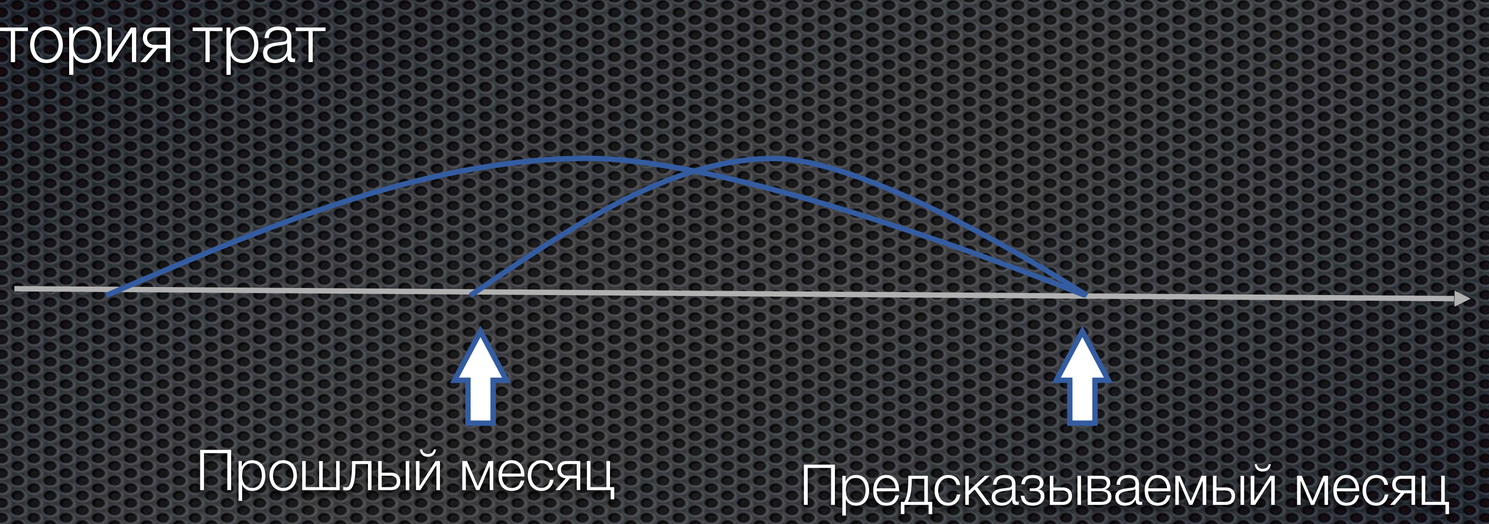
Сбербанк
Data Science Contest
Задача C

Дмитрий Алтухов

Описание задачи

- Дана история транзакций пользователей за 14 месяцев.
- Необходимо предсказать объем трат в следующем месяце в каждой из 184 категорий для 3000 пользователей.
- Мера качества - RMSLE со сдвигом 1.

Обучающая выборка



- В качестве обучающей выборки используем данные о прошлых тратах

Структура решения

1. Извлечение базовых признаков
2. Построение линейных моделей для каждого пользователя
3. Построение линейных моделей для каждого MCC кода
4. Объединение всех признаков в XGBoost
5. Легкий постпроцессинг

Базовые признаки

- Меры:
  1. Объем трат
  2. Количество трат
  3. STD трат
- Измерения:
  1. Месяц, Пользователь, MCC
  2. Месяц, Пользователь
  3. Месяц, MCC

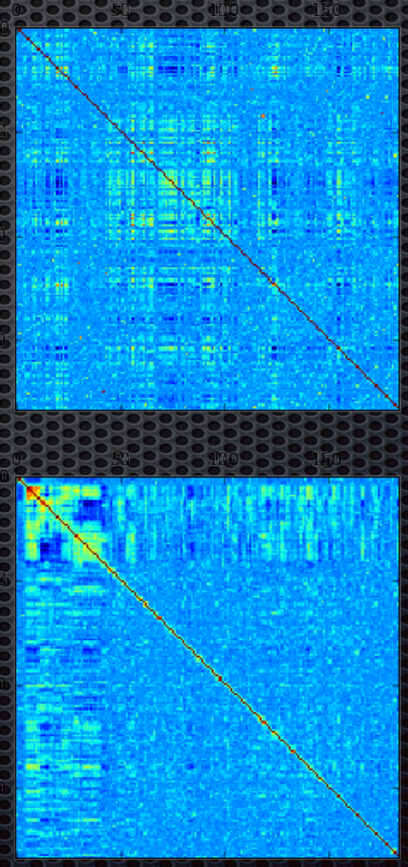
Для всех измерений считаем все меры за последние 5 месяцев

Модели в разрезе пользователя

- У каждого пользователя уникальный характер трат (регулярность, количество)
- Для каждого из них строятся сильно регуляризованные Ridge регрессии и KNN Regressor на базовых признаках (по последним 3, либо последним 5 месяцам)
- Подобные предсказания уже показывают неплохое качество на публичной части тестовый выборки 1.31

Модели в разрезе MCC

- MCC кодов не так много, как пользователей, однако структура трат по ним достаточно сложна для описания одной общей моделью
- Можно заметить, что траты по некоторым кодам связаны друг с другом, поэтому помимо базовых признаков в модель добавляются траты за предыдущий месяц по каждому из кодов
- В качестве модели используется BayesianRidge



Итоговая модель

- В итоговую модель включены как базовые признаки, так и предсказания, полученные с помощью пользовательских/MCC регрессий
- В качестве алгоритма машинного обучения используется XGBoost
- Если предсказание получается < 0, то заменяем его на 0

Используемые инструменты

- IPython ноутбуки. Стандартный стек из pandas +numpy+sklearn+xgboost
- Считал сначала на MBP Pro 16GB, затем переехал на амазоновский m4.10xlarge. Извлечение признаков занимает около 2 часов, примерно столько же обучение XGBoost’a

Спасибо за внимание!