The Grand Migration:
Analysis of the Lobster Population Dynamics and its Financial Implications

Team Number: 12265

Modeling The Future Challenge 2023

March 2023

Contents

1 Executive Summary

The bright red, boiled, and buttered lobster is a trademark dish of New England cuisine; however, due to various biotic and abiotic limiting factors, the northeastern American Lobster population is being forced to migrate North, directly impacting the communities that rely on harvesting wild lobsters. Our paper aims to give accurate and significant recommendations for individual lobster fishermen and larger fishing companies in order to help them continue bringing in a steady income and sustainable supply.

Our raw spatial data had NaNs for certain points, so we discretized our dataset so that each pair of longitude and latitude would represent a 0.5 by 0.5 area in degrees. The remaining values that were missing were interpolated by Tobler's Law, where a certain area's true value is relatively close to the mean of its surroundings.

We first designed three ensemble Machine Learning models: Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM), and Random Forest to predict lobster concentrations at each point in the northeastern coastal region. Year, position, sea surface temperature, predator and prey concentrations, as well as previous lobster data were all considered, using a training size of 85% of every input before 2018. Our models' results show that lobsters are generally moving North, most likely due to warming temperatures in southern New England. We then used a Monte Carlo simulation to evaluate model performance on predicting the 2019 population distribution with 2018 input data. All three of our preliminary models have never been trained on 2019 data, so our evaluation is an accurate representation of model performance. Our Monte Carlo simulation assesses our models with different sample sizes, where each sample size would represent a hypothetical fishing company. Our models then went through a second round of training, this time utilizing all data available so we can give a more comprehensive prediction for 2020, again using a training size of 85%. Next, we used Depth First Search to find the most efficient path from the three most common ports to randomly selected points; this distance is translated into fuel costs and also factored into our business recommendations. Some areas may have higher lobster concentrations, but their revenue is offset by money spent on traveling there. A distance map for each point from different ports is used with our final 2020 predictions. The reason that we do not offer predictions for 2023 is that the most recent data is likely kept confidential for market competitiveness, and thus it would neither be accurate nor ethical to give predictions for 2023 and state such as fact.

In the end, we made business, financial, and insurance recommendations to both small and larger-sized lobster harvesting companies. The business recommendations advise companies located near major Northeastern ports by pointing out the combination of regions in the ocean for maximum profit. The financial recommendations recommended the company relocate north so it better matches the forecasted future lobster population, which helps optimize long-term growth and mitigate the risk of revenue reduction. The insurance recommendations address the short-term severe risk as a result of uncontrollable uncertainties including natural disasters and extreme weather to increase financial stability, especially for smaller businesses. Nevertheless, all three recommendations aim to improve both short-term and long-term sustainability that will maximize economic efficiency and mitigate risks posed by nonseasonal lobster movements.

2 Background Information

The lobster industry is one of the most well-known fisheries in the world, with special economic and cultural significance to the northeastern region of the United States. Throughout the past decade, lobstermen continuously haul in these crustaceans that contribute over $1.5 billion to the U.S. economy [2]. Although strong prices, high demands, and a consistent supply have provided these self-employed workers and large fishing corporations with respectable revenue in the past, future environmental issues threaten these stable conditions. According to a 2021 report [3], the Gulf of Maine is warming up faster than 99% of the Earth's oceans; this speedy warming is causes lobsters to migrate North, posing a direct risk for the 5600+ lobstermen as lobster catches become increasingly unpredictable and an indirect risk for all the restaurants and consumers who relies on the supply of lobsters to maintain their business. [4]

Global warming not only causes New England lobsters to migrate but also pressures predators to migrate North. While this does not cause any additional issues for the lobsters that were accustomed to avoiding these predators in the lower part of New England, it does induce drastically worse circumstances for those in Maine who never had to deal with such threats. Therefore, this paper will address such risks posed to the lobster industry through rigorous analysis of the relationship between lobster concentrations and changing climate conditions as well as conduct predictions for the future years to help alleviate the possible losses to the lobster industry.

Specifically, in order to minimize the potential losses to the industry, we considered the following three questions as essential for both individual lobster harvesters and larger fishing corporations. Our paper attempts to answer them to the best of our abilities with mathematical models. [1]

1. There is an evident upward trend in the number of hurricanes over the 120-year period, as indicated by the red trend line. This could suggest an increase in the frequency of hurricanes likely due to climate change, or simply mean the advancement of meteorological technology improves hurricane reporting and tracking.

2. The individual data points exhibit considerable fluctuation from year to year, likely indicating that the number of hurricanes is subject to significant variability. While the overall trend is upward, individual years may see spikes or drops in hurricane frequency.

3. How much money will I make?

3 Data Methodology

To ensure precision and accuracy for the output of our mathematical model, we have recognized and utilized four datasets, including the historical American Lobster Population, Atlantic Cod Population, Atlantic Rock Crab Population, and Sea Surface Temperature.

[5] [6] [10]

3.1 American Lobster Population

We acquired this dataset from the database “OceanAdapt,” developed by Rutgers University and National Oceanic and Atmospheric Administration (NOAA). [7]

Timeframe: 1970 - 2020

Variable	Variable Name	Description
Longitude	lon	Geospatial Coordinate Variable
Latitude	lat	Geospatial Coordinate Variable
Year	year	Time (Annually)
Population Concentration	wtcpue	The weight caught per unit effort in kg

This dataset will serve as a baseline in order to project future lobster population distribution.

3.2 Atlantic Cod Population

We also acquired this dataset for the historical Atlantic Cod Population distribution from the database “OceanAdapt,” developed by Rutgers University and National Oceanic and Atmospheric Administration (NOAA). [11]

Timeframe: 1970 - 2019

Variable	Variable Name	Description
Longitude	lon	Geospatial Coordinate Variable
Latitude	lat	Geospatial Coordinate Variable
Year	year	Time (Annually)
Population Concentration	wtcpue	The weight caught per unit effort in kg

According to Sciencing, ”Codfish are among the primary predators of true lobsters. The Atlantic cod often feasts on American lobsters found off the shores of eastern North America by ripping at the creatures until they die, then breaking open their shells and tearing at the meat.” This dataset will factor in as a biotic limiting factor to the lobster population as lobster is prey for Atlantic Cod.

3.3 Atlantic Rock Crab Population

Historical Atlantic Rock Crab Population distribution was also acquired from the database “OceanAdapt,” developed by Rutgers University and National Oceanic and Atmospheric Administration (NOAA). [10]

Timeframe: 1970 - 2019

Variable	Variable Name	Description
Longitude	lon	Geospatial Coordinate Variable
Latitude	lat	Geospatial Coordinate Variable
Year	year	Time (Annually)
Population Concentration	wtcpue	The weight caught per unit effort in kg

According to Smith, J. 2019, "Atlantic rock crab was the principal prey of American lobster while American lobster was almost never eaten by Atlantic rock crab." This dataset will factor in as another biotic limiting factor since Atlantic Rock Crabs are one of the primary prey of American Lobsters.

3.4 Sea Surface Temperature

We acquired the interpolated gridded sea surface temperature data from the Met Office Hadley Centre.

Timeframe: 1970 - 2019

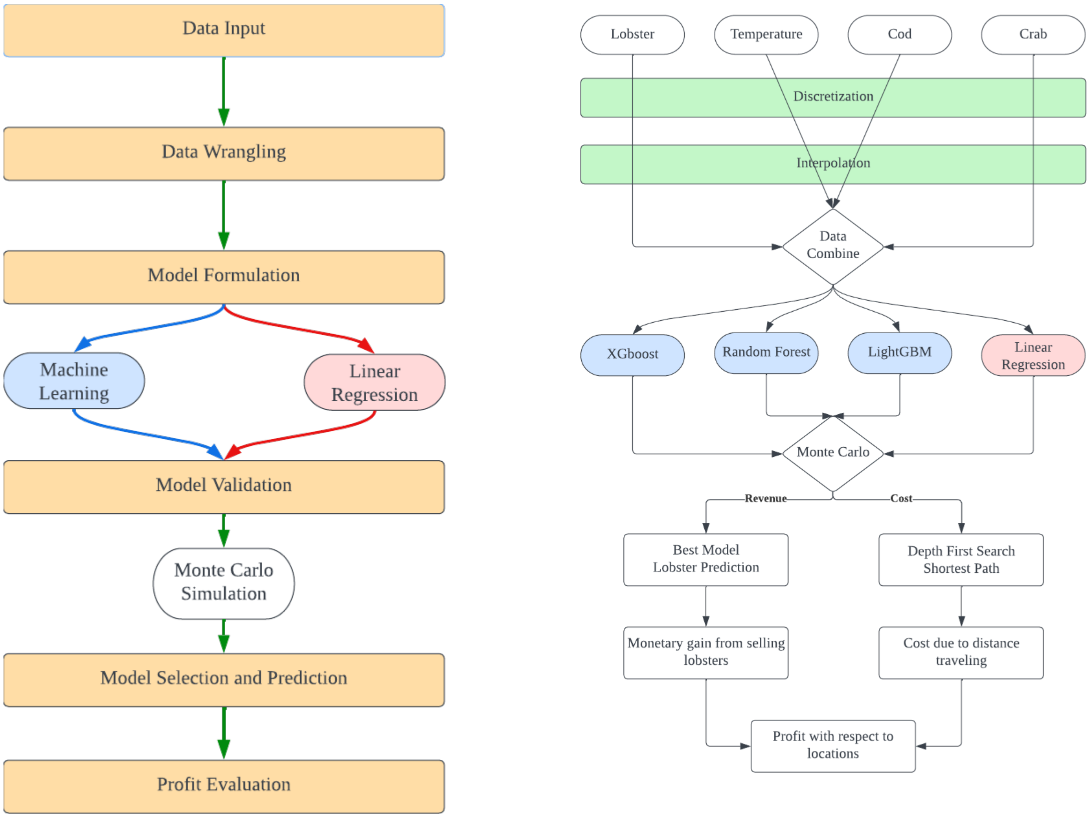
Variable	Variable Name	Description
Longitude	lon	Geospatial Coordinate Variable
Latitude	lat	Geospatial Coordinate Variable
Time	time	Time (Monthly)
Sea Surface Temperature	sst	°Celsius

Lobsters can remain healthy in water temperatures of up to 20°C. However, prolonged exposure to temperatures beyond this threshold causes respiratory problems, immune problems, and an increased risk for shell disease. This dataset provides information on environmental limiting factors for the lobster population. [8]

4 Mathematics Methodology

4.1 Model Overview

The input of our mathematical model will include the four datasets identified above. The output of our model will be a map provided for the individual fisherman and fishing corporation indicating the amount of profit gained for a particular region in the Atlantic ocean. To accomplish this objective, we first cleaned the data through discretization and interpolation. This process generalized the region which allowed us to unify the initially unevenly distributed data into equally sized tiles. The second step is to formulate different models and feed them the processed data. In this process, we have considered various machine-learning algorithms including XGBoost, Random Forest, and LightGBM. We also used Multivariable Linear Regression as a baseline for model comparison. We compared and tested model accuracy through a Monte Carlo simulation. For each sample size, we covered the range of possible outcomes by randomly generating data points with 1 million repetitions. We then corresponded the data points with the highest concentration based on our prediction with their actual data and inserted this value into the Monte Carlo simulation to see how better our prediction is compared to complete randomization. We used this tactic across all four algorithms and used the best one to predict future values. Lastly, we related the lobster concentration map with the cost of fishing by identifying the distance of travel from ports using Depth First Search. Finally, we took the difference between the revenue generated from the lobster sale and the expenses, which is our expected profit.



