CS 441 Applied Machine Learning
Spaceship Titanic Kaggle Challenge

Louis Sungwoo Cho

University of Illinois at Urbana-Champaign

Department of Civil and Environmental Engineering
(Transportation Engineering)

Department of Computer Science

Professor: Derek Hoiem

May 3rd, 2023

Approach

First the training and testing dataset are analyzed. If categorical data are null, then they are filled with "False" or "NA." For continuous data, the null values are filled with the median of the data. Standard Scaler was used to scale both the training and testing data. Encoded features were selected for feature engineering. The target of the data is the "Transported" column. Support Vector Machine will be used for dataset training and testing because this model provided the most optimal accuracy score out of the models that were used for testing. The Randomized Search CV will be used to determine the most optimal parameters for C and kernel while setting the gamma to 'auto'. Cross validation will be set to 5 with 5 iterations. Using the optimal C value and optimal kernel, the dataset will be trained and the tested ones will be stored to predictions.

Implementation Details

Libraries such as Numpy, Pandas, Seaborn, Matplotlib, SelectKBest, and chi2 were used for data visualization and feature engineering. For the machine learning libraries, sklearn's LogisticRegression, RandomForestClassifier, SVC, KNeighborsClassifier, MultinomialNB, DecisionTreeClassifier, GradientBoostingClassifier, GridSearchCV, and RandomizedSearchCV were used. To measure the performance of each model, sklearn's accuracy_score, precision_score, recall_score, f1_score and confusion_score were imported. In this project, Standard Scaler was used to scale the training and testing datasets [1]. Most of the feature engineering part was pre-implemented thanks to a blog written by Bohao Ning [0]. According to the feature engineering and dataset cleaning part, missing values in each column had to be filled up. For categorical variables, null values were filled up with either "NA" or False. For continuous variables, Null values had to be filled with the median values. Further data cleaning process was done by binary classification of ages (i.e. passengers older than the age of 18 are classified as "Adults") [0]. Standard Scaler was used to scale the training and testing datasets.

Experiments

First a Logistic Regression Classifier model was implemented with the parameters C = 100, penalty of l1 with a liblinear solver and a random state of 52. The accuracy score was 78.378%, the precision score was 76.421%, the recall score was 82.688% and the f1 score was 79.431%. A Random Forest Classifier model was implemented with the parameters of max depth = 10 and a random state of 42. The accuracy score was 79.011%, the precision score was 77.79%, the recall score was 81.777% and the f1 score was 79.733%. A Support Vector Machine Classifier (SVC) model was implemented with the parameters of gamma set to 'auto'. The accuracy score was 78.896%, the precision score was 76.98%, the recall score was 83.03% and the f1 score was 79.89%. A K-Nearest Neighbors Classifier (KNN) model was implemented with the parameters of k set to 7. The accuracy score was 77.918%, the precision score was 77.815%, the recall score was 78.702% and the f1 score was 78.256%. A Multinomial Naive-Bayes Classifier model was implemented. The accuracy score was 75.791%, the precision score was 77.106%, the recall score was 74.032% and the f1 score was 75.537%. A Decision Tree Classifier model was implemented with the parameters of max depths set to 5, min samples split to 10 and random state set to 42. The accuracy score was 76.251%, the precision score was 76.036%, the recall score was 77.335% and the f1 score was 76.68%. To further optimize the machine learning model, grid search and randomized search were performed to determine the most optimal parameters for each machine learning model. Once the optimal model with its respective parameters were determined, the optimized model was then trained and tested to predict the transported passengers with respect to the passenger id. The data was saved as csv file for Kaggle submission with an accuracy score of 80.476%.

Discussion

The Support Vector Machine (SVM) implementation with gamma set to "auto," C = 20 using the rbf kernel gave the most optimal accuracy score of 80.476%. Randomized Search CV was performed to determine the most optimal parameters that needs to be used to perform the support vector machine model. In the future, data cleaning and feature engineering will need to be investigated further and other machine learning classification models like Gradient Boosted Decision Trees or Extreme Gradient Boosting (XGBoost) will be experimented to optimize the accuracy score.

Citations/Acknowledgements:

[0] https://www.kaggle.com/code/doggypiggy/spaceship-titanic-top-10-0-80617

[1] https://github.com/kumod007/Titanic-Spaceship-Kaggle-Competition-End-To-End-Project/blob/main/titanic-spaceship-competition-end-to-end-project%20(2).ipynb