Sberbank Data Science Contest
Задача B

Василий Рубцов

12 ноября 2016

Задача B

- Прогнозирование временного ряда на 30 дней вперед.
- Всего 148 рядов.
- Метрика $\text{RMSLE}_{500} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 500) - \log(\hat{y}_i + 500))^2}$

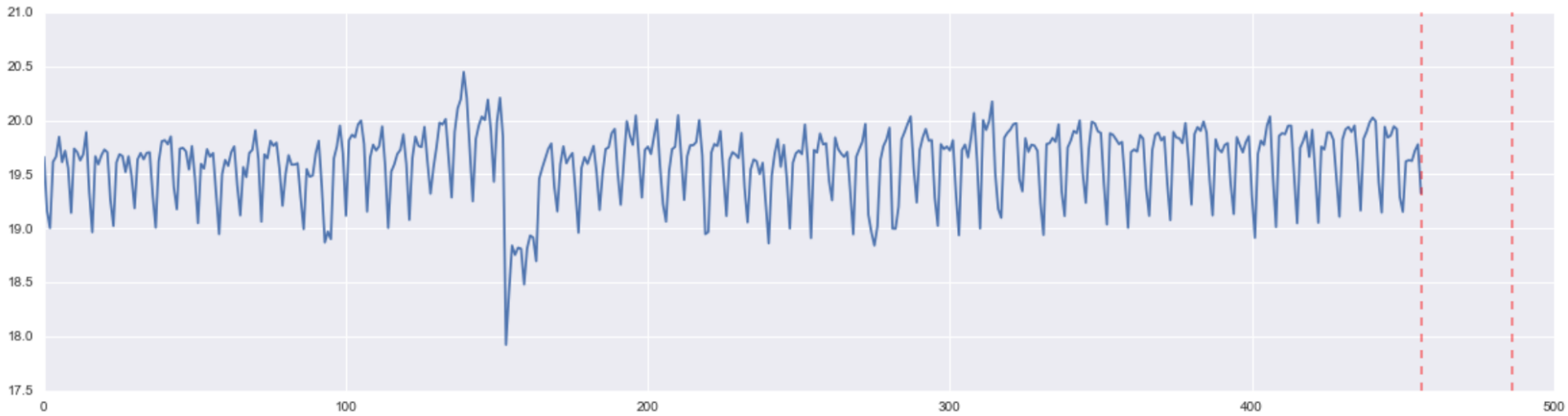

Рис. Код 6010 (финансовые институты — снятие наличности вручную)

