Survival Analysis of Salmon in the Salish Sea

Jenny Lee, Arturo Rey, Rafe Chang, Riya Eliza

26/06/2024

Contents

1. Executive Summary **3**

2. Introduction **3**

2.1 Motivation . . . 3

3. Data Products **3**

3.1 Survival Analysis . . . 4

3.1.1 Preprocessing . . . 4

3.1.2 Cormack-Jolly-Seber model with Bayesian approach . . . 4

3.1.3 Conclusion and Future Recommendations . . . 6

3.2 Outmigration Model . . . 6

3.2.1 Objective . . . 6

3.2.2 Data Collection . . . 6

3.2.3 Preprocessing . . . 6

3.2.4 XGBoost and Lasso . . . 6

3.2.5 Conclusion and Future Recommendation . . . 7

3.3 Species Imputation Model . . . 7

3.3.1 Objective . . . 7

3.3.2 Data Collection . . . 8

3.3.3 Preprocessing . . . 8

3.3.4 Deep Learning Model . . . 8

3.3.5 Result . . . 9

3.4 Species Prediction Model . . . 10

3.4.1 Objective . . . 10

3.4.2 About the data . . . 10

3.4.3 Preprocessing . . . 12

3.4.4 Ensemble modelling . . . 12

3.4.5 Voting mechanism 13
3.4.6 Final Product 13
3.5 Data Pipeline 13

4. Conclusion **14**

5. References **14**

1. Executive Summary

Our collaboration with the Pacific Salmon Foundation on the Bottleneck project will deliver visual analysis tools and statistical models to enhance biologists' understanding of salmon survival trends. By leveraging techniques from the Master of Data Science program, we aim to deepen our understanding of salmon survival probability, considering factors like predation and body size. This project is crucial due to the critical decline in returning adult salmon fish populations. Our work will provide biologists with the tools to make informed decisions and implement targeted conservation efforts, supporting the long-term sustainability of salmon populations and ecosystems.

2. Introduction

Salmons are critical to the ecosystem as they are food to 137 species (Rahr, 2023), such as grizzly bears. In British Columbia, there are over 9,000 distinct salmon populations ("State of Salmon", 2022). However, due to climate change and industrial development in the past 150 years, the population of Pacific salmon in BC has declined and their habitats have been facing unprecedented pressures.

Pacific Salmon Foundation is a non-profit organization committed to guiding the sustainable future of Pacific salmon and its habitat. The organization has a wide range of work such as community investments and salmon health. As a part of the organization's effort towards marine sciences, the Bottlenecks to Survival Projects investigate the survival bottlenecks, which refers to when a population size is reduced for at least one ("Understanding Evolution", 2024), for salmon and steelhead throughout the Salish Sea and southern BC regions.

2.1 Motivation

In ecological terms, a bottleneck refers to a specific event that results in a sharp decline in a population over a period of time (Pacific Salmon Foundation, 2021). Identifying these bottleneck points throughout the various stages of a salmon's life cycle is crucial, as it provides valuable insights into potential interventions to improve survival rates. Our study aims to provide comprehensive insights into the survival rates of salmon by observing the survival and detection probabilities across five critical stages of the salmon's outmigration-return path. This includes calculating the cumulative survival probability from the first stage to the last. By understanding these probabilities, we can identify where the most significant population declines occur and implement targeted measures to enhance survival rates.

In addition to the survival analysis, our study focuses on streamlining data processes to facilitate more accurate data retrieval and analysis. This includes developing machine learning models to predict the outmigration dates of salmon, which helps in planning and resource allocation. Accurate predictions can prevent financial losses by optimizing the timing of interventions and reducing the uncertainty associated with migration patterns. Furthermore, by improving the accuracy of data retrieval, we ensure that our findings are based on reliable and timely information, enhancing the overall effectiveness of conservation efforts.

In summary, our study not only aims to identify and address the critical bottlenecks in the salmon life cycle but also enhances the data processes to support more efficient and effective management practices. By combining survival analysis with advanced data modeling techniques, we provide a robust framework for improving the survival rates of salmon and ensuring the sustainability of their populations.

