Titanic classification report

Maria Musiał 156062
Wiktoria Szarzyńska 156058
Joanna Szczublińska 156070
Lidia Wiśniewska 156063

April 2024

Contents

Introduction

The sinking of Titanic remains one of the most tragic shipwrecks in history. It took place in 1912 during its maiden voyage. Considered "unsinkable", yet it collided with an iceberg, resulting in death of 1502 out of 2224 passengers. Surviving this tragedy was mostly a matter of luck, but we want to understand what factors might have influenced survival.

Our goal is to perform exploratory data analysis to understand the data. Later, we will perform preprocessing techniques in order to extract information from the data. Finally, we will check general classification result of survival.

This report is divided into several sections. First, we will introduce the dataset and explain what it contains. Then, we will perform exploratory analysis of the input and later on output features. We will explain preprocessing techniques used and motivation behind it. Lastly, we will see conclusions.

Description of the dataset

The Titanic dataset is a classic dataset used in machine learning and statistical modeling. It contains information about passengers aboard the RMS Titanic, which tragically sank on its maiden voyage on April 15, 1912. The dataset is often used for predictive modeling tasks, particularly binary classification, to predict whether a passenger survived or not based on various attributes.

Table 1: Titanic Dataset Example(Part 1)

Survived	Pclass	Name	Sex	Age
0	3	Braund, Mr. Owen Harris	male	22
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
1	3	Heikkinen, Miss. Laina	female	26
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
0	3	Allen, Mr. William Henry	male	35
0	3	Moran, Mr. James	male	
0	1	McCarthy, Mr. Timothy J	male	54
0	3	Palsson, Master. Gosta Leonard	male	2
1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27
1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14

Table 2: Titanic Dataset Example(Part 2)

Ticket	Fare	Cabin	Embarked
A/5 21171	7.25		S
PC 17599	71.2833	C85	C
STON/O2. 3101282	7.925		S
113803	53.1	C123	S
373450	8.05		S
330877	8.4583		Q
17463	51.8625	E46	S
349909	21.075		S
347742	11.1333		S
237736	30.0708		C

Here's a brief description of the Titanic dataset:

- **Source:** The dataset is derived from passenger records and historical data.
- **Features:**
  1. **Passenger ID:** A unique identifier for each passenger.
  2. **Survived:** This is the target variable. It indicates whether a passenger survived (1) or did not survive (0).
  3. **Pclass (Passenger Class):** The class of the ticket purchased by the passenger (1st, 2nd, or 3rd).
  4. **Name:** The name of the passenger.
  5. **Sex:** The gender of the passenger (male or female).
  6. **Age:** The age of the passenger in years.
  7. **SibSp:** The number of siblings/spouses aboard the Titanic.
  8. **Parch:** The number of parents/children aboard the Titanic.
  9. **Ticket:** The ticket number.
  10. **Fare:** The fare paid by the passenger.
  11. **Cabin:** The cabin number.
  12. **Embarked:** The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).
- **Target Feature:**
  - **Survived:** This binary variable indicates whether a passenger survived (1) the sinking of the Titanic or not (0).
- **Number of Instances:** The dataset contains 891 passenger records. We will divide them 60-40 to get training and testing datasets.
- **Missing Values:** The dataset contains missing values, particularly in the 'Age', 'Cabin' and some in 'Embarked' features. Handling missing values will be addressed later on.

Description of the input features

1. **Passenger ID:**
   - **Description**: A unique identifier assigned to each passenger.
   - **Type**: Discrete.
   - **Purpose**: It serves as an identifier and is not used for predictive modeling.
2. **Pclass (Passenger Class):**
   - **Description**: Indicates the class of the ticket purchased by the passenger.
   - **Type**: Discrete.
   - **Values**: 1st class, 2nd class, 3rd class.
   - **Purpose**: Class may be correlated with socio-economic status, potentially impacting survival chances. Higher classes may have had better access to lifeboats.
3. **Name:**
   - **Description**: The name of the passenger. Contains title, name, surname. Can contain second name or family name.
   - **Type**: Categorical.
   - **Purpose**: Name doesn't directly impact survival.
4. **Sex:**
   - **Description**: The gender of the passenger.
   - **Type**: Categorical (binary).
   - **Values**: Male, Female.
   - **Purpose**: Gender may have had a significant impact on survival, as "women and children first" was a priority during the evacuation.
5. **Age:**
   - **Description**: The age of the passenger in years.
   - **Type**: Continous.
   - **Purpose**: Age can be a critical factor in survival, as children and elderly passengers may have received priority during evacuation. However, there are some missing values that need to be addressed.
6. **SibSp:**

