HARVARD EXTENSION SCHOOL

EXT CSCI E-106 Model Data Class Group Project Template

Author: Dinesh Bedathuru Author: Brian Calderon
Author: Jeisson Hernandez Author: Hao Fu Author: Derek Rush
Author: Jeremy Tajonera Author: Catherine Tully

06 May 2025

In this project, our aim is to classify the probability of a passenger surviving the Titanic crash of 1912. We used a variety of linear and non-linear models to deduce the most accurate model and provide long-term stability in our predictions.

Table of contents


```
Size of entire data set: 1310
```

Classify whether a passenger on board the maiden voyage of the RMS Titanic in 1912 survived given their age, sex and class. Sample-Data-Titanic-Survival.csv to be used in the Final Project

Variable	Description
pclass	**Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)**
survived	**Survival (0 = No; 1 = Yes)**
name	**Name**
sex	**Sex**
age	**Age**
sibsp	**# of siblings / spouses aboard the Titanic**
parch	**# of parents / children aboard the Titanic**
ticket	**Ticket number**
fare	**Passenger fare**
cabin	**Cabin number**
embarked	**Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)**
boat	**Lifeboat ID, if passenger survived**
body	**Body number (if passenger did not survive and body was recovered**

Variable	Description
home.dest	**The intended home destination of the passenger**

1 Instructions:

0. Join a team with your fellow students with appropriate size (Up to Nine Students total) If you have not group by the end of the week of April 11 you may present the project by yourself or I will randomly assign other stranded student to your group. I will let know the final groups in April 11.

1. Load and Review the dataset named "Titanic_Survival_Data.csv" 2. Create the train data set which contains 70% of the data and use set.seed (15). The remaining 30% will be your test data set.

3. Investigate the data and combine the level of categorical variables if needed and drop variables as needed. For example, you can drop id, Latitude, Longitude, etc.

4. Build appropriate model to predict the probability of survival.

5. Create scatter plots and a correlation matrix for the train data set. Interpret the possible relationship between the response.

6. Build the best models by using the appropriate selection method. Compare the performance of the best logistic linear models.

7. Make sure that model assumption(s) are checked for the final model. Apply remedy measures (transformation, etc.) that helps satisfy the assumptions.

8. Investigate unequal variances and multicollinearity.

9. Build an alternative to your model based on one of the following approaches as applicable to predict the probability of survival: logistic regression, classification Tree, NN, or SVM. Check the applicable model assumptions. Explore using a negative binomial regression and a Poisson regression.

10. Use the test data set to assess the model performances from above.

11. Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model.

12. Create a model development document that describes the model following this template, input the name of the authors, Harvard IDs, the name of the Group, all of your code and calculations, etc..

Due Date: May 12 2025 1159 pm hours EST Notes No typographical errors, grammar mistakes, or misspelled words, use English language All tables need to be numbered and describe their content in the body of the document All figures/graphs need to be numbered and describe their content All results must be accurate and clearly explained for a casual reviewer to fully understand their purpose and impact Submit both the RMD markdown file and PDF with the sections with appropriate explanations. A more formal.

2 Executive Summary

This section will describe the model usage, your conclusions and any regulatory and internal requirements. In a real world scneario, this section is for senior management who do not need to know the details. They need to know high level (the purpose of the model, limitations of the model and any issues).

3 I. Introduction (5 points)

This section needs to introduce the reader to the problem to be resolved, the purpose, and the scope of the statistical testing applied. What you are doing with your prediction? What is the purpose of the model? What methods were trained on the data, how large is the test sample, and how did you build the model?

The Titanic was a British-registered ship that set sail on its maiden voyage on April 10th, 1912 with 2,240 passengers and crew on board. On April 15th, 1912, the ship struck an iceberg, split in half, and sank to the bottom of the ocean (National Oceanic and Atmospheric Administration (NOAA), 2023). In this report, we are going to analyze the data in the Titanic.csv file and use it to determine the best model for predicting whether someone on board would live or die. By creating this model, we hope to understand what factors a passenger could have taken into account in order to reduce their risk of death during the trip. We cleaned the data and split into into a train/test split in order to properly train our models. We created simple linear models, multivariate linear models, logistic models (both binomial and poisson), a regression tree, and a neural network model. The train sample size was 916 data points (70.03%) and the test sample size was 392 data points (29.97%). We built the models after examining the data and determining which predictor variables we thought would be most relevant for survival rate. Once we had our variables and training data, we created the models and examined the performance of the models on both training and testing data to determine if they were robust. We also examined if the model assumptions appeared to hold for each model.(Prepineer, 2025)

```
[1] 4
```

References

National Oceanic and Atmospheric Administration (NOAA). (2023). *RMS titanic – history and significance.* https://www.noaa.gov/office-of-general-counsel/gc-international-section/rms-titanic-history-and-significance

Prepineer. (2025). *Understanding the binomial distribution.* https://prepineer.com/s/binomial-distribution/