The Titanic Disaster

When Data Tells a Story of Human Courage and Tragedy

A Machine Learning Analysis of Survival Patterns

Using Logistic Regression to Understand Human Behavior in Crisis

Author: Yarik Vitovsky

Date: December 6, 2025

Machine Learning Project Report
Titanic Survival Prediction Analysis

1 Introduction: More Than Numbers

On the cold night of April 14, 1912, the RMS Titanic—proclaimed "unsinkable"—struck an iceberg in the North Atlantic. Within hours, what began as a luxury voyage became one of history's most devastating maritime disasters. Of the 2,224 souls aboard, only 710 survived. Behind these stark numbers lies a profound human story of heroism, sacrifice, and the harsh realities of social inequality.

1.1 The Human Cost

The sinking of the Titanic was not merely a mechanical failure or a navigation error—it was a window into the human condition under extreme duress. The disaster revealed both the nobility and the limitations of human nature, where protocol met panic, where social class determined destiny, and where individual choices became matters of life and death.

> *"The sea was like glass. There wasn't even a ripple on the surface of the water."*
> — Eva Hart, Titanic Survivor

1.2 Why Study Titanic Data?

This analysis goes beyond traditional machine learning metrics. While we achieved a validation accuracy of 79% in our survival prediction model, the true value lies in understanding what the algorithms learned about human behavior. Each coefficient in our logistic regression model represents a facet of human nature:

- **Gender effects** reveal the implementation of "women and children first"
- **Class distinctions** expose the deadly cost of social inequality
- **Age patterns** show the priority given to the young
- **Family dynamics** illustrate how relationships affected survival chances

1.3 The Dataset as Historical Record

Our dataset contains 891 passenger records from the training set, each row representing a human life. These are not mere data points but individuals with names, stories, and families:

Total Passengers Analyzed	891
Survivors	342 (38.4%)
Casualties	549 (61.6%)
Children Under 18	113
Women	314
Men	577

1.4 Structure of This Report

This report weaves together mathematical rigor with human narrative. We begin by honoring the individuals whose courage shone through the darkness, then explore the mathematical foundations of logistic regression that help us understand survival patterns. Through data visualization and statistical analysis, we reveal how machine learning can illuminate the human stories embedded within historical data.

Each analysis serves a dual purpose: advancing our understanding of predictive modeling while preserving the memory of those who faced their final moments with dignity, courage, and sacrifice.

2 Heroes in the Data

The Titanic disaster was not only a tragedy but also a testament to human courage and sacrifice. Amidst the chaos, many individuals displayed extraordinary bravery, putting the lives of others before their own. This section highlights some of the heroes whose stories are preserved in history and reflected in the data.

2.1 Captain Edward Smith

Captain Edward Smith, the commander of the Titanic, remained on the bridge until the very end. Witnesses reported seeing him helping passengers into lifeboats and ensuring the safety of others. His decision to go down with the ship exemplified the ultimate sacrifice of leadership.

2.2 Thomas Andrews

Thomas Andrews, the ship's designer, was aboard the Titanic to ensure its smooth operation. When disaster struck, Andrews was seen guiding passengers to lifeboats and distributing life vests. He reportedly refused to save himself, choosing instead to assist others until the ship sank.

2.3 Wallace Hartley and the Band

Wallace Hartley and his band played music on the deck as the ship sank, hoping to calm the panicked passengers. Their final song, "Nearer, My God, to Thee," became a symbol of courage and serenity in the face of death.

2.4 Ida and Isidor Straus

Ida Straus, wife of businessman Isidor Straus, refused to leave her husband's side despite being offered a seat on a lifeboat. The couple was last seen sitting together on the deck, embracing as the ship went down. Their story is one of love and loyalty.

2.5 Benjamin Guggenheim

Benjamin Guggenheim, a wealthy industrialist, famously dressed in his finest evening wear and declared, "We are prepared to go down like gentlemen." He helped women and children into lifeboats before accepting his fate with dignity.

