Titanic Survival Prediction: Machine Learning and Data Processing Based on Statistical Principles

Qianyou (Klay) Li [1, *], Jiahao (Joe) Song [2, a], Minghao (Terrence) Yu [3, b]

[1] D'Overbroeck's Oxford, Oxford, United Kingdom

[2] Shanghai Starriver Bilingual School, Shanghai, China

[3] Yew Chung International School of Beijing, Beijing, China

* Corresponding author: Qianyou (Klay) Li (Email: kli966124@gmail.com),

[a] joesong0612@163.com, [b] terrenceyu610@icloud.com

Abstract. On April 15, 1912, the RMS Titanic sank during its maiden voyage from Southampton to New York after colliding with a massive iceberg, resulting in the loss of 1,502 lives out of 2,224 people on board. This essay aimed to investigate potential factors that influenced the survival rate of passengers on the Titanic and introduced a statistical model incorporating logistic regression, decision trees, random forests, support vector machines, Cox regression, and hard voting to predict the survival probability of a randomly selected group of people in this incident. The modeling process involved imputing missing values, visualizing data, creating new variables through feature engineering, and fitting the aforementioned models to the dataset to achieve precise analysis and prediction. Overall, the model presented an accuracy of 88.20%. By blending classification, survival analysis, and ensemble learning, the model could be applied in several real-world fields, including healthcare, insurance, and automotive safety, by providing survival rate or life expectancy analysis and prediction based on various factors. However, there were certain limitations and substantial potential for improvement in terms of robustness and stability—particularly through the use of cross-validation to enhance generalizability and SHAP values to improve model interpretability in real-world applications.

Keywords: Titanic; Survival Rate; Prediction; Statistical Model.

1. Introduction

On April 15, 1912, the RMS Titanic, widely regarded as the epitome of early 20th-century maritime engineering and luxury, sank during its maiden voyage from Southampton to New York after colliding with a massive iceberg. According to relevant historical records, there were 2,224 passengers and crew from various regions and social classes onboard, among whom 1,502 lost their lives in the accident [1]. The sinking of the Titanic remained one of the most significant maritime tragedies in peacetime history, continuing to spark debate and inspire research aimed at further analyzing the existing data to gain deeper insights.

This essay drew upon previous research findings and constructed a predictive model based on the dataset provided. The model employed hard voting, logistic regression, decision trees, random forest, support vector machines, and Cox regression to estimate the probability of survival for a randomly selected group of individuals involved in the tragedy. Furthermore, it aimed to investigate how multiple factors influenced the likelihood of passenger survival aboard the Titanic.

The remaining sections of the essay were outlined as follows. Section 2 addressed the imputation of missing values and the creation of new variables as more interpretable indicators through feature engineering, which facilitated more accurate prediction. Data visualization was also conducted to identify relationships between individual factors and survival rates. Section 3 presented the assumptions underlying the models and provided a brief overview of the methodology adopted in the research. Section 4 encompassed data filtering, model construction and optimization, and the presentation of the final results.

2. Methodology

As shown in figure 1, the foundational concept of our method is to build a hard voting system involving logistic regression model, decision tree model, random forest, support vector machine, and cox regression model. We will first input training data set to each model and allow them to carry out individual prediction of the survival rate for a particular passenger. Each sub-model will have a different algorithm to make predictions. After this process is done, it will run through a hard voting process and eventually output the final prediction.

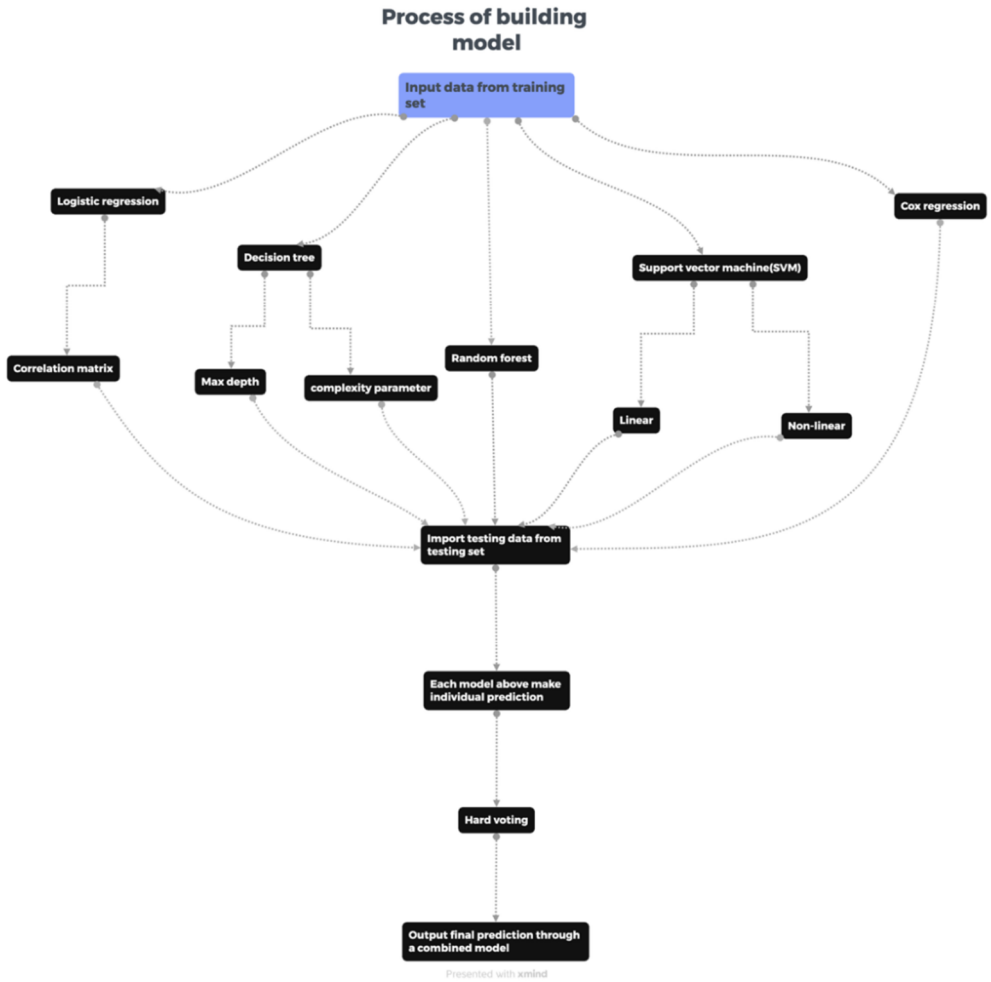


Figure 1. Processing of building model

2.1 Hypothesis

Assumption 1: Each data point is independent from each other.