3. Data Products

For this project, our goal is to develop four key models. First, a survival analysis model based on Bayesian modeling will be created to understand the survivability of salmon fish at various life stages. Second, an

outmigration model will be developed to predict and understand the outmigration patterns primarily of Chinook and Coho salmon. Third, a species imputation model will be implemented using machine learning to fill in missing or mislabeled species data for fish. Lastly, a species prediction model will be designed to predict the species of a fish in instances where there is ambiguity during fieldwork.

3.1 Survival Analysis

In this study, we employ the survival analysis model of salmon augmented with Bayesian modeling, to estimate the survival and detection probabilities of salmon across five stages of their outmigration-return path. This approach enables us to derive precise and accurate parameter estimates, helping to identify critical bottlenecks in the salmon life cycle and providing valuable insights for potential interventions to improve survival rates.

3.1.1 Preprocessing

The data was extracted using SQL queries, stage-wise from the Strait of Georgia data center. The 5 stages identified were facility (hatchery), downstream, estuary, field (or microtroll) and return. For each of these stages, we wanted data of all wild and hatchery origin fishes. SQL queries were used to extract the required data from the data center. The columns extracted were the unique tag_id of a fish, the date on which it was detected, the "stage" at which this data is derived from, the origin of the fish (hatchery/ wild), the fork length of the fish, action (tag/ detect), species of the detected fish.

tag_id	date	stage	origin	fork_length_mm	action	species
989.001041737092	2023-05-16	downstream	hatch	17.63	detect	co
989.001041757857	2023-05-16	downstream	hatch	17.63	detect	co

Figure 1: Preprocessed data

After data from all 5 stages are extracted in this manner, the datasets then go into a preprocessing.R file to be combined together. At this stage, we make any nomenclature changes needed to mitigate the variations in data that may have happened during the collection process. Finally, we output a CSV file (survival_analysis.csv) containing the data of all detections from all 5 stages.

3.1.2 Cormack-Jolly-Seber model with Bayesian approach

Our study implemented the Cormack-Jolly-Seber (CJS) model (Cormack, 1964; Jolly, 1965; Seber, 1965), which was originally designed for studying bird migration. However, due to the model's heavy reliance on recapture rates, we incorporated Bayesian modeling to address concerns arising from the lack of recapture events. By utilizing prior knowledge through Bayesian modeling, we aim to enhance the analytical power of our sparse recapture data.

The Bayesian approach improves the precision and accuracy of our parameter estimates by integrating prior knowledge, which stabilizes estimates when data is sparse. It acts as regularization to prevent overfitting, facilitates hierarchical modeling to borrow strength across different levels, and naturally quantifies uncertainty through credible intervals. This flexibility allows for complex model specifications that better capture the underlying biological processes, enhancing the reliability of our survival and detection probability estimates despite the limitations of sparse recapture data.

