Casualty on the Titanic based on Machine Learning Methods

Yang Liu *

Department of Computer Science, George Mason University, Virginia, the USA

* Corresponding author email: yliu51@gmu.edu

Abstract. The Titanic sank on April 15, 1914, with 2224 people on board, and only 32% survived. The survivors are somewhat random, but they are somewhat the same. Studying the types of people who are more likely to survive in a disaster will promote an understanding of the values and ideology of the society at the time. Therefore, this study collected data on some of the passengers on the Titanic and the survivors through Kaggle. This study predicts the survival of passengers through data cleaning, feature engineering, dimensionality reduction processing, and different models such as Random Forest, Decision Tree, KNN, and Logistic regression. In this study, it was found that Age, gender, representation of socioeconomic status, and whether they traveled alone were associated with survival rates. Among them, the decision tree and random forest models scored 98.65, which performed the best and had the highest prediction accuracy. After visualizing the data, this study concluded that survival rates were higher for females, children younger than ten years old first-class passengers, and non-solitary travelers.

Keywords: Machine Learning; Survival Rate; Feature Engineering; Feature Selection; Python.

1. Introduction

The advent of the Internet has made the dissemination of information more accessible, faster, and on a larger scale. The spread of information and people's connection to the Internet has created an enormous amount of data. People began to try to use the data and make predictions about future trends by analyzing it. In supervised machine learning, people collect data, analyze, and present statistical analysis, modeling, and prediction. The application area of this prediction is very wide, and it can be applied in the field of business, residential security, medicine, weather, historical research, and disaster. As a representative of the accidents of the 20th century, analyzing the characteristics of the Titanic's survivors can provide a better understanding of how limited survival resources would be allocated to the population in an accident. The main reason for the minimal number of survivors in the Titanic tragedy was the lack of adequate life rafts. After removing this essential factor, survival was the dependent variable in this study, and age, gender, social status, the number of parents and children, and the number of siblings and spouses were the independent variables. The analysis of the study explored how the independent variables would affect the survival rate.

The data about Titanic was obtained from Kaggle. Since the obtained data is raw, the dataset needs to be cleaned before modeling. Data cleaning is the final procedure to find and correct identifiable errors in the data files. Whether the data is "clean" or not has a direct impact on the efficiency and accuracy of the model training, so data cleaning is essential. Data cleaning includes the treatment of missing values. In this research paper, there are several ways to fill the null records, first is filling by mean, median, mode. Second is filling by unknow. This research paper uses two ways of encoding categorical features: one is one-hot encoding, the other is target encoding, and two kinds of reduction of dimensionality, correlation, and the other is feature selection. In constructing the models, eight different models were used for training and prediction. These eight models were SVC, Linear SVC, KNN, Decision Tree, and Random Forest. In the model evaluation, Random Forest and Decision Tree have higher scores, while Linear SVC has the lowest score of only 58.14. The remaining four models have similar scores in the range of 80-90.

2. Methodology

2.1 Data Collection

The data for this study were obtained from the Titanic - Machine Learning from Disaster competition provided by the Kaggle website. It contains a train dataset and a test dataset. The train dataset has 891 records and 12 features. The test dataset has 418 records and 11 features. This study's explanatory variable (dependent variable) is whether the passengers survived the Titanic tragedy. Since this variable has only two possibilities, it is assigned a value of 0 for non-survival and 1 for survival. In the train dataset, the survival rate was 38.38%. The independent variable in this paper is composed of three categorical features (Sex, Embarked, and Pclass), four numerical features (Age, Fare, SibSp, Parch.), and two mixed data types of features (Ticket, Cabin.) Pclass is the passenger's class of travel, with three classes: 1 for first class (216 records), 2 for second class (184 records), and 3 for third class (491 records); Embarked is the passenger's port of embarkation: S for Southampton, UK (646 records); C for Cherbourg, France (168 records); and Q for Queenstown, Ireland (77 records). Fare is the ticket code; SibSp is the total number of siblings and spouses on board Titanic; Parch is the total number of parents and children on board Titanic.

In this study, the null rate of each feature was calculated, and there were three features with null values in the training dataset, they were Cabin (77.10%), Age (19.87%), and Embarked (0.22%) from high to low. In the test dataset, there are two features with null values: Age (20.57%) and Fare (0.24%), from high to low.

2.2 Data Preparation

The data collected is raw data, which inevitably contains many null values. The current models and statistical methods need complete, no-vacancy data. Therefore, the data needs to be processed before it can be manipulated further. Several reasons usually cause missing values: e.g., participants abstaining from filling in information during statistical surveys, objective equipment failures causing missing collected data... [1] Researchers usually use the following methods for missing values:

- Ignoring the sample.
- Filling in the vacant values manually.
- Filling in with a global constant.
- Filling in with the mean of the attributes.
- Filling in with the mean of all samples in the same category as the given sample.
- Filling in with the plural.