- **Description**: The number of siblings/spouses of passenger aboard the Titanic.
- **Type**: Discrete.
- **Purpose**: Family size may have influenced survival, as individuals with family members aboard may have assisted each other during the evacuation.

7. **Parch:**
   - **Description**: The number of parents/children of passenger aboard the Titanic.
   - **Type**: Discrete.
   - **Purpose**: Similar to SibSp, family size may have played a role in survival, with parents prioritizing the safety of their children.
8. **Ticket:**
   - **Description**: The ticket number.
   - **Type**: Categorical.
   - **Purpose**: Ticket number does not have direct predictive value for survival.
9. **Fare:**
   - **Description**: The fare paid by the passenger.
   - **Type**: Continuous.
   - **Purpose**: Fare may correlate with class and socio-economic status, both of which could impact survival.
10. **Cabin:**
    - **Description**: The cabin number.
    - **Type**: Categorical.
    - **Purpose**: Cabin information may indicate the location of the passenger aboard the ship, which could be relevant to survival (e.g., proximity to lifeboats). However, there are many missing values in this feature.
11. **Embarked:**
    - **Description**: The port of embarkation.
    - **Type**: Categorical.
    - **Values**: C = Cherbourg, Q = Queenstown, S = Southampton.
    - **Purpose**: Embarkation port may be correlated with socio-economic status and could indirectly impact survival.

Exploratory analysis of the input features

First of all, we will start by saying that **PassengerId**, **Name**, **Ticket** and **Cabin** are features that won't give us much information, as they are almost unique or have too much missing values (as in case of Cabin). We will leave it out of deeper exploratory analysis.

Table 3: Summary Statistics for Numerical Features

	PassengerId	**Survived**	**Pclass**	**Age**	**SibSp**	**Parch**	**Fare**
count	891.00	891.00	891.00	714.00	891.00	891.00	891.00
mean	446.00	0.38	2.31	29.70	0.52	0.38	32.20
std	257.35	0.49	0.84	14.53	1.10	0.81	49.69
min	1.00	0.00	1.00	0.42	0.00	0.00	0.00
25%	223.50	0.00	2.00	20.12	0.00	0.00	7.91
50%	446.00	0.00	3.00	28.00	0.00	0.00	14.45
75%	668.50	1.00	3.00	38.00	1.00	0.00	31.00
max	891.00	1.00	3.00	80.00	8.00	6.00	512.33

Most important things to conclude:

- Passenger Id has unique values, indicating that it won't be useful for modeling purposes.
- Age's standard deviation suggests some variability between passengers in terms of age.
- Fare has a wide range, but the 75% percentile indicates that most of the values are low.

