Jason Chan

**Analysis Report**

Out of all the machine learning models in data science, we’ve chosen three of the most popular models. We use linear regression, more specifically logistic regression as our first model. Logistic regression is a generalised linear model that works well with inspecting the relationship between a dependent variable and one or more independent variables. Moreover, we’ve chosen this due to the fact our target value was a categorical value. Although it is not intended to be the most accurate titanic model, we’ve chosen the decision tree model to best visualize and understand the classification trees for the titanic dataset. Lastly, we’ve chosen the random forest model since it is one of the most flexible and easy to use algorithms. Random forest is a supervised learning algorithm which works by building multiple decision trees and merging them to get a more accurate and stable prediction. A very huge advantage of this model is that it can be used for both classification and regression models which form the majority of current machine learning systems.

We determined the accuracy of each model multiple ways. We would use the algorithm we’ve chosen and calculate the train and test split results of that respected algorithm. The next two methods would be methods that would essentially try to improve the model’s accuracy. The first way would be using cross validation on the current algorithm in usage. The second way would be using an ensemble method known as the BaggingClassifier. Afterwards, we would provide visualization for each model.

For example, our first selected model is the LogisticRegression model. We would predict the values for LogisticRegression. Then using its respected algorithm, we calculate the accuracy score of this model which would come out to be 0.78475. Afterwards, apply cross validation to it and show the results which would be 0.79351. And finally, we would use the ensemble method to further determine any improvements in the model’s accuracy. Using the BaggingClassifier for our chosen method would end up being 0.78027. So what we have determined so far is that the cross validation method did indeed improve our model’s accuracy, but the ensemble method did not. In terms of visualization, we provide a confusion matrix for this model.

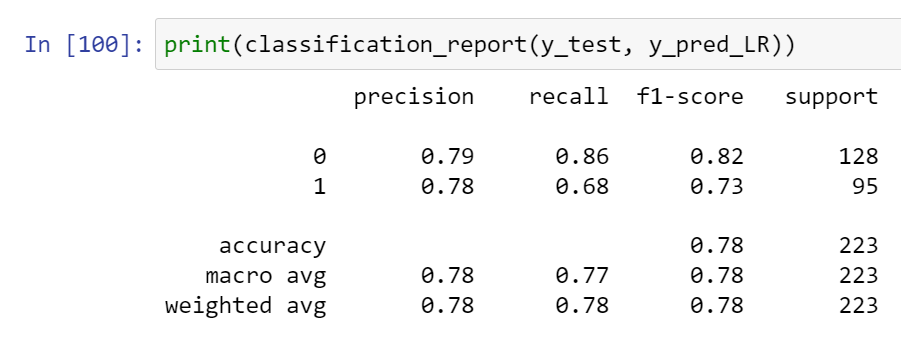
Our second selected model is the DecisionTreeClassifier model. We would predict the values for DecisionTreeClassifier. With a max depth of 5, we would then calculate the accuracy score of this model which would come out to be 0.78027. Afterwards, apply cross validation to it and show the results which would be 0.78906. Using the BaggingClassifier for our chosen method would end up being 0.77130. So what we have determined so far is that only the cross validation method improved our model’s accuracy whereas the ensemble method did not. As for visualization, we provide a decision tree for this model.

Our third selected model is the RandomForestClassifier model. We would predict the values for RandomForestClassifier. With a max depth of 5 and n\_estimators of 10, we would then calculate the accuracy score of this model which would come out to be 0.80269. Afterwards, apply cross validation to it and show the results which would be 0.81819. Using the BaggingClassifier for our chosen method would end up being 0.78027. So what we have determined so far shouldn’t even be a surprise at this point. Yet again, only the cross validation method was the one that impacted the model’s accuracy.. For visualization, we provide a confusion matrix for this model.

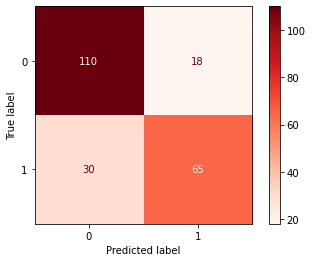
In conclusion, we see that the RandomForestClassifier model was the most accurate with an accuracy score of 0.80269. In second place, we have the LogisticRegression model with an accuracy score of 0.78475. Last but not least, we have our DecisionTreeClassifier model with an accuracy score of 0.78027. We can improve these models’ accuracies by building on the data and/or even enter missing age values as something more closer to the range of the mean values rather than just 100.

LogisticRegression Model

* Classification Report

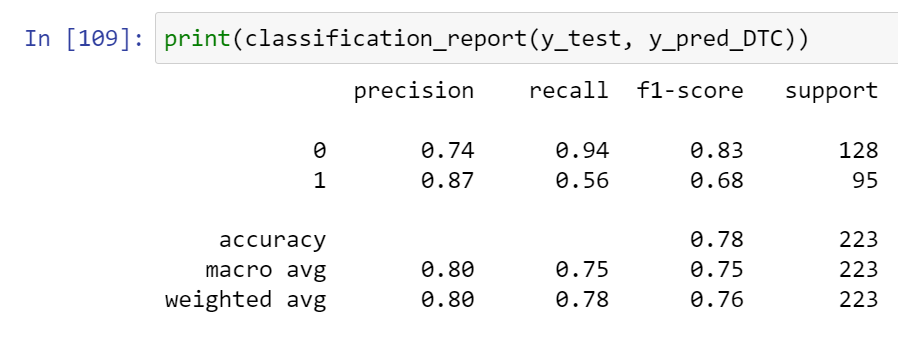


* Visualization (Confusion Matrix)

