Predicting bank campaign result

Piotr Żebrowski

Step 1. Exploratory data analysis (DataPreparationScript file)

- Data has 2 columns which have correlation coefficient of >0.9 ('nr.employed', ‘euribor3m')

- There are 250 rows which have no values (‘cons.price.idx’ column), which is less than 1% of the whole data set

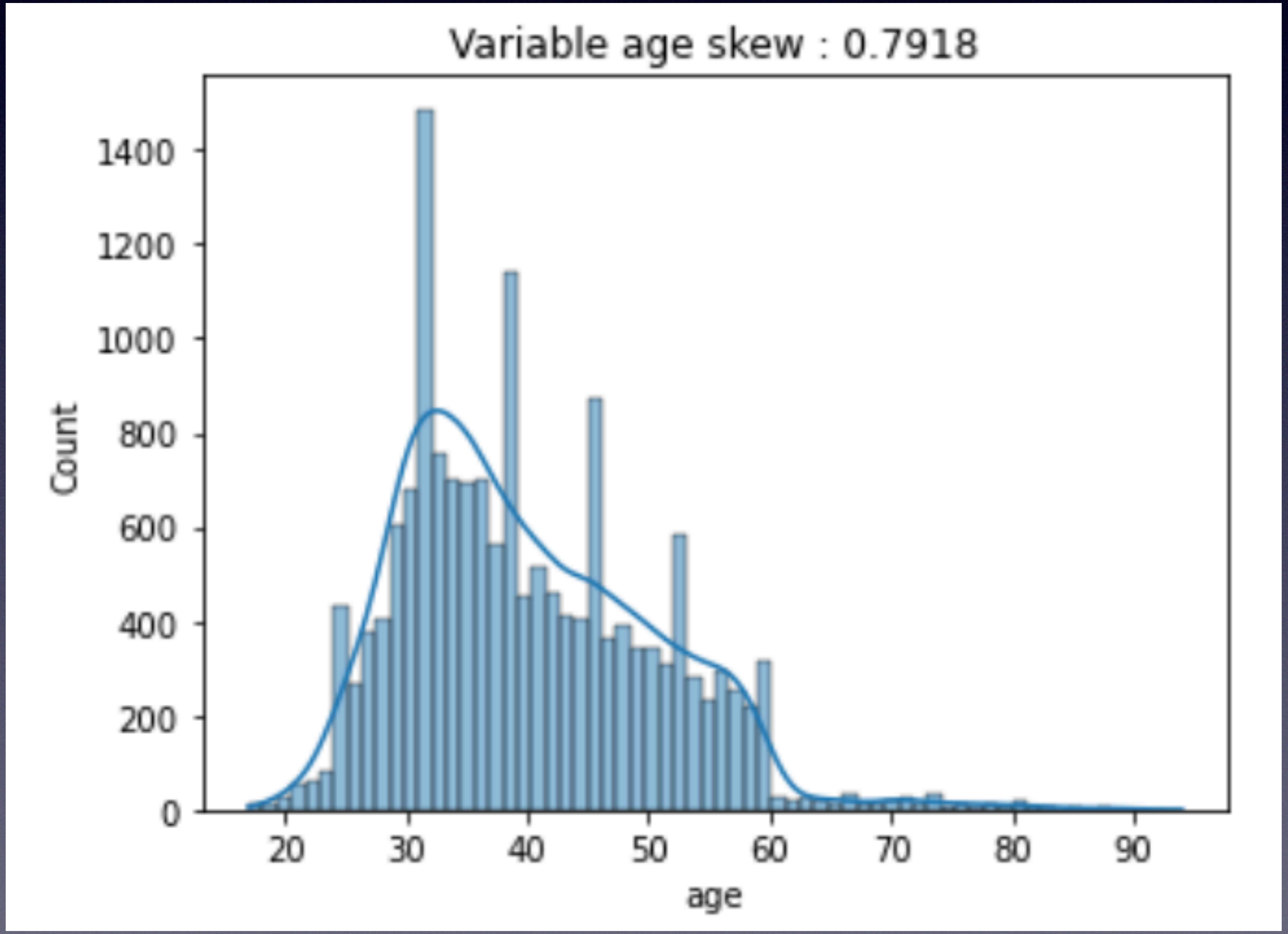
- 24712 out of 41118 clients were contacted

