Построение алгоритма определения вероятности подключения услуги пользователем

Выполнил студент гр. GU_AI_507 – Ермоленко Н. К.

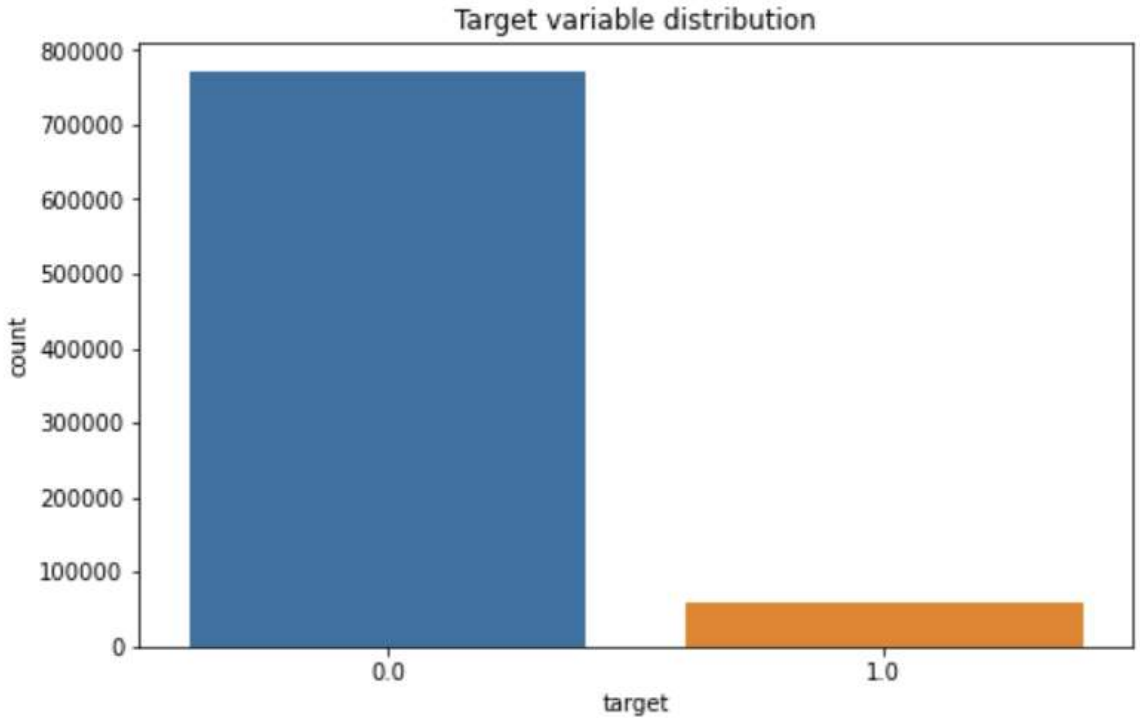
Цель и основные задачи курсовой работы

- Цель работы: построение алгоритма определения вероятности подключения услуги пользователем.

В рамках курсовой работы поставлены следующие задачи:

