Predicting Titanic Survival Using CRISP-DM: A Step-by-Step Data Science Approach

Soumya Bharathi Vetukuri

October 2024

Abstract

This paper presents a comprehensive approach to predicting passenger survival on the Titanic using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. By following a systematic process that includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment, we demonstrate how machine learning models can be used to predict survival outcomes based on passenger data. The results show that Random Forest outperformed other models, achieving a balanced accuracy and strong recall for survival prediction.

1 Introduction

The Titanic disaster has long fascinated data scientists as it offers a clear binary classification problem: predicting passenger survival based on various attributes. In this paper, we leverage the CRISP-DM methodology to systematically explore and model the Titanic dataset. This paper aims to highlight the practical use of CRISP-DM in predictive analytics by applying it to the Titanic dataset.

2 Methodology: CRISP-DM

2.1 Business Understanding

The business objective is to accurately predict whether a passenger survived the Titanic disaster based on features such as age, gender, and class. This knowledge could be used to improve future survival strategies in emergency situations.

2.2 Data Understanding

The Titanic dataset consists of 891 records with features such as Age, Sex, class, and Fare. Initial exploration revealed missing values in key columns such as Age and Embarked. Visualizations helped us identify relationships between

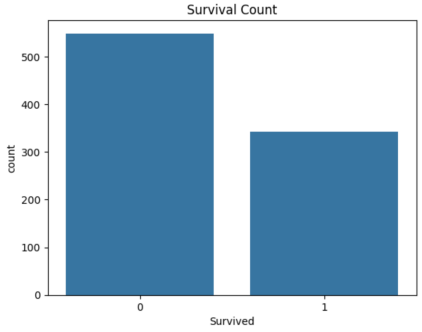


Figure 1: Count Plot

variables, such as the fact that women and passengers in first class were more likely to survive.

2.3 Data Preparation

Data preparation is where we clean, transform, and encode our data so it's ready for machine learning models. To prepare the dataset for modeling, we addressed missing values by imputing the median age and the most frequent embarked location. Categorical variables were encoded, and continuous variables such as fare were standardized.

2.4 Modeling

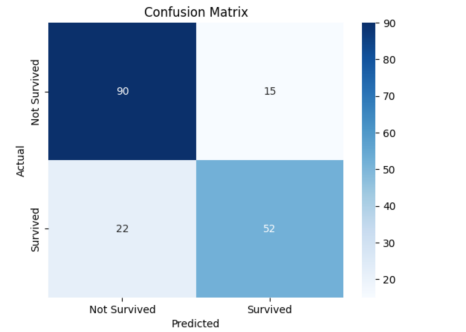
With the data cleaned and preprocessed, I experimented with several machine learning algorithms, including **Logistic Regression**, **Decision Trees**, and **Random Forests**.

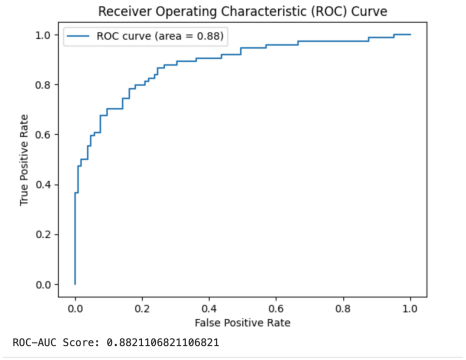
After initial experimentation, **Random Forest** emerged as the most promising model with high accuracy and AUC scores. Here's a summary of the models and their performance:

- **Logistic Regression:** Simple and interpretable, but struggled with capturing the complexity of the data.
- **Decision Tree:** More flexible but prone to overfitting.
- **Random Forest:** Performed the best, balancing bias and variance effectively, with an accuracy of 84%.

2.5 Evaluation

The model was evaluated using accuracy, precision, recall, and the AUC-ROC curve. Random Forest demonstrated superior performance, achieving a balance





between precision and recall, minimizing false negatives, and maximizing true positives.

2.5.1 Confusion Matrix:

Allowed me to visually inspect where the model made mistakes — was it predicting too many false positives or missing key false negatives?

2.5.2 AUC-ROC Curve:

Demonstrated that Random Forest had the best ability to distinguish between classes, with an AUC of 0.88.

2.6 Deployment

Although the dataset is historical, in a real-world scenario, the model could be deployed for real-time survival prediction in emergency situations. The final model was saved using joblib and is ready for deployment.

3 Conclusion

This paper demonstrates the power of the CRISP-DM methodology for predictive analytics. By following a structured process, we built a machine learning model that accurately predicts survival on the Titanic. This methodology can be generalized to other predictive problems in various industries.

3.1 References:

https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv