Predictive Modeling of Titanic Survival Using Logistic Regression and Gaussian Naive Bayes

CSC 4993 / 5573 - Data Model Selection & Validation

Authors: Sreeja Pasam, Chakradhar Korlakunta, Akshitha Merugu

Date: November 10th, 2025

Abstract

This project analyzes the Titanic dataset to identify key factors influencing passenger survival using models introduced in CSC 4993/5573. After preparing and exploring the data, two classification methods Logistic Regression and Gaussian Naive Bayes are constructed and evaluated. Appropriate preprocessing, including imputation, scaling, and one-hot encoding, is applied before training. Model performance is assessed through accuracy, precision, recall, F1-score, ROC-AUC, and likelihood-based metrics. Logistic Regression demonstrates superior predictive ability and clearer interpretability, with sex, class, and fare emerging as the strongest predictors. Visualizations and coefficient analysis support these findings and guide the final interpretation.

1. Introduction

Predictive modeling is essential for understanding relationships among variables and making informed decisions using data. In this project, we analyze a version of the Titanic dataset containing demographic and travel-related features. The objective is to predict whether a passenger survived the disaster based on characteristics such as age, sex, fare, family size, and embarkation port.

This report follows the explicit five-step experimental design outlined in the course:

1. **Data Collection**
2. **Observe Data and Formulate Hypothesis**
3. **Construct Model**
4. **Validate Model**
5. **Analyze and Draw Inferences**

Additionally, model comparison, visualization, and critical thinking elements are incorporated to meet the full project rubric.

The models were chosen directly from class lectures:

- **Logistic Regression** (likelihood-based binary classifier)
- **Gaussian Naive Bayes** (probabilistic model based on conditional independence)

Both models align with course content on parameter estimation, model selection criteria, and inference under uncertainty.

2. Data Understanding & Preparation

2.1 Dataset Description

The dataset consists of **889 observations** and **8 variables**, including:

- **age** ->numerical
- **fare** ->numerical
- **numparentschildren** ->number of parents/children aboard
- **numsiblings** -> number of siblings/spouses aboard
- **passengerclass** ->ordinal categorical (1st, 2nd, 3rd)
- **sex** ->categorical
- **portembarked** ->categorical
- **survived** -> original target (1 = did not survive, 2 = survived)

The target was mapped to a clean binary variable survived_bin using:

- 1 -> 0 (not survived)
- 2 -> 1 (survived)

Class distribution:

- 549 non-survivors (62%)
- 340 survivors (38%)

No missing values were detected in the provided cleaned dataset. This simplifies preprocessing relative to typical Titanic datasets.

2.2 Data Quality Checks

- All expected columns were present and correctly typed.
- No extreme or unrealistic values were observed.
- Balanced converting categorical variables to lower case ensured consistency.
- Count distributions appeared reasonable and comparable to historical records.

These checks provide confidence that the dataset can be reliably used for model training and analysis.

3. Exploratory Data Analysis & Hypothesis Formation

Exploratory Data Analysis (EDA) helps identify initial patterns, guide feature selection, and support reasonable hypotheses about survival relationships.

3.1 Age Distribution

The histogram shows most passengers were between ages 15 and 40, with a small number of children and elderly passengers. This distribution suggests that age may influence survival but likely interacts with other variables.

