hw2-Classification

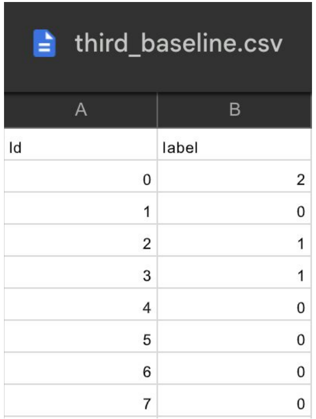
TAs
郭芳妤:clairekuo0217@gmail.com
楊依辰:sisia111062627@gapp.nthu.edu.tw
資電館743:若有問題請先來信預約

Hw2 Problem Discription

- Supervised multi-class classification problem (Credit Score)
- Given a data set
  - Training set with label
  - Testing set without label
- The dataset is transformed from a credit ranking dataset
  - 17 numeric features, 4 nominal features, 1 label
  - About 33643 cells become missing value
  - Our label is **CreditScore**
- Goal: predict the labels of testing data

Output Format

- Output your prediction to csv file with the following format and submit to kaggle
  Remember to output the first line



third_baseline.csv

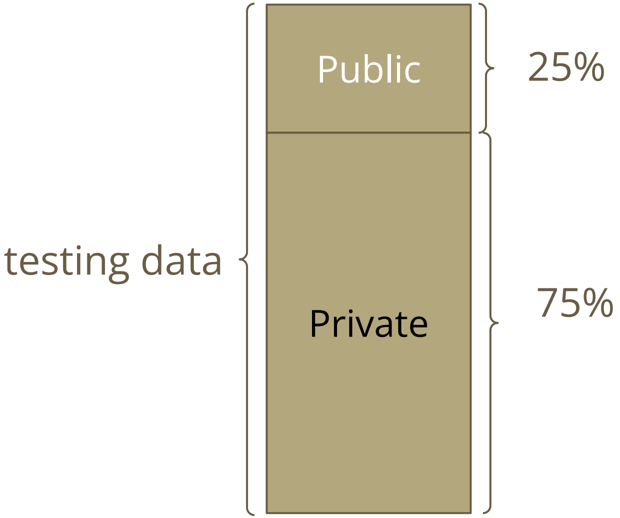
A	B
Id	label
0	2
1	0
2	1
3	1
4	0
5	0
6	0
7	0

Id	label
0	2
1	0
2	0
3	0
4	0
5	0
6	0

Please remember you need only Id & label columns with 6762 rows (Id starts from 0) !

Evaluation

- We use **F1-score** = 2×($precision$×$recall$)/($precision$+$recall$)
- **[update!!] We use macro-f1 in this homework**
- There are two leaderboards on Kaggle
  - Public: can be seen during competition
  - Private: can be seen after competition



Hw2 Submission

- HW2 will be held on Kaggle
  - Please register a Kaggle account first
  - hw2 link: https://www.kaggle.com/competitions/nthu-2024datascience-hw02-classification
- Kaggle is a platform of
  - Machine learning competition
  - Sharing dataset
- Hw2 deadline: **2024.04.16** Tue. 23:59 (3 weeks)
- We will use the result on Kaggle to score this homework
  - **No need to hand in any files on eeclass**
  - Remember to **fill your Kaggle name** in the google form: https://forms.gle/h6Co4wwWp5GZwCPEA

