Tom&Jerry Classification

Состав команды:

Морецкая Людмила

Украинцева Елена

Горбатова Екатерина

Куратор:

Козлов Кирилл

Ссылка на проект на гитхаб:

https://github.com/moretskayalv/MOVC_project_1

Данные для проекта:

tom_jerry_1

jerry

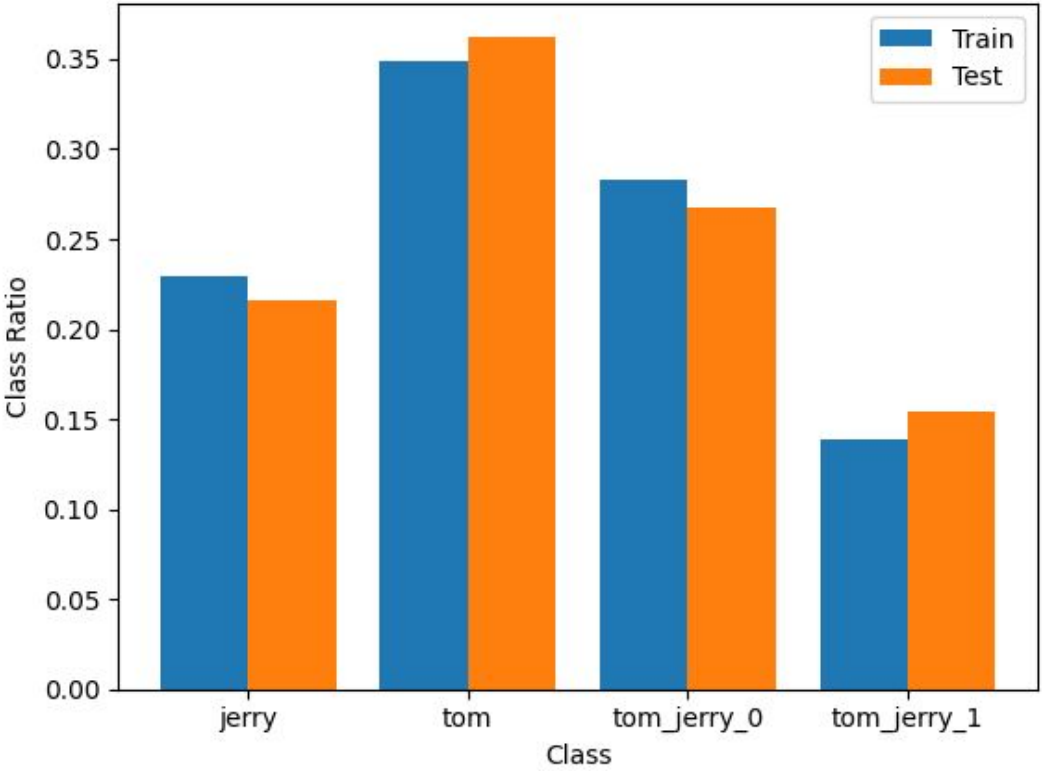
tom

tom_jerry_0

EDA

	Count(file)	Percent
No one	1528	27.893392
Jerry	1240	22.635999
Tom	1930	35.231836
Tom&Jerry	780	14.238773

ML



	Tom	Jerry	Tom&Jerry	No one
Train	34.88%	22.98%	13.85%	28.29%
Test	36.28%	21.61%	15.40%	26.72%

	accuracy test	*precision test*	*recall test*	*f1-score test*
LogReg	***0.69***	***0.71***	***0.69***	***0.68***
LogReg+PCA	***0.44***	***0.48***	***0.44***	***0.44***
SVC	***0.75***	***0.75***	***0.75***	***0.74***
SVC+ PCA	***0.75***	***0.75***	***0.75***	***0.75***

Алгоритмы и метрики (**PCA 500** компонент)

Выводы:

- PCA ухудшает показатели для LogReg, при этом незначительно (скорее случайно) улучшает показатели SVC
- PCA сильно ускоряет обучение модели. Без PCA на данных LogReg обучается и делает предсказания > 2 часов
- PCA имеет смысл использовать только для SVC, так как помимо ускорения получаем также неплохие показатели

Подбор количества компонент для PCA+SVC

	1	101	201	301	401	501	601	701	801	901	1000	1900
accuracy	0.32	0.71	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.75
precision	0.34	0.71	0.73	0.74	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
recall	0.32	0.71	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.75
f1-score	0.33	0.71	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.75	0.75	0.75

Изменения которые мы увидели на текущем чекпоинте

- В предыдущий раз, не получилось обучить LogReg, поэтому мы не увидели ухудшения при использовании PCA. Сейчас видим, что PCA сильно ухудшает наши метрики, не смотря на ускорение работы модели
- Опытным путем пришли к тому, что оптимальное количество компонент для использования алгоритма PCA+SVC около 600. До 600 рост метрик заметен достаточно явно. После рост очень идет очень медленно (от 1000 до 2000 рост в тысячных), а время обработки увеличивается гораздо больше.