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

Georgia Institute of Technology - MGT 6203 Spring 2023

Final Project Report - Team 55

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

Table of Content

1. Project context and research objective 2
2. Data overview 2
2.1 Sources 2
2.2 Cleaning 3
2.3. Exploratory Data Analysis 3
3. Modeling 4
3.1 Linear regression 4
3.2 Random Forest, Ridge regression and Lasso regression 7
3.3 Logistic Regression 8
3.4 K-Means Clustering 10
4. Discussion and key takeaways 10
4.1 Evaluation of models and challenges 10
4.2 Conclusion and future work 11
5. References 12

1. Project context and research objective

Salmon fishing is one of the most valuable industries in Alaska and a cornerstone of the state's economy. In 2007, 475,534 resident and nonresident licensed anglers collectively fished 2.5 million days in Alaska, and spent nearly $1.4 billion on licenses and stamps, trip-related expenditures, pre-purchased packages, equipment, and real estate used for fishing. The $1.4 billion angler spending in Alaska supported 15,879 jobs in Alaska and provided $545 million of income [1]. Additionally, recreational salmon fishing in the Upper Cook Inlet, which is the Kenai River system, generates 3,400 average annual jobs producing $104 million in income in 2006 [2]. As the anglers' dollars moved from business to business in the Alaska economy, government revenues were generated through personal income taxes, local property taxes, sales taxes, business taxes, and excise taxes. In total, $123 million in tax revenues were generated for state and local governments in Alaska and $123 million for the Federal government [1].

High variability in fish counts can lead to days of disappointed anglers and inefficient business operations. Currently there is no way to know or predict the current day's fish count. The only tools accessible for everyday tourists, business owners, and commercial fishermen is to view Alaska Fish & Game's website for the fish count of the previous day and make a guess at what todays fish count will be. Alaska Fish & Game has sonar equipment located at different points along the Kenai River where fish images are captured and counted by their technicians to better plan and target their salmon escapement goals. This fish count data has been captured since the mid-1980's and is documented on their website for each year making for a substantial dependent variable as our basis to perform tests and research.

Our objective was to see if by using data from different public agencies, we can create a fish count prediction model using factors that range from air and water temperature, precipitation, river flood stage, river discharge, lunar phase, and presence of commercial nets. The lunar phase is relevant because the phase of the moon affects the tides and there have been numerous studies recognizing that the phase of the moon influences the behavior of spawning salmon [3]. Creating such a tool – or better understanding of the factors that influence salmon runs would be hugely beneficial for the community, their economy, and those who depend on the salmon for financial or other reasons.

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: (1) Local businesses, such as fishing stores, tourism, and hospitality services will be able to make analytics-driven decisions to plan staffing and inventory levels; (2) 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; and (3) 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	[4]
Daily minimum, maximum and mean air temperatures at Kenai AP	Alaska Climate Research Center	Degree Fahrenheit	Integer	[5]
Daily Precipitation at Kenai AP	Alaska Climate Research Center	Inch	Integer	[5]
Daily river flood stage at Kenai River (Kenai Keys)	National Weather Service	Stage	Categorical	[6]
River discharge	USGS National Water Information System	ft3/s	Integer	[7]
Kenai River water temperature data at Soldotna	USGS National Water Information System	Degree Celsius	Integer	[7]
Moon Phase (split into 4 categories)	Papers with Code (Blog)	Moon Phase	Categorical	[8]
Set Net Location	Data from AF&G Biologist	Location	Categorical	
Drift Net Location	Data from AF&G Biologist	Location	Categorical	

Nets	Data from AF&G Biologist	Drift, Set, Both, no_nets	Categorical	

Table 2A: Data Sources

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 to 2022.

One dataset that required extensive 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. One extensive challenge we encountered was with the locations included with the commercial fishing data. Multiple locations had different variations in spelling across the years. We identified each of these locations and condensed them 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 more complete dataset.

2.3. Exploratory Data Analysis

We first hypothesized that we could use a CUSUM or change detection model to analyze the fish count 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 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 insight is meaningful for tourists to plan for fishing trips and helps local businesses and fishing guides in making adequate resource planning decisions in anticipation of the period of high demand.

