Who survived from Titanic?

Chloe Hur

Background

The sinking of the Titanic is one of the most known shipwrecks in history. The widely regarded "unsinkable" Titanic sank after striking an iceberg on April 15, 1912, while on her first voyage. Out of 2224 passengers and crew, 1502 perished because there were not enough lifeboats to go around.

Some people appeared to have higher survival rates than others, even though survival sometimes involved a certain amount of luck. Based on the provided passenger information, we'd love to explore the dataset and answer the question: what sorts of people were more likely to survive?

Dataset

The dataset we'll be using is from https://www.kaggle.com/competitions/titanic/data. Using this dataset, given a passenger on the Titanic, we'll predict whether they will survive the shipwreck. This dataset contains several features such as the passenger's sex, age, cabin number, and more:

Variable	Definition
survival	1 = survived, 0 = did not survive
pclass	Ticket class - a proxy for socio-economic status 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower)
sex	Sex (male or female)
age	Age in years - is fractional if less than 1 - if it's estimated, it's in the form xx.5
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

Exploratory Data Analysis (EDA)

The dataset is split into 70% training, a base dataset to train machine learning model, and 30% testing, a dataset to be used for predicting. There are 379 passengers in the training set that did not survive (roughly 61%) and 244 passengers that survived (roughly 39%). In addition, we can see that more females survived than males since there were more than 125 that survived compared to only about 70 males survived). Furthermore, more than 200 males did not survive compared to just under 50 females (*Fig1*).

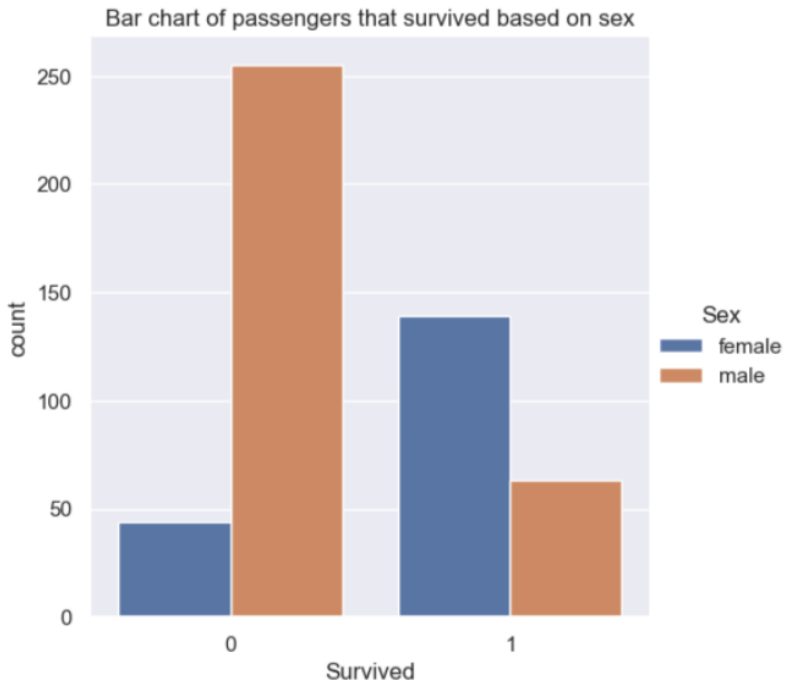

***Fig*.1 Comparing survival of passengers based on sex**

Next, let's see if the ticket class (a proxy for social economic status) plays a role in the survival rate.