In this study, Cabin has a missing value of 77.1%, resulting in many missing records if deleted directly. Since the sample is too small to guess based on other characteristics, it is supplemented by manually filling in the vacant values and introducing a new value 'N' to fill in.' N' stands for not known. And then Age should be filled after filling Cabin. For filling Age, the correlation between other features and Age is calculated, and the highest correlations are selected and calculated based on these features. Their correlations are as follows: Passenger Id: 0.037, Pclass: -0.37, SibSp: -0.31, Parch: -0.19, Fare: 0.096. Since Pclass and SibSp have the highest correlations, these two features are used to classify the Age and calculate the average value. Since Sibsp has seven different values in the training dataset, the sample distribution is calculated for the seven different values. Each value and distribution are: 0: 608 records, 1: 209 records, 2: 28 records, 3: 16 records, 4: 18 records, 5: 5 records, 8: 7 records. Because the sample size is small when Sibsp is larger than two and to avoid overfitting, only the cases with Sibsp of 0, 1, and 2 are considered in this study. The Pclass has three unique values they are 1, 2, and 3. The training data are grouped into ten classes according to Pclass and Sibsp, as shown in Table 1. The mean of Age in each class is calculated separately, and the total training data mean of Age is used when Sibsp is greater than 2.

Since Embark is a categorical feature, the mean cannot be used, so for missing values, this study uses the plural, which is the type with the most occurrences in the sample, for filling. In the null

detection of the test data, it was found that Fare was missing a value, and the global mean was used to fill the null record.

Table 1. Age guess array by Sibsp and Pclass

Pclass	Sibsp = 0	Sibsp = 1	Sibsp = 2	Sibsp > 3
Pclass = 1	39	37	37	30
Pclass = 2	31	27	19	
Pclass = 3	27	24	18	

2.3 Feature Engineering

2.3.1 New Numerical Variable Creation

For machine learning, sufficient informative features are very important, and the lack of features can cause the model to fail to perform the final task [2]. Therefore, in feature engineering, enough new features need to be added. The number of relatives on the Titanic might affect the survival rate of individuals. This is because if there are relatives with someone, he/she may know about the sinking earlier and have easier access to survival resources, which leads to an increased survival rate. So, at first the total number of relatives = SibSp + Parch. Some passengers are no relatives on board, with the value of total relatives being 0, resulting in a much lower correlation with survival. After changing the total number of relatives = SibSp + Parch + 1, where one was the passenger himself, the model accuracy improved. However, too many relatives will waste survival time by searching for each other, so to better determine the survival rate of a person who is alone on board and who has relatives on board, a new feature: Alone, is introduced. One person alone on board: Alone = 1, with relatives on board: Alone = 0. As well, due to the large gap in the distribution of Fare, Log (Fare) is introduced. There might be products among currently known numerical features that increase the correlation, the mutual product of three numerical features (Fare, Age, total_relative) are introduced.

2.3.2 Classification of Data Feature Extraction

Unlike numeric data, this study uses bag-of-words word counting for text data, which requires extracting special words from the text and counting them [2]. Because the more meaningful ones in Name are their Title, with the person's name separated by '.' separating them. So, all the words before '.' of all Name feature variables are counted [3]. This research paper refined the classification when looking at the isolated Titles and categorized them into seven categories: Mr., Mrs., Miss, Master, Others, Officer, Nobility. 'Others' includes 'Rev,' 'Jonkheer,' 'Dona'; 'Miss' includes 'Mlle,' 'Ms.,' and 'Miss'; 'Mme,' 'Mrs.' are categorized as 'Mrs.' 'Dr', 'Master' are categorized as 'Master'; 'Capt', 'Col', 'Don', 'Major' are categorized as ‘officer'; 'Lady', 'Countess’, and 'Sir' are categorized as 'mobility'. Just as mentioned in the previous data cleaning, Cabin Class is extracted the initials of Cabin, and there is nine Cabin Classes: A, B, C, D, E, F, G, N, T.

2.3.3 Encoding Categorical Fields

Since most machine learning models rely on numbers and categorical features are mostly string values, people use Encode Categorical Feature to convert the content of strings into integer format to facilitate machine learning and modeling.

2.3.3.1 One Hot Encoding

One-hot encoding is a separate column for each unique value in the variable. Each record contains a new column of 1 and the other columns are 0 [4][5]. Onehot performs well in most cases, but since the number of columns created is related to the number of unique values of the feature, if the feature has more unique values, too many columns will be created, leading to memory problems [6]. First using one hot to encode the four categorical features 'Embarked', 'Cabin_Class', 'Title', and 'Pclass'. ‘Embarked’ has 3 unique values; ‘Cabin_Class’ has 9 unique values; ‘Title’ has 7 unique values; ‘Pclass’ has 3 unique values. So, after one-hot encoding, the train dataset has added 18 new columns.