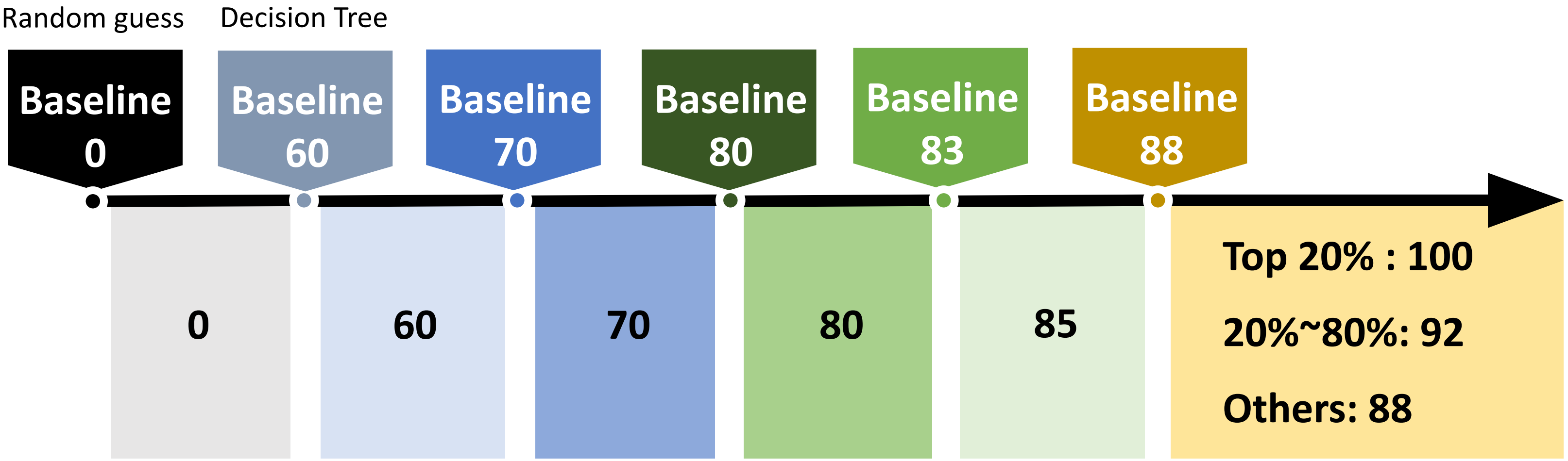
Scoring and Rules

Scoring

- Use private leaderboard result for final scoring
- Baseline scores: We will score according to given 6 baseline scores

baseline	public	private
baseline-0	0.30155	0.33979
baseline-60	0.53815	0.53698
baseline-70	0.60884	0.60554
baseline-80	0.66737	0.68570
baseline-83	0.70888	0.73306
baseline-88	0.77935	0.80414

Scoring



Random guess

Decision Tree

Baseline 0

Baseline 60

Baseline 70

Baseline 80

Baseline 83

Baseline 88

0

60

70

80

85

Top 20% : 100

20%~80%: 92

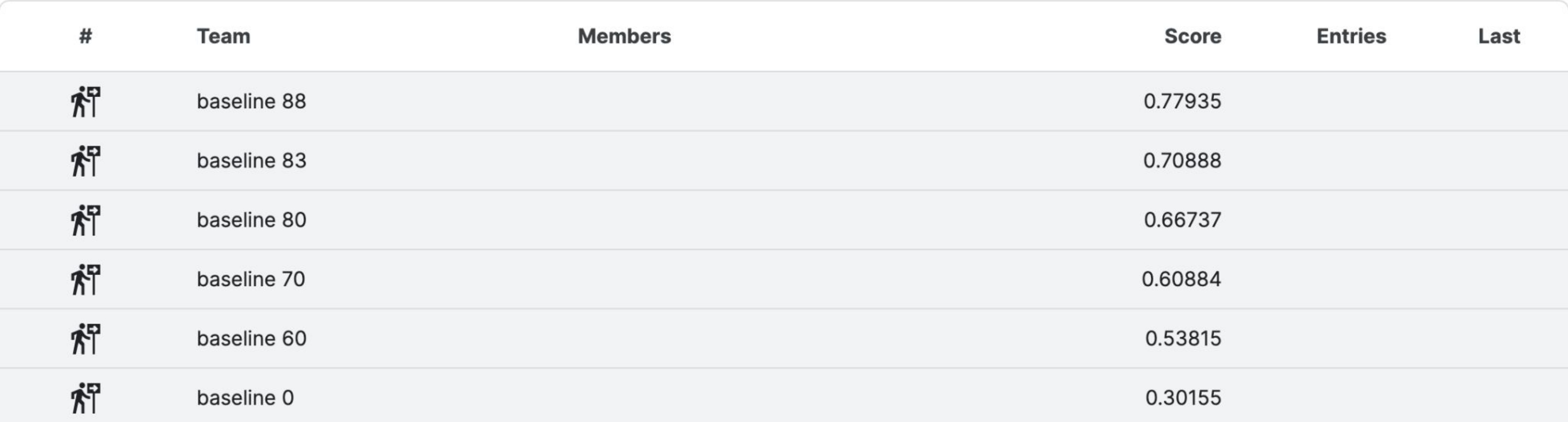
Others: 88

- You will get **0**, if your private score is between *baseline 0* and *baseline 60*
- You will get **60**, if your private score is between *baseline 60* and *baseline 70*
- You will get **70**, if your private score is between *baseline 70* and *baseline 80*
- And so on

