Titanic - Machine Learning from Disaster

DSCI 478
February 26, 2025
Jacy Werner

For this Kaggle Competition (Titanic - Machine Learning from Disaster), individuals and groups set out to solve a problem with the highest accuracy possible. The problem we are attempting to solve is to design a Machine Learning Algorithm to make predictions on whether or not an individual survived the Titanic based upon passenger data such as name, age, gender, and socioeconomic class. Using this information, competitors are given a "training dataset" known as "train.csv" which contains all passenger information. Using this dataset, we set out to design a model to make analytical predictions on trends between individuals who survived or are deceased. Additionally, we receive a "testing set" named "test.csv" which we use to analyze the accuracy of our models survival predictions. For the remainder of the write up, we will explore the strategies/methods used to complete this assignment and explore the accuracy of our results.

To begin, I imported the csv files into a folder where I could analyze the datasets through python. After brief analysis of the datasets. I found that there were multiple missing data points in the classifications of "age", "cabin" and "embarked". Using this code,

```
print("\nMissing Values in Train Dataset:")
print(train_df.isnull().sum())
```

I was able to find that "Age" was missing 177 data points, "Cabin" was missing 687 data points, and "Embarked" was missing 2. Additionally, I used plot visualization of each classification to find initial trends between the categories that I could implement into a machine learning model later. The next goal of this assignment was to "clean/prepare" the data to ensure the data's quality. I first began by merging the training and test datasets into a single data frame so the data cleaning was consistent between both the training and test sets. We can do this by implementing TrainFlag, which we can separate back later into their original sets. Additionally, I filled in the missing data from the "embarked" category with the most frequent port, as well as using the median age for the "age" columns. "Cabin" was simply missing too many values so I dropped it entirely. Originally, I had intended to remove the "embarked" category as I didn't think it had any significance to the chance of survival. After further research, I found that where an individual boarded the Titanic was a symbol of economic class and a significant indicator on their chances of survival. In order to simplify the name category, I extracted their titles (Mrs, Ms, etc.) which can be used as numeric values. Through encodement, I changed the categorical features of sex, embarked, and title into numeric values. Lastly, I dropped the non-predictive columns such as name and ticket and ultimately split the data back into their original test and training sets.

The next step was to decide what model to train. Under first impressions of the data, and under influence of encodement, every classification (such as “Age", “Sex”, “Survival”, ”Family Size”, “Parch”, “SibSP” and “Cabin") was a binary value. With this assumption, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), and Decision Trees would all be excellent choices, but I decided to use Logistic Regression as it is simple and effective to use for binary classifications. To begin this model, I split the training data into two subsets, a training split and validation split (80/20). I then implemented a higher iteration of 1000, to ensure the models convergence. To assess accuracy, I calculated the accuracy score and printed a classification report which displays the precision and f1-score of each class.

In conclusion, the Logistic Regression Model achieved a validation accuracy of 79.32% which is pretty good! After submission to the kaggle competition, I achieved a score of 77.75% which is additionally good. In the future, I would like to explore more binary machine learning models to see if any others would be a good fit for this data set. Additionally, I want to explore more variation of data inclusion such as family size and unique titles.

Citations:

“Titanic - Machine Learning from Disaster.” *Kaggle*, www.kaggle.com/competitions/titanic/overview. Accessed 25 Feb. 2025.

GitHub, github.com/. Accessed 25 Feb. 2025.

Orvakanti, Praveen kumar. “Surviving the Titanic.” *Medium*, Medium, 21 Oct. 2018, medium.com/@praveen.orvakanti/surviving-the-titanic-f28b39a7b10f.