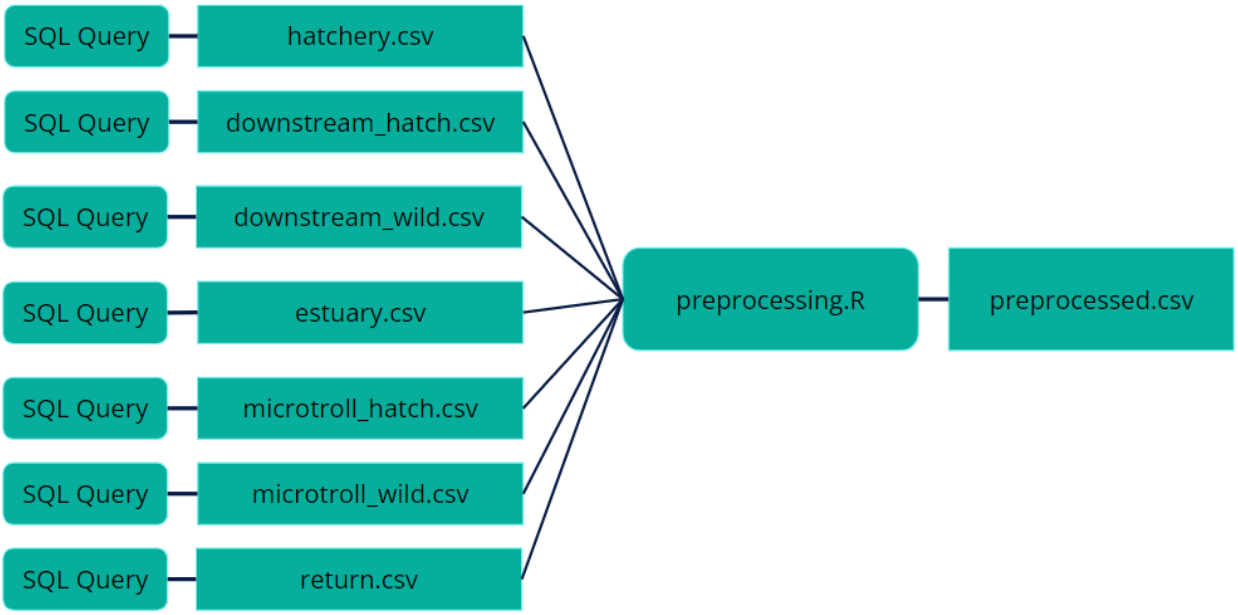

Figure 2: Proprocessing pipeline

Our data is modeled using hierarchical modeling to estimate survival and detection probabilities across multiple stages of the fish's outmigration-return path. The model combines elements of the CJS model with Bayesian techniques to handle sparse recapture data effectively.

Our prior distributions are as follows, where j indicates probabilities between stages. ϕ_j depicts the survival probability of salmon across different stages, and p_j depicts the detection probability of salmon across different stages.

$$\phi_j \sim \text{Beta}(1, 1)$$
$$p_j \sim \text{Beta}(1, 1)$$

Our likelihood distributions are as follows, where i indicates individual salmon. $z_{i,j}$ returns a binary value of either 0 or 1 to depict survival status of the fish, and $y_{i,j}$ depicts the binary value of either 0 or 1 to depict tagging status of the fish. A value of 0 indicates a negative status for the salmon, meaning it has either not survived at the stage or has not been detected. Conversely, a value of 1 indicates a positive status, meaning the salmon has survived the stage or has been detected.

$$z_{i,j} \sim \text{Bernoulli}(\phi_j, z_{i,j-1})$$
$$y_{i,j} \sim \text{Bernoulli}(p_j, z_{i,j-1})$$

Lastly, to capture the cumulative survival rates of salmon across all stages, we recursively compute the cumulative probability through each stage based on the survival probability of the past stage. Note that k represents a stage before the current stage j. The cumulative survival probability, Survship_j is calculated by taking the product of survival probabilities across all stages up to j. This approach provides an overall measure of the survival likelihood of salmon through multiple stages of their life cycle, from their origin to their return.

$$\text{Survship}_j = \prod_{k=1}^{j} \phi_k$$

This cumulative measure is crucial for understanding the overall survival dynamics of the salmon population, as it aggregates the individual stage-wise survival probabilities into a single metric that reflects the compounded likelihood of survival across the entire migratory journey.

3.1.3 Conclusion and Future Recommendations

We worked towards incorporating the origin of the salmon (e.g., hatchery versus wild) as a covariate in our models. This addition will help us understand the differential survival probabilities between hatchery-reared and wild salmon. By examining the impact of origin, we can gain insights into the factors that contribute to the survival disparities and inform targeted conservation efforts.

Mathematically, this can be represented by introducing an additional covariate origin_i for each salmon i. The model for survival probability $\phi_{i,j}$ at stage j can be modified as:

$$\log(\frac{\phi_{i,j}}{1-\phi_{i,j}}) = \beta_0 + \beta_1 \cdot \text{origin}_i$$

where β_0 is the intercept, and β_1 is the coefficient for the origin covariate.

3.2 Outmigration Model

3.2.1 Objective

The goal of this model is to accurately predict the outmigration timing of salmon. This model will take in river information including flow, level, and temperature, and then use these predictors to predict the outmigration timing of salmon. With this model, we hope to better inform the biologists at the Pacific Salmon Foundation on which day in the year that they should go out to the field to tag salmon.

3.2.2 Data Collection

The data used in this modeling is from three seperated source- the tagging history from the Pacific Salmon Foundation, the river flow and level is from the National Water Data Archive: HYDAT and the temperature data is from ECCC/MSC's National Climate Archive. The data used in for this model retrieved from five queries and stored in the data folder.

3.2.3 Preprocessing

After using SQL to retrieve the data from the Strait of Georgia Data Center, we used two functions to preprocess the data before using them to train the model. The first function is used to contact the 2014-2020 data and 2021-2023 data. The output, along with flow, level, and temperature will then be preprocessed with the preprocessing function that merges the data frames together; the function will also create the rolling mean of 30-35, 30-40, and 30-45 days prior to a given tagging date.

