Team 55 – MGT 6203 Group Project Proposal

TEAM INFORMATION (1 point)

Name: Hayden (GT id: hblackburn3) 2 years of experience as financial data analyst and 2 years of experience data quality engineer for large manufacturing company with a background in BS Finance/Marketing.

Name: Richard Li (GT id: rli465) 5 years of experience working in semiconductor manufacturing, currently working for a data science interviewing startup with a background in BS Industrial and system Engineering.

Name: Binh Vu (GT id: bvu38) 2 years of experience working in aerospace/manufacturing industry with background in BS Mechanical Engineering/Computer Science

Name Lijia Cheng (GT id: icheng94) 8 years of experience in public accounting specializing in financial statements and internal controls audit of companies with a background in BBA (Accountancy)

Name Alexandra (GT id: aprokhorova3) Risk analyst in a tech consulting company with a background in BA Linguistics, BS Electrical Engineering, and Data Science Bootcamp.

OBJECTIVE/PROBLEM (5 points)

Project Title: Improved prediction of salmon runs by analyzing of historical fish count and weather data to enhance business planning and resource allocation

Background Information:

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). High variability in fish counts can lead to days of disappointed anglers and inefficient business operations. Creating an accurate fish prediction model would provide invaluable insights for local businesses and anglers.

Using historical fish count data and weather data from the Kenai River, our goal is to see if we can create an accurate prediction model. We expect predictions of fish count to be a useful proxy for the expected demand for goods and services related to the Alaska fishing industry on a given day, which in turn allows local businesses to make better operating decisions to match supply with demand.

Problem Statement:

An accurate fish count prediction model will allow for better planning and allocation of resources for local businesses whose revenue depends on recreational fishing tourism.

Primary Research Question:

Using the data our team has collected, can we create an accurate fish count prediction model?

Supporting Research Questions:

1. Will a model providing a daily run count prediction or total season count prediction be more accurate?
2. Can we use change detection analysis to forecast an incoming large salmon run?
3. Can we use change detection model to analyze historical salmon run to help local businesses improve resource allocation?

4. What is the variable that has the most significant impact on sockeye salmon count?
5. Can a model be created to predict the number of salmon that will run on a given day?

Business Justification:

Understanding Alaska's salmon fishery is critical to the success of both the commercial fishing and tourism industries. Predicting fish count and insight into the key factors that influence salmon populations can allow businesses, tourists, and the local government to make more informed business decisions:

- Local businesses – Fishing stores, tourism, and hospitality services could benefit from accurate salmon predictions to manage supply and meet anticipated demand.
- Alaska Fish and Game (city/government) - Publicize relevant data to locals and visitors. This effort could also be used to manage the fishery, prevent overfishing, and maintain sustainable fish populations
- Visitors – Help visitors plan their trip to Alaska based on probabilities for fish run on a given day. Setting realistic expectations is the key to an enjoyable experience and increases the odds of returning to Alaska in the future.

DATASET/PLAN FOR DATA (4 points)

There are six different datasets that will be used containing data on fish counts, precipitation, temperature, water, and moon phases (proxy for tide data). See *Table (1)* in the appendix for full details. An example of the cleaned and joined dataset is below:

Date	Fish count	Precipitation (in)	Air Temperature (degF)	Stage	discharge (ft3/s)	water temp (F)	lunar
7/1/2022	5034	0	57.5	9.92	15000	59.72	3
7/2/2022	4212	0	56.5	10.02	15200	59	4
7/3/2022	5796	0	58.5	10.04	15300	57.2	5
7/4/2022	7314	0	57.5	10.08	15300	56.84	6

Key Variables:

(1) Dependent: Daily number of sockeye salmon. (2) Independent: Air temperature, water temperature, precipitation, river flood level in stages, river discharge, moon phase

For some models that we plan to use like logistic regression, we also need to create new indicator variables. (a) Large run: 0 when run is small, 1 when run is large (cutoff size to be determined). (b) River discharge: Low, medium, high categories. (c) Moon phase: new moon, first quarter, full moon, third quarter, and (d) MoonStage: an interaction term between moon phase and flood stage for insight into tide data.

We hypothesize precipitation and moon phase to be two important variables. Lunar data can tell us about tide levels and precipitation data could be considered a proxy for the muddiness of the waters which can be a deterrent for salmon to move upriver.