Categorical features

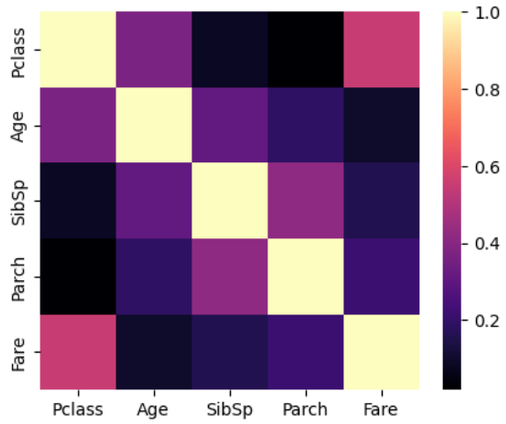

Figure 1: Correlation heat map

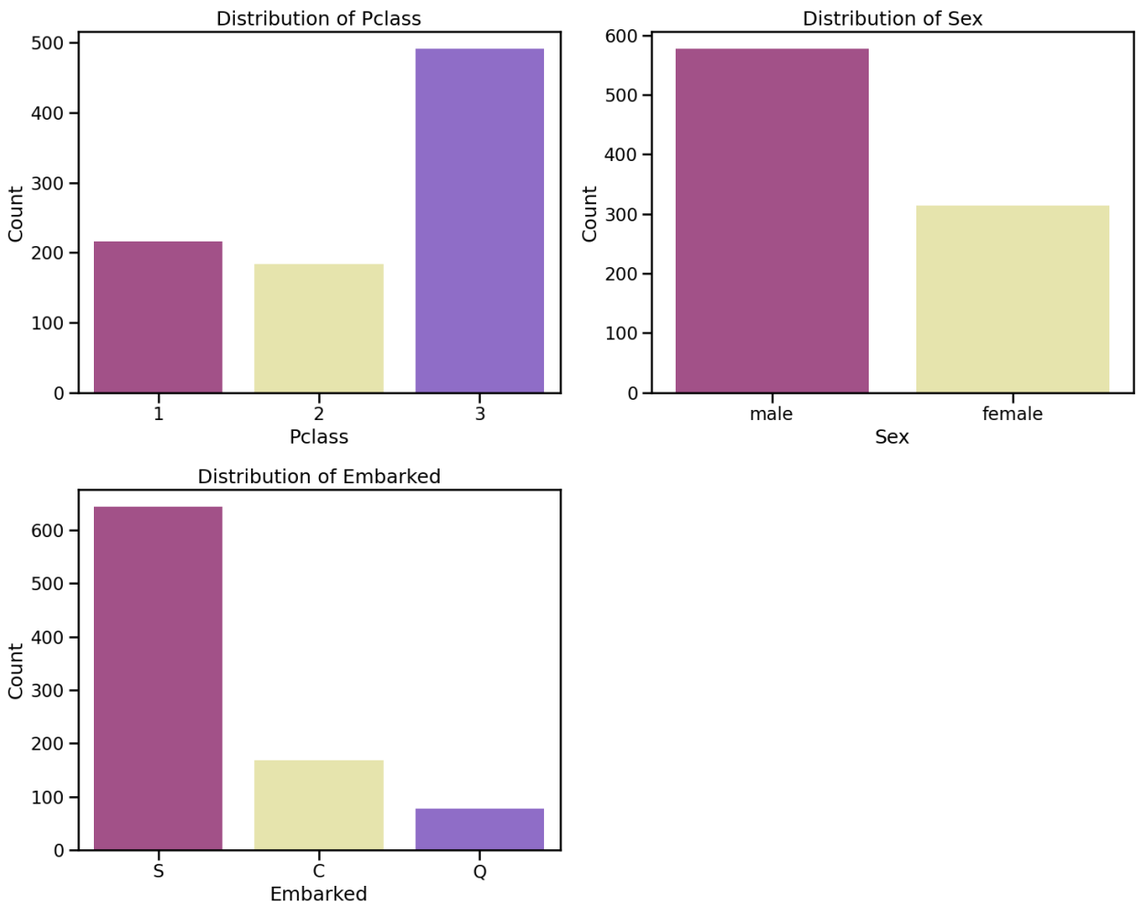

Figure 2: Categorical Histograms

- **Pclass**: Most of passengers were in class 3. Feature is highly correlated with Fare (as price reflects on class). It is also lightly correlated to Age.
- **Sex**: Majority of passengers were male. As we can see in figure below, survival rate for women was significantly higher than for men.
- **Embarked**: Most of passengers embarked in Southampton. From figure below, we can see that the highest probability of survival was for people who embarked in Cherbourg.