2.3.3.2 Target Encoding

Target Encoding, aka Risk Tables is done by calculating the average of different categories of y under this feature, which is also the average of the target column. For example, if f6 is categorical feature and it has three unique values they are A, B, C.$V_A = < AVE_y > |_{f6=A} V_B = < AVE_y > |_{f6=B} V_C = < AVE_y > |_{f6=C}$

Then encode f6 as

$$f6 \rightarrow \begin{cases} V_A, when\ f6 = A \\ V_B, when\ f6 = B \\ V_C, when\ f6 = C \end{cases} \quad (1)$$

Compared to one hot encoding, it does not need to add extra rows and direct encoding to what researchers are trying to predict [7]. However, to avoid overfitting, smoothing is required. The smoothing formula follow:

$$Y = Y_{low} - \frac{Y_{high} - Y_{low}}{1 + e^{-(n - n_{mid})/c}} \quad (2)$$

2.4 Dimensionality Reduction

2.4.1 The Curses of Dimensionality

After Encoding categorical features, two different data frames are output, one is the Onehot data frame, and the other is the Target data frame. In the Onehot data frame, there are 39 columns; in the Target data frame, there are 23 columns. The number of features in a sample is also known as dimensionality. Feature engineering aims to create as many features as possible; putting all the features into the model training would lead to a dimensional disaster. The dimensional catastrophe is when the researchers train the samples, the model effect decreases as the number of dimensions increases, while the number of samples remains the same. Not all features in all data collected by researchers will help in model training, but some features will reduce the model's accuracy. The impact of these poorer features grows as the variables increase [8]. So, to avoid the growth of latitude, some unhelpful features should be screened before modeling. In this paper, there are two types of screening, one for human screening by eradicating the correlation between each feature and the dependent variable and the other for machine screening using SelectKBest.

2.4.2 Correlation

The Pearson correlation coefficient (PCC) was used to calculate the correlation between each column and Survived in the two datasets, and the data with low correlation were excluded based on the correlation and the number of variables. Because it is essentially a normalization of the number of covariances, the PCC values always fall from 1 to -1, and only their linear correlations are calculated [9].

PCC(r) is given as follow:

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2].[n(\sum y^2) - (\sum y)^2]}} \quad (3)$$

2.4.3 SelectKBest

SelectKBest is a way to calculate feature scores and then select K features in descending order of scores [10]. This research paper uses Chi-squared stats of non-negative features for classification tasks to calculate the feature scores. Chi-squared is a statistical test used to examine the differences between categorical variables in a random sample to determine the goodness of fit between the expected and observed results [11]. The Chi-squared formula is as follow:

$$x_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

Where C represents degrees of freedom; O represents observed values; E represents expected values.

For the One hot data frame, the parameter k of SelectKBest is set to 20. Since the number of variables of target encoding is less than 20, k is 15 for the target data frame.

2.5 Modeling

2.5.1 Logistic Regression

Logistic regression is a way to solve dichotomous models. It is generated by combining linear regression and the Sigmoid function.

Linear Model formular is presented as below:

$$z = b_0 + b_1 x \tag{5}$$

Sigmoid formular is presented as below:

$$y = \frac{1}{1+e^{-z}} \tag{6}$$

Logistic Regression is presented as below:

$$y = \frac{1}{1+e^{b_0 - b_1 x}} \tag{7}$$

Logistic regression is theoretically supported by linear regression, and logistic regression introduces a nonlinear factor through the Sigmoid function. it can easily handle 0/1 classification problems. It determines the probability of something appearing as continuous, between 0 and 1. It then classifies something as 1 if it is >threshold and 0 if it is <threshold. Usually, the threshold is set to 0.5 [12].

2.5.2 SVC

SVC is the SVM search to find a hyperplane such that the point on the boundary is the farthest away from the plane. That is, the larger d is the better for SVM. Support vector machine is supervised learning in which a classifier divides the data into two classes [13]. The linear SVM the plane which can be presented as the formular (8):

$$P:\ a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + b\ =\ 0 \tag{8}$$

The distance between the hyperplane and support vectors could be presented as the formular (9)

$$d = \frac{|a_1 x_1^i + a_2 x_2^i + a_3 x_3^i + \cdots + b|}{\sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots}} \tag{9}$$

Matrix ω and x could be introduced to simplify the formular (8), ωcould represent the set of $a_1\ to\ a_n$and x could represent the set of $x_1\ to\ x_n$

$$\omega = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \end{pmatrix} \tag{10}$$

and the distance could be presented as the formular (11)

$$d\ =\ \frac{|\omega^T x + b|}{\|\omega\|} \tag{11}$$

for support vector at above the hyperplane, formular (11) is transformed to formular (13) by removing the absolute value and dividing $d\|\omega\|$left side and right side.

$$d\|\omega\|\ =\ \omega^T x\ +\ b \tag{12}$$

$$\frac{\omega^T}{d\|\omega\|} x\ +\ \frac{b}{d\|\omega\|}\ =\ 1 \tag{13}$$

for support vector at below the hyperplane, formular (11) is transformed to formular (15) by removing the absolute value

$$-\mathrm{d}\|\omega\|\ =\ \omega^{\mathrm{T}}\mathrm{x}\ +\ \mathrm{b} \tag{14}$$

$$\frac{\omega^{\mathrm{T}}}{\mathrm{d}\|\omega\|}\mathrm{x}\ +\ \frac{\mathrm{b}}{\mathrm{d}\|\omega\|}\ =\ -1 \tag{15}$$