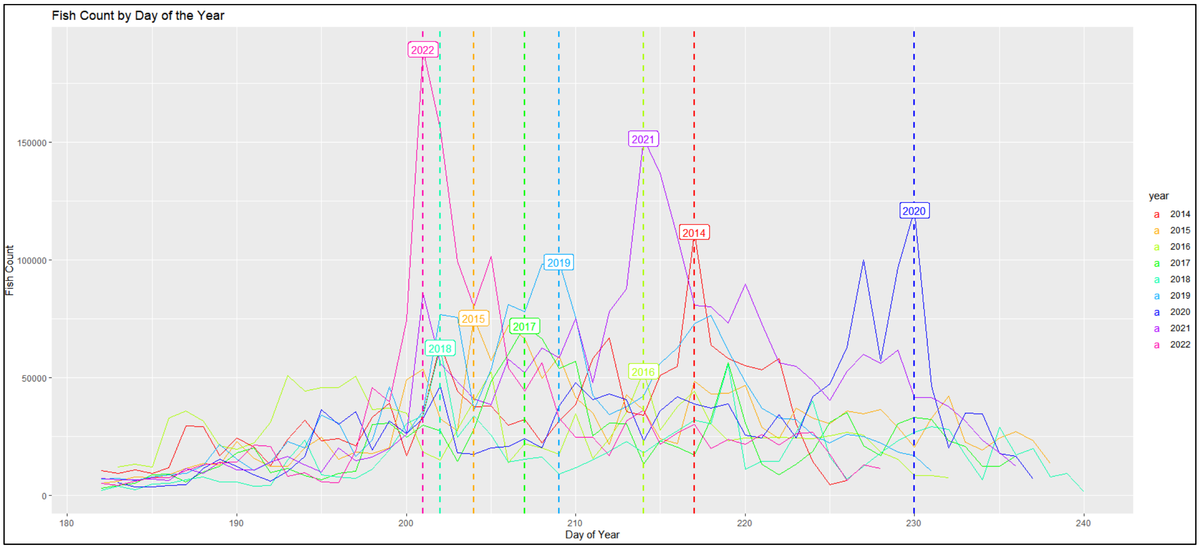


Figure 2A: Fish count peak analysis from 2014 to 2022

The combined distribution bar graphs of all variables (**Figure 2B**) show that distributions of most attributes are not normal, except for air temperature, river water discharge and water temperature. Fish count and precipitation exhibit a strong negative skew, while the presence of commercial nets, moon phase and river stage exhibit a multimodal distribution. Based on the distributions and magnitudes, log transformations were used to improve our R-squared value. When values are measured in different units – for example discharge in cubic ft per second – log transformations helped scale the large measurements to normality.

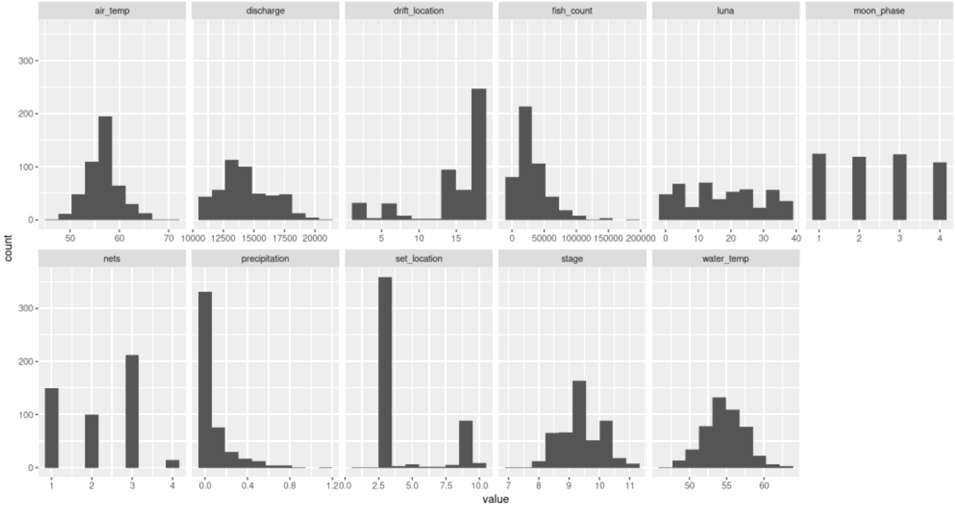


Figure 2B: Dataset Distribution Summary

When looking at the correlation between the numeric variables using a correlation plot (**Figure 2C**), we observed a strong correlation between stage and discharge. This should have been expected as both variables are measurements of how much water is moving through the river at a given time. When fitting models, discharge was shown to be more valid. Water temperature and air temperature had correlation because as the temperature rises or drops, they trend together.

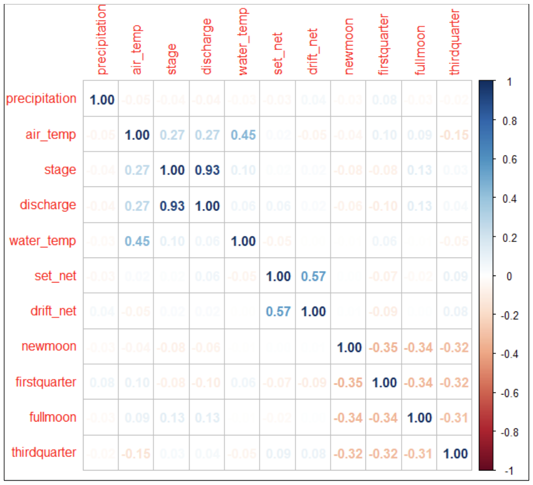
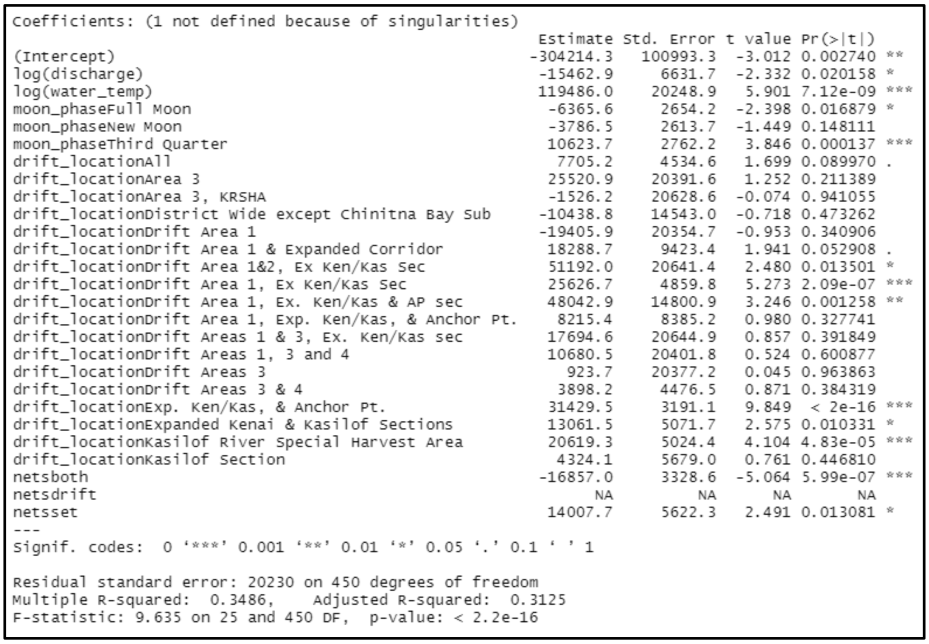


