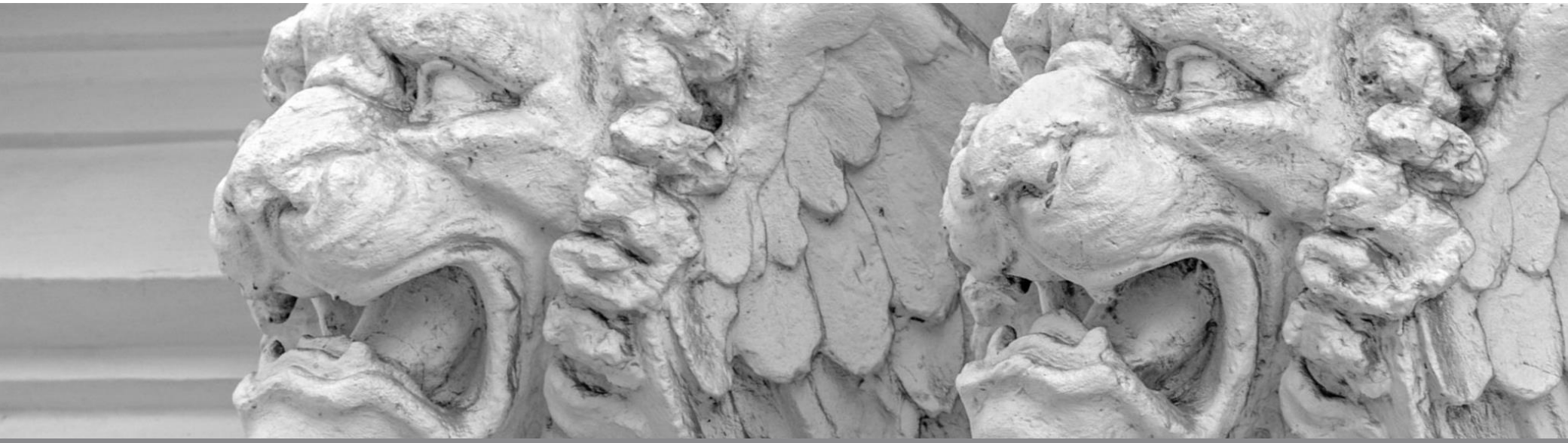
Банк России
Центральный банк Российской Федерации



А. Могилат

Департамент денежно-кредитной политики, Банк России

О методах оценки вероятности при работе с редкими событиями

Материал отражает личную позицию автора, которая может не совпадать с официальной позицией Банка России

Для чего может пригодиться оценка вероятности редких событий

1. Какая доля проблемных промышленных компаний ожидается в России через год?

2. Из-за чего возникают вспышки эпидемий отдельными заболеваниями?

3. Когда случится следующий мировой экономический кризис?

К чему могут привести стандартные подходы (1/1)

