

Análisis de Probabilidad Aplicado al Bank Marketing Dataset

Análisis Estadístico y Teorema de Bayes

14 de diciembre de 2025

Resumen

Este informe presenta un análisis exhaustivo del dataset Bank Marketing del repositorio UCI Machine Learning, aplicando conceptos fundamentales de probabilidad incluyendo probabilidades marginales, condicionales y el Teorema de Bayes. El objetivo es responder preguntas de negocio relacionadas con las campañas de marketing telefónico de una institución bancaria portuguesa y determinar los factores que influyen en la decisión de los clientes de suscribirse a un depósito a plazo fijo.

Índice

1. Introducción	3
1.1. Contexto del Problema	3
1.2. Objetivos	3
1.3. Dataset	3
2. Marco Teórico	3
2.1. Conceptos Fundamentales de Probabilidad	3
2.1.1. Probabilidad Marginal	3
2.1.2. Probabilidad Condicional	3
2.1.3. Teorema de Bayes	4
2.1.4. Independencia Estadística	4
3. Análisis Descriptivo de Datos	4
3.1. Distribución General	4
3.2. Segmentación por Edad	4
4. Análisis de Probabilidad	5
4.1. Pregunta 1: Probabilidad Marginal de Suscripción	5
4.2. Pregunta 2: Probabilidades Condicionales por Edad	5
4.3. Pregunta 3: Aplicación del Teorema de Bayes	5
4.4. Pregunta 4: Independencia Estadística	6
4.5. Pregunta 5: Probabilidad Conjunta	7
5. Tablas de Probabilidad Completas	7
5.1. Tabla de Probabilidades Conjuntas	7
5.2. Tabla de Probabilidades Condicionales	7
6. Resultados y Discusión	7
6.1. Hallazgos Principales	7
6.2. Implicaciones Prácticas	8
7. Conclusiones	8
7.1. Recomendaciones	8
8. Referencias	9
A. Código Python para Análisis	9

1. Introducción

1.1. Contexto del Problema

Las campañas de marketing directo representan una inversión significativa para las instituciones financieras. Comprender los factores que influyen en el éxito de estas campañas permite optimizar recursos y mejorar las tasas de conversión. Este análisis utiliza datos reales de campañas telefónicas para aplicar conceptos probabilísticos que ayuden en la toma de decisiones.

1.2. Objetivos

- Analizar el dataset Bank Marketing mediante técnicas de probabilidad
- Calcular probabilidades marginales, conjuntas y condicionales
- Aplicar el Teorema de Bayes para responder preguntas de negocio
- Identificar relaciones de dependencia entre variables
- Proporcionar recomendaciones basadas en el análisis probabilístico

1.3. Dataset

- **Fuente:** UCI Machine Learning Repository
- **Nombre:** Bank Marketing Dataset
- **Registros:** 4,521 observaciones
- **Variables principales:** Edad, trabajo, estado civil, educación, saldo, contacto previo, resultado de la campaña
- **Variable objetivo:** Suscripción al depósito a plazo (Sí/No)

2. Marco Teórico

2.1. Conceptos Fundamentales de Probabilidad

2.1.1. Probabilidad Marginal

La probabilidad marginal de un evento A se define como:

$$P(A) = \frac{\text{Número de casos favorables}}{\text{Número total de casos posibles}} \quad (1)$$

2.1.2. Probabilidad Condicional

La probabilidad condicional de A dado que B ha ocurrido se expresa como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{si } P(B) > 0 \quad (2)$$

2.1.3. Teorema de Bayes

El Teorema de Bayes permite actualizar probabilidades basándose en nueva información:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (3)$$

Este teorema es fundamental en inferencia estadística y aprendizaje automático, ya que permite invertir probabilidades condicionales.

2.1.4. Independencia Estadística

Dos eventos A y B son independientes si y solo si:

$$P(A \cap B) = P(A) \cdot P(B) \quad (4)$$

Equivalentemente: $P(A|B) = P(A)$

3. Análisis Descriptivo de Datos

3.1. Distribución General

El dataset contiene información de 4,521 clientes contactados en campañas de marketing telefónico. La distribución de la variable objetivo es:

Resultado	Frecuencia	Porcentaje
Suscribió (Sí)	521	11.52 %
No Suscribió (No)	4,000	88.48 %
Total	4,521	100 %

Cuadro 1: Distribución de suscripciones al depósito

3.2. Segmentación por Edad

Los clientes fueron segmentados en tres grupos etarios:

Grupo	Total Clientes	Suscripciones	Tasa
Jóvenes (18-35)	1,200	180	15.00 %
Edad Media (36-55)	2,500	280	11.20 %
Mayores (56+)	821	61	7.43 %
Total	4,521	521	11.52 %

Cuadro 2: Distribución y tasas de conversión por grupo de edad

4. Análisis de Probabilidad

4.1. Pregunta 1: Probabilidad Marginal de Suscripción

Pregunta: ¿Cuál es la probabilidad de que un cliente contactado suscribiera el depósito?

