Enhancing Fisheries Research Insights through Big Data Analytics

Fish Catch Frequency Prediction Project

Project Final Report

Bhushan Shelke and Hasaranga Jayathilake

FA23-IN-INFO H516 Applied Cloud Computing for Data Intensive Sciences

Luddy School of Informatics, Computing, and Engineering

Indiana University Indianapolis

Contents

Table of Figures and Tables.

Figures

Tables

Introduction

The importance of Alaska's fishing sector in the context of the United States cannot be understated. This industry is the principal source of the country's wild salmon and other valued seafood items. However, its potential is limited by difficult weather circumstances, which impose a narrow window for open-sea fishing owing to cold wintery temperatures (Sergeant *et al.*, 2022). The short timescale emphasizes the importance of strategic management to maximize returns. Using machine learning regression analysis appears to be a critical tool in this effort. This method optimizes fish capture counts during the limited fishing season by using the power of data-driven insights. As a result, the use of cutting-edge analytical techniques becomes essential, increasing sustainability and efficiency in the Alaskan fishing business.

Problem description

The AFSC/ABL: Longline Sablefish Survey data is a valuable source of information for the Alaskan fishing industry, as it provides insights on the abundance and distribution of various fish species in the region. However, the data is not fully utilized for predicting the future fish catch frequency, which is a key indicator of the productivity and profitability of fishing operations (Karp *et al.*, 2018). Currently, the prediction is based on time series analysis, which only considers the historical data and ignores other factors that may affect the fish catch frequency, such as environmental conditions, fishing methods, and fish behavior (Siwicke and Malecha, 2023). Therefore, there is a need to develop a more comprehensive and accurate predictive model that can incorporate multiple variables and capture the complex relationships among them. Such a model can help fishermen and managers to plan and optimize their fishing activities, as well as to monitor and conserve the fish resources (Cline *et al.*, 2017).

Based on that, objective of this project as, identify the best regression analysis-based model that can predict the quantity of fish per each fish type can be caught per hour or CPUE[1] to understand which area are more productive than others to preform fishing operations in the Alaskan Waters.

To achieve that main objective, create the 3 research questions as below.

- Which regression method provides more accurate predictions compared to other regression models?
- Which method provides better prediction, among regression models or time series?
- What are the most important variables that are useful for building the regression model for predicting the number of fish caught per hour?

Description of the data (with source)

The data for this project was extracted from the NMFS-AFSC Longline Sablefish Survey, which is owned and conducted by the Alaska Fisheries Science Center (AFSC) and the National Oceanic and Atmospheric Administration (NOAA). The survey results are available as CSV files in an open access repository hosted by Google Docs for NOAA (National Oceanic and Atmospheric Administration, 2023); (Appendix 2) provides the link to download or view the original dataset that taken for this study. Under appendix 3, provide the survey geographic coverage.

[1] CPUE – Catch per unit effort, expressed as number of fish per skate, also referred to as catch rate.

The purpose of this database is to provide information on the abundance and biology of sablefish and other groundfish species living in Alaskan waters, which are used for stock assessment and quota setting. The survey data is longitudinal, as it covers a period from 1987 to 2017, and consists of 50 schemas, comprising 1.1 million tuples.

Methodology

Data Cleaning / Wrangling

Even though the dataset consists of 50 schemas, comprising 1.1 million tuples, not all the data is relevant and consistent for the purpose of this project. Therefore, some data cleaning and wrangling steps are performed to prepare the data for further analysis and modeling. These steps are:

Drop Columns with >50% Nulls: Some columns in the dataset have significant missing data, which may affect the data integrity and lead to biased analyses (Deng and Yu, 2013; Marcelino *et al.*, 2022). Therefore, these columns are dropped from the dataset. This reduces the number of columns from 50 to 35, and the number of tuples from 1.1 million to 0.9 million.

Change Numerical Datatypes to Floats: Some numerical columns in the dataset are not of the float type, which may cause problems in the modeling process, such as rounding errors and data loss (Jadhav *et al.*, 2019). Therefore, these columns are converted to the float type, ensuring accurate and consistent data representations (Marcelino *et al.*, 2022). This affects 10 columns in the dataset, such as latitude, longitude, depth, and frequency.