Scoring

Baseline scores

- There are benchmarks on the leaderboard for reference



#	Team	Members	Score	Entries	Last
🚶	baseline 88		0.77935		
🚶	baseline 83		0.70888		
🚶	baseline 80		0.66737		
🚶	baseline 70		0.60884		
🚶	baseline 60		0.53815		
🚶	baseline 0		0.30155		

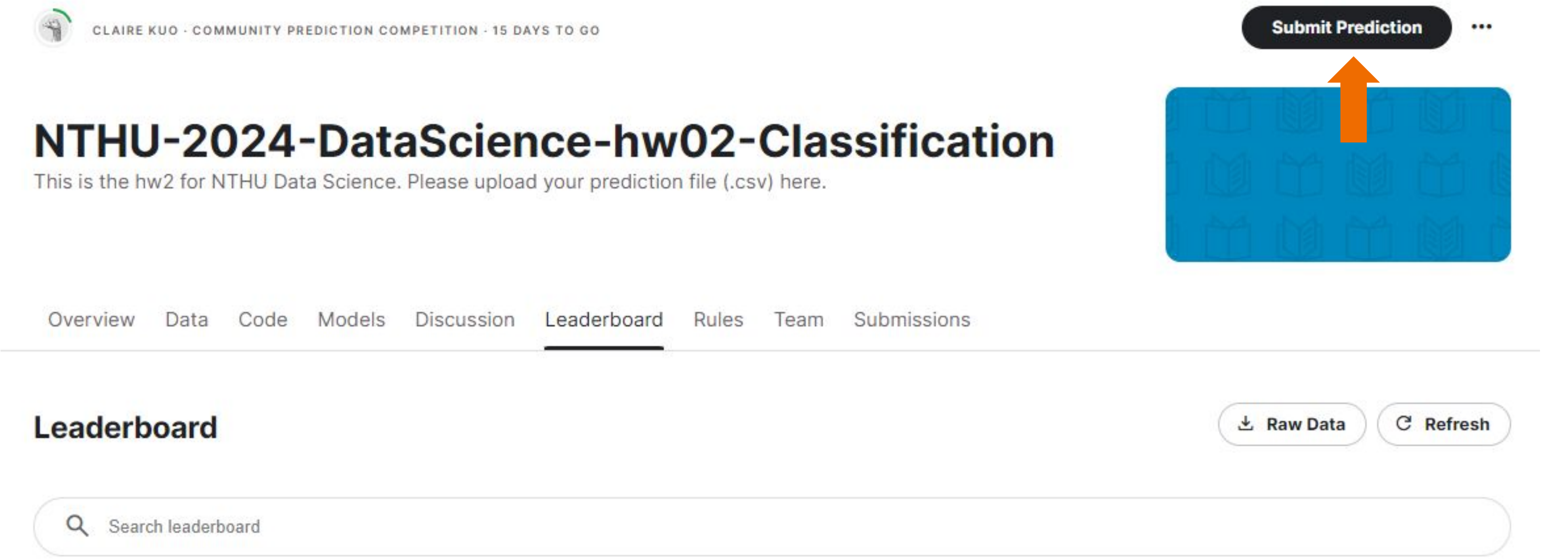
Rules

- You don’t have to submit the code!
- You can submit 20 times per day
- You can choose 4 predictions for final scoring
  - Kaggle will use the best one to be your final result
- No cheating!
- 若有問題，請寄信給負責助教。信件 title 請註明【DS-hw2-question】，並盡可能在信件內描述你遇到的困難，以方便我們協助你。
  - If you have any questions, please email the responsible teaching assistant. Please include "**[DS-hw2-question]**" in the email title and describe your difficulties as clearly as possible in the body of the email to facilitate our assistance.