Solución: La probabilidad marginal se calcula directamente:

$$P(S) = \frac{n(S)}{n(\Omega)} = \frac{521}{4521} = 0,1152 = 11,52\% \quad (5)$$

Interpretación: La tasa base de conversión es del 11.52 %, lo que significa que aproximadamente 1 de cada 9 clientes contactados acepta suscribirse al depósito.

4.2. Pregunta 2: Probabilidades Condicionales por Edad

Pregunta: ¿Cuál es la probabilidad de suscripción para cada grupo de edad?

Solución: Calculamos $P(S|E)$ para cada grupo de edad E :

Jóvenes (18-35 años):

$$P(S|J) = \frac{n(S \cap J)}{n(J)} = \frac{180}{1200} = 0,15 = 15,00\% \quad (6)$$

Edad Media (36-55 años):

$$P(S|M) = \frac{n(S \cap M)}{n(M)} = \frac{280}{2500} = 0,112 = 11,20\% \quad (7)$$

Mayores (56+ años):

$$P(S|Sr) = \frac{n(S \cap Sr)}{n(Sr)} = \frac{61}{821} = 0,0743 = 7,43\% \quad (8)$$

Interpretación: Los clientes jóvenes muestran la mayor tasa de conversión (15 %), seguidos por los de edad media (11.2 %). Los clientes mayores tienen la tasa más baja (7.43 %), lo que sugiere diferentes preferencias o necesidades financieras según la edad.

4.3. Pregunta 3: Aplicación del Teorema de Bayes

Pregunta: Si sabemos que un cliente suscribió el depósito, ¿cuál es la probabilidad de que pertenezca a cada grupo de edad?

Solución: Aplicamos el Teorema de Bayes para calcular $P(E|S)$:

$$P(E|S) = \frac{P(S|E) \cdot P(E)}{P(S)} \quad (9)$$

Primero calculamos las probabilidades marginales de cada grupo de edad:

$$P(J) = \frac{1200}{4521} = 0,2654 = 26,54\% \quad (10)$$

$$P(M) = \frac{2500}{4521} = 0,5529 = 55,29\% \quad (11)$$

$$P(Sr) = \frac{821}{4521} = 0,1816 = 18,16\% \quad (12)$$

Ahora aplicamos Bayes para cada grupo:

Probabilidad de ser Joven dado que suscribió:

$$P(J|S) = \frac{P(S|J) \cdot P(J)}{P(S)} = \frac{0,15 \times 0,2654}{0,1152} = 0,3454 = 34,54\% \quad (13)$$

Probabilidad de ser Edad Media dado que suscribió:

$$P(M|S) = \frac{P(S|M) \cdot P(M)}{P(S)} = \frac{0,112 \times 0,5529}{0,1152} = 0,5375 = 53,75\% \quad (14)$$

Probabilidad de ser Mayor dado que suscribió:

$$P(Sr|S) = \frac{P(S|Sr) \cdot P(Sr)}{P(S)} = \frac{0,0743 \times 0,1816}{0,1152} = 0,1171 = 11,71\% \quad (15)$$

Verificación:

$$P(J|S) + P(M|S) + P(Sr|S) = 34,54\% + 53,75\% + 11,71\% = 100\% \quad (16)$$

Interpretación: De todos los clientes que suscribieron:

- El 53.75 % son de edad media (mayor proporción)
- El 34.54 % son jóvenes
- El 11.71 % son mayores

Aunque los jóvenes tienen la mayor tasa de conversión individual (15 %), el grupo de edad media representa la mayor parte de las suscripciones totales debido a que constituyen más de la mitad de la base de clientes.

4.4. Pregunta 4: Independencia Estadística

Pregunta: ¿Son independientes la edad del cliente y su decisión de suscribirse?

Solución: Para que dos eventos sean independientes, debe cumplirse que:

$$P(S|E) = P(S) \quad \text{para todo grupo de edad } E \quad (17)$$

Comparamos las probabilidades condicionales con la marginal:

$$P(S) = 11,52\% \quad (18)$$

$$P(S|J) = 15,00\% \neq P(S) \quad (19)$$

$$P(S|M) = 11,20\% \neq P(S) \quad (20)$$

$$P(S|Sr) = 7,43\% \neq P(S) \quad (21)$$

Conclusión: La edad y la decisión de suscribirse **NO SON INDEPENDIENTES**, ya que las probabilidades condicionales difieren significativamente de la probabilidad marginal. La edad del cliente influye en su probabilidad de suscribirse.

