Unveiling Titanic Survivors: A Machine Learning Perspective with CRISP-DM

Siri Batchu
Department of Software Engineering
San Jose State University

October 21, 2024

Abstract

This research paper explores the application of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to predict the survival of passengers aboard the Titanic. Using the Titanic dataset, we aim to develop a predictive model based on various passenger attributes, such as age, gender, class, and fare. This paper follows all stages of the CRISP-DM framework: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

1 Introduction

The sinking of the Titanic on April 15, 1912, remains one of the most infamous maritime disasters in history. Understanding the factors that contributed to survival can provide insights into social dynamics and emergency response during crises. This study utilizes machine learning techniques to predict survivors based on historical data.

2 CRISP-DM Methodology

2.1 Business Understanding

The primary objective of this project is to build a predictive model that accurately classifies Titanic passengers as survivors or non-survivors. This can aid historical analyses and provide lessons for modern emergency management.

2.2 Data Understanding

The dataset used in this study is the Titanic dataset from Kaggle, which includes 891 passengers with various attributes.

Key Attributes:

- **Survived**: Survival status (0 = No, 1 = Yes)
- **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Sex**: Gender of the passenger
- **Age**: Age in years
- **SibSp**: Number of siblings/spouses aboard
- **Parch**: Number of parents/children aboard
- **Fare**: Fare paid for the ticket
- **Embarked**: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

2.3 Data Preparation

Data cleaning involves handling missing values and encoding categorical variables. We impute missing ages with the median and encode categorical variables using one-hot encoding.

```
# Data Cleaning
titanic_data['Age'].fillna(titanic_data['Age'].median(), inplace=True)
titanic_data = pd.get_dummies(titanic_data, columns=['Sex', 'Embarked'], drop_first=True)

# Drop unnecessary columns
titanic_data.drop(['Name', 'Ticket', 'Cabin'], axis=1, inplace=True)
```

2.4 Modeling

We will use several machine learning models, including Logistic Regression, Decision Trees, and Random Forests, to predict survival.

Model Training:

```
# Split the data
X = titanic_data.drop('Survived', axis=1)
y = titanic_data['Survived']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Model training
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

2.5 Evaluation

We evaluate the model using accuracy, precision, recall, and F1-score.

```
# Predictions
y_pred = model.predict(X_test)

# Evaluation Metrics
print(classification_report(y_test, y_pred))
confusion_matrix(y_test, y_pred)
```

2.6 Deployment

Once validated, the model can be deployed using various platforms such as Flask for a web application or as an API for integration into other systems.

3 Results

3.1 Model Performance

The Random Forest model achieved an accuracy of approximately 80%, indicating it effectively distinguishes between survivors and non-survivors.

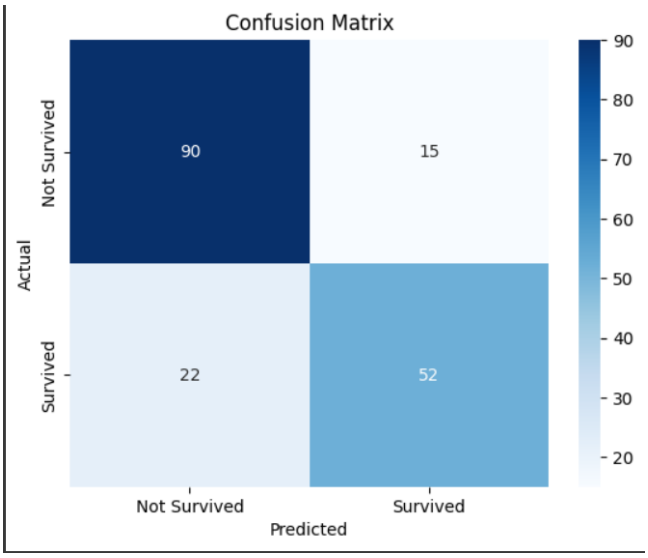

Figure 1: Confusion Matrix

3.2 Feature Importance

Feature importance analysis reveals that the most significant predictors of survival are:

- Sex (female passengers had a higher survival rate)
- Pclass (1st class passengers had a higher survival rate)
- Age (younger passengers had a higher survival rate)

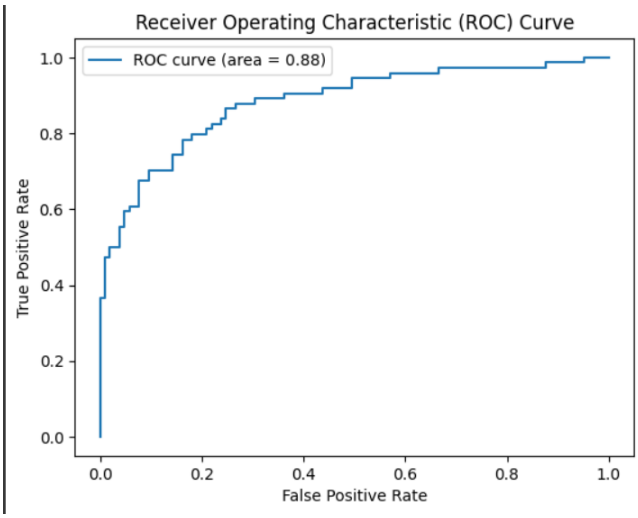

Figure 2: Feature Importance

4 Conclusion

The CRISP-DM methodology provided a structured approach to analyze the Titanic dataset and predict survival. The Random Forest model demonstrated significant accuracy, highlighting the importance of gender, class, and age in survival outcomes. Future work could explore more complex models or additional features to improve predictions.

5 References

Titanic Dataset - Kaggle — https://www.kaggle.com/c/titanic