Assignment 3 — Naive Bayes Classification
Author: Kristen Durkin
Date: 2025-10-12
GitHub Link: https://github.com/kdurkin5/64060-002-kdurkin5/tree/04b292f3e4314096a3a3bbeed45499f6ae031aa9/Assignment_3

### #Load & Select Required Columns

setwd("~/R_Assignments_Durkin/Assignment_3")

library(readr) UniversalBank <- read_csv("UniversalBank.csv", show_col_types = FALSE)

# Keep columns

bank <- data.frame( Loan = as.integer(UniversalBank$` Personal Loan` ), Online = as.integer(UniversalBank$Online), CC = as.integer(UniversalBank$CreditCard) )

# Validation checks

str(bank) colSums(is.na(bank)) # all should be 0 sapply(bank, range) # each should be 0..1

# Split 60% Training / 40% Validation

set.seed(123) n <- nrow(bank) idx_train <- sample(1:n, size = floor(0.6 * n)) train <- bank[idx_train, ] valid <- bank[-idx_train, ]

# Validation checks

nrow(train); nrow(valid) # ~3000 / ~2000 mean(train$Loan); mean(bank$Loan) # should be close (~0.09–0.10)

# 3 Way Pivot Table (CC × Online × Loan

pivot_A <- table(train$CC, train$Online, train$Loan) pivot_A sum(pivot_A) == nrow(train) # TRUE means all rows counted

## Empirical Conditional Probability

counts_C1_O1 <- pivot_A[“1”,“1”, ] prob_pivot <- counts_C1_O1[“1”] / sum(counts_C1_O1) prob_pivot # empirical probability ~0.10–0.12

## 2 Way Pivot Table and Six Probabilities

pivot_loan_online <- table(train$Loan, train$Online) pivot_loan_cc <- table(train$Loan, train$CC)

p_CC_given_L1 <- pivot_loan_cc[“1”,“1”] / sum(pivot_loan_cc[“1”, ]) p_On_given_L1 <- pivot_loan_online[“1”,“1”] / sum(pivot_loan_online[“1”, ]) p_L1 <- mean(train$Loan == 1) p_CC_given_L0 <- pivot_loan_cc[“0”,“1”] / sum(pivot_loan_cc[“0”, ]) p_On_given_L0 <- pivot_loan_online[“0”,“1”] / sum(pivot_loan_online[“0”, ]) p_L0 <- 1 - p_L1

round(c( `P(CC=1 | Loan=1)` = p_CC_given_L1, `P(Online=1 | Loan=1)` = p_On_given_L1, `P(Loan=1)` = p_L1, `P(CC=1 | Loan=0)` = p_CC_given_L0, `P(Online=1 | Loan=0)` = p_On_given_L0, `P(Loan=0)` = p_L0 ), 4)

## Naive Bayes Formula Estimate

num <- p_CC_given_L1 * p_On_given_L1 * p_L1 den <- num + (p_CC_given_L0 * p_On_given_L0 * p_L0) prob_nb <- num / den prob_nb # Naive Bayes formula probability

## Compare Empirical vs Naive Bayes

comparison <- data.frame( Source = c(“Pivot (empirical)”, “Naive Bayes (formula)”), `P(Loan=1 | CC=1, Online=1)` = c(prob_pivot, prob_nb) ) comparison

## Load Library - Naive Bayes Model (e1071

library(e1071)

train_nb <- transform( train, Loan = factor(ifelse(Loan==1,“Yes”,“No”), levels=c(“No”,“Yes”)), CC = factor(CC), Online = factor(Online) )

nb_model <- naiveBayes(Loan ~ CC + Online, data=train_nb)

## Model-based probability for (CC=1, Online=1)

newx <- data.frame( CC = factor(1, levels=levels(train_nbCC)), Online = factor(1, levels = levels(train$_n$bOnline)) ) prob_model <- predict(nb_model, newdata=newx, type="raw")[,"Yes"] prob_model

## Final Comparison

final_compare <- data.frame( Source = c("Pivot (empirical)", "Naive Bayes (formula)", "Model predict_proba"), `P(Loan=1 | CC=1, Online=1)` = c(prob_pivot, prob_nb, prob_model) ) final_compare

## Notes

The Naive Bayes model predicts P(Loan=1 | CC=1, Online=1) ~ 0.1106 (~11%).
This matches both the empirical estimate (≈10.7%) and the Naive Bayes formula result.
This confirms the model accurately captures the conditional relationships and that the independence assumption holds reasonably well for these predictors.

## Conclusion

The Naive Bayes model predicts P(Loan=1 | CC=1, Online=1) = 0.1106, matching the probability computed using the Naive Bayes formula. This agreement confirms that the model accurately captures the conditional relationships observed in the training data and that the independence assumption holds reasonably well for these predictors.

According to the Naive Bayes model, there's about an 11% chance that a customer who both has a credit card and uses online banking will accept a personal loan offer.