Figure 2C: Correlation plot of numeric variables in the dataset

3. Modeling

3.1 Linear regression

Linear regression is heavily used in machine learning and data analytics because of its ability to find the strength of relationships between a dependent variable and independent variables. One of our main objectives was finding the strength of relationships between our various variables and their effects on fish count. Another of our objectives was seeking to create an accurate fish count predictor model using linear regression. In creating a Linear Regression model, we worked to find which of our variables were shown to be statistically significant. Using fish_count as the dependent variable and using stepwise regression, the significant factors were found to be water_temp, moon_phase, discharge, drift_location, and nets.



```
Coefficients: (1 not defined because of singularities)
                                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                            -304214.3   100993.3  -3.012 0.002740 ** 
log(discharge)                                          -15462.9     6631.7  -2.332 0.020158 *  
log(water_temp)                                         119486.0    20248.9   5.901 7.12e-09 ***
moon_phaseFull Moon                                      -6365.6     2654.2  -2.398 0.016879 *  
moon_phaseNew Moon                                       -3786.5     2613.7  -1.449 0.148111    
moon_phaseThird Quarter                                  10623.7     2762.2   3.846 0.000137 ***
drift_locationAll                                         7705.2     4534.6   1.699 0.089970 .  
drift_locationArea 3                                     25520.9    20391.6   1.252 0.211389    
drift_locationArea 3, KRSHA                              -1526.2    20628.6  -0.074 0.941055    
drift_locationDistrict Wide except Chinitna Bay Sub     -10438.8    14543.0  -0.718 0.473262    
drift_locationDrift Area 1                              -19405.9    20354.7  -0.953 0.340906    
drift_locationDrift Area 1 & Expanded Corridor           18288.7     9423.4   1.941 0.052908 .  
drift_locationDrift Area 1&2, Ex Ken/Kas Sec             51192.0    20641.4   2.480 0.013501 *  
drift_locationDrift Area 1, Ex Ken/Kas Sec               25626.7     4859.8   5.273 2.09e-07 ***
drift_locationDrift Area 1, Ex. Ken/Kas & AP sec         48042.9    14800.9   3.246 0.001258 ** 
drift_locationDrift Area 1, Exp. Ken/Kas, & Anchor Pt.    8215.4     8385.2   0.980 0.327741    
drift_locationDrift Areas 1 & 3, Ex. Ken/Kas sec         17694.6    20644.9   0.857 0.391849    
drift_locationDrift Areas 1, 3 and 4                     10680.5    20401.8   0.524 0.600877    
drift_locationDrift Areas 3                                923.7    20377.2   0.045 0.963863    
drift_locationDrift Areas 3 & 4                           3898.2     4476.5   0.871 0.384319    
drift_locationExp. Ken/Kas, & Anchor Pt.                 31429.5     3191.1   9.849  < 2e-16 ***
drift_locationExpanded Kenai & Kasilof Sections          13061.5     5071.7   2.575 0.010331 *  
drift_locationKasilof River Special Harvest Area         20619.3     5024.4   4.104 4.83e-05 ***
drift_locationKasilof Section                             4324.1     5679.0   0.761 0.446810    
netsboth                                                -16857.0     3328.6  -5.064 5.99e-07 ***
netsdrift                                                     NA         NA      NA       NA    
netsset                                                  14007.7     5622.3   2.491 0.013081 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20230 on 450 degrees of freedom
Multiple R-squared:  0.3486,	Adjusted R-squared:  0.3125 
F-statistic: 9.635 on 25 and 450 DF,  p-value: < 2.2e-16
```

Figure 3A: Linear Regression Output

Water temperature

From the results of linear regression, our water temperature variable had a strong statistically significant effect on fish count, a temperature change of 1% increases our fish count by 119,486. Originally, we did not apply log transformation on this variable, but eventually determined that log transformation is necessary considering how the fish count and temperature are different units of measurement. This not only increased our R squared (**Table 4A**) but also created more realistic predictions when tested that were closer to the average.

