Progress Report for:

**Improved Prediction of Salmon Runs by Analysis of Historical Fish Count and Weather Data to Enhance Business Planning and Resource Allocation**

Hayden Blackburn, Richard Li, Binh Vu, Lijia Cheng, Alexandra Prokhorova

**Team 55**

# 1. Project context and research objective

Salmon fishing is one of the most valuable industries in Alaska and is a cornerstone of the state's economy. Recreational salmon fishing in the Upper Cook Inlet, which is the Kenai River system, generates 3,400 average annual jobs producing $104 million (2006 dollars) in income (Horton [6]). High variability in fish counts can lead to days of disappointed anglers and inefficient business operations.

Using data from different public agencies, our project plans to create a fish count prediction model using factors including air and water temperature, precipitation, river flood stage, river discharge, lunar phase and presence of commercial nets. Lunar phase is included because the phase of the moon affects the tides and there have been various studies recognizing that the phase of the moon influences the behavior of spawning salmon.[7]

The results of this analysis could impact a broad set of decisions across many different businesses with a focus on planning. We have identified the following key areas of decision-making that could benefit from the analysis:

* Local businesses, such as fishing stores, tourism, and hospitality services will be able to make analytics-driven decisions to plan staffing and inventory levels.
* Alaska Fish and Game (city/government) will be able to enhance their data offerings with more published insights into factors that impact fish count and provide predictive analysis. This effort could also be used by authorities to better regulate the fish population and help maintain ecological balance.
* Visitors will be able to plan their trip to Alaska based on external variables we find significant which will improve their satisfaction and increase the likelihood of them returning to Alaska to sustain the tourism economy.

2. Data overview

## 2.1 Sources

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Source** | **Unit** | **Data type** | **Link** |
| Daily sockeye salmon fish count at Kenai River (Late-Run Sockeye) | Alaska Department of Fish and Game | Count | Integer | [1] |
| Daily minimum, maximum and mean air temperatures at  Kenai AP | Alaska Climate Research Center | Degree Fahrenheit | Integer | [2] |
| Daily Precipitation at Kenai AP | Alaska Climate Research Center | Inch | Integer | [2] |
| Daily river flood stage at Kenai River (Kenai Keys).  *\*Proxy for river water level* | National Weather Service | Stage | Categorical | [3] |
| River discharge | USGS National Water Information System | ft3/s | Integer | [4] |
| Kenai River water temperature data at Soldotna | USGS National Water Information System | Degree Celsius | Integer | [4] |
| Moon Phase (split into 4 categories)  *\*Proxy to investigate impact of tides* | Papers with Code (Blog) | Moon Phase | Categorical | [5] |
| Set Net Location | Data from AF&G Biologist | Location | Categorical |  |
| Drift Net Location | Data from AF&G Biologist | Location | Categorial |  |
| Nets | Data from AF&G Biologist | Drift, Set, Both, no\_nets | Categorical |  |

