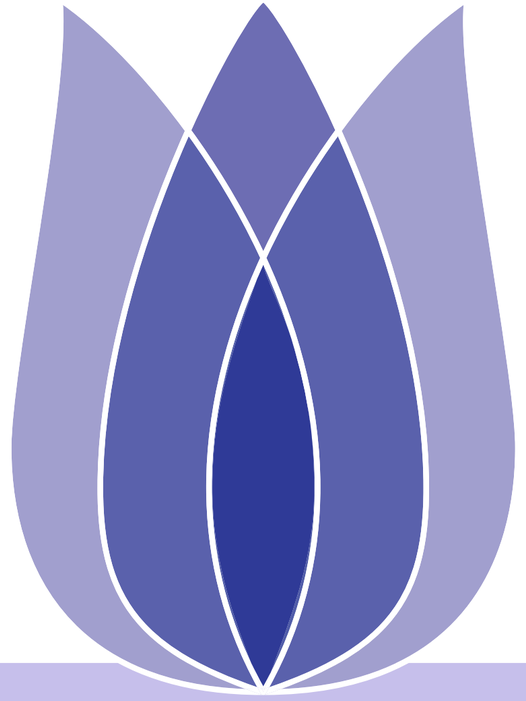
CDMC2020 Task Presentation

Rongxin Xu, Xiaoyan Wang

2020-10-25



Overview

Task 1

- Feature importance
- Result

Task 2

- NLP
- Feature engineering
- Data imbalance
- Cross-validation result
- Results on the validation set

Task 1

Feature importance

Defn There are many columns containing a large number of 0 values, these columns obviously can not provide more information. The following figure shows the importance of features using XGBoost.

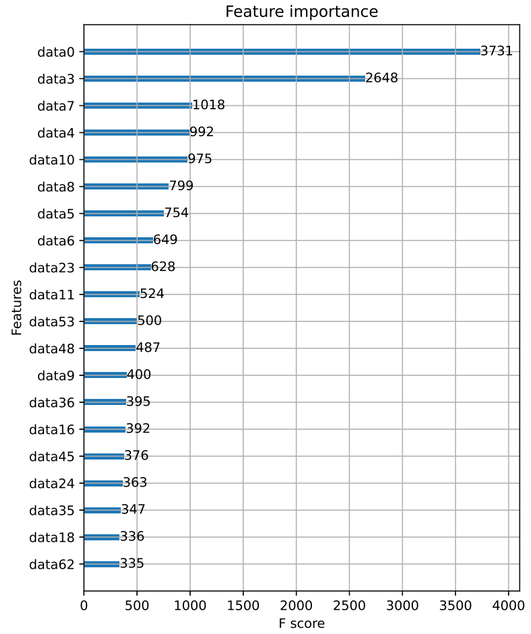


Figure 1: feature importance

Result

Defn We still use models such as XGBoost, LightGBM, and Randomforest for cross-validation, and observe the effect on the validation set.

- The average accuracy of each model is between 97% and 98%, and the average F1-score is between 95%-97%.
- The accuracy on the verification set is about 92%, and the F1-score is about 80%.

Task 2

NLP

Defn Simply use the punctuation in bytesquence for word segmentation.Then use BERT to make predictions.

Family	CPU	ByteSeq nce
Mirai	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Bashlite	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Mirai	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Mirai	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Dofloo	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Mirai	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV
Bashlite	armel	ZkaWFmaXlhaBy/C8YEyWtYFMXVmBZI1Zy2ggWYFoXV

Figure 2: Simply nlp

Feature engineering

- Convert String to ASCII
- One-hot encoding the attribute "CPU"

Data imbalance

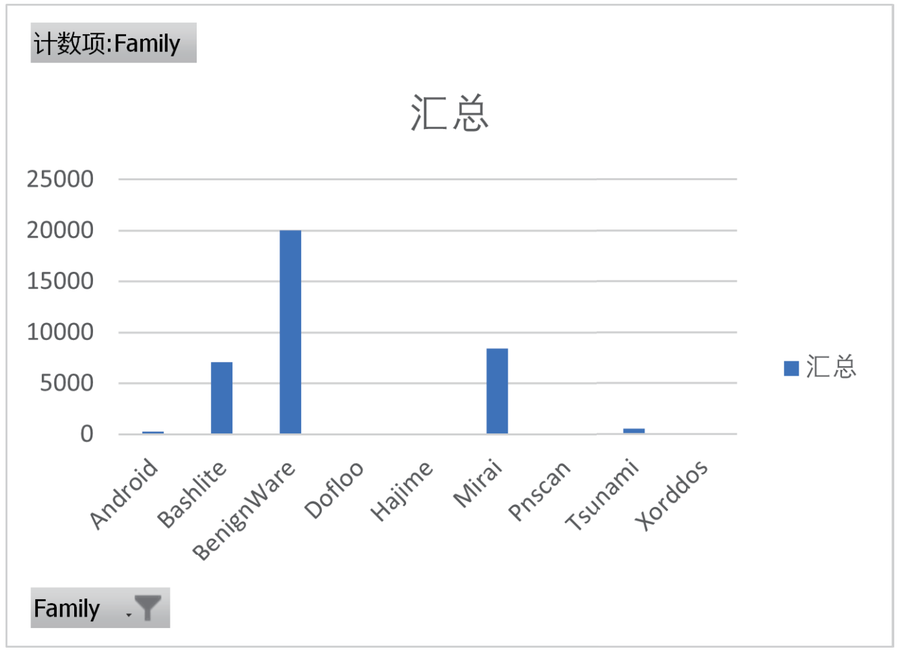


Figure 3: imblance

Cross-validation result

Defn Simply use xgboost to perform 5-fold cross-validation and observe the effect on the validation set.

- The average accuracy is about 87%
- The average f1-socre is about 51%

Results on the validation set

	precision	recall	f1-score	support
0	0.39	0.45	0.42	49
1	0.84	0.88	0.86	1418
2	0.93	0.90	0.92	3995
3	0.17	0.25	0.20	4
4	0.86	1.00	0.92	12
5	0.84	0.84	0.84	1684
6	0.00	0.00	0.00	1
7	0.33	0.54	0.41	101
8	0.00	0.00	0.00	2
accuracy			0.87	7266
macro avg	0.48	0.54	0.51	7266
weighted avg	0.88	0.87	0.88	7266

Figure 4: Results on the validation set

End

Thank you!