Project Report

Arman Nik Khah

University of Texas at Dallas

February 1, 2024

Introduction

- **Project Overview:**
    - Analyzing the Titanic dataset to predict survival outcomes.
    - Emphasis on the role of various socio-demographic factors.
- **Objectives:**
    - Develop predictive models to identify key determinants of survival.
    - Apply machine learning techniques for accurate predictions.
- **Approach:**
    - Comprehensive exploration of data (EDA).
    - Rigorous data preprocessing for model readiness.
    - Implementation of various machine learning algorithms.

Methodology

Our approach integrates several stages of data analysis and modeling. We begin with an extensive EDA to understand the data structure, followed by rigorous data preprocessing, including handling missing values and feature engineering. We then implement various machine learning algorithms to predict survival outcomes. Our methodology emphasizes the integration of statistical methods and machine learning techniques to derive meaningful insights and accurate predictions from the data.

Exploratory Data Analysis (EDA)

Objective: Uncover patterns, anomalies, and relationships within the Titanic dataset.
Impact: Guided data preprocessing and modeling strategies.

Data Insights and Interpretation

Passenger Demographics:

- Survival rate varied across age, gender, and socio-economic classes.
- Higher survival rates for younger passengers and women.

Socio-Economic Status:

- Clear divide in survival chances based on socio-economic classes.
- Higher survival likelihood for passengers in higher classes.

Visualization and Analysis

Distribution Plots:

- Visualize age and fare distributions.
- Insight into socio-economic status and age profile.

Correlation Analysis:

- Heatmaps and correlation matrices to explore feature interdependencies.
- Identification of features influencing survival.

Survival Rate Comparisons:

- Bar graphs and pivot tables to compare survival rates.
- Highlighting disparities based on gender and passenger class.

Anomalies and Patterns

Family Dynamics:

- Influence of family size on survival rates.
- Larger families had lower survival rates.

Embarkation Points:

- Survival rates varied based on embarkation points.
- Suggests differences in lifeboat allocation at different ports.

Statistical Summary

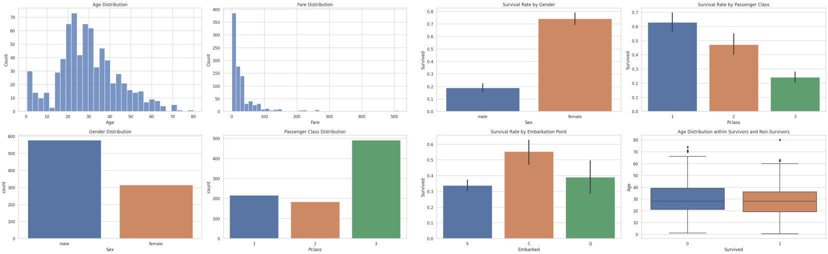
Descriptive Statistics:

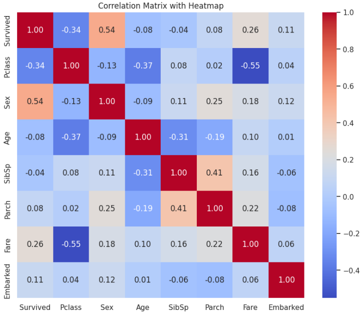
- Comprehensive summary with measures of central tendency and dispersion.

Interpretation:

- EDA insights into socio-economic, demographic, and familial factors.
- Crucial role of age, gender, and socio-economic status in survival.

EDA Visualizations




Correlation Matrix with Heatmap

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
Survived	1.00	-0.34	0.54	-0.08	-0.04	0.08	0.26	0.11
Pclass	-0.34	1.00	-0.13	-0.37	0.08	0.02	-0.55	0.04
Sex	0.54	-0.13	1.00	-0.09	0.11	0.25	0.18	0.12
Age	-0.08	-0.37	-0.09	1.00	-0.31	-0.19	0.10	0.01
SibSp	-0.04	0.08	0.11	-0.31	1.00	0.41	0.16	-0.06
Parch	0.08	0.02	0.25	-0.19	0.41	1.00	0.22	-0.08
Fare	0.26	-0.55	0.18	0.10	0.16	0.22	1.00	0.06
Embarked	0.11	0.04	0.12	0.01	-0.06	-0.08	0.06	1.00

Data Preprocessing

Handling Missing Values

Data preprocessing is a critical step in our analytical process, determining the effectiveness of our machine learning models. This section outlines the strategies employed to prepare the Titanic dataset for analysis, along with interpretations of the impacts of these methods.