Independence means that there is no hidden relationship and patterns between each variable that could skew the analysis or lead to over-estimation; it also prevents the model from memorizing instead of generalizing during predictions. We conducted a chi-square test to see whether the variables in the dataset are related. Some of the p-values are smaller than 0.05, which indicates that there are associations between the corresponding variables. This is potentially because those with high p-values are indicating similar information. For example, Pclass and Fare both represent the social status of passengers.

Assumption 2: There is no multicollinearity among explanatory variables.

This is particularly essential for logistic regression and Cox regression. Logistic regression estimates the effect of each predictor on the log odds of the outcome, while cox regression investigates how predictors affect the hazard rate over time. When variables are highly correlated, the model might struggle to distinguish the individual effects clearly, which could lead to misleading and unreliable

results. In contrast, tree-based models such as decision trees and random forest naturally handle correlated features by splitting on one and ignoring the rest, so they are not significantly influenced by multicollinearity.

Assumption 3: Explanatory variables and log odds of response variables have linear relationships. (Typically in logistic and cox regression)

Assumption 4: Sample size of the data is sufficient for models including logistic & cox regression and random forest to stabilize the results.

2.2 Logistic Regression

Logistic regression is a statistical model used for binary classification to make predictions and output probabilities between 0 and 1 based on the relationship between two or more input variables. In this case we apply the model first to the "train" dataset and make predictions of the survival rate of passengers in the "test" dataset based on the patterns learned before.

2.3 Decision Tree

Decision tree is a flowchart-like machine learning model that works by splitting data into branches based on feature values and ultimately provides results at the leaf nodes. We used decision trees for our prediction by choosing different variables as separate criteria and placing them in the internal nodes of the flowchart. Classification tree rather than regression tree was chosen, as the dependent variable is categorical.

2.4 Random Forest

Random forest is an ensemble machine learning model that builds multiple decision trees and combines their output. Each tree is trained on a random subset of the data and features, and the final prediction is made by majority voting, that is, retaining the class selected by most trees. Compared to decision trees, random forest offers higher accuracy & stability and reduces overfitting via averaging despite lower interpretability due to many trees.

2.5 Support Vector Machine

Support Vector Machine (SVM) is a type of algorithm that is commonly used in classification and regression tasks. This model seeks to find a boundary known as "hyperplane" that best divides two classes from each other, and new inputs are classified based on their position in comparison with the hyperplane [2].

2.6 Cox Regression

Cox regression is a type of model used in survival analysis. This model predicts the hazard ratio (HR) that is linked with risk factors and the passage of time [3]. We implemented this model in our research to predict the survival possibilities of passengers over time (Figure) and the HR of the passengers.

2.7 Hard Voting

For hard voting, each individual model will make their prediction, which counts as one "vote". In this case, each model takes in certain variables from testing data and predict whether a particular passenger will die or not (0 for died, 1 for survived). If the number of 1s is larger than the number of 0s, the passenger will eventually be counted as survived. Therefore, hard voting combines prediction and returns the mode, which provides a more comprehensive and precise prediction.

3. Data Processing

3.1 Data Sources

Three datasets were extracted from Kaggle [4] as the original source for this analysis: train, test, and gender submission. The train and test datasets provided personal information for a total of 1,309 passengers aboard the Titanic, with survival status available for 891 individuals and unknown for the remaining 418. The gender submission dataset presented a baseline prediction model that assumed all and only female passengers survived. Given the limited sample size, the train and test datasets were merged to maximize the available information. The predictions generated by the constructed model were subsequently compared to those in the gender submission dataset to evaluate relative accuracy.

3.2 Variables

There were originally 12 variables in the dataset given: *PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp, Parch, Ticket, Fare, Cabin, Embarked*. Meanwhile, there were also some missing values in several variables. 1 missing in *Fare*, 2 in *Embarked*, 263 in *Age* and 1014 in *Cabin*. In the next section, the methods we used to clean and organize the data will be described in detail.

3.3 Missing Value Treatment and Imputation Strategy

3.3.1 Fare

As mentioned above, there was only one missing value in the category 'Fare'. Firstly, we checked the personal information of this individual to see if there are any noticeable features (shown in Table1).

Table 1. Information of the passenger with ticket fare data missing

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1044	NA	3	Storey, Mr. Thomas	male	60.5	0	0	3701	NA	NA	S

As shown in Table 1, the individual was a male from the third class and had embarked from Southampton port. We attempted to predict his ticket fare by looking at the distribution of third-class passengers embarked from Southampton port, as passenger class is closely linked to ticket fare, and the embarkation indicated the social class and economic level of passengers to some extent.

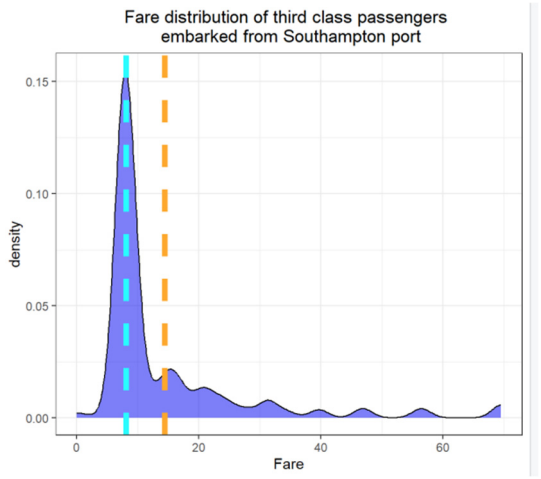


Figure 2. Fare distribution of third-class passengers embarked from Southampton port

In Figure 2, the cyan and orange lines represent the median and mean fare respectively. We chose to impute the missing value by the median of all fares, as the proportion of passengers with fare around median is very high (also since there are some extremely high values that drive the average upward) [4].

3.3.2 Embarked

In the previous section, we discussed that *'Fare'*, *'Pclass'* and *'Embarked'* are closely related, so it is reasonable to take advantage of this relationship again to help us impute the missing values of *'Embarked'*.

Table 2. Information of the passenger with embarked data missing

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
62	1	1	Icard, Miss. Amelie	female	38	0	0	113572	80	B28	NA
830	1	1	Stone, Mrs. George Nelson (Martha Evelyn)	female	62	0	0	113572	80	B28	NA

From Table 2 we observed that both the two ladies with missing *'Embarked'* data were from the first class, while they also had the same ticket fare of £80. This information helped us determine the underlying embarkation in the following boxplot, which categorizes the population by passenger class and ticket fare.

