Адаптивный кредитный скоринг

Астахов Антон

Московский физико-технический институт

24.04.19

Цель работы

Исследуются

Методы прогнозирования вероятности дефолта клиента банка в следущие 12 месяцев

Требуется

Предложить алгоритм аккумулирования раличных исторических промежутков выплат клиента с целью прогнозирования длинной переменной

Проблемы

- извлечение максимально информативных исторических данных
- учитывание извлеченных исторических данных

Литература

[1] Shweta Arya, Catherine Eckel, Colin Wichman. Anatomy of the credit score

[2] Engku Muhammad Nazri E. A. Bakar. Credit scoring models: techniques and issues

[3] Liran Einav, Mark Jenkins, Jonathan Levin. The impact of credit scoring on consumer lending.

Входные данные

Имеется информация (исторические данные) о парах (клиент, дата): ежемесячные платежы клиента банка.

Признаки клиента

Выделено признаковое подпространство из R^3

- $utilization\_3m\_0\_old\_1\_lin$
- $max\_overdue\_days\_cnt\_0\_old\_1\_lin$
- $sloppy\_0\_old\_1\_lin$

Входные данные

Информация о платежах

- $dYpX$ - флаг достижения Y просрочек за следующие X месяцев после $curr\_due\_dt$.
  $dYpX$ определен: $Y \in \{1..4\}$, $X \in \{1..12\}$, $Y <= X \Rightarrow$
  U - множество значений (Y, X)
- $curr\_due\_dt$ - дата на которую, строится прогноз.
  Пусть $finish\_dt$ - дата созревания целевой переменной $d4p12$. $(curr\_due\_dt = finish\_dt - 12 * len\_of\_month)$.
  $curr\_due\_dt \in [2008.06.05, 2019.03.07]$

Вектор логистической регресии

Для прогноза $dYpX$, как вероятности, определяется $curr\_due\_dt$ и $finish\_dt$, где $curr\_due\_dt = finish\_dt - x * len\_of\_month$.

Если мы имеем данные, то модель логистической регресии, обучаясь на клиентах из интрервала $(curr\_due\_dt - len\_of\_month,\ curr\_due\_dt)$, предсказывает вероятность $dYpX$ на $finish\_dt$.

Результат логистической регрессии

$W^t_{y,\,x} \in R^3$ и $B^{\,t}_{y,\,x} \in R$, где $t = finish\_dt$.

Обозначим $V^t_{y,\,x} = \begin{bmatrix} B^{\,t}_{y,\,x} \\ W^t_{y,\,x} \end{bmatrix}$.

Вектор логистической регресии

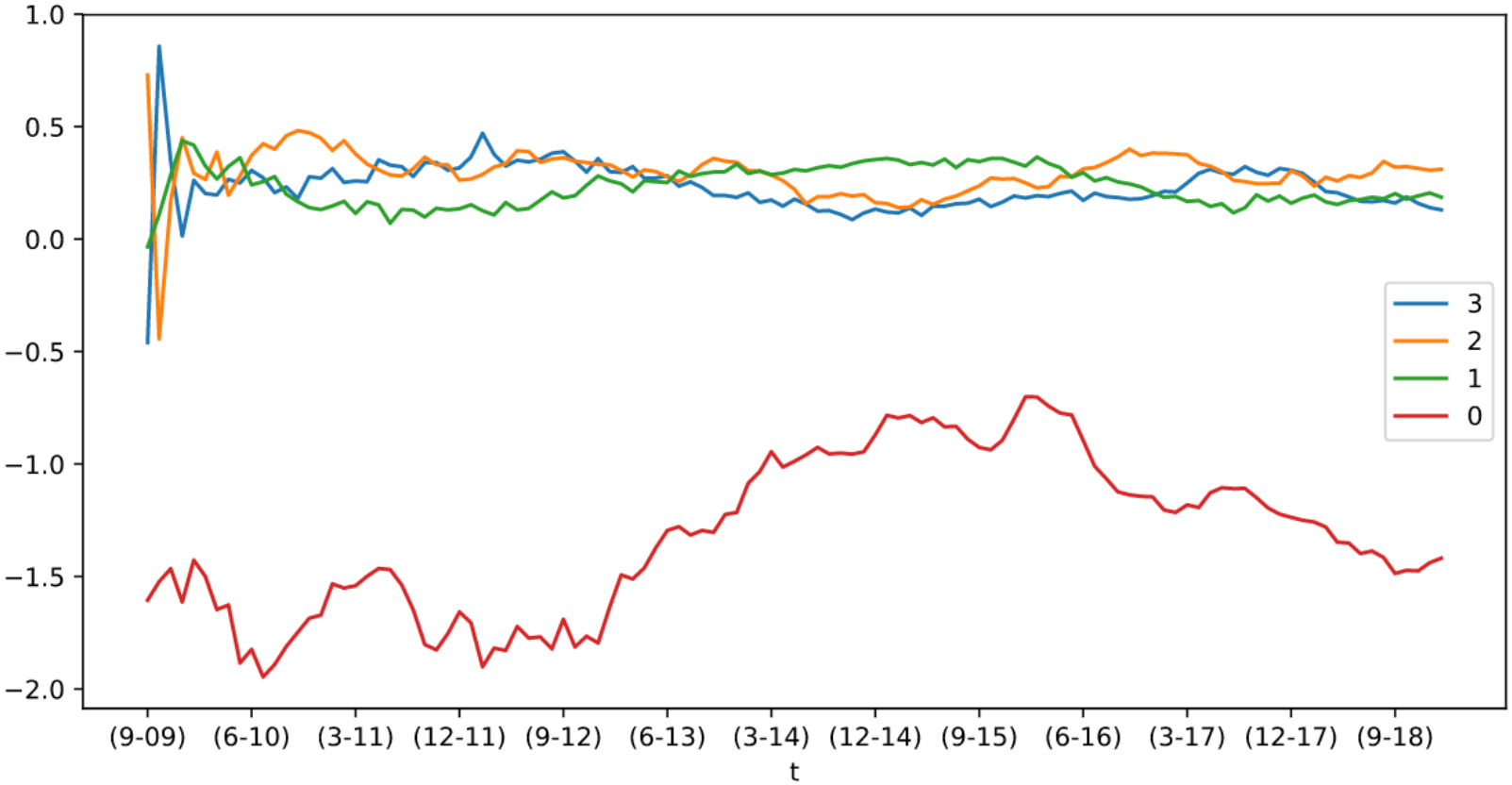


Рис.: $V^t_{4,12}$

Использование данных

Введем понятие "созревший"месяц для переменной $dYpX$, как месяц, включая X следующих, принадлежит множеству месяцев, о которых у банка есть информация.

$\exists V^t_{y,\,x} \Longleftrightarrow (t-x)$ созревший для переменной $dYpX$.

Базовый алгоритм

Условие

Необходимо спрогнозировать дефолт клиентов на периоде $[t, t+12]$, то есть вероятность просрочки $d4p12$

