ML Zoomcamp Midterm Project

Dataset: Adults

https://archive.ics.uci.edu/dataset/2/adult (downloaded and saved as file adult_income.csv)

Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Problem: predict in historical perspective people's wellness opportunities based on their life characteristics. Further can be adapted to current times to see in what way society wellness factors changed within 30 years.

- 14 features
- 48,842 rows.

Data preprocessing

1. Coding income.
2. Replace null values.
3. DictVectorizer.
4. StandardScaler.

Long story short: best model is XgBoost with ROC AUC = 93.37%

Model	Accuracy	Precision	Recall	ROC AUC
Logistic regression	85.77	74.83	60.68	91.1
Decision Tree	85.86	78.94	55.44	90.72
Random Forest	86.77	80.17	59.09	92.24
XgBoost	88.03	80	66.35	93.37

Decision Trees

	Before tuning, %	After tuning, %	Progress, p.p.
accuracy	81.44	85.86	4.42 ↑
precision	60.73	78.94	18.21 ↑
recall	62.53	55.44	-7.09 ↓
roc_auc	75.1	90.72	15.62 ↑

Optimal parameters:

- Max depth: 10
- Min leaves: 8

Random Forest

	Before tuning, %	After tuning, %	progress
accuracy	84.84	86.77	1.93 ↑
precision	68.75	80.17	11.42 ↑
recall	66.65	59.09	-7.56 ↓
roc_auc	88.54	92.24	3.7 ↑

Optimal parameters:

- N-estimators: 80
- Max depth: 20
- Min Sample Leaf: 3

XGBoost

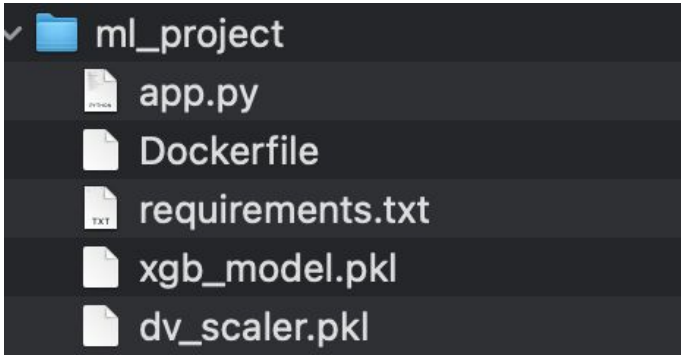
	Before tuning, %	After tuning, %	progress
accuracy	86.34	88.03	1.69 ↑
precision	75.66	80	4.34 ↑
recall	62.91	66.35	3.44 ↑
roc_auc	91.79	93.37	1.58 ↑

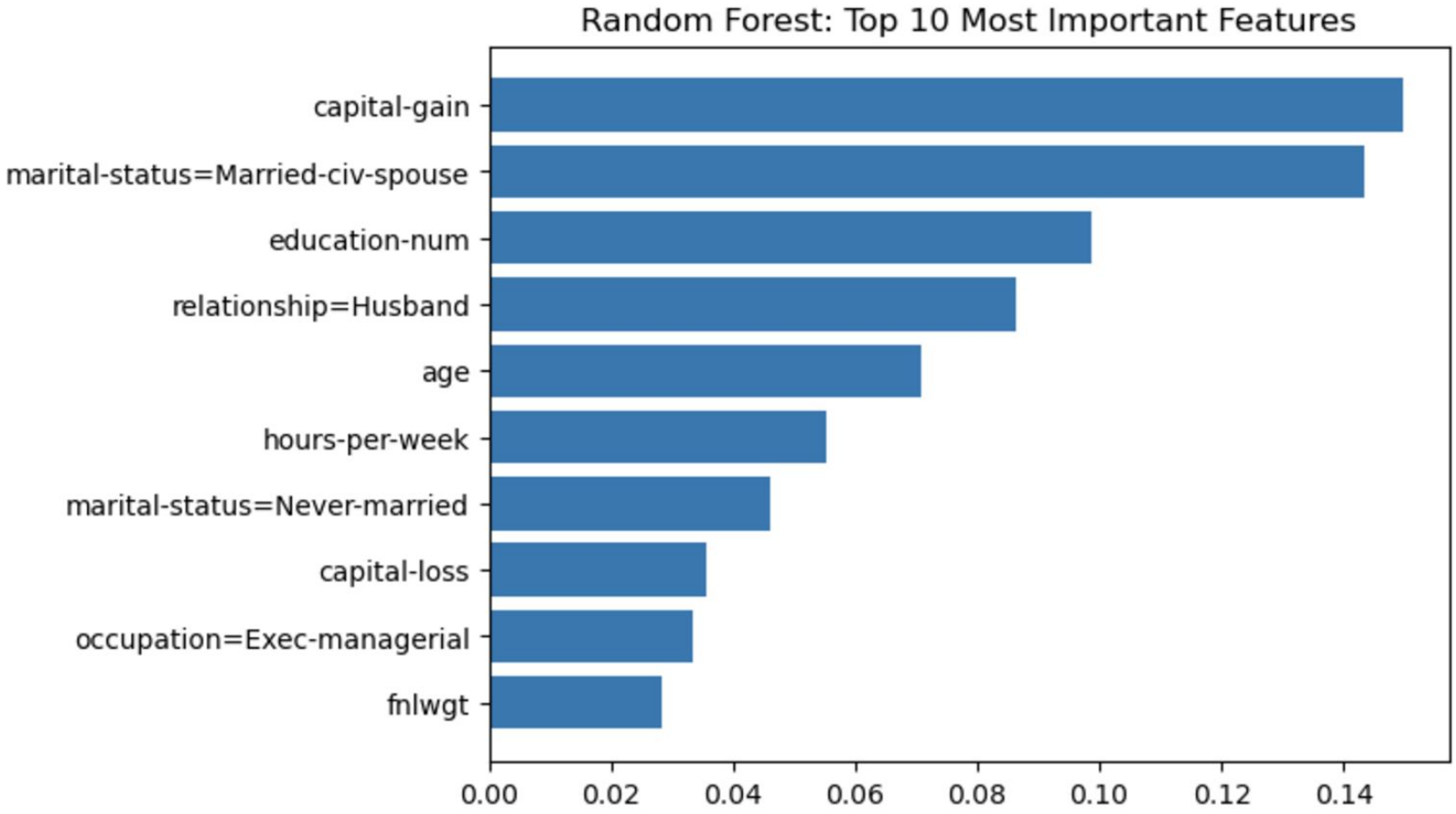
Optimal parameters:

