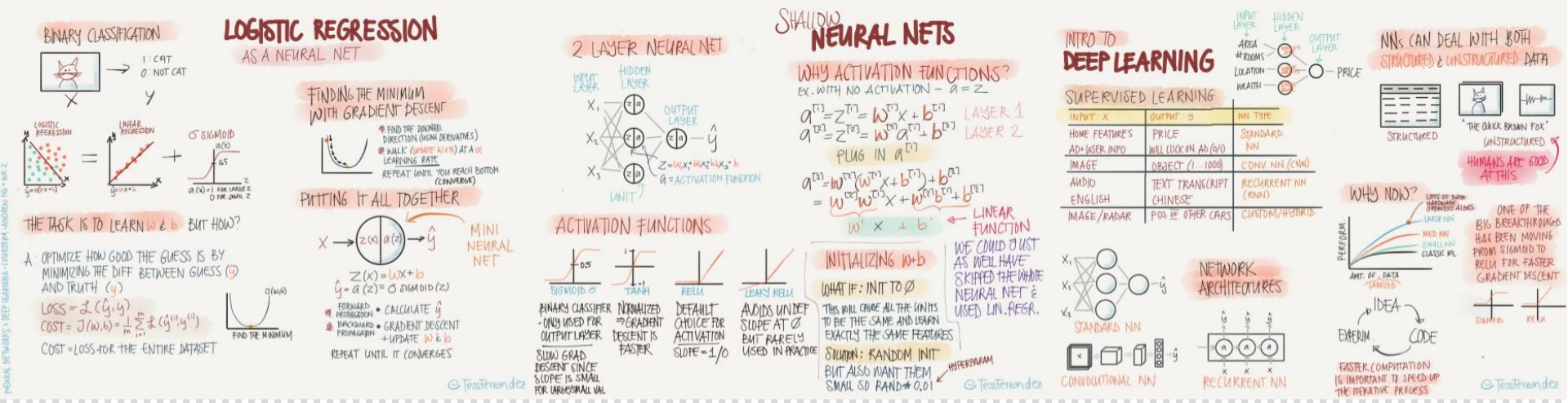


Two Forest Jump

Комбинированный отбор признаков с использованием двухлесового метода

Афанасьев Сергей
КБ «Ренессанс Кредит»

8 сентября 2020 г.
Москва

Для чего нужен отбор признаков?

1. **Избежать** проклятья размерности
2. **Снизить** переобучение модели
3. **Упростить** модель для интерпретации
4. **Сократить** время обучения модели

Фильтры

(Filter method)

- Матрица корреляций
- Таргет-корреляция
- PCA

...

Обертки

(Wrapper method)

- Stepwise Regression
- Random Forest
- RFE

...

Вложения

(Embedded method)

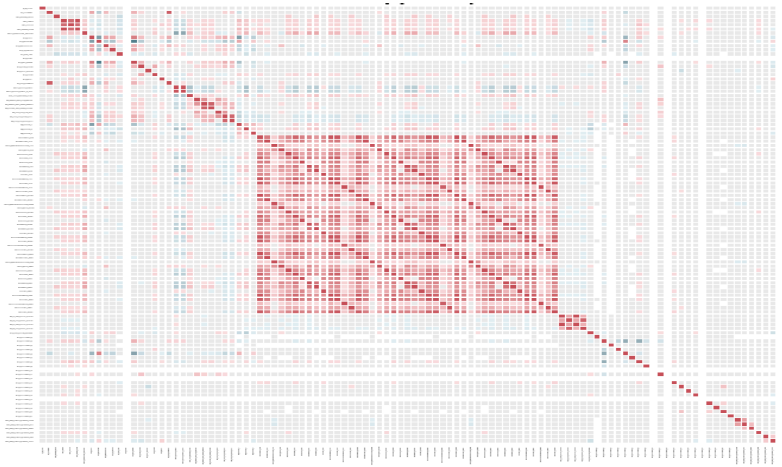
- LASSO (L1)
- Ridge regression (L2)
- ElasticNet

...

Mark A. Hall. Correlation-based Feature Selection for Machine Learning. — 1999.

Матрица корреляций

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}$$



Principal Component Analysis (PCA)

Метод главных компонент (PCA)

Постановка задачи PCA (principal component analysis):

$f_1(x),\dots, f_n(x)$ — исходные числовые признаки

$g_1(x),\dots, g_m(x)$ — новые числовые признаки, $m \le n$;

Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1,\dots,n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке $x_1,\dots,x_l$:

$$\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 \to \min_{\{g_s(x_i)\},\{u_{js}\}}$$

Матричные обозначения

Матрицы "объекты-признаки", старая и новая

$$\underset{l\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}; \quad \underset{l\times m}{G} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_l) & \dots & g_m(x_l) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^T \overset{хотим}{\approx} F.$$

Найти: и новые признаки G, и преобразование U

$$\sum_{i=1}^{l}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 = \left\|GU^T - F\right\|^2 \to \min_{G,U}$$

Основная теорема метода PCA

Если $m \le rk\ F$, то минимум $\left\|GU^T - F\right\|^2$ достигается, когда столбцы U — это собственные векторы матрицы F^TF, соответствующие m максимальным собственным значениям $\lambda_1,\dots,\lambda_m$, а матрица $G - FU$.

При этом:

- матрица U ортонормированна: $U^TU = I_m$;
- матрица G ортогональна: $G^TG = \Lambda = diag(\lambda_1,\dots,\lambda_m$;
- $U\Lambda = F^TFU$; $G\Lambda = FF^TG$;
- $\left\|GU^T - F\right\|^2 = \left\|F\right\|^2 - tr\ \Lambda = \sum_{j=m+1}^{n}\lambda_j$.

