Title: Titanic - Machine Learning from Disaster

Introduction (including problem definition and motivation)

We have chosen a Kaggle Competition dataset regarding the Titanic shipwreck. The Titanic was a historic disaster occurring the April of 1912, in which more than half of the passengers died. The 1997 “Titanic” disaster film sparked renewed international interest in this event, and it has been a cultural phenomenon and continues to be so. We analyzed who survived the shipwreck and what characteristics the survivors have in common.

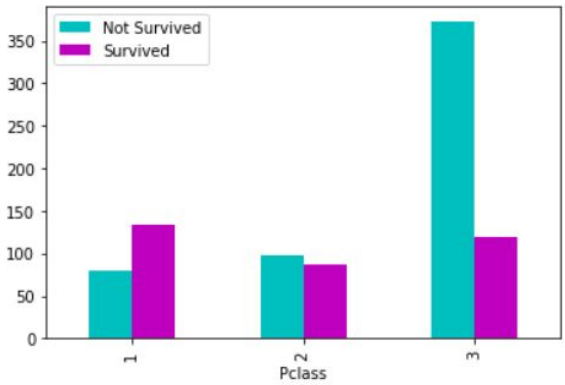
The target variable is “Survival”, a binary variable where 0=No and 1=Yes. The data provided by the competition includes 9 feature variables, which are: ticket class and number, sex, age, number of siblings and/or spouses, number of parents and/or children, the cost of the ticket, the cabin number and where the passenger embarked on the boat.

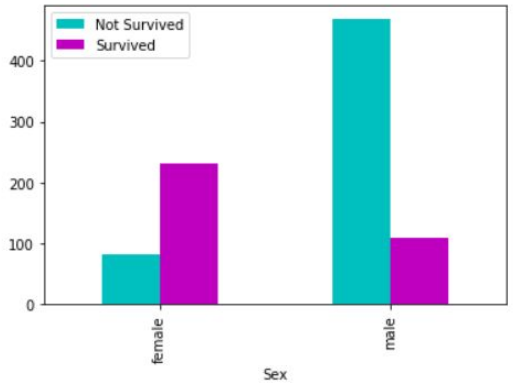
Proposed method and the idea behind it (e.g. why should the method work)

We used Python Jupyter notebooks to conduct this analysis. We utilized the pandas, numpy, SciPy, MatPlotlib, Seaborn, and sklearn packages.