Discharge

Discharge is the volume of water flowing through the river and since it is measured in cubic ft per second, we similarly applied log transformation on this variable. Based on linear regression results, the discharge variable had the largest negative correlation, where a 1% change in discharge decreases the fish count by 15,463.

Moon phase

We originally hypothesized moon phases would be significant and of the four moon phases tested, full moon and third quarter moon phases were statistically significant. The full moon phase had a negative coefficient while the third quarter phase had a positive coefficient, implying that spring tides (occurs during full moon) impacts fish count negatively, while neap tides (occurs during third quarter) impacts fish count positively based on results of linear regression.

Nets

This was a categorical variable for the nets that were out for a given day whether it was drift, set, both, and no nets. Based on the results of the linear regression, we found the expected observation that when variable nets = both it was statistically significant with a negative coefficient, where we could expect a decrease in the fish count of -16,857 when both nets are out.

Drift locations

One interesting takeaway is that drift nets, according to our linear regression model, have more of an impact than set nets and using the different stepwise methods found various drift locations to be statistically significant. Of the drift locations these were found to be statistically significant: (1) All, (2) Drift Area 1 & Expanded Corridor, (3) Drift Area 1&2, Ex Ken/Kas Sec, (4) Drift Area 1, Ex Ken/Kas Sec, (5) Drift Area 1, Ex Ken/Kas Sec & AP Sec, (6) Exp. Ken/Kas, & Anchor Pt. , (7) Expanded Kenai & Kasilof Sections, (8) Kasilof River Special Harvest Area. Having these locations be significant was encouraging because these are the fishing locations closest to the mouth of the Kenai River. Refer to **Figure 3B** below where locations found to be statistically significant are marked in red.

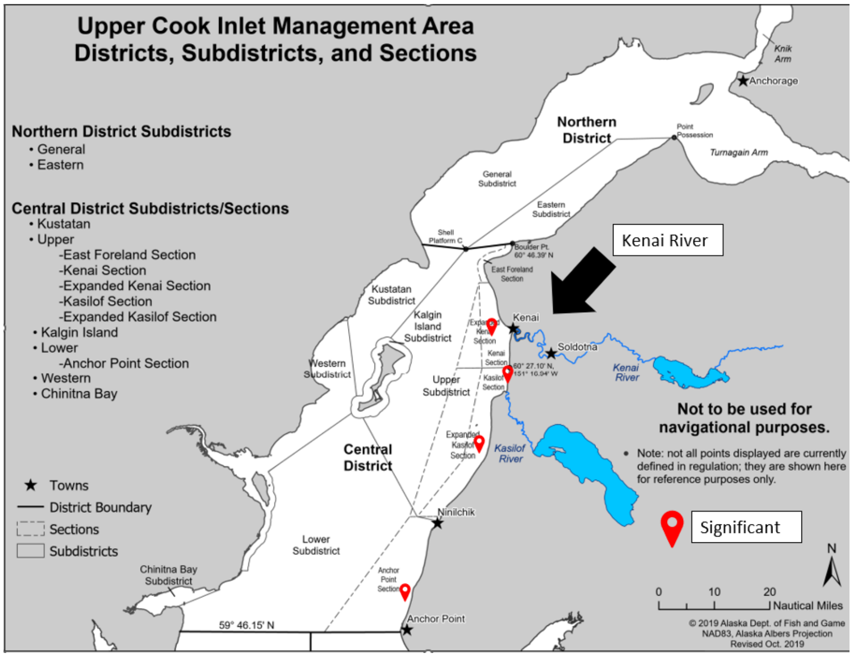


Figure 3B: Statistically significant drift locations

However, we found an unexpected observation that the regression coefficients were positive when we expected them to be negative. We expected them to be negative because when commercial fishermen are fishing in waters nearest to the Kenai, the number of salmon entering the Kenai River should decrease. An explanation for this could be that there is not enough historic data used in our model and our model finds them significant simply because they are among the most fished locations (see **Figure 3C**) which means they capture many of the high fish count days, thus making the coefficient positive.

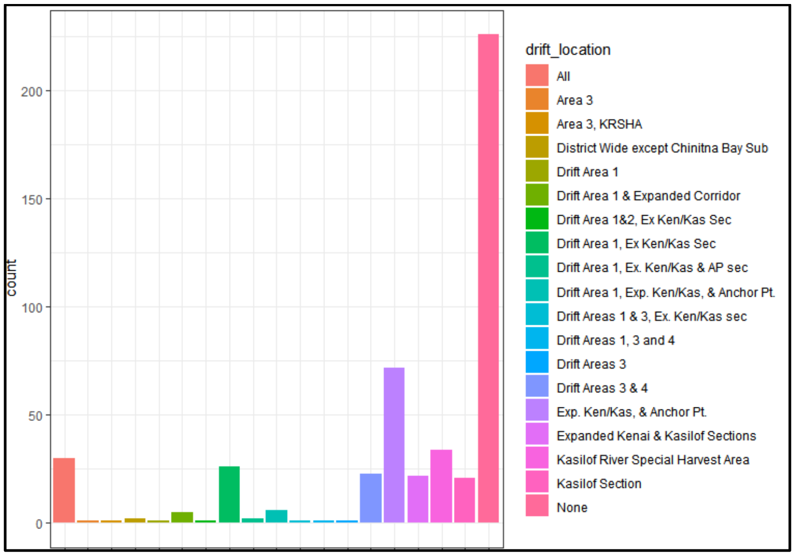


Figure 3C: Bar chart showing the count of drift fishing locations

Note that another way we tested our net data was using lag() from the dplyr library. We tested lagging the nets data to see if the nets had a lag effect on the fish_count. We tested lagging 2-5 days however, we found that this approach was not able to produce an R squared value greater than our original model, suggesting that the 'lag effect" is not significant based on our model.