Связь с сингулярным разложением

Если взять $m = n$, то:

- $\left\|GU^T - F\right\|^2 = 0$;
- представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$: $F = GU^T = V\sqrt{\Lambda}U^T$; $U^TU = I_m$; $V^TV = I_m$.
- линейное преобразование U работает в обе стороны: $F = GU^T$; $G = FU$.

Поскольку новые признаки некоррелированы $\left(G^TG = \Lambda\right)$, преобразование U называется декоррелирующим (или преобразованием Карунена-Лоэва).

Эффективная размерность выборки

Упорядочим с.з. F^TF по убыванию: $\lambda_1 \ge \dots \ge \lambda_n \ge 0$.

Эффективная размерность выборки — это наименьшее целое m, при котором:

$$E_m = \frac{\left\|GU^T - F\right\|^2}{\left\|F\right\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \le \varepsilon$$

Решение задачи НК для МЛР в новых признаках

Задача наименьших квадратов для МЛР: $\left\|F\alpha - y\right\|^2 \to \min_{\alpha}$.

Заменим $\underset{l\bullet n}{F}$ на ее приближение $\underset{l\bullet m}{G}\underset{m\bullet n}{U^T}$, предполагая $m \le n$:

$$\left\|GU^T\alpha - y\right\|^2 = \left\|G\beta - y\right\|^2 \to \min_{\beta}$$

Связь нового и старого вектора коэффициентов:

$$\beta = U^T\alpha; \quad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно β (единственное отличие — m слагаемых вместо n):

$$\beta^* = D^{-1}V^Ty; \quad \alpha^* = UD^{-1}V^Ty = \sum_{j=1}^{m}\frac{1}{\sqrt{\lambda_j}}u_j\left(v_j^Ty\right);$$

$$G\beta^* = VV^Ty = \sum_{j=1}^{m}v_j\left(v_j^Ty\right)$$

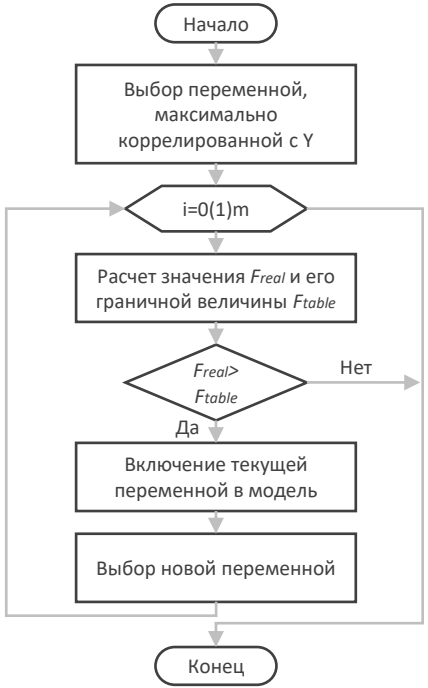
Резюме:

- Метод главных компонент позволяет приближать матрицу ее низкоранговым разложением;
- Для этого достаточно взять из SVD-разложения первые m сингулярных чисел и векторов матрицы;
- Этот прием широко используется в анализе данных — в задачах регрессии, классификации, сжатия данных и др.

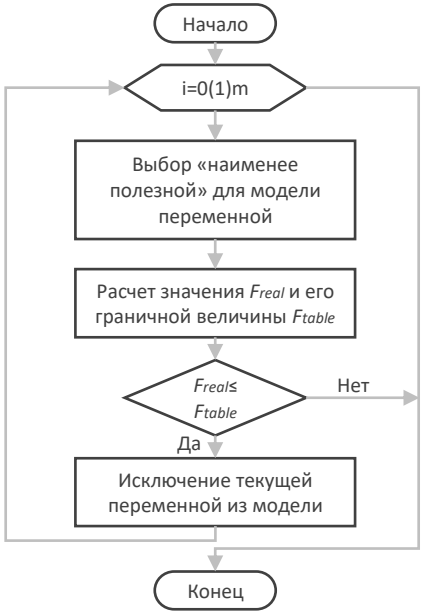
Источник: Coursera, К.В. Воронцов

Stepwise Regression

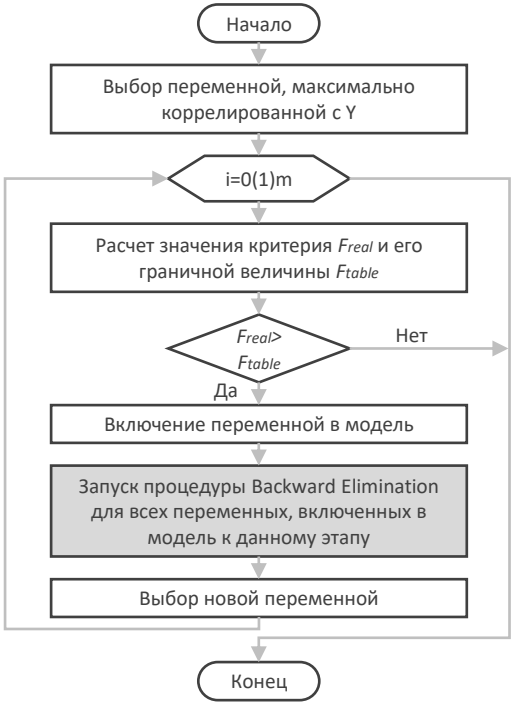
Forward Selection



Backward Elimination



Bidirectional elimination



Random Forest

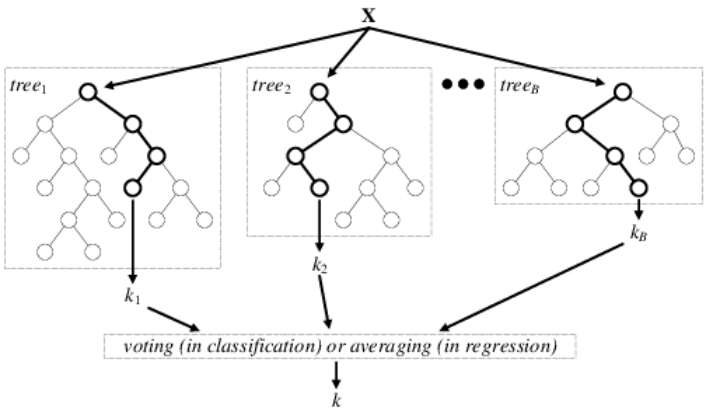
Подход 1. Важность на основе уменьшения неоднородности