Numerical features

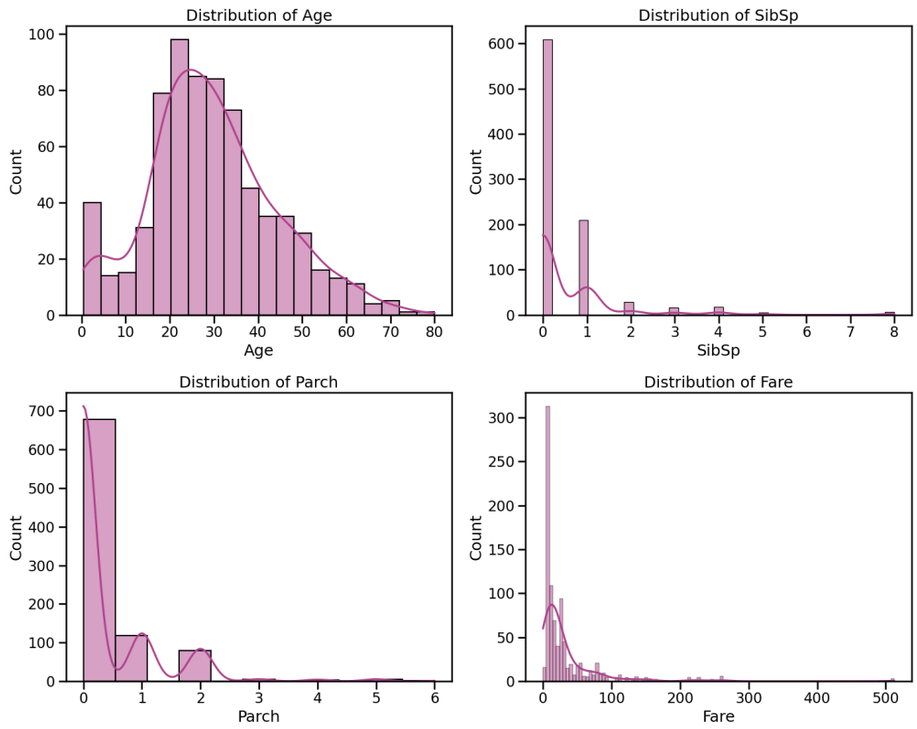

Figure 3: Distribution of numerical features

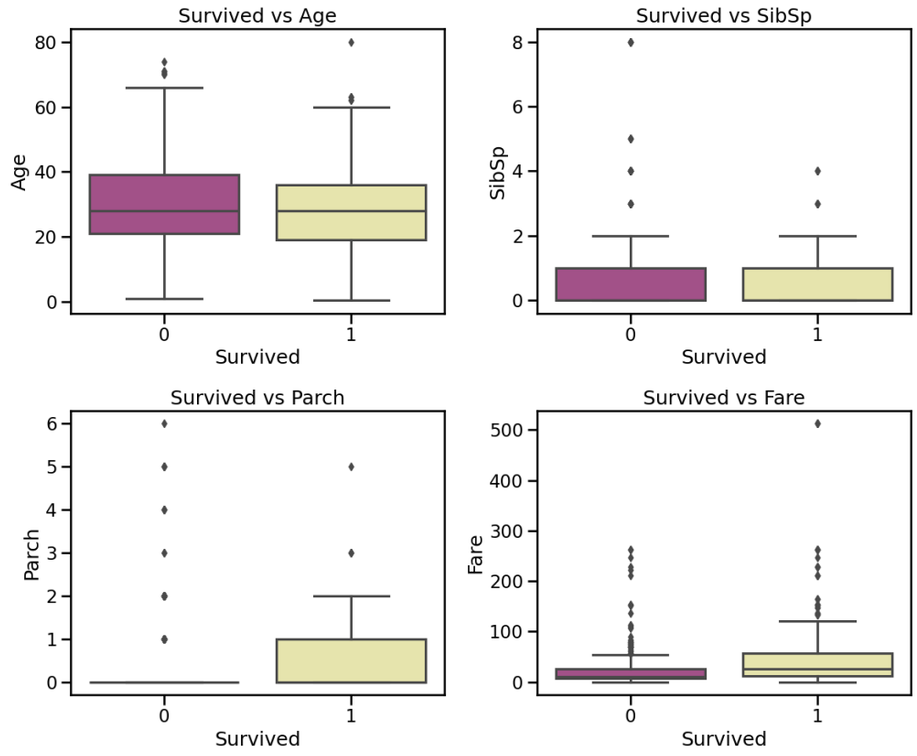

Figure 4: Survival vs Numerical features

- **Age**: Majority of passengers were of Age 20-40. The correlation low with Pcalss fature and with SibSp feature, meaning number of siblings/spouses. We can think that it's because its more likely for children to be on the ship with multiple siblings. In relation to survival we can see subtly that younger people a bit more likely to survive.
- **SibSp**: Majority of people travelled alone, some with one person. This feature has high correlation with Parch. It is probably because if someone had a sibling abroad its very likely they also had parents with them.
- **Parch**: Here we have the same story as with Siblings/Spouses. Majority of people travelled alone and high correlation with Siblings/Spouses feature.
- **Fare**: Highly correlated with Pcalss, as described earlier. Distribution is skewed, as almost all tickets were under 100 with highest price going up to 512. If we observe this feature in relation to survival, we can see that people with higher fare were more likely to survive.

Preprocessing techniques used in the assignment

Description

Data Cleaning

The methods we used include:

- Replacing null values with new values based on the data type of the column.
  - For numerical columns, replaces null values with the maximum value in the column plus 1.
  - For categorical columns, replaces null values with the string 'Null'.
- Changing the 'Name' column to contain only surnames. They can be the same for married people.
- Encoding categorical columns from dataset using LabelEncoder.
  - Encoding categorical labels as integer numbers. Perfect solution Sex feature. We were thinking about OneHot encoder, as it could be better, but we didn't want to add another columns.
