

Assignment 4s

Applied Machine Learning

We will develop a classification pipeline to predict if a passenger from the Titanic survived or not. Go to [Kaggle page for Titanic data](#) and download the training and testing data sets. (Verification: 891 data points for training and 418 data points for testing dataset files)

1. [70 pts] Preprocess the data, impute missing values as you see fit, and remove features that seem useless.
2. [30 pts] Submit your predictions to Kaggle for the test dataset and report your accuracy in your submission. (You will need a Kaggle account – use a dummy email address if you prefer to protect your school email address, etc.) For your reference, I achieved 79% using my preprocessing pipeline and a Random Forest classifier. This is not the best, as in Kaggle there are better results. Kaggle also has some results with 100% accuracy, which cannot be taken as honest submissions in my opinion.

Use the following code to export your predictions into a .csv format Kaggle will accept:

```
def save_preds(_fn, _y_pred, _df):
    import csv
    with open(_fn, 'w') as fout:
        writer = csv.writer(fout, delimiter=',', lineterminator='\n')
        writer.writerow(['PassengerId', 'Survived'])
        for yid, ypred in zip(_df['PassengerId'], _y_pred):
            writer.writerow([yid, ypred])

save_preds('predictions_erhan.csv', y_pred, df_test_org)
```

Note that in the above code the dataset `_df` has to include the `'PassengerId'`, which must not be used for the classification model. Kaggle needs it to compute a performance score.

