Analyzing Titanic Passenger Data to Predict Survival Outcomes: Methods, Findings and Recommendations

04/01/2025
Ippala Deepthireddy
di61232n@pace.edu
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Business recommendations and technical next steps

Executive summary

Business Problem:

- A cruise line company wants to **improve emergency preparedness** in case of a disaster. They are looking to **understand which types of passengers** are most likely to survive based on past patterns, so they can **design better safety plans and training** for future trips.

Solution:

- This project analyzes data from the Titanic to find out which passenger **characteristics affected survival**. These insights can help cruise companies **plan better for emergencies** and make decisions that could **save more lives** during future disasters.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/25/2025	Completed
Methods, Findings, and Recommendations	04/01/2025	Completed
Final Presentation	04/22/2025	Completed

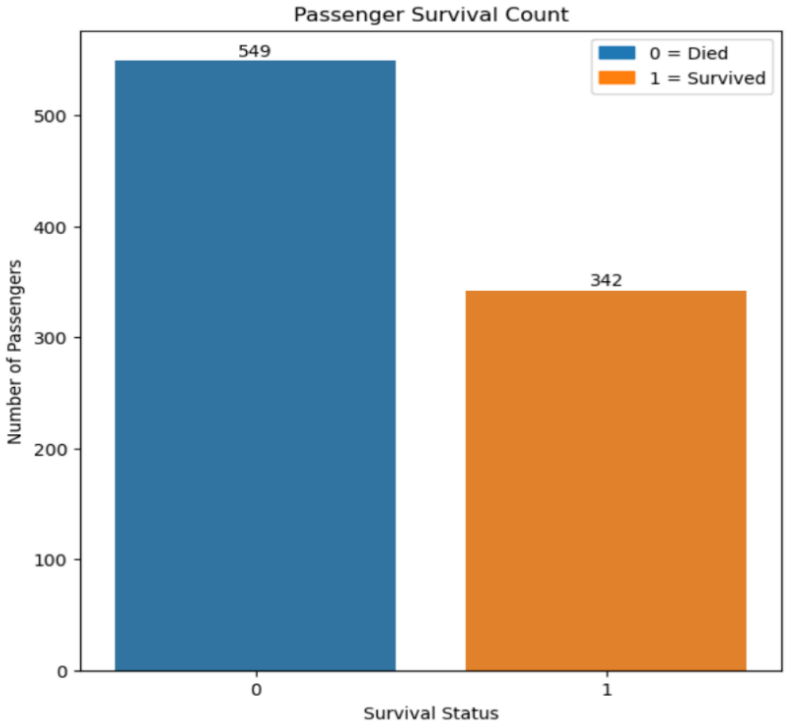
Data

Data

- **Dataset**: Titanic Dataset
- **Source**: Kaggle
- **Sample size**: 891 rows
    - Each row represents one passenger on the Titanic, including their personal and travel details.
- **Time period**: 1912-04-10 to 1912-04-15(Titanic Voyage)
- **Included data:** Passenger characteristics that may influence survival: gender, age, class, fare, family size, and embarkation point.
- **Excluded data**: Data with significant missing values or limited relevance were excluded such as cabin, ticket number, and name.
- **Assumptions:**
    - The dataset is assumed to include enough passenger information to fairly represent what happened to everyone on the Titanic.

Exploratory Data Analysis(EDA)

