Predicting Titanic Survival Using Logistic Regression and Gaussian Naive Bayes

CSC 4993/5573 Data Model Selection & Validation

Presented by:

Chakradhar Korlakunta

Sreeja Reddy Pasam

Akshitha Merugu

INTRODUCTION

Goal: Build a predictive model for survival outcomes in the Titanic dataset.

- Dataset contains passenger demographics, socio-economic status, and family information.

Task: Compare two statistical models and evaluate prediction performance.

Motivation: Survival depended on multiple interacting factors (class, sex, age).

Outcome: Identify strongest predictors + compare model reliability.

Dataset Summary

889 samples, 8 cleaned features + binary target

Key variables:

- Age, Fare
- Sex, Passenger Class
- Port of Embarkation
- Family Size (siblings, parents/children)

Target: survived → survived_bin (0 = died, 1 = survived)

No missing values after preprocessing

Key EDA Findings

Age Distribution

Younger passengers more frequent; few elderly survivors

Fare Distribution

Strong right-skew → richer passengers paid more

Survival Differences

- Females showed significantly higher survival rate
- 1st class passengers survived more than 3rd class

Early Hypothesis: Socio-economic status and gender are major drivers of survival

Formulated Hypotheses

H1 Women had higher survival probability

H2 Higher fare (proxy for wealth) correlates with higher survival

H3 Passenger class strongly influences survival

H4 Family size may show a non-linear effect(very high or low risky)

These hypotheses guided model choice and feature engineering

Why Logistic Regression & Gaussian Naive Bayes?

Aligned with course learning outcomes - Both models taught in lecture

Statistical interpretability - LR gives coefficients, GNB gives class-conditional means

Simple, robust, fast - Appropriate for structured tabular datasets

Different assumptions:

- LR: linear boundary, maximum likelihood
- GNB: independence assumption, generative approach

Great for comparison - Contrasting discriminative vs. generative methods

Both support probability outputs → ROC/AUC, AIC/BIC
Therefore: ideal pair for evaluation + explanation + validation

Preprocessing Pipeline

Numeric features → median imputation + StandardScaler

Categorical features → most frequent imputation + One-Hot Encoding

ColumnTransformer ensures consistent preprocessing

Train/test split: 80/20 stratified to preserve class balance

Output matrix used for LR + GNB, and later for AIC/BIC with statsmodels

MODEL A: LOGISTIC REGRESSION

Logistic Regression

- Solver = lbfgs, max_iter = 500(avoids convergence warnings)
- Produces probabilities → ROC, AUC, likelihood metrics

Key Coefficients:

- **Sex:** female strongest positive predictor
- **Class:** 1st class positive; 3rd class negative
- **Fare:** higher fare → more likely to survive

Strong interpretability → great for inference

MODEL B: GAUSSIAN NAIVE BAYES

- Assumes: features are **conditionally independent**
- Very fast to train; stable with small datasets
- Learns **class-wise Gaussian distributions**
- Performs well even when LR assumptions don’t hold
- Complements LR → helps validate robustness and model uncertainty

MODEL VALIDATION

Results & Metrics

Logistic Regression(Test Performance):

- Accuracy: ~0.79–0.82
- **AUC: 0.85**
- Strong precision/recall balance

Gaussian NB

- Accuracy: ~0.76–0.79
- **AUC: 0.82**
- Slightly lower but consistent

- LR → fewer false negatives
- GNB → more balanced FP/FN but slightly weaker overall

MODEL COMPARISON

AIC, BIC, ROC Comparison

ROC curves: LR > GNB with wider margin at medium thresholds

AIC/BIC (statsmodels Logit):

656

AIC ≈

702

BIC ≈

- Lower AIC/BIC → LR preferred over GNB in likelihood terms
- Both models validated → LR selected as final model

INFERENCE & CONCLUSIONS

Key Findings

- **Sex** was the strongest predictor (females far more likely to survive)
- **Passenger Class and Fare** → clear survival advantage for wealthier passengers
- **Logistic Regression** produced the best interpretation + best AUC
- **GNB** supported results, confirming reliability

Statistical modeling + validation metrics uncover real-world patterns

THANK YOU

Questions?