$\frac{\omega^{\mathrm{T}}}{\mathrm{d}\|\omega\|}$ and $\frac{\mathrm{b}}{\mathrm{d}\|\omega\|}$ are constant, so use $\omega_1 = \frac{\omega^{\mathrm{T}}}{\mathrm{d}\|\omega\|}$ and $\mathrm{b}_1\ =\ \frac{\mathrm{b}}{\mathrm{d}\|\omega\|}$ to replace above formular one. For every point at above the hyperplane: $\omega_1 \mathrm{x_i} + \mathrm{b}_1\ >=\ 1$ for every point at below the hyperplane

$\omega_1 x_i + b_1 <= -1$. y_i is -1 or 1, which means that point should belong to above the hyperplane or below the hyperplane. For all data sets satisfying the condition $y_i(\omega_1 x_i + b_1) >= 1$, the larger d is the better. When encountering a dataset that cannot be solved using linearity, the use of mapping the dataset to higher dimensions via kernel functions makes the samples linearly separable in the new space [14].

2.5.3 KNN

K Nearest Neighbors is a circle of k closest points with data points as the center [15]. The distance between two points is the Euclidean distance:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + \dots} \tag{16}$$

The type of the point is classified as the type to which most data in the k points belong [16].

First step is to point in x space for which people want to predict. Second step is to find the K nearest neighbors in the x space. Third step is to predict for y that is average of the nearest neighbors in the x space [17]. The advantages of this model are simple algorithm, high accuracy, insensitive to anomalies, easy to implement, no need to adjust parameters, and suitable for multiple classifications. The disadvantage is that it is computationally intensive and easy to misclassify when the samples are not balanced.

2.5.4 Decision Tree

Decision trees are predictive analytic models expressed through the data structure of a tree. In a decision tree, each internal node is a feature, and each leaf node is a representation of a class. The central idea of constructing a decision tree is that as the number of nodes in the tree increases, the entropy decreases as fast as possible. Entropy represents the degree of data confusion in this data set. The higher the entropy, the more chaotic the data set is and the lower the data purity. Entropy is usually calculated by formular (17).

$$H(x) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{17}$$

$$p_i: The\ probability\ of\ the\ event\ occurring$$

There are three ways to construct decision trees, ID3, C4.5, and Gini index. In ID 3, starting from the root node, the information gain of all possible features is calculated for the node, and the feature with the most significant information gain is selected as the next node. The information gain is the parent node's entropy minus the child node's [18]. The information gained is given by formular (18):

$$Gain(D, a) = H(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} H(D^v) \tag{18}$$

D: Data Set; a: Discrete feature; a has V possible values; H (D) is the parent node entropy; H (D^v) is the child node entropy

The Gini index can also be used as a calculation of purity, and like entropy, the smaller the value, the higher the purity. The Gini index is given by formular (19):

$$Gini(x) = 1 - \sum_{i=1}^{n} p_i^2 \tag{19}$$

Since, in order to avoid affecting the result, such as ID, each data has a specific value of features, the information gain rate can be used instead of information gain. The information gain rate is given in Eq. 20:

$$Ratio\left(\frac{D}{a}\right) = \frac{Gain(D,a)}{IV(a)} \tag{20}$$

$$IV(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} * log_2 \frac{|D^v|}{|D|} \tag{21}$$

Usually, the higher the number of possible values of attribute a (i.e., the larger V), the larger the value of IV(a) will usually be. Since decision trees are prone to overfitting, pruning is essential. Pre-pruning is done by limiting the tree's height, the minimum number of samples of the nodes, and other parameters to stop early in the process of building the tree. Post-pruning is a decision tree that starts to reduce branches after it has been constructed. According to Equation 22:

$$C_\alpha(T) = C(T) + \alpha(T_{leaf}) \tag{22}$$

$$C(T) = \sum_{t \epsilon leaf} N_t * H(t) \quad (23)$$

C(T) is the loss function, N_t *denotes the amount of data in the leaf, and H(t) denotes the entropy of the current leaf node.* T_{leaf} *denotes the number of leaf nodes. The magnitude of* α *determines the impact of the number of leaf nodes on the overall.*

A larger α value indicates that the number of leaves has a larger impact, while a smaller value indicates that the number of leaves has a small impact on the overall.

2.5.5 Random Forest

Random forest is a random selection of N data samples, and in these, N samples are put back sampled using Bootstrapping. In order to avoid special values of features, M features are randomly selected. The n trees are built in turn, and the test values are placed into the n trees to obtain the n predicted values, and the value with the number of occurrences among them is taken as the predicted value in the classification problem [12].