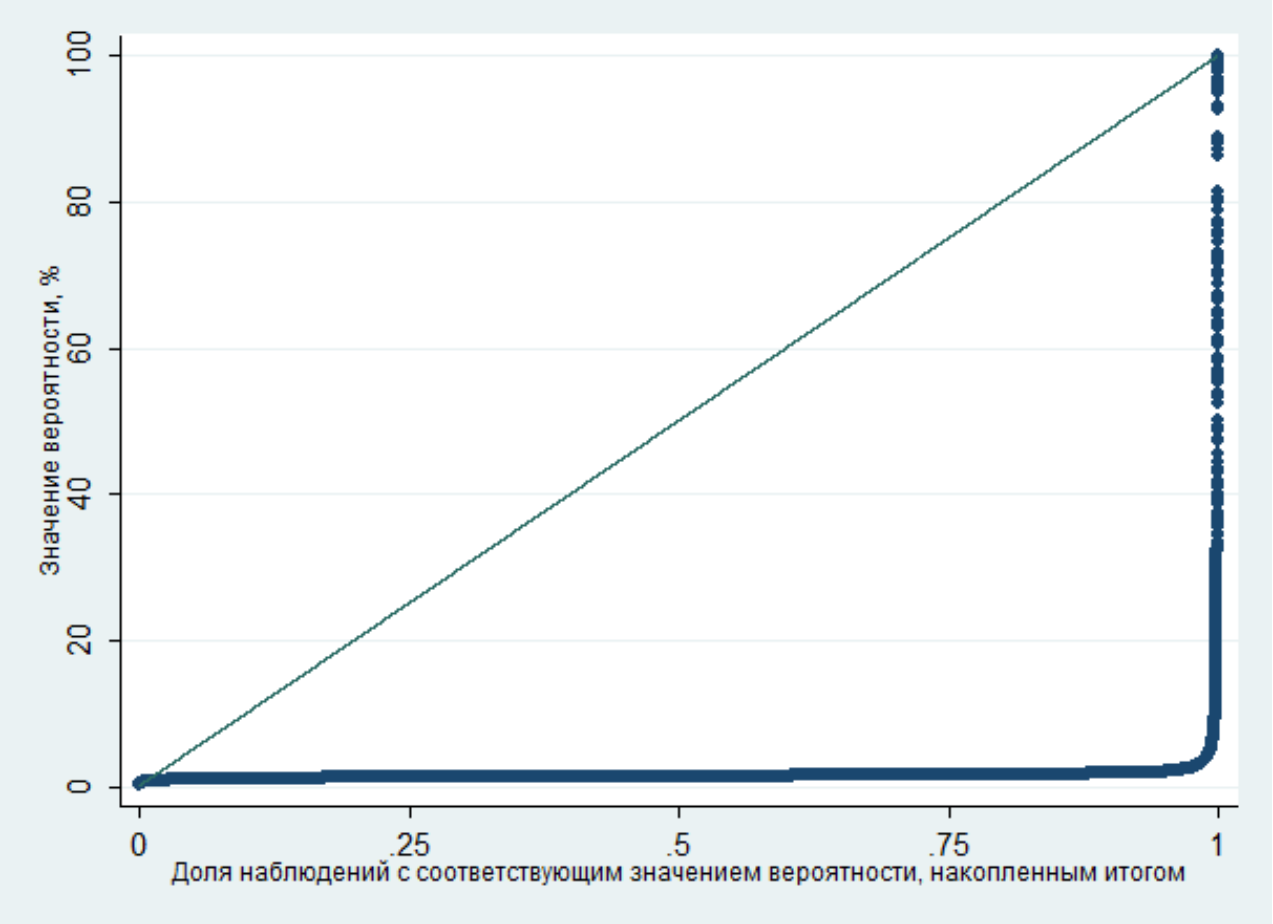
```
Logistic regression                                     Number of obs   =     185542
                                                        LR chi2(3)      =     865.12
                                                        Prob > chi2     =     0.0000
Log likelihood = -14050.838                             Pseudo R2       =     0.0299

------------------------------------------------------------------------------
         bnk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rent |  -1.463966   .0578399   -25.31   0.000    -1.577331   -1.350602
         z_a |   .3508561   .0257143    13.64   0.000     .3004571    .4012552
        dsr3 |   .0247035   .0126086     1.96   0.050    -9.04e-06    .0494159
       _cons |  -4.203102   .0200286  -209.85   0.000    -4.242357   -4.163846
------------------------------------------------------------------------------
```

- **Зависимая переменная – банкротство (1 – банкрот, 0 - НЕбанкрот);**
- **Rent** – рентабельность продаж; **z_a** - чистая кредиторская задолженность к активам; **dsr3** – долговая нагрузка (объем долга к выручке); **cons** – константа
- Данные российских промышленных компаний с 2006 по 2016 гг.
- Все переменные значимы, знаки соответствуют ожиданиям
- **НО**…

Подробнее о спецификации – см. (Донец, Могилат, 2017)

К чему могут привести стандартные подходы (2/1)



- Медианная вероятность ≈ 0;
- Очень длинный правый хвост распределения

- вероятность банкротства ≠ риск банкротства;
- недооценка рисков одних компаний и переоценка рисков других

