

HW # 3: Money Brings Satisfaction But Not Happiness: Income Prediction

- Kaggle In-class
- Competition Link: <https://www.kaggle.com/t/47afc87ea9fc4e9abe59962c8e6419f8>
- Deadline: 11/19/2018 11:59 PM

HW # 3: Money Brings Satisfaction But Not Happiness: Income Prediction

- Task is to predict whether income exceeds \$50K/yr based on census data. It is a **binary classification** problem.
Note that in the train and test data, salary >50K is represented as 1 and salary ≤50K is represented as 0
- The total number of instance is 45222, 70% for training, 30% for testing(50% for private and 50 % for public). In addition, the data set is imbalanced.
≤50k: 34014 (75%)
>50k: 11208 (25%)

Data

- There are 14 attributes for one instance. Below is a brief overview of type and values for various features in the data set.

Age: continuous.

Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous. (The number of people the census takers believe that observation represents.)

Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Education-num: continuous.

Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Data

Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Sex: Female, Male.

Capital-gain: continuous.

Capital-loss: continuous.

Hours-per-week: continuous.

Native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

Training data

train

37	Private	272950	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	0
31	Private	261943	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	0
29	Private	285419	12th	8	Never-married	Tech-support	Not-in-family	White	Male	0	0	40	United-States	0
40	Private	182217	Some-college	10	Married-civ-spouse	Other-service	Wife	White	Female	0	0	40	Scotland	0
52	State-gov	71344	Masters	14	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	United-States	0
49	Self-emp-inc	201080	Some-college	10	Divorced	Craft-repair	Not-in-family	White	Male	0	0	55	United-States	0
21	Private	122048	Some-college	10	Never-married	Adm-clerical	Own-child	White	Female	0	0	29	United-States	0
61	Self-emp-not-inc	221884	Some-college	10	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0	50	United-States	1
25	Self-emp-not-inc	108001	9th	5	Never-married	Craft-repair	Not-in-family	White	Male	0	0	15	United-States	0
47	Private	171751	HS-grad	9	Divorced	Sales	Not-in-family	White	Female	0	0	40	United-States	0
20	Private	227411	Some-college	10	Never-married	Other-service	Own-child	White	Female	0	0	20	United-States	0
26	Private	166301	Bachelors	13	Never-married	Tech-support	Not-in-family	White	Male	0	0	40	United-States	0
33	Private	208855	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	1
56	Private	235205	HS-grad	9	Widowed	Other-service	Unmarried	White	Female	0	0	40	United-States	0
35	Private	474136	HS-grad	9	Never-married	Craft-repair	Not-in-family	White	Male	0	1408	40	United-States	0

Testing data

33	Private	118551	HS-grad	9	Divorced	Adm-clerical	Unmarried	White	Female	0	0	40	United-States
34	Local-gov	198953	Some-college	10	Divorced	Protective-serv	Unmarried	Black	Female	0	0	40	United-States
27	Private	303954	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States
41	Private	194636	Assoc-voc	11	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States
43	Self-emp-not-inc	176069	HS-grad	9	Married-civ-spouse	Other-service	Wife	White	Female	0	0	16	United-States
24	State-gov	334693	HS-grad	9	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0	40	United-States
71	Self-emp-not-inc	238479	10th	6	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	8	United-States
19	Private	42069	HS-grad	9	Never-married	Transport-moving	Own-child	White	Male	2176	0	45	United-States
17	Private	106733	11th	7	Never-married	Craft-repair	Own-child	White	Male	594	0	40	United-States
23	Private	117767	HS-grad	9	Never-married	Other-service	Own-child	White	Male	0	0	35	United-States
28	Private	212588	Some-college	10	Married-civ-spouse	Handlers-cleaners	Husband	White	Male	0	0	40	United-States
41	Federal-gov	130760	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	24	United-States
60	Private	128367	Some-college	10	Divorced	Prof-specialty	Unmarried	White	Male	3325	0	42	United-States
23	Private	203203	Bachelors	13	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	40	United-States
34	Private	158420	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	70	United-States
54	Private	167770	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	1902	55	United-States
41	Private	171234	Some-college	10	Never-married	Machine-op-inspct	Not-in-family	White	Female	0	0	48	United-States

Evaluation

- The competition will take 50% of the test data to calculate the accuracy.
Final rank will show on E3 after the competition.
- The evaluation for this competition is F1 score.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Submission files

- The maximum number of daily submissions is **10**.
- The file should be **CSV file** contains two columns:
ID: the index
ans: salary >50K is represented as 1
salary <=50K is represented as 0











ID	ans
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Leaderboard](#)[Rules](#)[Team](#)[Host](#)[My Submissions](#)[Late Submission](#)

The private leaderboard is calculated with approximately 50% of the test data.

This competition has completed. This leaderboard reflects the final standings.

[Refresh](#)

#	△pub	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	—	purplearrow_nthu			0.85365	9	5mo
2	▲ 6	publicname_nctu			0.82926	34	5mo
3	▲ 1	wulala_nthu			0.82926	57	5mo
4	▲ 1	toma_nthu			0.82774	34	5mo
5	▼ 3	automl_nthu			0.82469	34	5mo
6	▲ 1	yee_nthu			0.82164	61	5mo
7	▼ 4	冰鳥_nthu			0.82012	83	5mo
8	▲ 9	kk123_nthu			0.81250	54	5mo
9	▲ 1	123321_nctu			0.81097	29	5mo
10	▲ 1	Jen_nctu			0.81097	8	5mo

Grading policy

- Below baseline (0.8): 0
top 10%: 100
top 25%: 90
top 50%: 80
top 75%: 75
Others: 70

Requirements

Please archive your code, testing result and submit on E3.

Deadline: 11/19/2018 11:59 PM

Submission folder (your team name on Kaggle) should contain 2 files:

- [Student ID].py
- answer.csv

EX.

Redman_nctu :

- 0310707.py
- answer.csv

Contact Information

- If you have any questions, please email 陳泓仁.
 - Gmail: 0226.hjc@gmail.com