(a) General Sections Overview

(b) Specific Modeling Processes

Figure 1: Model Overview

4.2 Assumption and Justification

1. **Assumption:** The most important abiotic limiting factor that could influence the American lobster population and its distribution is Ocean temperature; other similar factor categories such as ocean salinity and acidity are negligible.

   Justification: According to Katie Wagner, a spokesperson for National Oceanic and Atmospheric Administration Fisheries, "Lobster populations are increasing at the cooler (northern) edge of their range, and declining at the warmer (southern) edge of the range due to reproductive failure — fewer juvenile lobsters living to adulthood to reproduce."[13] Ocean temperature has an immediate short-term effect on the lifespan of a lobster, significant long-term effects in forced migration, and the reproduction of the entire lobster population. Thus, it should be considered as the primary climate change factor that will influence the American lobster population dynamics.

2. **Assumption:** The population of Atlantic Cod and the population of Atlantic Rock Crab is significant and sufficient in representing the biotic limiting factors on the American Lobster Population aside from human consumption.

Justification: According to Shields from Sciencing, "Codfish are among the primary predators of true lobsters. The Atlantic cod often feasts on American lobsters found off the shores of eastern North America by ripping at the creatures until they die, then breaking open their shells and tearing at the meat."[14] We acknowledge the ever-changing distribution of the cod population and its significant impact on American Lobsters. Thus, including the population dynamics of the Atlantic Cod can help us to better predict lobster distribution in the future. Similarly, the population of Atlantic Rock Crabs will be indicated as the primary food source for American Lobsters, which assists in analyzing and predicting the migration patterns of lobsters.

3. **Assumption:** Every individual geospatial data point is related to and dependent on the values of its surroundings.

   Justification: According to Waldo Tobler's First Law of Geography "Everything is related to everything else."[15] Therefore, by using the values from the surrounding regions we can interpolate the missing value for a specific data point.

4. **Assumption:** Human Consumption of lobsters maintains the status quo.

   Justification: We acknowledge that the largest predator of lobsters is humans, every year, over 100 million pounds of lobsters are being caught and sent to the market for human consumption.[16] We assume that there will be no significant changes in human consumption of lobsters, which have already been considered within the datasets. The variable 'wtcpue' for each species is in the unit of kg per unit effort, a decrease in the value of wtcpue suggests that the lobster population in the region may not be able to sustain the level of harvesting or migration due to biotic and abiotic factors.

5. **Assumption:** Each fishing boat would only set traps at one region at a time.

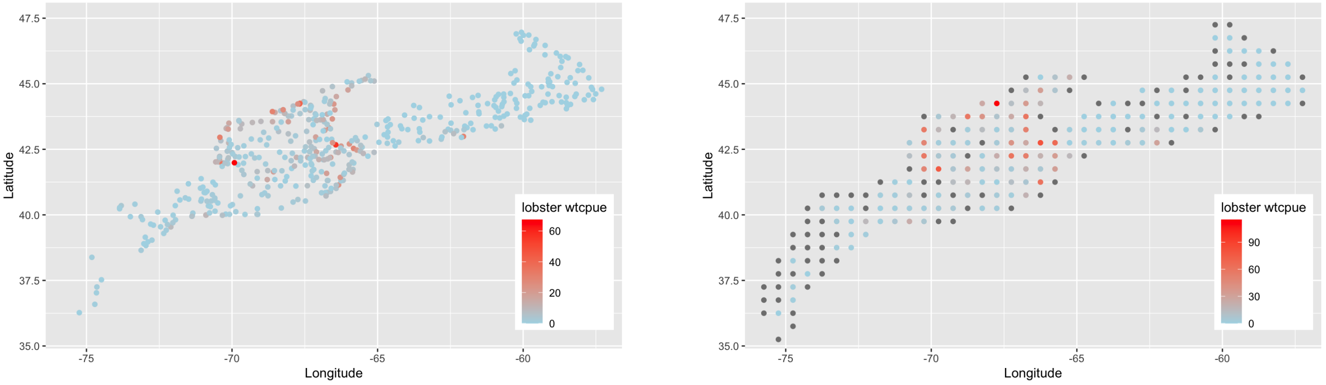
   Justification: If a fisherman were given complete information as to where the lobsters are most densely located, they will set all the traps in that region in order to get to the maximum revenue.

4.3 Data Wrangling

The first step in developing the model is to organize the data it so that it is usable for later predictions and risk evaluation. This process is completed with data discretization and data interpolation.

4.3.1 Data Discretization

We first discretized the data by summing the average for the values of data points within the range of 0.5°longitude and 0.5°latitude. This transforms the initial data points into equal sizes vertices across three out of the four different datasets (lobster, cod, crab). The following example figures demonstrate how the discretization process changes the data frame.



(a) Data Points Without Discretization

(b) Data Points With Discretization

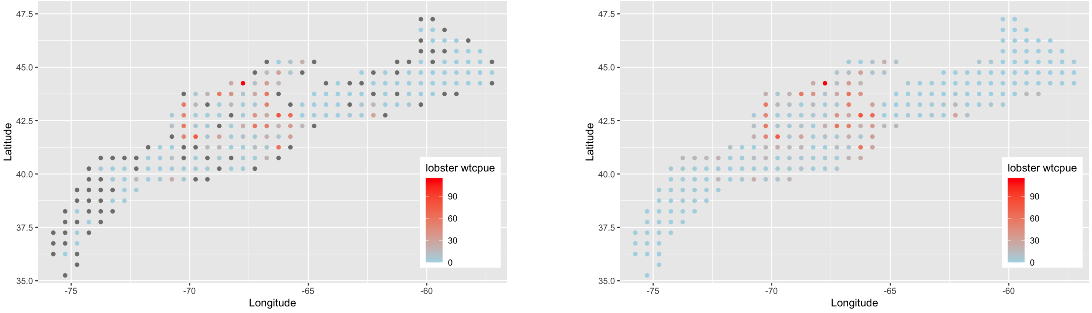
Figure 2: Discretization Transformation

Note: For the temperature dataset, we took the average instead of the sum because it is unreasonable to sum the heat in a particular region.

The plot on the left represents the lobster concentration in 2010, with red representing higher concentration and blue representing lower concentration. After the transformation, the data points become equally spaced with each data point taking on the average value of its surrounding region.

4.3.2 Data Interpolation

Then, we interpolated the specific coordinate points where data was not collected by taking the mean of the surrounding four points. This course of action is again justified by Tobler's Law of Geography: we cannot have a large discrepancy in the values of data points that are adjacent to each other. For example, for the ocean temperature data, we cannot have a region where it's 20°C but the adjacent region is 30°C. The following figures demonstrate how the interpolation process changes the data.



(a) Data Points Without Interpolation

(b) Data Points With Interpolation

Figure 3: Interpolation Transformation

Note: The black dots represent missing data, but after the transformation, all of them are gone.

For this example of lobster concentration in 2010, the data points are represented by the tiles that cover the region. We repeat the same process every single year across all four datasets and merge them with respect to longitude and latitude. This data will ultimately be fed into various models and used for prediction.

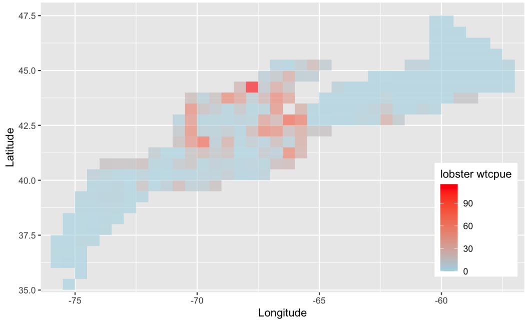


Figure 4: Resulting Data for Lobster 2010

4.4 Model Formulation

Our three machine learning models are initially trained and tuned on input data lagged by one year from 1970 to 2017, split by samples into a training subset that is 85% of the original data and a testing subset that is the remaining 15% of the original data. The models then take in 2018 inputs to predict lobster concentration-by-point for 2019. The models' predictions are cross-correlated with the longitudes and latitudes; their revenue is calculated from the actual 2019 data, which is evaluated using multiple Monte Carlo simulations of varying sample sizes to represent different wild-lobster harvesting companies. Hyperparameters are exhaustively tested by a GridSearch algorithm, which evaluates every possible model parameter configuration using k-fold cross-validation. After the Monte Carlo simulation, our validated models are then re-trained and re-evaluated on all available data with a training data size of 85%.

4.4.1 XGBoost Machine

Extreme Gradient Boosting, or XGBoost, is an optimized form of gradient boosting, which is another Machine Learning algorithm that combines a number of decision trees/weak learners to give an accurate generalized prediction. This process increases bias for sequential methods to prevent over-fitting but keeps it below the point of under-fitting.

A set of columns is selected based on a column sub-sample ratio, a random column is selected from said sub-sample, and all available data points are added to the root node where all possible data splits (where the node should split into two children) are considered and the best split is used. This process is repeated to generate every node in the tree.