3.2.4 XGBoost and Lasso

We employed an XGBoost model (extreme gradient boosting) to predict salmon outmigration on specific dates. XGBoost is used because of its high accuracy- it creates models to fit the error of the previous model prediction. After experiencing linear regression, logistic regression, and SARIMAX modeling, XGBoost produces the most desirable outcome as it is able to work with sparse data better than the other modeling methods.

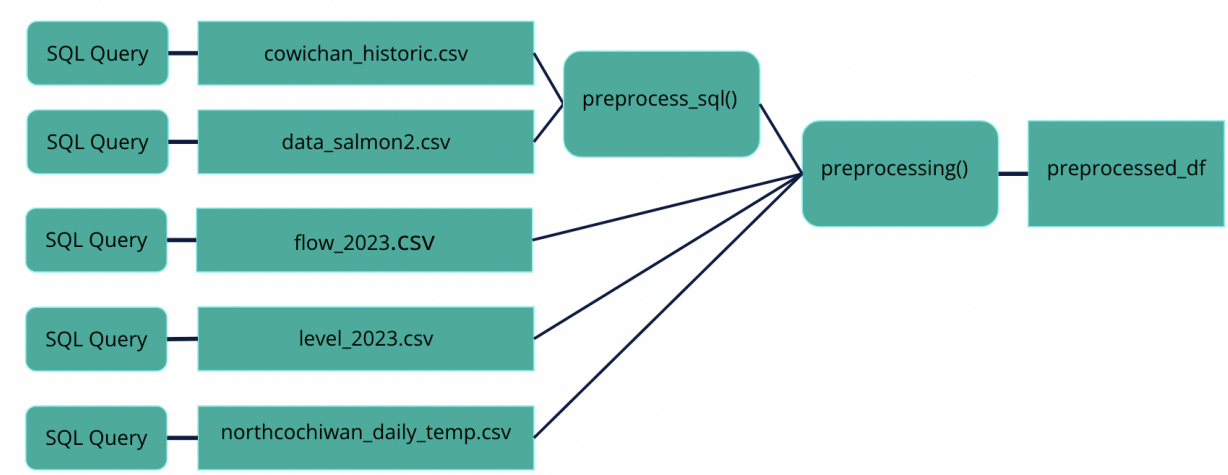

Figure 3: preprocessing

Since we have more than 20 predictors, the model is likely to overfit the training set. To avoid overfitting, we conducted feature selection where we used Lasso regression to identify the most significant variables, as lasso is the most widely used feature selection method. We also tune the alpha value (strength of penalty) in Lasso with validation on 5 different splits; the splits were created without shuffling, which fits the time series nature of the data.

The model will train based on the user's input on predicting the year (e.g. if the the user wants to predict for 2022, the model will use everything prior to 2022 as the training set.)

For prediction, we applied a function that will turn all negative predictions to 0, which means that there was no outmigration on that date- then the model will return a range of dates on the predicted year based on the lower and upper quartile user decides.

3.2.5 Conclusion and Future Recommendation

We hope to better inform biologists on the outmigration date with this model. As this model is created with time constraints and with only 10 years of data, there is rooms of improvement in future work such as exploring different modeling approaches and including different potential predictors.

3.3 Species Imputation Model

3.3.1 Objective

The goal of this model was to make the data stored in the data center as complete as possible by imputing the species of the fish wherever possible and/or necessary. This was to be done on data that had been collected in the past by experts doing field work.

1. To impute data in places where the species of a fish was not recorded
2. To detect and correct mislabeled species if any

3.3.2 Data Collection

The data for training had to be confirmed data. Out of the 57k data points (fishes) in the field table, 5000 of them had their species confirmed by the genetics lab. This became our training set. The extracted dataset (from the data center) had the following columns,

date	watershed	river	site	method	local	water_temp_start	fork_length_mm	annoted_species	confirmed_species	tag_id_long
2021-06-02	puntledge	puntledge	above tsolum	beach seine	in-river	13.4	85	ck	coho	989.001038864826
2021-06-02	puntledge	puntledge	above tsolum	beach seine	in-river	13.4	85	ck	chinook	989.001038864862

Figure 4: Training table

Where `annotated_species` was the field detection and `confirmed_species` was the result from the lab. The train set had 5014 rows.