- Discretizing continuous values in columns using KBinsDiscretizer.
  - Used for binning numerical features into discrete intervals (bins). Focuses on maximazing the distances between geometrical means. Changing continuous values which are in 'Age' and 'Fare' class makes our model more interpretable and roubstness to outlieres in these features. If there is any pattern it should be now easier to found.

Data Transformation

- Normalization of the data using MinMax Scaler
  - It's a simple transformation that can be applied to numerical features without requiring complex parameter tuning. We normalize features to be in range 0-1

$$v' = \frac{v - min}{max - min} \cdot (max' - min') + min'$$

Feature Extraction

- PCA- Principal Component Analysis - Transforms original, high dimensional dataset into smaller, lower dimensional space while preserving the information. New components are linear combinations of original features.
  - **Dimensionality reduction**: reduces the number of features in the dataset, making it easier to visualize and analyze high-dimensional data.
  - **Noise reduction**: helps reduce the effects of redundant information in the data.
  - **Feature Extraction**: extracts the most important information from the original features using linear combination of them. Represents it in a lower-dimensional space, allowing for more efficient computation and modeling.
  - **Helps avoiding overfitting**

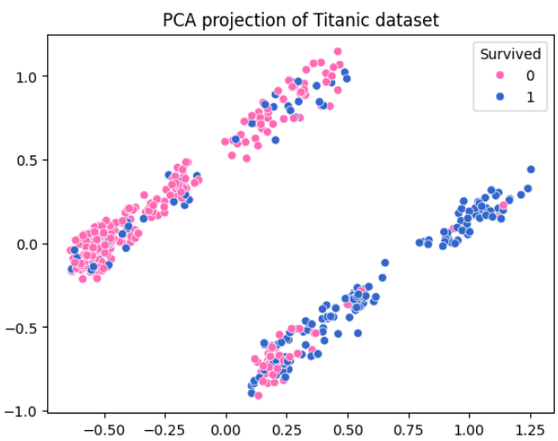

Figure 5: Visualisation first 2 principal components

Feature Selection

Now, we are selecting best k features using ANOVA score (f-classif)

ANOVA (Analysis of Variance) F-value is a statistical test used in supervised feature selection to assess the significance of the relationship between each feature and the target variable (output feature).

