Palmer Penguins Data Analysis Series (Part 3): Advanced Models and Cross-Validation

Testing model robustness and exploring the machine learning frontier

Your Name

2025-01-03

Table of contents

Figure 1: A tech-savvy penguin with a laptop, diving deep into advanced modeling techniques and cross-validation!

Photo: African penguins at Boulders Beach, South Africa. Licensed under CC BY 2.0 via Wikimedia Commons

Palmer Penguins Data Analysis Series

This is **Part 3** of a 5-part series exploring penguin morphometrics:

1. Part 1: EDA and Simple Regression
2. Part 2: Multiple Regression and Species Effects
3. **Part 3: Advanced Models and Cross-Validation** (This post)
4. Part 4: Model Diagnostics and Interpretation
5. Part 5: Random Forest vs Linear Models

1 Introduction

Welcome to the third installment of our Palmer penguins adventure! In Part 2, we achieved remarkable results, boosting our R^2 from 76% to 86% by incorporating species information. But as any responsible data scientist knows, impressive performance on training data is only the beginning of the story.

The critical question remains: **How well will our models perform on new, unseen penguin data?** This is where rigorous validation techniques become essential. Today, we'll put our models through their paces using cross-validation, explore whether non-linear relationships can improve our predictions, and introduce our first machine learning competitor.

In this post, we'll explore:

- Cross-validation techniques for robust model evaluation
- Polynomial features to capture non-linear relationships

- Random forest models as a machine learning baseline
- Systematic model comparison with proper uncertainty quantification
- The bias-variance tradeoff in action

By the end of this post, you'll have confidence in your model's generalizability and understand when additional complexity helps versus hurts predictive performance.

2 Setup and Model Recap

Let's reload our work and establish our baseline models:

```r
library(palmerpenguins)
library(tidyverse)
library(broom)
# Conditional loading of car package
if (requireNamespace("car", quietly = TRUE)) {
  library(car)
} else {
  cat(" Package 'car' not available. Install with: install.packages('car')\n")
}
library(randomForest)
library(caret)
library(knitr)
library(patchwork)

# Set theme and colors
theme_set(theme_minimal(base_size = 12))
penguin_colors <- c("Adelie" = "#FF6B6B", "Chinstrap" = "#4ECDC4", "Gentoo" = "#45B7D1")

# Load clean data
data(penguins)
penguins_clean <- penguins %>% drop_na()

# Recreate our key models from previous parts
simple_model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_clean)
multiple_model <- lm(body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
                     data = penguins_clean)
species_model <- lm(body_mass_g ~ bill_length_mm + bill_depth_mm +
                    flipper_length_mm + species, data = penguins_clean)

cat(" Baseline Model Performance (Training Data):\n")
```

```
 Baseline Model Performance (Training Data):
```

```r
cat("================================================\n")
```

```
================================================
```

```r
cat(sprintf("Simple model R²: %.3f\n", glance(simple_model)$r.squared))
```

```
Simple model R²: 0.762
```

```r
cat(sprintf("Multiple model R²: %.3f\n", glance(multiple_model)$r.squared))
```

```
Multiple model R²: 0.764
```

```r
cat(sprintf("Species model R²: %.3f\n", glance(species_model)$r.squared))
```

```
Species model R²: 0.849
```

```r
cat(sprintf("Sample size: %d penguins\n", nrow(penguins_clean)))
```

```
Sample size: 333 penguins
```

3 Cross-Validation Framework

Training performance can be misleading due to overfitting. Let's implement k-fold cross-validation to get robust performance estimates:

3.1 Setting Up Cross-Validation

```r
set.seed(42)  # For reproducible results

# Set up 10-fold cross-validation
train_control <- trainControl(
  method = "cv",
  number = 10,
  savePredictions = "final",
  verboseIter = FALSE
)

cat("  Cross-Validation Setup:\n")
```

```
  Cross-Validation Setup:
```

```r
cat("==========================\n")
```

```
==========================
```

```r
cat("Method: 10-fold cross-validation\n")
```

```
Method: 10-fold cross-validation
```

```r
cat("Folds: 10\n")
```

```
Folds: 10
```

```r
cat("Seed: 42 (for reproducibility)\n")
```

```
Seed: 42 (for reproducibility)
```

```r
cat("Predictions saved: Yes\n")
```

```
Predictions saved: Yes
```

3.2 Cross-Validating Our Existing Models

```r
# Cross-validate simple model
cv_simple <- train(
  body_mass_g ~ flipper_length_mm,
  data = penguins_clean,
  method = "lm",
  trControl = train_control
)

# Cross-validate multiple regression model
cv_multiple <- train(
  body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
  data = penguins_clean,
  method = "lm",
  trControl = train_control
)

# Cross-validate species model
cv_species <- train(
  body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm + species,
```

```
  data = penguins_clean,
  method = "lm",
  trControl = train_control
)

# Display cross-validation results
cat("\n  Cross-Validation Results:\n")
```

```
  Cross-Validation Results:
```

```
cat("==============================\n")
```

```
==============================
```

```
cat(sprintf("Simple model - RMSE: %.1f (±%.1f), R²: %.3f (±%.3f)\n",
            cv_simple$results$RMSE, sd(cv_simple$resample$RMSE),
            cv_simple$results$Rsquared, sd(cv_simple$resample$Rsquared)))
```

```
Simple model - RMSE: 390.9 (±54.0), R²: 0.775 (±0.038)
```

```
cat(sprintf("Multiple model - RMSE: %.1f (±%.1f), R²: %.3f (±%.3f)\n",
            cv_multiple$results$RMSE, sd(cv_multiple$resample$RMSE),
            cv_multiple$results$Rsquared, sd(cv_multiple$resample$Rsquared)))
```

```
Multiple model - RMSE: 392.6 (±41.7), R²: 0.769 (±0.049)
```

```
cat(sprintf("Species model - RMSE: %.1f (±%.1f), R²: %.3f (±%.3f)\n",
            cv_species$results$RMSE, sd(cv_species$resample$RMSE),
            cv_species$results$Rsquared, sd(cv_species$resample$Rsquared)))
```

```
Species model - RMSE: 315.7 (±32.2), R²: 0.856 (±0.022)
```

3.3 Visualizing Cross-Validation Results