Project Milestone 1

Harlan Wittlieff

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

July 9, 2022

Topic

This project is named Space Titanic. The goal is to predict which passengers of the spaceship have been transported to another dimension.

Business Problem

Kaggle unveils fascinating data science competition projects. For this instance, Kaggle developed a dataset for a futuristic scenario of an interstellar passenger liner, the *Spaceship Titanic*, collided with a hidden spacetime anomaly. This collision sent some of the passengers onboard into an alternate dimension.

Rescue crews must s+ave the lost passengers. Therefore, I must predict which passengers were transported using a recovered dataset from the spaceship's damaged computer system. Being able to accurately predict which passengers are transported will additionally give insight for future voyages into space. From these insights measures may be taken to reduce the likelihood of a passenger being transported to another dimension.

Datasets

Kaggle already separated this data into a training and test datasets featuring the following variables:

- Passenger ID – Unique Identifier
- Home Planet
- Cryo Sleep
- Cabin
- Destination
- Age
- VIP
- Room Service

- Food Court
- Shopping Mall
- Spa
- VR Deck
- Name
- Transported

Methods

The target variable for this project is “Transported,” as the model will predict whether a passenger was transported to the alternate dimension. If the classes are not balanced, a method such as SMOTE would need to be leveraged to address the issue. As this is a classification scenario, a logistic regression model will be investigated. Additional models, such as the random forest model may also warrant investigation pending the results of the logistic regression model.

Feature engineering may provide some useful information for the model. For example, engineering features from the Cabin variable may provide additional learning opportunity. Another potential feature will be creating family sizes. This can be extracted from the PassengerID variable as family members are sequenced after the initial ID.

Ethical Considerations

Since this is a fictional dataset, the real-life ethical impacts are minimal. However, if this was a real-life scenario care would need to be taken to ensure that no classes of people are treated in a biased fashion. Passenger transportation has the potential to create a life-or-death situation. Accurately predicting a specific subset of the population’s transportation status at the expense of another subset would be unacceptable. In other words, the VIP passengers should not be given any preferential treatment relative to the construction of the model than other class.

Challenges/Issues

The data set contains null values likely due to the damage sustained to the computer. Identifying the correct way to handle the nulls will be challenging.

Feature engineering has the potential to provide additional nuggets of information for the model. Engineering the best features will primarily be done via trial and error.

References

Kaggle (2022). *Spaceship Titanic*. Www.kaggle.com. https://www.kaggle.com/competitions/spaceship-titanic/overview