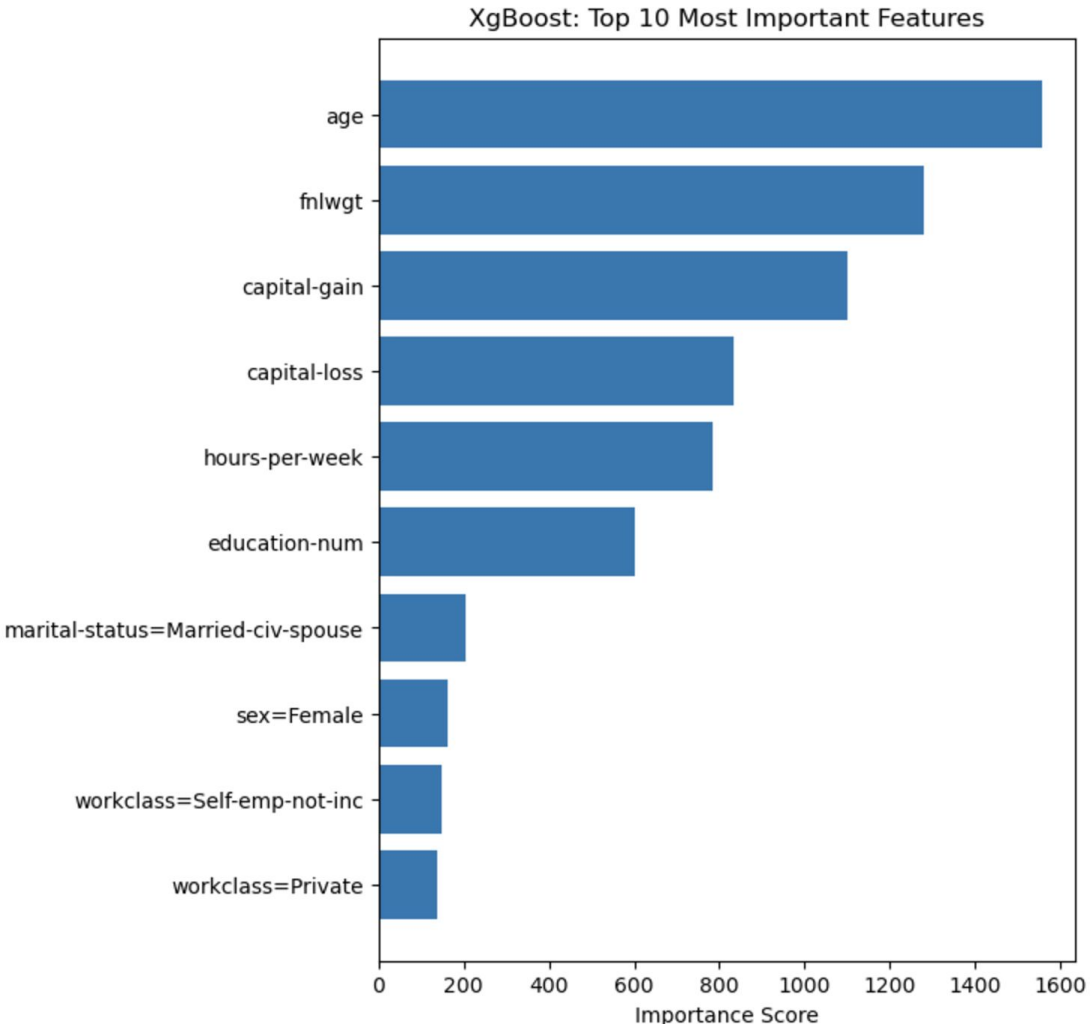
- eta: 0.05
- Max depth: 5
- Min Sample Leaf: 1
- Trees: 499