- **Age**: Median imputation was used due to the robustness of the median against outliers in age distribution.
- **Cabin**: The high proportion of missing values in 'Cabin' led to the extraction of deck information, transforming a largely incomplete feature into a categorical variable.
- **Embarked**: Missing values in 'Embarked' were filled with the mode, reflecting the most common embarkation point.

Data Preprocessing

Feature Engineering

Two significant features were engineered:

- **FamilyNameEncoded**: Captured family groupings, hypothesizing an impact of family size and structure on survival rates.
- **Deck**: Extracted from 'Cabin', aimed to explore the influence of passenger location on survival.

Data Preprocessing

Feature Transformation and Normalization

Categorical variables like 'Embarked', 'Deck', and 'Pclass' were processed using One-Hot Encoding to avoid introducing artificial ordinal relationships. Continuous variables such as 'Age', 'Fare', and 'Family Size' were normalized using MinMaxScaler to ensure equal contribution in model training.

Machine Learning Model Implementation - Overview

In this project, we employed a range of machine learning models, each with unique strengths and adaptabilities to our dataset characteristics. These models include:

- Logistic Regression
- Gaussian Naive Bayes
- K-Nearest Neighbors (KNN)
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- Ensemble methods
- Fully Connected Neural Network (FCN)

We also explored a Fully Connected Neural Network (FCN) for its capacity to capture non-linear patterns.

Machine Learning Model Implementation - Model Performance and Interpretation

Logistic Regression
Accuracy: 87.68%. Effective in class separation with a balance of precision and recall.

Random Forest
Accuracy: 86.23%. Offers depth in feature importance analysis, efficient in handling categorical and continuous variables.

Support Vector Machines (SVM)
Accuracy: 68.84%. Challenges in application to our specific dataset.

Ensemble Models
Accuracy: 86.96%. Combines strengths of individual models, offering comprehensive understanding.

Fully Connected Neural Network (FCN)
Accuracy: 85%. Comparable to traditional models, but with higher complexity and computational requirements.

Machine Learning Model Implementation - Insights and Implications

The diverse performances of these models underscore the complexity of the Titanic dataset. Key insights include:

- Logistic Regression's high performance suggests significant linear relationships.
- Success of Random Forest and Ensemble models indicates the presence of complex, non-linear patterns.
- Variance in model effectiveness highlights the importance of feature selection and engineering.

These insights are crucial in predictive modeling, especially in datasets with socio-economic and demographic nuances like ours.

Evaluation

Model Performance Metrics

In-depth evaluations were conducted for models including Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machines (SVM), Ensemble Models, and a Fully Connected Neural Network (FCN). Key metrics: Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

Evaluation

Model Comparisons and Interpretations

- Logistic Regression: High accuracy, balanced precision-recall, effective class differentiation.
- Gaussian Naive Bayes: Lower accuracy, high recall, effective in identifying true positives.
- K-Nearest Neighbors (KNN): Lower performance, sensitive to irrelevant features.
- Decision Tree: Moderate performance, less effective in class differentiation.
- Random Forest: High accuracy and precision, good recall, robust in classification.
- Support Vector Machines (SVM): Reasonable accuracy, limited interpretability due to test set characteristics.
- Ensemble Model: High AUC-ROC and recall, balanced precision and accuracy.
- Fully Connected Neural Network (FCN): Comparable accuracy, traditional models were sufficient.

Evaluation

Heatmap Visualization

A heatmap visualized the performance metrics, facilitating comparison and informed model selection.