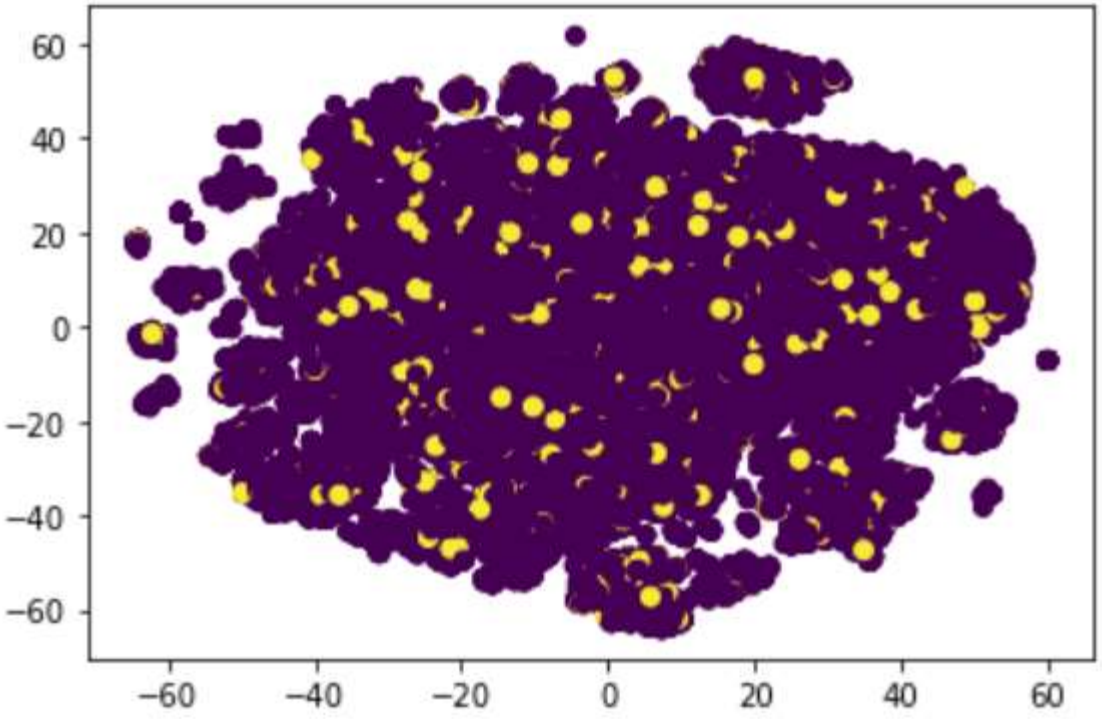
- Исследование набора данных.
- Обзор алгоритмов классификации.
- Оценка результатов работы алгоритмов для представленного набора данных.
- Улучшение результатов работы выбранного алгоритма.
- Предсказание вероятностей подключения услуг пользователями.

Распределение целевой переменной



Значения переменной целевого класса распределены неравномерно

Облако данных построенное при помощи t-SNE



По облаку данных невозможно выделить однородных хорошо различимых кластеров

Рассмотренные алгоритмы классификации

- Логистическая регрессия
- Дерево решений
- Случайный лес
- Адаптивный бустинг на логистической регрессии
- Адаптивный бустинг на деревьях решений
- Градиентный бустинг (catboost)

Результаты работы алгоритмов классификации

	f1-score (average='macro')
Логистическая регрессия	0.7028
Дерево решений	0.7078
Случайный лес	0.7107
AdaBoost на логистической регрессии	0.7065
AdaBoost на деревьях решений	0.7129
CatBoost	0.7168

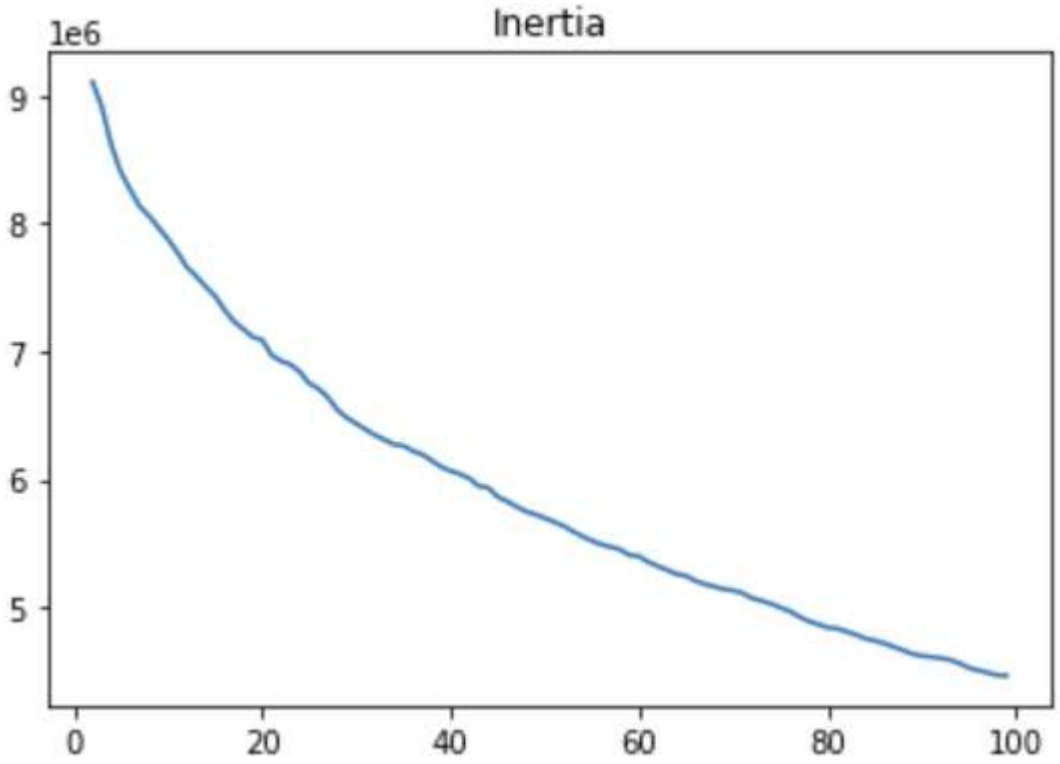
Все алгоритмы отработали со схожими результатами на кросс-валидации. Градиентный бустинг показал чуть более высокий результат по невзвешенной оценке f1. Он и принят за основу для дальнейших исследований.

