Отчёт №2
о проделанной работе по проекту

ФИО студента: Кашникова Анна Дмитриевна
Группа: 21-ПМИ-2
Образовательное учреждение: национальный исследовательский университет
«Высшая школа экономики»

1 Обзор библиотек: sklearn, xgboost, catboost

Было проведено изучение таких библиотек как sklearn, xgboost, catboost. Так как задачей являлось построение модели градиентного бустинга, стоял выбор какой лучше библиотекой воспользоваться. После изучения стало ясно, что на выбор библиотеки сильно влияет датасет. И в зависимости от задачи предпочтительно использовать что-то более подходящее.

- Scikit-learn - популярная библиотека машинного обучения, предоставляющая простые и эффективные инструменты для интеллектуального анализа данных. Он подходит для широкого спектра стандартных задач машинного обучения, включая классификацию, регрессию, уменьшение размерности и предварительную обработку и тп.

- XGBoost является вычислительно эффективной реализацией градиентного бустинга над решающими деревьями. Помимо оптимизированного программного кода, авторы предлагают различные улучшения алгоритма.

- Catboost - библиотека с реализацией градиентного бустинга над решающими деревьями, особенно удобна для работы с датасетами, содержащими категориальные признаки.

Таким образом можно сравнить данные библиотеки:
Тип задачи - для стандартных задач машинного обучения Scikit-learn часто является хорошим выбором благодаря своей гибкости и обширной библиотеке моделей и инструментов. Для потенциально сложных данных на практике используются XGBoost и CatBoost.
Производительность - XGBoost и CatBoost имеют высокую производительность.
Категориальные признаки - если набор данных содержит большое количество категориальных признаков, CatBoost будет удобнее всего для эффективной работы с ними.

Датафрейм был составлен следующим образом: из файлов с извлеченными признаками для каждого аудио, были посчитаны средние признаки на протяжении всей песни, и для каждой песни был определён свой класс исходя из статистических оценок песни.

В данном датасете нет категориальных переменных, и датасет достаточно объемный 1802 $\times$ 261, именно поэтому было решено использовать xgboost.

2 Составление модели

Для составления модели, был подготовлен датасет - было проведено label-encoding, нормализация с помощью standart scaler. Выборка была разбита следующим образом: тестовая выборка - составляет 0.3 от всей, тренировочная соответственно 0.7.

Была построена модель градиентного бустинга с параметрами по умолчанию из библиотеки xgboost и результативность модели была проверена с помощью accuracy-метрики и составила малую часть: 0.323.

Также была построена модель градиентного бустинга для сравнения из библиотеки sklearn, с параметрами по умолчанию и результативность модели была проверена с помощью accuracy-метрики и оказалась ещё меньше: 0.297.

3 Методы по уменьшению размерности

Были изучены такие методы, основанные на линейных преобразованиях, как PCA(Princip Component Analysis) и ICA(Independent Component Analysis). Также было попробовано применить их к датасету и оценить модели с новыми признаками. Применив PCA из библиотеки sklearn с величиной объяснимой дисперсии 0.95, точность модели, построенной с помощью xgboost, оцениваемой метрикой accuracy составила 0.227. Применив ICA с параметрами по умолчанию, точность модели, построенной с помощью xgboost, оцениваемой метрикой accuracy составила ещё меньше по сравнению с PCA: 0.205.