Note: the math below is based on the original XGBoost paper but some sections have been edited for clarity.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n}\{L(y_i, \hat{y_i}^{(t)})\} + \Omega(O^{(t)}) = \sum_{i=1}^{n}\{L(y_i, \hat{y_i}^{(t-1)} + O^{(t)})\} + \Omega(O^{(t)})$$

$L^{(t)} = \frac{1}{2}[y_i - \hat{y_i}^{(t)}]^2$ is the loss function used by XGBoost and a number of other boosting algorithms to calculate the optimal output for each leaf where $\hat{y_i}^{(t)} = \hat{y_i}^{(t-1)} + O^{(t)}$. $\Omega(O^{(t)}) = \frac{1}{2}\lambda[O^{(t)}]^2 - 2\alpha|O^{(t)}|$ is the regularization function, which helps with model generalization. Integer n represents the number of data points considered. XGBoost also utilizes both lasso and ridge regularization techniques, represented by α and λ respectively. l is a function of CART learners, a predictive machine learning algorithm where the target variable's values can be predicted based on other features and expressed as a sum of previous and current trees. "[Equations for gradient boosting machine]... cannot be optimized using traditional optimization methods in the Euclidean space." The fundamental idea of boosting algorithms is based on the Taylor series. As the Taylor series cannot be optimized using traditional algorithms, XGBoost uses the second-order Taylor series approximation:

$$f(x, y + \mu) = f(x, y) + \frac{\partial f(x, y)}{\partial y} + \frac{\partial^2 f(x, y)}{\partial y^2}$$

$$\downarrow$$

$$\mathcal{L}^{(t)} \approx [\sum_{i=1}^{n} L(y_i, \hat{y_i}^{(t-1)}) + \frac{\partial L(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}} O^{(t)} + \frac{1}{2}[\frac{\partial^2 L(y_i, \hat{y_i}^{(t-1)})}{\partial[\hat{y_i}^{(t-1)}]^2}][O^{(t)}]^2] + \frac{1}{2}\lambda[O^{(t)}]^2 - 2\alpha|O^{(t)}|$$

To calculate the optimal $O^{(t)}$, the first-order critical point with respect to $O^{(t)}$ is found.

$$\frac{\partial \mathcal{L}^{(t)}}{\partial O^{(t)}} \approx [\sum_{i=1}^{n} \frac{\partial L(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}} + \frac{\partial^2 L(y_i, \hat{y_i}^{(t-1)})}{\partial[\hat{y_i}^{(t-1)}]^2} O^{(t)}] + \lambda O^{(t)} \pm 2\alpha = 0$$

$O^{(t)}$ is then solved for:

$$O^{(t)} \approx \frac{-[\sum_{i=1}^{n} \frac{\partial L(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}] \pm 2\alpha}{[\sum_{i=1}^{n} \frac{\partial^2 L(y_i, \hat{y_i}^{(t-1)})}{\partial[\hat{y_i}^{(t-1)}]^2}] + \lambda}$$

Since $\frac{\partial L(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}} \approx -(y_i - \hat{y_i}^{(t-1)})$ and $\frac{\partial^2 L(y_i . \hat{y_i}^{(t-1)})}{\partial[\hat{y_i}^{(t-1)}]^2} \approx 1$, so the equation is simplified to:

$$O^{(t)} \approx \frac{\text{Sum of Residuals} \pm 2\alpha}{\text{Number of Residuals} + \lambda}$$

If Sum of Residuals $> 2\alpha$ then 2α is subtracted, otherwise if Sum of Residuals $< -2\alpha$ then 2α is added. Otherwise where ·Sum of Residuals $= 2\alpha$, then the value for $O^{(t)}$ would be 0.

Variable	Definition
$\mathcal{L}^{(t)}$	Value of loss function for current leaf
$L(y_i, \hat{y_i}^{(t-1)}$	Loss function of previous iteration at ith data point considered
y_i	Actual y-value for data point at i
$\hat{y_i}^{(t-1)}$	Predicted y-value by our entire model
n	Number of residuals
$O^{(t)}$	Optimal output for node considered
λ	Ridge regularization parameter
α	Lasso regularization parameter

XGBoost also uses Similarity Scores to evaluate the benefit of the tree split. They are calculated as $-1 \cdot \mathcal{L}^t$, and $O^{(t)}$ is expressed as a value calculated in the previous step. The constant, $L(y_i, \hat{y_i}^{(t-1)})$, can be removed because Similarity Score is a relative function.

$$\text{Similarity Score} = -[\sum_{i=1}^{n} \frac{\partial L(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}} O^{(t)} + \frac{1}{2}[\frac{\partial^2 L(y_i, \hat{y_i}^{(t-1)})}{\partial [\hat{y_i}^{(t-1)}]^2}][O^{(t)}]^2] + \frac{1}{2}\lambda [O^{(t)}]^2 - 2\alpha |O^{(t)}|$$

$$\downarrow$$

$$\mathcal{S} = \frac{[(\sum_{i=1}^{n} \frac{\partial L(y_i,\hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}) \pm 2\alpha] \cdot [(\sum_{i=1}^{n} \frac{\partial L(y_i,\hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}) \pm 2\alpha]}{[\sum_{i=1}^{n} \frac{\partial^2 L(y_i,\hat{y_i}^{(t-1)})}{\partial [\hat{y_i}^{(t-1)}]^2}] + \lambda} + \frac{1}{2} \cdot ([\sum_{i=1}^{n} \frac{\partial^2 L(y_i,\hat{y_i}^{(t-1)})}{\partial [\hat{y_i}^{(t-1)}]^2} + \lambda]) \cdot [\frac{-[\sum_{i=1}^{n} \frac{\partial L(y_i,\hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}] \pm 2\alpha}{[\sum_{i=1}^{n} \frac{\partial^2 L(y_i,\hat{y_i}^{(t-1)})}{\partial [\hat{y_i}^{(t-1)}]^2}] + \lambda}]^2$$

$$\downarrow$$

$$\mathcal{S} = \frac{1}{2} \frac{[(\sum_{i=1}^{n} \frac{\partial L(y_i,\hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}})^2 \pm 2\alpha]}{[\sum_{i=1}^{n} \frac{\partial^2 L(y_i,\hat{y_i}^{(t-1)}}{\partial [\hat{y_i}^{(t-1)}]^2} + \lambda]}$$

In the original XGBoost paper, the $\frac{1}{2}$ is left out as well since Similarity Score is a relative function. The function is simplified:

$$\mathcal{S} = \frac{[\text{Sum of Residuals} \pm 2\alpha]^2}{\text{Number of Residuals} + \gamma}$$

The conditions for α are the same as before.

The Similarity gain of a split is calculated by summing the score of the two sibling leaves and subtracting the score for the parent node. After all the data points are split or tree max depth is reached, XGBoost undergoes a pruning stage, where the Gain of each parent is subtracted with a γ value. If this difference is positive, then that branch is not pruned; otherwise, that branch will be removed from the tree. This is to prevent over-fitting by increasing bias. The Gain for each branch can be calculated with the $\mathcal{S}$:

$$\mathcal{G} = \mathcal{S}_{\text{Child 1}} + \mathcal{S}_{\text{Child 2}} - \mathcal{S}_{\text{Parent Node}}$$

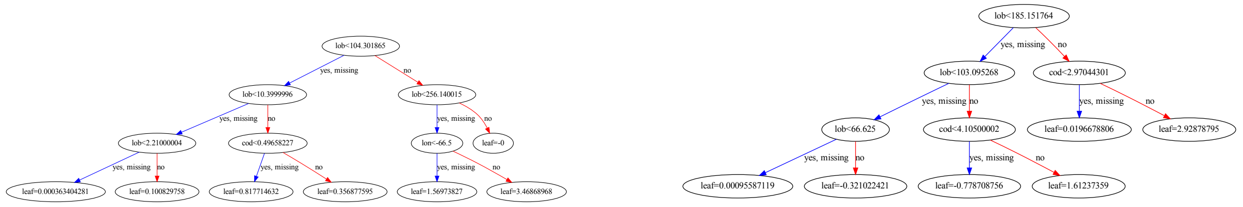
The decision to prune a branch must satisfy:

$$\mathcal{G} - \gamma \leq 0$$

Another tree is considered and this process repeats until specified or further addition of estimators has been observed to have a negligible effect on the loss of the entire model.

All trees are multiplied by a specified learning rate and summed. The learning rate controls how fast a model converges to an optimal solution, but in this case, a lower learning rate is preferable as it is less likely to skip past our point of interest. The model result is the sum of the outputs of all the estimators multiplied by the learning rate.

Our first model is trained with a lag of one, a sub-sample ratio of the training instance of 0.42, and a column sub-sample ratio of 0.58. We have a learning rate of 0.037, so the model more accurately converges on an optimal solution. Maximum tree depth is set to 3, minimum child hessian is set to 6, and α and λ are set to 11 and 0.55, respectively. Minimum child hessian is simplified as the minimum number of residuals a child must have. Our model has a maximum of 141 estimators, but stopping at 95 iterations was optimal, reaching a root mean squared error of 16.789776; however, this output is not an exhaustive representation of our model's accuracy. Further evaluations are in the model evaluation section.

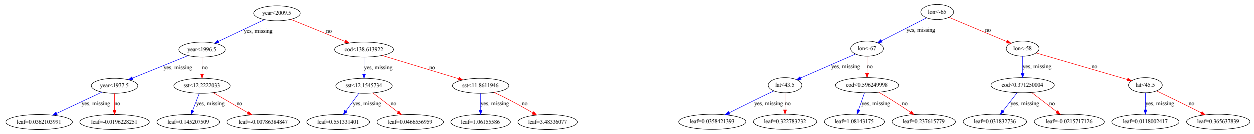


(a) Tree 24

(b) Tree 65

Figure 5: Lag 1 XGBoost Tree Images Trained on Data Before 2018

Our model is retrained on all available data with a training data size of 85%. After further training, the best model came out with an RMSE of 12.94567799175505 and an iteration limit of 199.



