Counterfactual Explanations for Titanic Survival Prediction

Shruthi Addagudi | 2303A52133 | Batch - 37

# 1. Objectives and Methodology

## Objectives

The main goal of this project is to:  
1. Predict whether a passenger survived the Titanic disaster based on passenger details such as age, sex, class, and fare.  
2. Generate counterfactual explanations to understand what minimal changes in passenger features would alter the model’s prediction from 'Did Not Survive' to 'Survived'.  
3. Provide interpretable insights into model behavior using the DiCE (Diverse Counterfactual Explanations) library.

## Methodology

The following steps were followed:  
1. Dataset Loading - The dataset was downloaded directly from Kaggle using KaggleHub.  
2. Data Preprocessing - Missing values were handled, irrelevant columns dropped, and categorical variables encoded.  
3. Model Training - Logistic Regression and Random Forest models were trained.  
4. Counterfactual Generation - Using DiCE, counterfactuals were generated for insights.

# 2. Dataset Description

The dataset represents passenger details from the Titanic disaster, containing 891 rows and 12 columns before preprocessing.

|  |  |
| --- | --- |
| Column | Description |
| PassengerId | Unique ID for each passenger |
| Survived | Target variable (0 = No, 1 = Yes) |
| Pclass | Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) |
| Name | Passenger's name (dropped) |
| Sex | Gender (male, female) |
| Age | Age of passenger |
| SibSp | Number of siblings/spouses aboard |
| Parch | Number of parents/children aboard |
| Ticket | Ticket number (dropped) |
| Fare | Price paid for the ticket |
| Cabin | Cabin number (dropped) |
| Embarked | Port of embarkation (C, Q, S) |

# 3. Model Performance Results

After training both models, their performance was evaluated using precision, recall, F1-score, and accuracy.

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression | Random Forest |
| Precision (Class 0) | 0.83 | 0.83 |
| Precision (Class 1) | 0.79 | 0.82 |
| Recall (Class 0) | 0.86 | 0.89 |
| Recall (Class 1) | 0.74 | 0.74 |
| F1-score (Class 0) | 0.84 | 0.86 |
| F1-score (Class 1) | 0.76 | 0.78 |
| Overall Accuracy | 0.81 | 0.83 |

# 4. Counterfactual Explanations

Using the Random Forest model, one test instance predicted as 'Did Not Survive' was selected. The DiCE library generated three counterfactuals showing what changes would lead to a prediction of 'Survived'.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | CF #1 | CF #2 | CF #3 |
| Pclass | -0.369365 | 0.827377 | 0.827377 |
| Sex | 0.737695 | -1.355574 | -1.355574 |
| Age | -1.563997 | -0.104637 | -0.104637 |
| Fare | -0.341452 | 3.451164 | 9.470311 |
| Embarked | -1.942303 | -1.942303 | -1.942303 |
| Survived | 1 | 1 | 1 |

# 5. Interpretations and Reflections

Model Performance:  
- Random Forest was better at capturing non-linear relationships between features and survival probability.  
- Logistic Regression was simpler and interpretable but slightly less accurate.  
  
Counterfactual Explanations:  
- Lowering passenger class (Pclass) and changing gender to female significantly increased survival chances.  
- Increasing fare, which may indicate better cabins or services, also contributed to survival.  
- Younger age slightly improved survival probability.

# 6. Conclusion

This study successfully built predictive models to estimate Titanic passenger survival. Random Forest performed best with 83% accuracy. Counterfactual explanations generated using DiCE provided interpretable insights into how small changes in passenger characteristics could have altered survival predictions. Key determinants of survival included passenger class, gender, and fare.