Создание выборки для обучения и для тестирования

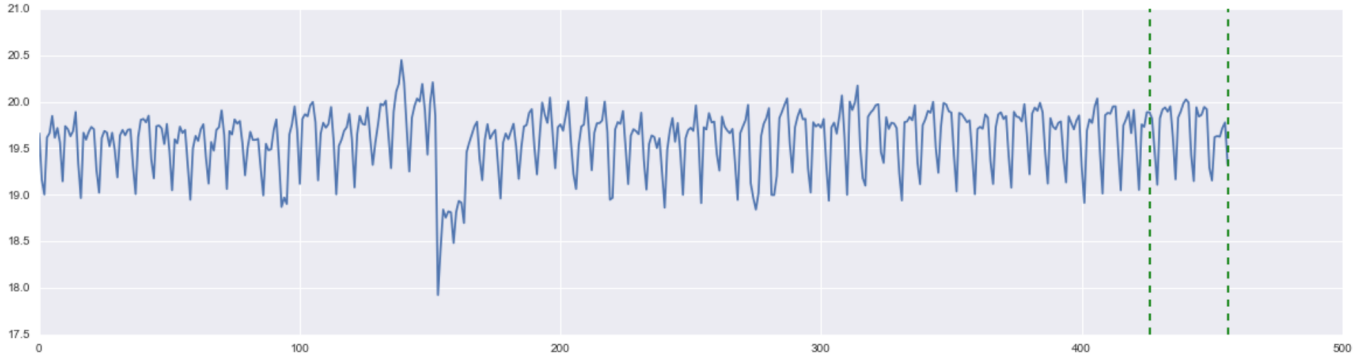

Рис. Промежуток времени для тестирования

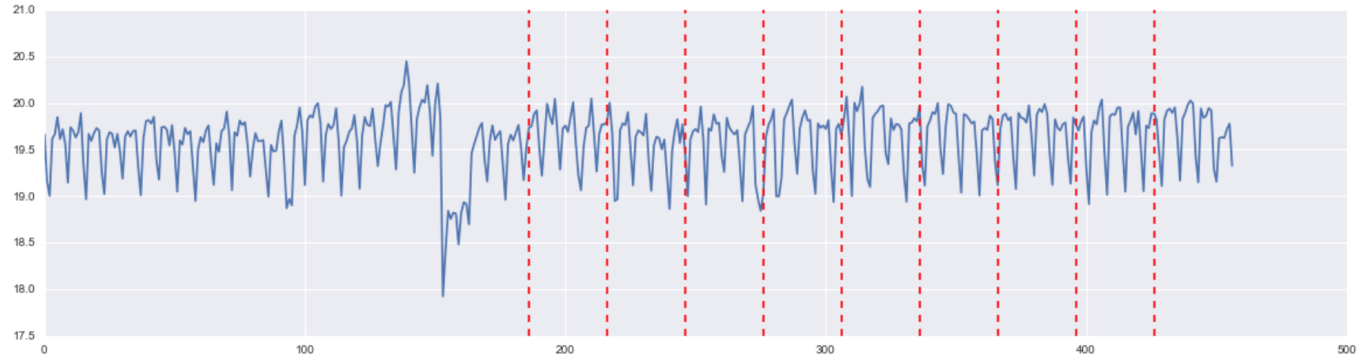

Рис. Промежутки времени для обучения

Обработка категориальных признаков

Label encoding vs. one-hot encoding

Mcc code
742
1711
1731
742
742
1731

Code 742	Code 1711	Code 1731
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1

Обработка категориальных признаков

Label encoding vs. one-hot encoding

Mcc code
742
1711
1731
742
742
1731

Code 742	Code 1711	Code 1731
1	0	0
0	1	0
0	0	1
1	0	0
1	0	0
0	0	1

Mcc code	Mean
742	7.74
1711	7.12
1731	6.69
1799	6.96
2741	6.56
3000	14.35

Mcc code	Mean	Encoding
2741	6.56	0
1731	6.69	1
1799	6.96	2
1711	7.12	3
742	7.74	4
3000	14.35	5

Базовые модели

Рис. Прогнозы базовых моделей

Базовые модели

Рис. Прогнозы базовых моделей

Простые модели:

- Средние за последнюю неделю/месяц/год.
- Средне взвешенное
- Среднее по дням неделям
- Линейная регрессия по всем данным / по дням неделям / за последний месяц

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$
- Блендинг: $w_1 y_{i_1} + ... + w_k y_{i_k}$,
  где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$
- Блендинг: $w_1 y_{i_1} + ... + w_k y_{i_k}$,
  где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.
- Обобщение блендинга: $w_{1,c,t}(e_{c,t})y_{i_1} + ... + w_{k,c,t}(e_{c,t})y_{i_k}$

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$
- Блендинг: $w_1 y_{i_1} + ... + w_k y_{i_k}$,
  где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.
- Обобщение блендинга: $w_{1,c,t}(e_{c,t})y_{i_1} + ... + w_{k,c,t}(e_{c,t})y_{i_k}$

Пусть $e_{c,t} = (e_{1,c,t}, ..., e_{k,c,t})$ — ошибки $y_{i_1}, ..., y_{i_k}$ при фиксированном коде c и времени t

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$
- Блендинг: $w_1 y_{i_1} + ... + w_k y_{i_k}$,
  где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.
- Обобщение блендинга: $w_{1,c,t}(e_{c,t})y_{i_1} + ... + w_{k,c,t}(e_{c,t})y_{i_k}$

Пусть $e_{c,t} = (e_{1,c,t}, ..., e_{k,c,t})$ — ошибки $y_{i_1}, ..., y_{i_k}$ при фиксированном коде c и времени t

$$\hat{w}_{i,c,t}(e_{i,c,t}) = \frac{1}{e_{i,c,t}^{\alpha}}, \quad \alpha > 1$$

Комбинации базовых моделей

Пусть $y_1, ..., y_n$ — набор базовых моделей.