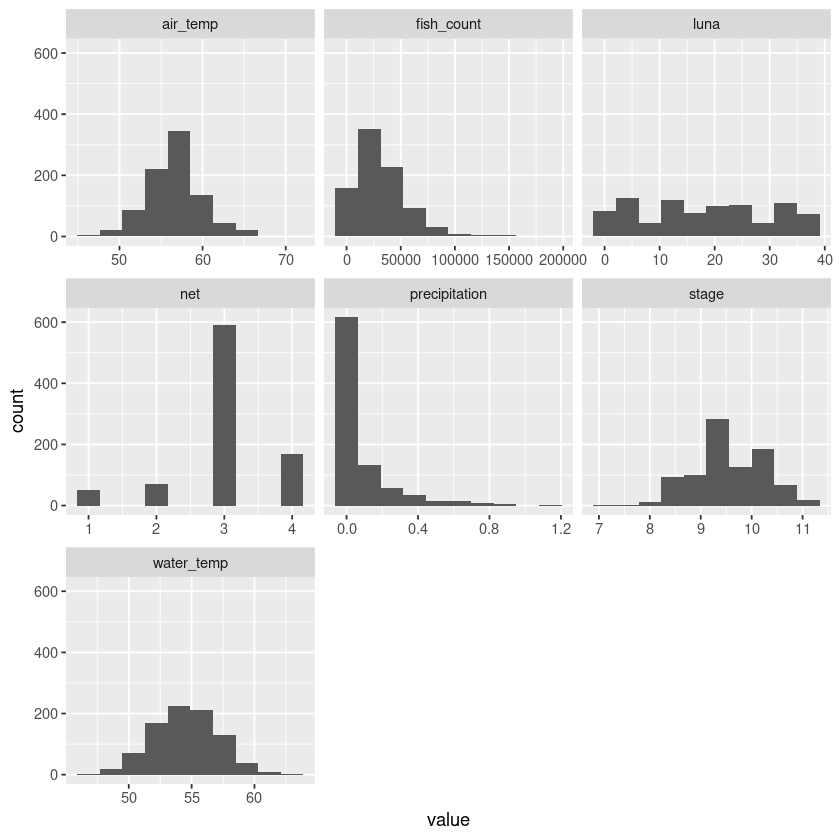
## 2.2 Cleaning

Each of the different data sources had to be joined into a single unified dataset. The key dataset was the daily sockeye salmon fish count at the Kenai River, the response variable we are looking to explain. This dataset was joined with all other data sources on date. Additionally, the precipitation, water temperature, air temperature, and discharge all had missing data which was imputed with linear interpolation except for water temperature. We went with this approach because there were few missing values for precipitation, air temperature, and discharge and it was a reasonable approach to use linear interpolation. However, eventually we decided because precipitation and water temperature had data missing for 13 years, between 2001 and 2013 we decided to scale our dataset to only include the years 2014-2022.

One dataset that required cleaning was the data provided by a biologist from Alaska Fish & Game, which contained historical commercial fishing data. We knew that commercial fishermen fished the waters outside the mouth of the river, so we had to include data related to when and where commercial fishing occurred. Individual CSV files were given to us for each year of commercial fishing data, indicating the locations and days when commercial fishing occurred. The location data had multiple variations in spelling for the same location, which we condensed into unique locations across the datasets.

Commercial fishing involves two types of "net" fishing: Set nets, which sit along the ocean floor and are stationary, and Drift nets, which are held up by buoys and connected to boats. Using this data, we set parameters for when nets were out and added a dataset that was joined on the date, where for a given day, a column called Nets was either Set, Drift, or Both. Unfortunately, this historical data did not include the number of commercial fishing boats that went out on a given day, which would have helped improve and enrich the data. However, combining our fish count and weather data with the historical commercial fishing data created a fuller dataset.

## 2.3. Exploratory Data Analysis

In Figure 2C, the combined distribution bar graphs of the variables show the distributions of most attributes are not normal. 'Precipitation' and 'fish\_count' exhibit a strong negative skew, while 'net' has a positive skew. 'stage' seems to show a bimodal distribution, and 'Luna' has a uniform distribution. We might have to perform some data transformation, such as a log transform, if our R-squared value is heavily impacted.

We had initially planned to use a CUSUM or change detection model to analyze the data. However, after several attempts and iterations, we found the resulting charts difficult to interpret and not particularly useful for our project. Instead, we focused on studying the peaks of each season (see graph below) and discovered that all peaks, except for one year (2020), occurred between July 19 and August 8. In the past 9 years, 8 have occurred within this period, while 7 out of the last 9 years saw peaks between July 19 and July 29. This information is meaningful for local businesses, fishing guides, and tourists. If I were looking to book a fishing vacation in Kenai, I would ensure it falls within those dates.

Graphical user interface, chart

Description automatically generated3. Modeling

## 3.1 Random Forest, ridge regression and lasso regression

Overall, we are looking for an R-squared value which is the percentage of variance that could be explained by the model. R-squared ranges from 0-1, where 1 indicates 100% variances explained. The higher the R-squared value the better. All three models are trained using 70/30 split of fish\_master dataset. The final coefficients are fish\_count, precipitation, air\_temp, stage, water\_temp, luna, and net. These coefficients are selected based on the result of correlation plot and exclusion of location because of null values.

Ridge regression limits the impact of multicollinearity by using a shrinkage penalty. The predictor variable with the least influence on the response will shrink toward zero but never reach zero. In Figure 3A and Figure 3B, the left graph shows how well each lambda to MSE – the lowest point in the curve indicates the optimal lambda. The right graph is a Trace plot which visualizes how coefficient estimates change as we increase lambda. In our case, lambda value of 1,731.7 is the optimal value. With that lambda, we trained optimal ridge regression. Both Ridge regression and Lasso regression are known as regularization methods because they both attempt to minimize the sum of squared residuals (RSS) along with some penalty term. Unlike ridge, lasso regression coefficients could go completely to zero as alpha gets sufficiently large. For lasso, the optimal lambda is 459.96. Lastly, we will explore random forest regression which combines the output of multiple regression trees to reach a single result.

|  |
| --- |
|  |
| Figure 3A: Ridge regression - MSE by lambda value and Trace plot |

From the resulting R-squared (Table 3C), we can see that RFR performed significantly better than the other two regression models. ~5% for both ridge and lasso could be a sign of error in calculation or model building. On the other hand, RFR seems to be much higher than the average r-squared from other regression models. More research will be conducted prior to the final report.

## 3.2 Linear regression

For Linear Regression we set up the Dependent variable to be fish count and began testing the different independent variables we had in our dataset.

We needed our data structure to consist of various data types such as date, numeric, int, and multiple categorical factor fields. Therefore, we had to ensure that the data was appropriately structured to be read into our model in the desired format. As we started running regressions, we found that our locations provided significant insights into how different locations affected Kenai salmon, and we could predict the fish count for a particular day based on the type of nets deployed and their location. This analysis will allow local businesses, tourists, fisheries, and fish guides to make informed decisions around fishing activities. This could also provide valuable insights into how the fishing industry could improve its practices in and around Kenai.

Our most significant factors after doing different types of variable selection including stepwise in both directions were as follows: Water\_temp, Moon\_phase, Discharge, Drift\_location and Nets.

Chart, scatter chart

Description automatically generatedWe are excited to continue to find ways to improve this initial regression analysis. The current model has an R2 of .347. By plotting the residuals against the independent variable (see plot below), we can check for any patterns in the data that might indicate the presence of nonlinear relationships or heteroscedasticity (unequal variances). Our data does cluster closely with the prediction line, however, there appear to be outliers we still need to address. We are encouraged by the linear pattern and constant variance which indicates a linear regression model is appropriate.

## 3.3 Feature engineering

Our team has decided on a multi-prong approach. That means each team member starting with the same dataset could utilize various combinations of techniques - log transformation, convert categorical data, and coefficient selections - that would produce the best individual model(s). Then, we would compare and select the best combination for the final models.

# 4. Project status and timeline

Our team has completed data cleaning, data exploratory analysis and is currently running the data through various models to determine which variables are significant and implementing a model that would be best at predicting fish count. The next steps will be to fine tune our models and decide which ones are most useful for prediction and drawing insights.

# 5. References

[1] Alaska Department of Fish and Game (n.d.). *Fish count data search*. Fish Counts - Sport Fish - ADF&G. Retrieved March 12, 2023, from <https://www.adfg.alaska.gov/sf/FishCounts/index.cfm?ADFG=main.displayResults>

[2] Alaska Climate Research Center. (n.d.). Daily air temperature and precipitation at Kenai AP datasets. Retrieved March 12, 2023, from <https://akclimate.org/data/data-portal/>

[3] US Department of Commerce, N. O. A. A. (2020, September 23). Historical river observations database. National Weather Service. Retrieved March 12, 2023, from <https://www.weather.gov/aprfc/rivobs>

[4] USGS water data. USGS water data for the nation. (n.d.). Retrieved March 12, 2023, from <https://waterdata.usgs.gov/nwis/>

[5] Mateos, L. (n.d.). Moon phases dataset. Moon Phases Dataset | Papers With Code. Retrieved March 12, 2023, from <https://paperswithcode.com/dataset/moon-phases>

[6] Horton, C. (2016, December 5). Economic impact of fishing the Kenai Peninsula up for debate. Alaska Journal. Retrieved March 12, 2023, from <https://www.alaskajournal.com/community/2008-05-25/economic-impact-fishing-kenai-peninsula-debate>

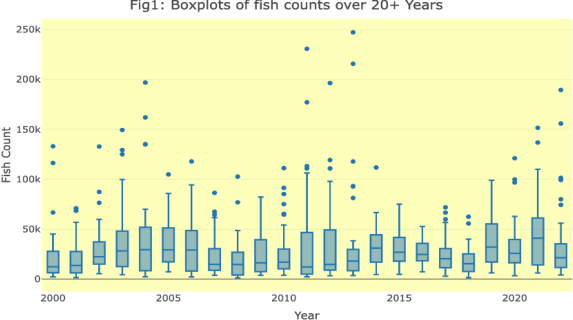
[7] Kramer, C. (2014, September 4). Lunar effects on salmon. Alaska Science Forum. <https://www.gi.alaska.edu/alaska-science-forum/lunar-effects-salmon>

# 6. Appendix

**Figure 2A: Correlation matrix** Chart, scatter chart, bubble chart

Description automatically generated

**Figure 2B: Fish counts over the years: mean, minimum, maximum and outliers**



**Figure 3B: LASSO regression - MSE by lambda value and Trace plot**

|  |
| --- |
|  |
| Inserting image... |

**Table 3C: Advanced Regression R-squared Values**

|  |  |
| --- | --- |
| **Model** | **R-squared** |
| Ridge Regression | 0.059 |
| Lasso Regression | 0.053 |
| Random Forest Regression (RFR) | 0.506 |