FUNDAMENTALS OF DATA SCIENCE

Final Project Report

Sapienza University of Rome

Spaceship Titanic: Binary Classification

Ambar Chatterjee, Himel Ghosh,
Paul Jèzèquel, Mayis Atayev

Abstract

This project report delves into the "Spaceship Titanic" dataset, focusing on a binary classification problem: predicting which passengers were transported to an alternate dimension when the spaceship collided with a spacetime anomaly. To address this problem, we use machine learning models, deploying a thorough data pipeline and evaluation metrics to identify the most effective model. The study guides you through exploratory data analysis, feature engineering, and model selection, culminating in a prediction model that highlights the intricate relationship between methodology and performance.

Introduction

The «Spaceship Titanic» dataset is a task set in the year 2912, when the spacecraft encounters a space disaster leading to the disappearance of passengers in an alternate dimension. The complexity of the dataset, combined with missing values and the need for careful feature engineering, makes it fertile ground for the application of various machine learning methods. Our goal is to solve this problem by achieving high prediction accuracy while simultaneously understanding the main factors affecting the performance of the models we used.

Related Work

There are a number of approaches within the Kaggle community, e.g. complete guides and end to project analysis, that emerge from this literature review. Our methodology, incorporating insights and proven solutions from previous work into the pipeline, is based on references like "Spaceship Titanic: A Complete Guide" and a wide variety of Kaggle code repositories.

Methodology: Data Processing and Model Selection

In order to understand the dataset's characteristics, our data pipelines start with an EDA. After that, feature engineering is conducted to generate important features which can potentially improve performance of our models. In order to ensure data integrity and comparability, the correction of missing values and scale is performed. Dimensionality reduction techniques are used in order to focus on the most relevant features. A selection of models, including Random Forest, Gradient Boosting, Neural Network, and a Stacked model, are considered. In order to optimize the model, each model is subjected to hyperparameter tuning, which is followed by an evaluation of the model's performance across the various metrics.

Dataset and Benchmarking

The dataset is obtained from Kaggle and consists of the files train.csv (8693x14) and test.csv (4277x13). It includes 14 features covering various aspects of passengers and their journey. Our benchmarking draws on the literature review, providing a comparative backdrop against existing solutions and their reported scores on Kaggle.

Results

The evaluation of our models has shown that the neural networks have the highest precision, the Gradient Boosting has the highest ROC AUC score, and the Stacked model has the highest accuracy, recall, and F1 score. We found out that not all models prioritize same features as important. Processing the missing values and reducing the dimension has been shown to have a negative effect on scores. In addition, the effectiveness of minimal feature engineering compared to more sophisticated techniques has been demonstrated which suggests a balance between data manipulation and model complexity. Possible reasons for the decrease of the accuracy metric after specific feature selection can be listed as follows:

1. Original model might have been overfitting to the training data, capturing noise as if it were a signal. After removing certain features, the model's ability to overfit decreased, which might have reduced performance if the model was relying on that noise to make predictions.
2. The features which might not have been so important individually, might possess some collective influence with certain other features. Hence, the assumption of features being independent of each other might not necessarily hold true and hence the ignorance of those features have led to decreased performance.

Conclusion and Future Work

The findings indicate a nuanced landscape where different models excel in various aspects of the prediction task. The choice of data preprocessing techniques and feature engineering has a significant impact on the outcome. In order to find robust strategies that improve model performance, the research journey will focus on more in depth exploration of innovative approaches for dealing with missing data. In addition, we're exploring the potential of ensembles techniques and looking at how combining several models can synergize with each other to enhance predictions. In our future work, as we strive to reach the limit of accuracy and reliability, continued refinement of features and model parameters will remain a central focus. To ensure the evolution of our Predictive Models with the latest innovation in this field, we will continue to rigorously assess new methods and developments in ML.

References

1. Samuel Cortinhas, "Spaceship Titanic: A Complete Guide", Kaggle Code

Repository.

2. KD Sharma, "Spaceship Titanic Competition: End-to-End Project", Kaggle Code

Repository.

3. Gusthema, "Spaceship Titanic with TF-DF", Kaggle Code Repository.