
# Project 3: Executive Report

ALY6040 – Data Mining Applications, Northeastern University

Professor Justin Grosz

04/27/2025

## Introduction

The goal of this project was to understand the dataset of the passengers travelling in the ship titanic having 891 rows/observations and 12 columns/variables containing various types information like the class of travel, the total fare they paid, travelling with the number of siblings/spouses /children, passenger's name, sex and age, whether they survived the incident or not, etc.

Here, first we have handled the missing data if present in any column and cleaned the dataset and built models using logistic regression and decision tree to predict survival chances of the passengers travelling in the titanic ship.

## Data Cleanup

The first step in our data cleaning process was to identify data variables, find missing data/null values present, if any, identify outliers and irrelevant features and apply appropriate techniques and convert categorical features into numerical encoding to achieve dataset of good quality.

The following variables were found to have missing data:

1. **'Age':** About 19.87% of its data was missing. To impute the missing values, group-based median imputation was chosen i.e., median age of passengers grouped by 'Pclass' and 'Sex'. This method was chosen because it accounts for correlation between 'Age', 'Pclass' and 'Sex' and it preserves the overall distribution of 'Age' within subgroups, avoiding oversimplification that single median would introduce and distort data distribution. Using grouped median here leverages domain knowledge(socio-economic status and age gap). Other methods like KNN could introduce complexity and overfitting when dataset is relatively small.
2. **'Cabin':** Had about 77% missing values. This column was removed because removing missing rows would eliminate substantial amount of data. Performing any imputation for such high number of missing values could introduce significant bias or noise into the model and might skew model's understanding of cabin-related patterns. Model can still perform well since 'Pclass' may capture similar socio-economic information indirectly.
3. **'Survived':** This is the target variable for prediction task. Rows with missing values cannot be used to train a supervised model. Any imputation in target variable could introduce biased evaluation or data leakage. Since, only 6 rows were missing which is less than 1% of the entire dataset, eliminating those rows would prevent any data leakage or bias in the model.
4. **'Parch':** With less than 1% missing data, impact of dropping these rows is minimal. Removing those missing rows avoids imputation thus preventing assumptions. Dropping would unlikely impact model performance.
5. **'Embarked':** Only had 2 rows(~0.23%) with missing data. Dropping these 2 rows is much more effective since performing any imputation is more complex and rather than imputing data which would be predicted constant value and might be incorrect one.

The second step in the data cleanup process is identifying irrelevant features and removing them from the dataset.

1. **'PassengerId':** This column is the unique identifier for each passenger. It serves no predictive purpose and has no correlation to 'survival'.
2. **'Name':** There is no purpose for 'name' in predicting survival chances other than it contains prefix 'Mr'/'Mrs' which provides gender information but that too is served by column 'sex'.
3. **'Ticket':** It represents the ticket number. Tickets are numbered in such alphanumeric way that represents the class and type of travel. Since, the class of travel is fulfilled by 'pclass' and there is no correlation of ticket number with survival chances and hence serves no purpose in the model building.

The third step in the data cleanup process is identifying outliers from numerical columns and applying appropriate techniques to handle those columns.

1. **'Age':** Upper bound age is 57.5 but people of age above 57.5 are valid and plausible for titanic passengers and relevant for survival prediction and hence are retained.
2. **'SibSp':** Upper bound is 2.5 but families having members more than 2.5(logically 3) are definitely possible and therefore outliers are retained. These values are not errors but reflect diversity of passenger groups.
3. **'Parch':** This column is heavily skewed and thus has 0 values for many rows. But values like 4, 5 or 6 are rare but real because parents travelling with their multiple children and cannot be removed since it may influence survival as families with children might be prioritized.
4. **'Fare':** Upper bound for fare is $65.6 and has maximum value of $512. Extremely high fares are rare and may indicate first-class travel or group fares. Since, this can disproportionately affect the model, capping them at upper bound preserves the scale and the fact that these passengers paid premium and may have higher survival rate.

The fourth step in the data cleanup process is to convert categorical features into numerical encoding.

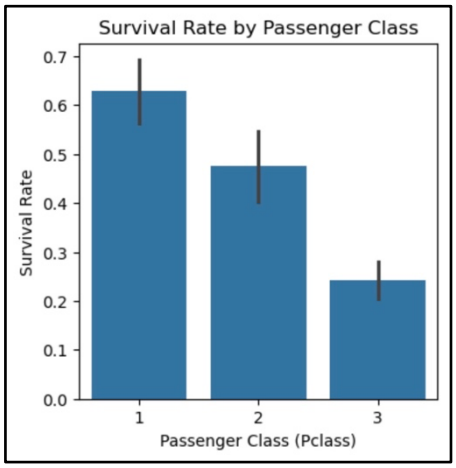
1. **'Sex':** Since model requires only numerical data type and not string data type, male is mapped to 0 and female is mapped to 1.
2. **'Embarked':** There are 3 ports of embarkation and hence therefore converted embarkation ports S to 0, C to 1 and Q to 2 for model to be built.