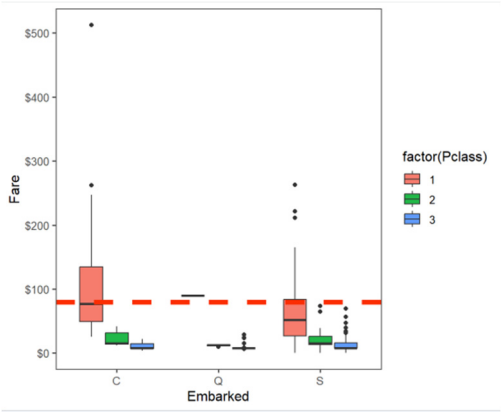


Figure 3. Relationship between embarked and fare grouped by passenger class

In Figure 3, the median fare for a first-class passenger departing from Cherbourg ('C') coincides nicely with the $80 paid by the two embarkment-deficient passengers, which provides enough evidence for us to replace the NA values with 'C' [5].

However, as we checked out the correct value on the official website [4], we found out that the two individuals with missing embark data actually went on board in 'S' (Southampton) (Table 3).

Our wrong prediction is likely to stem from the fact that fare and Pclass do not uniquely identify the embarkation point, namely, the correlation between either fare or Pclass and Embarked is not strong enough; The situation could vary if other variables tell a different story.

Table 3. The actual data of the two passengers with boarded showing 'S'

Surname	First Names	Age	Boarded	Survivors(S) or Victim (†)
Icard	Miss. Rose Amelie (Maid to Mrs. George Nelson Stone	39	Southampton	S
Stone	Mrs. Martha Evelyn	62	Southampton	S

Alternatively, we made use of the connection between ticket numbers and Embarkation. After converting ticket numbers into factors (which will automatically put similar format into close groups), we plotted a bar graph (Figure 4) to see whether there is a close relationship between the two variables. Afterwards, we look into the focus range (Figure 5) and based on the distribution in the graph, we concluded that the missing value is more likely to be S (3). However, we believed that this method

has its own limitations as well. It seemed like a correct approach just because its outcome is consistent with the actual data.

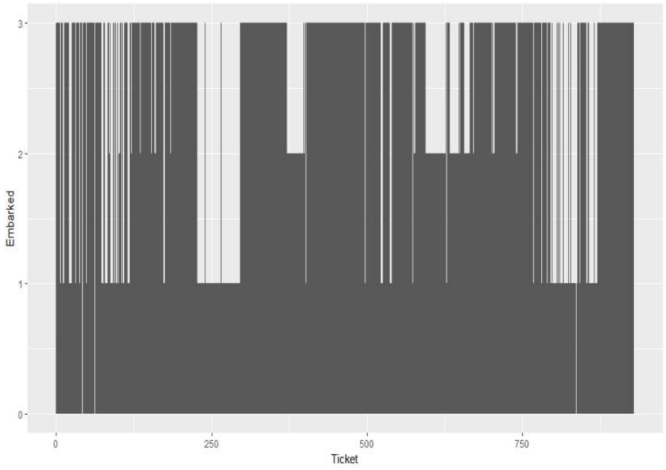


Figure 4. Bar graph of relationship between embarked and ticket number

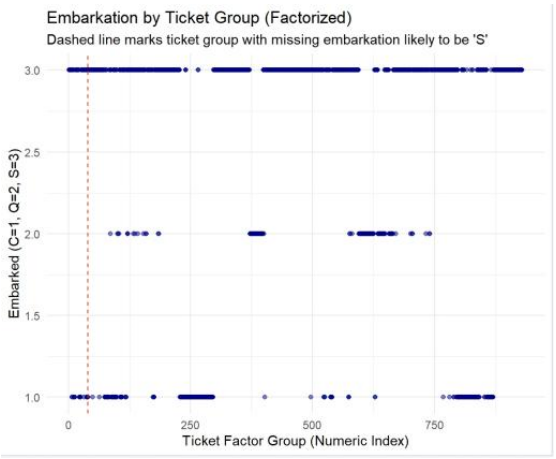


Figure 5. Distribution of embarkation by ticket group

3.3.3 Age

We decided to create a model to predict ages based on other variables in the dataset by applying a package called "mice" (*Multivariate Imputation by Chained Equations*) in RStudio. We then compared the results it generates with the original distribution of passenger age to examine its accuracy. Figure 6 below showcases a similar pattern of our prediction and the original data, indicating that the output from the mice model fits the existing distribution very well [5].

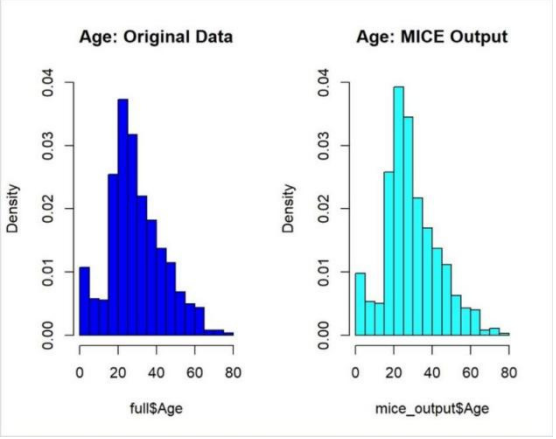


Figure 6. Comparison of age distribution before and after applying the model

3.3.4 Cabin

We revisited the table of missing values referenced at the beginning of the section and observed that the variable Cabin contained 1,014 missing entries. The data exhibited a semi-structured format of <Letter+ Number>, which posed considerable challenges for imputation. In theory, imputing missing Cabin values was highly problematic due to the lack of sufficient ground truth, which limited the capacity of statistical models to learn reliable patterns. Furthermore, the distribution of Cabin data was uneven across passenger classes: first-class passengers were more likely to have recorded cabin information, whereas third-class passengers were rarely assigned cabin identifiers.

Despite the high degree of uncertainty and the potential for bias in any predictive attempt, we proceeded with an imputation strategy. This decision was informed by the strong association between Cabin and several other variables, as well as the availability of supplementary information from external sources. According to the structural layout of the Titanic, Decks A through G represented distinct vertical levels. The location of a passenger's cabin was presumed to influence their likelihood of survival, given that individuals situated on upper decks would have had greater access to lifeboats located on the ship's topmost level.

To obtain more precise information, we consulted an academic research about Titanic's structural layout [6], which indicated that Decks A–C (sometimes T) were predominantly occupied by first-class passengers, Decks D–E by second-class passengers, and Decks F–G by third-class passengers. Although detailed cabin-passenger mappings were available online, the process of importing and aligning these records with our existing dataset proved to be both complex and redundant. Consequently, we opted to simplify the Cabin variable by extracting the deck letter and discarding the numerical component, thereby creating a new categorical variable named CabinLetter.