How to submit and choose predictions

How to submit

- Click ***'Submit Predictions'*** button on the navigation bar



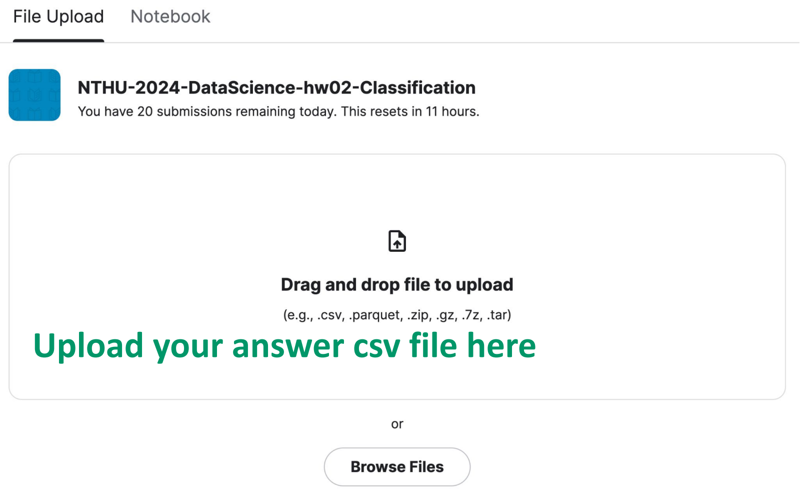
CLAIRE KUO · COMMUNITY PREDICTION COMPETITION · 15 DAYS TO GO

NTHU-2024-DataScience-hw02-Classification

This is the hw2 for NTHU Data Science. Please upload your prediction file (.csv) here.

Leaderboard

How to submit



File Upload Notebook

NTHU-2024-DataScience-hw02-Classification

You have 20 submissions remaining today. This resets in 11 hours.

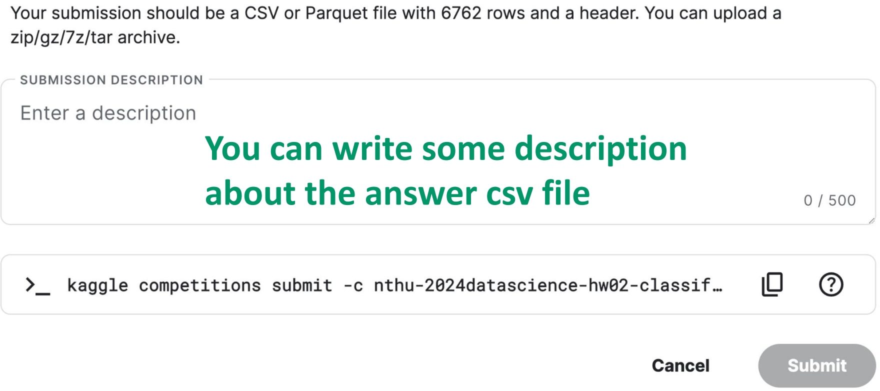
Drag and drop file to upload

(e.g., .csv, .parquet, .zip, .gz, .7z, .tar)

or

Browse Files

Upload your answer csv file here



Your submission should be a CSV or Parquet file with 6762 rows and a header. You can upload a zip/gz/7z/tar archive.

SUBMISSION DESCRIPTION

Enter a description

0 / 500

```
>_ kaggle competitions submit -c nthu-2024datascience-hw02-classif…
```