1. Для каждого дерева случайного леса вычисляем сумму уменьшений неоднородности на всех ветвлениях, связанных с данным предиктором;

2. Итоговую сумму уменьшений неоднородности делим на общее количество деревьев;

3. Повторяем шаги 1-2 для всех переменных.

Важность = частота использования переменной в качестве предиктора ветвления.



Подход 2. Важность на основе уменьшения качества прогнозирования при случайной перестановке (пермутации)

1. Обучить модель;

2. На тестовом/OOB множестве посчитать ошибку;

3. Зафиксировать переменную/группу переменных, случайно переставить значения на тестовом/OOB множестве;

4. По новой выборке посчитать ошибку;

5. Вычислить разность ошибок на исходном множестве и множестве с перестановкой.

Var 1	Var 2	Var 3	Target	Predict	Right
1	2	101	0	0	x
2	3	102	1	0	
3	5	103	1	1	x
4	7	104	0	0	x

accuracy 75%

Var 1	Var 2	Var 3	Target	Predict	Right
1	5	101	0	1	
2	7	102	1	0	
3	2	103	1	1	x
4	3	104	0	0	x

accuracy 50%

Регуляризация L1 (LASSO) и L2 (Ridge)

Регрессия LASSO

LASSO — Least Absolute Shrinkage and Selection Operator,
два эквивалентных варианта постановки задаци:

$$Q(\alpha) = \|F\alpha - y\|^2 \to \min_{\alpha} \quad \text{при} \sum_{j=1}^{n} |\alpha_j| \le \chi;$$

$$Q(\alpha) = \|F\alpha - y\|^2 + \tau \sum_{j=1}^{n} |\alpha_j| \to \min_{\alpha};$$

После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^=; \\ |\alpha_j| = \alpha_j^+ - \alpha_j^=; \end{cases} \quad \alpha_j^+ \ge 0; \quad \alpha_j^- \ge 0.$$

ограничения принимают канонический вид:

$$\sum_{j=1}^{n} \alpha_j^+ + \alpha_j^- \le \chi; \quad \alpha_j^+ \ge 0; \quad \alpha_j^- \ge 0.$$

Чем меньше χ, тем больше j таких, что $\alpha_j^+ = \alpha_j^- = 0$.

Резюме:

LASSO обнуляет веса и приводи к отбору признаков
в линейных моделях.

Источник: Coursera, К.В. Воронцов

Гребневая регрессия (Ridge Regression)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \frac{\tau}{2}\|\alpha\|^2,$$

где τ — неотрицательный параметр решуляризации.

Модифицированное МНК-решение (τI_n — "гребень")

$$\alpha_\tau^* = \left(F^T F + \tau I_n\right)^{-1} F^T y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения α_τ^* и МНК-аппроксимация целевого вектора $F\alpha_\tau^*$:

$$\alpha_\tau^* = U\left(D^2 + \tau I_n\right)^{-1} DV^T y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j \left(v_j^T y\right);$$

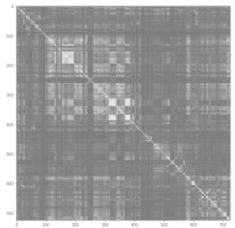
$$F\alpha_\tau^* = VDU^T \alpha_\tau^* = V\,diag\left(\frac{\sqrt{\lambda_j}}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j \left(v_j^T y\right);$$

$$\left\|\alpha_\tau^*\right\|^2 = \left\|\left(D^2 + \tau I_n\right)^{-1} DV^T y\right\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{\left(\lambda_j + \tau\right)^2} \left(v_j^T y\right)^2$$

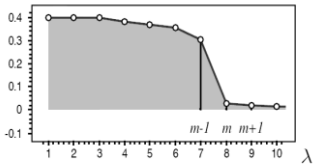
$F\alpha_\tau^* \ne F\alpha^*$, но зато решение становится гораздо устойчивее.

КЛАССИКИ

Matrix
(Scoring)



PCA
(ML)