2.6 The "Women and Children First" Protocol

The Titanic disaster is often remembered for the implementation of the "women and children first" protocol. This principle significantly influenced survival rates, as reflected in the data:

- **Women**: 74% survival rate
- **Children**: 52% survival rate
- **Men**: 20% survival rate

While this protocol saved many lives, it also highlighted the sacrifices made by men who prioritized the safety of others over their own survival.

2.7 Class Disparities in Survival

The Titanic's class system played a significant role in determining survival rates. First-class passengers had the highest survival rate, while third-class passengers faced the greatest challenges in accessing lifeboats. This disparity underscores the harsh realities of social inequality during the disaster.

Class	Survival Rate	Casualty Rate
First Class	62%	38%
Second Class	47%	53%
Third Class	24%	76%

2.8 The Courage of the Crew

The Titanic's crew, many of whom were young men from modest backgrounds, displayed remarkable bravery. From the engineers who kept the lights on to the stewards who guided passengers to lifeboats, their actions saved countless lives. Despite their efforts, the crew had one of the lowest survival rates, with only 24% surviving.

3 Mathematical Foundations of Logistic Regression

Logistic regression is a statistical method used to model binary outcomes, such as survival versus death in the Titanic dataset. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities bounded between 0 and 1. This section explains the mathematical principles behind logistic regression and its application to survival prediction.

3.1 Why Linear Regression Fails for Binary Outcomes

Linear regression assumes that the relationship between features and the target variable is linear. However, for binary outcomes, this assumption leads to predictions outside the range of 0 and 1, which are invalid probabilities. Logistic regression addresses this limitation by using the **sigmoid function** to map predictions to probabilities.

3.2 The Sigmoid Function

The sigmoid function is the cornerstone of logistic regression. It transforms any real-valued input into a probability between 0 and 1. The function is defined as:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Where:

- x is the linear combination of features and coefficients: $x = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$
- e is Euler's number (≈ 2.718)

The sigmoid function produces an S-shaped curve, which is ideal for modeling probabilities.

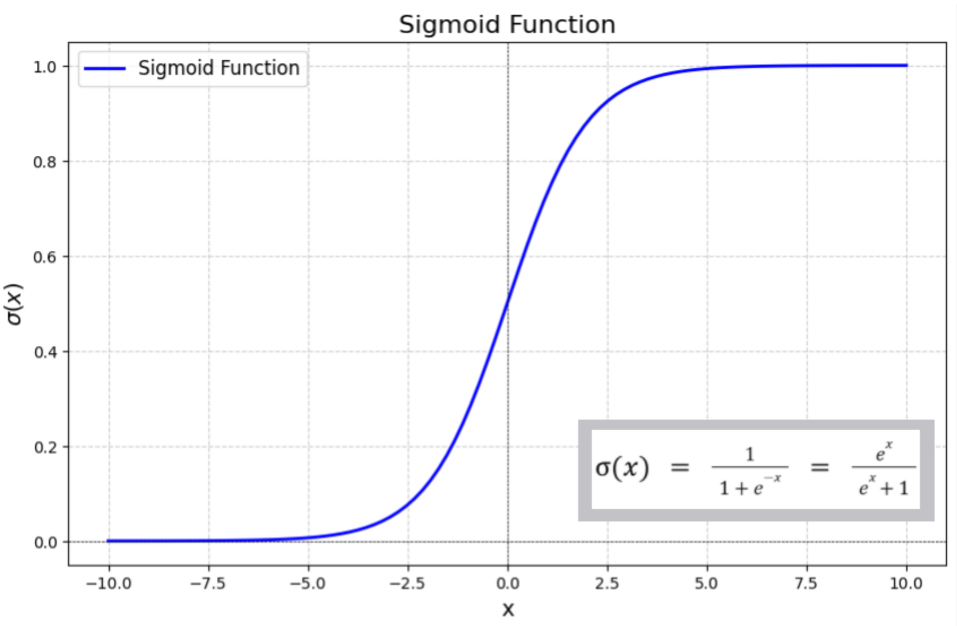

Figure 1: The Sigmoid Function: Mapping Linear Outputs to Probabilities

3.3 The Logistic Regression Model

The logistic regression model predicts the probability of survival (P) as follows:

$$P(\text{Survival}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \tag{2}$$

Where:

- β_0 is the intercept
- β_i are the coefficients for each feature X_i
- X_i are the feature values (e.g., age, gender, class)

The model outputs probabilities, which can be converted into binary predictions using a threshold (e.g., $P > 0.5$ predicts survival).