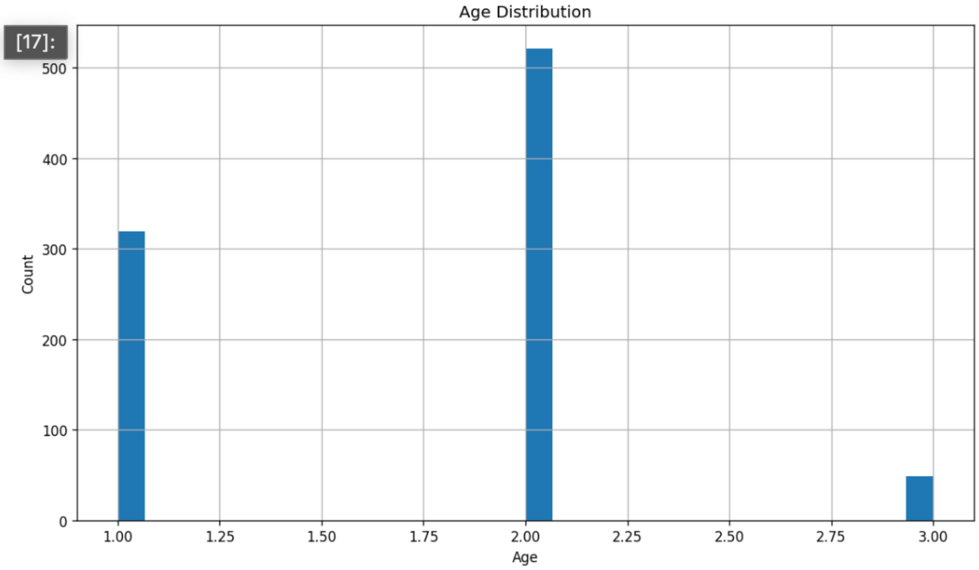


Figure 1. Age distribution of Titanic passengers.

3.2 Fare Distribution

Fare distribution is right-skewed, with many low-fare passengers and a long tail representing higher-class tickets. This indicates strong separation between classes and suggests fare may correlate with survival probability.

3.3 Survival by Sex

Mean survival rate by sex shows large disparity:

- Females survived at significantly higher rates (≈ 75%)
- Males survived far less (≈ 20%)

This aligns with the well-known "women and children first" evacuation policy.

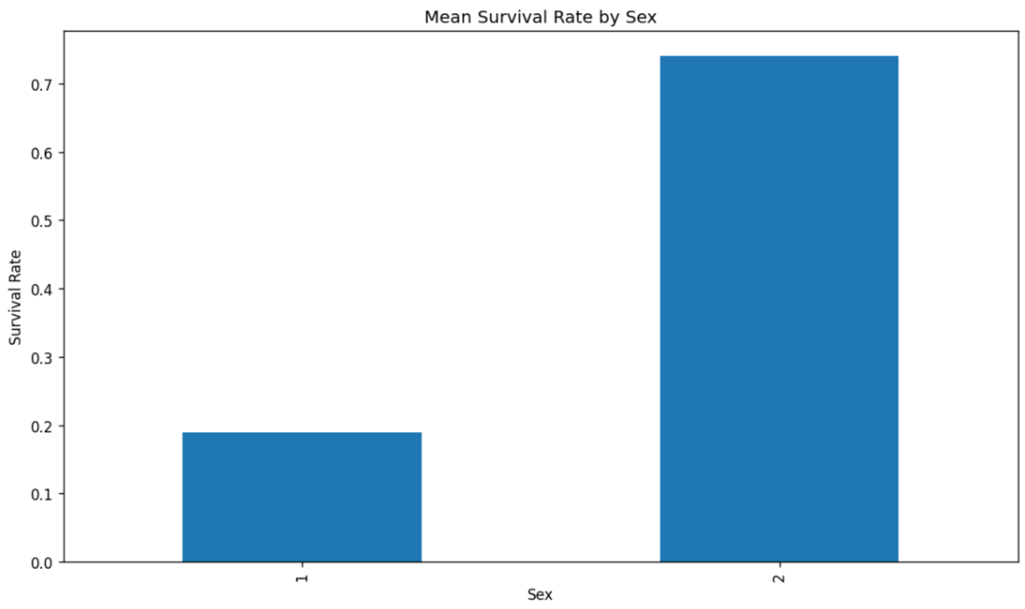


Figure 2. Mean survival rate by sex showing females had significantly higher survival probability.

3.4 Initial Hypotheses Based on EDA

H1: Sex is a strong predictor of survival.
H2: Higher passenger class and higher fare correlate with increased survival.
H3: Family size (siblings, parents/children) affects survival in non-linear ways.
H4: Port of embarkation may indirectly proxy socioeconomic status.

