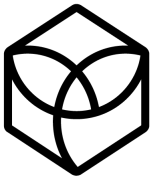

Medius
Custom Software & AI

Technical report
Job application

Titanic survival prediction

Anja Hrvatič

Abstract

This technical report presents an analysis of the Titanic dataset with the aim of predicting passenger survival. The survey involves data exploration, cleaning, feature extraction, and model training using machine learning algorithms including Gradient Boosted Trees (GBT), Logistic Regression, and Support Vector Machines (SVM). Key phases include identifying and handling missing values, redundant features, and exploring the relationship between various features and survival. The best-performing model, SVM, achieved an accuracy of 80.01% and a log loss of 0.4706. Additionally, insights into feature importance highlight factors such as passenger sex, age, and class, which significantly influenced survival predictions. The report underscores the importance of data quality and provides insights for further model optimization.

Keywords
Surviving Titanic, Machine learning, SVM

Introduction

This report outlines the process and results of analyzing the Titanic dataset. The main goal was to build a model to predict if a passenger survived the accident. We used different machine learning algorithms and compared them to find the best one. The report is structured around three primary phases:

1. **Data analysis and clean up:**
   Data exploration, finding anomalies and cleaning the data;
2. **Data preparation and feature extraction:**
   Inferring missing values, dealing with categorical features;
3. **Model training and evaluation:**
   The models we chose were the Gradient Boosted Trees Classifier, Logistic Regression, and SVM.

Methods

Data analysis and clean up

We were dealing with the Titanic dataset that included 891 Titanic passengers and information about them. Initially, the dataset had 15 columns containing information about age, sex, class, number of people they were travelling with, if they survived the journey and other details about their trip.

By a glance at the features, some of the features seemed to contain the same information. For example, there was a binary column *survived* with 0 if not and 1 if yes, and another column *alive* with values "yes" and "no". Both contain the same information. Some features seemed to be redundant as their information was included in a combination of other features. For machine learning, there must be no duplicated features, so we dropped the duplicated and redundant ones. Features with the same information were:

- *embarked* & *embark_town*
- *survived* & *alive*
- *pclass* & *class*

Redundant features were:

- *alone*: if columns *sibsp* (number of siblings and spouses travelling with the passenger) and *parch* (number of children and parents with the passenger) are both 0, the value in the *alone*column will be Ture, and False otherwise.
- *adult male*: will be False for children and women, this information is also included in column *who*, which has possible values woman, man, child.

Our target variable was *survived*, so we checked how the features might influence the survival status of a passenger. Figure 1 shows the relationship between values of different features and the target variable. Some patterns can already be seen. We can expect that the age and the sex will be important

for predicting if a passenger survived. This is following the policy of saving children and women first in case of an accident. Additionally, The class the passenger was travelling in also visibly influences the survival. The fare has some influence on survival as well, but it is not independent of passenger class (higher class costs more) so it makes sense.

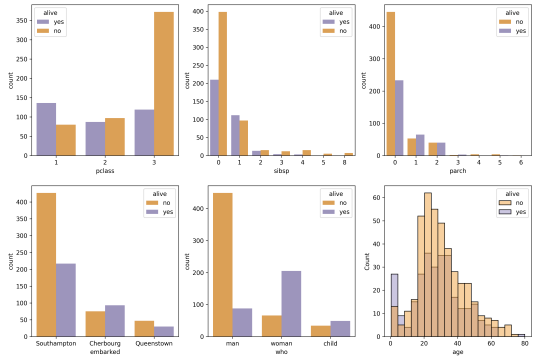

Figure 1. Relationship between features and survival.

We checked for missing values and found there are missing values in columns *age* (19.87%), *embarked* (0.22%), and *deck* (77.22%). There are so few missing values in embarked that they can be ignored. To decide what to do with the missing values in columns *age* and *deck* we have to inspect them further. Since our target variable is the survival status of the passenger we looked at the distribution of missing values for *deck* regarding *survival*. Figure 2 shows there are more values recorded for passengers who survived the accident.

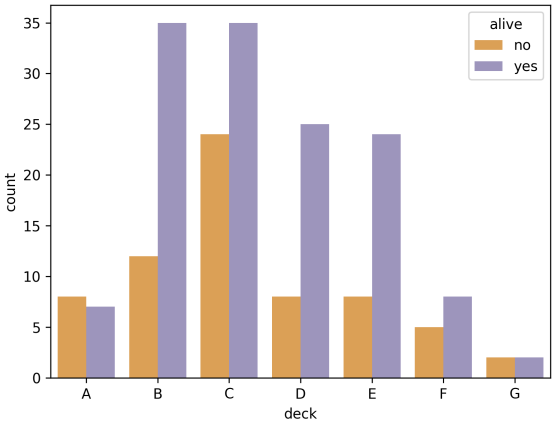

Figure 2. Relationship between *deck* and *survived*.

Class and *deck* might be correlated features as each class might be located on a specific deck. This would mean that *deck* would not bring any new information to our model, so we could ignore the whole feature. In figure 3 we can't see any connection between class and deck, as there are not enough values for the second and the third class. The values that are present still bring some information so we will not discard them.

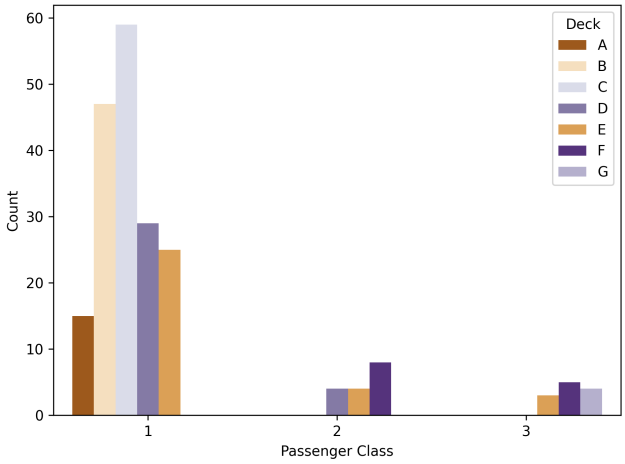

Figure 3. Relationship between *deck* and *class*.