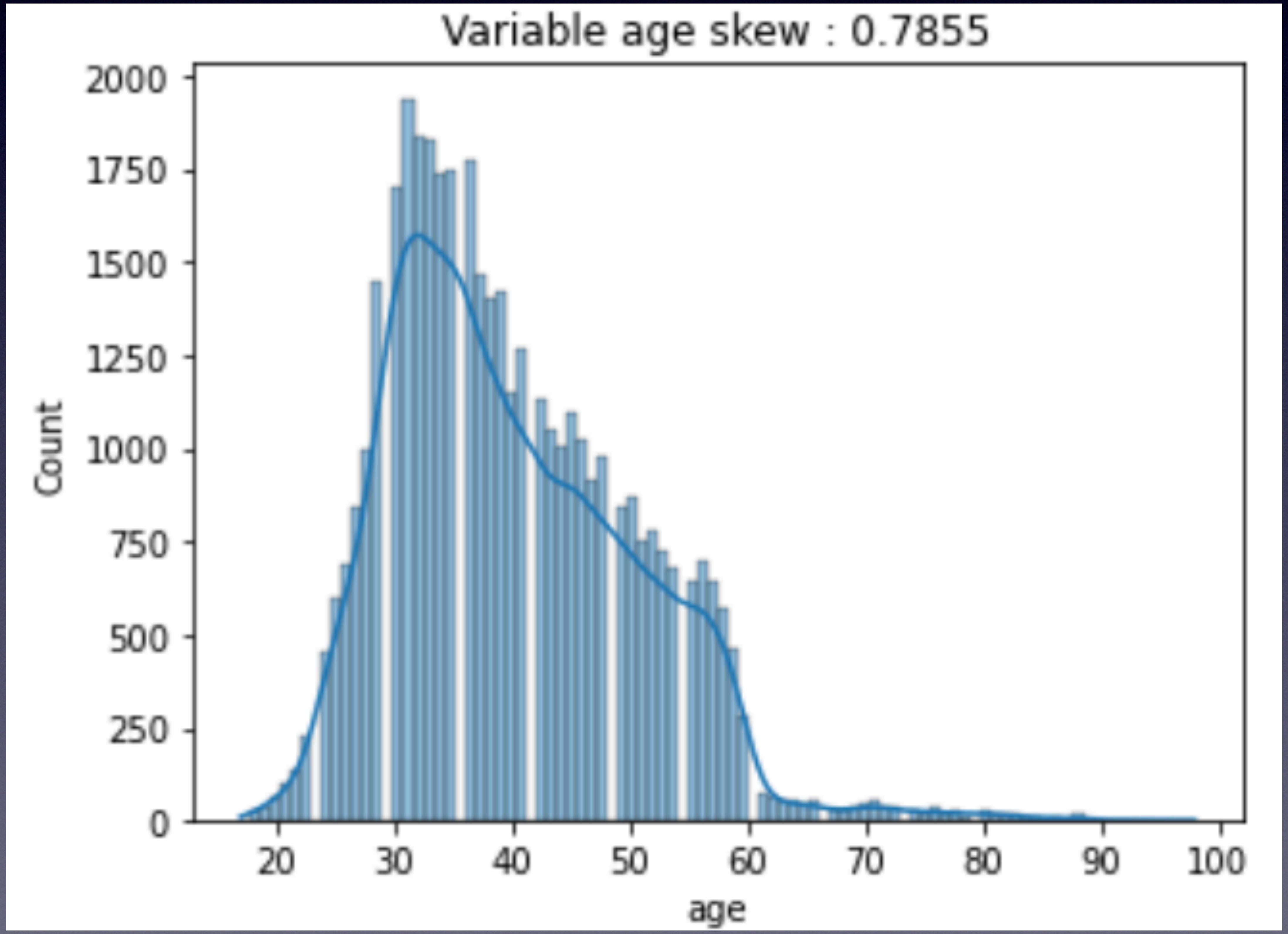
- Previous campaign effectiveness was 51.54%