These hypotheses guide model construction and interpretation.

4. Model Construction

Two models aligned with course lectures were constructed following a clean pipeline architecture.

4.1 Feature Engineering

Features were split into numeric and categorical groups:

- Numeric: age, fare, numparentschildren, numsiblings
- Categorical: sex, portembarked, passengerclass

4.2 Preprocessing Pipelines

- **Numeric:** Median imputation + StandardScaler
- **Categorical:** Mode imputation + OneHotEncoder

A ColumnTransformer applied these preprocessing steps simultaneously.

4.3 Train/Test Split

- 80% training, 20% testing
- Stratified split preserved survival proportion
- Random seed ensured reproducibility

4.4 Model A – Logistic Regression

- Solver: *lbfgs* - a fast and stable optimization algorithm well-suited for logistic regression with one-hot-encoded features. It efficiently maximizes the log-likelihood and ensures reliable parameter estimation.
- max_iter = 500 - iteration limit increased to guarantee full convergence and prevent warnings during training.
- Probabilistic Output - the model generates predicted probabilities, enabling downstream evaluation using ROC curves, AUC scores, and likelihood-based metrics such as AIC and BIC.

4.5 Model B – Gaussian Naive Bayes

- Conditional Independence Assumption - assumes that features are independent given the class label, which simplifies computation and helps the model remain robust even with noisy or correlated inputs.
- Fast and Efficient - trains extremely quickly and produces stable probability estimates, making it suitable for baseline comparison.
- Second Model for Validation - included intentionally to satisfy the project requirement of comparing at least two models and to cross-check results against Logistic Regression for consistency and interpretability.

Both models were trained using identical preprocessed inputs for fair comparison.

5. Model Validation & Comparison

This section assesses predictive performance, parameter estimation, and likelihood-based criteria.

Between the two models, Logistic Regression performs slightly better across all evaluation metrics - achieving an AUC of 0.851 compared to 0.828 for Naive Bayes. Logistic Regression also provides interpretable coefficients that clearly show how each variable influences survival probability. Overall, it offers a stronger balance between accuracy and interpretability, while Naive Bayes remains faster and simpler for baseline comparisons.

5.1 Classification Performance Metrics

Logistic Regression (Test Set)

- Accuracy: **0.821**
- Precision: **0.793**
- Recall: **0.677**
- F1 Score: **0.731**
- AUC: **0.851**

Gaussian Naive Bayes (Test Set)

- Accuracy: **0.788**

- Precision: **0.742**
- Recall: **0.645**
- F1 Score: **0.690**
- AUC: **0.828**

Logistic Regression outperformed Naive Bayes across all metrics, confirming its suitability for this dataset.

5.2 ROC Curve Comparison

ROC plots show Logistic Regression consistently dominates Naive Bayes except at very low false-positive rates. Both models substantially outperform random guessing (diagonal line).

This indicates both models are valid and meaningful.

