

Адаптивный кредитный скоринг

Задача:

предсказать вероятность дефолта $d4p12$ (столбец pd) определенного клиента

Условие:

имеем данные о ежемесячных платежах клиентов банка и его различные признаки:

1. $curr_due_dt$ - дата на которую, строится прогноз. Пусть $finish_dt$ - дата созревания целевой переменной ($curr_due_dt = finish_dt - 12month$).

Интервал времени $curr_due_dt$ с 2008-06-05 до 2018-03-07.

2. $utilization_3m_0_old_1_lin$, $max_overdue_days_cnt_0_old_1_lin$, $sloppy_0_old_1_lin$ - признаки клиента.

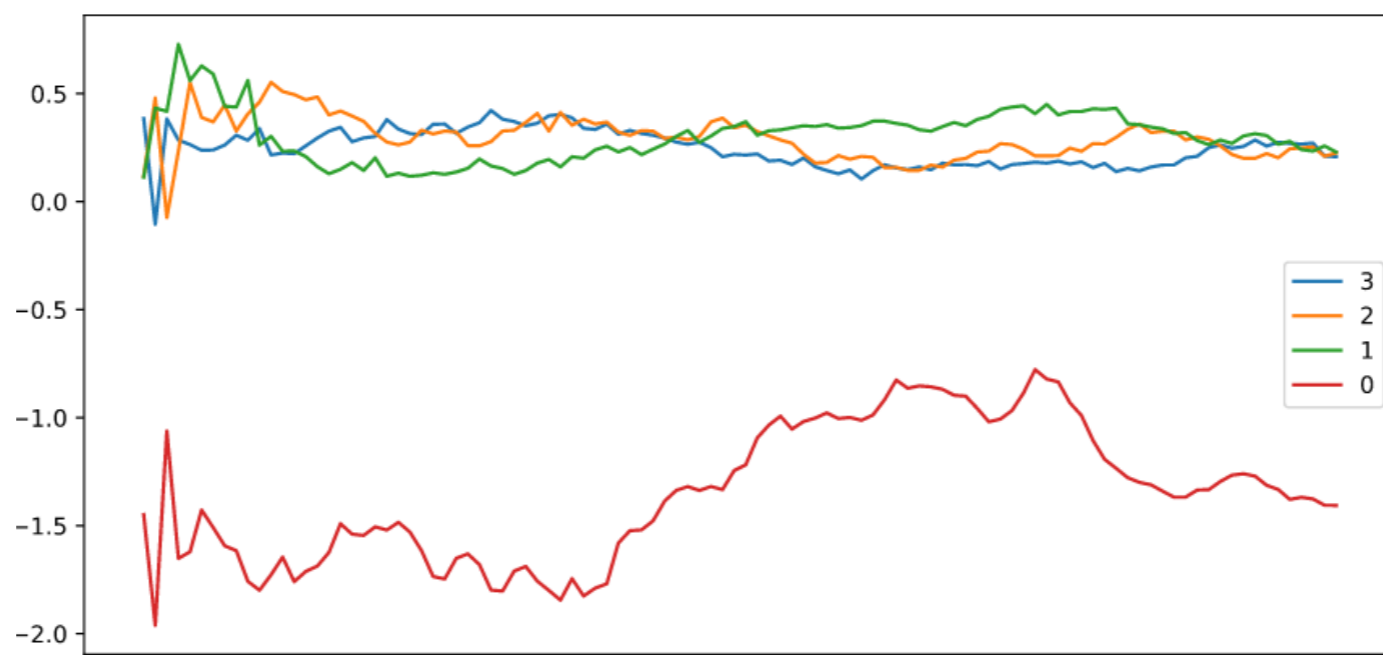
3. $dYpX$ - флаг достижения Y просрочек за следующие X месяцев после $curr_due_dt$.

$dYpX$ определен: $Y \in \{1..4\}$, $X \in \{1..12\}$, $Y \leq X \Rightarrow U$ - множество значений (Y, X) .