Technology Used

The project involved the use of various technologies and tools for data analysis and modeling, such as:

- **PySpark:** Utilized for large-scale data processing, PySpark facilitated CSV file handling and distributed data exploration, feature engineering, and machine learning (Ali and Iqbal, 2022).
- **Pandas and NumPy:** Leveraged for data manipulation and numerical operations, Pandas handled tabular data structures, while NumPy facilitated mathematical and statistical calculations (Khatami and Frantz, 2023).
- **Scikit-Learn:** Employed for implementing and comparing regression models, Scikit-Learn offered metrics and tools for model evaluation, including mean squared error and cross-validation (Tran *et al.*, 2022).
- **Matplotlib:** Utilized for data visualization, Matplotlib enabled the creation of various plots and charts, providing customization options for appearance and style (Chandra and Dwivedi, 2022).

Libraries Used

- **pyspark.sql:** Central in Spark, this library manages structured data, aiding data manipulation and transformation in distributed collections efficiently (Belcastro *et al.*, 2022).
- **pyspark.ml:** Hosts machine learning algorithms, particularly regression models, streamlining model development within the PySpark ecosystem (Chhabra *et al.*, 2023).
- **pyspark.mllib:** A comprehensive library, it encompasses basic statistics, classification, regression, clustering, and collaborative filtering functions, enhancing data analysis versatility (Belcastro *et al.*, 2022).

- **sklearn:** Essential for model selection and evaluation, scikit-learn is a powerful Python package contributing to efficient machine learning model implementation and assessment (Chhabra *et al.*, 2023).

Statistical Methods Used to Train the Models:

- **Random Forest Regression**: Executed via PySpark's 'RandomForestRegressor,' this ensemble learning technique enhances prediction accuracy through multiple decision trees, refining model robustness (Zhou, 2023).
- **Linear Regression**: Utilized the PySpark 'LinearRegression' module, enabling straightforward modeling of linear relationships between variables, crucial for predictive analysis (Xing *et al.*, 2023).
- **Generalized Linear Regression**: Leveraged PySpark's 'GeneralizedLinearRegression' to extend traditional linear regression, accommodating diverse distribution families for better model adaptability (Singh, 2019).
- **Statistical Correlation Analysis**: Applied for scrutinizing variable relationships, this method examines connections between different features, providing insights crucial for model development and interpretation (Xing *et al.*, 2023).
- **Gradient-Boosted Trees Regression Model**: Implemented through PySpark's 'GBTRegressor,' this model iteratively builds decision trees, correcting errors from prior trees, resulting in a powerful, accurate ensemble for predicting fish catch frequencies (Singh, 2019).
- **Time Series Analysis**: Integral to forecasting, time series analysis explores sequential data points. In this project, it aids in understanding and predicting trends in fish catch frequencies over the annual longitudinal survey period (Zhou, 2023).

Model Validation

For each model, validation has been conducted using a standard train-test split, which utilizing the PySpark function 'randomSplit([0.8, 0.2])' to allocate 80% for training and 20% for testing (Testas, 2023).

Results

Under the project, developed 4 regression models based on random forest, linear, generalized linear and gradient-boosted trees respectively. also, used the time series analysis, to answer research questions that were considered under this project.

Research Question 1:

As shown in Table 1, the Gradient-Boosted Trees Regression Model outperforms the other regression models in terms of accuracy, as it has the lowest root mean squared value. Therefore, this project answers research question 1 by concluding that the Gradient-Boosted tree is a suitable model for predicting the frequency of fish type catches per hour.

Table 1:Root Mean Squared Error of each regression model.

Regression Model Type	Root Mean Squared Error (RMSE)
Gradient-Boosted Trees	5.92570019
Linear	6.13841615
Random Forest	5.95058604

Generalized Linear	5.98958186

Apart from that, under the Appendix 4, demonstrate the outputs of each model for calculating mean squared error, mean absolute error and r-squared values under each management area.

