Project Documentation: Exploratory Data Analysis of Titanic Survivors Prediction

By Harshit Dwivedi:

Dataset: Titanic Dataset

Source: Kaggle [(https://www.kaggle.com/datasets/brendan45774/test-file?datasetId=826163&sortBy=voteCount)]((https:/www.kaggle.com/datasets/brendan45774/test-file?datasetId=826163&sortBy=voteCount))

Project Objective:

Conduct data analysis of the Titanic dataset to uncover insights and patterns that can inform the prediction of passenger survival in this historical event.

Data Cleaning:

* Handled missing values by imputing or removing rows/variables as appropriate.
* Removed duplicates to ensure data consistency.
* Addressed outliers that could affect model performance.

Exploratory Data Analysis (EDA):

* Explored the data structure, including the distribution of passenger classes, ages, genders, and fares.
* Investigated the relationship between variables and survival status using data visualization.
* Calculated and visualized correlations between variables.
* Predictive Model and Overfitting:
* Built predictive models, including logistic regression or decision trees, to forecast passenger survival.
* Addressed overfitting by evaluating model performance using a confusion matrix.

Key Insights:

* Insight 1: Passengers with higher socioeconomic status (1st class) had a significantly higher chance of survival.
* Insight 2: Women and children had higher survival rates compared to men.
* Insight 3: The "Embarked" location 'C' had a relatively higher survival rate, possibly indicating a correlation with socioeconomic status.

Data Visualization:

* Created survival rate plots for passenger class, gender, and age groups.
* Generated a correlation matrix heatmap to highlight relationships between variables.
* Produced histograms to show age and fare distributions.

Conclusion:

In a nutshell, we found that rich folks (1st class), ladies, and kids had a better shot at surviving on the Titanic. If you got on the ship at 'C' (Cherbourg), your odds improved too.

Documentation Date:

24-09-2023