DS-поток.

Продвинутое практическое задание.

Предсказание спроса на товары.

Описание задачи.

Задание представляет собой участие в соернованиях по предсказанию спроса (продаж) для магазинов Walmart в трех штатах: Калифорнии, Техасе и Висконсин. Имея информацию о предыдущих трех месяцах продаж товаров, нужно предсказать их спрос на них на каждый день следующего месяца. Ссылки на Kaggle:

1. `https://www.kaggle.com/c/m5-forecasting-accuracy`
2. `https://www.kaggle.com/c/m5-forecasting-uncertainty`

Эти два соревнования проводятся на одном и том же датасете. В первом нужно предсказать точное значение спроса на определенный товар в определенный день, а во втором соревновании – квантили уровня 0.5, 0.67, 0.95 и 0.99 для каждого товара вместо точного значения.

В описании соревнований вы найдете более подробное описание задачи, данных и сами файлы для скачивания. Также составители соревнования предоставили целую книгу участника: `https://mofc.unic.ac.cy/m5-competition`, обязательно ее прочитайте, перед тем, как приступать к выполнению. В ней можно узнать смысл всех полей в табличных данных, подробное описание используемых для проверки качества метрик и описание бейзлайнов на лидерборде.

В книге участника обратите особое вниманиена:

1. Датасет. Подробно описана схема агрегации продаж по магазинам в каждом штате и каждом магазине. Обязательно нужно иметь это в виду.
2. Бейзлайны составителей соревнования.
3. Метрики. В данном случае в первом соревновании используется Root Mean Squared Scaled Error, а не обычный RMSE:

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \widehat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(Y_t - Y_{t-1}\right)^2}}$$

Y_t – истинное значение, $\widehat{Y}_t$ – предсказание, n – число значений в обучающей последовательности, h – горизонт предсказания. Эта метрика помогает надежно оценивать предсказания для временных рядов разной длины. Во втором соревновании используется Scaled Pinball Loss, про который можно подробнее узнать уже в самой книге.

Формат выполнения задания

Каждая исследовательская задача по заданию должна состоять из следующих шагов.

1. Формулировка задачи на простом "пользовательском" языке.
2. Формулировка задачи на математическом языке.
3. Исследование на Питоне или на R. Весь необходимый код должен быть включен в решение задания.
4. Выводы как на математическом, так и на "пользовательском" языке.

Правила

1. См. общие правила продвинутых практических заданий.
2. Код вашего решения при проверке может быть запущен, поэтому он должен работать без ошибок.
3. Решение должно содержать подробное текстовое описания метода решения задачи, а также его обоснование.
4. Запрещено нарушать правила, оговоренные в соревновании на kaggle (`https://www.kaggle.com/c/m5-forecasting-accuracy`). Основные из них:
   (a) Только один аккаунт для каждого участника.
   (b) Максимум 5 сабмитов в день.
   (c) Для финального сабмита, по которому будет измеряться метрика на Private Leaderboard, может быть выбран только 1 сабмит.
   (d) Можно использовать сторонние данные, но в этом случае эти данные должны быть доступны всем участникам соревнования и ссылка на данные должна быть опубликована вами до 23 июня в официальном форуме соревнования (Discussion).
5. Необходимо отправлять результаты своих экспериментов в Kaggle на Public Leaderboard.
6. Ваше имя в лидерборде должно иметь формат : "mipt-stats-Фамилия"
7. Решение должно быть полностью воспроизводимо для получения в точности такого же score на Public и Private Leaderboard.