Решение:

0. Для прогноза $dYpX$, как вероятности, определяется $curr_due_dt$ и $finish_dt$. Модель логистической регрессии, обучаясь на данных из интервале $(curr_due_dt - month, curr_due_dt)$, предсказывает вероятность $dYpX$. Как результат, получаем веса логистической регрессии $W_{y,x} \in R^3$.

```
In [15]: from IPython.core.display import Image, display
display(Image('w4_12.png', width=700, unconfined=True))
```



1. Введем $\Delta W_{y,x}^{t,t+i} = W_{y,x}^{t+i} - W_{y,x}^t$ - приращение вектора весов для переменной $dYpX$ с $finish_dt = t$ до $finish_dt = t + i$, где t итерация по месяцам. Возьмем отрезок для t длины $num_month = 95$ с началом в 2010-01-01 ...

Переформулировка задачи(информация по клиентам до месяца t включительно):

предсказать $\Delta W_{4,12}^{t,t+12}$ ($i = 12$)

Решение:

... Введем понятие "созревший" месяц для переменной $dYpX$, как месяц, включая X следующих, принадлежит множеству месяцев, о которых у банка есть информация. То есть $\exists W_{y,x}^t$, если месяц $(t - x)$ созревший. А если $\exists W_{y,x}^t \Rightarrow \forall (i > 0) \exists \Delta W_{y,x}^{t-i,t}$

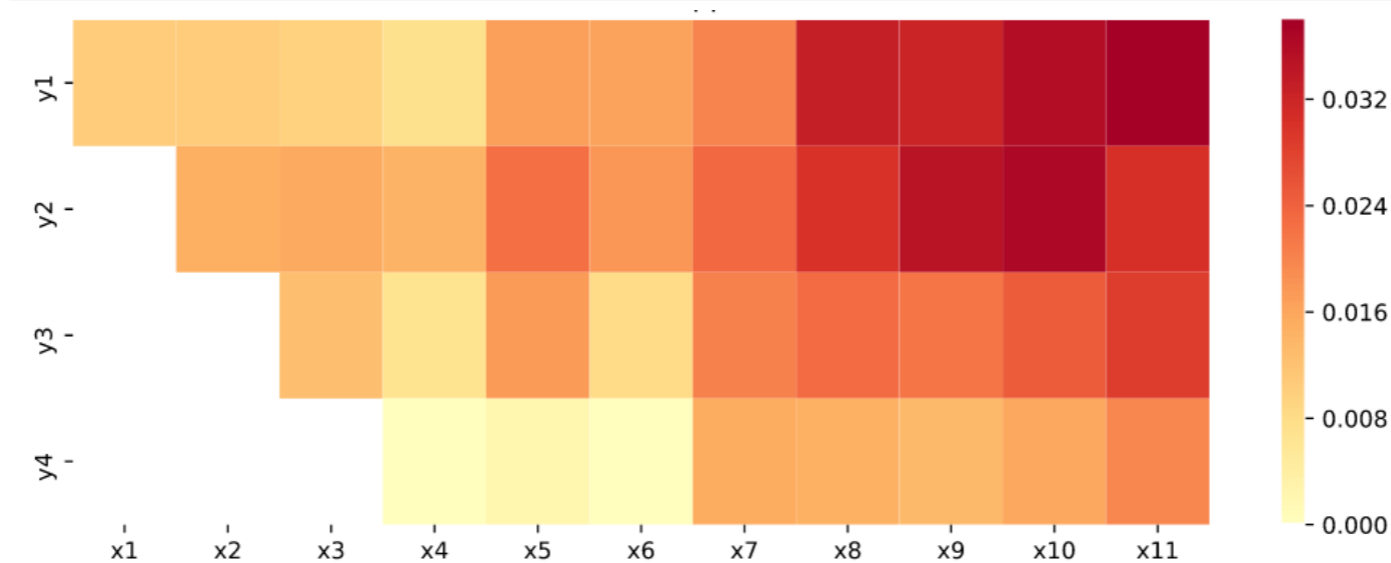
Выберем систему векторов, по которой мы будем раскладывать $\Delta W_{4,12}^{t,t+12}$. Для определение системы выберем:

