Exploration of Titanic Data

Keaton Whitehead, Angus MacDonald
University of Colorado Boulder

ABSTRACT

The project's goal aimed to convert a raw dataset that held information on Titanic passengers into an application for the user to explore on their own. The dataset holds personal information on each passenger along with other factors that may correlate with whether they survived. The primary goal was to hand over control to an arbitrary viewer. The application is a graph with multiple y-axes where the user decided which factors display on the graph. Survival is depicted by color. As a result, trends are revealed throughout the infographic for the user to explore and conduct further research on their own if they wish.

Keywords: Titanic, survival factors, correlation.

1 INTRODUCTION

The motivating problem behind the project was the limiting structure of the titanic dataset. Although a significant amount of research has already been uncovered and published freely across the internet and other media, we found that it was not easy for the user to search for correlations between attributes of the passengers aboard the titanic. With the dataset we used, the available attributes include ticket class, age, gender, number of siblings or spouses aboard, number of parents or children aboard, and the passenger's ticket fare. Additionally, each passenger's data is color-coded to show whether they survived, so that the user can see for themselves which attributes tended to influence survival with their own eyes, along to what degree those attributes held.

The unique aspect of this project came from the aim to hand the user the power to search through the data, rather than prepare static visualizations that the user is stuck with. If the user finds something that interests them, they may explore it through our application if fitting with the given attributes in the dataset.

2 GENERAL RESEARCH

On the 15th of April 1912, the Titanic sank early in the morning after hitting an iceberg. The ship was sailing from England to New York City. There were over 1,300 passengers on board, ranging from wealthy businessmen and families to impoverished emigrants in search of a better life, and plenty of middle-class individuals in the middle of the spectrum. The set of passengers was therefore very diverse and has led to many researchers investigating factors that influenced whether a passenger survived the sinking of the ship. Commonly referenced attributes include wealth and gender. With our dataset on the Titanic passengers, we were able to open trend searching to other variables outside of class and gender, such as ticket fare, number of family members on board with the passenger, and age.

*kewh9902@colorado.edu, anma2271@colorado.edu

To be clear, other work has been done regarding this same exact dataset. For example, Prateek Chanda has build a python script that displays a graph on the dataset as well [2]. The idea with Prateek's work was to incorporate machine learning into exploring the dataset with the intention of predicting which passengers would survive the tragedy (note that machine learning is beyond our project's scope). While interesting, this approach still doesn't necessarily hand the user control in researching the factors themselves with their own eyes so that they may draw conclusions on their own, depending on what interests them.

The two most seemingly significant attributes to determine survival probability were class and gender, as visualized by source three in our citations [3]. Over 350 deaths were from class III tickets alone, while less than 100 deaths were from in either class I or class II. Over 400 of those who perished were male, while less than 100 were female [3]. The statistical analysis continues, exploring several factors but only with static visualizations like pie charts and bar graphs that the viewer can only scroll along with, without the ability to change which factors to examine at once at their convenience.

Our fourth source directly dives into the question of "which people were likely to survive the sinking of the Titanic," [4]. It uses statistical analysis methods and algorithms including random forests, gradient boosting machine, and support vector machine (again, beyond the scope of this class and our project). However, the paper concludes that there is in fact deciding factors that determine predictable likelihoods of survival, and that those factors can be calculated using the same dataset we utilized for our application [4].

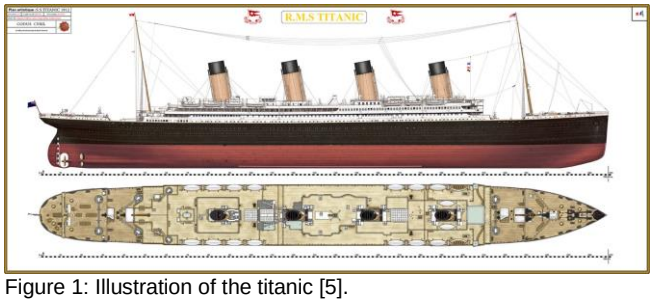

Figure 1: Illustration of the titanic [5].

Our fifth source went even deeper and listed example algorithms that can be used to explore the Titanic dataset to find trends and come to conclusions about survival probability in relation to different factors. Some of those methods include logistic regression, k-nearest-neighbor, and building decision trees [5].

Our sixth source found a particularly interesting fact exploring the idea of prioritizing the safety of women and children and getting them into lifeboats before all others. Although a few outliers exist, probably due to the chaotic nature of the sinking, the rule tended to be effectively applied for classes I and II, but not for class III [6]. It may be disputed on if this was a deliberate occurrence, i.e. the third class was purposefully discriminated against, but there is also the fact that the lower classes tended to be in the lower parts of the ship and had much more distance and hallways to maneuver to get to the upper decks. In general, they had less time and more to

overcome to have a shot for safety simply due to their relative location on the ship.

Source seven specifically explores the use of logistic regression in exploring the information within the Titanic dataset. It discusses the accuracy of the method in relation to this data and concludes an accuracy rate of 95% [7], which seems to be high especially for our general research purposes. This provides an idea of what kind of accuracy will be provided, speaking to the reliability of statistically analyzing the Titanic dataset (with a very general outlook of course).

The final source goes far less into the technical side of the data and focuses much more on social aspects to the results of survivorship on the Titanic. For example, one reason for the overall low rate of survival stemmed from difficulty in loading lifeboats, even when the people to load were present and ready to go. In some reports, people held on to the belief that the Titanic was unsinkable regardless of what they were told by crew, which probably delivered mixed signals to the passengers amidst the short time-frame and lack of available information. When faced with the decision to stay on board the "unsinkable" Titanic or to board a small lifeboat hovering 60+ feet over the ocean's depths, many decided it was better to stay on board. Also, some passengers heard that there was another ship nearby that could pick them up instead, so getting on board a life raft could end up being an unnecessary risk in their eyes. With an already severe lack of lifeboats on the Titanic, this attitude only solidified the number of casualties that would result in the sinking of the Titanic [8].