With respect to the missing values, we reiterated that deck assignment was strongly correlated with Pclass. Additional indicators—such as Fare, Ticket, and other contextual variables—also provided clues regarding the probable deck on which a passenger may have been located. These factors were explored in subsequent sections.

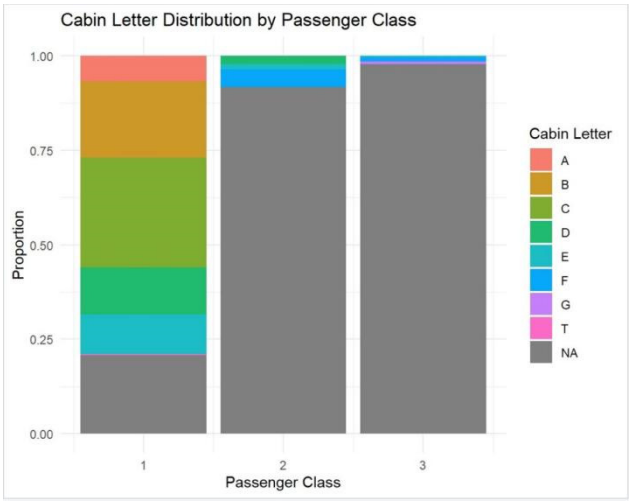


Figure 7. Distribution of Cabin letter grouped by passenger class

We plotted a bar graph (Figure 7) to illustrate the proportion of each deck level across the three passenger classes. It revealed that first-class passengers mainly resided on Decks A–E, second-class passengers on Decks D–F, and third-class passengers on Decks E–G. Subsequently, we conducted a chi-square test to evaluate whether Pclass and CabinLetter were statistically independent.

Null hypothesis (H_0): The variables are independent (Pclass doesn't influence cabin assignment)
Alternative hypothesis (H_1): The variables are related (Pclass does influence cabin location.)

The chi-square test yielded a p-value of 2.2e-16, which was significantly smaller than the significance level of 0.05, indicating a strong statistical association between Pclass and CabinLetter

(H_0 was rejected). We subsequently calculated the distribution of CabinLetter for each Pclass and imputed missing values by sampling from the respective distributions.

During this process, we observed that certain entries in the Cabin column contained multiple cabin identifiers. For instance, Miss Mabel Helen Fortune was assigned cabins C23, C25, and C27. We also noted that she had three siblings and two parents recorded, all of whom shared the same Cabin, Fare, and Ticket number. Similarly, Miss Amelie Icard and Mrs. George Nelson Stone, who were individual passengers, shared the same Cabin despite having no familial ties, but identical Ticket, Fare, and Embarkation values. These findings suggested that passengers with common Surname, Parch, SibSp, Ticket, and Fare were likely to reside on the same deck level. This rationale was used to refine the previously imputed values.

However, the number of second- and third-class passengers with recorded cabin information was limited, introducing potential randomness and bias into the predictions. The actual number of passengers assigned to cabins F and G was 574 and 172, respectively, revealing substantial error in the initial predictions.

To improve accuracy, we experimented with two additional methods: a constraint-based cabin assignment model in R using linear programming, and a multinomial logistic regression model incorporating variables such as Embarked, Fare, Sex, and Age. Both approaches produced unsatisfactory results that failed to align with real-world conditions.

Consequently, we compromised by predicting CabinLetter within the range A–G, and opted to group certain deck levels to enhance data stability and precision. Based on the CabinLetter distribution bar graph, Pclass 3 passengers were predominantly assigned to cabins F and G, while Pclass 2 passengers were predicted to occupy cabins D, E, and F. Given that a substantial portion of the existing Cabin data pertained to Pclass 1 passengers, the cabin distribution for this group, as derived from the bar graph, was considered relatively reliable.

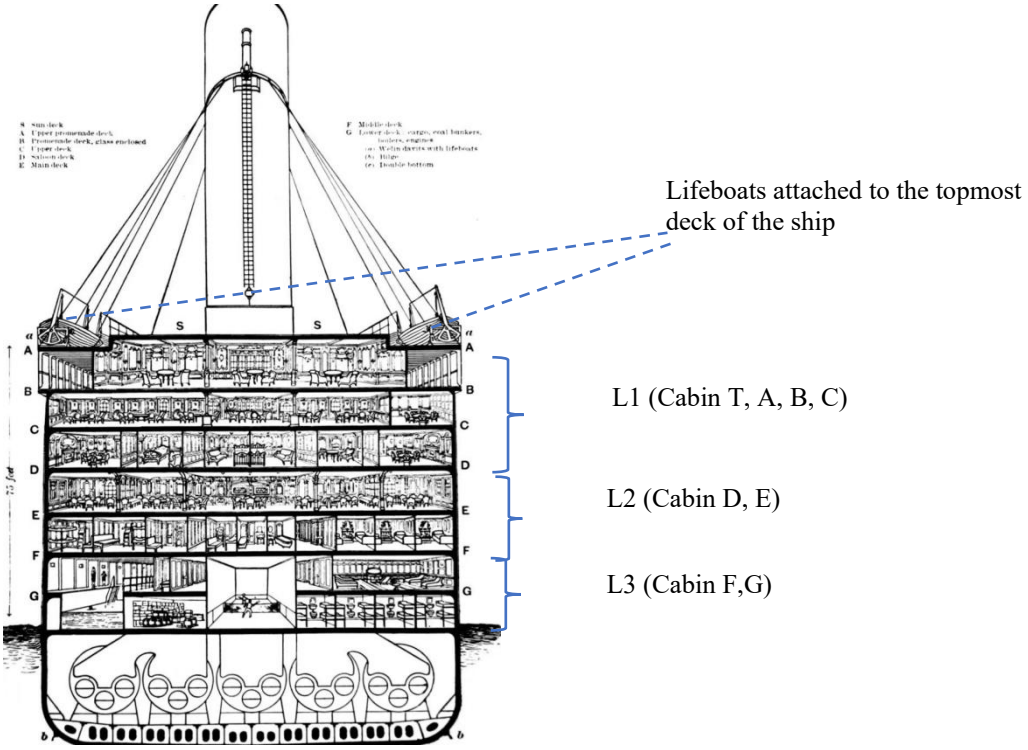


Figure 8. Titanic cutaway diagram [7]

Accordingly, we created a new variable that categorized Cabins A–G into three cabin levels: (T)A–C, D–E, and F–G, which we designated as L1, L2, and L3, respectively (Figure 8). It was evident that all Pclass 3 passengers were assigned to L3, and Pclass 1 passengers could be analyzed directly, as their cabin distribution was already known. The cabin level for Pclass 2 passengers remained a challenging issue, as it was unclear which individuals should be allocated to L3.