3.3.3 Preprocessing

In preprocessing the data, several steps were taken to prepare it for modeling. First, any null values present in the dataset were removed. Next, two new features were extracted from the date column: the day of the year, represented as a whole number between 1 and 365, and the year, which ranged from 2021 to 2023. The tag_id column, being unique to each entry, was removed from the dataset. Additionally, standard scaling was applied to the fork length and day of the year features to normalize their values, and the remaining categorical features were one-hot encoded to convert them into a format suitable for ingestion into a machine learning model.

3.3.4 Deep Learning Model

The model that we finally landed on was a Deep learning neural network on a tensorflow framework. The reason for finalizing this model was due to its validation accuracy (95%) and due to its ability to detect subtle nuances in the dataset.

The model had 4 layers; 1 input, 2 hidden and 1 output layer. The layers and parameters of the model is as shown below:

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               2816

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 3)                 195

=================================================================
Total params: 11,267
Trainable params: 11,267
Non-trainable params: 0
_________________________________________________________________
```

Figure 5: Parameters of the model

The model architecture consists of the input layer with a ReLU activation function that takes in all the features, followed by two hidden layers and finally a softmax layer to output prediction probabilities for the species of each fish.

The diagram below shows the model output given that the following features of one fish are identified and fed into the model.

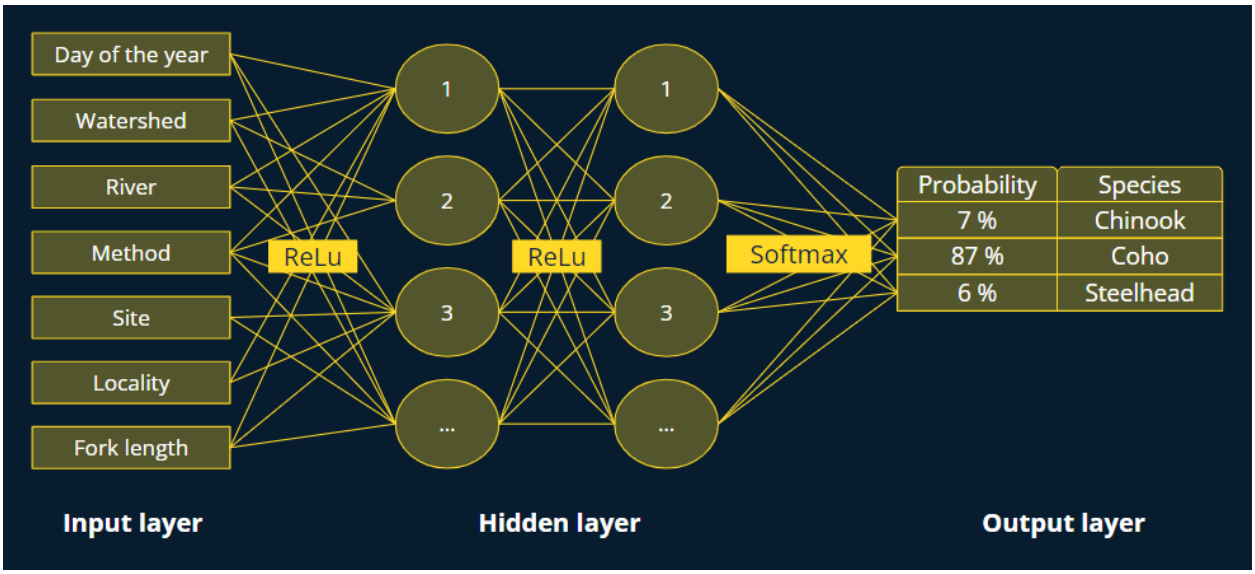

Figure 6: Deep learning prediction

The model tested with a 95% accuracy, where the accuracy is defined as the number of correct predictions in comparison to the results from the genetics lab. The model training was steady along 20 epochs, where an epoch is defined as one complete pass through the entire training dataset during machine learning model training. The accuracy and loss for both train and validation sets can be seen here.

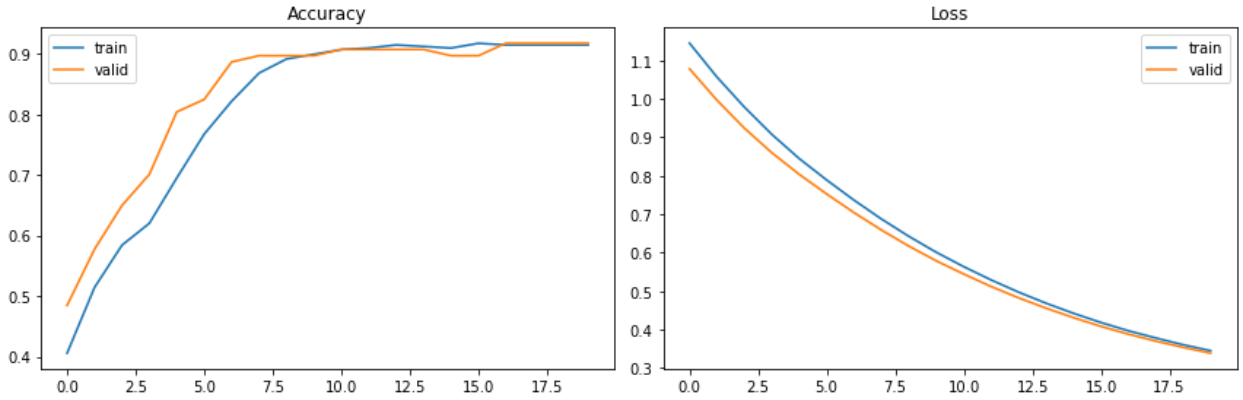

Figure 7: Accuracy and loss curves

3.3.5 Result

The results of this model are finally delivered as a CSV file with the tag ID of a fish, the field identified label, the predicted label and the predicted probability.

Tag ID	Species	Prediction	Predicted probability
989.0010419	ck	chinook	0.701667
989.0010389	ck	chinook	0.846776
989.0010419	ck	chinook	0.685648
989.0010449	co	chinook	0.497474
989.001042	ck	chinook	0.700291

Figure 8: Prediction table

3.4 Species Prediction Model

