Classwork 1: Building Your First Predictive Model

Introduction to Predictive Analytics in R

Prof. Asc.Endri Raco, Ph.D.

Department of Mathematical Engineering
Polytechnic University of Tirana

November 2025

Section 1

Classwork Overview

Today's Challenge: Bank Marketing Campaign

Will the client subscribe to a term deposit?

Scenario: You work for a bank's analytics team. The bank wants to run a phone marketing campaign for term deposits. Your task: build a model to identify clients most likely to subscribe.

Time: 30 minutes

Work: Individually or in pairs

Learning Objectives

By the end of this classwork, you will be able to:

1. Load and explore a real dataset
2. Create an analytical basetable
3. Fit a logistic regression model
4. Make predictions
5. Evaluate basic model performance
6. Visualize results

Dataset: Bank Marketing

Source: UCI Machine Learning Repository (simplified version)

Target Variable: `subscribed` (yes = 1, no = 0)

Features Available:

- `age`: Client's age
- `job`: Type of job
- `marital`: Marital status
- `education`: Education level
- `balance`: Average yearly balance (euros)
- `previous`: Number of previous contacts
- `campaign`: Number of contacts in this campaign

Getting Started

Step 1: Setup Your Environment

Open RStudio and create a new R script

```r
# Load required libraries
library(tidyverse)
library(ggplot2)

# Set working directory (adjust as needed)
setwd("~/classwork1")

# Load the data
bank_data <- read_csv("bank_marketing.csv")
```

Section 2

Part 1: Data Exploration (8 minutes)

Task 1.1: Initial Exploration

```r
# View first few rows
head(bank_data)

# Check structure
glimpse(bank_data)

# Get dimensions
cat("Rows:", nrow(bank_data), "\n")
cat("Columns:", ncol(bank_data), "\n")
```

Question 1: How many clients are in the dataset?

Task 1.2: Target Variable Analysis

```r
# Check target distribution
table(bank_data$subscribed)

# Calculate subscription rate
subscription_rate <- mean(bank_data$subscribed)
cat("Subscription rate:",
    round(subscription_rate * 100, 2), "%\n")
```

Question 2: What percentage of clients subscribed? Is this balanced or imbalanced?

Task 1.3: Missing Values Check

```r
# Check for missing values
colSums(is.na(bank_data))

# Percentage of missing per column
missing_pct <- colMeans(is.na(bank_data)) * 100
print(round(missing_pct, 2))
```

Question 3: Are there any missing values? If yes, which variables?

Task 1.4: Summary Statistics

```r
# Numeric variables summary
summary(bank_data %>%
          select(age, balance, previous, campaign))

# By subscription status
bank_data %>%
  group_by(subscribed) %>%
  summarise(
    mean_age = mean(age),
    mean_balance = mean(balance),
    mean_previous = mean(previous)
  )
```

Task 1.5: Quick Visualization

```r
# Age distribution by subscription
ggplot(bank_data,
       aes(x = age, fill = factor(subscribed))) +
  geom_histogram(bins = 30, alpha = 0.6,
                 position = "identity") +
  scale_fill_manual(
    values = c("0" = "red", "1" = "green"),
    labels = c("No", "Yes")
  ) +
  labs(title = "Age Distribution",
       x = "Age", y = "Count",
       fill = "Subscribed") +
  theme_minimal()
```

Section 3

Part 2: Build Logistic Regression (10 minutes)

Task 2.1: Simple Model with Age

```r
# Fit univariate logistic regression
model_simple <- glm(
  subscribed ~ age,
  data = bank_data,
  family = binomial
)

# View summary
summary(model_simple)

# Extract coefficients
coef(model_simple)
```

Question 4: Is age positively or negatively associated with subscription?

Task 2.2: Multiple Predictors Model

```r
# Fit multivariate model
model_multi <- glm(
  subscribed ~ age + balance +
               previous + campaign,
  data = bank_data,
  family = binomial
)

# View summary
summary(model_multi)
```

Task 2.3: Coefficient Interpretation

```r
# Extract and display coefficients nicely
coefs <- coef(model_multi)
coefs_df <- data.frame(
  Variable = names(coefs),
  Coefficient = round(coefs, 5),
  Sign = ifelse(coefs > 0, "Positive", "Negative")
)

print(coefs_df)
```

Question 5: Which variable has the strongest effect (largest absolute coefficient)?