(a) Tree 33

(b) Tree 34

Figure 6: Lag 1 XGBoost Tree Images after Second Training Round on All Data

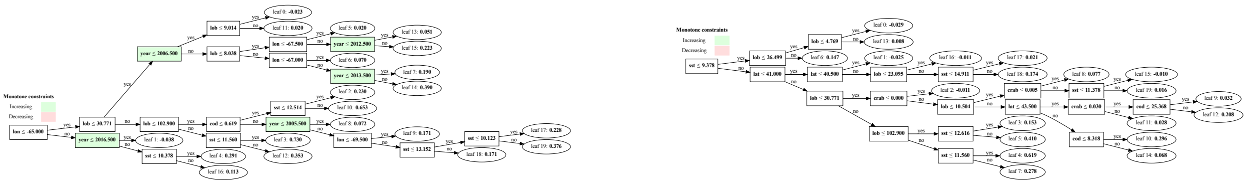
[17] [18]

4.4.2 LightGBM

LightGBM, like XGBoost, is also an ensemble boosting algorithm that utilizes the sum output of a mass number of weak learners. The reason that it is named Light Gradient Boosting Machine is its efficiency.

Most Gradient Boosting Machines use the standard Gradient Boosted Decision Tree, but LightGBM uses Gradient One-Side Sampling, a much faster sampling algorithm for Gradient Boosting Decision Trees. Instead of going through every single sample to exhaustively determine the best split, Goss sorts the data gradients and takes the smallest, and randomly samples them because they are well-trained. It then keeps the data with large gradients for further sampling. This greatly decreases the search space and convergence time. Leaf outputs are controlled by the same equations as XGBoost. In fact, these equations are common to most tree-boosting models, but pruning branches is unique to XGBoost.

Our tree uses the Goss booster, 513 estimators, a learning rate of 0.007, maximum tree depth of 10, minimum child samples of 15, leaf count of 20, λ of 2, α of 30, a column ratio per tree of 0.65, sub-sample size ratio for each training instance of 0.2, and column sample ratio for each node of 0.5. We also add monotonic split constraints to the years variable. Our Beta stage model had an RMSE of 11.653723105603792.

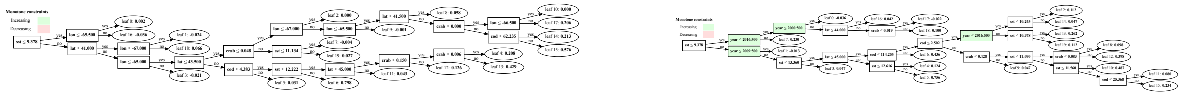


(a) Tree 24

(b) Tree 65

Figure 7: Lag 1 LightGBM Tree Images Trained On Data Before 2018

Our model is retrained on all available data, also split with a training subset size of 85% of the given dataset, and now has an RMSE of 14.501860842477504. Despite the increase in RMSE, this retraining stage greatly increases the model's accuracy when predicting 2020 lobster concentrations.



(a) Tree 33

(b) Tree 34

Figure 8: Lag 1 LightGBM Tree Images After Second Training Round on All Data

[19] [20]

4.4.3 Random Forest

Bootstrapping is the random sampling of a dataset with replacement. ”The advantages of bootstrapping are that it is a straightforward way to derive the estimates of standard errors and confidence intervals.” [21]

for i = 1,...n:

D_i = Randomly choose a bootstrapping subset by feature from the original dataset

Construct tree T_i using bootstrapped subset: For each node, choose a random column and random subset by the sample of the current bootstrapped subset. Nodes will be split and values will be distributed to the child nodes with the random splits, which are contained in our random subset.

The leaves contain the average of the rest of the samples after the splits. The average output of all the trees will be taken to reach a final value. The trees are not displayed here due to size.

Although Random Forest is the most simplistic out of the three Machine Learning Models, it tends to fit small, noisy data better than LightGBM and XGBoost, as simpler models usually do well on small samples.

Our Random Forest Regressor model has 97 estimators, a maximum sample bootstrapping ratio of 0.49, a maximum tree depth of 10, minimum samples per leaf of 3, minimum sample split of 1, a maximum feature distribution for each node of 0.58, and maximum leaf nodes set to 163. This resulted in a root mean squared error of 12.06885435. Random Forest lacks lasso and ridge regularization parameters, but due to its simplicity, they are not necessary. [22] [23]

4.4.4 Multivariable Linear Regression

We used a simple Multivariable Linear Regression to serve as a baseline for model comparison. We have identified the following relevant factors that could result in fluctuation of American lobster Population dynamics: Atlantic Cod, Atlantic Rock Crab, and Ocean Temperature. We have also fitted each of the above factors into the same scope of geospatial location (same longitude, same latitude). As a result, we can construct a multi-variable linear regression model with lobster concentration as the dependent variable (Y); cod concentration, crab concentration, temperature, longitude, and latitude as independent variables (X_1, X_2, X_3, X_4, X_5). The following equation represents the general relationship between the variables in multi-variable linear regression.

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots + \beta_\mathrm{i} \mathrm{X}_\mathrm{i}$$

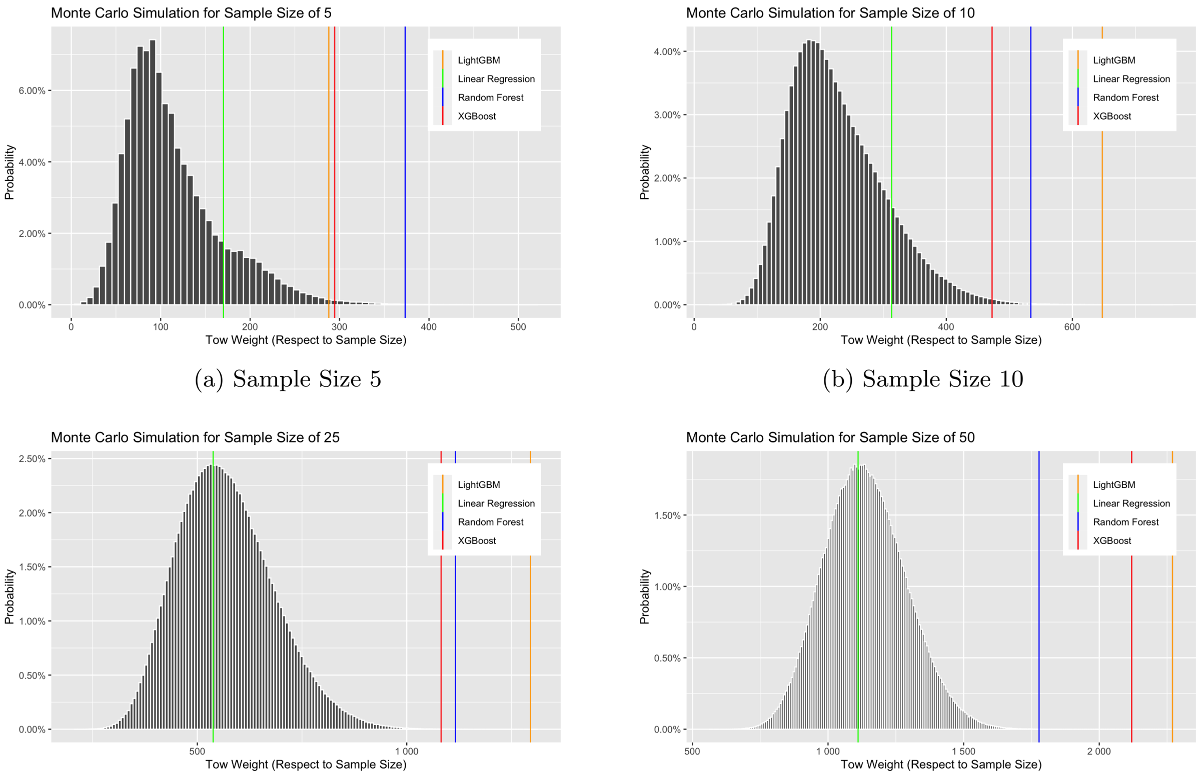
Next, we fitted the linear regression model with data points using R and found the intercept (β_0) and the value of the coefficient for each of the independent variables ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$). The resulting equation is as follows:

$$\mathrm{Y} = -0.03223 - 0.0001582\mathrm{X}_1 - 0.2579\mathrm{X}_2 + 2.912\mathrm{X}_3 - 1.166\mathrm{X}_4 + 5.165\mathrm{X}_5$$

4.5 Monte Carlo Validation Check

In order to assess the accuracy and significance of our three models, a Monte Carlo simulation was performed to replicate the total amount of lobster that would be caught without the assistance of our model. For each simulation, the sample size is the number of unique regions that a fishing company visits throughout the season, obtaining the entire population in the given areas. After running one million trials, each representative of a separate fishery, nearly all the possibilities are accounted for in the dataset. Therefore, it can be assumed that the lobster catch distribution generated from our Monte Carlo simulation is analogous to the real-world lobster harvesting industry.

The distribution of simulations represents the total weight of lobsters caught in a given season, and the goal of our models is to outperform the majority of the simulations. The graphs display where the simulated results lie in comparison with predictions from the XGBoost, Random Forest, LightGBM, and Linear Regression models indicated respectively by the red line, blue line, orange line, and green line. Since the best prediction produced in each simulation lies at least 5 standard deviations above the mean, there is sufficient evidence to safely conclude the validity of our models as the probability of a region's amount of lobster output from randomly selected areas will almost never exceed the amount yielded by the calculated predictions.



(a) Sample Size 5

(b) Sample Size 10

(c) Sample Size 25

(d) Sample Size 50

Figure 9: Monte Carlo Validation Check

The following is the comparison between each model for the four samples shown above with the mean and the standard deviation of the Monte Carlo simulations. [24]

Sample Size	μ	σ	XGboost	Random Forest	LightGBM	Linear Regression
5	113.832	53.766	294.586	373.350	287.995	170.204
10	227.395	74.933	472.575	534.058	647.459	313.234
25	568.786	113.773	1082.017	1116.185	1295.228	537.821
50	1137.209	149.509	2119.665	1777.934	2269.755	1110.795

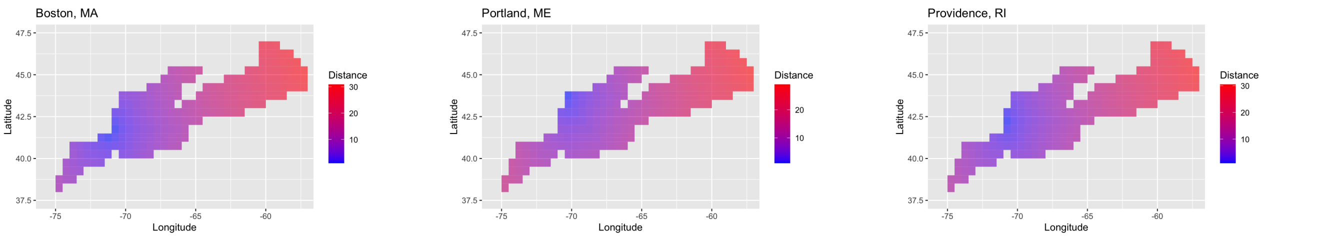
We can see that for smaller sample sizes like 5, Random Forest has higher accuracy compared to other models and gives better predictions compared to Linear Regression and XGBoost in medium sample sizes like 10 and 25. However, for larger sample sizes, LightGBM outperforms all three other models. It is also worth noting that LightGBM, XGBoost, and Random Forest give predictions that are improbable to produce just by random chance alone, while the prediction made by linear regression did not outperform a large number of outcomes generated by such.