This method calculates the F coefficients and p-values for each feature in the context of classification. The higher the F coefficient, the more significant the feature is for classification.

As we can see in figure below, feature 0 has the most information, therefore we can drop all others, that carry little to none information.

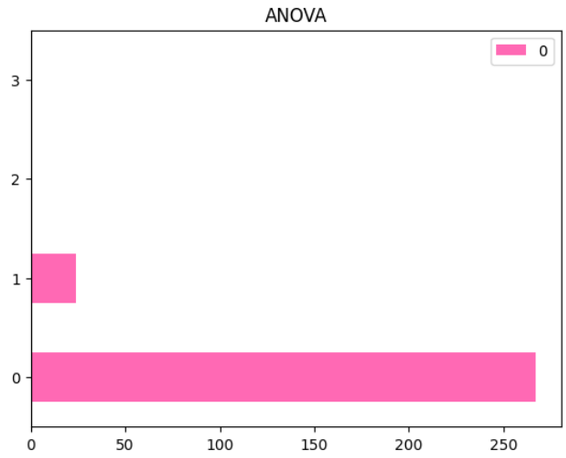

Figure 6: ANOVA for our features after PCA

Motivation

Motivation for using PCA was expressed earlier. All advantages for our methods are expressed in their description.

We tried the using ANOVA and LDA, but the data isn't normally distributed which i think is the reason for lower accuracy. We also experimented with chronology of dimentionality reduction and feature selection and decided that reduction then selection works best for our data.

Description of the output features

- **Survived (Output Feature):**
  - **Description**: Indicates whether a passenger survived the Titanic disaster.
  - **Type**: Categorical (binary).
  - **Values**: 0 = Did not survive, 1 = Survived.
  - **Purpose**: This is the target variable for predictive modeling. It indicates the outcome of interest-whether a passenger survived or not and is used to train model to predict survival based on other features.

Exploratory analysis of the output features

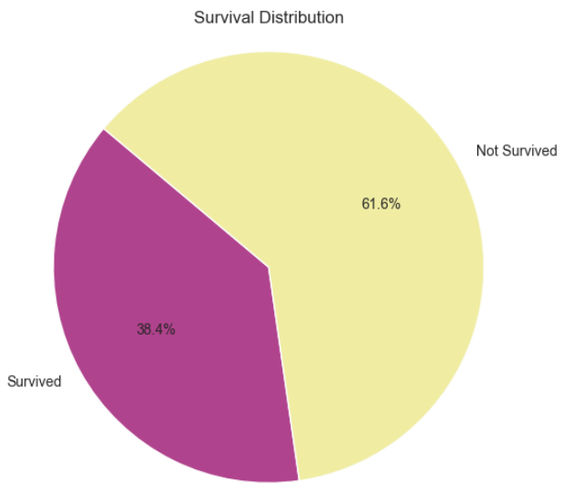

Figure 7: Distribution of survival

The pie chart illustrates the distribution of survival among passengers aboard. As we can see, the majority of passengers (61.6%) did not survive.

Conclusions

Future enhancements

There is space for future exploration of this dataset that could enhance our performance.

- **Feature engineering:** there is space for future exploration of this dataset that could enhance our performance.
- **More advanced handling of missing data:** Additional research into history of our dataset and its origin could give us idea for more adequate handling of missing values. But it required domain knowledge.

Chronology of techniques

Regarding chronology between dimentionality reduction and feature selection we think there are several reasons to choose reduction first for our data:

- **High dimentionality**: Our dataset has a lot of features (high dimentionality). Performing PCA first allows us to reduce computational complexity.
- **Correlated Features**: Identifying and removing muilticollinearity. It helps in improving model performance.
- **Noise reduction**: Filtering out irrelevant features first makes feature selection more effective.
- **No need for interpretability**: Selecting feature selection first would be important if we want to achieve interpretability. In our case, we don't need this information, our goal was to achieve high accuracy of classification.