Probabilistic Statistics of Titanic Dataset

Rajput, Kaustubh Rajendra

University of the Cumberlands

MSDS-530-A01: Fundamentals of Data Science

Dr. Afshin Zarenejad

January 19, 2025

Probabilistic Statistics of the Titanic Dataset

The Titanic dataset, a classic resource for statistical and machine learning exploration, provides insights into the survival rates of passengers based on various demographic and socioeconomic factors. This report delves into three key probabilistic tasks: basic probability calculations, conditional probabilities, and applying Bayes' Theorem. The chosen variables include qualitative (Sex, Survived) and quantitative (Age, Fare) factors, analyzed using Python and visualized for clarity. Below, we interpret the methodology, present the calculations, and reflect on the implications of these findings in decision-making.

Basic Probability Rules and Concepts

For the qualitative variable *Sex*, the probability of each category was calculated. Of the total 891 passengers, 577 were male, and 314 were female. The probabilities are as follows:

$$P(\text{Male}) = 577 / 891 \approx 0.647$$

$$P(\text{Female}) = 314 / 891 \approx 0.353$$

For the quantitative variables, we computed probabilities related to age and fare:

$$P(\text{Age} < 18) = 0.1583$$

$$P(\text{Fare} > 100) = 0.0672$$

Interpretation: These probabilities highlight the gender distribution and the economic stratification among Titanic passengers. The low probability of high fares reflects the limited proportion of first-class travelers.

Conditional Probability

We further explored conditional probabilities for both qualitative and quantitative variables:

Qualitative Variable (Sex): P(Female | Survived) = 0.2615 / 0.3838 = 0.6813

Interpretation: Among survivors, 68.13% were female, reflecting the prioritization of women during evacuation.

Quantitative Variable (Age): P(Age < 18 | Fare > 100) = 0.0098 / 0.1583 = 0.0619

Interpretation: Among passengers under 18, 6.19% paid fares above 100, indicating a subset of wealthier young passengers.

Bayes' Theorem

Bayes' Theorem was applied to examine survival probabilities given high fares. Using the following probabilities:

P(Survived) = 0.3838

P(Fare > 100) = 0.0672

P(Fare > 100 | Survived) = 0.1140

The updated probability of survival given a fare > 100 was calculated as:

P(Survived | Fare > 100) = 0.6511

Calculations Summary

Metric	Probability
Male Passengers	0.6476
Female Passengers	0.3524
Passengers Under 18	0.1583
Passengers with Fare > 100	0.0672
Survived Female Passengers	0.2615
Female Passengers Given Survived	0.6813
Fare > 100 Given Age < 18	0.0619
Survival Given Fare > 100	0.6511

The analysis yields key insights. Women were significantly more likely to survive, with 68.13% of survivors being female. This reflects historical prioritization of women and children during emergencies. Younger passengers (<18 years) were less likely to pay high fares, though a small subset of wealthy young passengers existed. Passengers who paid higher fares (>100) had an elevated survival probability, calculated as 65.11% using Bayes' Theorem. This suggests socioeconomic status influenced survival chances.

The findings underscore the power of probabilistic analysis in understanding historical patterns. The prioritization of women during rescue efforts highlights the influence of social norms on survival. Additionally, the link between socioeconomic status and survival emphasizes economic disparities' role in shaping outcomes. By integrating conditional probabilities and Bayes' Theorem, this analysis provides a nuanced understanding of survival determinants. Such methodologies have practical implications, offering tools for predictive modeling, policy-making, and resource allocation in disaster management.

Reference

Fundamentals of Machine Learning for Predictive Data Analytics. Algorithms, Worked Examples, and Case Studies by Kelleher et al. (ISBN: 978-0262044691)

Practical Statistics for Data Scientists: 50 Essential Concepts by Bruce and Bruce (ISBN: 978-1491952962)

Favero, L.P. and Belfiore, P., Data Science for Business and Decision Making, Academic Press, 2019.