Passenger survival count



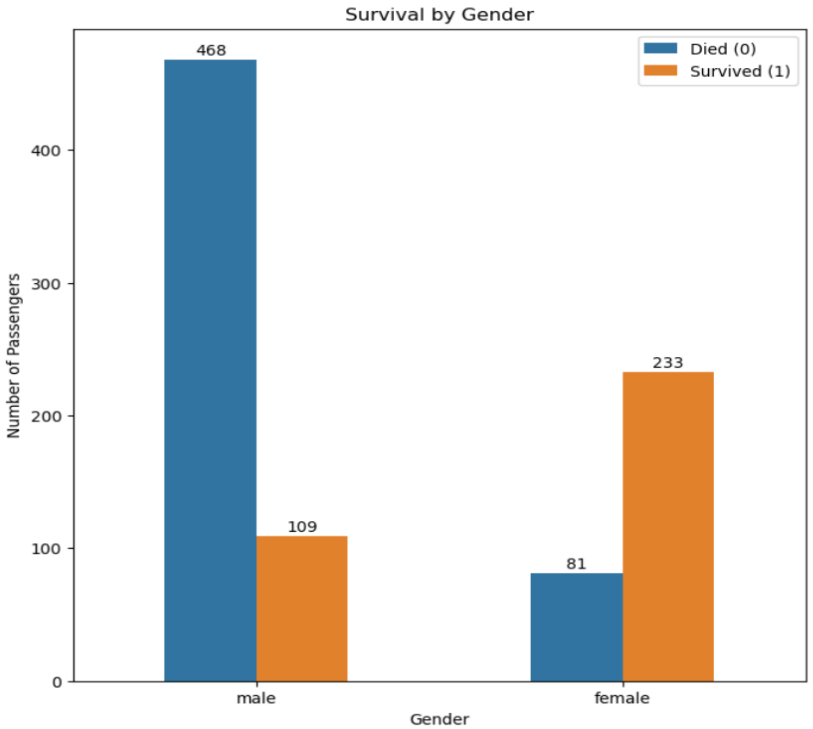
Key Takeaways:

- Out of 891 passengers on the Titanic, 549 passengers (62%) did not survive, and 342 passengers (38%) survived.
- This shows the importance of identifying which factors gave passengers a better chance. Cruise companies must learn from past patterns to avoid a similar large-scale loss of life during emergencies.

Data Notes:

- Survival Status:
  - 0 = Did Not Survive
  - 1 = Survived

Survival rate by gender



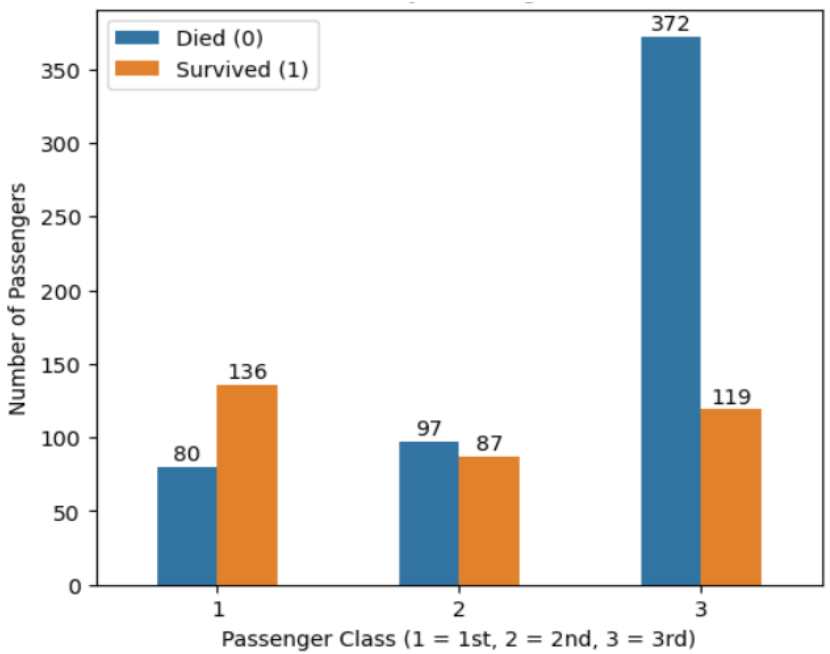
Key Takeaways:

- About 74% of women survived, while only 19% of men survived.
- Women were far more likely to survive than men. This suggests that gender played a major role during evacuation, likely due to the "women and children first" policy followed during the disaster.
- Cruise companies should ensure that **evacuation plans are fair and inclusive**, providing **equal access to safety** for all passengers regardless of gender. Emergency procedures must focus on protecting everyone, not just women and children.

