Theran Meadows
Project 2: Titanic
Milestone 3
July 21, 2024

Business Problem

The story of the Titanic is tragically famous and still studied today even 100 plus years later. The last survivor died at 97 years old back in 2009. She was just a 3 month old baby when her family boarded the ship along with 2239 other passengers. 706 people survived once the ship finally sank.

My project will be to take on the famous Titanic data set and see who of the 2240 passengers would have lived. Or at least to make a predictive model to see who of the test data would have lived and what my accuracy was.

My question I will be attempting to answer is:

What sorts of people were more likely to survive?

Background/History

Currently the show has had a consistent number of contestants per season at 20. The show will then filter through one or two challenges per episode usually eliminating at least one person but sometimes two or more can also be eliminated. There may be partner challenges or group team challenges as well. The single or team winner of a challenger will then be given immunity from being eliminated and will therefore usually sit the following challenge out. The final 3 are then considered the finalists and complete through 3 final rounds cooking. The first round is the appetizer, the second is the entree and the third is the dessert. After the appetizer or the entree is usually when one of the finalists is eliminated. The winner is then chosen after both have completed the challenge

Data Explanation (Data Prep/Data Dictionary/etc)

Kaggle.com is providing the data set I will be using. It provides two data sets: one to be used for training and one to be used for testing. The columns and definitions of values used are:

- Survival - whether or not the passenger survived or not
- Pclass - the class of the passenger
- Age - how old the passenger was
- Sibsp - whether or not the passenger was a brother/sister (including steps-) and/or a husband or wife.
- Parch - whether or not the passenger was a mother/father or a daughter, son(including steps-)
- Ticket - Ticket number
- Fare - fare amount
- Cabin - cabin number
- Embarked - what port the passenger embarked from

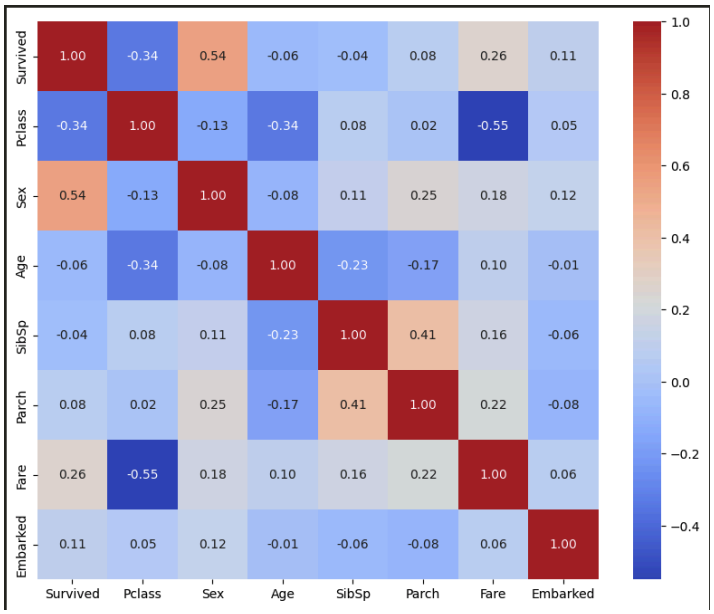
Methods

The target value for this data set will be whether or not the passenger survived, which is a simple ‘yes’ as 1 or ‘No’ as 0. The rest of the columns will be used to see what values are good predictors of the survival column.

There will also be exploratory data analysis to identify any patterns or possible correlations. One thing I did want to investigate is the ratio of men to women and children. From the 1997 film, women and children were the first to board the rescue

boats because there wasn't enough. A lot of the men were left behind. I want to investigate if more women and children did in fact survive.

Analysis



Starting with a correlation matrix, it was clear to see that gender was the most impactful attribute on survivability. What about class, age or family though? From the matrix, it did really seem to affect it all that much. We can build models to help determine if that is true. For this analysis, I will be using two models: a logistic regression and a random forest model.

I used a logistic regression model because of its simplicity and readability with making relationships linear. I used the random forest model to handle the more complex relationships between some of the variables. After using a logistics regression model and a random forest model, the accuracy of both models was within 2% of each other.

Both models were able to determine pretty clearly that Gender, Age, Socio-economic class played major parts in the odds of someone surviving the sinking

of the Titanic. Gender had the biggest of the three with more than 3x as many women surviving than men did. If you were a rich female, you had the best chance of survival as you probably had priority getting on one of the 20 lifeboats. Unfortunately, for the men, even high class ones, did not have a good survival rate.

Conclusion

The sinking of the Titanic is a tragic disaster in history. With more than 2000 passengers on board and only 20 lifeboats with a 60 person capacity, any disaster that struck was already dooming nearly 50% of the passengers. In the state of emergency, it was determined that women and children had priority to board lifeboats with some men able to board if they were the sole parent or guardian of children that had boarded. Sadly, if they had a wife that had boarded with their kids, they were left behind most likely to go down with the ship or succumb to the freezing waters.