Step 2. Defining 3 data sets

- Data set 1: Whole data - 250 NaN rows - 2 columns with high correlation
- Data set 2: subset of data set 1 with campaign_group-only rows
- Data set 3: data set 1 with category columns reduced to 0/1 values (deleted divorced marital status etc.)
- Each data set has it’s own Jupyter Notebook file for analysis

Distribution of age column in campaign group (left) & whole dataset (right)



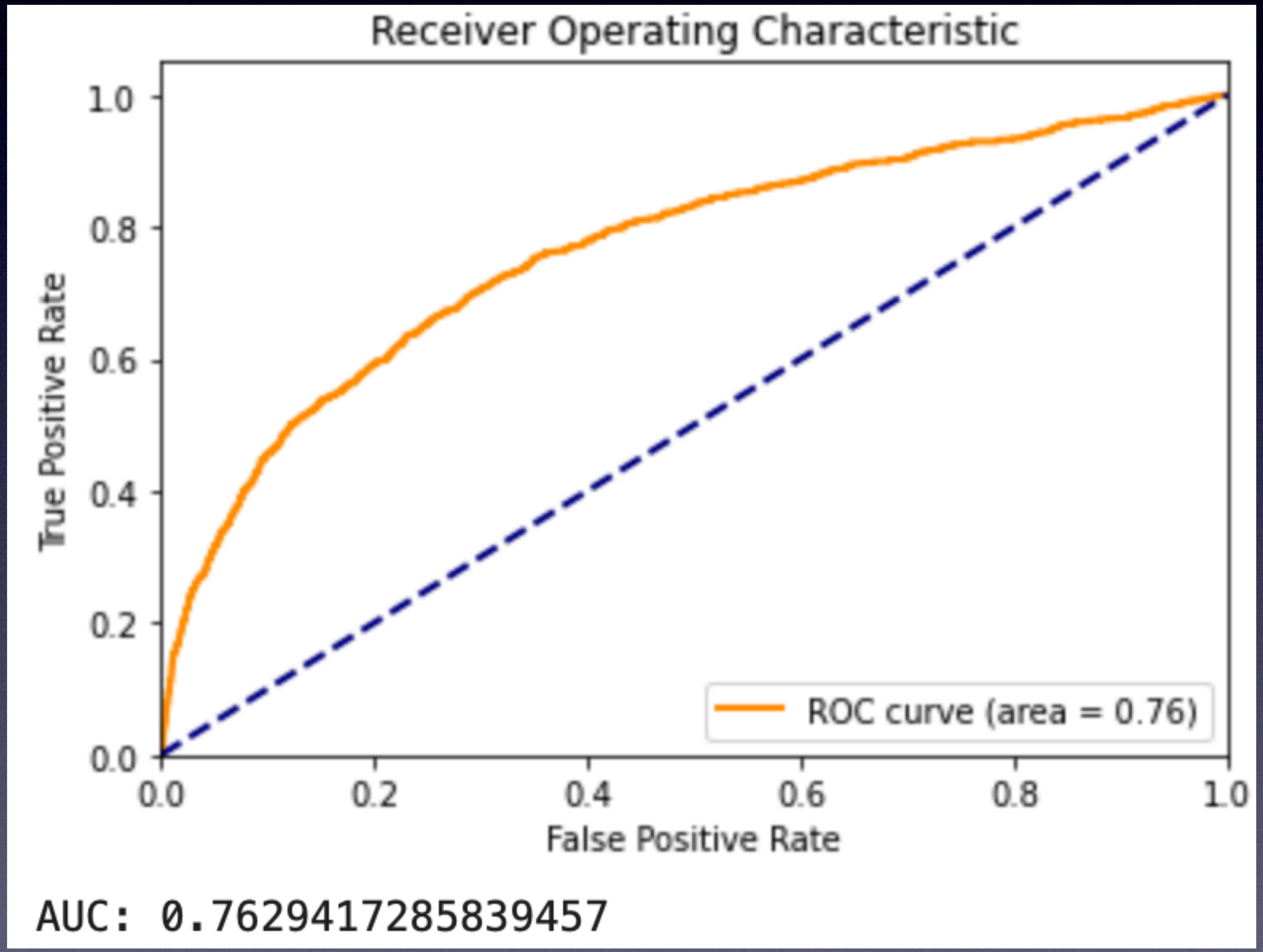


(More charts in Jupyter Notebook files)