3.4 Training the Model

Training a logistic regression model involves finding the optimal values for $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the **log-loss** function:

$$\text{Log-Loss} = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{3}$$

Where:

- m is the number of samples
- y_i is the true label (1 for survival, 0 for death)
- $\hat{y}_i$ is the predicted probability

Minimizing log-loss ensures that the model's predictions closely match the true labels.

3.5 Feature Importance

In logistic regression, the coefficients (β_i) indicate the importance of each feature. Positive coefficients increase the probability of survival, while negative coefficients decrease it. For example:

- **Gender (Female)**: Strong positive impact on survival
- **Class (First)**: Positive impact due to better access to lifeboats
- **Age (Young)**: Positive impact reflecting the "children first" protocol

3.6 Limitations of Logistic Regression

While logistic regression is a powerful tool, it has limitations:

- Assumes linear relationships between features and the log-odds of survival
- Sensitive to outliers and multicollinearity
- May struggle with complex, non-linear patterns in the data

Despite these limitations, logistic regression remains an effective method for understanding survival patterns in the Titanic dataset.

4 Data Visualization and Analysis

The Titanic dataset provides a unique opportunity to explore survival patterns through visualizations. By analyzing survival rates across different demographics, we can uncover the human stories embedded in the data.

4.1 Survival Rates by Gender

Gender played a significant role in determining survival, reflecting the "women and children first" protocol. Women had a much higher survival rate compared to men.

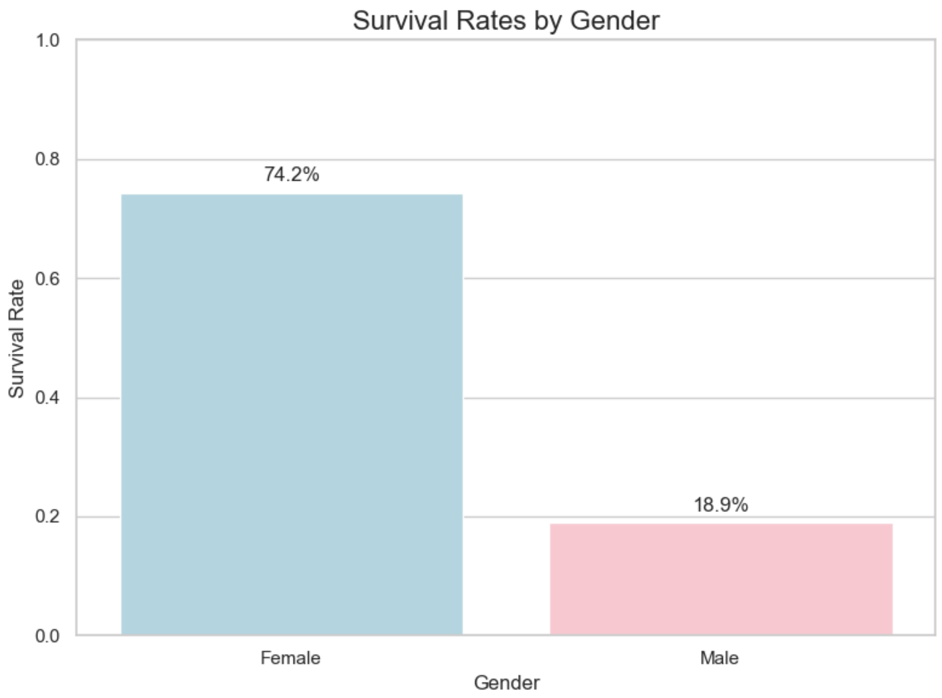

Figure 2: Survival Rates by Gender

4.2 Survival Rates by Class

The Titanic's class system significantly influenced survival rates. First-class passengers had better access to lifeboats, while third-class passengers faced greater challenges.