Прогнозирование

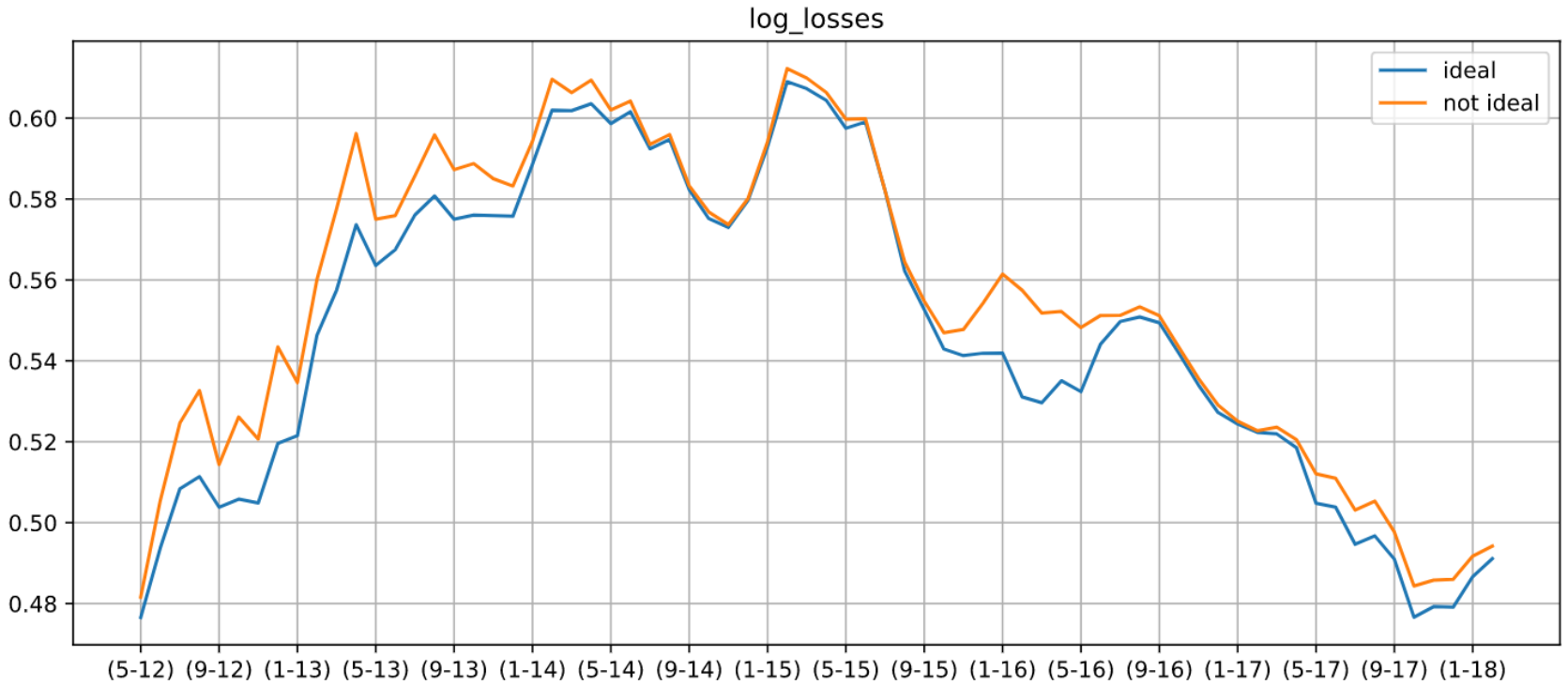
Имеем $V_{4,12}^{t}$. С помощью этого вектора предскажем вероятность $d4p12$, которую должнен был бы описывать $V_{4,12}^{t+12}$

Модели

- ideal - модель $V_{4,12}^{t+12}$
- not_ideal - модель $V_{4,12}^{t}$(базовый алгоритм)

