THE SPACE CADETS

Carson Wallace
Treyson Grange
Thadeus Ferney

Spaceship Titanic

Predicting which of 13,000 passengers on board were transported by the anomaly

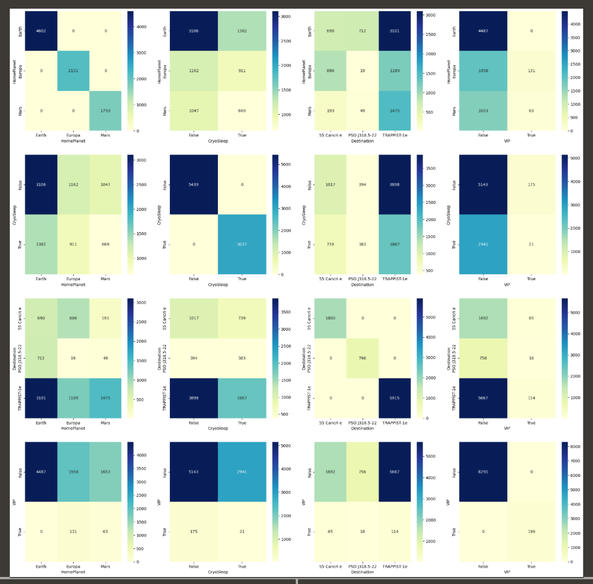
UtahState University

01. Feature Engineering

Initial Features: PassengerID, HomePlanet, CryoSleep, Cabin, Destination, Age, VIP, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck, Name, and Transported
Target Column: Transported
Feature processing:

- First we split PassengerID into a GroupID and an IndividualID
- Cabin could be split into the a Deck, Side, and CabinNumber

Dropped Columns: Name, PassengerId, Cabin



02. Data Imputation

Some columns could be used to indicate the value of other columns. Using this, we were able to fill in some of the unknown values with their correct data.

- First, no passengers under 18 are VIPs
- Also no passengers from Earth and VIPs
- Passengers in the same group will have the same homeplanet
- Those in the same group will also be in the same Cabin group
- Passengers in CryoSleep will not spend money any where on board and those spending money are not in CryoSleep

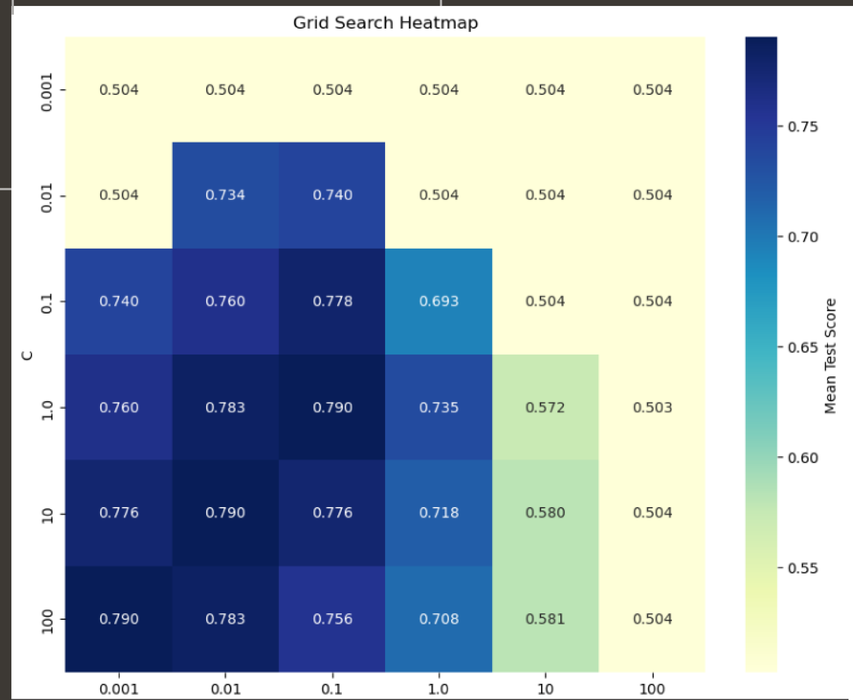
03. Transformer

First, of the available and imputed features, Name, PassengerId, and Cabin are dropped. PassengerId and Cabin were split into more useful features. Name is clearly unrelated to transported.

The remaining features were now only numeric and categorical. The numeric features were scaled. Any remaining values were filled with the most frequent value in each column.

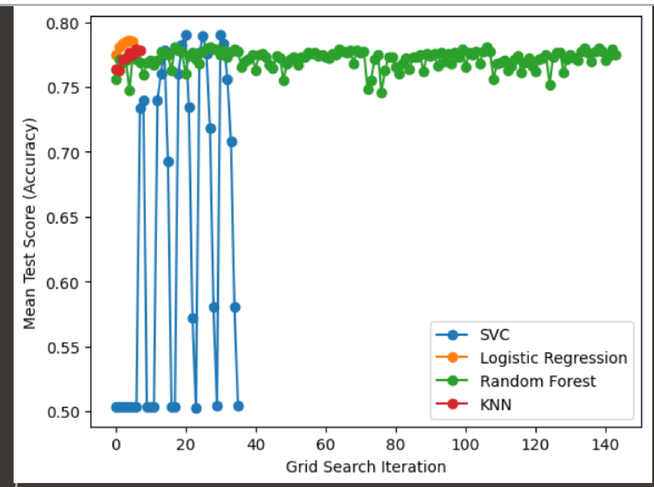
04. Model Testing

Multiple models were tuned and compared. These models included an SVC, logistic regressor, random forest, and KNN. To compare these results, we performed a grid search on each one. Once the best hyperparameters were chosen, we then tried creating an ensemble with these for models. A stacking and voting ensemble were tested. For the voting ensemble, we performed a search on a variety of weights. Finally, for the stacking ensemble, another logistic regressor was used as the meta classifier. Another grid search was performed to tune the hyperparameters of this model.



05. Results

Model	Cross Validation Score (cv=5)
DummyClassifier	0.5036
KNN	0.7789
RandomForest	0.7806
LogisticRegressor	0.7856
SVC	0.7903
VotingEnsemble	0.7958
StackingEnsemble	0.7949



06. Evaluation and Future Work

We took our best-performing single model, the SVC, and the Voting ensemble, and submitted both to the competition. Overall the SVC ended up being the better of the two despite the slightly lower validation accuracy. We finished with a 0.8073 testing accuracy on the public test set.

To continue improving, better data imputation could further increase accuracy. Also creating an ensemble that better averages out its faults could provide much better results.