3.2 Random Forest, Ridge regression and Lasso regression

To find the model that produces the best R squared value, we want to utilize advanced regression models and compare the result. These models improve on different aspects of linear regression. Both ridge regression and lasso regression lessen the impact of any influential factor and prevent overfitting. Similarly, random forest regression will give us the average R squared of hundreds of decision trees which should give us a more accurate result. All three models are trained using 70/30 split of the dataset.

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. To generate the two models in R, we used glmnet() method where we would pass in the training dataset and a shrinking coefficient lambda. Additionally, we used cross-validation to determine the optimal shrinking coefficient for the respective models based on the lowest MSE.

In **Figure 3D** and **Figure 3E**, 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 decrease lambda. Green color indicates an increase in coefficient value and red indicates a decrease. From the trace plot, all predictors seem to approach zero around log lambda of 14. In our case, lambda value of 1183.4 is the optimal value for ridge regression. With that lambda, we trained optimal ridge regression.

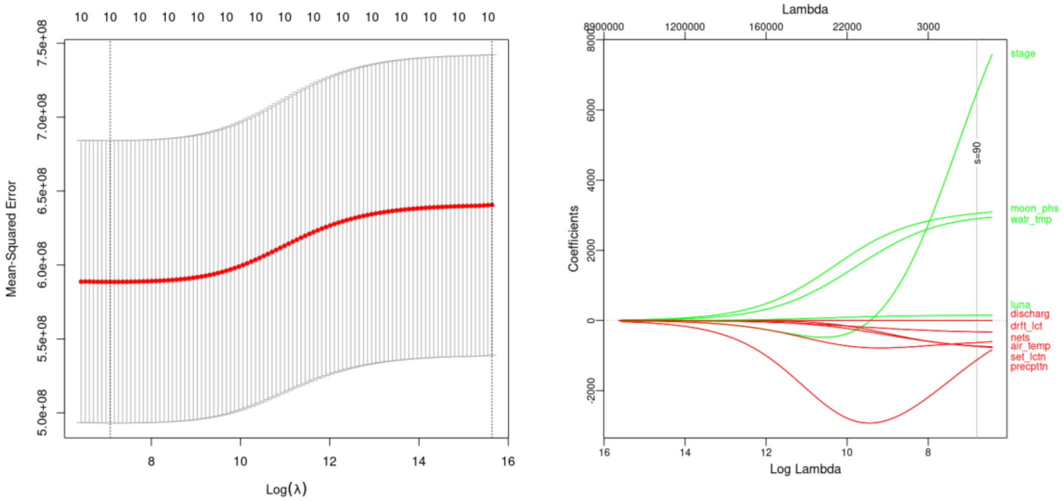


Figure 3D: Ridge regression - MSE by lambda value and Trace plot

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 lambda gets sufficiently large. For lasso, the optimal lambda is 136.1. From the trace plot in **Figure 3B**, stage is the first predictor to be eliminated and water_temp is the last. Lastly, we will explore random forest regression which combines the output of multiple regression trees to reach a single result.

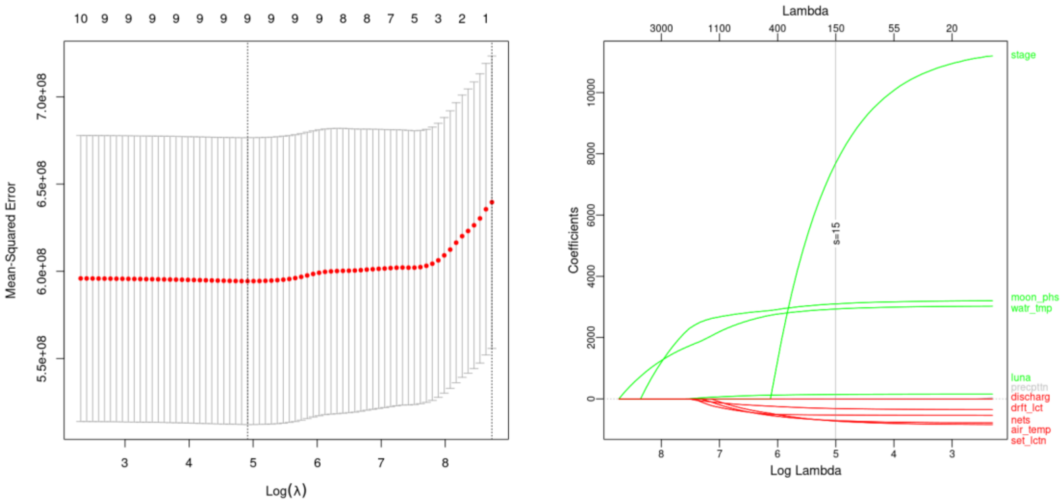


Figure 3E: LASSO regression - MSE by lambda value and Trace plot

From the coefficients output summary (**Table 3A**), both ridge regression and lasso regression demonstrate that precipitation is insignificant in determining the fish count. Ridge regression also minimizes the predictor stage. In terms of negative/positive factors, both models are aligned with each other.