Классическая модель бинарного выбора

Функция правдоподобия:

$$L_{ML}(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i} \xrightarrow[\beta]{} \max$$

Лог-преобразование:

$$\log L_{ML}(\beta) = \sum_{i=1}^{n} \left[y_i \log \pi\left(x_i, \beta\right) + \left(1-y_i\right) \log\left(1-\pi\left(x_i, \beta\right)\right)\right]$$

ММП-оценка:

$$q_{ML} : \left(\frac{\partial \log L(\beta)}{\partial \beta} \right) = 0$$

Подходы к учету редкости события

Эконометрические методы (1)

В работе (McCullagh, Nelder, 1989) показано, что смещение оценки в логит-модели имеет вид:

$$bias(\hat{\beta}) = \left(X'WX\right)^{-1} X'W\xi$$

где: $\xi_i = 0{,}5\, Q_{ii}\left[(1+w_1)\hat{\pi}_i - w_1\right]$

Q_{ii} – диагональные элементы матрицы

$W = diag\left\{\hat{\pi}_i\left(1-\hat{\pi}_i\right)w_i\right\}$ – матрица весов $Q = X\left(X'WX\right)^{-1}X'$

$w_i = w_1 Y_i + w_0\left(1-Y_i\right)$ – вес наблюдения i (элемент матрицы весов W), при этом $Y_i \sim Bernoulli\left(Y_i \mid \pi_i\right)$

$$\pi_i = \frac{1}{1+e^{-x_i\beta}}$$

Оценка с коррекцией на величину смещения имеет вид: $\tilde{\beta} = \hat{\beta} - bias(\hat{\beta})$

Эконометрические методы (2)

- **Модификация функции правдоподобия**
  - Метод штрафов (Firth, 1993)
  - Метод двойных штрафов (Gao, Shen, 2007)
- **Корректировка константы**
  - «эндогенная»
  - «экзогенная»

Методы модификации функции правдоподобия

1. Метод штрафов (Firth, 1993)

- Функция правдоподобия:

$$L_{PML}(\beta) = L_{ML}(\beta)\left|i(\beta)\right|^{1/2}$$

- Лог-преобразование:

$$\log L_{PML}(\beta) = \log L_{ML}(\beta) + (1/2)\log\left|i(\beta)\right|$$

- ММП-оценка:

$$q_{PML} = q_{ML} + (1/2)tr\left[i^{-1}\left(\frac{\partial i}{\partial \beta}\right)\right]$$

$\left|i(\beta)\right|^{1/2}$ – инвариантный к параметризации модели приор (Jeffreys, 1946);

i – определитель информационной матрицы Фишера

Методы модификации функции правдоподобия

2. Метод двойных штрафов (Gao, Shen, 2007)

- Функция правдоподобия: $L_{PML}(\beta) = L_{ML}(\beta)\left|i(\beta)\right|^{1/2} e^{-\lambda\|P\beta\|^2}$
- Лог-преобразование: $\log L_{PML}(\beta) = \log L_{ML}(\beta) + (1/2)\log\left|i(\beta)\right| - \lambda\|P\beta\|^2$
- ММП-оценка: $q_{PML} = q_{ML} + (1/2)tr\left[i^{-1}\left(\frac{\partial i}{\partial \beta}\right)\right] - 2\lambda\beta$

P – матрица линейных ограничений для параметров β;

λ – параметр, отвечающий за жесткость ограничений;

|| || – евклидова норма

Апробация на российских данных о банкротствах (1/2)