Section 4

Part 3: Make Predictions (7 minutes)

Task 3.1: Single Prediction

```r
# Create a new client profile
new_client <- data.frame(
  age = 35,
  balance = 1500,
  previous = 2,
  campaign = 1
)

# Make prediction
pred_prob <- predict(model_multi,
                     newdata = new_client,
                     type = "response")

cat("Predicted probability:",
    round(pred_prob, 3), "\n")
```

Question 6: What is the probability this client subscribes?

Task 3.2: Batch Predictions

```r
# Predict for all clients in dataset
bank_data$predicted_prob <- predict(
  model_multi,
  newdata = bank_data,
  type = "response"
)

# View distribution
summary(bank_data$predicted_prob)

# Top 10 most likely to subscribe
bank_data %>%
  arrange(desc(predicted_prob)) %>%
  select(age, balance, previous,
         campaign, predicted_prob) %>%
  head(10)
```

Task 3.3: Set Decision Threshold

```r
# Use 0.5 as threshold
bank_data$predicted_class <- ifelse(
  bank_data$predicted_prob > 0.5,
  1,  # Predict "will subscribe"
  0   # Predict "will not subscribe"
)

# How many predicted to subscribe?
table(bank_data$predicted_class)
```

Question 7: How many clients are predicted to subscribe?

Section 5

Part 4: Evaluate Performance (5 minutes)

Task 4.1: Confusion Matrix

```r
# Create confusion matrix
conf_matrix <- table(
  Actual = bank_data$subscribed,
  Predicted = bank_data$predicted_class
)

print(conf_matrix)

# Calculate accuracy
accuracy <- sum(diag(conf_matrix)) /
            sum(conf_matrix)
cat("Accuracy:", round(accuracy * 100, 2), "%\n")
```

Question 8: What is the model accuracy?

Task 4.2: Additional Metrics

```r
# Extract values from confusion matrix
TN <- conf_matrix[1,1]  # True Negative
FP <- conf_matrix[1,2]  # False Positive
FN <- conf_matrix[2,1]  # False Negative
TP <- conf_matrix[2,2]  # True Positive

# Calculate metrics
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)

cat("Precision:", round(precision, 3), "\n")
cat("Recall:", round(recall, 3), "\n")
```

Task 4.3: Visualize Predictions

```
# Histogram of predicted probabilities
ggplot(bank_data,
       aes(x = predicted_prob,
           fill = factor(subscribed))) +
  geom_histogram(bins = 30, alpha = 0.6,
                 position = "identity") +
  geom_vline(xintercept = 0.5,
             color = "blue",
             linetype = "dashed", size = 1) +
  scale_fill_manual(
    values = c("0" = "red", "1" = "green"),
    labels = c("No", "Yes")) +
  labs(title = "Predicted Probabilities",
       x = "Predicted Probability",
       fill = "Actually Subscribed") +
  theme_minimal()
```

Section 6

Bonus Challenges

Challenge 1: Different Threshold

```
# Try threshold = 0.3
bank_data$pred_class_30 <- ifelse(
  bank_data$predicted_prob > 0.3, 1, 0
)

# New confusion matrix
conf_matrix_30 <- table(
  Actual = bank_data$subscribed,
  Predicted = bank_data$pred_class_30
)

print(conf_matrix_30)
```

Bonus Question: How does changing the threshold affect precision and recall?

Challenge 2: Add More Variables

```r
# Include categorical variables
# First, check unique values
unique(bank_data$job)
unique(bank_data$education)

# Fit model with categorical predictors
model_full <- glm(
  subscribed ~ age + balance + previous +
               campaign + job + education,
  data = bank_data,
  family = binomial
)

summary(model_full)
```

Bonus Question: Does including job and education improve the model?