After testing several variables as potential determinants, we decided to assign Pclass 2 passengers with the lowest Fare values to Cabin Level L3. Given that there were 323 passengers in first class, 277 in second class, and 709 in third class, and that Cabin Level L3 contained 746 passengers, it was inferred that 37 of them originated from Pclass 2. We selected 37 individuals from Pclass 2 with the lowest Fare and reassigned their Cabin Level to L3.

This approach yielded a satisfactory result, with the final distribution being L1 = 234, L2 = 329, and L3 = 746. Overall, it represented a worthwhile attempt.

3.4 Feature Engineering and Derived Variables

3.4.1 Title

We observed that the names in the dataset comprised multiple components, including first name, surname, and title. We considered titles to be a meaningful indicator for predicting survival rate, as they served as a strong proxy for social status, age, and gender. Accordingly, we separated the title from the Name variable and created a new categorical variable, Title.

Given that some titles were repetitive and others were uncommon, we grouped all rare titles under the category Rare_title and consolidated the repetitive ones for improved classification. The final variable contained only the following categories: Mr., Miss., Mrs., Master, and Rare_title.

3.4.2 Family Size

The size of passengers' families was also considered a potential determinant of survival rates, so we created a new variable based on the number of siblings/spouses (SibSp) and children/parents (Parch). We named this variable Fsize and categorized families into three subgroups.

Singleton --- Fsize=1; Small --- Fsize = [2:4]; Large --- Fsize>4 (5 and more)

3.4.3 Underage & Adult

Considering the policy of "women and children first", a new variable called "Child" was created by separating individuals into child and non-child based on their age.

3.4.4 Mother & Not Mother

Whether a passenger is a mother or not could possibly influence a person's survival rate as well, because crew and fellow passengers may prioritize assisting mothers, viewing them as needing protection.

3.5 Exploratory Data Visualisation

3.5.1 Age, Sex & Survival

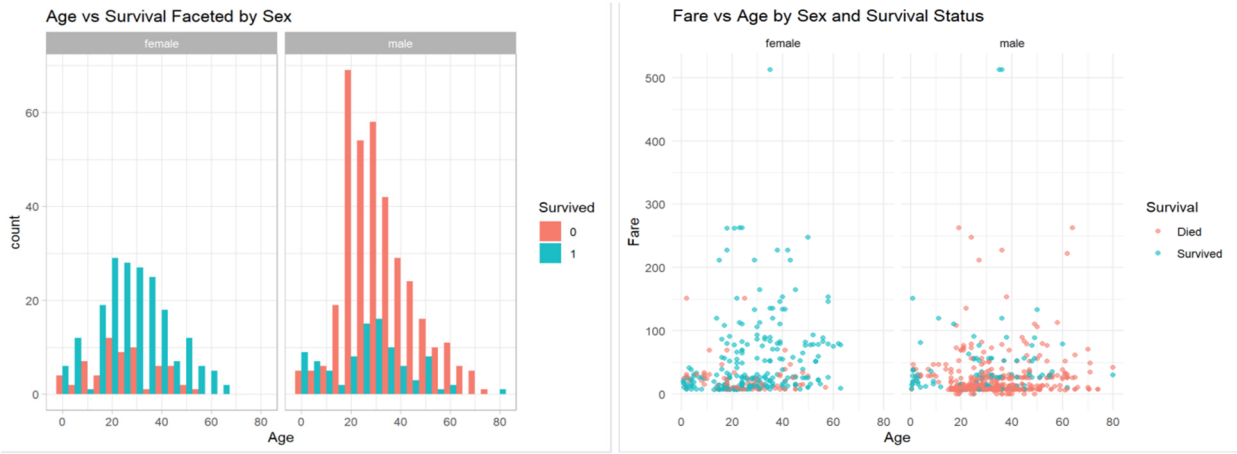


Figure 9. Survival count by age and sex; **Figure 10.** Distribution of passengers by fare, age and gender

As shown in Figure 9, the number of male fatalities in the incident was significantly higher than that of females. Additionally, passengers aged between 15 and 50 who survived the accident outnumbered those in other age groups. This trend was partly attributable to their more mature cognitive abilities compared to children, and greater physical capacity to escape relative to elderly individuals. However, another non-negligible factor was the large population base of passengers within the 15 to 50 age range.

3.5.2 Fare, Age, Sex & Survival

Figure 10 presented a scatter plot illustrating the distribution of passengers by fare, age, and gender. The graph clearly demonstrated the influence of fare on survival outcomes. It was observed that red dots (representing deceased individuals) among female passengers were predominantly concentrated at the lower end of the fare spectrum, most notably around £10. In contrast, very few male passengers survived when their ticket fare was within this range.

Despite the overall low survival rate among males—evidenced by a significantly higher number of red dots compared to cyan ones—a substantial proportion of children aged between 0 and 10 successfully evacuated. This pattern indicated that age also played a pivotal role in determining survival likelihood.

3.5.3 Family Size, Cabin Level& Survival

From Figure 11, we observed that the likelihood of survival gradually decreased as passengers' cabin level declined. This trend was attributed to the limited access to lifeboats for individuals residing on lower decks, as lifeboats were attached to the ship's topmost deck. Furthermore, the high concentration of passengers on the lower decks likely contributed to congestion during evacuation, further impeding escape.

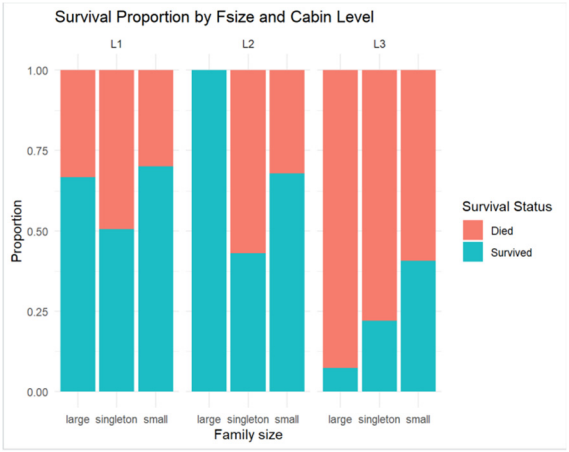


Figure 11. Survival proportion by family size and cabin level

Upon closer examination of Cabin Level L3, we found that large families exhibited a survival probability of less than 10%, single passengers around 20%, and small families above 40%. This phenomenon was likely attributable to the tendency of small families to remain together and offer mutual support, in contrast to singletons who lacked assistance and large families who may have faced difficulties securing sufficient space on lifeboats.