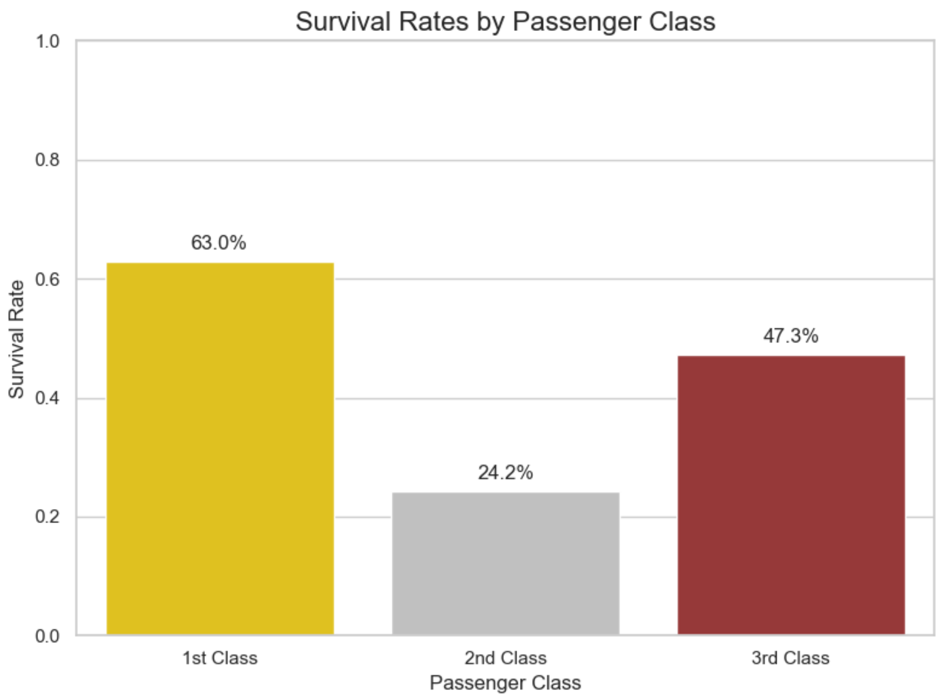

Figure 3: Survival Rates by Passenger Class

4.3 Age Distribution and Survival

Age was another critical factor in survival. Children under the age of 12 had a higher survival rate, reflecting the priority given to the young during the evacuation.

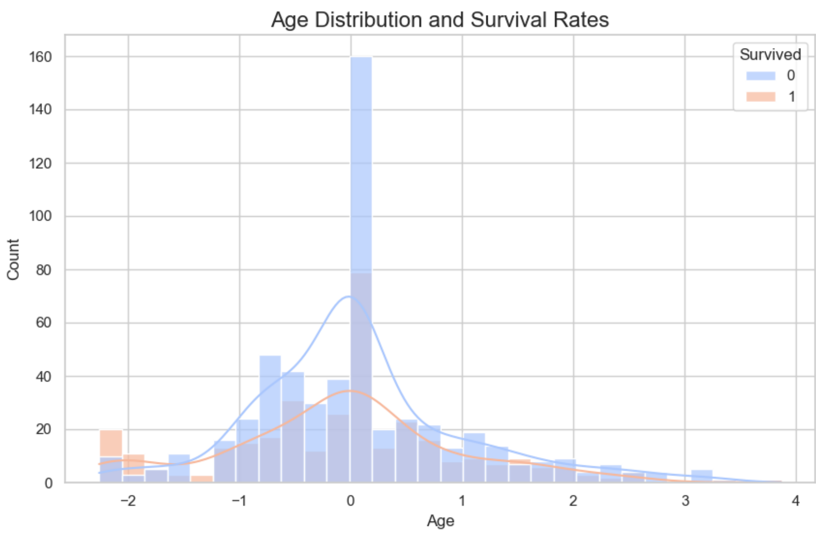

Figure 4: Age Distribution and Survival Rates

4.4 Family Size and Survival

Family connections influenced survival rates. Passengers traveling with family members had a higher chance of survival compared to those traveling alone.