Column *who*, can help us approximate the missing values of *age*, as it tells us if the passenger is an adult or child. It will not be precise but this information might help the model. Figure 4 shows that there is no missing values of *age* for children.

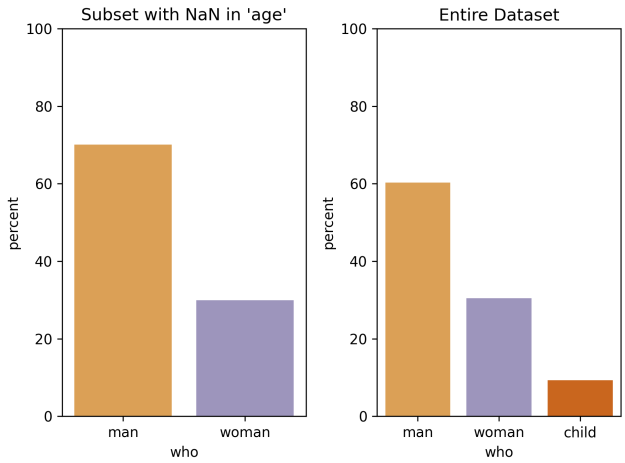

Figure 4. Count of women, men and children for the subset where the *age* is missing and for the whole dataset.

Data preparation and feature extraction

Categorical features need to be turned into dummy features. Even if 77% values in the column *deck* are missing, dummy variables will enable using information about it where it is available, which can benefit the model. For this reason, the column will not be dropped. There are no missing values of *age* for children. We can infer age values for women and men using the mean age for each sex separately. There are two missing values for *embarked*. This is not significant and if there is some information as to why these two values

are missing dummy variables will retain it. Duplicated rows were dropped, because certain machine learning algorithms assume that the observations are independent and identically distributed. Duplicates violate this assumption and may lead to incorrect conclusions.

Model training and evaluation

For Titanic survival prediction we chose to train Gradient Boosted Trees, Logistic Regression and SVM classifiers. The baseline was the majority classifier, which meant we predicted the passenger died. The models were evaluated with 10 times 5-fold cross-validation. The success metrics were log-loss and classification accuracy. The parameters of each of the models were:

- **Gradient Boosted Trees (GBT):**
  - *number of estimators (trees):* 100
  - *max depth of the trees:* 5
  - *learning rate:* 0.1
  - *colsample_by_tree* (fraction of features (randomly selected) that will be used to train each tree): 0.2
- **Logistic Regression:**
  - *preprocessing:* standard normalization of features
  - *max iterations:* 1000
  - *fit intercept:* False
- **SVM:**
  - *preprocessing:* standard normalization of features
  - *kernel:* rbf

Results

Table 1 shows classification accuracy and log loss evaluation through cross-validation of the models. All of them outperformed the majority classifier baseline. They performed similarly well, but the best was SVM with a classification accuracy of 80.01 and log loss of 0.4706.

Table 1. Model performance.

Model	CA	Log-loss
Baseline	$58.80\% \pm 0.0$	$14.85 \pm 5.92e-16$
GBT	$79.95\% \pm 0.0015$	0.4535 ± 0.0017
Logistic Regression	$78.57\% \pm 0.0011$	0.4750 ± 0.0011
SVM	$80.01\% \pm 0.0019$	0.4706 ± 0.0016

For the GBT model, we also looked at feature importance to see what had the main influence on the survival of the passenger. In figure 5 we can confirm that the combination of sex and age was important for survival. If the passenger was an adult male the model tended to a negative prediction.

To see how the logistic regression model made predictions we checked the weights for each of the features. The bigger the absolute value of the weight is, the more influence it has

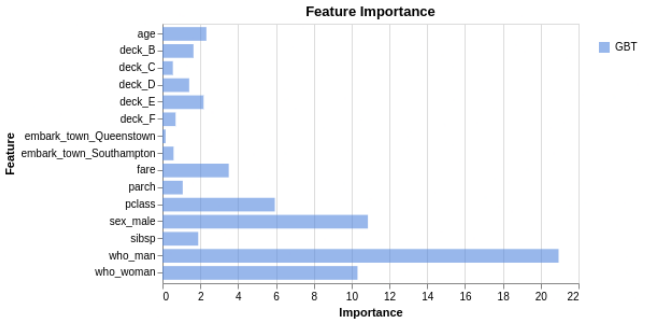

Figure 5. Relationship between *deck* and *class*.

on the prediction. As expected the weight for *adult_male* was negative and had by far the biggest absolute value.

For both models, there was another interestingly important feature. This was the passenger *class*. The reason is, that the lifeboats were positioned in the area of the first and second class. It was more difficult for people in the lower class to get out on the deck to the lifeboats.

Discussion

We analysed the data and built models that performed better than the chosen baseline. The best model was SVM with accuracy of $80.01\% \pm 0.0019$ and log loss of 0.4706 ± 0.0016. Both GBT and logistic regression followed closely. GBT model could be improved by increasing the number of trees. We could increase the maximum depth, but this might result in overfitting to the test set, reducing the performance on the unseen data.

It is important to have data of good quality. There were 20% of missing values for *age* and we decided to replace them with mean age for women and men separately. This method introduces bias as the variance will be underestimated, resulting in a less accurate model.