Title: Predicting Passenger Survival on the Titanic Using Machine Learning Techniques

Author:

Prashik Gajanan Bhimte

Abstract:

This study explores the application of machine learning algorithms to predict passenger survival on the RMS Titanic. We utilize a publicly available dataset containing information about passengers on the fateful voyage. Different classification models, including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest, and an Artificial Neural Network (ANN), are evaluated to determine the most effective approach for this task.

Keywords: Machine Learning, Survival Prediction, Titanic, Classification

1. Introduction

The RMS Titanic's tragic sinking in 1912 remains a haunting reminder of the importance of safety regulations in maritime travel. This event has also captured the imagination of researchers, leading to numerous studies exploring various aspects of the disaster. Here, we leverage machine learning techniques to shed light on factors that might have influenced passenger survival.

2. Dataset and Preprocessing

We employ the well-known Titanic dataset from Kaggle, which comprises information on passengers, including their Pclass (passenger class), Sex, Age, SibSp (number of siblings/spouses aboard), Parch (number of parents/children aboard), Fare (ticket price), Cabin, and Embarked location. The dataset contains 891 training instances and a separate test set with 418 instances for prediction.

Data preprocessing is crucial for effective machine learning. We address missing values using a SimpleImputer with the Mean strategy. Categorical features like Pclass, Sex, and Embarked are encoded using OneHotEncoder. Additionally, features that are unlikely to contribute significantly to survival prediction, such as PassengerId, Name, Ticket, and Cabin, are dropped. Finally, feature scaling is applied using StandardScaler to ensure consistent data distribution.

3. Methodology

We evaluate several classification algorithms:

- **Logistic Regression:** A linear model that calculates the probability of survival based on a weighted combination of input features.
- **Support Vector Machine (SVM):** A powerful algorithm that finds a hyperplane in the feature space that best separates the surviving and deceased passengers with a maximum margin. We consider both linear and RBF (Radial Basis Function) kernels.
- **K-Nearest Neighbors (KNN):** A non-parametric method that predicts the survival of a passenger based on the majority vote of its K-nearest neighbors in the training data.
- **Random Forest:** An ensemble method that builds multiple decision trees on random subsets of features and aggregates their predictions for improved accuracy and robustness.
- **Artificial Neural Network (ANN):** A layered model inspired by the human brain, capable of learning complex relationships between features and survival. We construct a simple ANN with two hidden layers of 10 neurons each with the ReLU (Rectified Linear Unit)

activation function. The output layer uses a sigmoid activation function to predict survival probability.

For each model, we perform grid search or random search to identify hyperparameters that optimize performance on a validation set derived from the training data. We assess model performance using accuracy, a metric that measures the proportion of correct predictions.

4. Results and Discussion

Table 1 summarizes the accuracy achieved by each model:

Model	Accuracy
Logistic Regression	0.77033
SVM (Linear)	0.76555
SVM (RBF)	0.77990
KNN	0.75837
Random Forest	0.75358
Artificial Neural Network (ANN)	0.78229

The Artificial Neural Network exhibits the highest accuracy of 0.78229, demonstrating its capability to learn complex relationships between features and survival. While the other models also achieved respectable accuracies, the ANN offers a slight edge.

5. Conclusion

This study has explored the use of machine learning algorithms for predicting passenger survival on the Titanic. The Artificial Neural Network emerged as the most effective model, achieving an accuracy of 0.78229. These findings highlight the potential of machine learning for historical analysis and suggest that incorporating non-linear relationships between features can improve predictive accuracy.

6. Future Work

This work lays the foundation for further exploration. Future research could:

- Investigate the use of feature engineering to create new features that capture more meaningful relationships between variables.
- Incorporate additional datasets containing information on weather patterns, ship design, and evacuation procedures to potentially improve prediction accuracy.
- Experiment with more