Challenge 3: ROC Curve (Advanced)

```r
# Install and load pROC package
# install.packages("pROC")
library(pROC)

# Calculate ROC
roc_obj <- roc(bank_data$subscribed,
               bank_data$predicted_prob)

# Plot ROC curve
plot(roc_obj, main = "ROC Curve",
     col = "blue", lwd = 2)

# Calculate AUC
auc_value <- auc(roc_obj)
cat("AUC:", round(auc_value, 3), "\n")
```

Section 7

Summary & Deliverables

What You Accomplished

1. ✓ Loaded and explored real banking data
2. ✓ Built univariate and multivariate logistic regression models
3. ✓ Made predictions on new data
4. ✓ Evaluated model performance
5. ✓ Visualized results

Key Insight: You can now build a basic predictive model from scratch!

Submission Requirements

What to Submit (via Email/LMS)

1. Your completed R script (.R file) 2. Brief report answering all 8 questions 3. At least 2 visualizations (saved as PNG) 4. Optional: Bonus challenge solutions

Deadline: Before next lecture

Format: PDF or Word document

Email: nele.verbiest@pythonpredictions.com

Answer Key Template

Question 1: Number of clients = _____

Question 2: Subscription rate = _____%
Is it balanced? _____

Question 3: Missing values: _____

Question 4: Age coefficient sign: _____

Question 5: Strongest predictor: _____

Question 6: Prediction for new client: _____

Question 7: Predicted subscribers: _____

Question 8: Model accuracy: _____%

Tips for Success

Debugging Tips

- Check data types with `str()` - Use `head()` to preview data - Read error messages carefully - Check variable names (case-sensitive!)

Best Practices

- Comment your code - Use meaningful variable names - Test each step before moving forward - Ask for help if stuck $>$ 5 minutes

Common Issues & Solutions

Issue 1: "Object not found"

Solution: Check spelling and make sure you ran previous code

Issue 2: Model won't converge

Solution: Check for missing values or perfect separation

Issue 3: Visualization doesn't show

Solution: Make sure ggplot2 is loaded

Getting the Data

Option 1: Download from course website

- URL: www.pythonpredictions.com/data/bank_marketing.csv

Option 2: Use built-in sample data

```
# Generate sample data if needed
set.seed(123)
n <- 1000
bank_data <- data.frame(
  age = rnorm(n, 40, 12),
  balance = rnorm(n, 1500, 3000),
  previous = rpois(n, 0.5),
  campaign = rpois(n, 2),
  subscribed = rbinom(n, 1, 0.12)
)
```

Help Resources

During Classwork:

- Raise your hand for TA assistance
- Check with your neighbor
- Refer to lecture slides

Online Resources:

- R Documentation: `?glm`
- Stack Overflow
- Course forum

Office Hours:

- Tuesday 2-4 PM, Room 305
- Thursday 10-12 PM, Room 305

Learning Outcomes Check

After completing this classwork, you should be comfortable with:

- Using `glm()` for logistic regression
- Interpreting coefficients
- Using `predict()` function
- Creating confusion matrices
- Calculating basic metrics
- Making business recommendations from model results

Next Steps

In Next Lecture:

- Model evaluation in depth (ROC, AUC, lift charts)
- Feature engineering techniques
- Handling categorical variables
- Cross-validation

Homework:

- Complete this classwork if not finished
- Read Chapter 2 on Model Evaluation
- Prepare questions

Reflection Questions

Think about these questions (no submission required):

1. How would you explain your model to a non-technical bank manager?
2. What are the business implications of false positives vs. false negatives?
3. What additional data might improve the model?
4. How would you deploy this model in production?

Quick Reference: Key R Functions

```r
# Data loading
read_csv("file.csv")

# Model fitting
glm(y ~ x1 + x2, data = df, family = binomial)

# Predictions
predict(model, newdata = new_df, type = "response")

# Evaluation
table(actual, predicted)
mean(actual == predicted)

# Visualization
ggplot(df, aes(x, y)) + geom_histogram()
```

Grading Rubric

Component	**Points**
Code runs without errors	30
All 8 questions answered	40
Visualizations included	15
Code quality (comments, organization)	10
Bonus challenges	+5 each
Total	**100**

Ready to Start!

Let's Begin!

You have 30 minutes

Remember: Focus on understanding, not just completion

Good luck!

Additional Support Slide

Need Help?

Emergency: email endri81@gmail.com

Don't hesitate to ask questions!