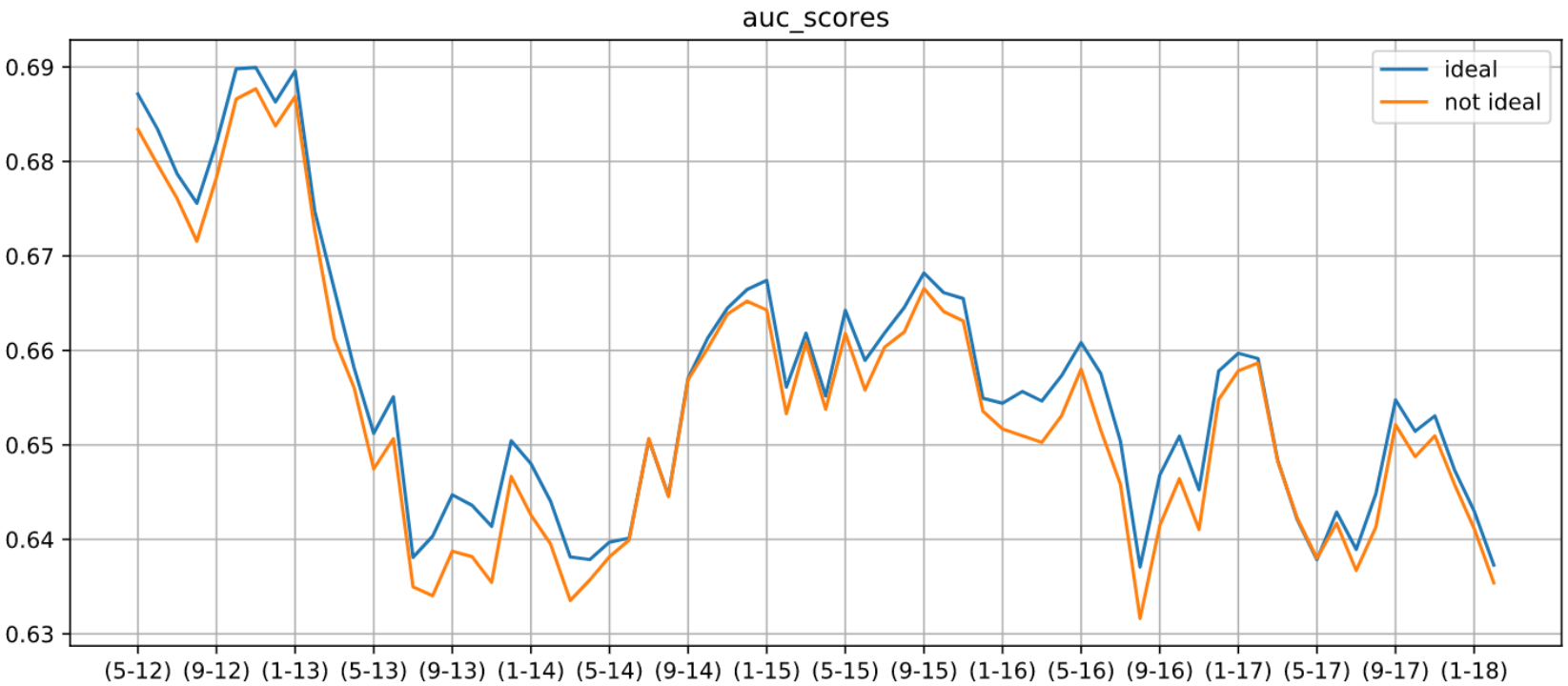
Рабочая область

Качество на тестовых месяцах



Рабочая область

Качество на тестовых месяцах



Изменение вектора логистической регресии

Введем $\Delta V_{y,\,x}^{t,\,t+i} = V_{y,\,x}^{t+i} - V_{y,\,x}^{t}$ - приращение вектора весов для переменой $dYpX$ с $finish\_dt = t$ до $finish\_dt = t + i$, где t итерация по месяцам.

$$\exists V_{y,\,x}^{t} \Longleftrightarrow (\forall i > 0, j \geq 0)\ \exists\ \Delta V_{y,\,x}^{t-i,\,t-j}$$