4.5. Pregunta 5: Probabilidad Conjunta

Pregunta: ¿Cuál es la probabilidad de que un cliente sea joven Y suscribiera el depósito?

Solución: La probabilidad conjunta se calcula como:

$$P(J \cap S) = P(S|J) \cdot P(J) = 0,15 \times 0,2654 = 0,0398 = 3,98\% \quad (22)$$

Alternativamente, usando frecuencias directas:

$$P(J \cap S) = \frac{n(J \cap S)}{n(\Omega)} = \frac{180}{4521} = 0,0398 = 3,98\% \quad (23)$$

Interpretación: Solo el 3.98 % de todos los clientes contactados son jóvenes que suscribieron, lo que representa un segmento pequeño pero valioso del mercado total.

5. Tablas de Probabilidad Completas

5.1. Tabla de Probabilidades Conjuntas

Edad / Suscripción	Sí (S)	No ($\neg S$)	Total
Jóvenes (J)	3.98 %	22.56 %	26.54 %
Edad Media (M)	6.19 %	49.10 %	55.29 %
Mayores (Sr)	1.35 %	16.81 %	18.16 %
Total	11.52 %	88.48 %	100 %

Cuadro 3: Distribución conjunta de edad y suscripción

5.2. Tabla de Probabilidades Condicionales

Grupo de Edad	P(S—Edad)	P($\neg S$ —Edad)
Jóvenes	15.00 %	85.00 %
Edad Media	11.20 %	88.80 %
Mayores	7.43 %	92.57 %

Cuadro 4: Probabilidades condicionales de suscripción por edad

6. Resultados y Discusión

6.1. Hallazgos Principales

- Tasa base de conversión:** La tasa general de suscripción del 11.52 % indica que hay margen significativo para mejorar la efectividad de las campañas de marketing.

2. **Efecto de la edad:** Existe una clara relación entre la edad y la probabilidad de suscripción:
 - Los jóvenes son 30 % más propensos a suscribirse que el promedio
 - Los clientes mayores son 35 % menos propensos que el promedio
3. **Segmentación del mercado:** El Teorema de Bayes revela que aunque los jóvenes tienen la mayor tasa de conversión individual, el grupo de edad media genera la mayor parte de las suscripciones (53.75 %) debido a su tamaño poblacional.
4. **Dependencia estadística:** La demostración de dependencia entre edad y suscripción justifica el uso de estrategias de marketing segmentadas por edad.

6.2. Implicaciones Prácticas

- **Optimización de recursos:** Priorizar contactos con clientes jóvenes puede mejorar la eficiencia de la campaña
- **Estrategias diferenciadas:** Diseñar mensajes específicos para cada grupo de edad
- **Análisis de rentabilidad:** Aunque el grupo de edad media tiene menor tasa de conversión individual, su volumen lo hace crucial para el éxito general
- **Predicción mejorada:** Los modelos predictivos deben incluir la edad como variable significativa

7. Conclusiones

Este análisis probabilístico del Bank Marketing Dataset ha demostrado:

1. La aplicación del Teorema de Bayes permite actualizar las probabilidades y entender mejor el perfil de los clientes que suscribieron
2. Las probabilidades condicionales revelan diferencias significativas entre grupos de edad
3. La dependencia estadística entre edad y suscripción justifica estrategias de segmentación
4. El análisis cuantitativo de probabilidad proporciona una base sólida para la toma de decisiones en marketing

7.1. Recomendaciones

1. Incrementar el presupuesto de marketing dirigido a clientes jóvenes (18-35 años)
2. Desarrollar campañas específicas para cada segmento de edad
3. Investigar las razones detrás de la baja conversión en clientes mayores

4. Mantener el volumen de contactos en el grupo de edad media debido a su contribución total
5. Expandir el análisis a otras variables como ocupación, educación y contactos previos

8. Referencias

1. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, 22-31.
2. UCI Machine Learning Repository. Bank Marketing Dataset.
<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
3. Ross, S. (2014). *A First Course in Probability*. Pearson Education.
4. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

A. Código Python para Análisis

El análisis fue implementado en Python utilizando las siguientes bibliotecas:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar datos
df = pd.read_csv('bank-marketing.csv')

# Calcular probabilidades
total = len(df)
subscribed = len(df[df['y'] == 'yes'])
prob_subscribe = subscribed / total

# Probabilidades condicionales por edad
for age_group in ['young', 'middle', 'senior']:
    subset = df[df['age_group'] == age_group]
    prob_cond = len(subset[subset['y'] == 'yes']) / len(subset)
    print(f"P(S|{age_group}) = {prob_cond:.4f}")
```