3. Result

3.1 Distribution of Passengers

The first is the distribution of passengers on board.2/3 of the people on board are men and 1/3 are women. The most significant number of people on board embarked from the port of Southampton, probably because Southampton served as the starting point for the voyage. The difference between the oldest and youngest people on board was about 80 years, and the most significant number of people on board was between the ages of 20 and 40, probably because young adults were more willing to venture out and had more financial support at the time. Most passengers on the boat got tickets for less than 100, however, there were a few 500+ tickets. The distribution of SibSp and Parch shows that most passengers came alone or with non-immediate family members.

3.2 Distribution of Survivors in Sex, Embarked, and Alone

In figure 1, blue bars are used to indicate survivors and red bars to indicate fatalities, looking at their distribution by gender, port of embarkation, cabin class, and whether they traveled alone, respectively. Although women boarded only 1/3 of the total number of passengers, the chart clearly shows that women survived in much higher numbers than men. In the port of embarkation distribution, Southampton had the highest number of passengers who survived, but Cherbourg had a much higher survival rate than the mortality rate, and Queenstown had the lowest number of passengers who survived, with a similar survival rate and mortality rate. For passengers who did not travel alone, the number of survivors was higher, and the survival rate was also higher than the mortality rate. This result is consistent with the previously mentioned hypothesis that passengers who do not travel alone have a relatively more remarkable ability to access the limited resources available for survival.

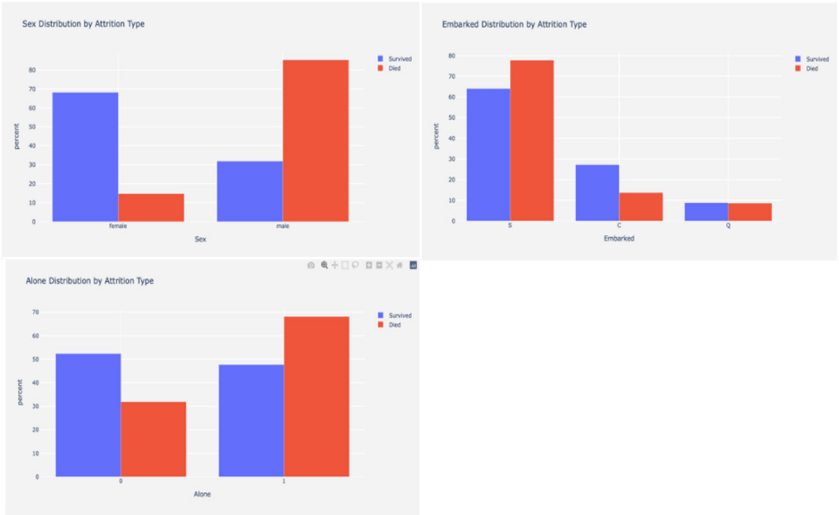


Fig 1. Distribution of survivors in Sex, Embarked, and Alone

3.3 Distribution of Survivors in Pclass

The Table 2 shows the highest fatalities in the third class, followed by the second class. Although the number of first and second-class survivors is similar, the survival rate is more reliable for a more objective comparison. The survival rate for the first class is the highest. It is about three times higher than the survival rate for the third class. Secondly, the third class has the highest number of passengers, 55% of the total number of passengers.

Table 2. Distribution of survivors in Pclass

Pclass	# of passenger	% in total	Survivial rate
1	216	24.2424%	62.96%
2	184	20.6510%	47.28%
3	491	55.1066%	24.23%

3.4 Distribution of Survivors in Age

Figure 2 demonstrates that the age span on the Titanic is extensive, from 0-80. The population is concentrated in 18-40, where the mortality rate is much higher than the survival rate until 18-32. Among children younger than ten years old, the survival rate is much higher than the mortality rate. In the age range of 32 to 38 years, the survival rate is higher than the mortality rate, and after 40 years, the ratio of survival to mortality is close. This result differs from the prediction in that the hypothesis is that children are less mobile than adults and would have difficulty surviving a disaster, but the results show that children have a higher survival rate. Instead, the survival rate was lower in the 20-30 age group.