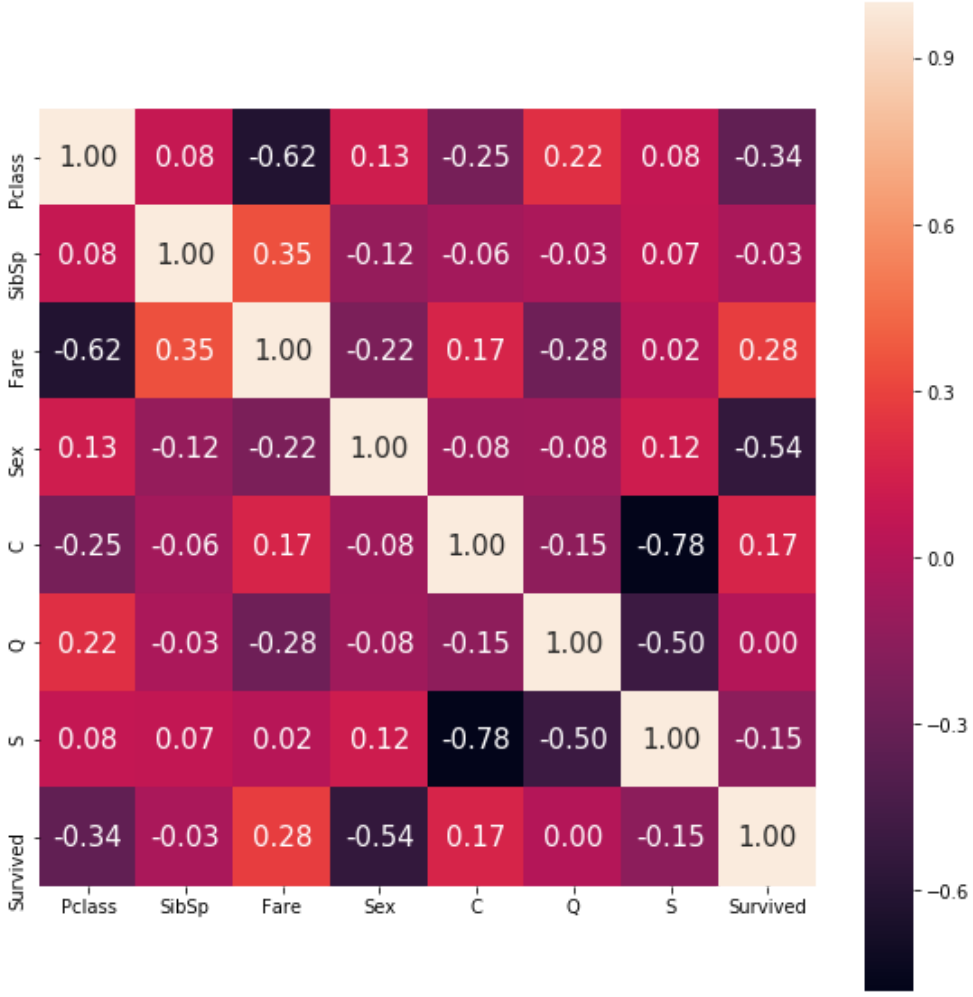
The data from kaggle is given as two datasets, train and a test csv. The test csv does not include the target feature. In order to check the final accuracy of the model a csv must be created from the final model that has PassengerId as the first column and the predicted class labels as the second column. This is imported into Kaggle, where a random sample is taken and compared to the actual results.

The first step in the method was to clean the data and perform exploratory data analysis. The EDA allows for understanding of the feature and target variables as well as how they relate to one another. Some visualisation of this EDA are histograms and a correlation matrix. The below graph show survived and not survived with the class of 1, 2 or 3 and the other graph shows Survived or not by gender. We can see that more females survived and there were more males on board. We can also see that the richer patrons had a higher percentage survive with the poorest having the most people on board and the most to not survive.





The correlation matrix shows that the feature with the highest correlation to Survived, is Gender.



This relates back to a very famous line from the “Titanic” movie where women and children were given first priority on lifeboats. This appears to have happened in real life and not just the movie; more females survived than males.

The next step of the analysis was to create a pipeline to find the most accurate model. The pipeline standardized the data and ran it through three models with many different parameter options. The pipeline used Cross Validation to find the accuracy of each model with the various parameters we listed. The models used were random forest, k-nearest neighbors (KNN), and support vector clustering (SVC). Linear regression and other regression models were not used because the target variable is discrete and not continuous. Due to the nature of the target, this is a classification problem. Logistic regression was not used, because the class labels are not linearly separable. We realized that if we used SVC, we could get a better result than logistic regression. We choose to include SVC in the pipeline and to leave out LR. Decisions trees were not used, because there is a similarity in how decision trees and random forest operate. We realized that random forest would give us a better chance of getting an accurate model than decision trees. Therefore, we choose to include random forest in the pipeline and to not include decision tree. Perceptron was also not used, as it tends to work better on problem where there is a way to linearly separate the classes. This is not possible on this particular dataset.

Experimental results and analysis (e.g. why the results look like this)

Support Vector Clustering was the model that provided the highest accuracy from the grid search. The parameters that produces this output was C=4, gamma =0.1 and kernel = 'rbf' with all else set to the default. The accuracy came in at 82%. The clustering methods group samples or observations by how alike they are. The features used in the model are that of groups. For example, the individuals aboard the titanic are either male or female, old or young, higher class or lower class, etc. Using the groups as a way to think about how the individuals end up provides a good algorithm. SVC is able to separate classes by introducing additional features, thus transforming lower dimensional spaces to higher dimensions. SVC can handle the complexity that many features bring. SVC also had the highest accuracy when predicting the test data. SVC had a 79.9% accuracy when submitted to Kaggle.

