Titanic, a social class tragedy

CS3573 – Ruan

Fabian Aguilar Gomez

November 24, 2020

Contents

Abstract

The purpose of this project is to use a dataset in order to test the skills that have been taught in the class throughout the semester. These skills include and not limited to data cleaning and loading, preprocessing, analysis, plotting, and machine learning. With these sets of skills that have been taught and reinforced through the homework assignments the project will serve as a steppingstone into analyzing and creating useful machine learning models on real world data. In this specific project the purpose is to take the dataset from the Titanic and analysis it in order to determine the probabilities of surviving based on factors such as the class, sex, and age , just to name a few. Using the libraries and tools we have learned in class such as *matplot, numpy, pandas,* and *sklearn* and the programming language of python we will perform the fundamental steps in a data science project in order to achieve what we set out to do see the probabilities of people surviving and why as well as to reinforce and show off our skills that we learned in the class. The sex of a person, number of siblings / spouses on board, number of parents / children on board, and the ticket class in played a role in determining if a person survived the Titanic disaster or not. When looking at age the trend shown was those that were younger or older than the average age had a higher probability of surviving. Females had almost twice the chance of surviving than males, and the number of siblings one had on board with them had good correlation with the outcome, with the sex being the biggest contributor to the outcome of a person surviving or not. This project was able to show that individuals who were female, young/elderly, and had a ticket which placed them in a higher class in the Titanic had a higher chance of surviving the Titanic. Those, who saw themselves in the opposite side of the spectrum that being a male, in their mid-20 or mid-30, and were in the lower classes did not have a good chance of surviving the Titanic.

***Keywords**:* Titanic, data science, machine learning, Kaggle, python, sklearn

Introduction

The Titanic was an infamous ship that was so called as “unsinkable”. It set sail on April 10, 1912 and on sank on April 15, 1912 only 5 days after the it had set sail. This disaster was showcased everywhere when news broke out and to this day it is still remembered as the famous shipwreck. What made this disaster famous was the fact that it did not have enough lifeboats for everyone on board of the ship since it was claimed to be “unsinkable”. Because of this negligence many of the passengers did not make it out alive and those who did were at the time some of the richest people in the world. The total amount of casualties is estimated to be over 1,500 and only 705 passengers survived that is about only 32% of the passengers that survived while about 68% died. I chose to study this incident to showcase how even back in 1912 the status of a person mattered more even in times of crisis and how having a higher status in society would give an unfair advantage over the others in the lower class. The goal at the end of this project is to back up my claim stated in the previous sentence and to shed on a light even more by being able to show the chances one had to survive the disaster if they were to be a part of it. This will be done by creating a model which will use the dataset and learn from it in order to be able to predict based on the one’s parameters if they would have survived the Titanic or not. The takeaway that after this report is that we can attempt to showcase how these “classes” in society give an unfair advantage to those that are blessed enough to be part of them and to shed some light on the hassles and challenges those in the lower bracket have to endure. There is no better way to show this than by visualizing the casualties and survivors in the Titanic disaster to validate such claim. With the tools that have been learned in the course and the topics we have covered thus far about machine learning this task should be a good challenge to not only prove our claim but also showcase the skills we have learned by analysis some real data and being able tell a story with the analysis.

Materials and Methods

In order to analyze the Titanic disaster a dataset which is valid and makes sense will be needed. The dataset which was used in this project was the Kaggle Titanic dataset which can be found on their website. The dataset consists of 891 rows and 12 columns. The columns represent a feature in the dataset such as age whereas a row in the dataset represents a passenger aboard the ship. The dataset is also stored as a Comma Separated File (CSV) which means that the features for each passenger are separated by commas. This is useful since we know how the dataset is structured, we can use the function *read_csv* from the *pandas* library in python in order to read the dataset and store the information in a data frame.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Table 1. Data frame of the Titanic dataset

Each row in the data frame contains a featured that belongs to each passenger. Some of the values in the features can be the same in between the passengers as for example multiple passengers could be female or male.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
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

Table 2. Variable definition for each of the columns in the dataset

The dataset contains many features which are not useful to use when we attempt the machine learning aspect of the project. Such features include the Name, Ticket, and Cabin as these features do not help to tell a story with numbers, as each person probably has a unique name and the name more than likely won't be a contributor to the passenger's survival. Additionally, rows that contain missing information or NaN values in certain columns need to be dropped from the dataset as it will just cause inconsistencies in the data analysis.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

Table 3(a). Data frame after some pre-processing

After some simple pre-processing, the data frame shape is that of 712 rows and 8 columns. This data frame will be used to do some simple data analysis on the dataset and see some trends that occurred. One final pre-processing does need to occur when we are preparing the data frame we need to convert the values in the 'Sex' and the 'Embarked' columns to unique integers as we will need to encode them as such to have a successful model and process. This was achieved by using the *LabelEncoder* function from *sklearn.preprocessing*

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	2
1	1	1	0	38.0	1	0	71.2833	0
2	1	3	0	26.0	0	0	7.9250	2
3	1	1	0	35.0	1	0	53.1000	2
4	0	3	1	35.0	0	0	8.0500	2

Table 3(b). Final data frame after pre-processing

The models that will be used when creating the Machine Learning algorithms are Logistic Regression, K-Neighbors, SVC linear, SVC RBF, Gaussian Naïve Bayes, Decision Tree Classifier, and Random Forest Classifier. These are found in the *sklearn* library and the reason for that many is to able to have to options to choose the best model for our dataset.

Results and discussion

Some interesting trends were seen after doing an analysis on the data and seem to back up the original hypothesis stated.

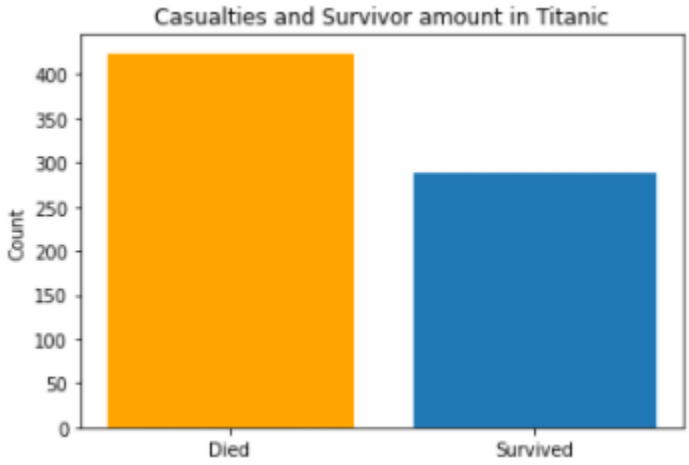

Figure 1. Total number of casualties and survivor in the Titanic

Before jumping into specific features, we need to look at the overall number of casualties and survivors in the Titanic. Figure 1 visualizes the number of passengers that died and those that survived and to no surprised almost twice the number of passengers aboard the ship died in the disaster.

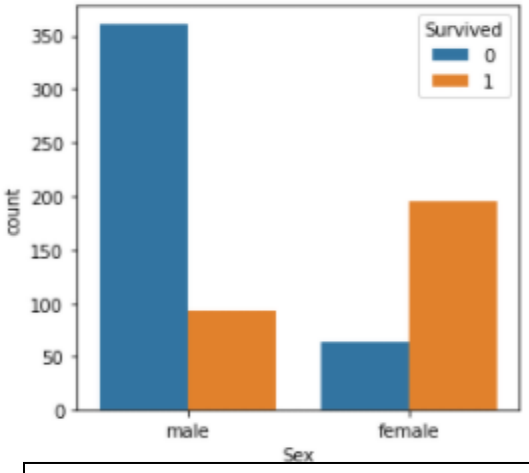

Figure 2(a). Number of casualties and survivors by sex

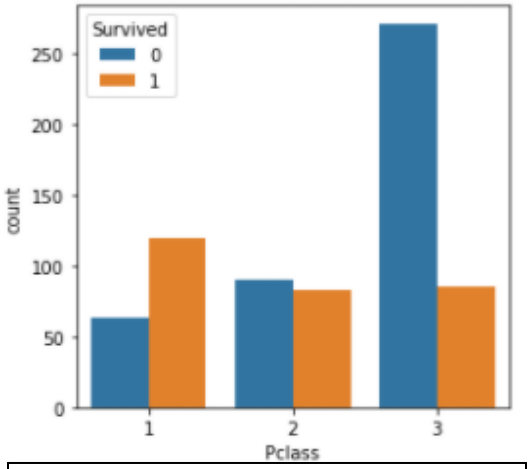

Figure 2(b). Number of casualties and survivors by class

Figures 2(a) and 2(b) show that those passengers who were in the third class had almost little chance of surviving as we can see the number of passengers that died is almost triple the number who survived. Compared to the other two classes the discrepancy is not that big specially with first class as almost double the number of people survived that were in the first class. Looking at 2(a) the number of males that died was almost four times the amount than those who survived. In contrast, the