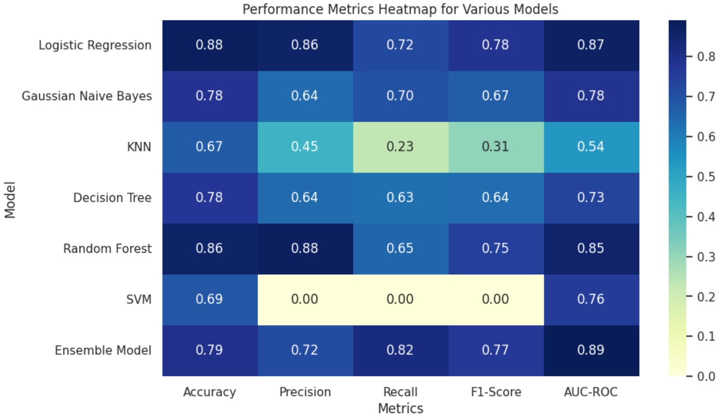


Figure: Heatmap of Model Performance Metrics

Evaluation

Interpretation of Results

- Effective models captured both linear and non-linear relationships.
- Challenges in KNN and SVM performance due to feature space and distribution.
- Ensemble Model's balanced performance highlighted the benefit of algorithmic diversity.

The evaluation highlighted the complexity of the Titanic dataset and the need for model-specific application requirements.

Feature Importance Analysis

- Analysis conducted for both Logistic Regression and Random Forest models.
- Key features influencing survival rates:
  - Socio-economic status (Pclass)
  - Gender (PersonType)
  - Age
- Other notable features:
  - Deck
  - FamilyNameEncoded
- These features highlight the impact of cabin location and family dynamics on survival.

Feature Importance Analysis - Visuals

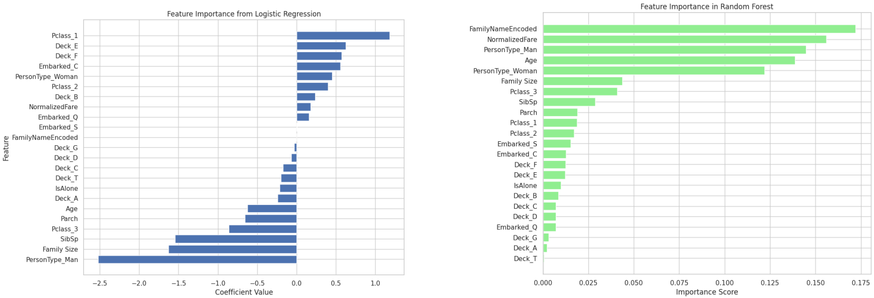


Figure: Feature importance in Logistic Regression and Random Forest models.

Comparative Model Analysis

- **Models Compared:**
  - AdaBoost
  - GradientBoosting
  - Random Forest
  - XGBoost
- **Top Performers:** XGBoost and GradientBoosting showcased the highest accuracy and F1-Scores, indicating a balanced precision-recall trade-off.
- **Performance Metrics:**
  - XGBoost marginally outperformed other models in terms of precision.
  - XGBoost showed slightly higher AUC-ROC scores.
- **Key Insight:** The models' performance varied, but XGBoost was a consistent front-runner.

Conclusions

- Comprehensive analysis using diverse machine learning models provides insights into the Titanic disaster's survival dynamics.
- **Gender, Age, and Socio-Economic Status:** Key determinants of survival.
  - Logistic Regression: Negative impact of being a man on survival chances.
  - Importance of passenger class and normalized fare, highlighting survival disparity based on class.
- **Family Dynamics:** Random Forest model underscores the role of family affiliations in survival.
- **Model Performance Comparison:**
  - Logistic Regression and Ensemble Models excel in accuracy.
  - XGBoost shows high precision and AUC-ROC, indicating its effectiveness in identifying true positives.
- **Cabin and Deck Analysis:** Certain deck locations, like Deck_D and Deck_E, associated with higher survival rates.
- **Embarkation Points:** Influence on survival rates could suggest logistical or administrative differences based on boarding locations.