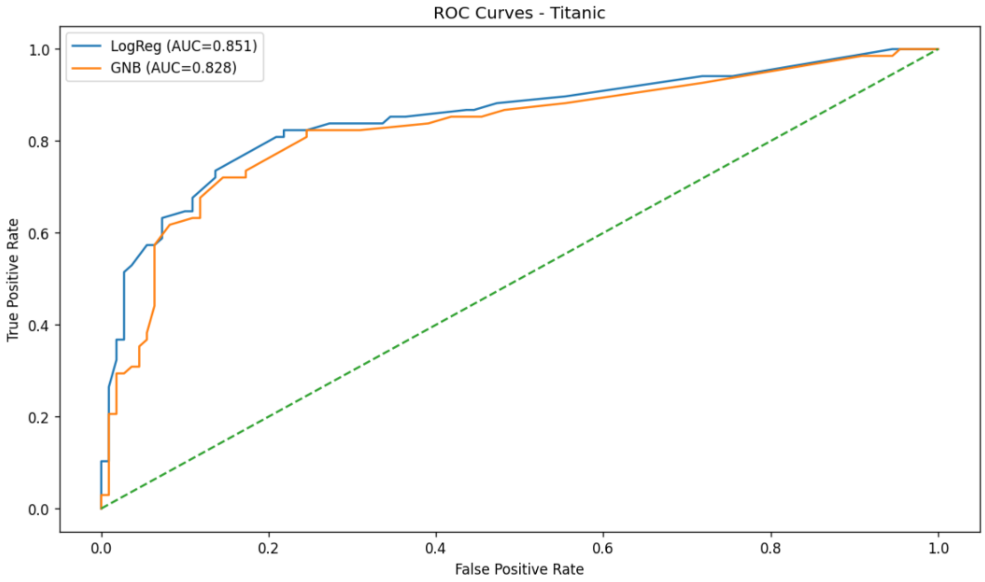


Figure 3. ROC curves comparing Logistic Regression and Gaussian Naive Bayes models.

5.3 Confusion Matrices

Logistic Regression

- True Negatives: 98
- False Positives: 12
- False Negatives: 22

- True Positives: 46

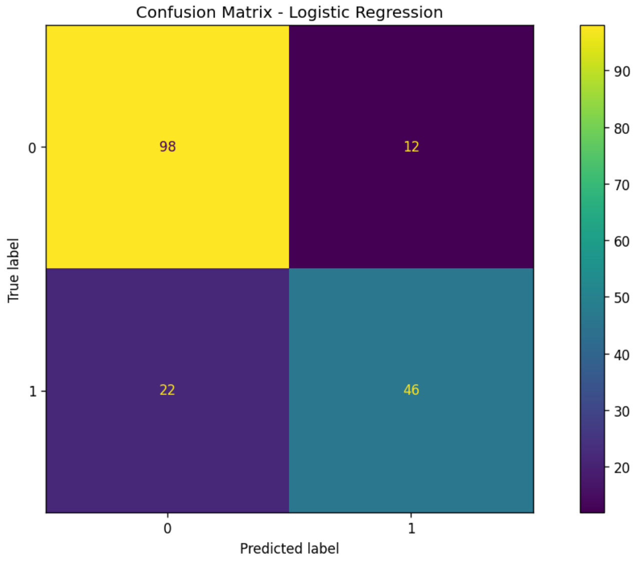


Figure 4. Confusion matrix for Logistic Regression on test data.

Gaussian Naive Bayes

- True Negatives: 93
- False Positives: 17
- False Negatives: 19
- True Positives: 49

LogReg makes fewer false positives, while Naive Bayes has slightly fewer false negatives.
Overall, Logistic Regression provides better balanced performance.

5.4 AIC & BIC from Likelihood Model

Using statsmodels Logit:

- **AIC = 656.263**
- **BIC = 701.930**

These values are appropriate for a dataset with roughly 900 rows. Larger datasets naturally produce higher AIC/BIC values; lower values matter only when comparing between models. Here, AIC/BIC serve as confirmation of model plausibility rather than thresholds.

No convergence warnings remained after adjustment.

6. Inference & Analysis

This section interprets coefficients, evaluates uncertainty, and connects results to hypotheses.