Analysis of each dataset

- For all Data sets, the best models are logistic regression, random forest classification and XGBoost (86% to over 88% accuracy).
- The following charts will reference the logistic regression model
- For all data sets, the worst model is decision tree classifier (about 82% accuracy

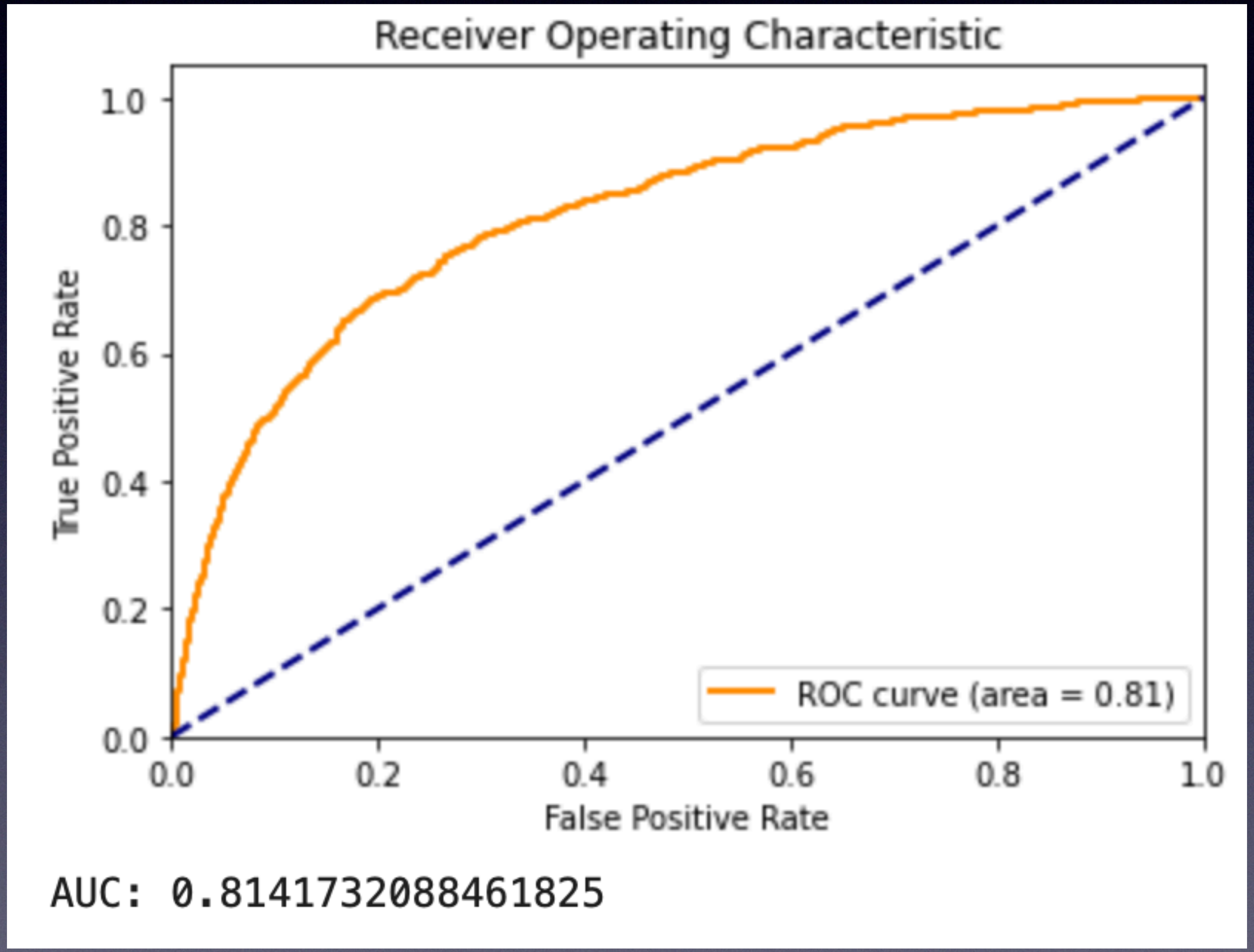
ROC & AUC of model on data set 1 (all data)



```
AUC: 0.7629417285839457
```

AUC of 0.76 is considered a decent score.