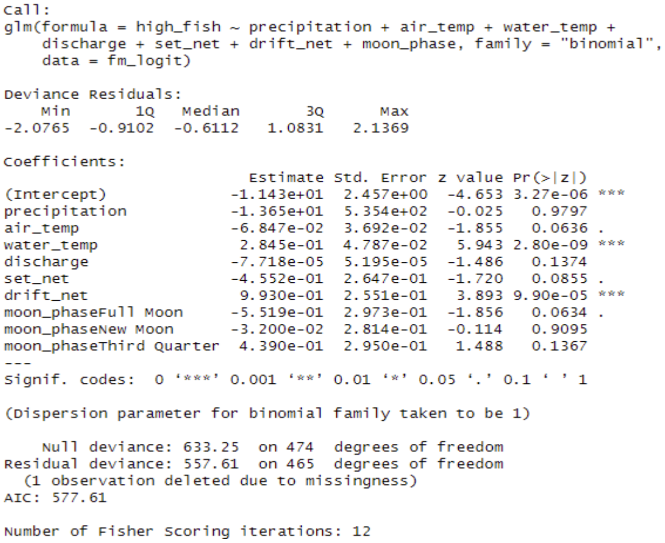
	Ridge Regression	Lasso Regression
Intercept	-71123.40	-110811.77
Precipitation	~ 0	0
Air_temp	-303.44	-708.45
Stage	~ 0	7679.49
Discharge	-0.95	-3.59
Water_temp	2360.96	2932.61
Lunar	82.37	142.34
Moon_phase	2741.20	3101.13
Set_location	-215.04	-725.54
Drift_location	-162.37	-316.29
Nets	-351.49	-530.68

Table 3A: Optimal Advanced Regression Coefficients

3.3 Logistic Regression

We used logistic regression as a method to predict the probability of a high fish count day. A high fish count day was defined using the mean fish count of the entire dataset, which was 30,737. A new feature, “high_fish”, was created where values were set to 1 when the fish count was above the mean and 0 when it was not. Using this new feature as the response variable, we fit the model.

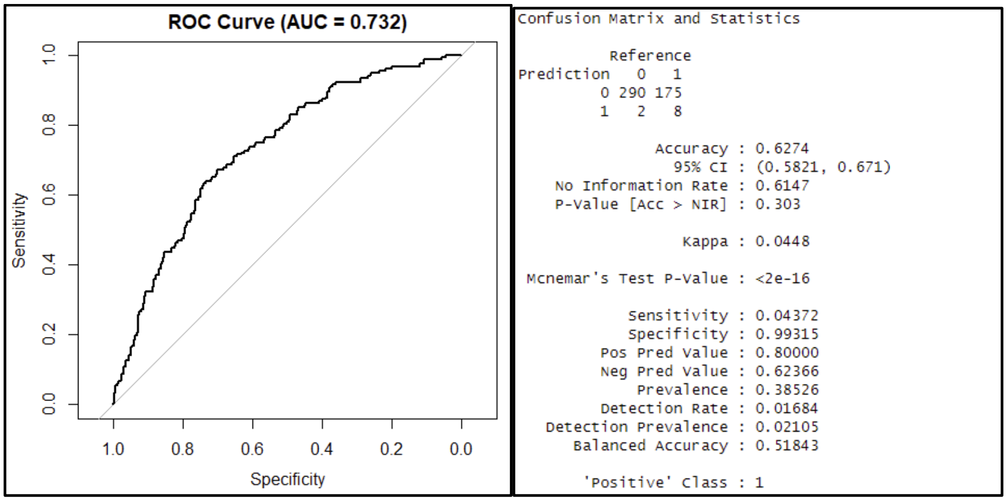
From the logistic regression output summary (**Figure 3F**), we observed that presence of drift nets and water temperature were the two most significant factors identified and they had a positive coefficient, implying a positive correlation with the log odds of a high fish count. We validated this result by looking at the mean fish count when drift nets were put out, which was 34,557. Another interesting finding was that the mean water temperature when fish counts were ‘high’ was 55.6 degrees, which is higher than the mean of the dataset of 54.7 degrees.



```
Call:
glm(formula = high_fish ~ precipitation + air_temp + water_temp +
    discharge + set_net + drift_net + moon_phase, family = "binomial",
    data = fm_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0765  -0.9102  -0.6112   1.0831   2.1369

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.143e+01  2.457e+00  -4.653 3.27e-06 ***
precipitation           -1.365e+01  5.354e+02  -0.025   0.9797
air_temp                -6.847e-02  3.692e-02  -1.855   0.0636 .
water_temp               2.845e-01  4.787e-02   5.943 2.80e-09 ***
discharge               -7.718e-05  5.195e-05  -1.486   0.1374
set_net                 -4.552e-01  2.647e-01  -1.720   0.0855 .
drift_net                9.930e-01  2.551e-01   3.893 9.90e-05 ***
moon_phaseFull Moon     -5.519e-01  2.973e-01  -1.856   0.0634 .
moon_phaseNew Moon      -3.200e-02  2.814e-01  -0.114   0.9095
moon_phaseThird Quarter  4.390e-01  2.950e-01   1.488   0.1367
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 633.25  on 474  degrees of freedom
Residual deviance: 557.61  on 465  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 577.61

Number of Fisher Scoring iterations: 12
```