Параметры модели

- eval_metric = “F1” – так как модель оценивается по метрике f1
- cat_features = “vas_id” – единственный категориальный признак в обучающем наборе данных
- auto_class_weights = “Balanced” – так как значения целевой переменной распределены неравномерно
- reg_lambda = 4.0, max_depth = 7 – были подобраны эмпирически

Использование дополнительного признака с информацией о разбиении на кластеры

Идея состоит в том, чтобы разбить исходный набор данных на кластеры, и использовать информацию о кластерной принадлежности при обучении выбранного алгоритма классификации. Для этого необходимо выбрать количество кластеров на которое будет производиться разбиение. Попробуем сделать это по методу «локтя».



По графику Inertia, построенному для разбиений до 100 кластеров, не удаётся определить точку резкого снижения суммы квадратов расстояний объектов кластеров до их центров. Это происходит либо из-за того, что оптимальное количество разбиений на кластеры для представленного набора данных больше 100 (что не подойдёт для модели, так как и без этого большой набор данных станет ещё больше), либо из-за того, что представленный набор данных невозможно разделить на однородные обособленные кластеры (что согласуется с визуальным представлением при помощи t-SNE). Таким образом, от идеи использования информации о кластерном разбиении лучше отказаться для представленного набора данных.

Использование метода главных компонент для снижения размерности

Оценка кумулятивной дисперсии показала, что 90% всей дисперсии данных приходится на 122 компоненты, а 80% на 90 компонент.

	f1-score (average='macro')	computation time
without PCA	0.7168	3m 6s
PCA(122 comp)	0.7166	3m 16s
PCA(90 comp)	0.716	2m 42s

Запуск на кросс-валидации на тестовой выборке показал, что использование метода главных компонент не позволяет существенно снизить время обучения модели без потери качества.

Использование составной модели

Идея в том, что для улучшения результатов, при обучении модели можно использовать результаты работы других алгоритмов классификации.
Для обучения итоговой (самой сильной) модели с результатами работы, были выбраны:

- Случайный лес
- Адаптивный бустинг на логистической регрессии
- Адаптивный бустинг на деревьях решений

	f1-score (average='macro')
CatBoost	0.7168
Complex model	0.7158

Использование составной модели не дало прироста качества на кросс-валидации.

Выводы

Все алгоритмы отработали на представленном наборе данных со схожими результатами. Чуть лучший результат, по отношению к другим алгоритмам, на кросс-валидации получился для градиентного бустинга (catboost). Попытки улучшить результаты путём использования информации о кластерном разбиении, использования метода главных компонент, использования составной модели не привели к улучшению результата. Для итоговых предсказаний использовалась модель catboost с представленными на слайде 7 параметрами.

По оценке precision-recall кривой, получается, что наилучший результат по метрике f1-score достигается, при использовании порога 0.62 для принятия решения в классификаторе.

Конец

Спасибо за внимание