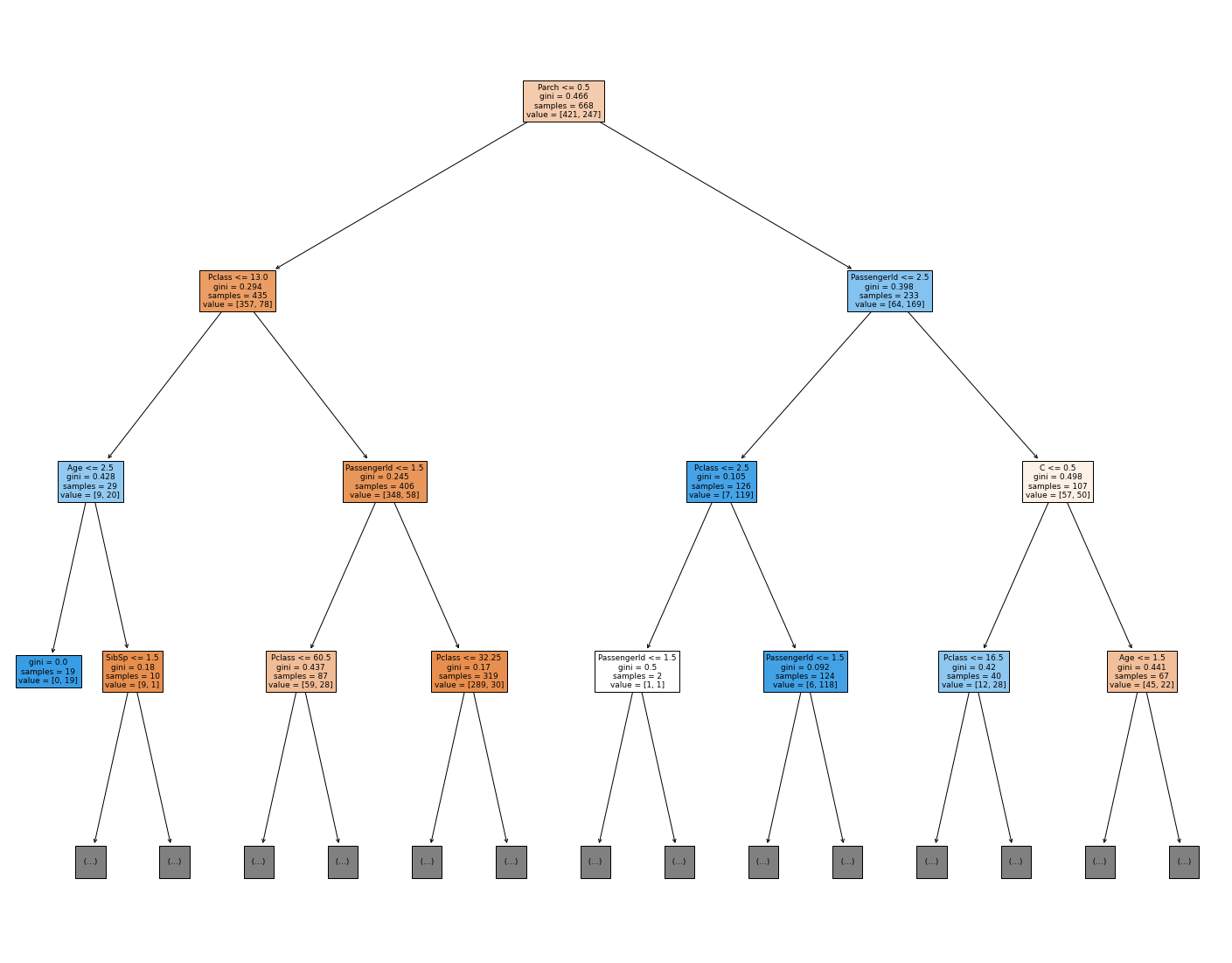


DecisionTreeClassifier Model

* Classification Report

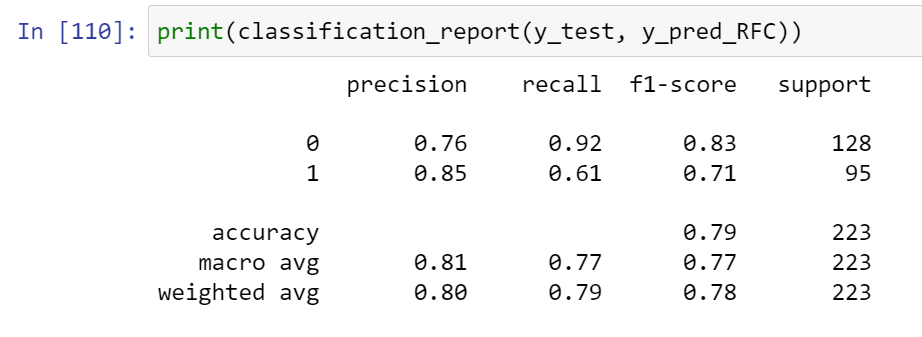


* Visualization (Decision Tree)



RandomForestClassifier Model

* Classification Report



* Visualization (Confusion Matrix)