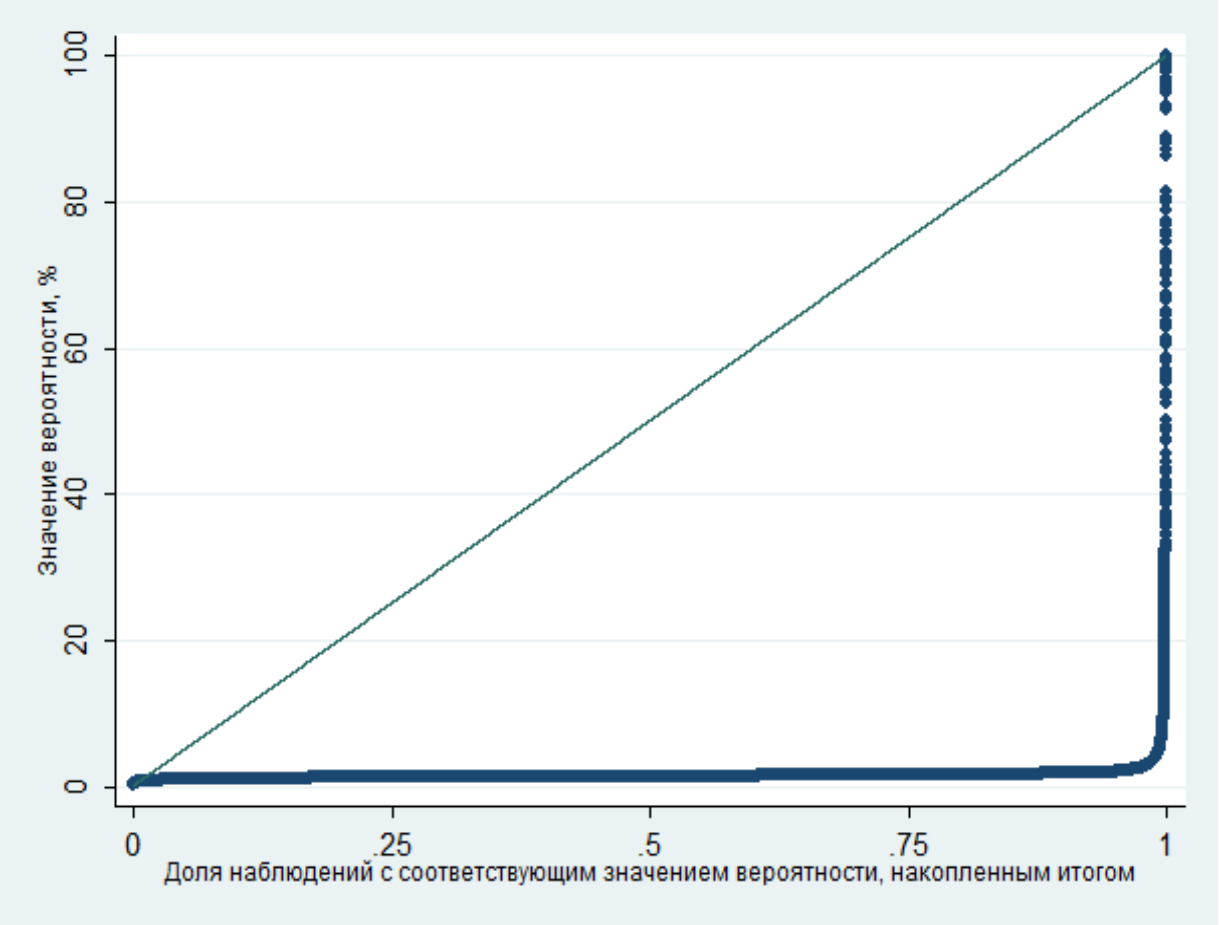
```
                                Number of obs     =     185542
                                                  Wald chi2(3)      =     852.36
Penalized log likelihood = -14083.734             Prob > chi2       =     0.0000

------------------------------------------------------------------------------
         bnk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rent |  -1.546058   .0577104   -26.79   0.000    -1.659169   -1.432948
         z_a |   .0024582    .000843     2.92   0.004     .0008058    .0041105
        dsr3 |   .0344062   .0119021     2.89   0.004     .0110786    .0577339
       _cons |  -4.176444   .0194831  -214.36   0.000     -4.21463   -4.138258
------------------------------------------------------------------------------
```