3.4.1 Objective

The goal of this model is to predict the species of a fish given its physical features, location, and site where it was detected.

3.4.2 About the data

The data needed to train this model is a combination of deterministic features (the physical features of the fish) and non-deterministic features (the location, site method, and locality of the fish).

The deterministic features for each species of fish are as follows:

1. Eye size
2. Snout shape
3. Parr marks
4. Parr marks length
5. Spotting density
6. Fin type
7. Parr marks spacing
8. Spotting characteristic

The non-deterministic features are as follows:

1. Location
2. Site method
3. Locality

These two sets of features are merged to create a complete dataset that is processed and fed into the probabilistic models. For the deterministic decision trees, only the deterministic features of the fish were used.

	species	eye_size	snout_shape	parr_marks	parr_marks_length	spotting_density	fin_type
0	ck	large	pointy	slightly faded	long	medium	anal fin
1	co	large	short and blunt	slightly faded	long	medium	anal fin
2	cm	medium	NaN	faded	short	medium	caudal fin
3	pink	medium	NaN	NaN	NaN	NaN	caudal fin
4	so	very large	NaN	slightly faded	irregular	NaN	caudal fin
5	stl	small	short and rounded	faded	short	high	caudal fin
6	ct	small	long and pointy	faded	short	high	caudal fin
7	rbt	small	short and rounded	NaN	short	high	caudal fin

Figure 9: Deterministic features

	watershed	river	site	method	local	water_temp_start	fork_length_mm	species
0	englishman	center creek	center creek	smolt trap	in-river	10.2	85.0	co
1	englishman	center creek	center creek	smolt trap	in-river	10.2	85.0	co
2	englishman	center creek	center creek	smolt trap	in-river	10.2	85.0	co
3	englishman	center creek	center creek	smolt trap	in-river	10.2	86.0	co
4	englishman	center creek	center creek	smolt trap	in-river	10.2	87.0	co

Figure 10: Non-deterministic features

3.4.3 Preprocessing

For the probabilistic models, the following preprocessing techniques were carried out:

1. **Numerical Features**: The numerical features were kept as they were for the decision trees and random forest.
2. **Categorical Features**: All the categorical features were one-hot encoded.
3. **Features Count**: Finally, we had 61 features for the probabilistic models.
4. **Handling NA**: Random forest cannot handle NAs, so those rows are dropped.

And for the deterministic features, since all the deterministic features are categorical and non-ordinal, the only required processing is to transform the features into binary features using the OneHot encoding technique.

3.4.4 Ensemble modelling

The idea for the ensemble model was to create a voting classifier. As illustrated in the figure, there are two branches: the deterministic branch and the probabilistic branch.