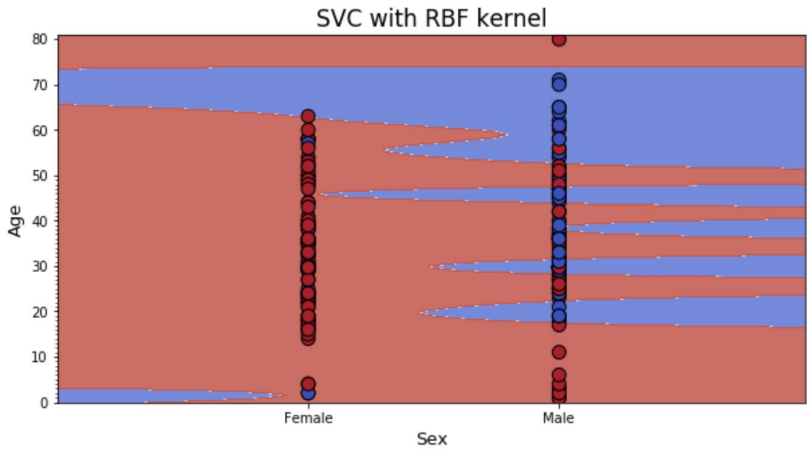
K nearest Neighbors had the second highest accuracy from the grid search with 81%. The parameters for this model are algorithm=brute, n_neighbors=10 and all else set to default. KNN also uses the idea of grouping the observation based on what features the individual has. The accuracy for the KNN model on the kaggle test dataset was at 77%. This is slightly lower than that from the grid search output.

Random forest was the model with the lowest accuracy of the three on the train set. The parameters that the pipeline found to be the best are minimum leaf samples at 2, minimum split samples at 10, n_estimators at 40 and all else set to default. Random forest is a popular classification algorithm, which worked particularly well on this data set. Random forest can handle large data sets, high dimensionality, maintain proportionality accuracy of data. A common problem when using a decision tree on a

dataset is overfitting the model. A random forest has many trees and therefore overfitting the model on a tree is a less likely issue to arise and the results may be more accurate. The benefits of random forest worked well with this dataset. The accuracy provided by the grid search for Random Forest was at 81%. However, the accuracy on the test data was slightly lower. Random Forest had 77% accuracy.

Conclusion

The Titanic Dataset is a great tool for practice with machine learning algorithms. This kaggle competition provided us with the ability to test our knowledge and understanding of how to work with data to make predictions. We were able to work as a cohesive team and compare our results to the other kaggle submissions. This allowed us to see where we ranked and then discuss how to better our end result. Our best accuracy on the test set landed us in the top 17% of over 10,000 submissions with accuracy on the testing set at approximately 80%. The best model of SVC has a plot to the right. In this plot, red is survived and blue is not survived. This shows that SVC is predicting that more older folks did not survive than younger folks and that more men do not survive. This means that children were given priority in the life rafts, as well as women.



For this data it can be said that those that are most likely to survive does follow the same pattern that anyone who is familiar with the movie might guess. Women, children and those of upper class are more likely to have survived. The common knowledge of the historic situation turns out to be based in fact. The model with the highest accuracy did utilize gender, age and class. It also used embarked location, number of family members the individual is traveling with as well as the cost of the ticket. The best model for determining the survival of titanic travelers is a random forest classifier.

It is interesting to note that the ranking of best accuracy from the grid search matched the ranking of best accuracy from the prediction for the test dataset. Not only that but the differences between adjacent models were quite similar between the rankings. This shows that the train and test datasets are similar in nature. The models held up in the test data, but where slightly less accurate than when used on the train dataset.