Machine Learning from Titanic Disaster

Kaggle Competition Project

Introduction

To effectively use Kaggle, I first researched its purpose and functionality. Once I familiarized myself with the platform, I learned how to import data from Kaggle's data sets. After importing the data, I processed it accordingly and implemented a Logistic Regression model to make predictions. Finally, I generated a submission file using these predictions and uploaded it to Kaggle for review.

Approach

- Unnecessary data columns which are Ticket, Cabin, Name and Passenger Id removed from training and test data with using drop method by Pandas.
- Missing data was identified in the SibSp, Parch, Fare, and Age columns. To address this, I used the "Fill Na" method in Pandas and replaced the missing data with the median value of all the data in the respective columns.
- Sex and embarked data were in string form I replaced them with numeric values with using preprocessing "Label Encoder" method e.g., male:0, female:1.
- For my project's predictions, I imported the Logistic Regression model. I also imported the "Train Test Split" method to test the predictions.
- The test size has been set to 0.2, while the random state has been defined as 44.
- The classifier has been configured to use the Logistic Regression model.
- To assess the accuracy of the model, I imported the accuracy score from Scikit Learn.
- Test data is used to make predictions.
- Keys and values defined with using Data Frame method from pandas for the submission.
- Submission file created using the “To csv” method from pandas Data Frame.

Libraries and Methods Used

- **Pandas**
    1. Data Frame – Drop
    2. Data Frame – Fill Na
- **Scikit Learn**
    1. Preprocessing
    2. Logistic Regression
    3. Train Test Split
    4. Accuracy Score

Conclusion

For predictive and classification problems, Logistic Regression models are commonly employed. To enhance the accuracy of my machine learning model, I pre-processed the data and eliminated some columns. While not ideal, I did everything possible within my capacity. As a result, my model's performance is approximately "0.77."

Kaggle Score

Submission and Description	Public Score
submission.csv Complete · 4h ago	**0.76315**
submission.csv Complete · 15h ago · This is my first Machine Learning Exercise.	**0.77033**