Figure 3F: Logistic Regression Summary

To evaluate the fit of the model and quality of predictions, we looked at the ROC curve and the AUC metric. We also made predictions with the original dataset at a threshold of 0.8 and constructed a confusion matrix to quantify the prediction results.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 290 175
         1   2   8

               Accuracy : 0.6274
                 95% CI : (0.5821, 0.671)
    No Information Rate : 0.6147
    P-Value [Acc > NIR] : 0.303

                  Kappa : 0.0448

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.04372
            Specificity : 0.99315
         Pos Pred Value : 0.80000
         Neg Pred Value : 0.62366
             Prevalence : 0.38526
         Detection Rate : 0.01684
   Detection Prevalence : 0.02105
      Balanced Accuracy : 0.51843

       'Positive' Class : 1
```

Figure 3G: ROC plot and Confusion Matrix for Logistic Regression model

The logistic regression model had a sensitivity of 0.044 and specificity of 0.993, which means it could predict true negatives very well but true positives poorly. The AUC metric of 0.732 indicates that there is about a 73.2% chance that the model will be able to accurately distinguish between a positive and negative class. In addition, the binned residual plot (**Figure 3H)**, used to assess the overall fit along with the assumptions of a regression model with respect to the binary outcome, showed that all the residuals fall within bounds, which implies that the model performed well.

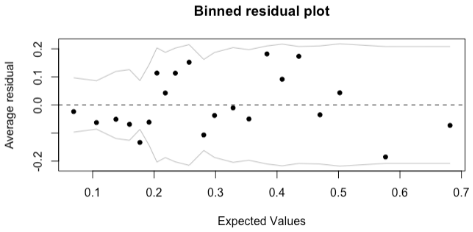


Figure 3H: Binned residual plot

3.4 K-Means Clustering

We used K-means clustering to identify clusters of data within the dataset and then further analyzed each cluster to check if there are any similar characteristics for clusters with high fish count. As part of the model selection process, we tested several values for the hyperparameter K, which is the number of cluster centers used in the algorithm.

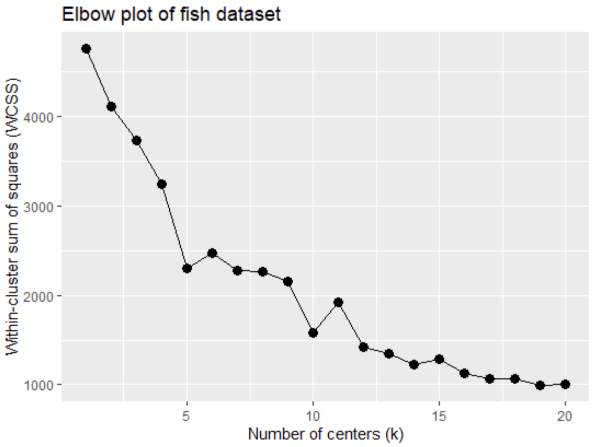
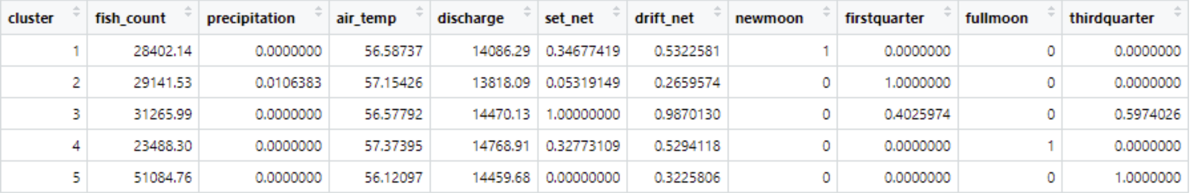


Figure 3I: Elbow plot of K-means Clustering ran with 1-20 centers

This elbow plot shows us the Within Cluster Sum of Squares (WCSS) metric for each cluster which allows us to identify the marginal benefit of adding another cluster. When the benefit of adding additional clusters is small, the number of clusters that the algorithm ran with is optimal. In the elbow plot, we can see this is the case at 5 clusters.



cluster	fish_count	precipitation	air_temp	discharge	set_net	drift_net	newmoon	firstquarter	fullmoon	thirdquarter
1	28402.14	0.0000000	56.58737	14086.29	0.34677419	0.5322581	1	0.0000000	0	0.0000000
2	29141.53	0.0106383	57.15426	13818.09	0.05319149	0.2659574	0	1.0000000	0	0.0000000
3	31265.99	0.0000000	56.57792	14470.13	1.00000000	0.9870130	0	0.4025974	0	0.5974026
4	23488.30	0.0000000	57.37395	14768.91	0.32773109	0.5294118	0	0.0000000	1	0.0000000
5	51084.76	0.0000000	56.12097	14459.68	0.00000000	0.3225806	0	0.0000000	0	1.0000000

Table 3J: Mean statistics of each cluster

Looking at the statistics for each of these clusters, we found that cluster 5 had the highest mean fish count at an average of 51,084. For this cluster, the air temperature and discharge both had values that were around the middle which tells us that these factors did not significantly impact the fish count. The average set_net value was 0, the lowest, and the drift_net value was 0.32, which is lower than 3 of the other 4 clusters. This tells us that that not having the nets could have potentially resulted in a higher fish count.

4. Discussion and key takeaways

4.1 Evaluation of models and challenges

From the R-squared values for all regression models (**Table 4A)**, we observed that R-squared values for ridge regression and lasso regression models were approximately 12%, suggesting that they fit poorly to the data. Comparatively, random forest regression ("RFR") and linear regression models performed significantly better than the other two advanced regression models with R-squared of approximately 35%. While RFR gave slightly better fit as compared to linear regression, RFR model does not give us insights into significance of factors on fish count and therefore, RFR would only be useful as a guiding model for comparing with other models.

Model	**R-squared**
Ridge Regression	0.1238
Lasso Regression	0.1272

Random Forest Regression (RFR)	0.3548
Linear Regression	0.3486

Table 4A: Regression Models R-Squared Values