Переформулировка задачи

предсказать $\Delta V_{4,\,12}^{t,\,t+12}$

Изменение вектора логистической регресии

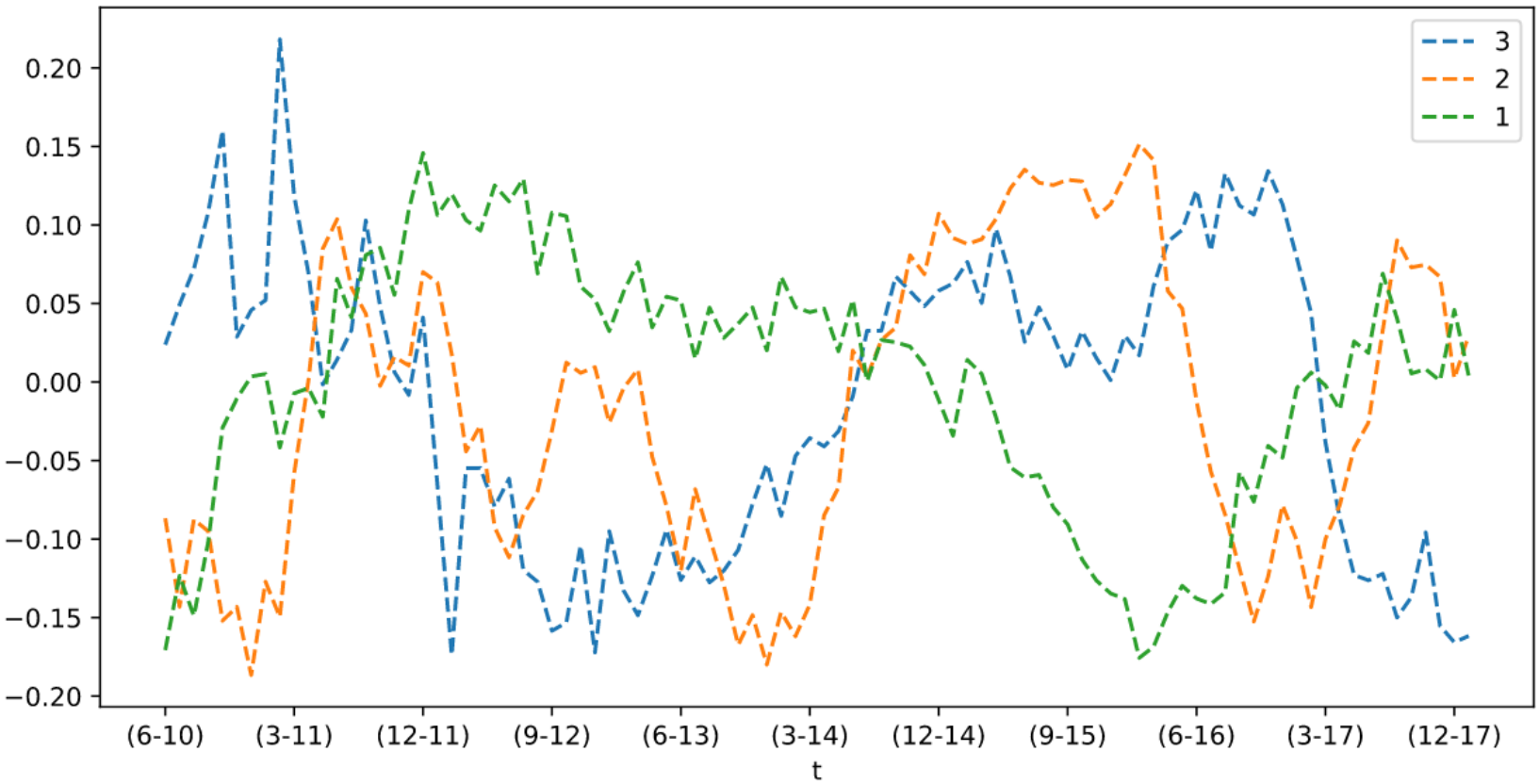


Рис.: $\Delta V^t_{4,12}$

Система векторов

Для прогнозирования вектора $\Delta V_{4,12}^{t,t+12}$ будем использовать систему векторов $\{\Delta V_{y,\,x}^{t-i,\,t-j}\}$ $(Y,X) \in U, \forall i > 0, j \geq 0$

Разложение по системе

$$\Delta V_{4,\,12}^{t,\,t+12} = \sum_{y,\,x}^{U} C_{y,\,x} * \Delta V_{y,\,x}^{t-i,\,t-j}$$

Выбор системы

Подходы

- Будем разбивать нашу историю с помощью коротких просрочек: $U^{'} = (Y, X) \in U \& X \leq 6$
- В нашу систему войдут все последение данные, которые мы имеем: $j = 0$
- Приращение вектора коротких просрочек состоит из векторов, которые вычисляются на непересекающихся интревалах времени: $\Delta V_{y,\,x}^{t-x,\,t} = V_{y,\,x}^{t} - V_{y,\,x}^{t-x}$ и $\Delta V_{y,\,x}^{t-2x,\,t} = V_{y,\,x}^{t-x} - V_{y,\,x}^{t-2x}$
- Делать не одно разложение для всего вектора: покомпонентно $\Delta V_{y,\,x}^{i}$, по подпрастроству(отдельно свободный член $\Delta B_{y,x}$ и $\Delta W_{y,x}$)

Выбор системы

Итоговое разложение

$$\Delta V_{4,\,12}^{t,\,t+12} = \sum_{y,\,x}^{U'} C_{y,\,x} * \Delta V_{y,\,x}^{t-x,\,t} + \sum_{y,\,x}^{U'} C'_{y,\,x} * \Delta V_{y,\,x}^{t-2x,\,t-x}$$

Параметризация

$$C_{y,x} = a * X + b * Y + c * Y * X + d$$

Метрики обучения

- $\|\Delta V_{4,\,12}^{t,\,t+12} - \Delta V(predict)_{4,\,12}^{t,\,t+12}\|$
- $LogLoss(d4p12, \sigma(X \cdot (\Delta V(predict)_{4,\,12}^{t,\,t+12} + V_{4,12}^{t}))^T)$