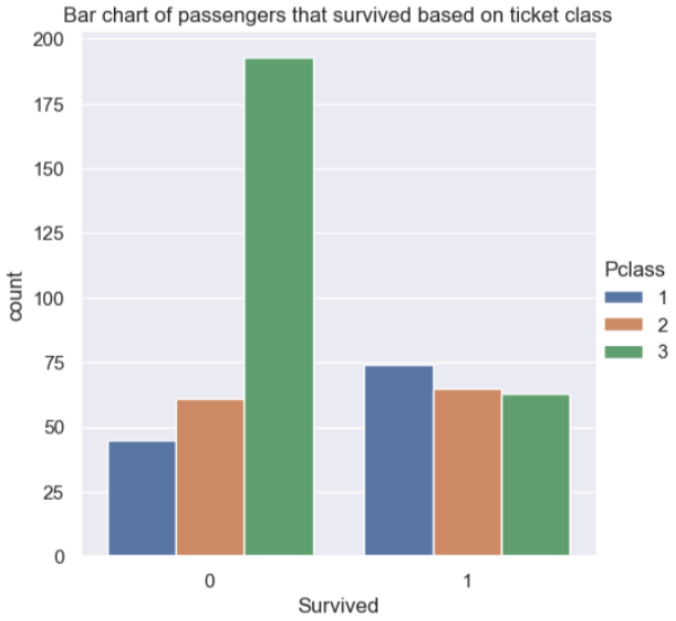

***Fig*.2 Comparison of number of survived passengers based on ticket class (a proxy for socio-economic status 1 = 1st (Upper), 2 = 2nd (Middle), 3 = 3rd (Lower))**

From this figure 2, we can see that most people who failed to survive are from the 3rd class, whereas those that survived the most are from the 1st class.

Next, let's compare the survival rate with the price of the cabin. We create a histogram with the x-axis as the passenger fares grouped into buckets and the y-axis showing the percentage of passengers in the buckets that survived.

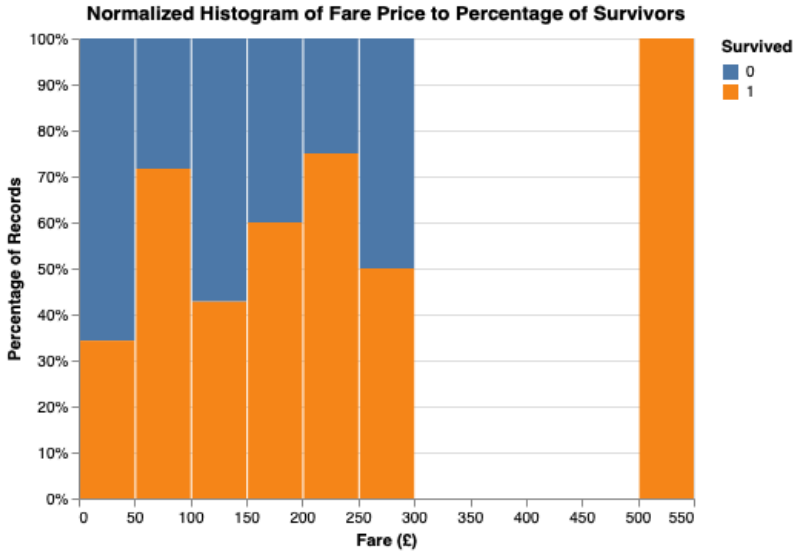

***Fig*.3 Distribution of passenger survival based on fare price of ticket**

Interestingly, we can observe from figure 3 that those who paid very cheap fares did not survive as much compared to those who paid more expensive fares greater than or equal to £50. We can see an outlier with a 100% survival rate for those that paid a fare of more than £500.

Predictive Modeling

To conduct data analysis, we used KNN (K-Nearest Neighbors) for our prediction. We chose this model since it is a simple classification model that requires few assumptions of what the data should look like and it's fast in terms of training.

To determine which features to use for prediction, as shown below(*fig4*), we found that the two features Sex and PClass have the highest p-value and score, thus they have the most influence on whether the passenger survived. Therefore, we'll use Sex and PClass as our main features for the KNN model.