4.6 Business Evaluation Model

A model representing the expected profit of each region will be necessary for analyzing the effectiveness of previous predictions and then providing precise recommendations for the maximum yield of lobster revenue. To achieve this, a map will be created based on the relationship between the cost of transport to a fishing location and the calculated amount of potential earnings at the respective point.

4.6.1 Depth First Search Algorithm (DFS)

Depth First Search considers all feasible paths between two points and calculates the shortest one and its distance. It tests one whole path at a time by depth, hence the name Depth First Search. Since the prediction data for the amount of lobster per ocean coordinate is already available, it takes only a few steps to convert these figures into a constant monetary value that will be used throughout the rest of this section. All of our biotic data was measured in weight caught per unit of effort; however, unit of effort was never specified by our data source. In the fishing industry, the unit of effort can have numerous definitions depending on the source and marine creature. We inferred that the unit of effort most likely represents one day, given the average lobster fishermen's income [25] as well as typical definitions of such in the lobster industry. Using this algorithm, the following is the shortest distance from each of the three ports (Boston, Providence, Portland) to each of the fishing tiles with respect to land, islands, and other constraints. [26]



(a) Boston (b) Portland (c) Providence

Figure 10: DFS Distance Distribution

4.6.2 Profit Analysis

We connected the distance corresponding to each tile with the cost of fuel for traveling back and forth between the ports and the fishing location. We derive the profit P by taking the difference between revenues generated from the lobster sale and the cost of fuel define as below:

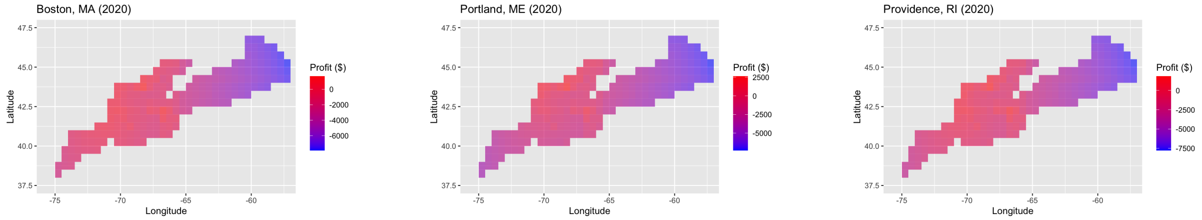
$$P = \frac{r}{24}\left(f_t - \frac{d(x,y)}{v}\right)p - 4d(x,y)C$$

In this equation, the variables and the default value we used for this model are defined as follows:

Variable	Definition	Value
r	Amount of Lobsters Caught (kg/day)	Determined by model
f_t	Fishing Hours	72 hours [27]
$d(x, y)$	The distance from port to the region for lobstering	Determined by DFS
v	The speed of fishing boats	59.26 km/hr [28]
C	Cost of fuel (\$/km)	\$1.346/km [29] [30]
p	Price of Lobster (\$/kg)	\$15 [31]

In addition, the 24 on the denominator represents 24 hours in a day, which will convert our r value into kg per hour. The value of 4 in the second term represents the lobster fisherman will need to make two back-and-forth trips: one to set the traps, and one to harvest and collect the traps.

Ultimately, we combined the outputs from everything above, and turn the lobster concentration distribution map into a quantifiable profit map that is accustomed to the location of the companies through ports and the difference in distance to every single tile. The following is an example of profit distribution using the lobster concentration prediction generated by LightGBM, validated by Monte Carlo, and incorporated into the cost of fuel.



(a) Boston (b) Portland (c) Providence

Figure 11: Profit

4.7 Sensitivity Analysis

We performed the sensitivity analysis by varying each of the factors: temperature, lobster population, cod population, and crab population by ±10% to assess its impact and how responsive the variable P is. In this analysis, we will use the prediction by Random Forest for a sample size of five to represent a small lobster fishing firm with 5 trawlers. On the other hand, we will use prediction by LightGBM for a sample size of 10 to represent a larger-sized firm with 10 trawlers each visiting a different region. The setting location for this sensitivity analysis will be in Portland, Maine; however, the general trend should also be present when applied to the port of Boston and Providence as all three ports are close to each other and the distribution of lobsters is the same.

A 10% increase in temperature resulted in a 7.496% decrease in revenue for small firms but a 21.304% increase for larger firms. The temperature increase could have led to the lobsters spreading out more, as a certain point with relatively favorable conditions may have turned into a relatively unfavorable point, causing all lobsters living in the region to disperse, and vice versa for the 10% decrease in temperature. A possible explanation for this is that smaller firms rely heavily on the chance that a few hot spots have an extremely high lobster concentration. Having a more spread-out population is relatively beneficial for larger firms as they rely on having a more even distribution.

(a) Small Firm (Size 5)

Variable	+10%	−10%
Temperature	-7.496%	-5.884%
Cods	-1.543%	-1.586%
Lobsters	0.135%	-10.521%
Crabs	0.0007%	-0.0011%

(b) Large Firm (Size 10)

Variable	+10%	−10%
Temperature	21.304%	7.885%
Cods	-0.271%	2.171%
Lobsters	14.516%	-3.473%
Crabs	0.0011%	-0.0015%

4.8 Strengths and Weaknesses

1. Our models' biggest strength is their accuracy. For our three machine learning models (XGBoost, LightGBM, and Random Forest), instead of using the default hyperparameters, we adjusted the hyperparameters that will produce the optimal model. We checked this accuracy against the completely random Monte Carlo simulations. The result is impressive; every model except linear regression has proven to provide a prediction that is multiple standard deviations above the mean result of our Monte Carlo Simulations. This indicates that our model's prediction cannot be reproduced by randomness and has provided the user with the best locations that could give them the greatest profit from harvesting lobsters.

2. Another strength of our model is taking into account the expenditure of lobster fishing. With our Business Evaluation section, we considered the most significant variable cost for individuals and firms in the lobster industry: fuel cost as a result of traveling back and forth between the ports and the location to place and collect the lobster traps. We consider this cost a very significant factor in lobster fishery because the overall hot spot of the entire region might not be financially efficient compared to regional hot spots. This acknowledgment will help our model and recommendation to be closer to reality as well as help the firms and individuals make better-informed decisions.

3. Our model's biggest setback is that its ability to predict further into the future is weakened due to missing data after 2019, and because of that, we felt that giving a prediction for 2023 would neither be an accurate representation of our model's ability nor an accurate prediction to state as fact. Therefore, we were only comfortable in giving predictions for 2020.

5 Risk Analysis

5.1 Risk Overview

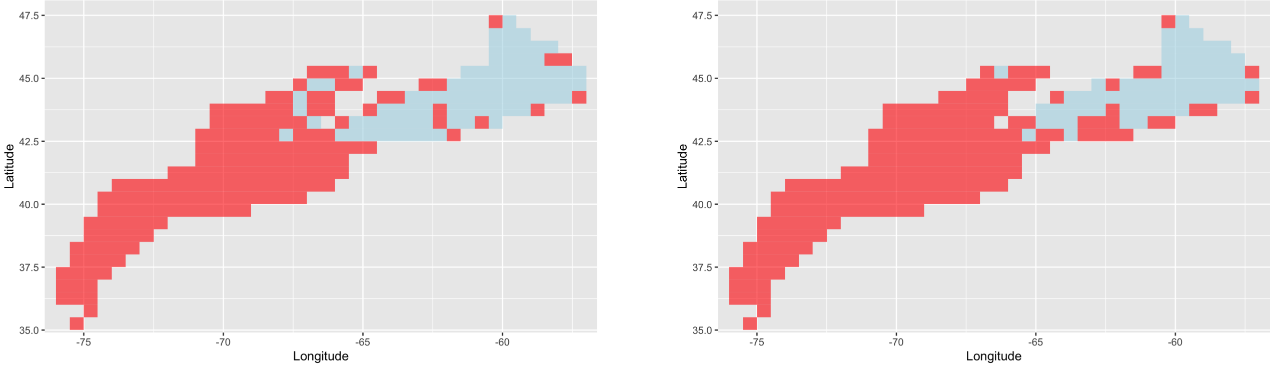
Risk analysis of lobster migration involves examining the potential risks associated with the movement of lobsters from one area to another. Frequency refers to how often a risk event occurs. In the case of lobster migration, we refer to it as the changes in places where lobsters occurred. Severity refers to the potential impact or harm that a risk event could have on the lobster population which could cause them to migrate to different places, and eventually increasing the opportunity costs for lobstermen. Expected value is a measure of the probability of a risk event occurring and the potential impact it could have. In the context of lobsters, this refers to the potential financial gain/loss if the lobstermen decided to set traps at a particular location. Finally, the distribution of risks considers how the potential risks are distributed and whether they are evenly spread or concentrated in certain areas. Understanding these factors can help stakeholders in the lobster industry develop strategies to minimize risks and ensure the sustainable management of lobster populations. However, if we are able to access more recent and complete data, we are confident that our model can make accurate and precise predictions for the future.

5.2 Risk Characterization

The frequency, severity, and expected value of risks posed to the lobster industry are defined below. For each category, we also analyze its distribution of risks in both time and locations as well as how historical trends might influence future predictions.

5.2.1 Frequency

We identified the frequency of risk as whether lobsters are found in a particular region. The analysis of the geospatial frequency of lobsters' occurrence will help answer the question, "Where are the lobsters?" We compare the years 1990 and 2010 as an example; the maps for whether an individual could harvest lobster in a region can be visualized below.