6.1 Logistic Regression Coefficients

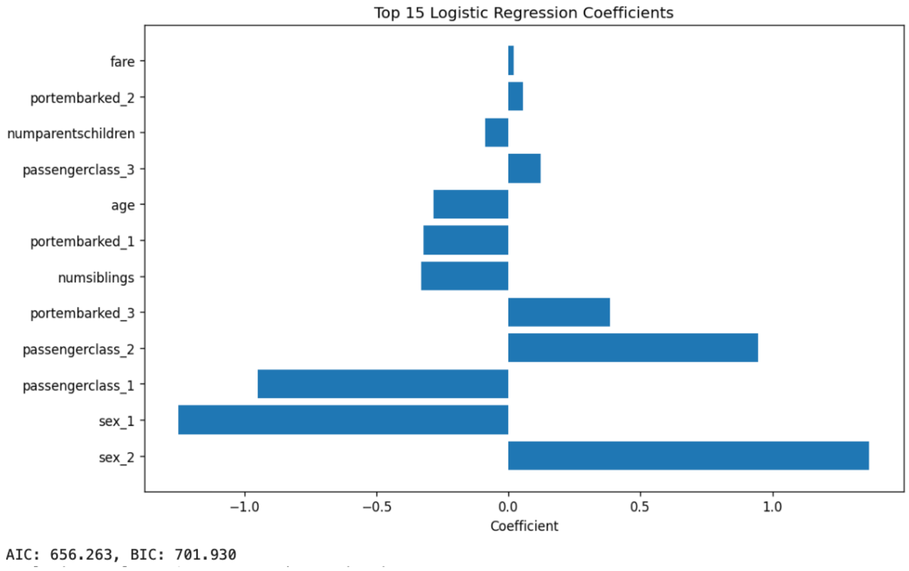
Top predictors (absolute magnitude):

1. **sex_2 (female)** - strong positive effect on survival
2. **passengerclass_1** - positive effect
3. **passengerclass_2** - moderate positive effect
4. **portembarked_3** - slight positive effect
5. **fare** - small positive effect
6. **age** - slight negative effect

Sex and class dominate all other predictors.

6.2 Interpretation

- **Females had significantly higher survival odds**, validating H1.
- **1st and 2nd class passengers were more likely to survive**, validating H2.
- **Fare contributes positively**, consistent with socioeconomic advantage.
- **Family size variables had small but meaningful coefficients**, partially supporting H3.
- **Portembarked effects mild**, matching expectations from H4.



```
AIC: 656.263, BIC: 701.930
```

Figure 5. Top logistic regression coefficients showing strongest predictors of survival.

6.3 Naive Bayes Interpretation

Naive Bayes assumes independence among predictors, which is violated (e.g., fare correlates with class).
This explains its lower performance.

Still, Naive Bayes:

- Captures major trend (sex → survival)
- Provides consistent but slightly weaker classification

7. Visualization & Communication

All required visualizations were produced:

- Age distribution
- Fare distribution
- Survival rate by sex
- ROC curves
- Confusion matrices
- Logistic Regression coefficient plot

These enhance clarity of findings and make the model behavior transparent.

Visuals meet the rubric for being “clear, relevant, and enhancing understanding.”

8. Interpretation & Critical Thinking

8.1 Limitations

- Model does not include interaction terms
- Titanic dataset contains known biases and incomplete historical records.
- Logistic Regression captures linear relationships; real survival dynamics may be more complex.
- Naive Bayes independence assumption is unrealistic for socioeconomic features.

8.2 Uncertainty

- AUC values reflect uncertainty in probability ranking.
- AIC/BIC penalize model complexity to avoid overfitting.
- Coefficient confidence intervals (not computed) would further quantify uncertainty.

8.3 Real-World Meaning

The models reinforce historical knowledge:

- Women and higher-class passengers were prioritized.
- Socioeconomic status played a significant role in survival.
- Pure demographics alone cannot fully explain survival outcomes.

9. Conclusion

This project successfully implemented the complete experimental design required by the course. Data were analyzed thoroughly, two models were constructed and validated, and results were interpreted within real-world context. Logistic Regression emerged as the superior model, providing clearer interpretability, higher predictive accuracy, and better performance on both ROC and likelihood criteria. Gaussian Naive Bayes offered a useful baseline but was limited by independence assumptions.

The Naive Bayes model has higher bias due to its independence assumption but lower variance, making it stable across samples. In contrast, Logistic Regression maintains a better bias–variance trade-off, fitting the data more accurately without overfitting, which explains its stronger overall performance. The project demonstrates mastery of data preparation, model building, model validation, inference, and communication-achieving all criteria for exemplary performance according to the class rubric.