а) $W_{y,x}^-$, то есть короткие переменные

б) i и j вектора $W_{y,x}^{t-i,t-j}$ для каждой переменной $W_{y,x}^-$, то есть паре (y, x) сопоставим пару (i, j)

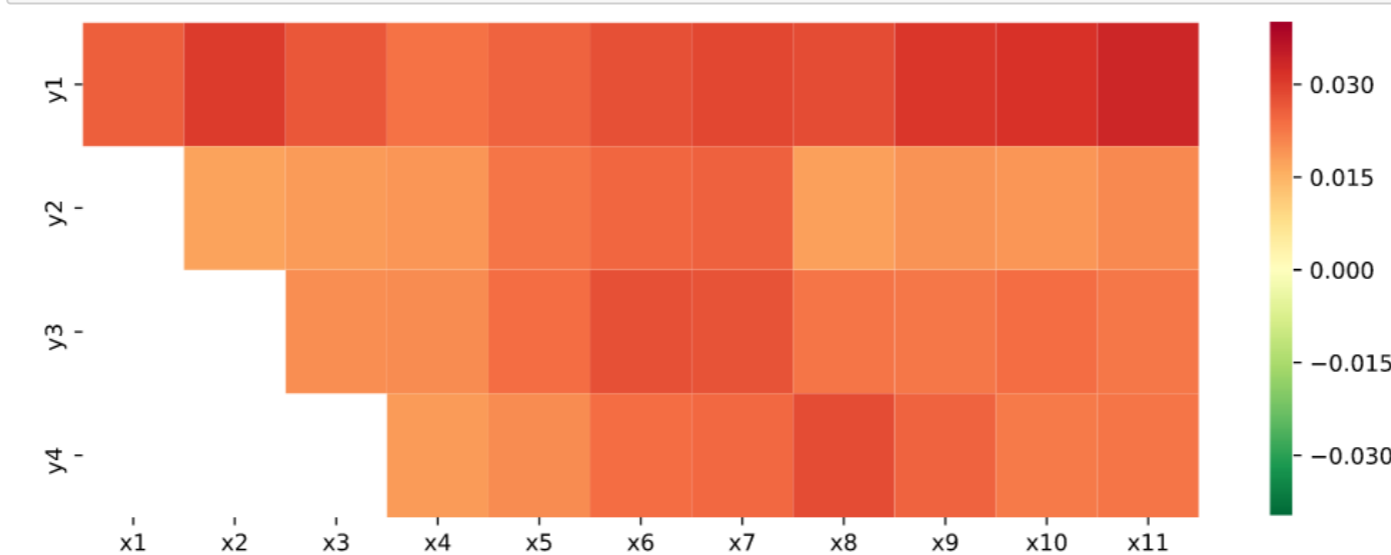
1.1 $\Delta W_{4,12}^{t,t+1} = \sum_{y,x}^V C_{y,x} * \Delta W_{y,x}^{t-1,t}$, то есть разложение реализуют различные переменные, использующие самые новые данные, иначе говоря, с совпадающей $finish_dt$, в плане месяца.

```
In [16]: display(Image('c1.png', width=700, unconfined=True))
```



1.2 $\Delta W_{4,12}^{t,t+1} = \sum_{y,x}^V C_{y,x} * \Delta W_{y,x}^{t-(12-x), t-(12-x+1)}$, то есть разложение реализуют различные переменные с совпадающей $curr_due_dt$ между собой и переменной $d4p12$, в плане месяца.

```
In [18]: display(Image('c2.png', width=700, unconfined=True))
```



Будем обучать коэффициенты $C_{y,x}$, на всей нашей выборке. Из-за недостаточного размера выборки создадим 10 "банков". Для каждого клиента инициализируем "банк рождения" по начальному pd , он же будет для клиента "банком сметри". Внутри "банка" каждый месяц будет считаться модель независимо от других "банков", то есть обучающая выборка для коэффициентов $C_{y,x}$ увеличится в 10 раз.