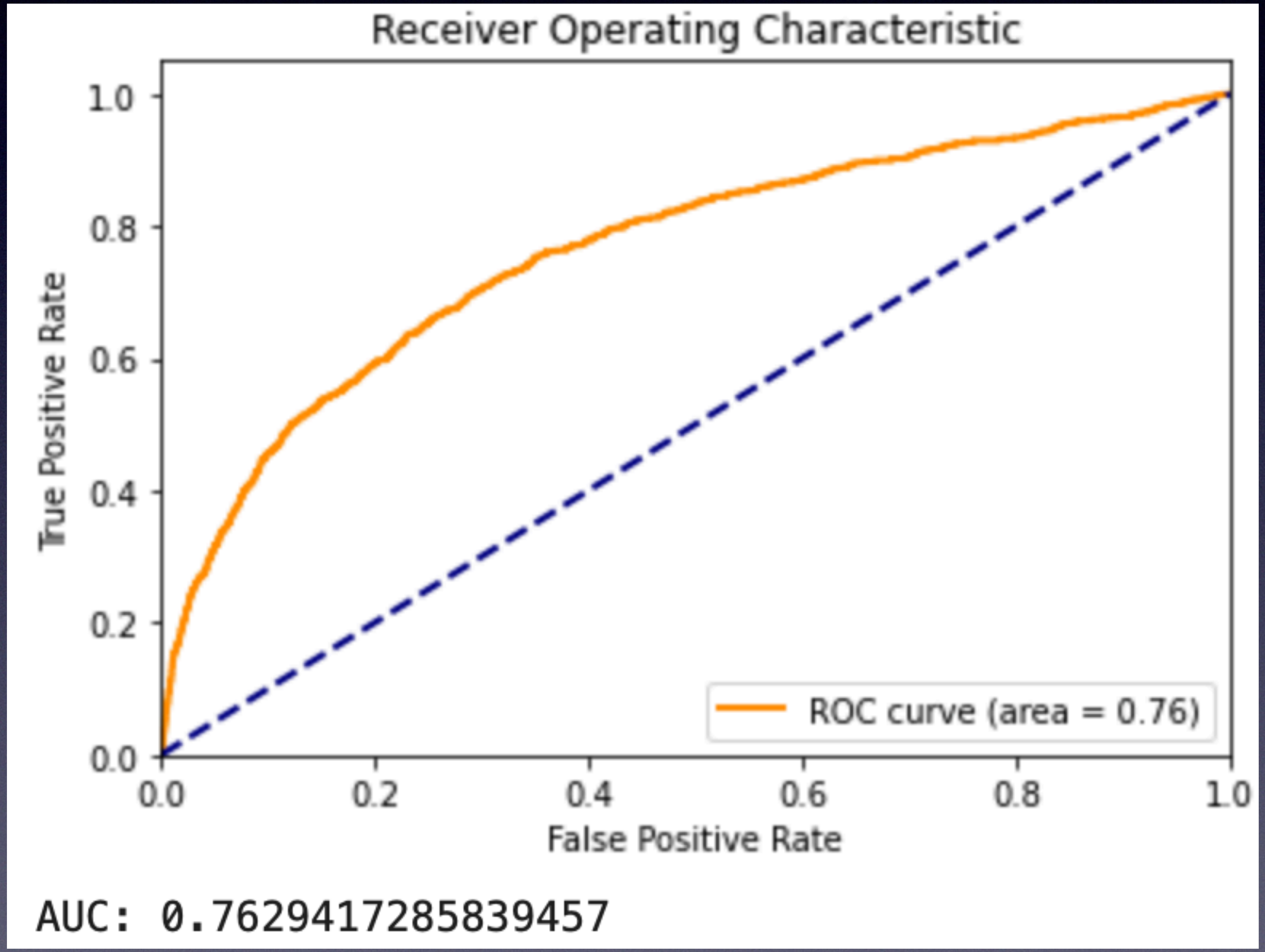
ROC & AUC of model on data set 2 (campaign group)



```
AUC: 0.8141732088461825
```

AUC of 0.81 is considered a good score. Campaign group data help with prediction.

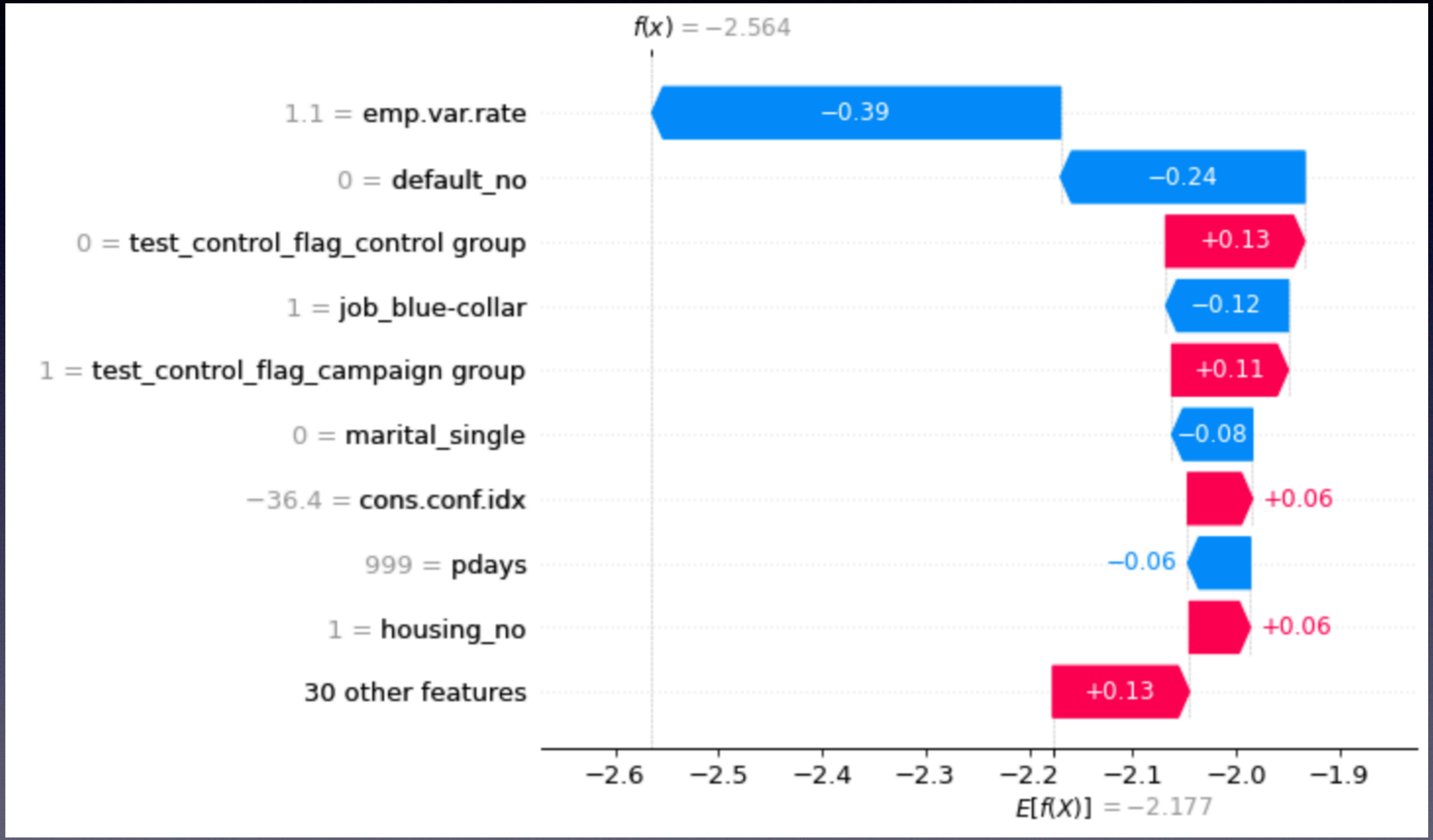
ROC & AUC of model on data set 3 (reduced data)



```
AUC: 0.7629417285839457
```