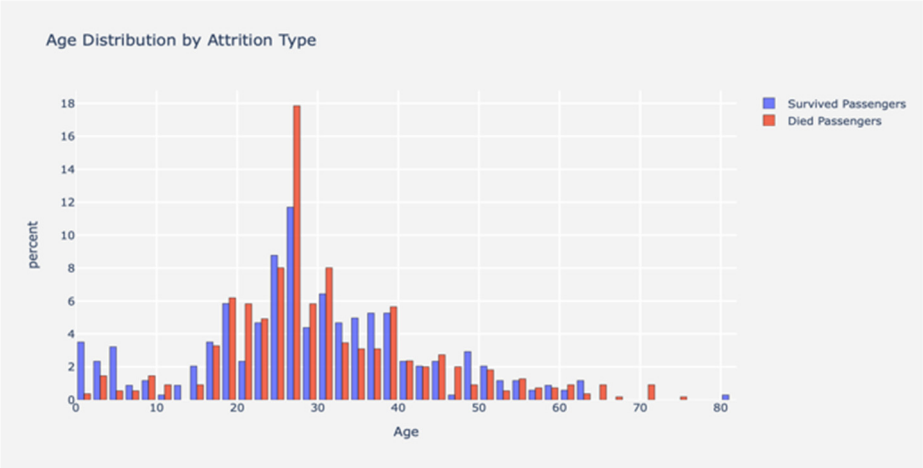


Fig 2. Distribution of survivors in Age

3.5 Distribution of Survivors in Fare

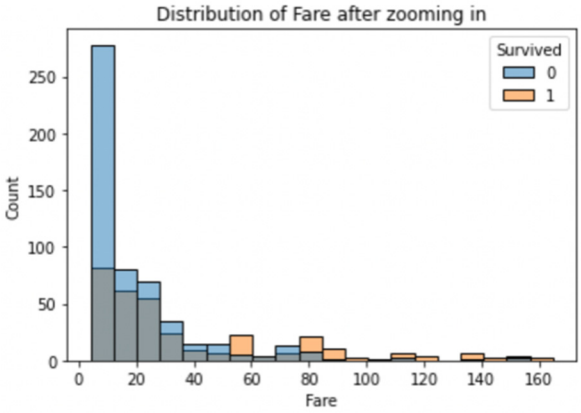


Fig 3. Distribution of survivors in Fare

Most of the Fare is concentrated in 0 - 200. To better understand the distribution of Fare, the 0 - 200 part is enlarged to produce the histogram on the right. Figure 3 clearly shows that the number of passengers with Fare in 0-10 is the highest, and the survival rate is also the lowest. In the range 10-50, the survival rate gradually increases; when Fare > 50, most survival rates are higher than the mortality rates.

3.6 Distribution of Survivors in Title

Figure 4 shows the survival rates of passengers with different Titles. The number of survivors is higher for passengers with Titles Miss and Mrs than for other passengers, and the proportion of deaths in the Mrs category is smaller than that of Miss, which can be interpreted as a higher survival rate for married women than for unmarried women. The figure also shows that the survival rate of those with Master or Doctor is higher than that of the general male population. More interesting is the absence of deaths in the Nobility cohort, however, it cannot be generalized in most cases due to the small sample size.

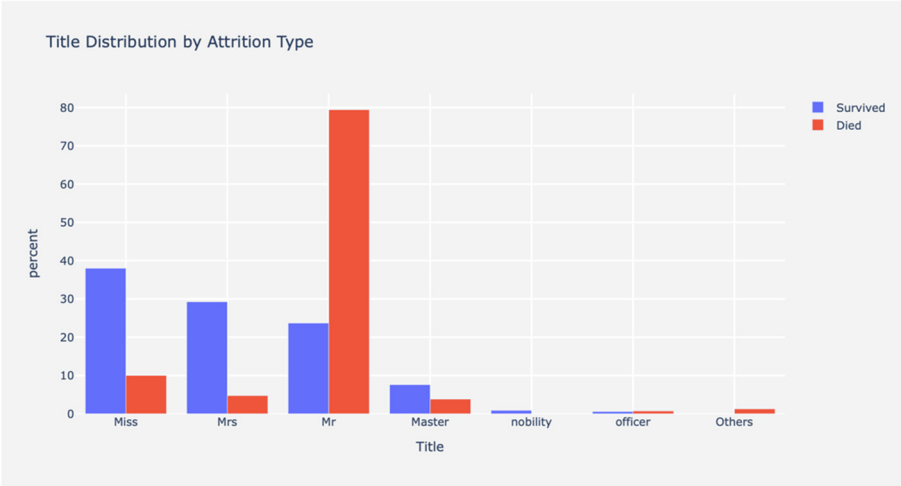


Fig 4. Distribution of survivors in Title

3.7 Distribution of Survivors in Relative Amount

Table 3. Distribution of passenger in SibSp

SibSp	Survived	# of passenger
0	0.345395	608
1	0.535885	209
2	0.464286	28
3	0.250000	16
4	0.166667	18
5	0.000000	5
8	0.000000	7

Figure 5 illustrates the total number of sibling partners, the number of parent-children, and their impact on survival. Table 3 shows the survival rates for different numbers of sibling partners, the

number of passengers, and the proportion of passengers. Since the sample size is too small when Sibsp = 3, 4, 5, and 8, only Sibsp = 0, 1, and 2 are considered in the analysis. 68% of the passengers have no siblings or spouses on board, and their survival rate is significantly higher for those with siblings or spouses on board. Table 4 indicates the distribution of passengers with different numbers of parent-children and the survival rate of passengers. The same as SibSp. Since Parch has a smaller sample greater than 2, it is also considered only for the 0, 1, and 2 cases. For Sibsp, 76% of the passengers were without parents/children on board, and their survival rate improved to 50% for passengers with parents/children on board. Also, figure 5 shows that if the number of passengers and relatives is > 4, the mortality rate is higher than the survival rate.