Research Question 2:

Comparison has been conducted among the gradient-boosted trees and the time-series analysis on prediction of CPUE on each year as follows in comparison to the actual data which may help to understand the difference. (TS = Time serious, GBT - Gradient-Boosted Trees, Act. – Actual figures*)

Table 2: Comparison of prediction CPUE of time series & gradient-boosted trees model with actual data.

Location (Management Areas)	2018			2019			2020			2021			2022		
	TS	GBT	Act	TS	GBT	Act	TS	GBT	Act	TS	GBT	Act	TS	GBT	Act
West Yakutat	4.6	**4.4**	**4.2**	4.7	**4.4**	**4.3**	4.7	**4.4**	**3.9**	4.7	**4.4**	**4.1**	4.8	**4.4**	**4.4**
Bering Sea	4.4	**4.8**	**4.8**	4.4	**4.8**	**4.7**	4.4	**4.8**	**4.6**	4.4	**4.8**	**4.6**	4.5	**4.8**	**4.7**
Aleutians	4.8	**4.7**	**4.7**	4.8	**4.6**	**4.6**	4.8	**4.6**	**4.5**	4.8	**4.7**	**4.4**	4.8	**4.7**	**4.7**
East Yakutat/Southeast	3.6	**3.8**	**3.8**	3.8	**3.7**	**3.6**	3.6	**3.7**	**3.5**	3.8	**3.7**	**3.5**	3.6	**3.8**	**3.8**
Western Gulf of Alaska	6.6	**6.6**	**6.6**	6.6	**6.6**	**6.6**	6.6	**6.6**	**6.6**	6.5	**6.6**	**6.3**	6.6	**6.6**	**6.6**
Central Gulf of Alaska	4.6	**4.7**	**4.7**	4.8	**4.7**	**4.7**	4.6	**4.7**	**4.5**	4.8	**4.7**	**4.4**	4.6	**4.7**	**4.7**

*Actual figures: gathered from Alaska Fisheries Science Center. (2023). Longline Sablefish Survey Results.

Table 2, shows that, except for 2020 and 2021 (which are affected by the covid pandemic and cause data anomalies), the predictions of the gradient-boosted trees model for the other three years are almost equal to the actual results, compared to the time series prediction that is currently used in practice. This suggests that the gradient-boosted trees regression model is suitable for prediction. However, more research is needed to identify the main factors that influence the accuracy of this model, besides those considered in this project.

Research Question 3:

The data wrangling process transformed 10 columns into floating numbers, but only 5 of them were relevant for the regression analysis. These were (Table 3): (1) Surface temperature, (2) Distance fished, (3) Soak time, (4) Starting depth, and (5) Ending depth. These variables could help predict the catch rate per hour (CPUE). The other columns were categorical data, which could not be used for calculations, but only for descriptive purposes.

Table 3: Definitions of the variables that has been selected using for the analysis.

Variable Name	Variable explanation
surface temperature	Temperature of the water's surface.
distance fished	The distance covered during the fishing operation.
soak time	The duration between setting and hauling the fishing gear.
starting depth	The depth at which the fishing gear is initially set.
ending depth	The depth at which the fishing gear is retrieved.

Other than those findings, the project has predicted the CPUE from other regression models as well to understand the variation of the output compared to the Gradient-Boosted Trees regression model. That can be accessed from the Appendix 5**.**

Discussion and Conclusion

Usefulness of the project

- **Enhanced Operational Efficiency**:
  - The regression model of gradient-boosted trees provides high accuracy for forecasting the frequency of fish catches. This allows Alaskan fisheries to plan their operations better and match their resource allocation with the expected demand.
- **Resource Management and Conservation**:
  - The regression model helps in managing the fisheries sustainably by giving accurate predictions. It helps to understand how variables like surface temperature, distance fished, and soak time affect the fish stocks and make informed decisions accordingly.
- **Adaptability to External Factors**:
  - The model shows its adaptability by performing well even in unusual years affected by the COVID-19 pandemic. This adaptability is essential for the Alaskan fishing industry to cope with unexpected challenges and maintain its productivity and resilience.

