White Paper

Harlan Wittlieff

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

July 24, 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 save 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.

Data Exploration

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

Overall, the training dataset contains information for 8,693 passengers. Of these 4,378 ended up being transported to an alternate dimension.

Data Preparation

To begin preparing and exploring the data, I first checked for null values. Several features contained nulls. The features RoomService, FoodCourt, ShoppingMall, Spa, and VRDeck all had their nulls replaced with zeros as these features represented amounts spent on various amenities aboard the ship. Nulls in the VIP & CryoSleep features were replaced with False. Lastly the nulls in the Age feature were replaced by the mean age of the dataset. Nulls in categorical features were left in the dataset.

Three additional features were created from the dataset. Deck and Side (port or starboard) were extracted from the Cabin feature to provide richer information on where the passenger was staying to the model. The number of passengers in each specific group was added to the features. This feature was extracted from the PassengerId field.

Lastly, True and False values within the data were replaced by the binary values 0 and 1. The dataset was split into training and test sets with 80% of the data reserved for training the model.

Methods

The target variable for this project is “Transported,” as the model will predict whether a passenger was transported to the alternate dimension. No class balancing was needed as the data

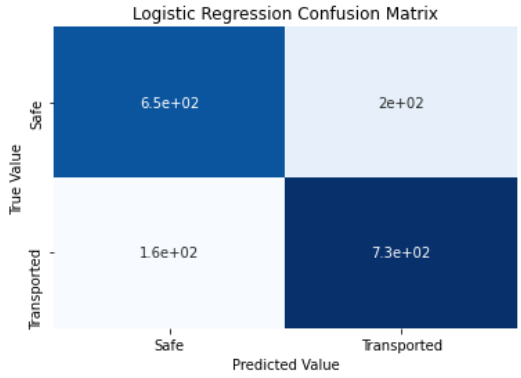
contained equal members in both transported and not transported. Overall, three models were constructed. These models included logistic regression, random forest, and XGBoost.

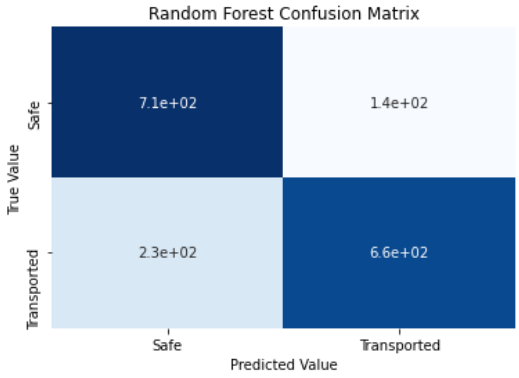
Analysis

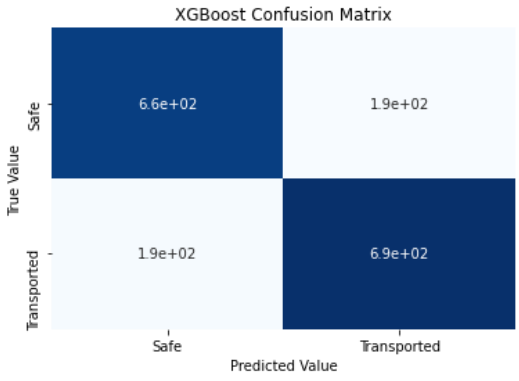
Since the data classes are balanced, the main method of analyzing the models is the accuracy of the predictions on the test dataset which are as follows:

1. Logistic Regression: 79.4%
2. Random Forest: 78.5%
3. XGBoost: 78.0%

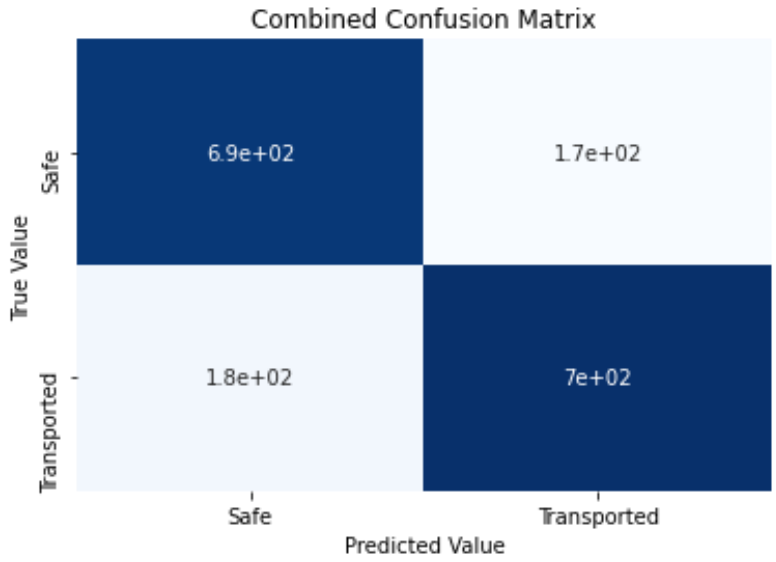
The logistic regression model performed the best on the test dataset with an accuracy of 79.4%. Confusion matrices were then built for each model to further evaluate the results.







With the individual model test results completed, a final model was built that combined the results from each individual model. The average prediction between the Logistic Regression, Random Forest, and XGBoost models was taken and used as the predicted value in this final model. This combined model was able to achieve an accuracy of 79.9%. That's an increase of 0.6% from the best individual model. The combined model's confusion matrix is pictured below.



Conclusion

Based on the model evaluations, the logistic regression model was the best individual model. However, combining the three models proved to be the best solution overall with an accuracy of 79.9%.

Assumptions & Limitations

The greatest assumptions were made around the null values existing in the dataset. Additional methodologies for handling the variety of null values contained within the features of the data may warrant further investigation.

Challenges

The dataset contained null values likely due to the damage sustained to the computer. Identifying the correct way to handle the nulls was challenging. Due to the fact that this dataset is fictitious, identifying the proper way to evaluate the data proved difficult.

Recommendations & Additional Applications

Although this was a fictional dataset, the methodology may have applications in real-life accident classification problems. Adding a neural network model both individually and as a voting member of the combined model may bear additional fruit.

Implementation Plan

Each of the four models had their predictions on the test dataset built and entered into the Kaggle competition. The final results for each model are listed below.

1. Combined – 0.79658
2. XGBoost – 0.79448
3. Logistic Regression – 0.79191
4. Random Forest – 0.78723

As expected, the combined model proved to be the most accurate in the competition. Surprisingly during our initial model testing, the XGBoost model performed the worst. However, during the competition it performed the best out of the three individual models.

Upon completion of the competition, the results will be followed up on for additional learning opportunities.

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.

10 Questions from the Audience

1. Is the model's accuracy score high enough to be reliable?
2. When will the model be implemented?
3. What additional steps can be taken to boost the model's accuracy?
4. How will we know if the model is working as intended?
5. What additional information may be useful to the model in the future that we should begin tracking now?
6. Which feature(s) provided the greatest predictability into the model?
7. Does any bias have the potential to exist in the model?
8. How do these models rank among the other models created for this Kaggle competition?
9. Are there any follow up steps after the competition concludes?
10. Would investigating any other model types be beneficial?

References

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