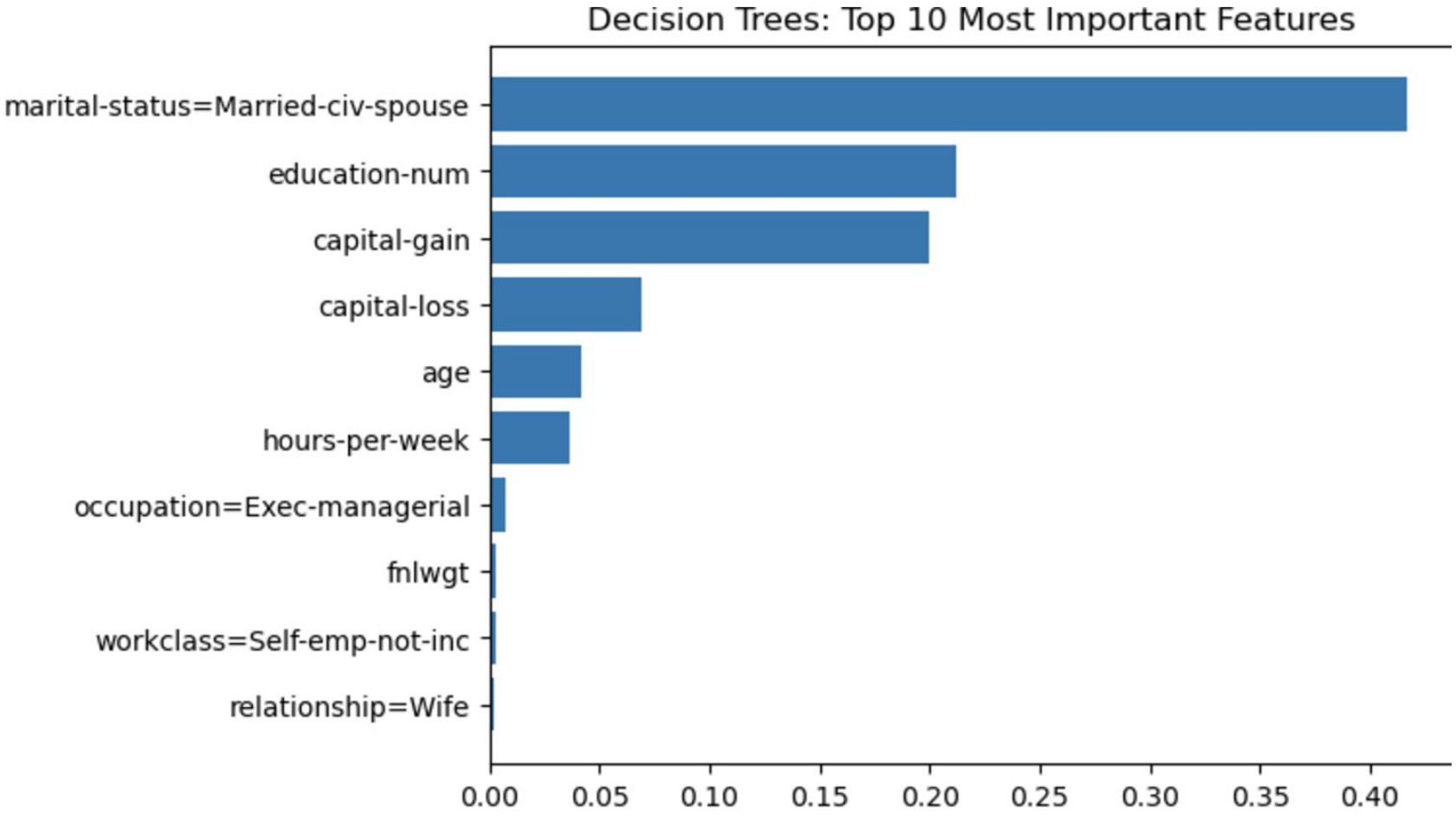
How to run project

1. Place in folder 5 files like on the screenshot.
2. Build image: docker build -t ml_project .
3. Run container: docker run -p 9696:9696 ml_project
4. Run python code from run_example.py.




Random Forest: Top 10 Most Important Features
capital-gain
marital-status=Married-civ-spouse
education-num
relationship=Husband
age
hours-per-week
marital-status=Never-married
capital-loss
occupation=Exec-managerial
fnlwgt
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


XgBoost: Top 10 Most Important Features
age
fnlwgt
capital-gain
capital-loss
hours-per-week
education-num
marital-status=Married-civ-spouse
sex=Female
workclass=Self-emp-not-inc
workclass=Private
0
200
400
600
800
1000
1200
1400
1600
Importance Score


Decision Trees: Top 10 Most Important Features
marital-status=Married-civ-spouse
education-num
capital-gain
capital-loss
age
hours-per-week
occupation=Exec-managerial
fnlwgt
workclass=Self-emp-not-inc
relationship=Wife
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40