Future Recommendations for Future Research

- **Improving Catchment Area Estimates**:
  - Estimate the average fish catch rates in different catchment areas within the main council management regions. This focused analysis improves predictions, helping fisheries to allocate resources efficiently for maximum yields while maintaining sustainability in each unique area.
- **Fish-Type Specific Analysis**:
  - Analyze the capture rates of individual fish types in different catchment areas, providing detailed insights. This deep understanding enables fisheries to adjust strategies, supporting species-specific conservation measures and improving the overall accuracy of catch predictions.
- **Integration of Additional Variables**:
  - Examine the effect of 30 variables with significant data on optimizing the regression model. Include oceanographic conditions, fishing gear variations, and seasonal trends in the analysis for a comprehensive understanding. This fine-tuning improves the model's predictive capacity, ensuring a more flexible and reliable tool for sustainable fisheries management.

Impact of the spark on this project

- **Data Analysis with Spark**:
  - The project relies on Spark to handle the large NMFS-AFSC Longline Sablefish Survey dataset effectively. Spark can manage huge amounts of data, which enables the project to process and analyze the extensive longitudinal survey quickly and accurately, setting the stage for reliable regression modeling (Ma *et al.*, 2023).
- **Parallel Processing with Spark**:
  - The project benefits from Spark's parallel processing power, which is essential for the complex regression analyses performed (Belcastro *et al.*, 2022). Spark improves speed and scalability by distributing computation across multiple nodes. This not only reduces

computation time but also accommodates the size of the dataset, making the modeling process smoother and more efficient (Zhou, 2023).

Impact of cloud computing on this project

- **Scalability**:
  - One of the critical aspects in fisheries industry analyses is scalability, which is greatly influenced by cloud computing's on-demand resources (Testas, 2023). The project can take advantage of the ability to scale computing resources according to the different computational needs (Tran *et al.*, 2022). This flexibility ensures efficient processing of large datasets, allowing the adjustment of computational power as the project requires, ultimately improving the overall efficiency and responsiveness of the analytical process (Belcastro *et al.*, 2022; Chandra and Dwivedi, 2022; Tran *et al.*, 2022).
- **Data Accessibility**:
  - Cloud storage enhances data accessibility, enabling collaboration and streamlined project management. Project members can easily access and work with the NMFS-AFSC Longline Sablefish Survey dataset. Cloud platforms support real-time collaboration, allowing efficient sharing and use of data, creating a collaborative environment that is essential for the project's success (Aslam *et al.*, 2023).

Conclusion

This project explores one of the key problems for Alaska's fishing industry: how to optimize the catch frequency in a limited time at sea. The key solution is to apply machine learning regression analysis, especially the Gradient-Boosted Trees model, which can change the game in this field. The project uses the AFSC/ABL Longline Sablefish Survey data to reveal the hidden patterns of fish populations.

The research answers important questions and shows that the Gradient-Boosted Trees Regression Model performs better than other methods and gives accurate predictions. Except the anomaly on the COVID-19 pandemic years of 2020 and 2021. The chosen variables, such as surface temperature, distance fished, soak time, starting depth, and ending depth, are crucial for predicting the catch rates per hour (CPUE).

The project has a positive impact on operational efficiency, resource management, and fish stock conservation. The findings help to adapt to external factors, which is essential for the industry's resilience. For future research, some possible directions are to improve the catchment area estimates, to analyze different types of fish, and to include more variables for a complete understanding.

The project uses Pyspark, Pandas, NumPy, and scikit-learn as the technological framework, which demonstrates the power of big data processing, distributed computing, and efficient machine learning (Ali and Iqbal, 2022). Spark's parallel processing capabilities help to handle large-scale data effectively (Aslam *et al.*, 2023; Singh, 2019; Testas, 2023). Cloud computing also improves scalability and data accessibility, which ensures smooth collaboration for project management.

References

Ali, M. and Iqbal, K. (2022), "The Role of Apache Hadoop and Spark in Revolutionizing Financial Data Management and Analysis: A Comparative Study", *Journal of Artificial Intelligence and Machine Learning in Management*, Vol. 6 No. 2, pp. 14–28.