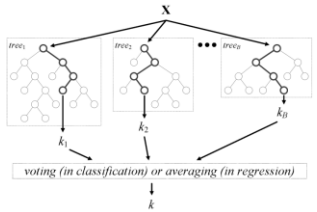


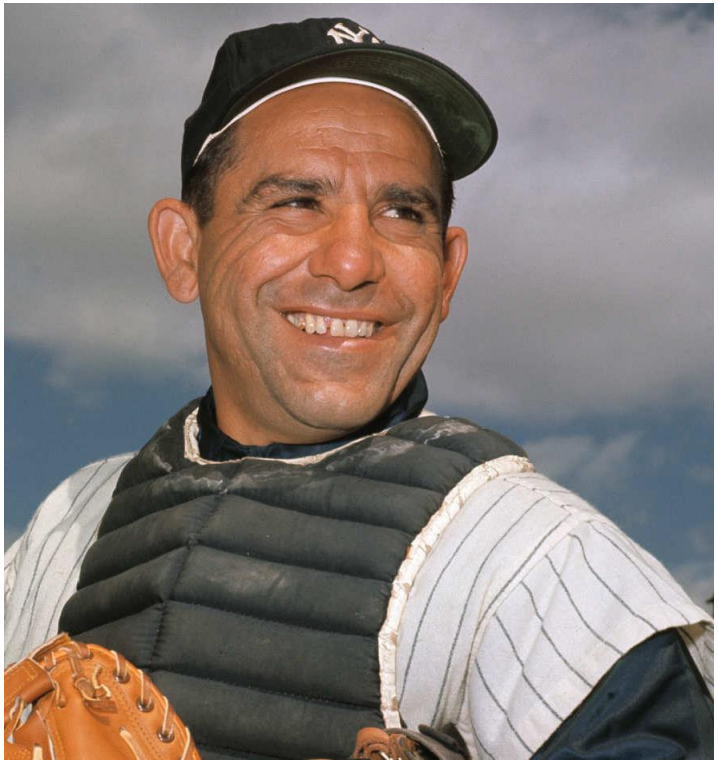
$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \cdots + \lambda_n}{\lambda_1 + \cdots + \lambda_n} \leqslant \varepsilon$$

НЕОФИТЫ

Boosting
(Kaggle)

Forest
(Science)

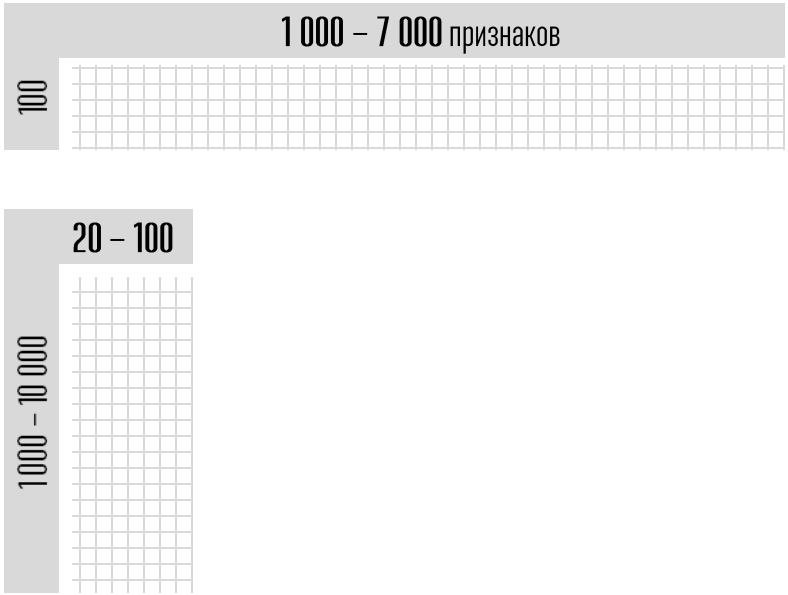




– В теории нет разницы между теорией и практикой. А на практике есть.

Йоги Берра

В научных статьях



В реальности

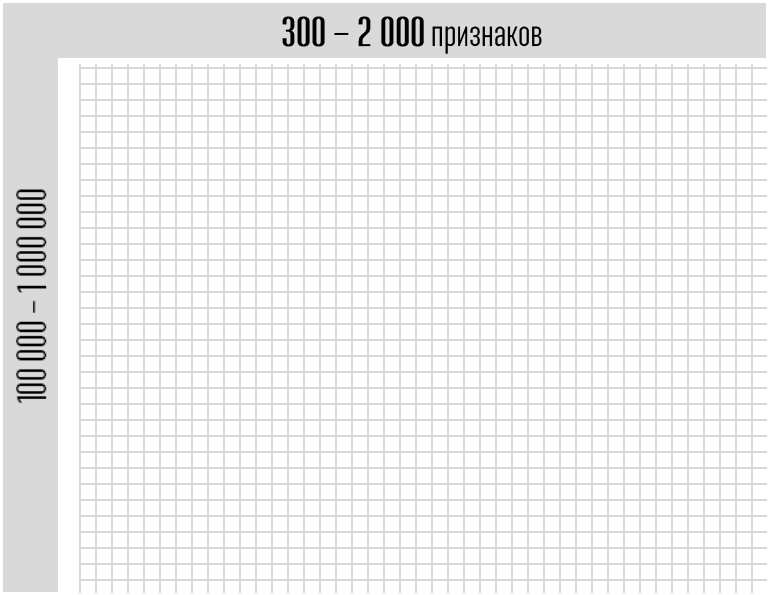
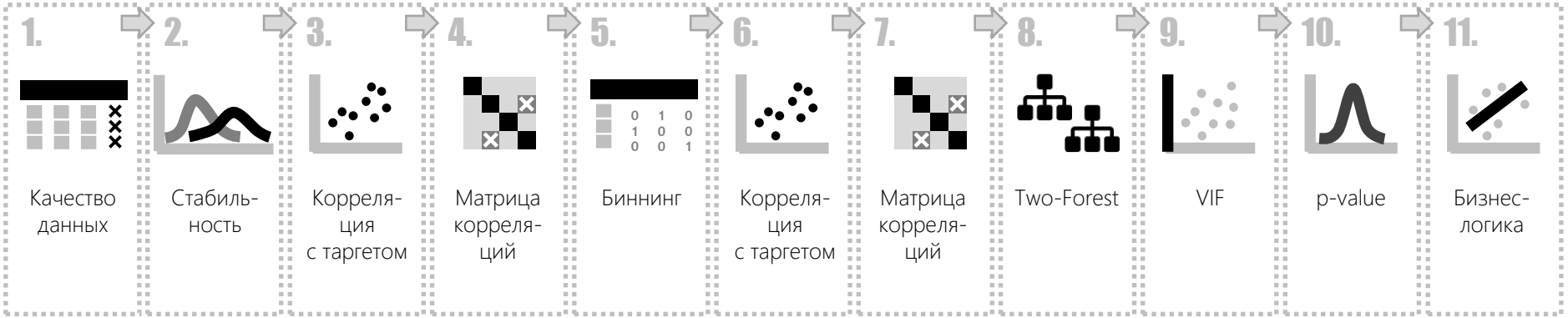


Схема отбора переменных



Шаг 1. Анализ качества данных

Критерии качества данных

- Точность (accuracy)
- Полнота (completeness)
- Согласованность (consistency)
- Достоверность (credibility)
- Правильность (correctness)
- Доступность (availability)

...