Data Notes:

- Gender groups: Male and Female
- "Survived" indicates passengers who made it to safety and were rescued.

Survival by passenger class



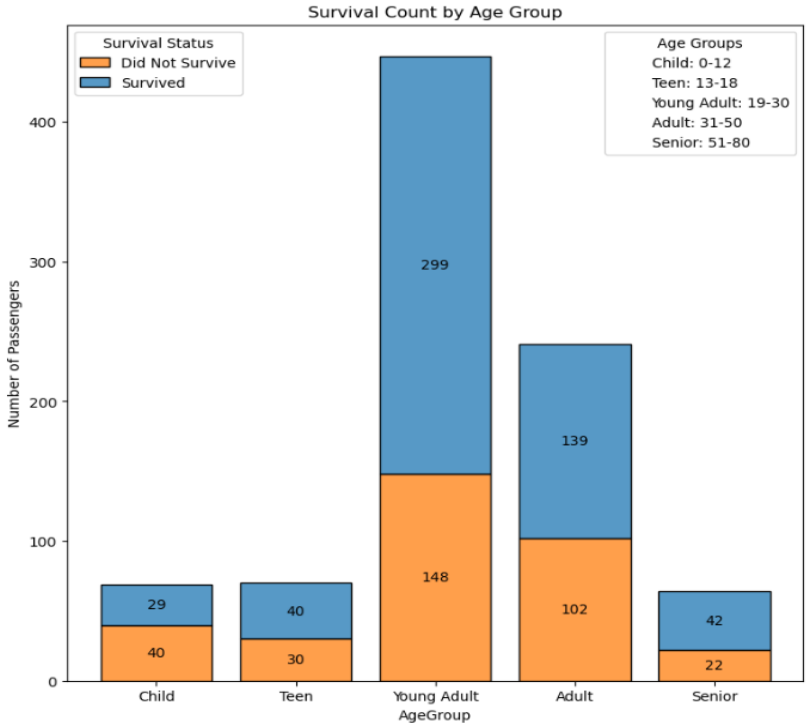
Key Takeaways:

- First-class passengers had the highest survival rate, with about 62% surviving. Only 25% of third-class passengers survived.
- This shows that wealth and social status made a big difference in survival chances, possibly because first-class passengers were closer to lifeboats and got priority in evacuation.
- Access to safety resources should not depend on travel class. Cruise lines need to design emergency plans that **ensure equal access to lifeboats and exits for all passengers**, no matter where their cabins are or how much they paid.

Data Notes:

- Passenger Class:
  - 1st Class = Wealthier passengers
  - 3rd Class = Lower-income passengers