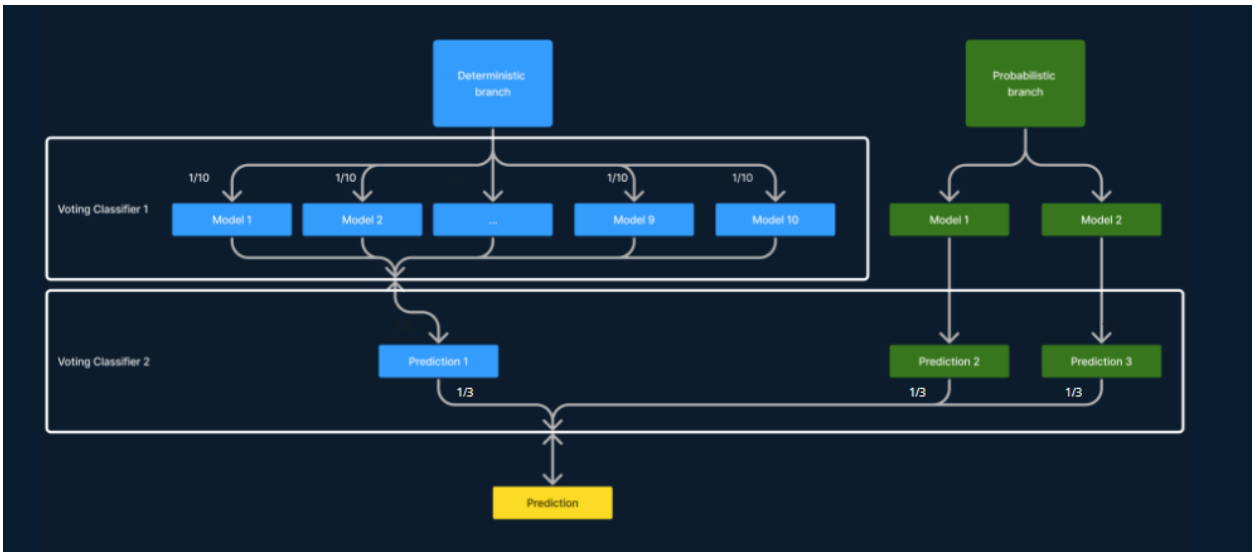

Figure 11: Voting classifier

Here, the deterministic branch consists of 10 decision trees, each randomly trained on different physical features of the fish. The prediction from each tree is combined using a majority vote, and the resulting prediction is used as the output of the deterministic branch.

The probabilistic branch includes two models:

1. **Decision Tree Classifier**: A decision tree is a tree-like model used in machine learning that makes predictions by recursively partitioning the input space into regions and assigning a label to each region based on the majority class of the training examples in that region. This model is trained on the entire feature set, including both deterministic features (physical features of the fish) and non-deterministic features (location, water temperature, site, method, etc.).
2. **Random forest**: Random Forest classifier with 100 decision trees (n_estimators=100) and a fixed random state for reproducibility is trained. This model is also trained on the complete feature set.

3.4.5 Voting mechanism

The final prediction is made by the voting classifier, which combines the outputs from both branches:

- The output from the deterministic branch contributes 1/3 of the total vote.
- The outputs from the two probabilistic models (the decision tree and random forest) each contribute 1/3 of the total vote.

The final prediction is determined by taking the majority vote from these three contributions, ensuring a balanced consideration of both deterministic and probabilistic features. This approach leverages the strengths of both branches to improve the accuracy and robustness of the species prediction.

3.4.6 Final Product

The final product delivered includes the pipeline code and a cron file. The pipeline code will be delivered in two scripts:

1. Processing File: An SQL file where all the processing mentioned in the document is performed.
2. Predict File: An SQL file where all the species predictions are performed. The cron file, however, is not mandatory due to its nature; it is an orchestration file that the partners will work on internally since only they have the knowledge of when to schedule each of the pipelines.

3.5 Data Pipeline

The primary objective of implementing a pipeline is to streamline the execution of an ETL (Extract, Transform, Load) procedure. ETL pipelines follow the sequential principle of extracting the data from the raw source, in this case the Strait of Georgia database. After this transforming the data to the desired format, and then loading it into a cleaner database, or table within the same database. By doing this, the pipeline automates the flow of data from extraction all the way to transformation to loading, ensuring an efficient process. The pipeline reads data from a database, processes it, and then utilizes a model to make predictions based on the processed data. After that, it dumps the prediction results into a specified table. This automation not only enhances efficiency of the process but also allows for up-to-date data processing, providing timely insights and reducing the need for manual intervention.