Этапы оценки качества данных

1. На стадии сбора данных
2. **На стадии разработки модели**
3. На стадии эксплуатации модели

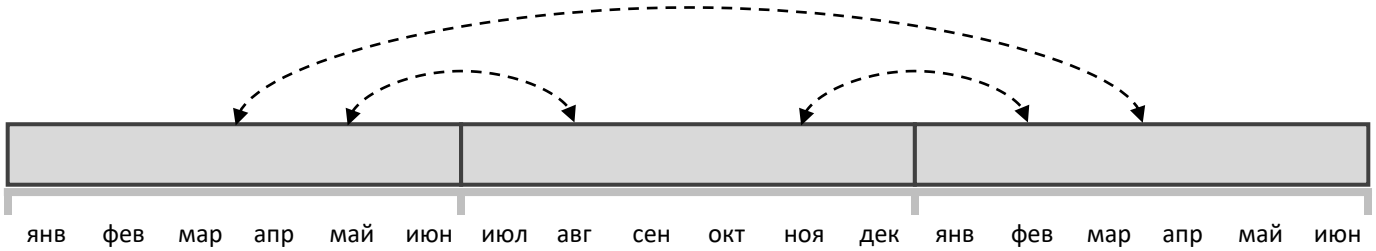
Проблемы снижения качества данных

- Пропуски или неполнота;
- Орфографические ошибки;
- Аномалии;
- Фиктивные значения;
- Логические несоответствия;
- Закодированные значения;
- Несоответствие форматов;
- Дублирование;
- Избыточность информации;

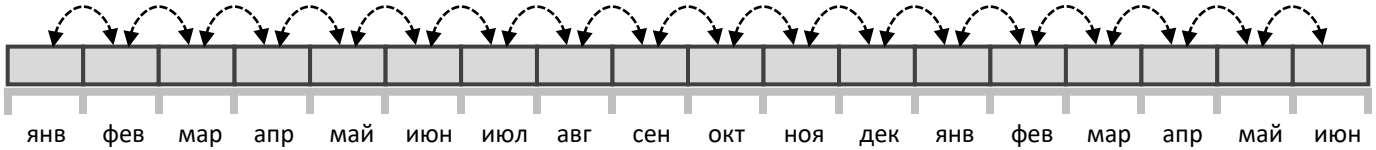
...

Шаг 2. Проверка стабильности

Большие периоды



Маленькие периоды



Шаг 2. Проверка стабильности

	Стабильна	Слегка нестабильна	Нестабильна
KS	[0%; 10%)	[10%; 25%)	[25%; 100%]
S	[0%; 10%)	[10%; 25%)	[25%; 100%]
PSI	[0; 10)	[10; 25)	[25; ∞)
Присвоеный вес	**0**	**0,5**	**1**
Большие периоды: $Big=max(KS_i)+max(S_i)+max(PSI_i)$	[0; 1]	[1; 2)	[2; 3)
Маленькие периоды: $Small=Avg(KS_j+S_j+PSI_j)$	[0; 1]	[1; 1,5)	[1,5; 3)
Присвоеный вес	**0**	**0,5**	**1**
Максимум (консервативный подход): Max(Big; Small)	{0}	{0,5}	{1}
Усреднение (лояльный подход): 0,5*Big+0,5*Small	[0; 0,2)	[0,2; 0,6)	[0,6; 1]
Присвоеный вес	**0**	**0,5**	**1**

Шаг 3. Корреляция с целевой переменной

Бинарная целевая переменная

- Для непрерывных признаков рассчитываются **статистика Стьюдента** (если признак распределен нормально) или **тест Манна-Уитни**;
- Для категориальных и бинарных признаков рассчитывается **Хи-квадрат** критерий Пирсона.

Категориальная целевая переменная

- Для непрерывных признаков проводится тест **ANOVA**;
- Для категориальных и бинарных признаков рассчитывается **Хи-квадрат** критерий Пирсона.

Непрерывная целевая переменная

- Для непрерывных признаков рассчитывается **корреляция Пирсона**;
- Для категориальных признаков проводится тест **ANOVA**;
- Для бинарных признаков рассчитываются **статистика Стьюдента** (если признак распределен нормально) или **тест Манна-Уитни**.

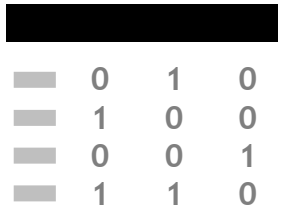
Шаг 4. Матрица корреляций

$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}$$

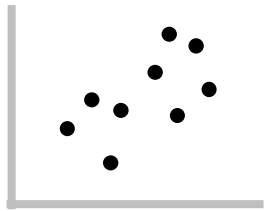
- Для непрерывных признаков рассчитывается **корреляция Пирсона**;
- Для категориальных и бинарных признаков рассчитывается **корреляция Спирмена**

