3. Прогнозирование временных рядов на карте

На этой неделе вам предстоит попробовать добавить в вашу регрессионную модель дополнительные признаки. Дайте волю своему воображению! Такие моменты — одни из лучших в работе специалиста в науке о данных.

Задание, оцениваемое сокурсниками: Дополнительные признаки

Срок сдачи прошел июль 2, 11:59 вечера PDT

Отправить сейчас
Выполненное задание необходимо отправить как можно раньше, чтобы сокурсники могли оценить вашу работу. В противном случае может не оказаться достаточного количества сокурсников для его оценки. Сдавайте работы как можно раньше!

1. **Инструкции**
2. **Моя работа**
3. **Обсуждения**

Инструкции

На этой неделе вам предстоит попробовать добавить в вашу регрессионную модель дополнительные признаки.

Во-первых, для прогнозирования можно использовать информацию, содержащуюся в сырых данных:

- средняя длительность поездок
- среднее количество пассажиров
- среднее расстояние по счётчику
- доли географических зон, в которые совершаются поездки
- доли поездок, совершаемых по тарифам каждого из типов
- доли способов оплаты поездок
- средняя стоимость поездок
- доли провайдеров данных

Все эти признаки можно использовать только с задержкой, то есть, при прогнозировании $y^{\wedge}_{T+i|T}$ эти признаки должны быть рассчитаны по данным не позднее момента времени T. Каждый из этих признаков можно использовать по-разному: как сырые значения за последние несколько часов, так и средние за последний день, неделю, месяц и т. д.
Во-вторых, чтобы улучшить качество прогнозов в аномальные периоды, вы можете найти информацию о потенциально влияющих на количество поездок событиях, таких, как государственные праздники. Проанализируйте, как именно поведение пассажиров меняется во время этих событий, и создайте признаки, отражающие эти изменения. Как показывает наш опыт, правильный учёт праздничных дней часто позволяет существенно уменьшить среднюю ошибку прогноза.

В-третьих, можно использовать признаки, связанные с географией. Например, скорее всего, суммарное количество поездок, совершаемых из географической зоны, пропорционально площади этой зоны. Для зон, прилегающих к аэропорту, может быть характерен специфический паттерн дневной сезонности, связанный с тем, что спрос на такси будет повышаться в те часы, когда общественный транспорт перестаёт работать. В деловом центре максимальное количество поездок будет приходиться на начало и окончание рабочего дня, на Бродвее — на время начала и окончания спектаклей. Все эти идеи не обязательно верны, мы приводим их здесь только для того, чтобы продемонстрировать принцип рассуждений. Ещё один пример географического признака: можно попробовать добавить идентификатор боро, который можно найти в файле https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv. Кроме того, нам кажется перспективным использование в качестве фактора количества поездок, совершённых за прошлый час/день и т. д. из соседних географических зон, или количества поездок, совершённых за прошлый час/день **в** текущую географическую зону.

Много примеров других признаков, которые можно использовать при регрессионном прогнозировании, можно найти в лекции Вадима Стрижова.

Чтобы сдать задание, выполните следующую последовательность действий.

1. Загрузите обучающие выборки прошлой недели, перечислите используемые в моделях признаки и посчитайте Q_{may} — качество прогнозов моделей, настроенных на данных до апреля 2016, в мае 2016.
2.
3. Попробуйте добавить признаки. Используйте идеи, которые мы предложили, или какие-то свои. Обучайте обновлённые модели на данных до апреля 2016 включительно и считайте качество новых прогнозов на мае. Удаётся ли вам улучшить качество? Не нужно ли увеличить сложность регрессионной модели? Если добавляемый признак не улучшает качество, всё равно оставьте доказательства этому в ноутбуке, чтобы ваши коллеги это видели при проверке.
4. Когда вы примете решение остановиться и перестать добавлять признаки, постройте для каждой географической зоны и каждого конца истории от 2016.04.30 23:00 до 2016.05.31 17:00 прогнозы на 6 часов вперёд; посчитайте в ноутбуке ошибку прогноза по следующему функционалу:

$$Q_{may} = \frac{1}{R * 739 * 6} \sum_{r=1}^{R} \sum_{T=2016.04.3023:00}^{2016.05.3117:00} \sum_{i=1}^{6} \left| \hat{y}^r_{T|T+i} - y^r_{T+i} \right|.$$

Убедитесь, что среднее качество прогнозов увеличилось.

5. Переобучите итоговые модели на данных до мая 2016 включительно, постройте прогнозы на июнь для каждого конца истории от 2016.05.31 23:00 до 2016.06.30 17:00 и запишите все результаты в один файл в уже знакомом вам формате: *geoID, histEndDay, histEndHour, step, y*
6. Загрузите полученный файл на kaggle: https://inclass.kaggle.com/c/yellowtaxi. Добавьте в ноутбук ссылку на сабмишн.
7. Загрузите ноутбук в форму.

Review criteriaменьше

В качестве ответа в этом задании вам нужно загрузить ноутбук; убедитесь, что ход анализа, который вы провели, описан достаточно подробно для того, чтобы ваши сокурсники поняли, что вы делали и почему.