Table 4. Distribution of passenger in Parch

Parch	Survived	# of passenger
0	0.343658	678
1	0.550847	118
2	0.500000	80
3	0.600000	5
4	0.000000	4
5	0.200000	5
6	0.000000	1

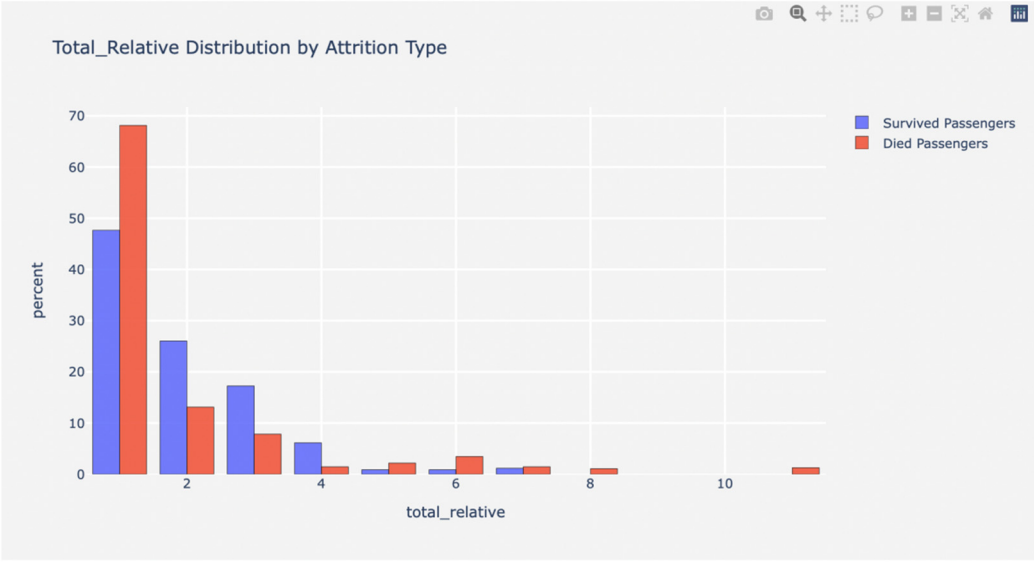


Fig 5. Distribution of survivors in total relative

3.8 Model Performance

Table 5 shows the performance of the six datasets processed differently for the six models trained. Both Random Forest and Decision Tree have the highest score of 98.65. In this study, the Random Forest model was finally used.

Onehot after Scale: the choice of Onehot Encoding when the dataset is digitized in categorical variables, using correlation when performing dimensionality reduction, manually deleting features with low relevance, and Scaling the data after processing.

Target after Scale: the choice of Target Encoding when the dataset is digitized in categorical variables, using correlation when performing dimensionality reduction, manually deleting features with low relevance, and Scaling the data after processing.

Onehot without Scale: the choice of Onehot Encoding when the dataset is digitized in categorical variables, using correlation when performing dimensionality reduction, manually deleting features with low relevance, but not scaling the data after processing.

Target without Scale: the choice of Target Encoding when the dataset is digitized in categorical variables, using correlation when performing dimensionality reduction, manually deleting features with low relevance, but not scaling the data after processing.

Onehot after Scale: the choice of Onehot Encoding when the dataset is digitized in categorical variables, using SelectKBest when performing dimensionality reduction, manually deleting features with low relevance, and Scaling the data after processing.

Target after Scale: the choice of Target Encoding when the dataset is digitized in categorical variables, using SelectKBest when performing dimensionality reduction, manually deleting features with low relevance, and Scaling the data after processing.

Table 5. Model performance

Model	One hot after Scale	Target after Scale	One hot without Scale	Target without Scale	One hot with selectKBest	Target with selectKBest
SVC	85.86	84.51	66.67	66.55	85.19	84.62
KNN	88.33	88.33	81.48	83.05	88.10	88.10
Logistic Regression	82.04	82.60	77.78	67.45	83.28	98.65
Random Forest	98.65	98.65	98.65	98.65	98.65	98.65
Decision Tree	98.65	98.65	98.65	98.65	98.65	98.65

4. Discussion

On the Titanic, only 1/3 of the total number of women boarded the ship, but 2/3 of the final survivors were women, with the highest survival rate for married women. Firstly, the reason for the higher survival rate of women is that in the early 20th century, it was believed that men should have the responsibility to protect women, so when faced with limited resources for survival, men voluntarily gave up their resources to women. Second, the higher survival rate of married women than unmarried women may be that married women who are mothers have a higher survival rate because they must take care of their children in the life raft. Among the survivors aged less than 10, the probability that their Parch is greater than 0 is 97.37%, and since the young Parch defaults to the number of their parents, it can be shown that the children were basically accompanied by their parents on the Titanic, while 40% of the married women who survived had a Parch greater than 0. The survival rate of children aged less than 10 is as high as 61%. Probably this is because children are vulnerable, and the hope of the future, and because 97.37% of them were accompanied by their parents, who would make more efforts to keep their children alive.