Шаг 5-7. Бинаризация и корреляции

Шаг 5
Бинаризация признаков



Шаг 6
Корреляция с таргетом



Шаг 7
Матрица корреляций

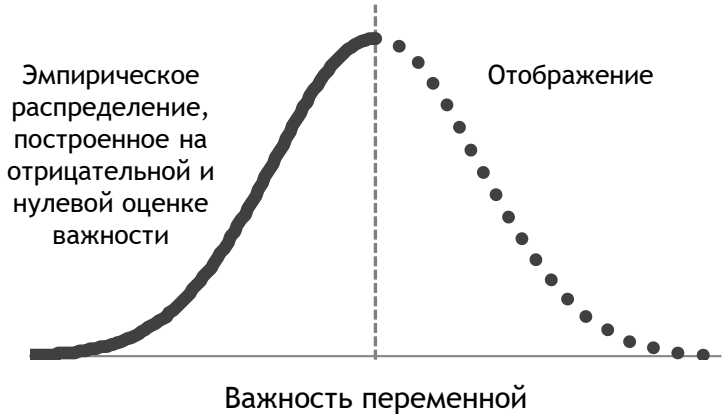
$$R_x = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_n} \\ \dots & \dots & \dots & \dots \\ r_{x_nx_1} & r_{x_nx_2} & \dots & 1 \end{pmatrix}$$

Шаг 8. Two-Forest

Важность – уменьшение качества прогнозирования при случайной перестановке

$$VI_j = P\left(Y \neq f\left(X_1,\ldots,X_j^*,\ldots,X_p\right)\right) - P\left(Y \neq f\left(X_1,\ldots,X_j,\ldots,X_p\right)\right)$$

1. Исходная выборка делится на две подвыборки;
2. На каждой из двух подвыборок строится случайный лес;
3. Для построенных моделей для каждой *j-ой* переменной вычисляется важность при перестановке VI_j на множестве, которое не использовалось для построения модели;
4. Определяются множества:
   M1 = {все отрицательные важности},
   M2 = {все нулевые важности},
   *M3 = {все отрицательные важности * (-1)}*;
5. На *M = M1 U M2 U M3* строится плотность распределения *F* ;
6. Для каждой *j*-ой переменной рассчитывается *p-value= 1– F(VI$_j$)*



Шаг 9. Мультиколлинеарность (VIF)

1. Для каждого признака *Xi* обучается линейная регрессия, где *Xi* является функций от всех остальных признаков:

$$X_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_j, \qquad i \neq j$$

2. Для каждого обученного признака рассчитывается коэффициент VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

3. Признаки со значением VIF>10 относятся к мультиколлинеарным. Из всех мультиколлинеарных признаков удаляется признак с максимальным значением VIF;

4. Шаги 1-3 итерационно повторяются до тех пор, пока максимальное значение VIF по всем оставшимся признакам не станет меньше или равно 10.

Шаг 10. Статистическая значимость

Проверку статистической значимости признаков можно делать с помощью различных тестов, среди которых тест отношения правдоподобия, тест Вальда и тест множителей Лагранжа (все три теста асимптотически эквивалентны).

Тест отношения правдоподобия: для проверки нулевой гипотезы сравниваются функции правдоподобия полной модели и укороченной модели без тестируемого признака. Для этого рассчитывается статистика отношения правдоподобия:

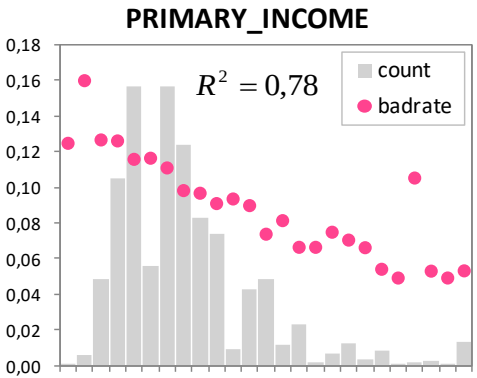
$$LR = 2 \cdot (L_l - L_s) = 2 \cdot \ln \frac{L_l}{L_s}$$

Где L_l – значение логарифмической функции правдоподобия полной модели,

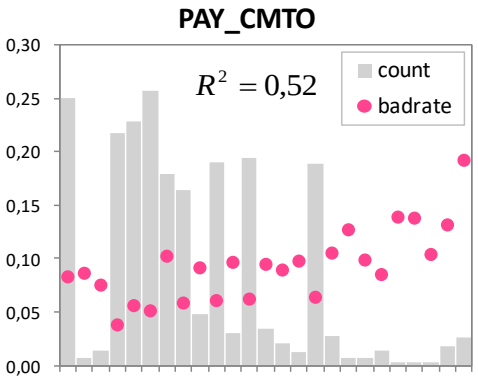
L_s – значение логарифмической функции правдоподобия укороченной модели.

