a "failure." Nonetheless, when comparing the accuracy and AUC of all models, two stuck out as being particularly impressive: the Random Forest model and the Two-Layer Neural Network.

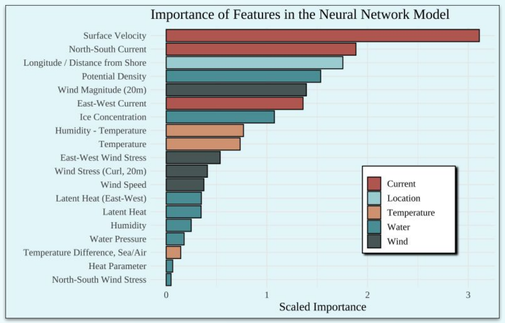
Model	Accuracy	AUC
Logistic	0.9959	NA
Decision Tree	0.9878	0.6135
Random Forest	0.9985	0.9450
SVM	0.9959	NA
One-Layer NN	0.9973	0.9879
Two-Layer NN	0.9981	0.9888
Three-Layer NN	0.9972	0.9888

Since we are primarily concerned with the features that cause shipwrecks, we can actually look at both the features that are most important in both the Random Forest and the two-layer neural network.

Feature Importance



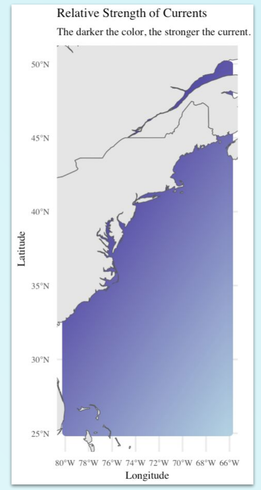
On the right, we have the positively important features in the neural network. On the left, the random forest. Both are colored by the 'type' of feature, such as the temperature, the wind, or the water. We can see that the brick red, which indicates features associated with currents, are far and away the most important in both models. We don't see this for any other category of feature.



Results

This study analyzed factors contributing to shipwrecks on the eastern US coast using two primary models, random forest, and a two-layer neural network. The results showed that the strength of currents is the most significant predictor variable in understanding the causes of shipwrecks. The other factors, such as water density and wind speed, were found to play a secondary and tertiary role in predicting shipwrecks. The random forest and neural network models produced consistent results, suggesting that current-based predictor variables significantly enhance the accuracy of shipwreck prediction models.

The study also considered multicollinearity and found that, while the exact same variables may not appear as predominant predictor variables in each model, the strength of currents consistently emerged as the most important variable. Therefore, incorporating the strength of currents as a predictor variable can help improve the accuracy of shipwreck prediction models specifically for the eastern US coast.



Conclusion

In conclusion, this study underscores the importance of incorporating current-based predictor variables in shipwreck prediction models on the eastern US coast. The analysis showed that the strength of currents is the most critical factor in predicting shipwrecks, followed by water density and wind speed. The study also highlights the need to consider multicollinearity in predictive models to accurately identify the most significant predictor variables.

Future research could further explore additional factors contributing to shipwrecks on the eastern US coast and investigate their relative importance in predicting shipwrecks. For example, it could be valuable to explore the impact of human factors, such as human error or equipment failure, on the occurrence of shipwrecks. Additionally, research could focus on regional variations in the predictors of shipwrecks, as different geographic areas could have unique factors contributing to shipwrecks on the eastern US coast.

Overall, the findings of this study provide valuable insights into the factors contributing to shipwrecks on the eastern US coast and can aid in developing better predictive models to reduce the risks associated with maritime navigation in this region.