APPROACH/METHODOLOGY (8 points)

Planned Approach

Our approach will be to use a combination of R and Python to do the following: (1) Data preparation: six different datasets will be cleaned and joined. Converting any unit mismatches (Celsius to Fahrenheit) (2) Data exploration: we will plot and inspect the datasets to detect any potential issues such as correlations, multi-collinearity, and outliers. For highly correlated variables, we will perform a feature selection process to retain only the most relevant variables. In

addition, we will investigate the trends in salmon runs over time and identify any seasonal patterns or irregularities. If the Q-Q plot and residuals vs fitted plot show non-linearity patterns, we will use log transformations. (3) Model training: our data will be split into training, validation, and test datasets. Then we plan to fit several types of ML models such as linear regression, logistic regression, k-means clustering, and CUSUM. Each of these models will give us different insights into the dataset such as identifying important variables and detecting trends in the data. (4) Model optimization: We will use cross-validation as a resampling technique and stepwise regression for variable selection/hyperparameter optimization. All regression models will be compared to a baseline R-squared/adjusted R-squared value (linear regression).

Anticipated Conclusions/Hypothesis

Through our modeling efforts – our change detection model will be the most accurate model because it will only use the historical fish count data.

With these insights and the data to back up the conclusions, we can describe characteristics of good/bad fishing days to local businesses, and they can use that data to appropriately plan staffing and inventory levels. Tourists would also be able to plan their fishing trips based on the variables that we find significant which will improve their satisfaction and increase the likelihood of them returning to Alaska to sustain the tourism economy.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

The fishing industry is one of the key drivers of tourism and business in Alaska. The ability to forecast the incoming fish population will allow the authorities to better regulate the fish population and help maintain ecological balance. Local businesses, such as hotels, restaurants, and local fishermen, can calibrate their business needs in accordance with the fish population forecast. In general, Alaska as a state could better allocate funding and resources based on our model prediction.

PROJECT TIMELINE/PLANNING (2 points)

- March 11 (Sat) - Data cleaning and joining datasets
- March 18 (Sat) - Data exploration and visuals
- March 25 (Sat) - Model training and optimization, Record project plan video presentation
- April 1 (Sat) - Project progress report
- April 8 (Sat) - Project report and slides
- April 15 (Sat) - Record final video presentation

Appendix:

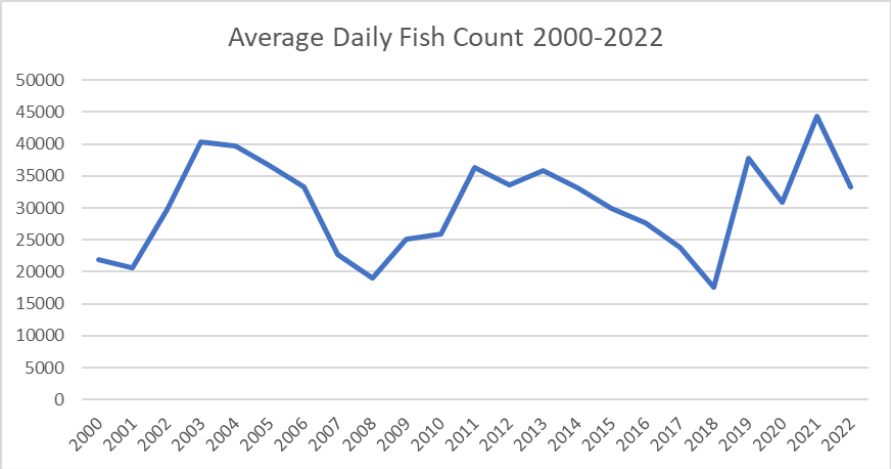

Figure (1): Average daily fish count for late-run sockeye salmon in the Kenai River from 2000-2022

Description	Period	Data Source
Daily sockeye salmon fish count at Kenai River (Late-Run Sockeye)	2000 – 2022 (July and August)	https://www.adfg.alaska.gov/sf/FishCounts/index.cfm?ADFG=main.displayResults
Daily minimum, maximum and mean air temperatures at Kenai Airport	1.1.2000 to 12.31.2022	https://akclimate.org/data/data-portal/
Daily Precipitation at Kenai Airport	1.1.2000 to 12.31.2022	https://akclimate.org/data/data-portal/
Daily river flood stage at Kenai River (Kenai Keys). It acts as proxy for river water level.	7.27.1981 to 2.9.2023	https://www.weather.gov/aprfc/rivobs
Kenai River water temperature data, Kenai River discharge data	1.1.2000 to 12.31.2022	https://waterdata.usgs.gov/nwis/
Lunar cycle data. Moon phases have a direct impact on tide levels.	1.4.1992 to 12.20.2027	https://paperswithcode.com/dataset/moon-phases

Table (1): Dataset description and data source

Citation:

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

Alaska Climate Research Center. (n.d.). Retrieved March 12, 2023, from https://akclimate.org/data/data-portal/

Hedstrom. (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

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

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

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