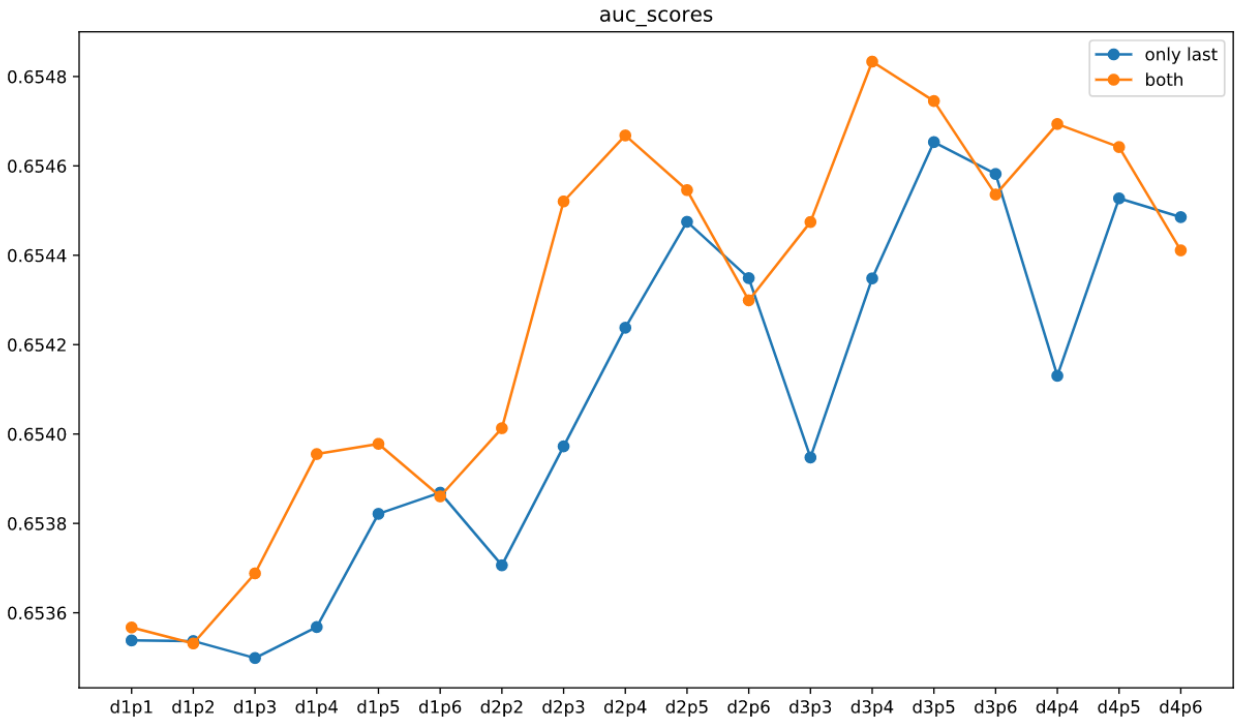
Короткие вектора

Будем прогнозировать приращение вектора отдельно для $\Delta B_{y,x}$ и $\Delta W_{y,x}$.

Модели

- *only_last*
  - $\Delta W_{4,\,12}^{t,\,t+12} = C_{y,\,x}^{w} * \Delta W_{y,\,x}^{t-x,\,t}$
  - $\Delta B_{4,\,12}^{t,\,t+12} = C_{y,\,x}^{b} * \Delta B_{y,\,x}^{t-x,\,t}$
- *both*
  - $\Delta W_{4,\,12}^{t,\,t+12} = C_{y,\,x}^{w} * \Delta W_{y,\,x}^{t-x,\,t} + C_{y,\,x}^{w'} * \Delta W_{y,\,x}^{t-2x,\,t-x}$
  - $\Delta B_{4,\,12}^{t,\,t+12} = C_{y,\,x}^{b} * \Delta B_{y,\,x}^{t-x,\,t} + C_{y,\,x}^{b'} * \Delta B_{y,\,x}^{t-2x,\,t-x}$

Only_last, Both



Вывод

Вся полезная информация содержится в 15 месяцах от текущей даты

Сравнение моделей

	roc_auc_score	log_loss
all_shorts(both)	0.653319	0.562852
all_shorts(only last)	0.65512	0.558857
d3p4(both)	0.654833	0.55493
d3p5(only last)	0.654653	0.555803
separately(only last)	0.654506	0.559405
from d3p6	0.654435	0.563361
ideal	0.656476	0.547124
not ideal	0.653658	0.554932

Заключение

Итог

- исследованы методы поправки базовой модели
- найдено окно максимальной информативности
- получена параметризация коротких векторов
- сравнены методы обучения вектора