Домашнее задание 1

April 2022

1 Введение.

В этом задании вам предлагается решить задачу предсказания продаж различных категорий товаров известной сети быстрого питания по исторически собранным данным. Вам предстоит узнать, как на основе статистической информации о связи между продажами в ресторанах быстрого питания и погодными условиями, можно сделать предсказания по поводу объемов продаж. Данные, с которыми вы будете работать, включают следующие признаки:

- date - дата в формате yy-mm-dd
- city_name - город, в котором находится данный ресторан
- store_id - идентификатор ресторана
- category_id - идентификатор категории обозначенного товара
- product_id - идентификатор продукта
- weather_desc - описание погоды в виде одной из фиксированного набора фраз
- humidity - влажность воздуха
- temperature - температура воздуха
- pressure - атмосферное давление
- sales - целевая метрика - продажи товара в штуках

Вы можете пользоваться любыми изученными алгоритмами, а также исследовать и реализовывать любые статистические модели, модели машинного обучения по своему усмотрению. Вы также можете проводить любую предобработку признаков. Вам предоставлена полная свобода.

Ваша цель - минимизировать метрику MAE (Mean Absolute Error) ваших предсказаний. MAE представляет из себя сумму модулей отклонений

предсказаний вашей модели от истинных значений. Подробнее об этой метрике можно прочитать тут: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html`

Вам предоставлены 3 файла:

- sample_submission - пример того, в каком виде вы должны предоставить своё решение
- train.csv - файл содержит тренировочный датасет вместе с колонкой ответов sales
- test.csv - файл содержит тестовый датасет, для которого вы должны сделать свои предсказания, и отправить их в ккачестве решения

2 Рекомендации

Несмотря на то, что вы можете предпринимать любые действия при обработке данных по своему усмотрению, мы предлагаем вам использовать некоторые из следующих идей в рамках вашей работы. Они с высокой долей вероятности помогут вам добиться лучшего качества вашей модели.

2.1 Генерация признаков

Несмотря на то, что модели машинного обучения способны зачастую находить сложные и неочевидные связи в данных, явная помощь в выделении некоторых зависимостей может существенно улучшить качество ваших предсказаний. Сделать это предлагается при помощи генерации новых признаков, которые вы можете добавить в свой датасет.

2.1.1 Дни недели

Мы рекомендуем обратить внимание модели на зависимость заказов от дня недели (вы можете проверить гипотезу о существовании такой зависимости при помощи отображения графика распределения числа заказов по дням недели). Вы также можете добавить отдельный признак, отображающий факт выходного дня, ведь ни одна модель по умолчанию не знает, что именно суббота и воскресенье признаны человечеством лучшими днями для отдыха.

2.1.2 Признаки с "задержками"(лагами)

Зачастую бывает полезно сообщить модели информацию о динамике процесса, который она пытается анализировать. Для этого можно добавить специальные признаки, которые отображают значение некоторого из уже имеющихся в датасете признака за n дней до текущего момента. Не стесняйтесь сгенерировать много таких дополнительных признаков!

В варианте по умолчанию мы предлагаем вам сгенерировать такие признаки с "лагами"для признака sales (да-да, ведь предсказываемое значение за предыдущие промежутки времени также является в некотором смысле признаком для нашей модели!), сгруппированного по ресторанам и продуктам, взятым с n = 7...22. Так вы сможете использовать этот признак также для тестовой выборки (ведь значения sales за неделю до любого из тестовых дней вы сможете взять из тренировочного датасета).

Для технической реализации этой идеи рекомендуем обратить внимание на следующие функции:

`https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.groupby.html`

`https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.shift.html`

2.1.3 Категориальные признаки

Чаще всего преобразование категориальных переменных может быть двух типов:

1. Замена категорий числовыми лейблами (поэтому такое преобразование называют label encoding).

   Например, преобразование строковых значений признака «Город» из «Сургут», «Москва», «Санкт-Петербург» в числовые 0, 1 и 2, соответственно. Для этого предназначен класс LabelEncoder() из модуля sklearn.preprocessing. Теперь признак «Город» — числовой, а не категориальный. Но у него есть существенный недостаток — он устанавливает между категориями отношения типа «больше»/«меньше». Для моделей, основанных на линейной взаимосвязи (например, для линейной регрессии), это не очень хорошо. Поэтому чаще используются другие подходы.

2. Представление одного категориального поля в множество бинарных полей (такое преобразование называют one-hot encoding).

   В этом случае вместо поля «Город» появятся поля «Москва», «Санкт-Петербург», «Сургут» и другие, а эти новые признаки будут принимать только значения 0 или 1. Для этого существует, например, функция pandas.get_dummies(). Ей на вход передаётся весь датафрейм, а она сама выделяет категориальные переменные, преобразует их в новые признаки (с удобными новыми названиями), которые называют dummy-переменными, и возвращает обновлённый датафрейм. Для такого преобразования также можно использовать класс OneHotEncoder модуля sklearn.preprocessing, но он чуть менее удобен.Бинарная замена категориальных признаков хороша почти всегда. Но если их слишком много или у какого-то одного из них много уникальных значений (или сразу то и другое), преобразование сильно раздувает матрицу (датафрейм). Почти у всех моделей машинного обучения с этим большие

проблемы (не говоря уже о том, что это увеличивает время всех расчётов) — тут вспоминают про label encoding или просто избавляются от отдельных признаков.

Мы рекомендуем воспользоваться one-hot encoding, например, для таких признаков как: city_name, weather_desc, product_id, store_id.

2.1.4 Другие возможные признаки

Суммарное количество заказов за день/неделю/2 недели/несколько дней назад. Не бойтесь подсказать вашей модели, сколько примерно заказов она должна предсказать, основываясь на историческом опыте.

Среднее количество заказов неделю и 2 недели назад. Можете также добавить веса для этого признака.

Различные статистики. Скользящие средние, медианы, стандартные отклонения и любые другие статистики, которые придут вам в голову! Подробнее про скользящие окна: `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.rolling.html`

2.2 Использование моделей

Мы рекомендуем вам попробовать различные модели машинного обучения для решения этой задачи. Помимо тех, которые мы изучили в рамках курса, вы можете также обратить внимание на следующие известные фреймворки, использование которых может быть полезно:

`https://catboost.ai/`
`https://lightgbm.readthedocs.io/en/latest/index.html`

3 Заключение

Мы желаем вам удачи при выполнении этого задания! Оно составлено так, чтобы вы могли раскрыть свой творческий потенциал при его выполнении, познакомиться с новыми способами применения изученных алгоритмов на практике, а также погрузиться в реалистичную задачу, решение которой может украсить ваше резюме. Мы не бросим вас на произвол судьбы один на один с этой задачей. Вы всегда можете задать ваши вопросы или попросить совета в наших каналах, равно как и на лекциях, а также во время консультаций.

Некоторые полезные ссылки:

`https://www.kaggle.com/code/matleonard/feature-generation/notebook`
`https://studme.org/72673/ekonomika/modeli_lagovymi_peremennymi`