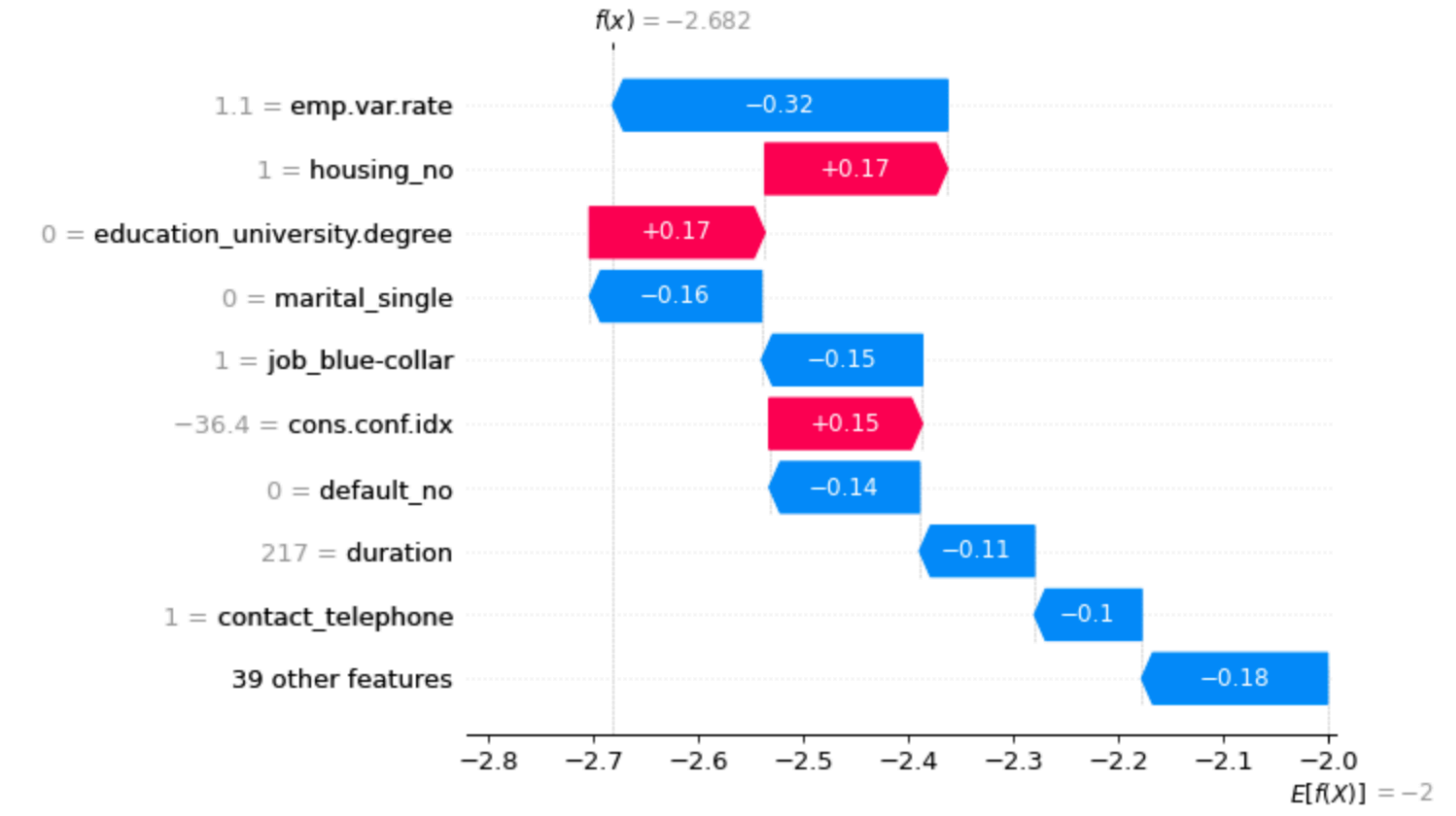
This AUC is almost the same as the AUC for the data set 1. Reduction of columns didn’t help much.