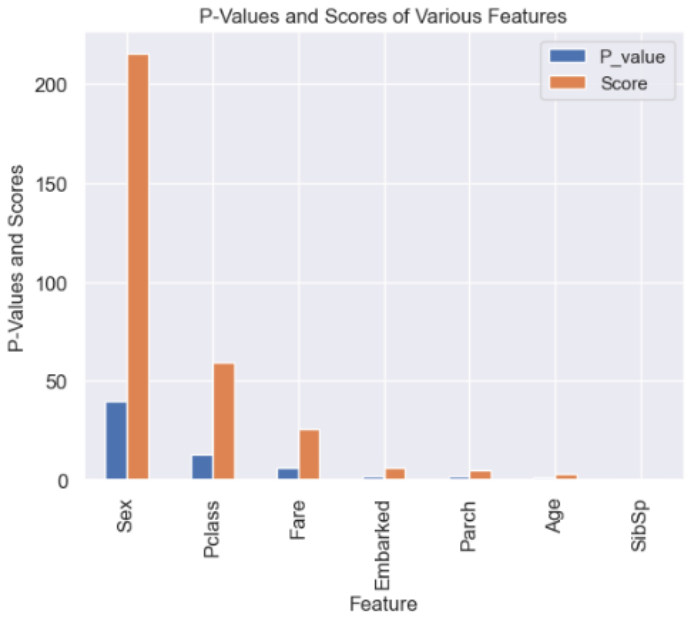

***Fig*.4 Comparisons of P-values and Scores between variables**

Prediction and Result

We evaluated our model on the test dataset and made predictions as shown below table1. We found that we got 0.806 accuracy of prediction on test set.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	predicted
0	179	0	2	Hale, Mr. Reginald	male	30.0	0	0	250653	13.0000	NaN	S	0
1	458	1	1	Kenyon, Mrs. Frederick R (Marion)	female	NaN	1	0	17464	51.8625	D21	S	1
2	17	0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.1250	NaN	Q	0
3	96	0	3	Shorney, Mr. Charles Joseph	male	NaN	0	0	374910	8.0500	NaN	S	0
4	121	0	2	Hickman, Mr. Stanley George	male	21.0	2	0	S.O.C. 14879	73.5000	NaN	S	0

***Table.1* Predictions whether passengers survived on test dataset**

Discussion

Limitations

As we explored the dataset, we found that there were some missing values in several categories, such as Age, Cabin, and Embarked. Thus, we used “imputer” to impute missing values. This could result in errors that might be different from actual values. In addition, since there was class imbalance between the number of people who survived and not survived, it could lead a misleading accuracy metric. As shown in figure 5 below, we can see that our model has more false negatives than false positives. Namely, it tends to predict that a passenger does not survive given that the passenger did survive.

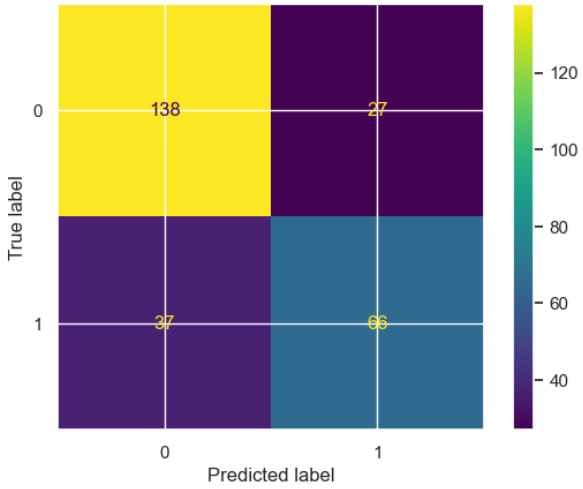

***Fig*.5 Confusion Matrix between true label and predicted label where 1=Survived and 0=Did not survive**

Another caveat that we need to consider is that there were models that performed better than KNN, such as RBF SVM, which had better test score than KNN. Thus, we can reconduct a model with RBF SVM to see better prediction.

Impact of Findings and Future Questions

Through data analysis with classification, we want to figure out what sort of people were more likely to survive from Titanic shipwrecks. As we classify the training data set, we could find specific variables/features such as “sex” or “ticket class” that influence the likelihood of survival in the Titanic disaster. For example, we would expect that passengers with higher ticket classes or those who lived in higher cabin numbers might have survived more than other groups of people. The impact of these findings can help us better understand which groups of people were more likely to survive than others. This could lead to future questions such as whether these groups shared similarities to survivors of other large-scale boat accidents or natural disasters that also resulted in a large number of deaths. Lastly, we could use these findings and further research methodologies to maximize the number of survivors in case an event like this were to happen again.