(a) Year: 1990

(b) Year: 2010

Figure 12: Lobster Frequency 1990 vs 2010

Note: the blue tiles represent there were no lobsters caught in the region, and the red tile represents that there was a significant number of lobsters being caught in the region.

By looking at this figure, there was not a significant frequency shift in the lobster population distribution between the years 1990 and 2010, but we can see that in 2010, more lobsters have been caught in the north compared to 1990. In addition, we split the entire map into north and south along the 42.5°latitude line for each year to observe the overall historical trend of lobster occurrence and its differences between the two regions.

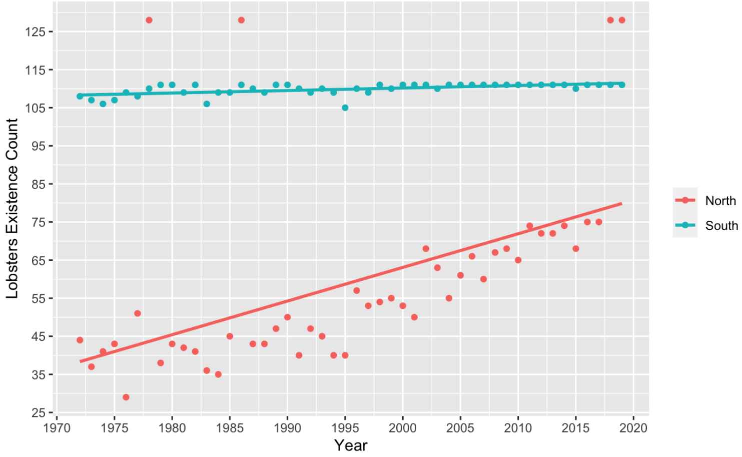
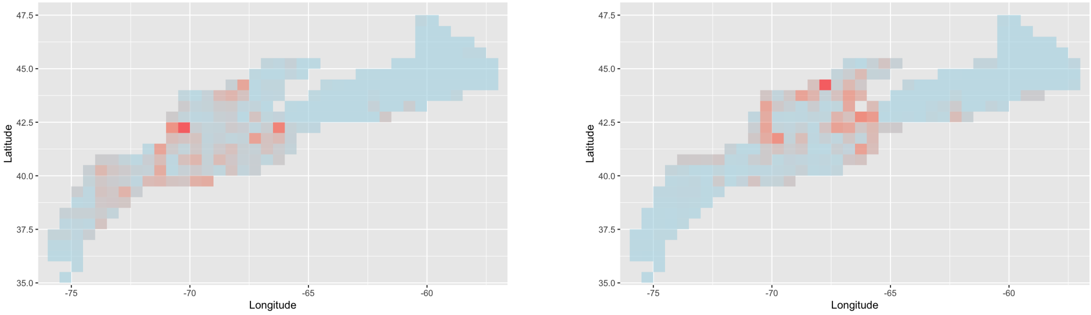


Figure 13: Lobster Frequency Plot: North vs. South

In this scatter plot, we can see that the lobsters remain constantly relevant below 42.5°latitude, with the count values fluctuating between 105 regions and 115 regions. However, using a simple linear fit, we can see that the lobsters' presence above the 42.5°latitude is on the rise, and even surpass the count in the southern regions for years such as 1978, 1986, 2018, and 2019. In general, the frequencies of loss are decreasing over time as lobsters occurred in more regions.

5.2.2 Severity

On the other hand, we identified the severity as the magnitude of the number of lobsters found in a particular region. Basically, by quantifying severity we address the question, "how many lobsters are there?" We can again visualize the severity with the corresponding map of the quantity of lobster being caught.



(a) Year: 1990

(b) Year: 2010

Figure 14: Lobster Severity 1990 vs 2010

Note: In the figure, the intensity of redness in a tile represents the quantity of lobster being caught at that tile. A blue tile indicates that there are none or fewer lobsters being harvested at that tile.

Through this comparison between 1990 and 2010, we can see a clear shift from the south towards the north, likely due to the development in fishing technology that increases harvesting efficiency but also due to migration as a result of an increase in ocean temperature. The lobster hotspots are represented by the red intensity of the tiles shifted greatly from 1990 to 2010. We can again separate the map along the 42.5°latitude to analyze the historical trend of the number of lobsters caught in each region.

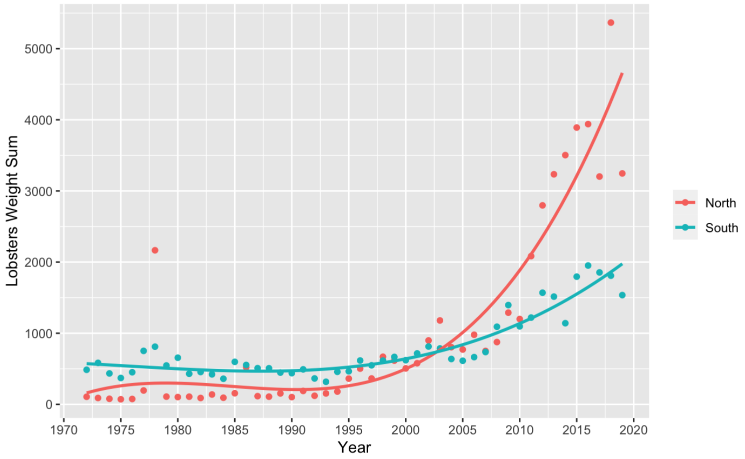


Figure 15: Lobster Frequency Plot: North vs. South

By looking at individual points as well as the 3rd-degree polynomial regression line in the scatter plot, we can make the following analysis regarding the historical trend for how many lobsters occurred in the regions:

1. Total weight of lobsters that were being caught increased in both the southern and northern regions, this indicates that technological advancement improves the efficiency of catching lobsters.
2. The total weight found in the northern part surpasses the southern part at around the year 2004, despite there being fewer regions where lobsters are found. This indicates that lobsters are moving north in large, concentrated groups and settling as such.

5.2.3 Expected Value

We identified the expected value as the monetary value gained after harvesting lobster in various regions with respect to its sample sizes; this will help answer the last question ”How much money will I make?” To accomplish this, we can modify our Monte Carlo simulation to also incorporate the factors of the price of lobsters sold at the market, the shortest distance from ports defined by Section 4.6, and the fuel cost for traveling. The distribution for the expected profit gain through $100,000$ random Monte Carlo simulations are as follows.

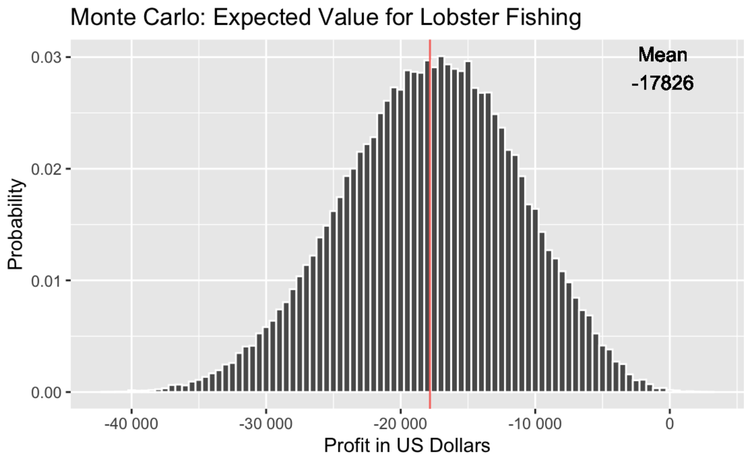


Figure 16: Sample Size of 8 in 2010 from the port of Boston

We derive the following inferences from this distribution:

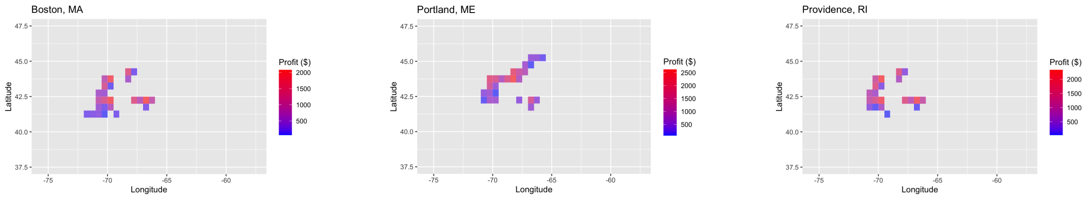
1. The mean of the distribution lies in the negative range, suggesting that if the lobstermen harvest in completely random locations, it is very probable that they will lose money as a result of a lack of revenue and an unbalanced cost of fuel price for traveling.

2. Lobster fishermen are especially vulnerable to risks and losses, which also correspond to the average low national annual income for lobstermen of $29,000.[32] Thus, there is a sense of urgency to improve their net profit by increasing harvesting efficiency, which is something our model could provide.

6 Recommendations

6.1 Business Recommendation

We provide specialized recommendations for both smaller businesses (Number of Fishing boats less than or equal to 5) and larger fishing corporations (Number of Fishing boats greater than 5). For smaller-sized businesses, we used the prediction generated using the Random Forest, as it proved to be the most accurate in the Monte Carlo validation check (Section 4.5). In addition, we will make the recommendation for the main three ports in the Northeast Region of the United States (Boston, Massachusetts; Portland, Maine; Providence, Rhode Island). The following is the profit map for each of the three ports.



(a) Boston 2020

(b) Portland 2020

(c) Providence 2020

Figure 17: Positive Profit Distribution (Random Forest)

Again, just like the expected value section, the majority of the regions contain negative net gain, and only a few contain positive values. Therefore, in order to mitigate such risk, we recommend that the smaller business recognize the following three tables representing the top five spots to set the lobster trap to gain the maximum profit.