Survival by age group



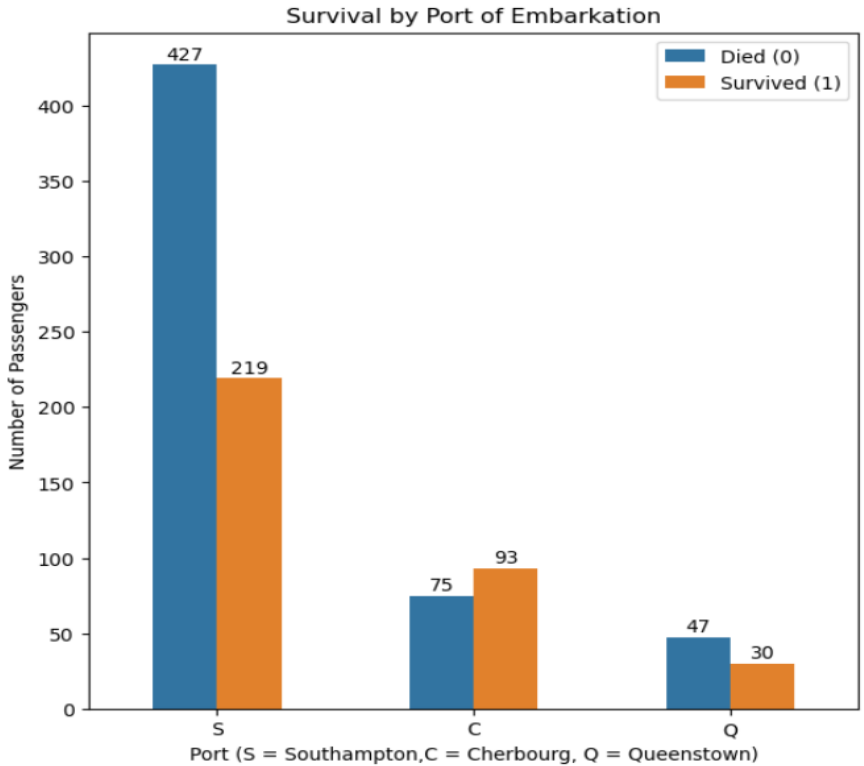
Key Takeaways:

- Young children and young adults had better survival chances. Where older adults were less likely to survive.
- This suggests that younger people were prioritized or had an easier time getting to safety during the evacuation.
- Evacuation strategies should consider age-related needs. Cruise lines must plan for **age-inclusive safety procedures**, ensuring elderly passengers receive the same attention and assistance as younger ones.

Data Notes:

- Age groups created for easier comparison

Survival by port of embarkation



Key Takeaways:

- Passengers who boarded at Cherbourg (France) had a higher survival rate (55%) compared to those who boarded at Southampton and Queenstown.
- This is likely because many first-class passengers boarded at Cherbourg, giving them better access to lifeboats.
- Cruise companies should recognize that **boarding and cabin location may impact access to safety equipment**. Safety drills and resource placement should be designed so that **all passengers can access help quickly**, regardless of where they board.

Data Notes:

- Embarkation Ports:
  - Southampton (S), Cherbourg (C), Queenstown (Q)

Modeling Methods

Modeling methods

Outcome variable - The goal of this model is to predict whether a **Titanic passenger survived or not**. This is the core question of the project. Understanding who was more likely to survive helps cruise companies learn from past patterns and make better decisions about **who may need more support during emergencies**.

Features used and rationale: The model uses the following passenger details to make its predictions:

- **Gender -** It is used to understand if men and women may have been treated differently during evacuation. This can help cruise teams review how emergency support is given across all passengers, regardless of gender.
- **Age -** Younger and older passengers may have different mobility needs. Age feature helps to explore whether certain age groups were more likely to be helped or needed more support in an emergency.
- **Passenger Class –** It determines cabin location, which may affect how quickly someone can reach safety. It helps to see if access to safety areas or lifeboats varied depending on the ticket class.
- **Fare Paid -** Fare may be linked to cabin location or comfort level. This feature helps us check whether paying more offered better access to safety or faster assistance during evacuation.
- **Family Onboard -** Passengers traveling with family may behave differently during an emergency, such as staying behind to help others. So, this helps to understand whether group travel impacts survival outcomes.
- **Embarkation Port -** Where someone boarded might reflect their route, social status, or cabin assignment. This feature helps to explore whether boarding location had any link to survival chances.

These features reflect personal, social, and travel-related factors that likely influenced survival during the Titanic disaster. They help the model find meaningful patterns that can guide safety planning today.

Model Type and Rationale (Non-Technical Version)

Model Used: A logistic regression model was trained to predict whether a titanic passenger survived or not.

How the Model Works: The model looks at each passenger's details such as their age, gender, and travel class and predicts if they were likely to survive.

