Машинное обучение, DS-поток.

Задание 6.

- Дедлайн **27 марта 02:00**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru`, указав тему письма "`[ml] Фамилия Имя - задание 6`". Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- По задачам 2, 3 прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `6.N.ipynb` и `6.N.pdf`, где `N` — ваш номер из таблицы с оценками.
- Задачу 1 необходимо оформить в tex'е и прислать `pdf` или же прислать фотку в правильной ориентации рукописного решения, где все четко видно.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Не забывайте делать пояснения и выводы.

Задачи

1. **(6 баллов)** В модели XGBoost запишите задачу оптимизации при построении нового дерева, критерий информативности и оптимальные ответы в листьях в следующих случаях.
   (a) Задача регрессии, квадратичная функция потерь $\mathcal{L}(y, z) = (y - z)^2$.
   (b) Задача классификации, экспоненциальная функция потерь $\mathcal{L}(y, z) = e^{-yz}$.
   (c) Задача классификации, логистическая функция потерь $\mathcal{L}(y, z) = \log\left(1 + e^{-yz}\right)$.

   В задачах классификации классификатор предсказывает степень уверенности принадлежности классу из $\mathscr{Y} = \{-1, 1\}$. Решающее правило имеет вид $f(x) = \text{sign}\left(\widehat{y}(x)\right)$.

2. **(10 баллов)**

   В файле `houses_train.csv` представлен набор данных про квартиры в городе Сиэтл, штат Вашингтон. Задача — предсказать цену на жилье по имеющимся данным.

   Данные имеют следующие столбцы:

   - `id` — идентификационный номер жилья
   - `date` — дата продажи дома
   - `price` — цена
   - `bedrooms` — количество спален
   - `bathrooms` — количество ванных комнат, где `.5` означает комнату с туалетом, но без душа

- `sqft_living` — площадь жилья
- `sqft_lot` — площадь участка
- `floors` — количество этажей
- `waterfront` — видна ли набережная
- `view` — насколько хороший вид
- `condition` — индекс от `1` от `5`, отвечающий за состояние квартиры
- `grade` — `1` до `13`, `1-3` соответствует плохому уровню строительства и дизайна, `3-7` — средний уровень, `11-13` — высокий.
- `sqft_above` — жилая площадь над уровнем земли
- `sqft_basement` — жилая площадь под уровнем земли
- `yr_built` — год постройки жилья
- `yr_renovated` — год последней реконструкции жилья
- `zipcode` — почтовый индекс
- `lat` — широта
- `long` — долгота
- `sqft_living15` — средняя площадь жилья ближайших 15-и соседей
- `sqft_lot15` — средняя площадь участка ближайших 15-и соседей

Исследуйте зависимость качества по метрике MAPE для моделей XGBoost, LightGBM, CatBoost, а так же градиентного бустинга из sklearn, в зависимости от количества деревьев, их максимальной глубины, шага обучения, а также различных регуляризаций. При исследовании одного гиперпараметра рисуйте один график для всех моделей. Тщательно подберите цвета и тип линий так, чтобы картинка была легко читаемой. При проведении исследований не забывайте писать подробные комментарии и выводы.

Совет. Используйте код с семинаров.

3. **(3 балла + бонусы)** В продолжении исследований из предыдущей задачи выберите некоторое количество хороших на ваш взгляд моделей, постройте для них предсказание на тестовой выборке. Полученные предсказания отправьте в тренировочное соревнование на Kaggle.

   Ссылка: https://www.kaggle.com/c/f9fmt2f5hk1sjv

   Инвайт: https://www.kaggle.com/t/1beba2c61e0c44869d39b24770927d18

   Правила:

   - В Kaggle в данное тренировочное соревнование можно отправлять не более 7 решений в день (8-ю система не позволит).
   - Решения индивидуальные.
   - Качество считается по метрике MAPE.
   - До окончания соревнования доступны значения качества, посчитанные только на случайных 30% тестовых данных. Значения отображаются в Public Leaderboard
   - После окончания соревнования становится доступным Private Leaderboard, в котором значения качества посчитанны на оставшихся 70% объектов.

- Для включения в Private Leaderboard можно выбрать две посылки.
- В Leaderboard должны отображаться ваши реальные имя и фамилия. В противном случае решение может быть не зачтено.
- Все файлы, которые вы отправляете в соревнование, видны организаторам соревнования. Файлы должны иметь понятное имя, при отправке файла в систему необходимо написать краткое описание решения.
- В решении, отправляемом на почту, должно быть отображено, результаты каких моделей вы отправляете в соревнование.
- Пользоваться можно любыми пройденными в курсах моделями.
- Код студентов, занявших первые 3 места, будет запускаться. Также выборочно может запускаться код и остальных студентов
- Не забывайте сделать пояснения к своему решению.

Бонусы:

- Выдаются только при соблюдении всех правил и суммируются.
- 1 балл — ваше решение лучше чем `sample_submission.csv` на Private Leaderboard.
- 1 балл — ваше решение имеет ошибку не более 12.8% по метрике MAPE на Private Leaderboard.
- 1 конфета — попадание в топ-10 на Private Leaderboard.
- 1 конфета — попадание в топ-3 на Private Leaderboard.
- 1 конфета — попадание в топ-1 на Private Leaderboard.