The observed 100% survival rate of large families in Cabin Level L2 was presumed to result from a small sample size and potential inaccuracies introduced during the imputation and feature engineering of the *Cabin* variable.

3.5.4 Pclass, Time & Survival

In addition to the preceding variables, we added the concept of *time* into the visualization, which could directly display the passengers' danger over time.

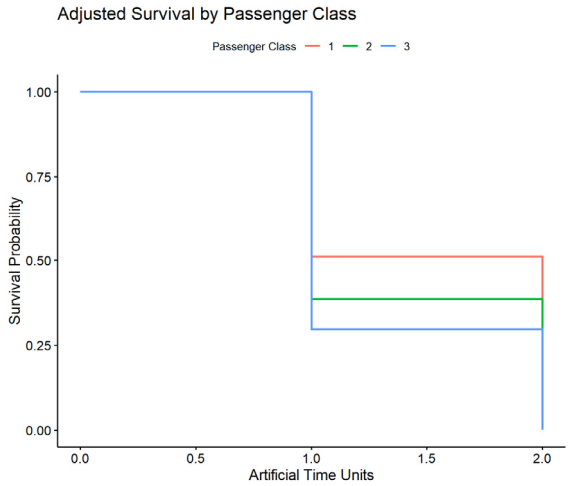


Figure 12. Survival rate by passenger class and time

As shown in Figure 12, although the survival probabilities of all passengers declined over time, those in higher passenger classes ultimately exhibited greater chances of survival compared to those in lower classes. Passengers from Class 1 had an ultimate survival probability of approximately 52%, while those from Class 2 had around 37%, and Class 3 passengers only reached about 30%.

4. Analysis of Results

4.1 Logistic Regression

To begin with, we chose the factors that we concluded to be most influential in the previous steps and implemented them into the model. The result of the model has an accuracy of 84.27%. Moreover, we noticed that the p-values between certain variables were concerning.

Table 4. P-value matrix of 5 variables

	P class	Sex	Embarked	Title	Cabin Level
Pclass	0	2.743620 e-04	7.863121e-21	1.603004e-09	1.329840e-200
Sex	2.743620e-04	0	1.419766e-02	2.342819e-151	8.532655e-05
Embarked	7.863121e-21	1.419766e-02	0	4.100311e-04	2.402933e-18
Title	1.603004e-09	2.342819e-151	4.100311e-04	0	1.015139e-10
Cabin Level	1.329840e-200	8.532655e-05	2.402933e-18	1.015139e-10	0

As highlighted in Table 4, the p-values between Cabin Level and Pclass, Sex and Title are extremely low compared to the others, which implies that these variables are strongly connected to each other. This is understandable, because the variable *Title* is correlated to the gender of the passengers, whereas *Pclass* and *CabinLevel* are equivalent for most cases. Thus, we decided to delete one from each pair to simplify our model.

Table 5. VIF of the logistic regression

	GVIF	Df	GVIF^(1/(2*Df))
Pclass	1.999863e+01	2	2.114704
Sex	5.626813e+06	1	2372.090413
Age	1.879122e+00	1	1.370811
SibSp	1.581561e+00	1	1.257601
Embarked	1.334785e+00	2	1.074862
Title	1.184543e+07	4	7.659384
CabinLevel	1.774883e+01	2	2.052544

Table 5 displays the Variance Inflation Factor and related data of the logistic model. The VIF of *Sex* is evidently the greatest, followed by *Title*. This suggests that these two factors have high multicollinearity, meaning that they are highly correlated. This also happens for *Pclass* and *CabinLevel*, for their GVIF^(1/(2×df)) is greater than 2, indicating that they have high multicollinearity. Due to having the greatest VIF, the variable *Sex* is decided to be deleted first.

Table 6. Summary of logistic regression without the variable Sex

	Estimate	Std. error	z value	P(r>|z|)
Intercept	4.086558	0.665348	6.142	8.15e-10
Age	-0.028352	0.009274	-3.057	0.00224
SibSp	-0.595756	0.132282	-4.504	6.68e-06
EmbarkedQ	-0.094286	0.416901	-0.226	0.82108
EmbarkedS	-0.506297	0.276730	-1.830	0.06731
TitleMiss	-0.367633	0.536004	-0.686	0.49279
TitleMr	-3.187535	0.578979	-5.505	3.68e-08
TitleMrs	0.240029	0.597660	0.402	0.68797
TitleRare Title	-2.192502	0.786834	-2.786	0.00533
CabinLevelL2	1.195867	0.465500	2.569	0.01020
CabinLevelL3	0.938572	0.750470	1.251	0.21106

Table 7. VIF of the logistic regression without the variable Sex

	GVIF	Df	GVIF^(1/(2*Df))
Pclass	20.024436	2	2.115388
Age	1.892119	1	1.375543
SibSp	1.575753	1	1.255290
Embarked	1.329332	2	1.073763
Title	2.421077	4	1.116866
CabinLevel	17.793655	2	2.053839

Table 8. Confusion Matrix of the logistic regression without the variable Sex

	Actual: 0	Actual: 1
Predicted: 0	99	11
Predicted: 1	17	51

With Sex removed, most of the variables' VIF returned to normal, resulting in an accuracy of 85.39%., which slightly improves the previous accuracy. Moreover, we chose to remove the variable *Pclass* from the model, so that the VIF of *CabinLevel* can return to normal.

Table 9. Summary of logistic regression without Sex and Pclass

	Estimate	Std. error	z value	P(r>|z|)
Intercept	3.224110	0.692892	4.653	3.27 e-06
Age	-0.018178	0.008641	-2.104	0.03540
SibSp	-0.609033	0.137915	-4.416	1.01e-05
Parch	-0.252453	0.154804	-1.631	0.10294
Fare	0.006524	0.003316	1.968	0.04909
TitleMiss	-0.556738	0.550468	-1.011	0.31183
TitleMr	-3.411810	0.592139	-5.762	8.32e-09
TitleMrs	-0.057471	0.600382	-0.096	0.92374
TitleRare Title	-2.436133	0.787148	-3.095	0.00197
CabinLevelL2	-0.109180	0.349184	-0.313	0.75453
CabinLevelL3	-1.661581	0.380261	-4.370	1.24e-05

Table 10. VIF of the logistic regression without Sex and Pclass

	GVIF	Df	GVIF^(1/(2*Df))
Age	1.676879	1	1.294944
SibSp	1.690754	1	1.300290
Parch	1.460056	1	1.208328
Fare	1.758736	1	1.326174
Title	2.535778	4	1.123347
CabinLevel	2.046565	2	1.196069

Table 11. Durbin-Watson Test of logistic regression without Sex and Pclass

lag	Autocorrelation	D-W Statistic	p-value
1	0.02589857	1.947721	0.466

Table 12. Confusion Matrix of the logistic regression without Sex and Pclass

	Actual: 0	Actual: 1
Predicted: 0	101	9
Predicted: 1	16	52

The resulting accuracy is improved to 85.96%, which is marginally better than the preceding versions. The following graph depicts the ROC curves of the three versions, which shows the accuracy of our model. The true positive rate is the proportion of predictions that correctly predict to be true, whereas the false positive rate is the proportion of predictions that predict to be true yet false. Our final model has the most Area Under Curve (AUC) of 0.883, which means that it has the best accuracy.