- If the model sees that most women in first class survived, it learns that these traits are linked to survival.
- If it sees that most men in third class did not survive, it learns that these passengers were at higher risk.

So, when the model sees a new passenger who is a man in third class, it may predict that he was less likely to survive. The model uses patterns like these to make a simple decision: yes or no — did this person likely survive?

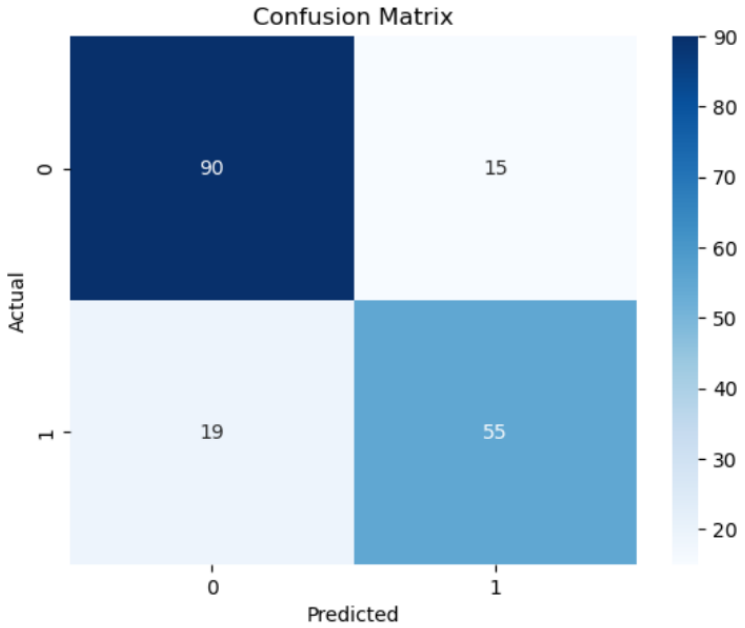
Why This Model Was Chosen:

- It gives **clear and simple predictions**, which makes the results easy to explain.
- It can show **which passenger details were most important**, helping us understand why certain people were more likely to survive.
- It works well for **yes-or-no problems** like this one.
- It helps cruise companies **identify risk factors** and make **better safety decisions** for future emergency planning.

Note - A more technical explanation is included in the appendix.

Findings

The model accurately predicted who survived & who didn’t



Key Findings:

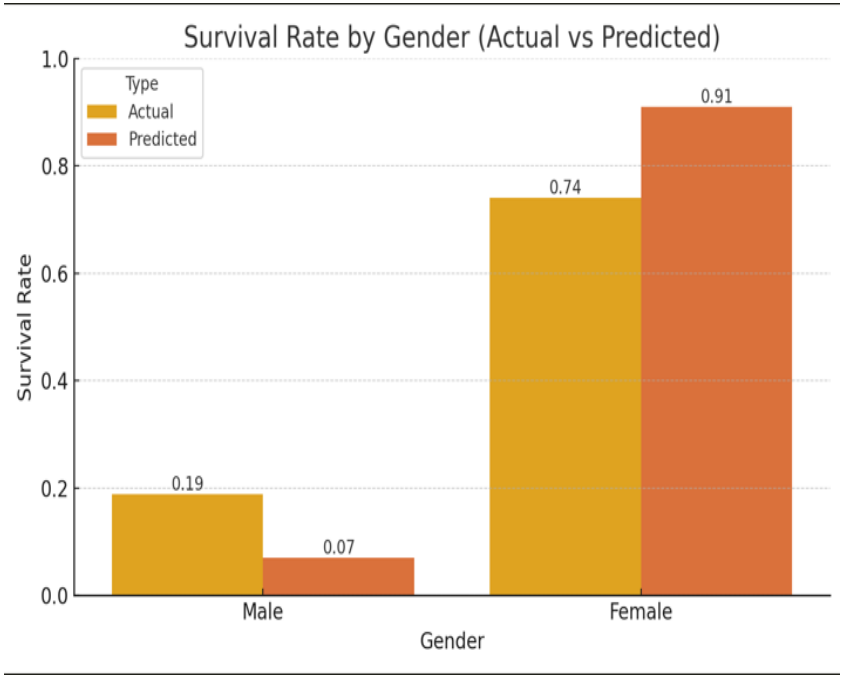
- The model was **right 81% of the time** when predicting if someone survived or not.
- Out of **105 passengers who died**, the model **correctly predicted 90** of them.
- Out of **74 passengers who survived**, the model **correctly predicted 55** of them.
- This means the model is better at predicting who **did not survive** but still does a good job overall.