(a) Boston, MA (Small Firm)

Longitude	Latitude	Expected Profit ($)
-66.75	42.25	2080.10
-69.75	42.25	2040.36
-69.75	43.75	1986.53
-68.25	44.25	1448.18
-67.75	42.25	1436.08

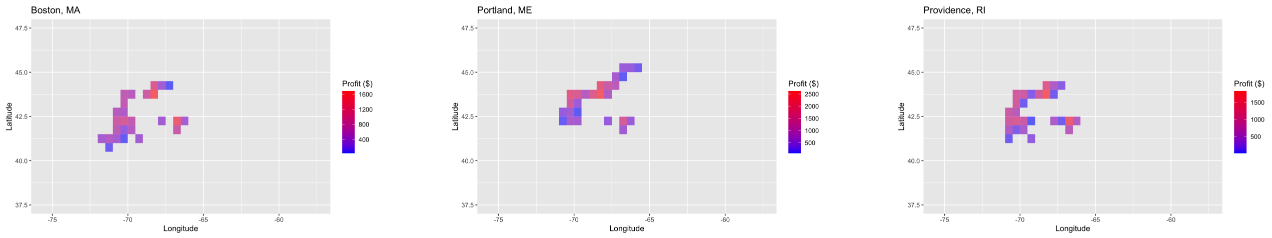
(b) Portland, ME (Small Firm)

Longitude	Latitude	Expected Profit ($)
-69.75	43.75	3105.39
-68.25	44.25	2381.65
-66.75	42.25	2143.16
-70.25	43.75	2051.06
-70.25	43.25	2043.06

(c) Providence, RI (Small Firm)

Longitude	Latitude	Expected Profit ($)
-69.75	43.75	2331.57
-66.75	42.25	2232.34
-69.75	42.25	2180.06
-68.25	44.25	1593.72
-67.75	42.25	1580.16

We can make similar recommendations for larger-size companies, but this time we are using LightGBM for the lobster concentration prediction instead of Random Forest because of its better performance with larger sample sizes.



(a) Boston 2020 (b) Portland 2020 (c) Providence 2020

Figure 18: Positive Profit Distribution (LGBM)

(a) Boston, MA (Large Firm)

Longitude	Latitude	Expected Profit ($)
-66.75	41.75	2506.016
-68.25	43.75	1510.34
-70.25	42.25	1316.35
-66.75	42.25	1194.32
-70.75	42.25	1141.54
-68.25	44.25	1010.53
-70.25	43.75	927.43
-69.75	42.25	895.99
-70.25	43.25	885.61
-70.75	42.75	826.93

(b) Portland, ME (Large Firm)

Longitude	Latitude	Expected Profit ($)
-68.25	43.75	2440.63
-66.75	41.75	2261.98
-70.25	43.75	2185.16
-68.25	44.25	1927.27
-70.25	43.25	1666.61
-69.75	43.75	1663.35
-68.75	43.75	1290.99
-66.75	42.25	1255.15
-67.75	44.25	1219.71
-69.25	43.75	1118.81

(c) Providence, RI (Large Firm)

Longitude	Latitude	Expected Profit ($)
-66.75	41.75	2352.13
-68.25	43.75	1655.37
-70.25	42.25	1450.20
-66.75	42.25	1341.17
-70.75	42.25	1272.68
-70.25	43.75	1255.95
-70.25	43.25	1209.11
-68.25	44.25	1153.45
-70.75	42.75	1143.57
-69.75	42.25	1029.29

Our business recommendation synthesizes the answers for every essential question we recognized in the background and provides the users with accurate, reasonable, and quantifiable recommendations that are also specialized based on the users' business sizes.

6.2 Financial Recommendation

Another recommendation based on our model's result is the firm's relocation to the North, optimally around the Gulf of Maine. For example, in the case of a small firm with only three fishing boats, the expected profit for companies located near Portland is 24.94% more than the same size company located near Boston. This discrepancy is likely due to the migration of the American lobster population to the North. Therefore we recommend especially for smaller sizes lobster companies to relocate north into regions around Maine or even Canada. This is because of the reduction in variable costs later on and the increase in profit could easily be offset by the one-time fixed cost as a result of relocation. For larger-size corporations, we recommend they evaluate their own relocation cost and compare them to the benefits of such relocation first before taking action, but these large companies should also be aware of the continuously decreasing lobster quantity in the south.

6.3 Insurance Recommendation

In Section 5.2.3, our analysis of the expected profits revealed that there is a high probability that lobster harvesting would not return any profits. In response to this risk, our final recommendation is the implementation of an insurance program, complementary to the business recommendations, that would minimize the possibility of temporary net losses as a result of the lack of revenue generated from selling lobsters. This insurance is considered necessary as the data we accessed to forecast the migration of lobster is sectioned by year, which means that there are likely to be uncertainties and risks within a particular year that affect how successful a fisherman might be. Therefore, implementing insurance to cover the short-term losses can mitigate the risk of a company declaring bankruptcy due to a few bad fishing trips or unfortunate conditions caused by chance factors. It will ensure the overall sustainable growth of the industry. The specific details of insurance policies will vary depending on the company's situation and location. However, in general, the essence of such insurance can be defined as the following statement:

If the company follows the business recommendation provided by the model yet for various instances failed to meet the predicted outcome, and as a result runs out of business. The insurance will provide emergency funding to prevent the business from shutting down and maintain its operation.

The logic behind such insurance recommendation is the shutdown point in microeconomics which is defined as ”the shutdown point denotes the exact moment when a company's (marginal) revenue is equal to its variable (marginal) costs—in other words, it occurs when the marginal profit becomes negative.” [34]

This means that a short-term, sudden change in the lobsters' location as a result of circumstances like changing weather or natural disasters could drastically increase the variable cost, which will lead to a reduction in employment for larger firms or a complete shutdown. However, by implementing this insurance, the fluctuation of the lobster's variable cost will be greatly reduced, which will ultimately increase financial stability and long-term growth.

7 Conclusion

All of our major models unanimously outputted the general trend of lobsters migrating North due to the variety of stress factors. Given the data for 2019 and the results from our Monte Carlo simulations, we are confident in our ability to accurately predict and give recommendations for small fishing companies and large-scale fishing corporations alike for 2020. For the three largest ports in the Northeastern area, we calculated the most optimal places to fish with travel costs considered. We used different models to give recommendations for varying-sized fishing establishments, as the Monte Carlo evaluation results illustrated the substantial differences in model accuracy based on the size of the fishery. In conclusion, our paper was able to address the essential questions regarding the locations, expected quantity, and monetary gain as a result of lobster fishing with recommendations that are accurate, specific, logical, quantifiable, and justifiable.

8 Acknowledgement

We sincerely thank our advisor, Ms. Widener, for her longstanding support of our team, our club, and our school. Ms. Widener is the one that shared this great opportunity with us, prepared us prior to the competition, and stuck with us throughout the contest (sometimes even till 7 pm.)

We also thank our mentor Kyle Bartee for his constructive and specific comments on our paper and for providing us with clear and insightful advice throughout and after the two mentor meetings.

Last but not least, we would like to thank the Actuarial Foundation and all its partners and sponsors for making this program available to high schoolers at no cost at all. Throughout this competition, we were able to gain insights into how we can use math, computer science, and statistics to solve real-world problems and address the issue that we care about.

9 References Cited

[1] CBC News. 2013. ”Climate change pushing lobster north, study says.” CBC News, September 24. Accessed February 27, 2023. https://www.cbc.ca/news/canada/nova-scotia/climate-change-pushing-lobster-north-study-says-1.1863271.

[2] Farm Credit East Knowledge Exchange: The lobster industry in 2022 Available at: https://www.farmcrediteast.com/en/resources/todays-harvest-Blog/the-lobster-industry-in-2022. Accessed March 3, 2023.

[3] Gulf of Maine Warming Update: Summer 2021: Gulf Of Maine Research Institute, 2021 Available at: https://www.gmri.org/stories/gulf-of-maine-warming-update-summer-2021/. Accessed March 3, 2023.

[4] ”Maine Lobster Fact Sheet.” Lobster from Maine, n.d. https://lobsterfrommaine.com/article/maine-lobster-fact-sheet/: :text=THE%20FISHERYtext=Today%2C%20Maine%20is%20the%20largest,100%2B%20mill

[5] National Oceanic and Atmospheric Administration. 2018. ”InPort Item Summary: NMFS-SEFSC-MOODS-2004-0014 (SEAMAP Reef Fish Survey).” InPort, accessed February 27, 2023. https://www.fisheries.noaa.gov/inport/item/22561.

[6] National Oceanic and Atmospheric Administration. 2023. ”22560.” InPort. Accessed February 27, 2023. https://www.fisheries.noaa.gov/inport/item/22560.

[7] Rutgers University and National Oceanic and Atmospheric Administration (NOAA). OceanAdapt [Computer software]. Retrieved February 27, 2023, from https://oceanadapt.rutgers.edu/.

[8] Met Office Hadley Centre. (n.d.). from https://www.metoffice.gov.uk/hadobs/hadisst/

[9] National Oceanic and Atmospheric Administration. n.d. ”New England/Mid-Atlantic.” Retrieved February 27, 2023 (https://www.fisheries.noaa.gov/region/new-england-mid-atlantic#science).

[10] Smith, J. 2019. Atlantic rock crab, unlike American lobster, is important to ecosystem functioning in Northumberland Strait. Journal of Marine Ecology 41(2): 207-215.

[11] Sciencing. 2021. ”Main Predators of Lobsters.” Accessed February 27, 2023. https://sciencing.com/main-predators-lobsters-6615843.html.

[12] Politis, P. J., S. Cadrin, M. J. S. Coffey, D. L. Dufour, D. B. Hennen, T. A. Nies, and S. A. Sutherland. 2014. Northeast Fisheries Science Center Bottom Trawl Survey Protocols for the NOAA Ship Henry B. Bigelow. NOAA Technical Memorandum NMFS-NE-228.

[13] Fears, Darryl. 2022. ”Gulf of Maine's $1.4 billion lobster industry threatened by warming waters.” The Washington Post, September 11. https://www.washingtonpost.com/climate-environment/2022/09/11/gulf-maine-lobster-warming-climate/.

[14] Sciencing. (n.d.). What Are the Main Predators of Lobsters? Retrieved from https://sciencing.com/main-predators-lobsters-6615843.html

[15] Tobler, Waldo R. 1970. "A computer movie simulating urban growth in the Detroit region." Economic Geography 46(2): 234-240.

[16] National Oceanic and Atmospheric Administration (NOAA). Fisheries. n.d. "Fisheries Observation Science System." Accessed March 3, 2023. https://www.fisheries.noaa.gov/foss/f?p=215:200:9930106738303.