Moreover, at the beginning of the hypothesis, only the child's strength is considered, without considering factors such as people's morality and parents' love. This study showed that once the fare was greater than 50, the survival rate was higher than the mortality rate because the high fare corresponded to first class. the higher survival rate of the first-class passengers may be because the first-class cabin is in the uppermost part of the whole ship, and they have more time to escape. Passengers who were accompanied on the boat also had a higher survival rate because it may have been easier for more people to secure survival resources and learn about the need to leave in a life raft earlier than one person. However, once the number of relatives exceeds four, they may miss the opportunity to leave because they are looking for each other. by looking at the Title, it appears that

the group with nobility has a 100% survival rate, but the sample size is too small to determine if the theory holds.

5. Conclusion

In this study, it was found that in the Titanic tragedy, people who were female, younger than ten years old, traveling in first class or from wealthier families, with a Master or Doctor degree, and accompanied by 1-2 family members had a higher survival rate. Thus, it was found that men were willing to take up the obligation to protect the vulnerable in the early 20th century, and they were willing to sacrifice themselves to leave the hope of survival to women and children. Moreover, the rich or those with higher social status would have easier access to survival resources than the general population because they had more leverage in their hands or higher social value. It also reflects that family members will be more united in case of disasters. During model training, combining the Random Forest model and manual filtering of features based on correlation resulted in the best prediction performance. This study can better predict the survival of people who were lost in the early 20th century due to disasters. This study is beneficial for future analysis of 20th-century society and culture and disaster prediction. However, the sample size of this study is insufficient and should be expanded. Moreover, the study has limitations in using models, and new training models can be added in the future to tune further the existing models, which will lead to a further increase in prediction rate.

References

[1] Osborne, J. W. (2013). Six: dealing with missing or incomplete data: debunking the myth of emptiness. In Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data (pp. 105-138). SAGE Publications, Inc., https:// dx.doi.org/ 10.4135/ 97814522 69948.n6.

[2] Zheng, &Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists (First edition.). O'Reilly.

[3] Sherlock, Muniswamaiah, M., Clarke, L., &Cicoria, S. (2018). Classification of Titanic Passenger Data and Chances of Surviving the Disaster.

[4] F. Farahnakian and J. Heikkonen, "A deep auto-encoder based approach for intrusion detection system," 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 178-183, Doi: 10.23919/ICACT.2018.8323688.

[5] Rodríguez, Bautista, M. A., Gonzàlez, J., &Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing, 75, 21–31. https://doi.org/ 10.1016/ j. imavis.2018.04.004.

[6] Jeff Hale (Towards Data Science). 2018. Smarter Ways to Encode Categorical Data for Machine Learning. https:// towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of -3-6dca2f71b159.

[7] Nazyrova, Chaussalet, T. J., &Chahed, S. (2022). Machine Learning Models for Predicting 30-Day Readmission of Elderly Patients Using Custom Target Encoding Approach. In Computational Science – ICCS 2022 (pp. 122–136). Springer International Publishing. https://doi.org/10.1007/978-3-031-08757-8_12.

[8] Altman, N., &Krzywinski, M. (2018). The curse(s) of dimensionality. Nature Methods, 15(6), 399-400. Doi: https: // doi-org.mutex.gmu.edu/10.1038/s41592-018-0019-x.

[9] Sedgwick. (2012). Pearson's correlation coefficient. BMJ (Online), 345(jul04 1), e4483–e4483. https:// doi.org/ 10.1136/bmj. e4483.

[10] Pedregosa et al., 2011. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/ stable/ index. html.

[11] Adam Hayes. 2022. Chi-Square ($\chi 2$) Statistic.https://www.investopedia.com/terms/c/chi-square-statistic. asp.

[12] Couronné, Probst, P., &Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(1), 270–270. https://doi.org/10.1186/s12859-018-2264-5.

[13] Vapnik. (2013). Nature of Statistical Learning Theory. Springer.

[14] Kanevski. (2009). Machine Learning for Spatial Environmental Data: Theory, Applications, and Software. EPFL Press. https://doi.org/10.1201/9781439808085.

[15] Xing, & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. IEEE Access, 8, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754.

[16] Yigit. (2015). ABC-based distance-weighted kNN algorithm. Journal of Experimental & Theoretical Artificial Intelligence, 27(2), 189–198. https://doi.org/10.1080/0952813X.2014.924585.

[17] Stephen Coggeshall (July 30,2022). Session I-3. Machine Learning for Business Applications, Online, California, United States.

[18] Huang Ming, Niu Wenying, & Liang Xu. (2009). An improved Decision Tree classification algorithm based on ID3 and the application in score analysis. 2009 Chinese Control and Decision Conference, 1876–1879. https://doi.org/10.1109/CCDC.2009.5192865.