number of females that survived was almost four times, two complete opposites of each other. Looking at the other features we can see that ‘Fare’ and ‘Age’ seem to have a positive correlation which could mean they would have increased the chances of

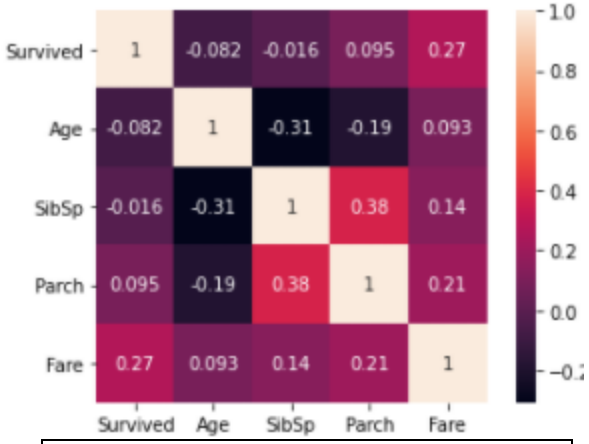

Figure 3. Correlation between age, # of siblings, # of parents/children, fare, and survived

surviving as opposed to the other features Figure 3 shows this by means of a correlation heatmap. Moving on to the machine learning results after running our dataset through the models that were stated in the previous section Random Forest Classifier yielded the highest accuracy and thus it was the chosen model moving forward. Using the trained model of choice determining the priority of each of the features was crucial since determining which feature was important to the survival of passenger would help in understand how the model will predict the outcome.

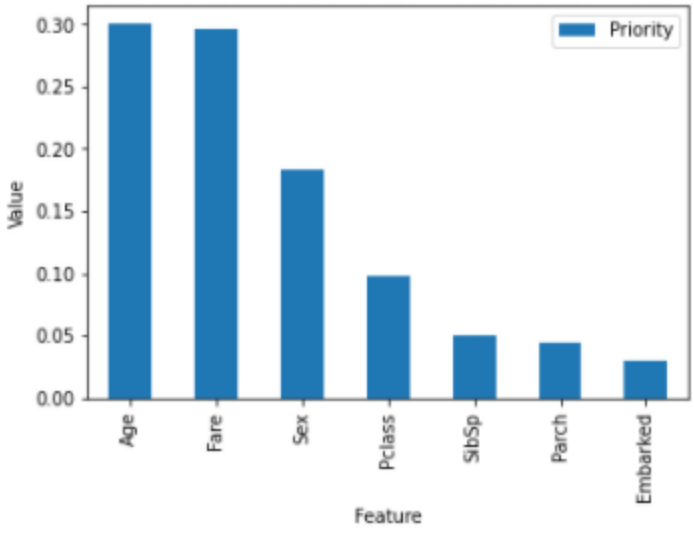

Figure 4. Priorities of features based on a Random Forest Classifier model

Figure 4 shows that ‘Age’, ‘Fare’, ‘Sex’, and ‘Pclass’ were the top features that played a role on the survival of each passenger and it makes sense as to why Fare is so high. A person of high status has the means and wealth to pay for a good ticket in the Titanic whereas a person of low status probably does not have access or means to pay for a

high-priced ticket. Figure 5 can visualize the discrepancy between the classes even further by showing the fare paid by each passenger based on their class.

