Project Proposal - Titanic: Machine Learning from Disaster

Maeki Kashana, Miguel Melo Ochoa, Alex Hayet, and Francisco Gomez

San Diego State University

CS549 - Machine Learning

Professor Xin Zhang

April 4, 2025

Project Proposal - Titanic: Machine Learning from Disaster

Contents

Problem Statement

All details for this project will be provided by a competition titled "Titanic - Machine Learning from Disaster" introduced by kaggle.com (Cukierski, 2012).

Project Description and Statement

This project will build a predictive model that aims to find the type of people who are going to survive the sinking of the Titanic, a well renowned ship that was believed to be "unsinkable" until it hit an iceberg in 1912, resulting in the death of 1502 passengers/crew.

This project will utilize passenger information, such as name, age, gender, socio-economic class, etc., to determine what sorts of groups were more likely to survive than others during the sinking of the Titanic.

Project Motivations

What motivated this group to take on the challenge of creating a predictive model that will dictate the type of individuals that were the most likely to survive the sinking of the Titanic is that this project provided an adequate challenge from a machine learning perspective.

To give more detail, this project will provide a lot of testing data in order to make predictions on who will survive. In consequence of this data, this team will be able to create a predictive machine learning model that utilizes the data provided for this project to make predictions on who will survive. Moreover, this project will enable the team to practice the machine learning concepts previously learned in order to find a solution for this problem.

In addition, another motivation that this team had for undertaking this specific project is that this project requires the primary use of the "numpy" and "pandas" python packages in order to find a solution. As a result, this project will enable this team to become more skillful in using these packages to manipulate data and apply machine learning concepts.

Dataset Description

Of this project of the Titanic, the data that we will be using contains many features that will be used to train a predictive model in order to predict whether a passenger could survive the saddening occurrence of this Titanic tragedy. Of the many features, one of them is PassengerId, which is used to identify which passenger is which and the number of passengers in order to be tracked in the Titanic dataset. The next feature is Survived. This feature is a true or false identification; this tells the model that the passenger in the respective row has either survived or not survived in this incident. The feature Pclass is an identification of the respective passenger's class, either as Upper, Middle, or Lower class status in the Titanic. The features Name, Sex, and Age are the names, sex, and age of the passenger of the respective row which can be important to identify when the model is predicting if they could have survived. The feature Sibsp identifies if the passenger was part of a family relation, such as being a sibling or spouse. Some examples of being siblings or spouses are being a brother, sister, stepbrother, stepsister, husband, or wife. This would be an important factor because during the Titanic incident, this would have meant more family to worry about while trying to survive. The Parch feature defines the family relation in terms of being a parent and child. Examples of this family relation are if the passenger was a father or mother and they had either a daughter, son, stepdaughter, or stepson. An important note for the model is that some children have aboarded the Titanic with their Nanny, but that is not a father or mother, which would cause Parch to be 0 then. Lastly, the features ticket, fare, cabin, and embarked are the tickets, fares, cabin number, and where they embarked from for each passenger which they had owned during their time on the Titanic.

Planned Methodology

Since this is a classification based problem and our goal is to make one of two possible predictions based on a set of given independent variables, logistic regression would be the optimal choice. In this case, the two possible outcomes are 1 (survived) and 0 (died). Of all the features that are provided, the independent variables that could be used to properly train a model could be things like gender, class, age, etc. Other features such as id, ticket, and fare are not relevant enough in helping us predict someone's survival. Another model to consider for this would be the random forest model. If we were to do that, the model could be trained by having it find patterns in gender, class, parch, and family relations. While it is possible that we may switch to a different model in the future, it is very unlikely due to how everything is already set up to create a logistic regression model.

Evaluation Metrics & Expected Outcomes

When looking at the evaluation metrics of the Titanic problem, the problem uses binary classification in order to show if a passenger survives the tragedy. In this instance, if a passenger is marked with a 1 this means they survived and a 0 would mean that they did not survive. When wanting to make a prediction on whether or not a passenger will survive there are a few metrics that we could use in order to find out the most likely outcome for the passenger. The first metric to look at would be the gender of the passengers. When looking at the data it seems that female passengers had a much larger survival rate than men passengers. Only about 20% of men survived aboard the Titanic while around 75% of women survived. This is largely due to the fact that women and children were the number one priority to get on the lifeboats first. Passenger class is another metric that affects the survival rate of passengers. First class passengers were more likely to survive in the event of sinking due to being higher up in the ship than second or third class passengers. Age is another factor that helped in the survival rate for passengers on the Titanic. As stated earlier, women and children were the highest priority passengers to get onto the lifeboats. This had a large impact on the survival rate for children aboard the Titanic and gave them a better chance at surviving. The ports in which passengers had boarded the Titanic from do show a correlation to survival. This is a socio-economic factor that does link back to the class of passengers as we can see a connection between the ports passengers boarded from to a class level on the boat. For example, Cherbourg has a lot of passengers that boarded into the first class.which means that they had a higher chance of surviving than passengers that boarded from Queenstown which had more third class passengers than Cherbourg. These are just some of the metrics that can be used to find expected outcomes in the event of the Titanic.

Work Distribution

Maeki Kashana

1. Describe the problem statement and motivation
2. Download data required for project
3. Set up coding environment for the project
4. Help with finding a solution for the problem with other team members and work on sections of the code to implement the solution.

Miguel Melo Ochoa

1. Describe the training dataset that will be used.
2. Help other team members to understand obstacles together.
3. Work on sections of the code and help train and test the predictive model.

Alex Hayet

1. Describe relevant evaluation metrics and expected outcomes
2. Assist fellow members of the team with issues that might arise
3. Work on code with the team in order to develop a proper solution to the issue

Francisco Gomez

1. Describe optimal machine learning model and techniques
2. Help team with any problems they might come across
3. Work on code to create an accurate ML model

References

Cukierski, W. (2012). Titanic - machine learning from disaster [Kaggle].