So What?

Cruise companies can use this information to:

- **Identify passengers** who may be at higher risk in emergencies.
- **Design safety drills and training** that support those groups.
- **Make sure evacuation plans help everyone equally**, not just certain groups.
- **Improve placement of life jackets, signage, and crew assignments** based on who may need help the most.

Women were more likely to survive



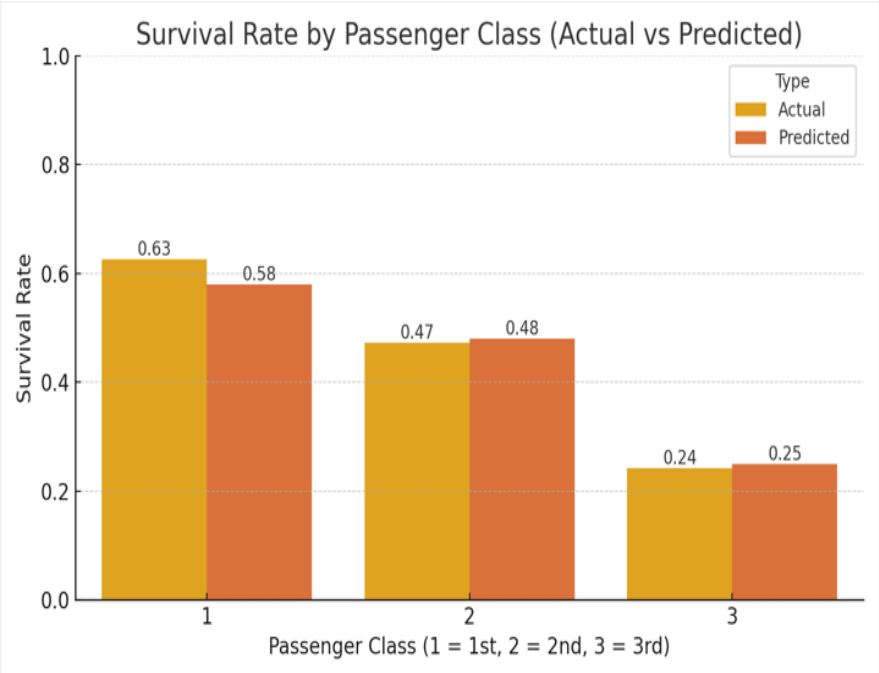
Findings:

- This chart shows the actual vs predicted survival rates for men and women.
- The model correctly picked up that women were far more likely to survive than men. It predicted a survival rate of **91% for women** and only **7% for men**, which aligns with the actual data.
- This shows evacuation policies favored women, which cruise lines must now address with equal-access planning.

So what?

- Make sure **emergency drills and safety messaging reach everyone**, not just groups historically prioritized.
- Provide **equal access to life-saving equipment** for all passengers, regardless of gender.
- Ensure staff are trained to assist **all passenger types equally** in real emergencies.