- Оценки изменились, но не кардинально;
- Все переменные остались значимыми – это хорошая новость;
- **НО**…

Подробнее о спецификации – см. (Донец, Могилат, 2017)

Апробация на российских данных о банкротствах (2/2)



- Медианная вероятность ≈ 0;
- Очень длинный правый хвост распределения

- вероятность банкротства ≠ риск банкротства;
- недооценка рисков одних компаний и переоценка рисков других

Методы корректировки константы

1. «Эдогенная» (King, Zeng, 2001)

- Функция правдоподобия и веса:

$$\ln L_w(\beta \mid y) = w_1 \sum_{[Y_1=1]} \ln(\pi_i) + w_0 \sum_{[Y_0=1]} \ln(1-\pi_i)$$

$$w_1 = \tau/\bar{y},\ w_0 = (1-\tau)/(1-\bar{y})$$

- Корректировка вероятности:

$$\Pr(Y_i = 1) \approx \tilde{\pi}_i + C_i,\ где$$

$$C_i = (0{,}5 - \tilde{\pi}_i)\tilde{\pi}_i(1-\tilde{\pi}_i)x_0 V(\tilde{\beta})x_0'$$

- Дисперсия оценки:

$$V(\tilde{\beta}) = \left[\sum \tilde{\pi}_i(1-\tilde{\pi}_i)x_i' x_i\right]^{-1}$$

τ – доля событий в генеральной совокупности;

$\bar{y}$ – параметр, отвечающий за жесткость ограничений

Апробация на российских данных о банкротствах (1/3)

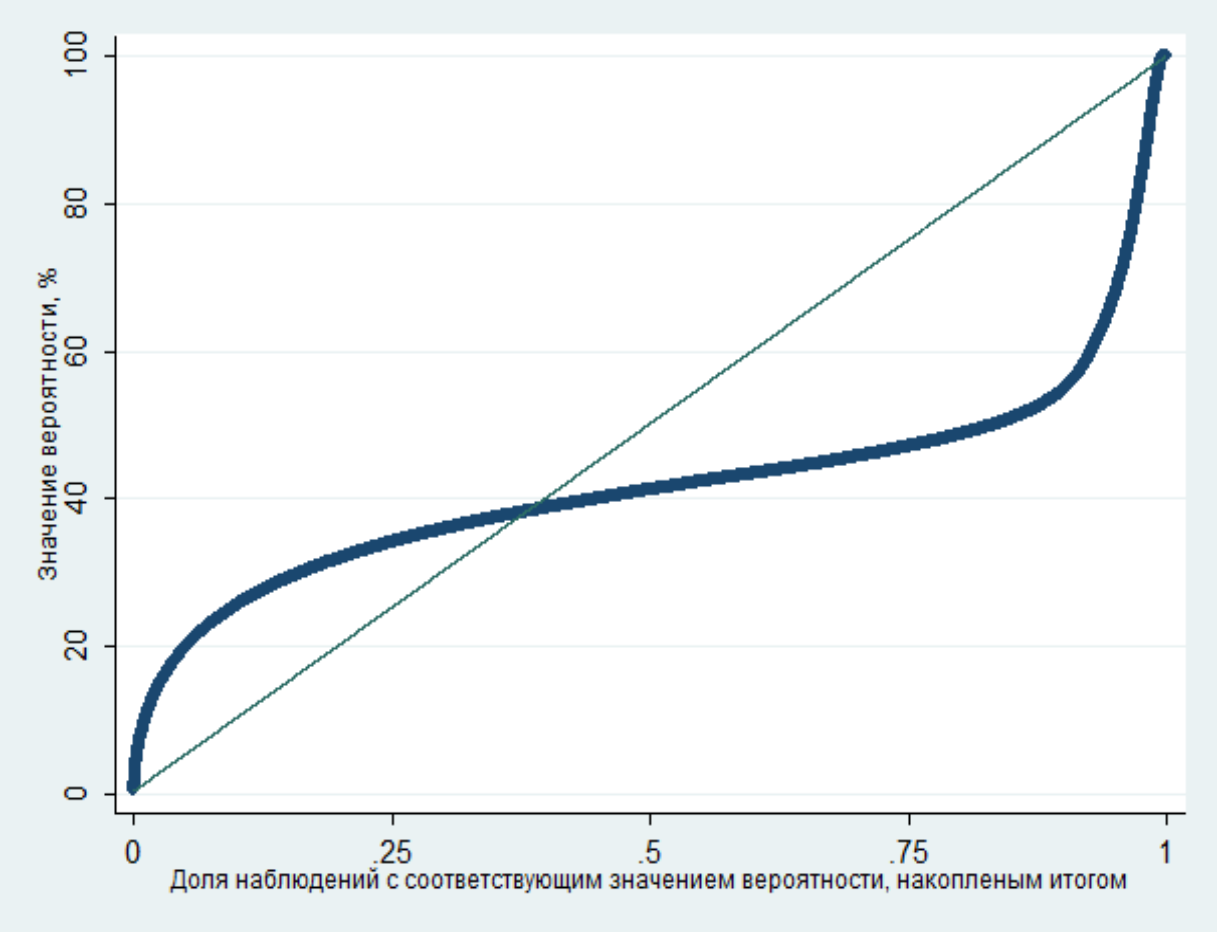
```
Corrected logit estimates                          Number of obs =    185542

------------------------------------------------------------------------------
                                           |               Robust
         bnk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rent |  -4.798896   .2592478   -18.51   0.000    -5.307012   -4.290779
         z_a |   .5163711   .1546848     3.34   0.001     .2131945    .8195478
        dsr3 |   .6582202   .0727999     9.04   0.000     .5155351    .8009054
       _cons |  -.2584053   .0435575    -5.93   0.000    -.3437765   -.1730341
------------------------------------------------------------------------------
```