Assumptions/ limitations

The data set is closest in number to other comparable data sets outside of Kaggle. The count of passengers is within 10 passengers of each other.

Although passenger data is given, the time of the impact to the iceberg and where passengers were located at the time might also play a part in who survived. Certainly those that were first onto the rescue boats had the most time to get away from the ship while it was still floating horizontally before it took on a lot of water.

Challenges

Other than no being a subject matter on this topic, some challenges I anticipate are validation problems. I am unfamiliar with what models I want to use so I will continue research on what the best fit model would be before validating.

Future Uses/Additional Applications

This model would not be wise to implement for future uses other than for curiosity. The event in the history of cruise ships gave a massive overhaul on rules, regulations, and protocols for what happened if this happened again.

Since the year 2000, there have been around 200 people lost at sea due to the cruise ship sinking. Around 15 cruise ships have sunk since 2000 as well. Safe to say though that more than enough lifeboats are provided for on each ship.

Recommendations

There are no real recommendations that I can give now that haven't already been implemented. This event in history has changed how ships are prepared when such disaster strikes. Modern cruise liners today now carry both lifeboats and life rafts for emergencies. The lifeboats on the Titanic could only hold around 60 people while modern lifeboats can hold 300-400 people. With inflatable life rafts being implemented, there is more than enough room for all passengers to make it to safety if time allows.

Ethical Assessment

No ethical considerations to consider since the data I am using is open to the public. No survivors of the Titanic are living today but respect is due for those that were lost.

References

Titanic - machine learning from disaster. Kaggle. (n.d.).

https://www.kaggle.com/c/titanic/data

Wikimedia Foundation. (2024b, July 2). *Titanic*. Wikipedia.

https://en.wikipedia.org/wiki/Titanic

Encyclopædia Britannica, inc. (2024, July 13). *Titanic*. Encyclopædia Britannica.

https://www.britannica.com/topic/Titanic

10 Questions

1. **What were the most significant factors determining survival on the Titanic?**

Factors such as passenger class (Pclass), sex, and age played significant roles. Historical records and analysis often show that women and children had higher survival rates, and passengers in higher classes were more likely to survive due to better access to lifeboats.

2. **How did you handle missing data in your analysis, especially for variables like Age and Cabin?**

Missing data were addressed through techniques like imputation. For age, common methods include using the median or mean age, whereas for cabin, due to the high percentage of missing values, it was dropped though having it derived into a simpler form such as 'Cabin Known' vs 'Cabin Unknown' could have been possible.

3. **Can you explain how you chose the model for your analysis? Why did you prefer it over others?**

The choice of model could be based on its suitability for binary classification problems and the dataset size. Logistic regression is a common choice due to its interpretability and efficiency. More complex models like random forests or support vector machines might be chosen for their higher accuracy in more complex scenarios.

4. **Were there other performance metrics of your predictive model?**

Performance metrics likely include accuracy, precision, recall, and the F1-score. These metrics help evaluate the model's ability to correctly predict survival, balance between sensitivity and specificity, and the harmonic mean of precision and recall, respectively.

5. **How did the socio-economic status (represented by the passenger class) influence the survival rates?**

Socio-economic status, indicated by the passenger class, significantly influenced survival chances. Higher-class passengers often had better access to lifeboats and emergency resources, leading to higher survival rates.

6. **Did the embarkation point affect the survival chances, according to your model?**

The embarkation point might show some influence due to socio-economic factors related to the demographics of passengers boarding at different ports. However, its impact is generally less significant compared to other variables like sex and class.

7. **How did you ensure that your model was not overfitting the training data?**

Techniques such as cross-validation, where the data is split into multiple subsets to validate the model against different parts of the dataset, help prevent overfitting. Regularization methods in model training also reduce the risk of overfitting by penalizing overly complex models.

8. **Could you expand on how feature engineering influenced your model's performance?**

Feature engineering, such as creating new features like family size from SibSp and Parch, or extracting titles from names, could help the model by introducing new relevant information or simplifying existing information, thus potentially increasing predictive accuracy.

9. **Are there ethical considerations in using demographic data (like sex and age) for predictive modeling in this context?**

Ethical considerations arise when using demographic data, as it involves sensitive information. In historical analysis like the Titanic, it's primarily for understanding disparities in survival rather than decision-making. However, in modern applications, such usage requires careful ethical review to avoid reinforcing biases.

It should also be noted that this disaster happened over 100 years ago. The last survivor died in 2009 at the age of 97. She was 3 months old when she boarded with her family.

10. **What further data would you like to have to improve the predictive accuracy of your model?**

Additional data like the physical health of passengers, exact location on the ship at the time of the disaster, or detailed crew actions during the evacuation could provide deeper insights and improve model accuracy by highlighting other survival factors.