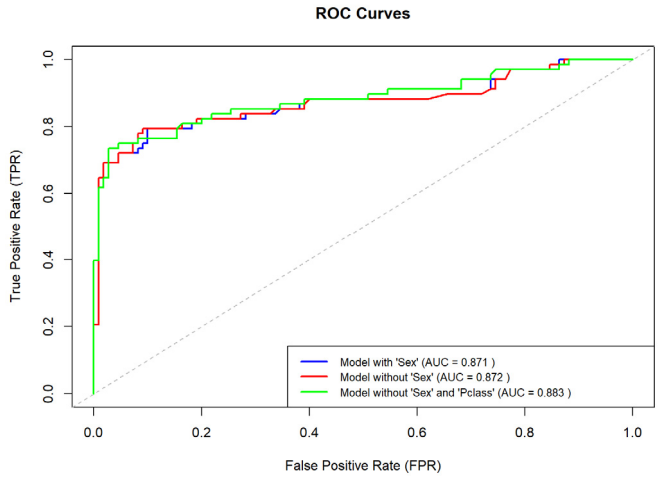


Figure 13. ROC curves of the logistic regression models

4.2 Decision Tree

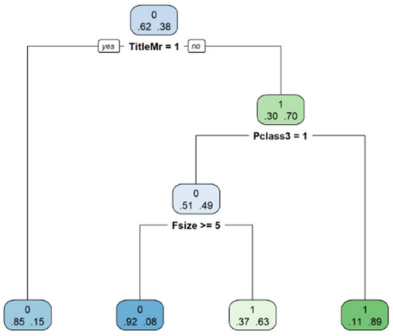
Initially, the default code is used to predict the results. The accuracy of prediction is 80.34%, which is not ideal. Thus, we altered the maximum depth of the tree and found out that at a maximum depth of 5, the model reaches the highest accuracy of 83.15%.

Table 13. Confusion Matrix of the default code

	Actual: 0	Actual: 1
Predicted: 0	94	16
Predicted: 1	19	49

Table 14. Confusion Matrix with maxdepth = 5

	Actual: 0	Actual: 1
Predicted: 0	100	10
Predicted: 1	20	48



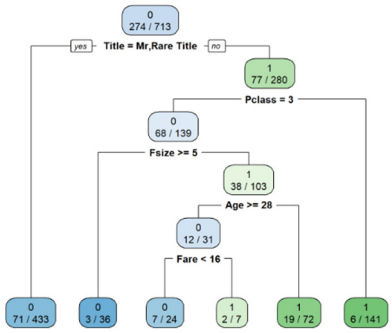


Figure 14. Default decision tree; **Figure 15.** Decision tree with maxdepth = 5

The initial tree shown in Figure 14 has a depth of 3 and 4 leaf nodes, whereas the latter one has a depth of 5 and 6 leaf nodes. This suggests that the default model has insufficient splits to make accurate predictions. Furthermore, we decided to change the complexity parameter (cp) of our model, and the best result is when cp is equal to 0.001. The model now has a depth of 9, and an accuracy of 84.62%.

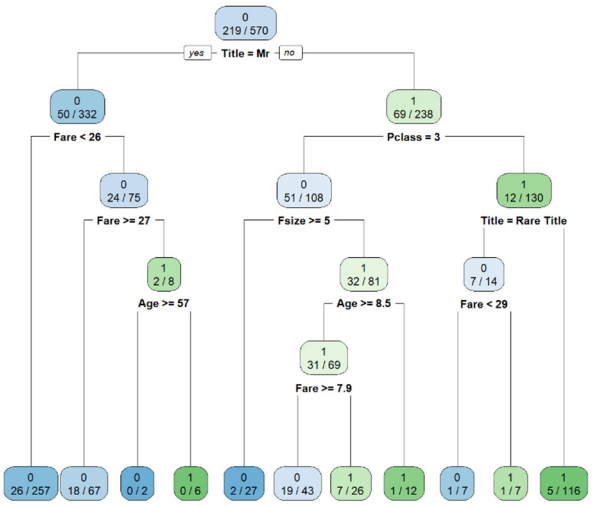


Figure 16. Decision tree with cp = 0.001

Table 15. Confusion Matrix with cp = 0.001

	Actual: 0	**Actual: 1**
Predicted: 0	247	22
Predicted: 1	19	130

4.3 Random Forest

Table 16. Confusion Matrix of the Random Forest model

	Actual: 0	**Actual: 1**
Predicted: 0	247	22
Predicted: 1	19	130

Random forest is applied aiming to mitigate the issue of overfitting in decision trees and hence improve precision and stability. We first use the default value of ntrees= 500 and ended up achieving an accuracy of 90.19%. Since the gradient of the curve tends to reach constant at 200-500 trees and

the model itself is already robust, we decided to not change the value in case of any unpredictable changes.

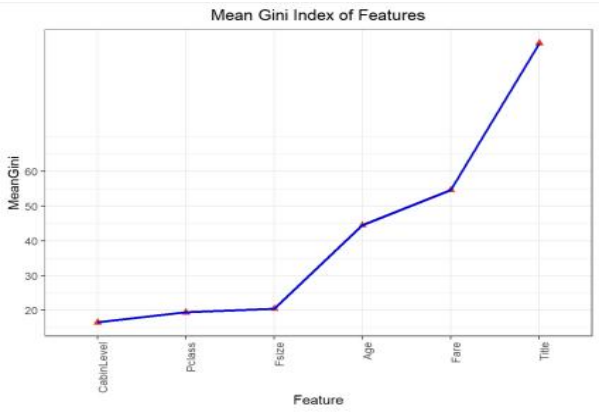


Figure 17. Mean Gini Index of Features of variables in random forest

4.4 Support Vector Machine

For this model, we mainly chose variables with lower variance, such as *Sex*, *Pclass*, *CabinLevel*, etc. However, there are two exceptions: *Age* and *Fare.*

Table 17. Variance of Age, SibSp, Parch, and Fare

Age Variance	SibSp Variance	Parch Variance	Fare Variance
217.354968	1.265124	0.604594	2291.5143

The variances of *Fare* and *Age* are enormous in comparison with *SibSp* and *Parch*, thus the method of scaling is used to transform all variances into 1. The default code has linear and non-linear support vector machine models, and the result shows that the accuracy of the linear model is 86.52%, whereas the non-linear model has an accuracy of 88.20%, as shown in the table below.