## Exploratory Data Analysis and Feature Engineering

For feature engineering, two new columns are created to perform EDA and prove/disprove hypotheses results.

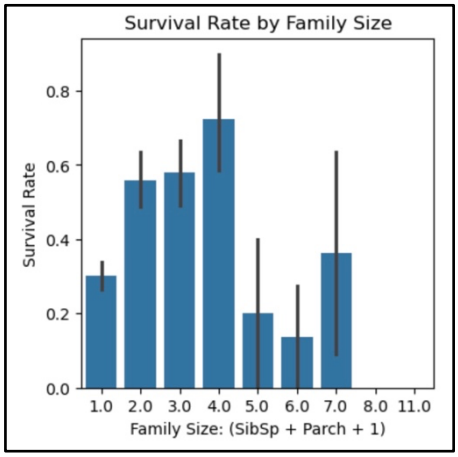
1. **'FamilySize':** Combine 'SibSp' and 'Parch' to create count of family size feature and add 1 including for passenger themselves.
2. **'IsAlone':** Created binary feature 'IsAlone':1 if 'FamilySize' is 0. Passengers travelling alone might have different survival chance as family with children might be prioritized.

**Hypothesis 1**: Purchasing ticket of higher class can lead to higher survival rates as crew may give them priority during evacuation as they paid premium for it.




From the visualisation result obtained in the left image, it proves that higher class correlates with higher survival rates. Hence, this proves that paying a premium and opting for higher class during travel can provide an advantage during any emergency situation and increases chance of surviving.

**Hypothesis 2:** Passengers traveling alone had higher survival rate than those with small families since they just need to run without having any responsibility of another passenger.



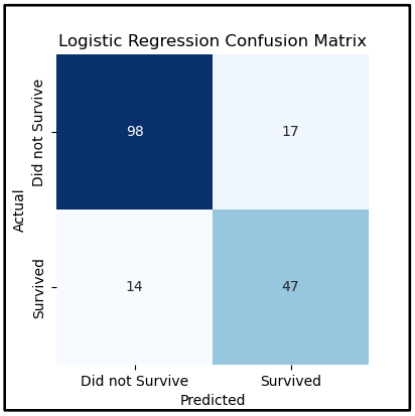

Using feature engineering, a new column was created to calculate family size which combined number of siblings to parents, parents to children and added 1 for the passenger themselves.

From the image result obtained in the left, our hypothesis is disproved that passengers traveling alone/solo had higher survival rate.

It is clear that family size ranging from 2 to 4 members are prioritized by the crew. This is possibly due to families having children. Also, family sizes above 4 are difficult to accommodate in the same lifesaving boats altogether and difficult for family members themselves to safely and responsibly take all their family members towards evacuation zone.

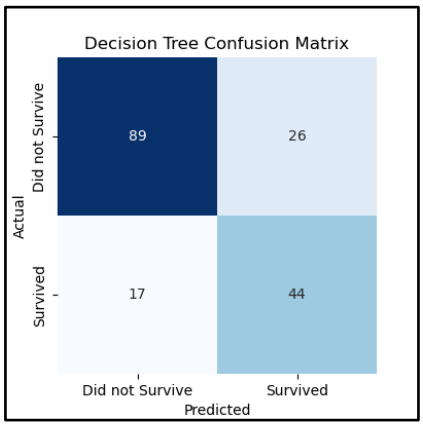
## Model Development and Evaluation

Logistic regression model and decision tree models are built to predict survival chance of passengers traveling in the titanic ship using original dataset variables: pclass, sex, age, sibsp, parch, fare and embarked but did not take into account of the features ‘FamilySize’ and ‘IsAlone’ generated using feature engineering.




Logistic regression model has 98 true negatives i.e., non-survivors as non-survivors and 47 true positives i.e., correctly predicted survivors as survivors.

While the model has 17 false positives meaning it predicted actual non-survivors as survivors and has 14 false negatives meaning those passengers who actually survived were predicted as did not survive.




Decision tree model has 89 true negatives i.e., non-survivors as non-survivors and 44 true positives i.e., correctly predicted survivors as survivors.

While the model has 26 false positives meaning it predicted actual non-survivors as survivors and has 17 false negatives meaning those passengers who actually survived were predicted as did not survive.