- Разность некоторых моделей: $y_i - y_j$
- Блендинг: $w_1 y_{i_1} + ... + w_k y_{i_k}$,
  где $w_1 + ... + w_k = 1$ и $w_i \geqslant 0$ для $i = 1, ..., n$.
- Обобщение блендинга: $w_{1,c,t}(e_{c,t})y_{i_1} + ... + w_{k,c,t}(e_{c,t})y_{i_k}$

Пусть $e_{c,t} = (e_{1,c,t}, ..., e_{k,c,t})$ — ошибки $y_{i_1}, ..., y_{i_k}$ при фиксированном коде c и времени t

$$\hat{w}_{i,c,t}(e_{i,c,t}) = \frac{1}{e_{i,c,t}^{\alpha}}, \quad \alpha > 1$$

$$w_{i,c,t}(e_{c,t}) = \frac{\hat{w}_{i,c,t}(e_{i,c,t})}{\hat{w}_{1,c,t}(e_{1,c,t}) + ... + \hat{w}_{k,c,t}(e_{k,c,t})}$$

Улучшение качества

1) Код.
2) День, на который прогнозируем (от 1 до 30).
3) Доля нулевых значений в истории данного кода.
4) Базовые модели:
   - y_1 — Среднее за все время
   - y_2 — Среднее за последние пол года
   - y_3 — Среднее за последний месяц
   - y_4 — Среднее за последнюю неделю
   - y_5 — Значение в последний известный день
   - y_6 — Средне взвешенное за последние 50 дней
   - y_7 — Среднее по дням неделям
   - y_8 — Среднее по дням неделям за последние пол года
   - y_9 — Линейная регрессия
   - y_{10} — Линейная регрессия по дням неделям
   - y_{11} — Линейная регрессия на данных за последние пол года
   - y_{12} — Линейная регрессия на данных за последний месяц

Улучшение качества

1) Код.
2) День, на который прогнозируем (от 1 до 30).
3) Доля нулевых значений в истории данного кода.
4) Базовые модели:
   - y_1 — Среднее за все время
   - y_2 — Среднее за последние пол года
   - y_3 — Среднее за последний месяц
   - y_4 — Среднее за последнюю неделю
   - y_5 — Значение в последний известный день
   - y_6 — Средне взвешенное за последние 50 дней
   - y_7 — Среднее по дням неделям
   - y_8 — Среднее по дням неделям за последние пол года
   - y_9 — Линейная регрессия
   - y_{10} — Линейная регрессия по дням неделям
   - y_{11} — Линейная регрессия на данных за последние пол года
   - y_{12} — Линейная регрессия на данных за последний месяц

RMSLE $= 1.538$

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

1.5284

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284
6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284

6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
   1.5263

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284
6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
   1.5263
7) Переобозначение кодов.

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284
6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
   1.5263
7) Переобозначение кодов.
   1.5188

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284
6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
   1.5263
7) Переобозначение кодов.
   1.5188
8) $y_{14} = w_1(e)y_7 + w_2(e)y_{10} + w_3(e)y_6 + w_4(e)y_2$

Улучшение качества

5) Некоторые разницы
   - $d_1 = y_1 - y_7$
   - $d_2 = y_1 - y_2$
   - $d_3 = y_5 - y_{10}$

   1.5284
6) $y_{13} = w_1 y_8 + w_2 y_3 + w_3 y_{10}$
   1.5263
7) Переобозначение кодов.
   1.5188
8) $y_{14} = w_1(e) y_7 + w_2(e) y_{10} + w_3(e) y_6 + w_4(e) y_2$
   1.51259

Корректировка предсказаний некоторых кодов

Ноябрь 2014 года!

- 3501 (жилье — отели, мотели, курорты)
- 4722 (туристические агентства и организаторы экскурсий)
- 6211 (ценные бумаги: брокеры/дилеры)

Спасибо за внимание!