- Оценки поменялись сильнее;
- Все переменные остались значимыми – это хорошая новость;
- **Неужели распределение снова все испортит?..**

Подробнее о спецификации – см. (Донец, Могилат, 2017)

К чему могут привести стандартные подходы (2/3)



- Медианная вероятность ≈ 40%;
- Правый хвост распределения длиннее левого, но не сильно

- вероятность банкротства ≈ риск банкротства

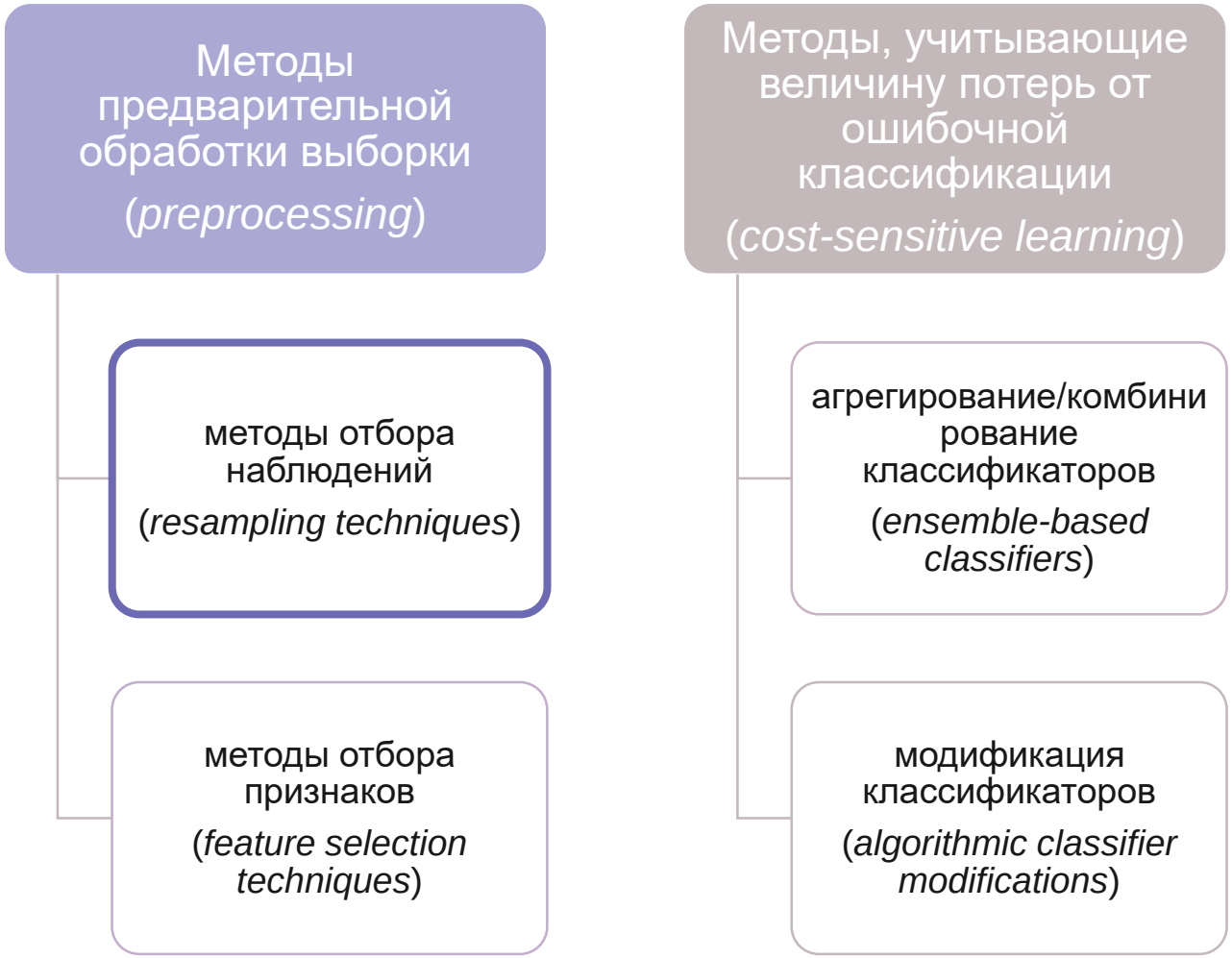
Методы корректировки константы

1. «Экзогенная» (King, Zeng, 2001)

- Корректировка константы: $$\tilde{\beta}_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

- Остальные параметры – стандартные ММП-оценки логит-модели

Методы машинного обучения (1)



Методы машинного обучения (2)
Отбор наблюдений

Метод сокращения

удаление наблюдений преобладающей группы

Бывает:
- случайное
- комбинирование редких событий с равными подвыборками
- кластерный подход

Метод расширения

повторное воспроизведение элементов меньшей группы или их комбинации

Бывает:
- случайное
- «искусственное»
- кластерный подход

Комбинированные методы

наблюдения одного класса добавляем, другого - удаляем

Бывает:
- случайное
- определенное автором сочетание
- кластерный подход

Подробнее о – см. (Galar et al, 2012), (Shaza et al, 2013), (Haixiang et al, 2017)

Так на чем же остановиться?

Литература

1. Донец С.А., Могилат А.Н. (2017). Кредитование и финансовая устойчивость российских промышленных компаний: микроэкономические аспекты анализа // Деньги и кредит. №7
2. Firth D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38
3. Galar M., Fernandez A., Barrenechea E., Bustince H., Herrera F. (2011). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews.
4. Gao, S., Shen, J. (2007): Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. In: Statistics and Probability Letters 77: 925-930
5. Haixiang G., Yijing Li, Shang J., Yuanyue H., Bing G. (2017). Learning rom class-imbalanced data: Review of methods and applications. Expert Systems with Applications. 73. Pp. 220-239
6. King G., Zeng L. (2001). Logistic кegression in rare events data. Political Analysis, 9, 137–163
7. Shaza M., Ajith A. (2013). A Review of Class Imbalance Problem. Journal of Network and Innovative Computing. ISSN 2160-2174, Volume 1 (2013) pp. 332-340

Спасибо за внимание!