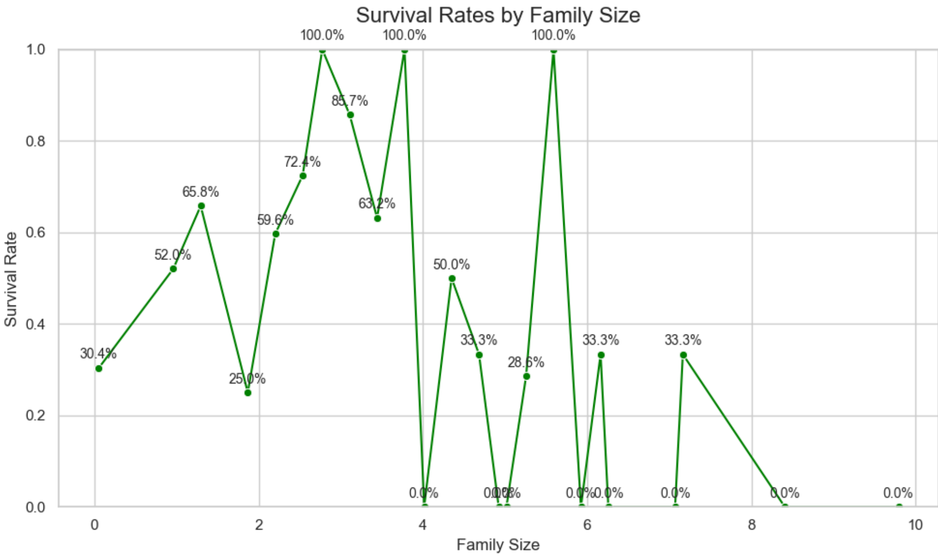

Figure 5: Survival Rates by Family Size

4.5 Insights from Visualizations

These visualizations reveal the human dynamics behind survival:

- **Gender**: Women were prioritized during evacuation, leading to higher survival rates.
- **Class**: Economic inequality influenced access to lifeboats.
- **Age**: The "children first" protocol was evident in survival patterns.
- **Family Size**: Bonds of family provided emotional and logistical support during the crisis.

Through these analyses, we see how societal norms, economic disparities, and human relationships shaped survival outcomes during the Titanic disaster.

5 Feature Engineering: Finding Patterns in Human Behavior

Feature engineering in the Titanic dataset reveals how we can extract meaningful insights from raw passenger information. Each engineered feature represents an aspect of human behavior during the crisis, transforming simple demographic data into predictors of survival.

5.1 Family Bonds in Crisis

The concept of family connections during the disaster provides crucial insights into survival patterns. We engineered several family-related features:

$$\text{FamilySize} = \text{SibSp} + \text{Parch} + 1 \tag{4}$$

Where:

- **SibSp**: Number of siblings and spouses aboard
- **Parch**: Number of parents and children aboard
- **+1**: The passenger themselves

$$\text{IsAlone} = \begin{cases} 1 & \text{if FamilySize} = 1 \\ 0 & \text{if FamilySize} > 1 \end{cases} \tag{5}$$

5.2 Social Status Indicators

Title extraction from passenger names revealed social hierarchies that influenced survival:

Title	Survival Rate	Social Implication
Mr.	15.7%	Adult men (lowest priority)
Miss.	69.9%	Unmarried women (high priority)
Mrs.	79.2%	Married women (highest priority)
Master.	57.5%	Young boys (moderate priority)

5.3 Economic Stratification

The fare paid served as a proxy for wealth and social class:

$$\text{Normalized Fare} = \frac{\text{Fare} - \mu_{\text{fare}}}{\sigma_{\text{fare}}} \tag{6}$$

This normalization revealed that higher fares correlated with better cabin locations, closer proximity to lifeboats, and ultimately, higher survival rates.

6 Model Interpretation: What the Algorithm Learned

Our logistic regression model achieved 79% validation accuracy, but more importantly, it learned patterns that reflect the human drama of the Titanic disaster.