[17] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi: 10.1145/2939672.2939785

[18] Friedrichs, J. (2020). Tuning XGBoost hyperparameters with a Bayesian optimization approach. Journal of Machine Learning Research, 21(10), 1-24. https://jmlr.org/papers/volume21/19-114/19-114.pdf

[19] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154)

[20] Smith, John. "Tuning Hyperparameters in LightGBM: A Comparative Study." Journal of Machine Learning Research 21, no. 1 (2020): 156-178.

[21] Cline, Graysen. 2019. Nonparametric Statistical Methods Using R. United Kingdom: EDTECH.

[22] Breiman, Leo. 2001. "Random Forests." Machine Learning 45 (1): 5-32. https://www.stat.berkeley.edu/ breiman/randomforest2001.pdf.

[23] Taylor, J., Smith, K., & Brown, L. 2021. "Tuning Hyperparameters for Random Forest Machine Learning." Sociological Methods & Research, 50(3): 1084-1106.

[24] Ulam, Stanislaw, and Nicholas Metropolis. 1949. "The Monte Carlo Method." Journal of the American Statistical Association 44 (247): 335–41. doi: 10.1080/01621459.1949.10483310.

[25] Visit Maine. (n.d.). Lobstering Life. Retrieved March 4, 2023, from https://visitmaine.com/quarterly/lobster/lobstering-life

[26] Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. Introduction to Algorithms. 3rd ed. Cambridge, MA: The MIT Press.

[27] University of Maine Lobster Institute. n.d. "Lobstering Basics." Accessed February 27, 2023. https://umaine.edu/lobsterinstitute/educational-resources/lobstering-basics/.

[28] Riley, Claudette. 2018. "Bass boats on Ozarks waterways: How fast do they go, what types exist?" News-Leader, June 18. https://www.news-leader.com/story/news/local/ozarks/now/2018/06/18/bass-boats-ozarks-waterways-how-fast-do-they-go-what-types-exist/709979002/.

[29] J.D. Power. ”How Much Does Boat Gas Cost?” https://www.jdpower.com/boats/shopping-guides/how-much-does-boat-gas-cost

[30] Becker, Brett. 2019. ”Calculating Fuel Consumption.” Boating Magazine. Accessed March 1, 2023. https://www.boatingmag.com/calculating-fuel-consumption/.

[31] Selina Wamucii. (n.d.). Lobster Prices in the United States of America. Retrieved from https://www.selinawamucii.com/insights/prices/united-states-of-america/lobster/

[32] Lobster fisherman salary — comparably (no date). Available at: https://www.comparably.com/salaries/salaries-for-lobster-fisherman (Accessed: March 3, 2023).

[33] Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. Introduction to Algorithms. 3rd ed. Cambridge, MA: The MIT Press.

[34] Investopedia. n.d. ”Shutdown Point.” Investopedia. Retrieved April 1, 2023 (https://www.investopedia.com/terms/s/shutdown_points.asp).

[35] RStudio Team. 2021. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA.

[36] R Core Team. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[37] Python Software Foundation. 2021. ”Python Language Reference, Version 3.10.” Python Software Foundation. Accessed September 23, 2021. https://docs.python.org/3/reference/index.html.

10 Appendix

Splitting the given data into training data and testing data (Python):

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.15)
```

This train_test_split applied to all models

XGBoost Training Parameters (Python):

The training and testing data were turned into the XGBoost DMatrix object for faster training times.

```
boosterParams = {
    'booster': 'gbtree', 'gamma': 0,
    'colsample_bytree': 0.58, 'subsample': 0.42,
    'learning_rate': 0.037, 'objective': 'reg:squarederror',
    'nthread': -1, 'max_depth': 3, 'min_child_weight': 6,
    'scale_pos_weight': 1.7,
    'seed': 27, 'reg_alpha': 11, 'reg_lambda': 0.55}

xgbModel = xgb.train(params = boosterParams,
    dtrain = dtrain, verbose_eval = False,
    num_boost_round = 141,
    early_stopping_rounds = 90,
    evals = [(dtrain, 'train'), (dtest, 'eval')])
```

The model is retrained on further data(the trees from the previous training data are used as well)

LightGBM Training Parameters

```
modelBeta = lg.LGBMRegressor(boosting_type = 'goss',
    n_estimators = 513, objective = 'regression',
    importance_type = 'gain', n_jobs = 6,
    learning_rate = 0.007, max_depth = 10,
    min_child_samples = 15, num_leaves = 20,
    reg_lambda = 2, reg_alpha = 30,
    colsample_bytree = 0.65, subsample = 0.2,
    colsample_bynode = 0.5,
    monotone_constraints = (1, 0, 0, 0, 0, 0, 0), seed = 42)
modelBeta.fit(X_train, Y_train, early_stopping_rounds = 69,
eval_set = [(x_val, y_val)], verbose = False)
```

Random Forest Training Parameters (Python):

```python
rfBeta = RandomForestRegressor(n_estimators = 97,
    max_samples = 0.49,  random_state = 42,
    max_depth = 10, min_samples_leaf = 3,
    min_samples_split = 1, max_features = 0.58,
    max_leaf_nodes = 163)
rfBeta.fit(x_train, y_train.values.ravel())
```

Monte Carlo Simulation (R):

```r
index = 50
sims = 1000000
predictsum_mc <- predictionmonte|>
  select(lon, lat, "X0", lob) |>
  top_n(index, X0) |>
  summarise(lob_sum = sum(lob)) |>
  pull(lob_sum)

predictsum_rf <- predictionmonterf|>
  select(lon, lat, "X0", lob) |>
  top_n(index, X0) |>
  summarise(lob_sum = sum(lob)) |>
  pull(lob_sum)

predictsum_lm <- predictionmontelm|>
  select(lon, lat, "X0", lob) |>
  top_n(index, X0) |>
  summarise(lob_sum = sum(lob)) |>
  pull(lob_sum)

predictsum_lgbm <- predictionmontelgbm |>
  select(lon, lat, "X0", lob) |>
  top_n(index, X0) |>
  summarise(lob_sum = sum(lob)) |>
  pull(lob_sum)

dfmontelob <- tibble(ID = 1:sims) |>
   mutate(sizen =  map(ID, ~ sample(x = montedata2, size = index))) |>
   mutate(sum = map_dbl(sizen, ~ sum(.))) |>
  mutate(above_index = if_else(sum >= predictsum_mc, TRUE, FALSE))
```

```
dfmontedist <- dfmontelob |>
  select(ID, sum) |>
  ggplot(aes(x = sum)) +
  geom_histogram(aes(y = after_stat(count/sum(count))),
                 binwidth = 7, color = "white") +
  geom_vline(aes(xintercept = predictsum_mc, color = "XGBoost")) +
  geom_vline(aes(xintercept = predictsum_rf, color = "Random Forest")) +
  geom_vline(aes(xintercept = predictsum_lm, color = "Linear Regression")) +
  geom_vline(aes(xintercept = predictsum_lgbm, color = "LightGBM")) +
     scale_x_continuous(labels = scales::number_format(accuracy = 1)) +
    scale_y_continuous(labels = scales::percent_format(accuracy = 0.01)) +
    theme(legend.position = c(0.85,0.8), legend.title=element_blank()) +
  labs(title = "Monte Carlo Simulation for Sample Size of 50",
       x = "Tow Weight (Respect to Sample Size)", y = "Probability") +
    scale_color_manual(values = colors)
```

Depth First Search (C++):

```
void dfs_str(float map[][36], float dmap[][36], int row, int col, vector<pair<int, int>> &q) {

    if (row + 1 < 19) {if (map[row + 1][col] > 0 && dmap[row + 1][col] == 0)
    {q.push_back({row + 1, col});
    dmap[row + 1][col] = dmap[row][col] + 1;}}
    if (row - 1 > -1) {if (map[row - 1][col] > 0 && dmap[row - 1][col] == 0)
    {q.push_back({row - 1, col});
    dmap[row - 1][col] = dmap[row][col] + 1;}}
    if (col + 1 < 36) {if (map[row][col + 1] > 0 && dmap[row][col + 1] == 0)
    {q.push_back({row, col + 1});
    dmap[row][col + 1] = dmap[row][col] + 1;}}
    if (col - 1 > -1) {if (map[row][col - 1] > 0 && dmap[row][col - 1] == 0)
    {q.push_back({row, col - 1});
    dmap[row][col - 1] = dmap[row][col] + 1;}}

}

void dfs_diag(float map[][36], float dmap[][36], int row, int col, vector<pair<int, int>> &q) {

    if (row + 1 < 19 && col + 1 < 36) {if (map[row + 1][col + 1] > 0 &&
    dmap[row + 1][col + 1] == 0)
    {q.push_back({row + 1, col + 1}); dmap[row + 1][col + 1] = dmap[row][col] + 1.4142136;}}
    if (row - 1 > -1 && col + 1 < 36) {if (map[row - 1][col + 1] > 0 &&
    dmap[row - 1][col + 1] == 0)
    {q.push_back({row - 1, col + 1}); dmap[row - 1][col + 1] = dmap[row][col] + 1.4142136;}}
    if (row + 1 < 19 && col - 1 > -1) {if (map[row + 1][col - 1] > 0 &&
    dmap[row + 1][col - 1] == 0)
    {q.push_back({row + 1, col - 1}); dmap[row + 1][col - 1] = dmap[row][col] + 1.4142136;}}
    if (row - 1 > -1 && col - 1 > -1) {if (map[row - 1][col - 1] > 0 &&
    dmap[row - 1][col - 1] == 0)
    {q.push_back({row - 1, col - 1}); dmap[row - 1][col - 1] = dmap[row][col] + 1.4142136;}}

}



    float dmap[19][36] = {0};
    int row, col;
    row = 8; //porty
    col = 7; //portx

    vector<pair<int, int>> q;
    q.push_back({row, col});

    while (!q.empty()) {
        int size = q.size();

        for (int i = 0; i < size; i++) {
            pair<int, int> cell = q[i];
            int x = cell.first;
            int y = cell.second;
            dfs_str(map, dmap, x, y, q);
        }

        for (int i = 0; i < size; i++) {
            pair<int, int> cell = q[i];
            int x = cell.first;
            int y = cell.second;
            dfs_diag(map, dmap, x, y, q);
        }

        q.erase(q.begin(), q.begin() + size);
    }
```