The paper goes on to discuss factors regarding survival statistics regarding ticketing class and gender. With ticketing class, some arguments have been laid down stating that third class passengers suffered the brunt of the casualties due to their physical location on the ship on the lower decks, which mandated an excessive amount of time to maneuver and find their way to the top decks of the ship. Additionally, since many of the third class passengers were emigrants who wanted to establish a new life in the U.S. or Canada, there were also language barriers and reluctance on behalf of the passengers to leave the few belongings they had behind as it may have felt as though they were abandoning everything they had if they were to leave the ship. This further contributed to hesitation, loss of time, increased confusion and a result of more deaths among third class passengers [8].

3 Project Description

Our project revolved around the development of an application which took in the Titanic dataset with each passenger and their attributes in an Excel .csv file. The application then displays each row as a line, meaning each passenger is represented by a single line in the graph, as displayed below.

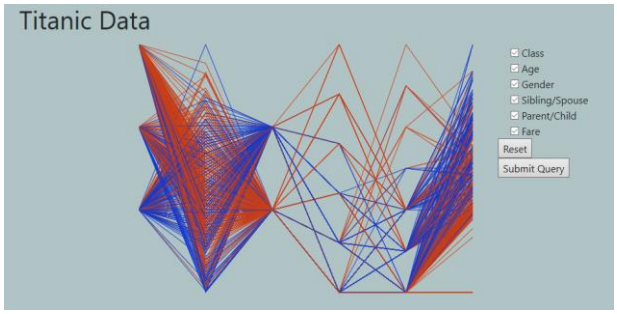

Figure 2: Our application with all attributes displayed.

With the above picture, we have yet to successfully implement labels for each y-axis, along with tick marks to show the user the values being represented at each point of the lines. Also, a legend is not yet present. This is being mentioned to make it clear that we did not overlook these elements, but instead came across technical challenges. Despite these shortcomings, the overall application demonstrates the role of providing the user the ability to search the dataset themselves.

In the above picture, red lines represent survivors, while blue lines show casualties. The user can conveniently choose exactly which axes are displayed, but not the order in which they are displayed (again, due to technical limitations). With the current design, the attributes are displayed in order left-to-right according to the checklist from top-to-bottom, depending upon which of the attributes are checked. The "Reset" button clears out any existing checkmarks on behalf of the user.

4 Conclusion

The findings discovered from use of our application verified the statistical conclusions that each of our references came to, strictly in terms of general trends. The stark color changes, where color represents survivorship, between class and gender verify that indeed social factors heavily influenced what occurred on the Titanic that fateful morning. Additionally, the user can see just how diverse the passenger set was in terms of age, fare, and how many family members they had on the ship since there appears to be a relatively even distribution overall along those respective axes.

Though we failed to achieve every technical goal we aspired for in the project, the overall aim was accomplished in developing an application that a user can interact with to control exactly which factors get shown at a given time.

References

[1] "Passengers of the RMS Titanic." *Wikipedia*, WikimediaFoundation, 2Dec. 2019, en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic.

[2] Chanda, Prateek. (2018). Predicting Passenger Survival Rates on the Titanic. 10.5281/zenodo.1098228.

[3] Jatturat. "Finding Important Factors To Survive Titanic." *Kaggle*, Kaggle, 23 Feb. 2018, www.kaggle.com/jatturat/finding-important-factors-to-survive-titanic.

[4] Barakat, Mohammed. "Titanic Tragedy: Survival Prediction with Machine Learning." *Rstudio,* Rstudio, 9 Fed. 2018, http://rstudio-pubs-static.s3.amazonaws.com/358312_6fcb6eff7f214d0f8630fbfe59c37e29.html.

[5] Donges, Niklas. "Predicting the Survival of Titanic Passengers." *Medium*, Towards Data Science, 15 May 2018, towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8.

[6] Oleg Leyzerof. "Titanic: Factors to Survive." *Oleg Leyzerov*, 23 Apr. 2016, olegleyz.github.io/titanic_factors.html.

[7] Kshirsagar, Vaishnav, and Nahush Phalke. "Titanic Survival Analysis Using Logistic Regression." *Irjet*, Aug. 2019, www.irjet.net/archives/V6/i8/IRJET-V6I815.pdf.

[8] Hall, Wayne. "SOCIAL CLASS AND SURVIVAL ON THE S.S. TITANIC. *Espace*, espace.library.uq.edu.au/data/UQ_152940/HallSSM2261986.pdf?Expires=1576351729&Key-Pair-Id=APKAJKNBJ4MJBJNC6NLQ&Signature=ZjzJA4bzqCJXOD-sVixjTnwtyuOFbgqnIXa6UbQE0NIO6QRzl8-I8HirUPAzqsf7nIwAlIp40PydKyfYRoxiClYH2Y0wc3n~Tz-oLQKGpPh3AmApZOJVAfmmfJASkaYrW-qUIxZXcJkt50I7bURXj44KTH8NM4NYNaJ0qZIsDL5sIibotDFfyaIFF96kGgtij-goaGrr5oujSK6HrgA6OWMEsbXyfePcIEeFHPu6P9nGBLD9l3eIQfTCFWCKv57zg-RZwwBsLSZDXh3fMG~i1cG7ZQH4APTfoKHfkWOxbL0lnGSwOUXf9hoc5wEPrG-Y2yq4G3z3iv-2~gGPCU75EQ__.