

Predicting the future space mission status

Abrar Alsalman

Abstract

The goal of this project was to use classification models to predict the status of the next mission in order to help improve operations and maintenance planning of these units. I worked with data provided by Kaggle (Space_Corrected.csv from <https://www.kaggle.com/agirlcoding/all-space-missions-from-1957>) including all the space missions since the beginning of Space Race (1957). After refining a model, I built an interactive dashboard to visualize and communicate my results using PowerPoint.

Design

This project originates from the latest need to have accommodation/entertainment trips to space. The data is originally provided by (<https://nextspaceflight.com/launches/past/?page=1>). It presents for classes for Classifying statuses accurately via machine learning models would have to make a decision whether this mission is going to succeed or fail.

Data

It contains **9 rows** and **4323 observations** which are:

- **Unnamed 0:** Index.
- **Unnamed 0.1:** Index.
- **Company Name.**
- **Location:** Location of the Launch.
- **Datum:** Datum and Time of Launch.
- **Detail:** Rocket Name.
- **Status Rocket:** Status of the rocket whether it is active or retired.
- **Rocket:** Cost of the mission: in \$ million.

- **Status Mission:** Status of the mission whether it is successful or not.

Algorithms

- **Feature Engineering**

1. Splitting some columns to more detailed columns:

Datum -> date, month, year.

Location -> country.

2. Converting categorical features to binary dummy variables
3. Combining particular dummies and ranges of numeric features to highlight the mission status.

- **Models**

Logistic regression, k-nearest neighbors, and random forest classifiers were used before settling on random forest as the model with strongest performance.

- **Model Evaluation and Selection**

The entire training dataset was split into 70/30 train vs. test. All scores reported below were calculated on the test portion only.

Final random forest scores:

- **Accuracy:** 100%
- **F1 score:** 1.00
- **precision:** 1.00
- **recall:** 1.00

This is the accuracy for all the models:

1) Logistic regression Score = 92%

2) KNN Classifier Score = 91%

3) Random Forest Classifier Score = 100%

Tools

- Numpy and Pandas for data manipulation.
- Scikit-learn for modeling.
- Matplotlib and Seaborn for plotting.
- PowerPoint for interactive visualizations.

Communication

PowerPoint with visuals is used to deliver the outcome.