Data-scientist-exercise01: Report

I started off this exercise by flattening the database by continually create a new table every time I would perform a join. (flattenDatabase2TableVersion1.sql) I created a more efficient solution by simply altering the table given. (flattenDatabase2TableVersion2.sql) This resulted in the creation of flattenedRecords.csv. For the exploratory analysis step, I utilized R where I discovered information about the data. (abstractFlattened.R) In order to create test, train, and validation data sets, I originally explored R, but soon after switched to Python. Next, I developed 3 models to predict whether individuals, based on the census variables provided, make over $50,000/year. (test.ipynb)

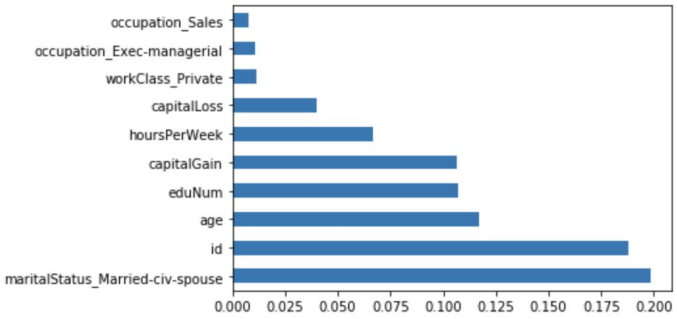
1. General Decision Tree Classifier

```
Accuracy: 0.8116490940730884

Report :
              precision    recall  f1-score   support

           0       0.88      0.87      0.88      7430
           1       0.60      0.63      0.61      2339

    accuracy                           0.81      9769
   macro avg       0.74      0.75      0.74      9769
weighted avg       0.81      0.81      0.81      9769
```



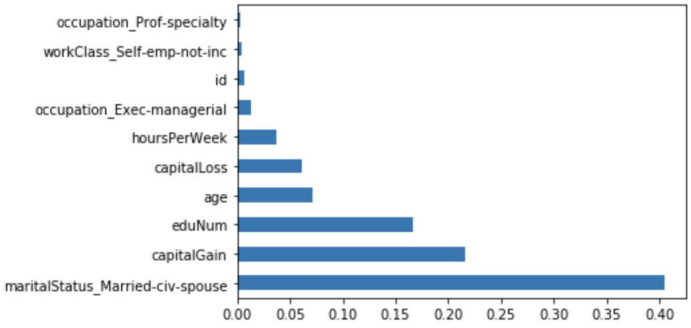
2. Improved Decision Tree Classifier

```
Accuracy: 0.8591462790459617

Report :
              precision    recall  f1-score   support

           0       0.88      0.95      0.91      7430
           1       0.78      0.58      0.66      2339

    accuracy                           0.86      9769
   macro avg       0.83      0.76      0.79      9769
weighted avg       0.85      0.86      0.85      9769
```



3. Random Forest Classifier

```
Accuracy: 0.8531067663015662

Report :
              precision    recall  f1-score   support

           0       0.89      0.93      0.91      7430
           1       0.73      0.62      0.67      2339

    accuracy                           0.85      9769
   macro avg       0.81      0.77      0.79      9769
weighted avg       0.85      0.85      0.85      9769
```