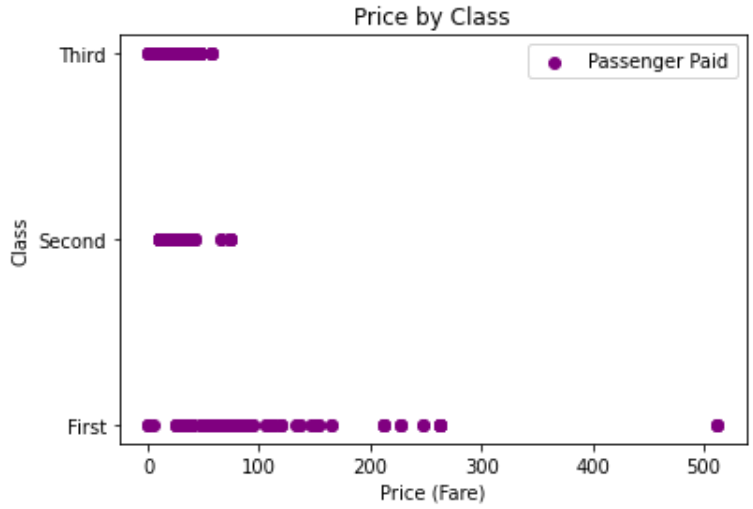

Figure 5. Fare paid by everyone based on class

Using what has been analyzed and the trained machine learning model predictions can be made, and they would make sense. Using the file provided by Kaggle called "test.csv" testing the to see if the model makes sense was made simple since the file format is the same as the training data set.

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	0
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	0
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q	0
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S	1
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S	1
5	897	3	Svensson, Mr. Johan Cervin	male	14.0	0	0	7538	9.2250	NaN	S	0
6	898	3	Connolly, Miss. Kate	female	30.0	0	0	330972	7.6292	NaN	Q	0
7	899	2	Caldwell, Mr. Albert Francis	male	26.0	1	1	248738	29.0000	NaN	S	0
8	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18.0	0	0	2657	7.2292	NaN	C	0
9	901	3	Davies, Mr. John Samuel	male	21.0	2	0	A/4 48871	24.1500	NaN	S	0
11	903	1	Jones, Mr. Charles Cresson	male	46.0	0	0	694	26.0000	NaN	S	0
12	904	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	23.0	1	0	21228	82.2667	B45	S	1
13	905	2	Howard, Mr. Benjamin	male	63.0	1	0	24065	26.0000	NaN	S	0
14	906	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance...	female	47.0	1	0	W.E.P. 5734	61.1750	E31	S	1
15	907	2	del Carlo, Mrs. Sebastiano (Argenia Genovesi)	female	24.0	1	0	SC/PARIS 2167	27.7208	NaN	C	1
16	908	2	Keane, Mr. Daniel	male	35.0	0	0	233734	12.3500	NaN	Q	0
17	909	3	Assaf, Mr. Gerios	male	21.0	0	0	2692	7.2250	NaN	C	1
18	910	3	Ilmakangas, Miss. Ida Livija	female	27.0	1	0	STON/O2. 3101270	7.9250	NaN	S	0
19	911	3	Assaf Khalil, Mrs. Mariana (Miriam")"	female	45.0	0	0	2696	7.2250	NaN	C	0
20	912	1	Rothschild, Mr. Martin	male	55.0	1	0	PC 17603	59.4000	NaN	C	1

Figure 6. Testing data frame after using the data set on the trained model

Figure 6 shows the result of the testing data after it was put through our model and on the last column, we could see the predicted value to see if they survived or not. The

common trend still holds valid that is those who were of higher class had a higher class did indeed tend to survive as opposed to those in lower classes.

Conclusion

From the findings in this project the hypothesis that those of higher status had an advantage over those who were of lower status is solidified. This is shown by demonstrating that those in higher classes were prioritized in the Titanic and had access to the limited number of lifeboats in the ship. Thus, those in higher classes had a higher probability to survive the disaster. This study is limited in scope as this disaster occurred over 100 years ago and to demonstrate that the statement still holds true newer data from another disaster or related to the topic would have to be used. Some improvements for future projects would be to get data that is more recent to today's time and probably today's issues but nonetheless this project still holds some validity since the disaster is still being talked about today. From this project I hope others learn to use data and current events to shed some light on these topics and demonstrate how those in higher classes have an advantage using numbers.

References

1. "Titanic." *Encyclopædia Britannica*, Encyclopædia Britannica, Inc., 15 Oct. 2020, www.britannica.com/topic/Titanic.
2. "Titanic:Machine Learning from Disaster." *Kaggle,* www.kaggle.com/c/titanic/data.