Explaining model on data set 1



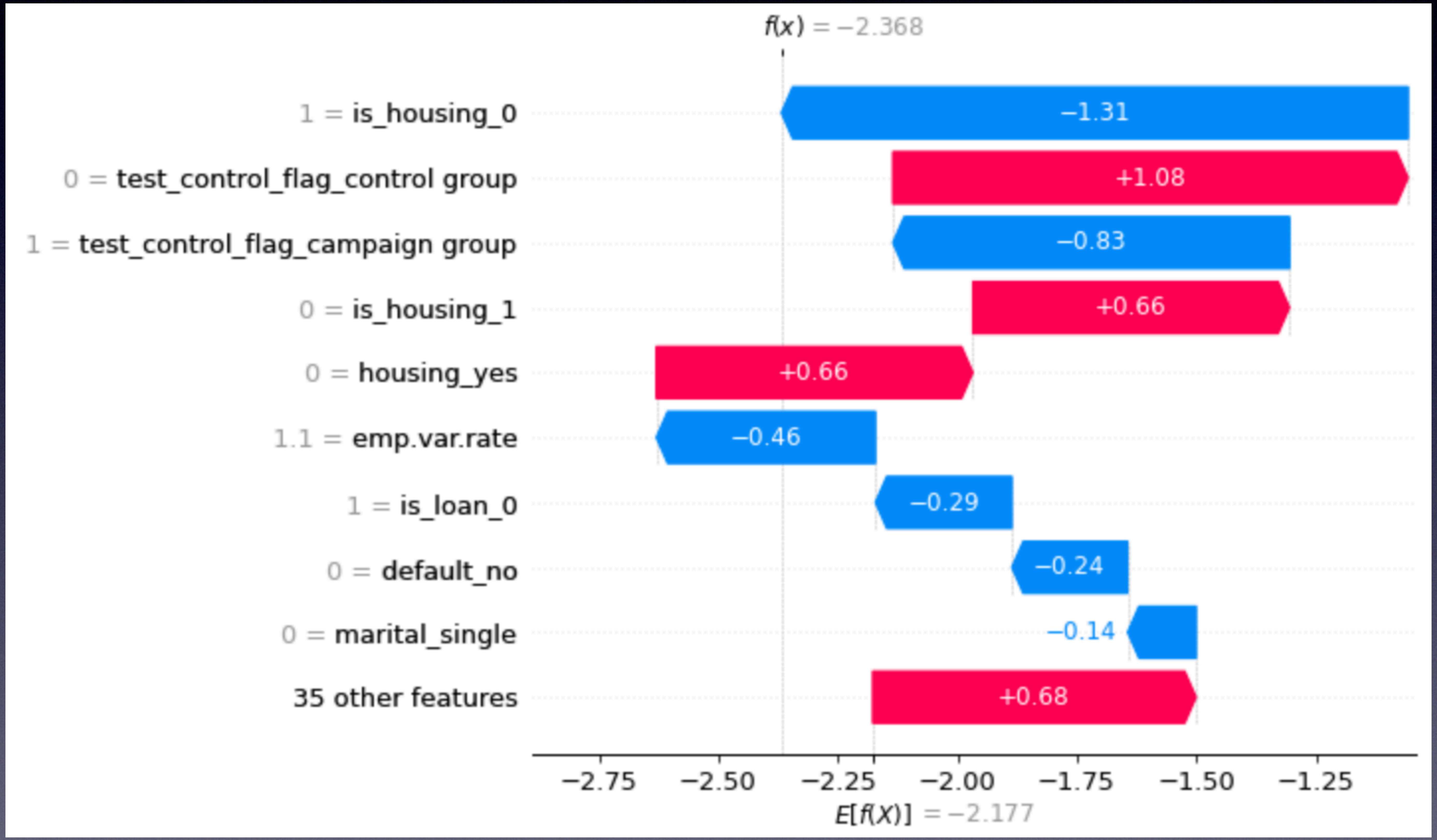
SHAP Waterfall plot for an example instance of index = 2. Most important feature is emp.var.rate

Explaining model on data set 2



SHAP Waterfall plot for an example instance of index = 2

Explaining model on data set 3



SHAP Waterfall plot for an example instance of index = 2. Different variables play the most role.

Possible improvements

- More data preprocessing and exploration (e.g. exploring the distribution of variables to understand class imbalances)
- Collection of more data with more ‘y’ = 1 cases
- Domain expertise (consultation with domain experts who might provide insights on which features are most relevant and why)
- Hyperparameter tuning of the model (e.g. choosing the best solver)