Cancel Submit

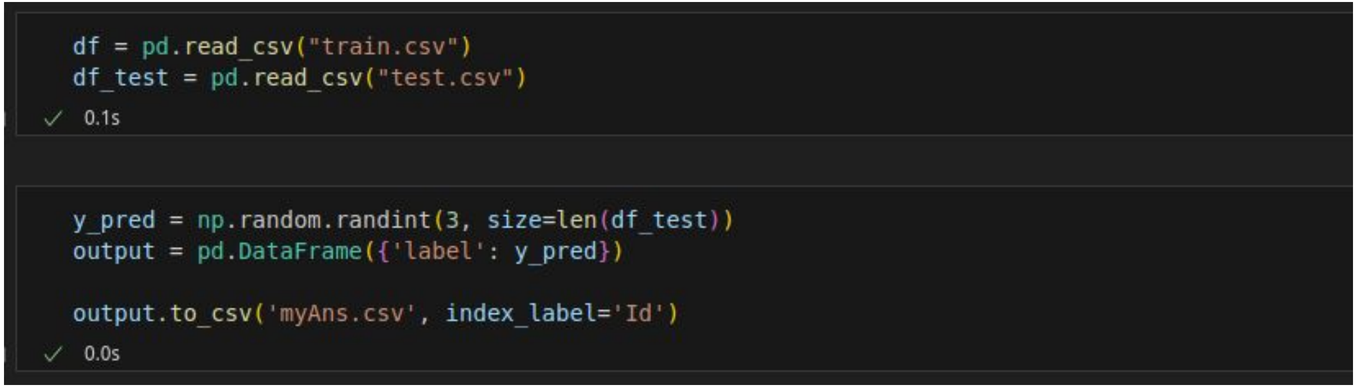
You can write some description about the answer csv file

Click to submit

Hints

Hints

- How to read/write the file?



```
df = pd.read_csv("train.csv")
df_test = pd.read_csv("test.csv")
```

✓ 0.1s

```
y_pred = np.random.randint(3, size=len(df_test))
output = pd.DataFrame({'label': y_pred})

output.to_csv('myAns.csv', index_label='Id')
```

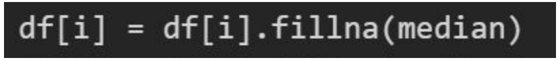
✓ 0.0s

Hints - Categorical features

`col_2`, `col_13`, `col_17`, `col_20` are categorical variables.
Please use the correct methods to handling those columns during training.

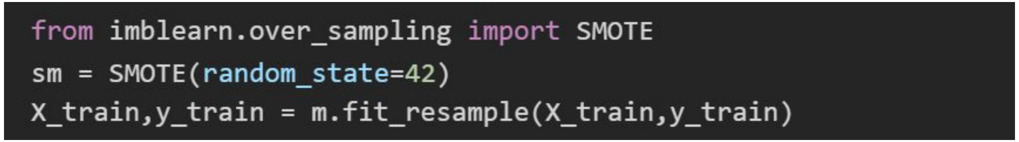
Hints

- Fillna with median in numeric features instead of 0



```
df[i] = df[i].fillna(median)
```

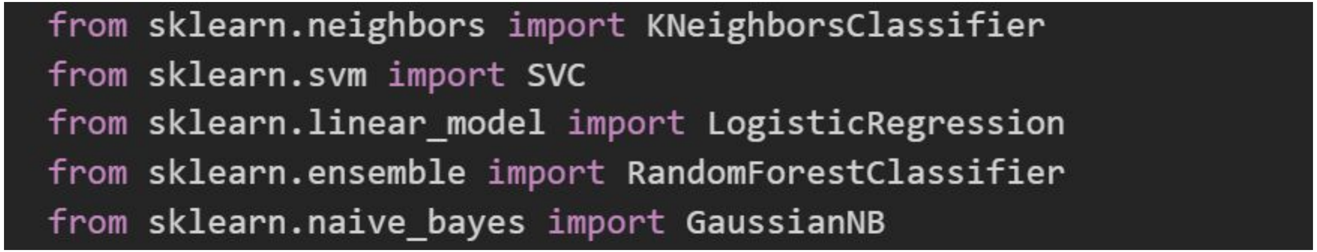
- Deal with data imbalance



```
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state=42)
X_train,y_train = m.fit_resample(X_train,y_train)
```

Hints

- Try different models
  - KNN, SVM, Logistic Regression, Random Forest ...



```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
```

Finetune the model may achieve higher effect than the baseline 70 and 80

Hints

- **More techniques for better performance**
  - Feature selection (MI score)
  - Normalization
  - Dimension reduction (PCA, TSNE)
  - Try other different models
  - ...
- **We use private leaderboard as the final score**
  - Use public score to choose your model is dangerous
  - It’s better to perform validation

Packages you may use

- Scikit-learn
  - https://scikit-learn.org/stable/index.html

- Pandas
  - https://pandas.pydata.org/pandas-docs/stable/

- Imbalance learn (for over sampling and down sampling)
  - https://imbalanced-learn.org/stable/