Шаг 11. Экспертный анализ

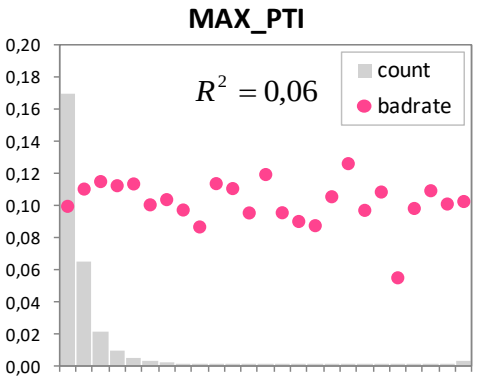
Сильная зависимость



Умеренная зависимость



Слабая зависимость



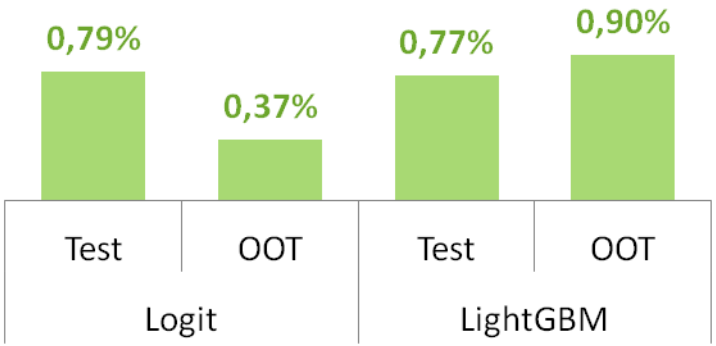
Matrix	Forward	Two-Forest
1 Качество	1 Качество	1 Качество
2 Стабильность	2 Стабильность	2 Стабильность
3 Таргет-корреляция	3 Таргет-корреляция	3 Таргет-корреляция
4 Матрица корреляций	4 Матрица корреляций	4 Матрица корреляций
5 Биннинг	5 Биннинг	5 Биннинг
6 Таргет-корреляция	6 Таргет-корреляция	6 Таргет-корреляция
7 Матрица корреляций	7 Матрица корреляций	7 Матрица корреляций
8 Gini	8 Forward	8 Two-Forest
9 VIF	9 VIF	9 VIF
10 P-value	10 P-value	10 P-value

Бизнес-задачи

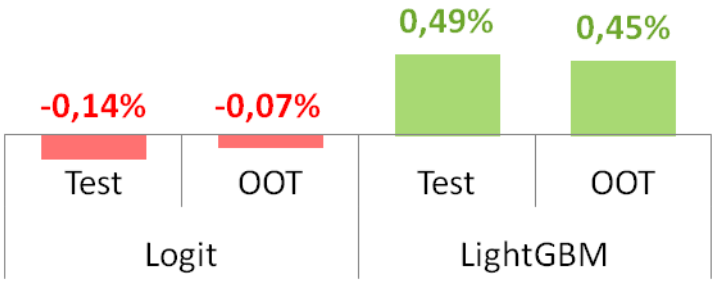
	Кол-во наблюдений	Кол-во признаков
PTB (CRM)	303 220	1 222
Application PD (Scoring)	497 063	423
Behavioral PD (Scoring)	588 385	1 087
Allocation (Collection)	172 250	162

Результаты (разница точности, Gini)

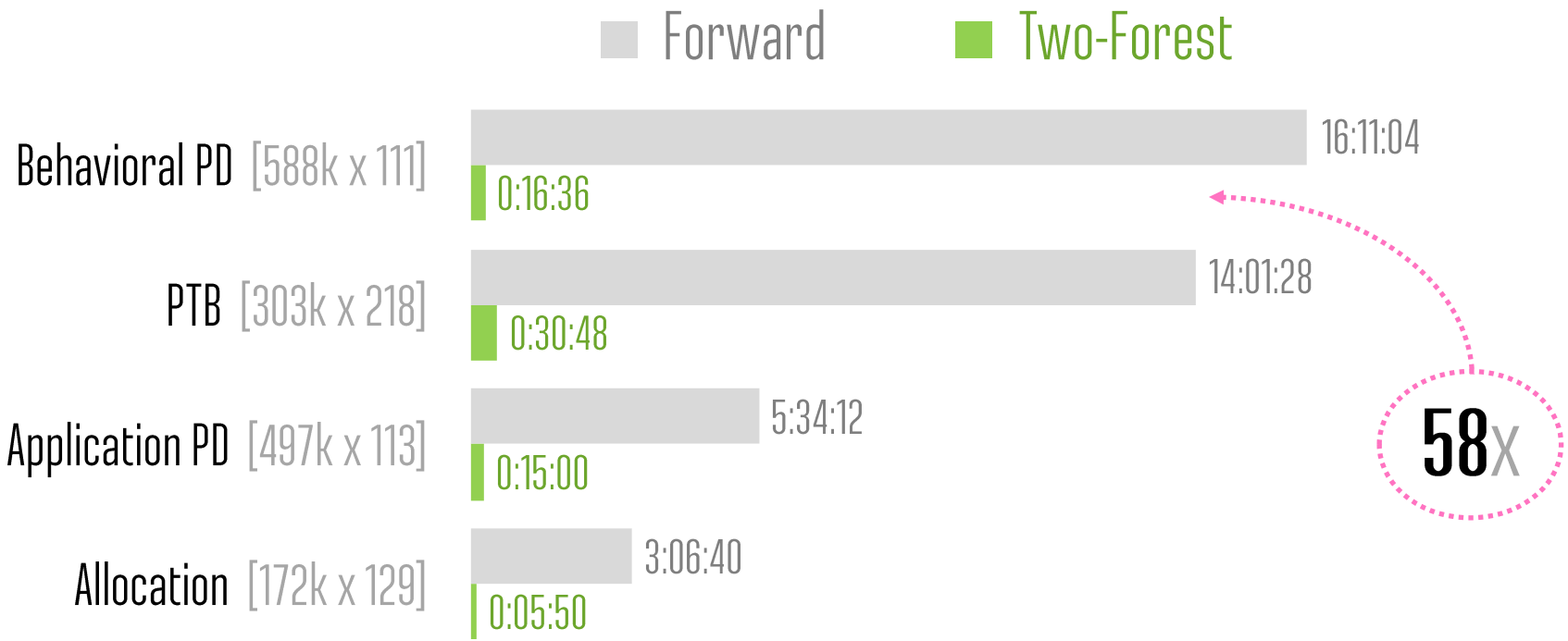
Δ(TwoForest–Matrix)