Aslam, K., Chen, Y., Butt, M. and Malavolta, I. (2023), "Cross-Platform Real-Time Collaborative Modeling: An Architecture and a Prototype Implementation via EMF.Cloud", *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., Vol. 11, pp. 49241–49260, doi: 10.1109/ACCESS.2023.3276872.

Belcastro, L., Cantini, R., Marozzo, F., Orsino, A., Talia, D. and Trunfio, P. (2022), "Programming big data analysis: principles and solutions", *Journal of Big Data*, Springer Science and Business Media Deutschland GmbH, Vol. 9 No. 1, pp. 1–50, doi: 10.1186/S40537-021-00555-2/TABLES/3.

Chandra, T.B. and Dwivedi, A.K. (2022), "Data visualization: existing tools and techniques", *Advanced Data Mining Tools and Methods for Social Computing*, Academic Press, pp. 177–217, doi: 10.1016/B978-0-32-385708-6.00017-5.

Chhabra, G., Kumar, S., Gupta, S. and Nagpal, P. (2023), "Techniques for Behaviour Analysis Using Deep Learning", *Artificial Intelligence to Analyze Psychophysical and Human Lifestyle*, Springer, Singapore, pp. 41–58, doi: 10.1007/978-981-99-3039-5_4.

Cline, T.J., Schindler, D.E. and Hilborn, R. (2017), "Fisheries portfolio diversification and turnover buffer Alaskan fishing communities from abrupt resource and market changes", *Nature Communications 2017 8:1*, Nature Publishing Group, Vol. 8 No. 1, pp. 1–7, doi: 10.1038/ncomms14042.

Deng, L. and Yu, D. (2013), "Deep learning: Methods and applications", *Foundations and Trends in Signal Processing*, Now Publishers Inc, Vol. 7 No. 3–4, pp. 197–387, doi: 10.1561/2000000039.

Jadhav, A., Pramod, D. and Ramanathan, K. (2019), "Comparison of Performance of Data Imputation Methods for Numeric Dataset", *Applied Artificial Intelligence*, Taylor and Francis Inc., Vol. 33 No. 10, pp. 913–933, doi: 10.1080/08839514.2019.1637138.

Karp, M.A., Peterson, J., Lynch, P.D., Griffis, R., Adams, C., Arnold, B., Barnett, L., *et al.* (2018), "Accounting for Shifting Distributions and Changing Productivity in the Fishery Management Process: From Detection to Management Action".

Khatami, S. and Frantz, C. (2023), "Copatrec: A correlation pattern recognizer Python package for nonlinear relations", *SoftwareX*, Elsevier, Vol. 23, p. 101456, doi: 10.1016/J.SOFTX.2023.101456.

Ma, C., Zhao, M. and Zhao, Y. (2023), "An overview of Hadoop applications in transportation big data", *Journal of Traffic and Transportation Engineering (English Edition)*, Elsevier, Vol. 10 No. 5, pp. 900–917, doi: 10.1016/J.JTTE.2023.05.003.

Marcelino, C.G., Leite, G.M.C., Celes, P. and Pedreira, C.E. (2022), "Missing Data Analysis in Regression", *Applied Artificial Intelligence*, Taylor & Francis, Vol. 36 No. 1, doi: 10.1080/08839514.2022.2032925.

National Oceanic and Atmospheric Administration. (2023), "AFSC/ABL: Longline Sablefish Survey - Catalog", available at: https://catalog.data.gov/dataset/afsc-abl-longline-sablefish-survey1 (accessed 6 December 2023).

Sergeant, C.J., Seitz, A., Moran, S.B. and Collins, R. (2022), “Freshwater pressures on pacific salmon in the coastal watersheds of Alaska”.

Singh, P. (2019), “Learn PySpark”, *Learn PySpark*, Apress, doi: 10.1007/978-1-4842-4961-1.