Furthermore, we also ran two other analytical models: logistic regression model and k-means clustering. Logistic regression provided insights into the significant factors that impact fish counts, and it could be a useful, easy-to-use tool for predicting the probability of a good salmon run with fish count exceeding a set threshold. However, we see that our current model has very low accuracy and sensitivity, therefore suggesting that our model still needs to be further refined (e.g., adding more relevant variables and removing irrelevant variables) to be useful.

On the other hand, k-means clustering model was interesting to explore as an unsupervised learning model, although the results were difficult to interpret in a useful way. The optimal k value of 5 clusters, found with the elbow plot (**Figure 3I**), was not reliable in distinguishing well separated clusters. Additional clusters could be added to reduce the error metric and different initial clusters would have also given different results. Looking at the mean statistics for each cluster (**Table 3J**), cluster 5 had higher than average fish counts with the set_net feature equal to 0. Cluster 2's set_net value was similar at 0.05 but the difference in fish counts for the two clusters was ~22,000 fish. This difference makes it hard to conclude whether set_nets were correlated with higher fish counts.

Overall, we determined that the R-squared values across all models are too low to be used confidently and we noted a few inconsistencies across models. For example, linear regression and logistic regression had contradicting insights with regards to the impact of the presence of nets on fish counts, where linear regression suggests that nets would result in lower fish count (consistent with our expectation), whereas logistic regression suggests the opposite.

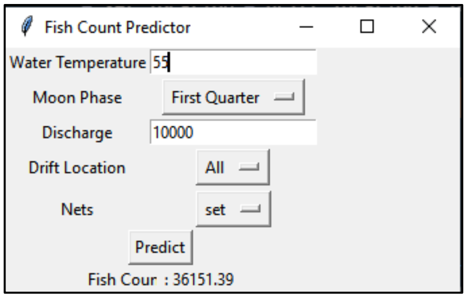
Like linear regression, both lasso regression and ridge regression selected similar predictors as statistically significant. Furthermore, all models found precipitation to be an insignificant predictor in predicting fish counts. However, unlike linear regression, the two advanced regression models also include three other predictors: air_temp, lunar, and set_location.

We believe that if we could remove less relevant variables and collect more data surrounding the commercial fishing locations, such as the number of commercial fishing boats went out on given day, their catch count, and how long they fished at the location, the R-squared value would be significantly improved.

4.2 Conclusion and future work

To make our project more applicable to the parties utilizing our research, we set up a python tkinter GUI as if our model had a dependable R squared value so that tourists, fishing guides, business owners, and Alaska Fish & Game could have a tool that helped in determining the days fish count. The reason we did this exercise was to understand the best way people could interact with our model and what that would look like. Users can input the water temperature, moon phase, discharge, drift location, and nets data points. When the user hits ”Predict” it then uses our linear regression model to give the predicted fish count. We limited this GUI to only include the statistically significant factors – but it could be expanded to include more.

Although we were unable to create an accurate, dependable prediction model, we were able to find relationships between certain factors and their effect on fish count. We were also able to develop a working tool that could be used once a more accurate model is established and have created a solid foundation for future research on this topic.



Water Temperature 55

Moon Phase First Quarter

Discharge 10000

Drift Location All

Nets set

Fish Cour : 36151.39

Figure 4A: Fish Count Predictor GUI

5. References

[1] Southwick Associates, Inc., Willian J. Romberg, Allen E. Bingham, Gretchen B. Jennings and Robert A. Clark (2008, December). Economic Impacts and Contributions of Sportfishing in Alaska, 2007. Retrieved April 14, 2023, from Alaska Department of Fish and Game

[2] 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

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

[4] 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

[5] 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/

[6] 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

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

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