**Model Comparison**

| Model/Results | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic Regression | 0.8239 | 0.7344 | 0.7705 | 0.7520 |
| Decision Tree | 0.7557 | 0.6286 | 0.7213 | 0.6718 |

Logistic regression makes impressive prediction with an accuracy of 82.39%, striking good balance between precision(0.73) and recall(0.77). It correctly identifies 77.05% of survivors(high recall) but has some false positives meaning that 17 passengers actually did not survive but were predicted by model that they survived, leading to lower precision.

The model is effective at detecting survivors but occasionally predicts non-survivors as survivors.

Decision tree model has a lower accuracy of about 75.57% compared to logistic regression. It has low recall(0.7213) meaning it identifies less survivors than logistic regression(only 17 false negatives) and its precision (0.6286) is lower due to more false positives of about 26 passengers predicted as survived but actually did not survive.

The model is better at predicting survivors but at the cost of misclassifying more non-survivors as survivors, leading to lower overall accuracy and f1-score.

Logistic regression model performs best overall as it shows high accuracy and better f1-score indicating more balanced performance between precision and recall.

The F1-score for logistic regression is 0.7520 higher than decision tree's 0.6718 indicating that logistic regression achieves better balance between precision and recall

compared to that of decision tree. Decision tree has lower recall, its lower precision due to more false positives results in a lower F1-score. Thus, logistic regression outperforms the decision tree in terms of overall predictive balance.

| Model/Features | Sex | Pclass | Age | SibSp | Fare |
|---|---|---|---|---|---|
| Logistic Regression | 1.22 | -0.82 | -0.59 | -0.40 | 0.23 |
| Decision Tree | 0.3 | 0.1 | 0.27 | 0.06 | 0.21 |

For logistic regression model, the features 'sex', 'pclass', 'age', 'sibsp' and 'fare' are top contributors for model training. 'Sex' is the strongest positive predictor. One standard-deviation increase in sex significantly increases likelihood of survival, aligning with women and children first policy. The 'pclass' is strongest negative predictor, meaning that 1st class passengers are preferred than 2nd class and 3rd class passengers, which is consistent with our EDA hypothesis. Also, 'fare' is positively correlated and hence, higher fares(higher classes) lead to higher survival chances.

For decision tree model, the features 'sex', 'age', 'fare' and 'pclass' are top contributors. Here, 'sex' is most important feature, consistent with logistic regression. The decision tree frequently splits on sex, reflecting strong survival disparity between genders. Age, fare and pclass are then other relevant features in contributing survival prediction.

## Business Focus and Recommendations

From the analysis of titanic dataset, the two key variables with the highest impact on survival are 'sex' and 'pclass'. In both the logistic regression and decision tree models, 'sex' emerged as the most influencing feature. 'pclass' was second most impactful variable with first-class passengers having much higher survival rate followed by second and third class passengers.

1. Prioritize equitable safety protocols across demographics: Future safety protocols should ensure equitable treatment for all genders. For example- maritime safety drills could focus on balanced evacuation strategies ensuring equal access to lifeboats and medical kits for injured persons without any gender bias.
2. Address to socio-economic disparities in access to safety resources: Cruise lines and maritime regulators can improve survival predictions by modelling passenger distribution and ensuring that lifeboats access is uniform across all the passengers by implementing mandatory evacuation plans that prioritize equal access.

## Conclusion

In the analysis of titanic dataset of the passengers travelling in the ship, first we identified the features having missing data and were then cleaned using appropriate techniques to obtain dataset of good quality. The feature 'age' had missing data which were imputed with grouped based median by 'pclass' and 'sex' to preserve the overall distribution of age within subgroups. After cleaning, irrelevant features were removed from the dataset which were of no use in training and building of the model. After removing irrelevant features from the model, outliers were identified and appropriate techniques were applied to

handle the outliers. In the last step of data cleanup process, categorical features were converted into numerical encoding.

Our dataset is now sound and ready to perform EDA, feature engineering and to build models. Two new columns were created ‘FamilySize’ and ‘IsAlone’ to perform exploratory data analysis and perform EDA. As a result, survival rate for passengers travelling in higher class is more than passengers travelling in lower classes. Also, passengers travelling with families with 2 to 4 members in the family had higher survival chances than passengers travelling solo. Logistic regression and decision tree models are built, evaluated and compared out of which we found that logistic regression model is more accurate and had higher F1-score. Survival chances were strongly influenced by ‘sex’ and ‘pclass’ for the passengers travelling in the titanic ship.

## References

Canvas:

https://northeastern.instructure.com/courses/210219/pages/lesson-3-1-methods-overview-classification?module_item_id=11870594

https://northeastern.instructure.com/courses/210219/pages/lesson-2-6-logistic-regression?module_item_id=11870595