The pipeline adheres to the following schema:

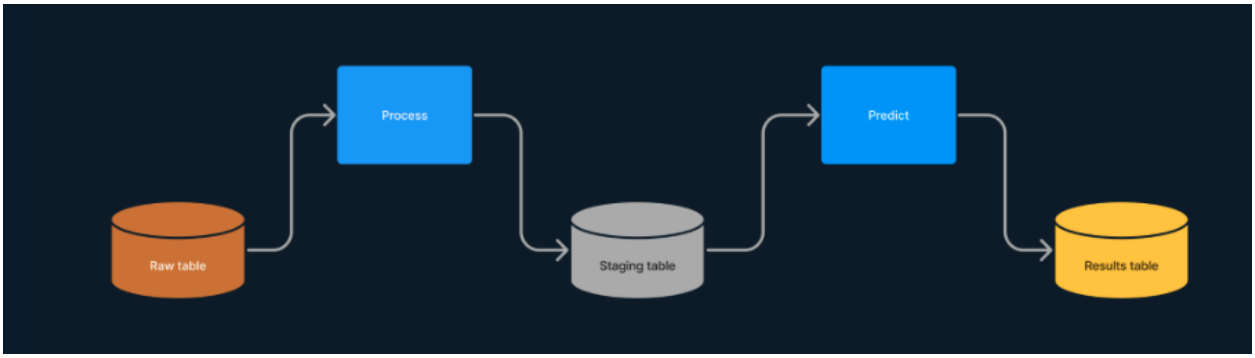

Figure 12: Flow chart of the pipeline

where:

1. Raw table corresponds to the tables the pipelines will be reading from. These tables come from the Marine Science database.

2. Process corresponds to the processing script where, depending on the pipeline, different transformations of the data will be performed.

3. Staging table corresponds to the table where all the transformed data will be stored, in order to be read again by the Predict script. Having a staging table is very standard in the industry because it helps separate the extraction and transformation from the loading part of the pipeline, ensuring an easier debugging process.

4. Predict corresponds to the script that will be performed on the transformed data. This will run the models and get a prediction which will then be stored in the Results table.

5. Results table corresponds to where all the results from the model prediction will be stored.

4. Conclusion

The number of adult salmon returning based on observed and recorded data is critically low, posing a significant threat not only to the species that rely on the nutritional value of salmon but also to British Columbia's seafood industry, which exports $1.38 billion annually. Through survival analysis, we examine the survivability across different stages of the salmon's lifetime and understand at which stage the bottleneck may occur. With the outmigration model, we aim to reduce time, cost and efforts for the team at PSF when deciding when to come to the field and observe fishes and finally with the species prediction model, we enable the completion and accuracy of the database to ensure clear future analysis. We aim to empower scientists with robust data to better confirm their hypotheses and make efficient decisions.

5. References

- Rahr, Guido. "Why Protect Salmon." Wild Salmon Center, 7 Nov. 2023, wildsalmoncenter.org/why-protect-salmon/.
- "State of Salmon." Pacific Salmon Foundation, 13 Apr. 2022, psf.ca/salmon/.
- "Understanding Evolution", https://evolution.berkeley.edu/bottlenecks-and-founder-effects/. Accessed 7 May 2024.
- "Breaking the Bottlenecks: A PSF Initiative Seeks to Identify the Danger Zones in the Salmon Life Cycle". Pacific Salmon Foundation, 2021, https://psf.ca/salmon-steward/breaking-the-bottlenecks/#:~:text=A%20salmon%20survival%20bottleneck%20is,time%2C%20ultimately%20limiting%20future%20production.