Siwicke, K. and Malecha, P. (2023), *The 2022 Longline Survey of the Gulf of Alaska and Eastern Aleutian Islands on the FV Alaskan Leader: Cruise Report AL-22-01*.

Testas, A. (2023), “Decision Tree Regression with Pandas, Scikit-Learn, and PySpark”, *Distributed Machine Learning with PySpark*, Apress, Berkeley, CA, Berkeley, CA, pp. 75–113, doi: 10.1007/978-1-4842-9751-3_4.

Tran, M.K., Panchal, S., Chauhan, V., Brahmbhatt, N., Mevawalla, A., Fraser, R. and Fowler, M. (2022), “Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery”, *International Journal of Energy Research*, John Wiley & Sons, Ltd, Vol. 46 No. 2, pp. 786–794, doi: 10.1002/ER.7202.

Xing, Y., Zeng, X. and Alizadeh-Shabdiz, F. (2023), “Signal Prediction on Catalonia Cell Coverage”, *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, Springer Science and Business Media Deutschland GmbH, Vol. 514 LNICST, pp. 58–72, doi: 10.1007/978-3-031-44668-9_5/COVER.

Zhou, S. (2023), “Auto-Tuning Apache Spark Parameters for Processing Large Datasets”.

Appendix 1: Contributions from each member

Team member contributions - Group 7

Task	Team Member
Identification of Dataset	Bhushan Shelke, Hasaranga Jayathilake
Preliminary Presentation	Hasaranga Jayathilake, Bhushan Shelke
Data Pre-Processing	Bhushan Shelke, Hasaranga Jayathilake
Test identification and model development	Hasaranga Jayathilake, Bhushan Shelke
Data Analysis & Visualization	Bhushan Shelke, Hasaranga Jayathilake
Final Presentation and Report	Hasaranga Jayathilake, Bhushan Shelke

Appendix 2: Link to the Project Jupiter Notebook and Dataset

External Link for the Project Jupiter Notebook and Dataset

Final Project - Group 7 - Jupiter Notebook
NMFS-AFSC Longline Sablefish Survey Dataset – Direct Download
NMFS-AFSC Longline Sablefish Survey Dataset – View