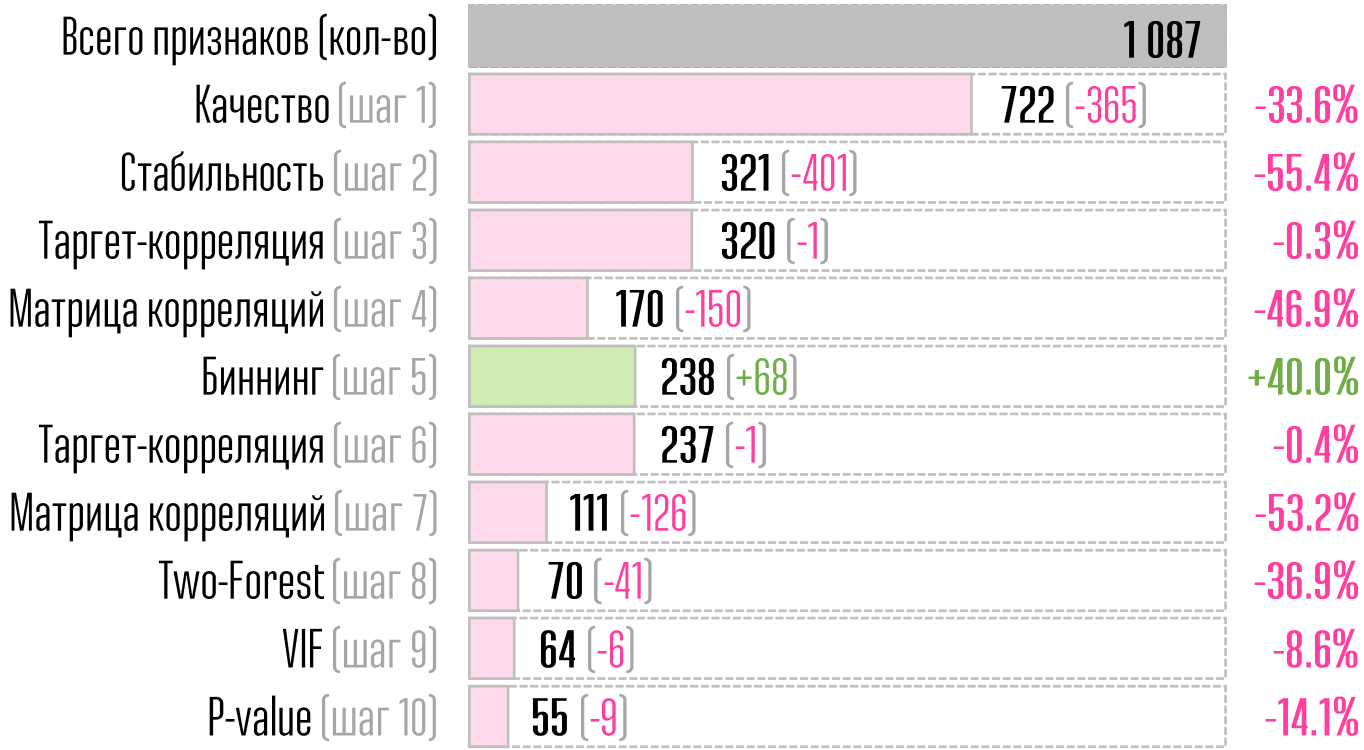
Δ(TwoForest–Forward)



Результаты (скорость работы методов)



Воронка отбора признаков (Behavioral PD)



Выводы

1. **Комбинированный отбор хорошо работает с большим количеством признаков:** фильтры отсеивают мусорные признаки, обертки – учитывают многомерные зависимости

2. Проверка качества и стабильности переменных **позволяет использовать один большой универсальный набор признаков** для разных типов задач.

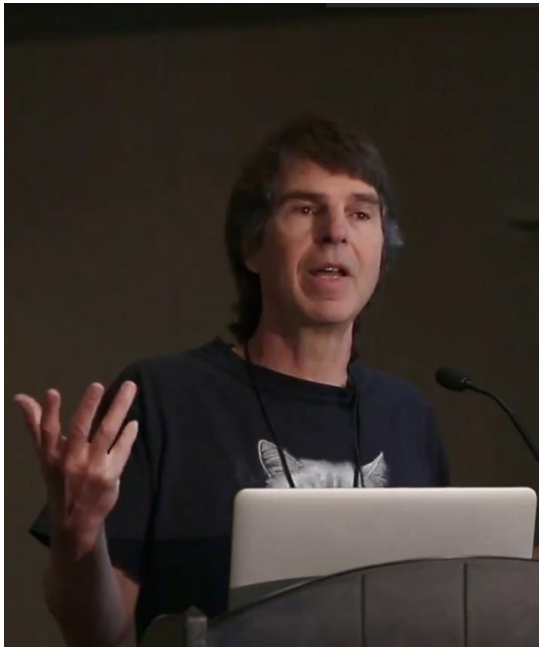
3. **Two-Forest** в сравнении с Forward **лучше по качеству для нелинейных моделей** и сопоставим по качеству для линейных моделей. По скорости **Two-Forest в десятки раз быстрее** регрессионных методов.

4. **Включение** в комбинированную схему отбора **нескольких оберток** позволяет контролировать корректность работы методов на предыдущих шагах

– С точки зрения академической науки, когда я начинал заниматься машинным обучением 25 лет назад, тот научный коллектив, в который я пришел еще студентом, в общем-то жил с полной уверенностью (и она была основана на примерно 30-летнем опыте предыдущих исследований), что задачу можно решать любым методом.

Константин Воронцов, 2018

Теорема о бесплатных завтраках



В среднем по всем возможным порождающим определениям у любого алгоритма классификации частота ошибок классификации ранее не наблюдавшихся примеров одинакова. Самый изощренный алгоритм, который мы только можем придумать, в среднем (по всем возможным задачам) дает такое же качество, как простейшее предсказание: все точки принадлежат одному классу.

David H. Wolpert, 1996

Спасибо за внимание!

Афанасьев Сергей

Исполнительный директор
Начальник управления статистического анализа

КБ «Ренессанс Кредит»