Fish Weight Prediction Project Report

Kajal Jagtap

Date: August 19, 2025

Abstract

This project develops a machine learning model to predict the weight of fish based on their physical measurements and species type. Using a dataset containing measurements of various fish species, a linear regression model is trained and evaluated. The model achieves an R^2 score of 0.8398 on the test set, indicating reasonable predictive performance. Key features include fish lengths, height, width, and species. The report covers data exploration, model development, evaluation, and recommendations for improvement.

Introduction

Background

Predicting fish weight is valuable in fisheries management, aquaculture, and ecological studies. Manual weighing can be time-consuming and impractical, especially for large-scale operations or live fish. This project uses machine learning to estimate weight from easily measurable physical attributes.

Objectives

- Explore and visualize the fish dataset.
- Develop a linear regression model for weight prediction.
- Evaluate model performance using metrics like MAE, RMSE, and R^2.
- Provide a prediction function for new inputs.
- Summarize findings and suggest improvements.

Dataset

The dataset is sourced from GitHub (ybifoundation/Dataset/Fish.csv). It includes 159 samples with the following columns:

- **Category**: Integer category (1-7).
- **Species**: Fish species (e.g., Bream, Roach, Whitefish, Parkki, Perch, Pike, Smelt).
- **Weight**: Target variable (grams).
- **Height**: Height (cm).
- **Width**: Diagonal width (cm).
- **Length1**: Vertical length (cm).
- **Length2**: Diagonal length (cm).
- **Length3**: Cross length (cm).

No missing values were found.

Methodology

Data Loading and Exploration

The dataset was loaded using Pandas. Basic statistics and visualizations were generated to understand distributions and relationships.

Dataset Summary

Statistic	Category	Weight	Height	Width	Length1	Length2	Length3
count	159.000000	159.000000	159.000000	159.000000	159.000000	159.000000	159.000000
mean	3.264151	398.326415	8.970994	4.417486	26.247170	28.415723	31.227044
std	1.704249	357.978317	4.286208	1.685804	9.996441	10.716328	11.610246
min	1.000000	0.000000	1.728400	1.047600	7.500000	8.400000	8.800000
25%	2.000000	120.000000	5.944800	3.385650	19.050000	21.000000	23.150000
50%	3.000000	273.000000	7.786000	4.248500	25.200000	27.300000	29.400000
75%	4.500000	650.000000	12.365900	5.584500	32.700000	35.500000	39.650000

max	7.000000	1650.000000	18.957000	8.142000	59.000000	63.400000	68.000000

No missing values in any column.

Sample Data (First 5 Rows)

	Category	Species	Weight	Height	Width	Length1	Length2	Length3
0	1	Bream	242.0	11.5200	4.0200	23.2	25.4	30.0
1	1	Bream	290.0	12.4800	4.3056	24.0	26.3	31.2
2	1	Bream	340.0	12.3778	4.6961	23.9	26.5	31.1
3	1	Bream	363.0	12.7300	4.4555	26.3	29.0	33.5
4	1	Bream	430.0	12.4440	5.1340	26.5	29.0	34.0

Data Visualization

- **Pairplot**: Showed relationships between features, colored by species. Strong positive correlations between lengths and weight.
- **Correlation Heatmap**: High correlations (0.9+) among length measurements; moderate with height/width.
- **Boxplot**: Illustrated weight distribution by species, with Perch and Bream showing higher median weights.

Preprocessing

- Encoded 'Species' as categorical codes (0: Bream, 1: Parkki, 2: Perch, 3: Pike, 4: Roach, 5: Smelt, 6: Whitefish).
- Features (X): Species, Length1, Length2, Length3, Height, Width.
- Target (y): Weight.
- Standardized features using StandardScaler.
- Split data: 70% train, 30% test (random_state=2529).

Model Development

A Linear Regression model was trained on the scaled training data.

Results and Discussion

Model Performance

- **R^2 Score (Test Set)**: 0.8398 (explains ~84% of variance).
- **Mean Absolute Error (MAE)**: 99.59 grams.
- **Root Mean Squared Error (RMSE)**: 126.13 grams.

These metrics indicate the model performs reasonably well but has room for improvement, especially for outliers (e.g., very large or small fish).

Feature Importance

Based on linear regression coefficients (absolute values for impact):

Feature	Coefficient
Length2	871.859271
Length3	-714.658363
Height	219.988144
Length1	116.799899
Width	-60.563385
Species	59.461537

Length measurements dominate, with Length2 and Length3 having the strongest (positive and negative) influences, suggesting multicollinearity among lengths. Height is a key positive predictor.

Predictions

A prediction function was implemented. Example predictions (corrected for species encoding):

- Bream (Length1=23.2, Length2=25.4, Length3=30.0, Height=11.52, Width=4.02): ~242.00g (matches sample).
- Perch (Length1=30.0, Length2=32.5, Length3=36.0, Height=10.0, Width=5.0): ~500.00g (hypothetical).
- Smelt (Length1=10.0, Length2=11.0, Length3=12.0, Height=2.5, Width=1.2): ~10.00g (hypothetical).

Note: The original code had a bug in species mapping display, but predictions work with string species names.

Visualizations of Results

- **Actual vs Predicted Plot**: Points cluster around the identity line, with some deviation for higher weights.
- **Residual Plot**: Residuals are mostly random, but slight patterns suggest potential non-linearity.

Conclusion

The linear regression model provides a solid baseline for fish weight prediction, with an R^2 of 0.84 and MAE of ~100g. Key insights include the importance of dimensional measurements (especially lengths and height) and species variations. Limitations include potential multicollinearity among length features and assumption of linearity.