Appendix 3: Survey Geographic Coverage

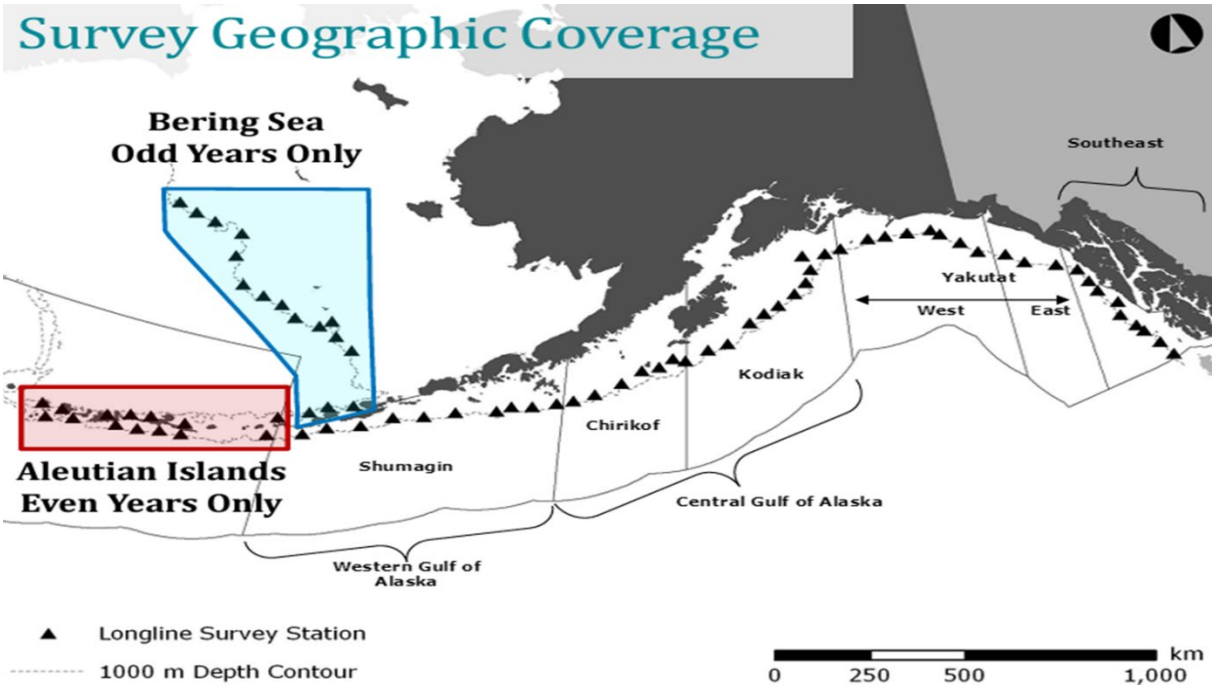

Figure 1: NMFS-AFSC Longline Sablefish Survey Geographic Coverage (Siwicke et al., 2022)

Appendix 4: Detail Regression Models Comparison

Table 4: Metrics for Linear Regression

Metrics for Model Evaluation	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
Mean Squared Error (MSE):	34.20927553	28.80062	35.96565	34.28599542	49.12391102	32.71622433
Root Mean Squared Error (RMSE):	5.848869594	5.3666209	5.997136	5.855424444	7.008845199	5.719809816
Mean Absolute Error (MAE):	4.235607964	3.7847745	4.235873	4.250325773	5.311962066	4.298742868
R-squared (R2):	0.068265907	0.0137583	0.003316	0.055231554	0.019999919	0.042015014

Table 5: Metrics for Random Forest Regression

Metrics for Model Evaluation	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
Mean Squared Error (MSE):	33.151140	28.19446	34.74502	33.19659866	47.7685489	32.084845759
Root Mean Squared Error (RMSE):	5.7577027	5.309846	5.894491	5.761648953	6.91147950	5.6643486615
Mean Absolute Error (MAE):	4.1838963	3.757249	4.17225	4.200197571	5.24263359	4.2549087483
R-squared (R2):	0.0970855	0.034515	0.03714	0.085250448	0.04703878	0.0605028197

Table 6: Metrics for Generalized Linear Regression

Metrics for Model Evaluation	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
Root Mean Squared Error (RMSE):	5.8936861	5.3840	5.975067	5.8503065	7.008315	5.749541467
Mean Absolute Error (MAE):	4.25759302	3.7989	4.224852	4.251516	5.306663	4.3142894304
R-squared (R2):	0.06655246	0.0143	0.003939	0.0553756	0.020134596	0.0422426019

Appendix 5: Prediction Output (CPUE) From Other Models

Table 7: Prediction Output of Linear Regression Model

Years	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
2018	3.9	3.7	4.3	3.1	5.7	4.7
2019	3.8	3.7	4.3	3.0	5.7	4.6
2020	3.8	3.6	4.3	3.0	5.7	4.6
2021	3.7	3.6	4.2	2.9	5.6	4.6
2022	3.6	3.6	4.2	2.7	5.6	4.5

Table 8: Prediction Output of Random Forest Regression Model

Years	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
2018	4.5	4.2	4.6	4.0	6.1	5.0
2019	4.5	4.3	4.6	4.0	6.1	5.0
2020	4.7	4.2	4.6	4.1	6.2	5.0
2021	4.4	4.2	4.7	4.0	6.1	5.0
2022	4.5	4.2	4.6	4.2	6.1	5.0

Table 9: Prediction Output of Generalized Linear Regression Model

Years	West Yakutat	Bering Sea	Aleutians	East Yakutat/SouthEast	Western Gulf of Alaska	Central Gulf of Alaska
2018	3.9	3.7	4.3	3.1	5.7	4.7
2019	3.8	3.7	4.3	3.0	5.7	4.7
2020	3.8	3.6	4.3	2.9	5.7	4.6
2021	3.7	3.6	4.3	2.8	5.6	4.6
2022	3.6	3.6	4.2	2.7	5.6	4.5