First-class passengers had the best survival odds



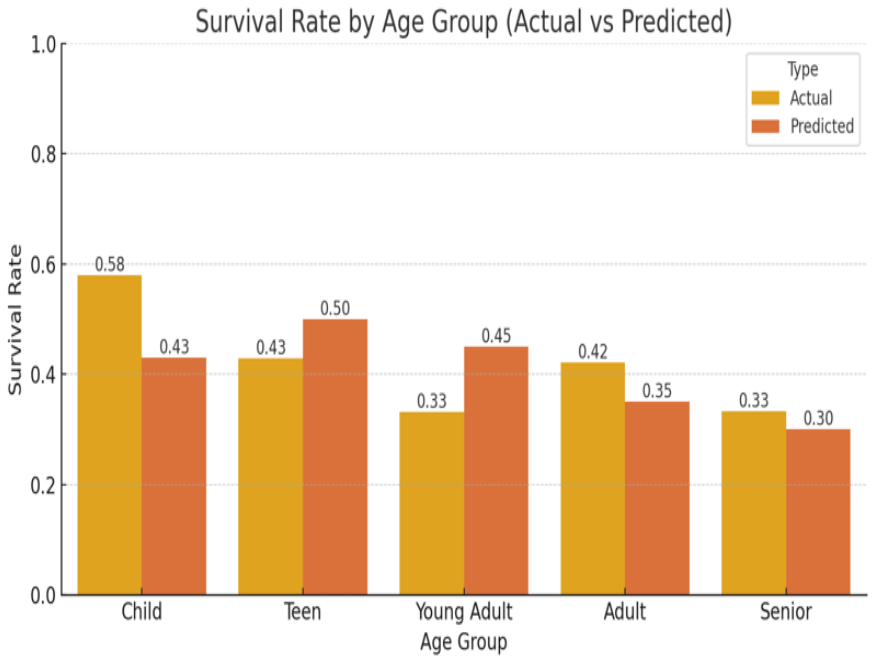
Findings:

- The model found that **1st class passengers** had the highest survival rates (63% actual), while **3rd class passengers** had the lowest (24%).
- The model closely matched these patterns in its predictions.
- Passengers in lower classes had less access to safety equipment or were physically farther from lifeboats, reducing their chances of survival.

So what?

- Every passenger cabin, regardless of fare, should offer equal access to exits and life-saving tools reachable by all passenger types.

Children had higher survival rates than seniors



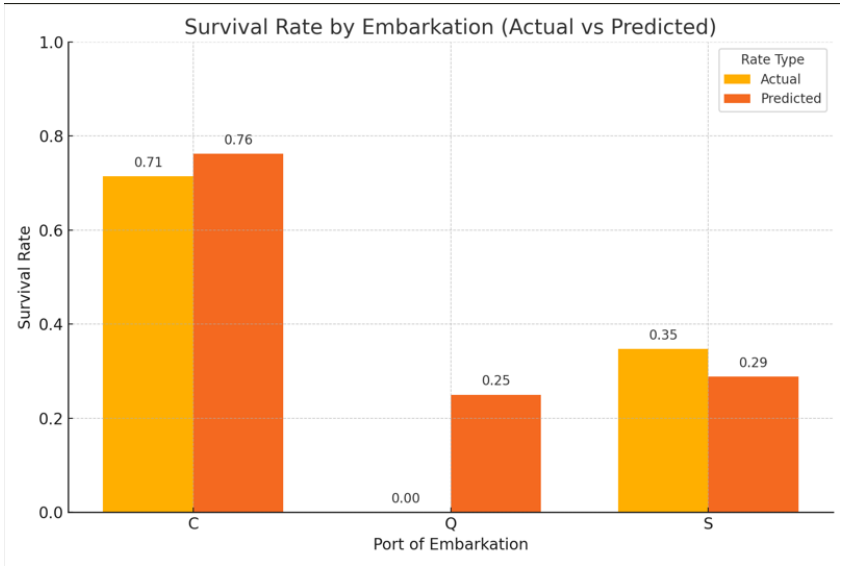
Findings:

- Children had the highest actual survival rate at **58%**, while seniors had the lowest at **33%**. The model followed this trend well, predicting higher survival for children and lower for older passengers.
- In emergencies, younger passengers may be given priority, or may be more easily assisted, while seniors could face mobility challenges.