Table 18. Confusion Matrix of the linear SVM model

	Actual: 0	Actual: 1
Predicted: 0	102	8
Predicted: 1	16	52

Table 19. Confusion Matrix of the non-linear SVM model

	Actual: 0	Actual: 1
Predicted: 0	105	5
Predicted: 1	16	52

4.5 Cox Regression

As previously mentioned, the Cox Regression model is used to predict the possibility of survival and HR. Thus, we picked four main variables in the dataset: *Pclass, Sex, Age,* and *FsizeD.*

Figure 18 displays the hazard ratios for the passengers under several factors. For *Pclass*, we take Passenger Class 1 as reference, and the result shows that passengers at Passenger Class 2 have a 1.45 times higher chance of death. Furthermore, passengers at Passenger Class 3 are almost 2 times more likely to die than at Passenger Class 1. Another factor is *Sex*, where females are taken as reference, and the result shows that male passengers have 2.15 times more chance to die than females. As for *FsizeD*, single passengers are taken as reference, and interestingly, small families seem to have a lower chance of dying compared to singletons, whereas large families are 1.68 times more likely to die.

4.6 Hard Voting

In conclusion, we used hard voting to receive the best result out of our models. The accuracy of hard voting is 88.20%, which is over our expectations. Our model is highly accurate in identifying

passengers who did not survive, with the accuracy of 93.64%, while the identification of surviving passengers is a little low, scoring 76.47%. This is due to the lack of surviving passenger data, with most of the passengers not surviving.

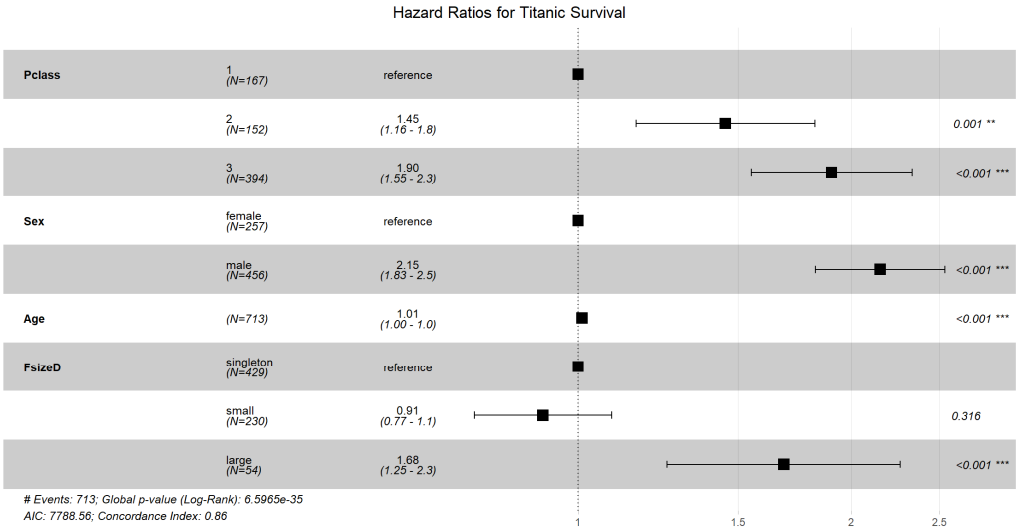


Figure 18. HR for Passengers under the factors of Pclass, Sex, Age, and Family Size

Table 20. Confusion Matrix of the final prediction

	Actual: 0	**Actual: 1**
Predicted: 0	105	5
Predicted: 1	16	52

Table 21. Summary of the final prediction

Accuracy	**Sensitivity**	**Specificity**
0.882	0.7647	0.9364

5. Conclusion and Reflection

This essay aims to make predictions of the Survival rate of Titanic passengers based on historical data of their personal information through data processing and machine learning models. In the section of data cleaning, we utilized imputation of missing values, feature engineering and data visualization to ensure the dataset was complete, interpretable, and optimized for downstream modeling. In the subsequent stage we built a comprehensive model based on logistic regression, decision tree, random forest, support vector machine, cox regression and hard voting and ultimately reached a high accuracy of 88.20%. This model can be adapted to many fields including healthcare, insurance and automotive safety, where classification or survival analysis is needed. With the approaches and processes given, government and enterprises could gain insights into how various factors influence certain events, which would contribute to a holistic resolution to tackling social issues.

There are some limitations to the model as well. For instance, some of the methods incorporated in the model, such as random forest and logistic regression, rely on sufficient sample size to achieve high precision and stability. Meanwhile, there are inevitably some associations between certain variables, as they are either deduced from each other or indicators of similar social patterns, which could potentially affect the accuracy of our prediction. This could be improved by taking use of regularization, dimensionality reduction and stepwise algorithm.

References

[1] Wikipedia contributors. (2025, October 31). Titanic. Wikipedia. https://en.wikipedia.org/wiki/Titanic.

[2] Gandhi, R. (2018, June 5). Support Vector Machine — Introduction to Machine Learning Algorithms | Towards Data Science. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47/.

[3] Abd ElHafeez, S., D'Arrigo, G., Leonardis, D., Fusaro, M., Tripepi, G., & Roumeliotis, S. (2021). Methods to Analyze Time-to-Event Data: The Cox Regression Analysis. Oxidative Medicine and Cellular Longevity, 2021(1), 1–6. https://doi.org/10.1155/2021/1302811.

[4] Mrisdal. (2017, December 26). Exploring survival on the Titanic. https://www. kaggle. com/ code/ mrisdal/exploring-survival-on-the-titanic.

[5] Thilaksha Silva. (2017, December 14). Predicting Titanic Survival using Five Algorithms. Kaggle.com; Kaggle. https://www.kaggle.com/code/thilakshasilva/predicting-titanic-survival-using-five-algorithm.

[6] Titanic Passenger List • Titanic Facts. (2020, July 13). Titanic Facts. https://titanicfacts.net/titanic-passenger-list/.

[7] Ts, S. (2020). RMS TITANIC CASE STUDY. Cusat. https://www. academia. edu/ 44571681/ RMS_ TITANIC_CASE_STUDY.