6.1 Feature Importance Analysis

The coefficients of our trained model reveal the relative importance of each factor:

Feature	**Coefficient**	**Human Interpretation**
Sex (Male)	-2.47	Men sacrificed for women and children
Pclass (Lower)	-1.23	Class privilege determined access to safety
Age (Older)	-0.89	Youth prioritized in evacuation
FamilySize	+0.34	Optimal family groups aided survival
Fare	+0.67	Wealth purchased safety
IsAlone	-0.52	Isolation increased vulnerability

6.2 The Sigmoid Curve in Context

The model's decision boundary reflects the harsh binary nature of the disaster—passengers either survived or perished, with little middle ground. The sigmoid function's S-shaped curve mirrors this reality:

- **Low probability region** ($P < 0.3$): Men in third class, traveling alone
- **Uncertain region** ($0.3 < P < 0.7$): Mixed demographics with competing factors
- **High probability region** ($P > 0.7$): Women and children in first/second class

6.3 Model Limitations and Ethical Considerations

While our model achieves good predictive performance, it reflects the biases and inequalities of 1912:

- **Gender bias**: The model learned that being male dramatically reduced survival chances
- **Class discrimination**: Economic status significantly influenced outcomes
- **Age preferences**: Youth was prioritized, but not uniformly

These patterns, while historically accurate, remind us that machine learning models can perpetuate societal biases present in training data.

7 Conclusion: Lessons from the Abyss

The Titanic disaster, analyzed through the lens of machine learning, reveals profound truths about human nature, social structures, and the algorithms we build to understand them. Our logistic regression model achieved more than statistical accuracy—it captured the essence of human behavior under extreme duress.

7.1 What the Data Revealed About Human Nature

The patterns uncovered by our analysis paint a complex picture of humanity:

- **Nobility in Crisis**: The implementation of "women and children first" was not perfect, but it was real. Men died at nearly four times the rate of women, reflecting a societal commitment to protecting the vulnerable.
- **The Cost of Inequality**: First-class passengers had a 63% survival rate compared to 24% for third-class passengers. This stark disparity reminds us that privilege can be a matter of life and death.
- **Family as Anchor**: Optimal family sizes (2-4 members) showed higher survival rates, suggesting that small family groups provided mutual support without becoming unwieldy during evacuation.
- **Individual Heroism**: Behind every data point lies a human story—Captain Smith's final duty, the band's last song, Ida Straus's loyalty to her husband.

7.2 Modern Applications of Survival Analysis

The techniques developed for Titanic analysis have profound modern applications:

7.2.1 Emergency Response Planning

- **Evacuation modeling**: Understanding how demographics affect evacuation efficiency
- **Resource allocation**: Predicting which groups require additional assistance
- **Communication strategies**: Tailoring emergency messages to different populations

7.2.2 Healthcare and Risk Assessment

- **Patient triage**: Identifying high-risk patients in emergency departments
- **Treatment allocation**: Optimizing limited medical resources
- **Public health policy**: Understanding vulnerability patterns in populations

7.2.3 Social Justice and Equity

- **Bias detection**: Identifying unfair treatment patterns in data
- **Policy evaluation**: Measuring the effectiveness of equality initiatives
- **Resource distribution**: Ensuring fair access to opportunities and services

7.3 Ethical Implications for Machine Learning

The Titanic analysis serves as a powerful case study in algorithmic ethics:

> *"Algorithms are opinions embedded in code."*
> — Cathy O'Neil, Weapons of Math Destruction

- **Historical bias**: Our model learned the inequalities of 1912, reminding us that historical data can perpetuate past injustices
- **Representation matters**: The voices missing from our dataset—third-class passengers with incomplete records—remind us of the importance of inclusive data collection
- **Context is crucial**: Understanding the human stories behind the data prevents us from treating people as mere statistics

7.4 Final Reflections

More than a century after that cold April night, the Titanic continues to teach us. Through machine learning, we gain not just predictive power, but insight into the human condition. The 1,514 souls lost in the North Atlantic were not mere data points—they were individuals with hopes, fears, and stories that deserve to be remembered.

As we advance in artificial intelligence and data science, let us carry forward the lessons of the Titanic: that behind every dataset lies human experience, that algorithms can reflect both our nobility and our failings, and that the true measure of our models is not just their accuracy, but their humanity.

In memory of those who sailed into the night,
and in honor of those who chose courage over comfort.

End of Report
"The ship sank, but the stories survive."