So what?

- Tailor safety drills to include extra support for seniors and those with mobility concerns.
- Provide visual aids and crew guidance targeted to the needs of different age groups.

Survival rates varied slightly based on boarding port



Findings:

- Passengers who boarded at **Cherbourg (C)** had higher survival rates (71%) than those from Southampton (S) or Queenstown (Q). The model successfully learned these differences.
- Embarkation point might reflect differences in passenger demographics or cabin assignments.

So What?

- Use this insight to understand group-based risks and offer pre-boarding safety guidance tailored to regional passenger needs.
- Reexamine how passenger onboarding locations influence crowd distribution and safety response times.
- Cruise companies can also collect insights on passenger groups by region to ensure safety planning meets cultural and communication needs.

Business Recommendations & Technical Next Steps

Business Recommendations

Model Finding #1:

- **Gender Had the Strongest Impact -** Women were significantly more likely to survive than men, even after accounting for other factors like age and ticket class.
- **Why it matters -** This suggests that evacuation protocols historically prioritized women, which left some passengers (especially men) at greater risk.
- **Actionable Recommendation -** Cruise companies should ensure that modern evacuation plans prioritize **equal treatment for all passengers**, regardless of gender. Safety drills should train staff to **support all groups fairly** and avoid assumptions that certain people will be helped first.

Model Finding #2:

- **Passenger Class Strongly Affected Survival -** Passengers in lower classes had lower survival rates, likely due to **reduced access to lifeboats or exits**.
- **Why it matters -** This highlights a potential gap in access to emergency help based on a passenger's physical location on the ship.
- **Actionable Recommendation -** Cruise companies should conduct a **safety access audit** of all decks and ensure that **lifesaving resources are equally distributed**. All cabins—regardless of price—should have clear, fast access to exits and safety equipment.

Technical Next Steps

1. Explore More Complex Models - While logistic regression gave useful insights, more advanced models like **Random Forest or Gradient Boosting** could uncover additional patterns and improve prediction accuracy.

2. Collect More Detailed Data - This project only used basic passenger details. For future improvements, by collecting more data related to

- Cabin location on the ship
- Mobility or health information
- Group travel status (solo, family, tour group)

Could help better understand who faces challenges during evacuation and make survival predictions more precise.

3. Recommendation for Deployment - This version of the model is not ready for real-world use with limited training data and simple features. But it provides valuable direction for improving cruise safety planning.

Appendix

Project Materials

- Git Repo: <link>

Model Type and Rationale (Technical Overview)

Model Type - Logistic regression was used as a binary classification model to predict the probability that a titanic passenger survived. This model is commonly used when the outcome has only two possible values (e.g., survived or not survived).

Why Logistic Regression Was Chosen:

- It provides **probabilities** and **clear coefficients** that show how each feature affects survival.
- It's easy to interpret, which makes it ideal for communicating results and it performs well with **structured, tabular data** like this Titanic dataset.

How the Model Was Applied:

- The dataset was split into **training (80%)** and **validation (20%)** sets.
- Passenger features were cleaned and preprocessed:
  - Missing values in **Age** and **Fare** were filled with the median.
  - **Sex** was encoded as 0 (male) and 1 (female).
  - **Embarked** was one-hot encoded, and unused fields were dropped.
- The model was trained using **Scikit-learn's Logistic Regression** with default settings.

Performance of the Model:

- Validation Accuracy: 81%
- The model was able to correctly identify most passengers who survived or did not survive.
- A **confusion matrix** and **classification report** were used to evaluate performance.
- The model performed better at predicting non-survivors but still had reasonable recall for survivors.
- **Feature importance (based on model coefficients)** showed that **Gender, Class, and Age** were the most influential factors.

Note: Click here to go back to the main slide.