

# Машинное обучение

Лекция 5. Оценка качества на исторических данных



# На этой лекции

- I. Метрики качества в задачах регрессии
- II. Метрики качества в задачах классификации
- III. Выбор метрик качества: пример
- IV. Устойчивость моделей

# I. Метрики качества в задачах регрессии

## Метрики качества

- MAE
- RMSE
- MAPE
- SMAPE
- logloss

## MEAN AVERAGE ERROR

- Отклонение прогноза от исходного значения
- Усредненное по всем наблюдениям

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

# ROOT MEAN SQUARED ERROR

- Корень из среднего квадратичного отклонения прогноза от исходного значения
- Сильнее штрафует за большие по модулю отклонения

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

## MEAN AVERAGE PERCENTAGE ERROR

- Ошибка прогнозирования оценивается в процентах

$$M = \frac{100}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

# SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- Ошибка оценивается в процентах

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

# SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- Ошибка оценивается в процентах

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{|A_t| + |F_t|}$$

# SYMMETRIC MEAN AVERAGE PERCENTAGE ERROR

- По-разному штрафует за перепрогнозирование и недопрогнозирование

- Перепрогнозирование:

$$A_t = 100, F_t = 110 \sim \text{SMAPE} = 4.76\%$$

- Недопрогнозирование:

$$A_t = 100, F_t = 90 \sim \text{SMAPE} = 5.26\%$$

## LogLoss

- Логарифмическая ошибка
- Хорошо оценивает вероятность

$$\text{LogLoss} = - \frac{1}{n} \sum_{i=1}^n \left[ y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

## II. Метрики качества в задачах классификации

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC

Accuracy

Доля правильных ответов при классификации

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 1 0 0

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 1 1 0

# Accuracy

Доля правильных ответов при классификации

target: 1 0 1 0 0 0 0 1 0 0

predicted: 0 0 1 0 0 0 0 1 1 0

accuracy = 8/10 = 0.8

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC

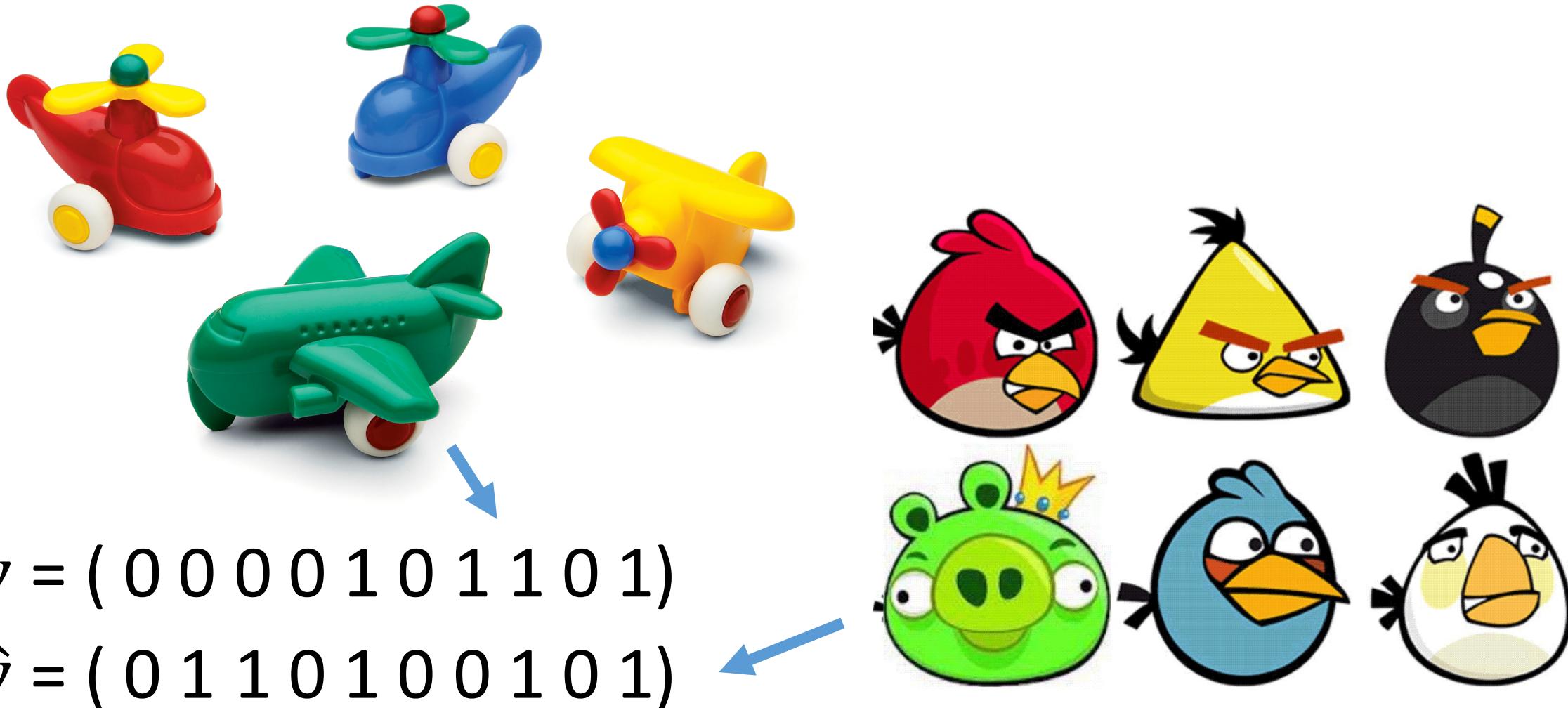
# Precision & Recall

- Precision – точность
- Recall - полнота

# Сбитые самолеты



# Сбитые самолеты



Precision

Precision – точность выстрелов:

Количество сбитых самолётов

---

Количество выстрелов

$$y = (0\ 0\ 0\ 0\ \textcolor{red}{1}\ 0\ 1\ \textcolor{red}{1}\ 0\ 1)$$

$$\hat{y} = (\textcolor{blue}{0}\ \textcolor{blue}{1}\ \textcolor{blue}{1}\ 0\ 1\ 0\ 0\ \textcolor{blue}{1}\ 0\ 1)$$



# Recall

Recall – «полнота» сбивания самолетов:

**Количество сбитых самолётов**

---

Общее количество самолётов

$$y = (0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 1)$$

$$\hat{y} = (0 \ 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1)$$



# Связь с True Positive, False Positive и др.

		Actual Class	
		Yes	No
Predicted Class	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

## F-measure (F-score, F1)

- Среднее геометрическое между precision и recall
- Значение F-measure ближе к меньшему из precision, recall

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

## Метрики качества

- Accuracy
- Precision
- Recall
- F-measure
- ROC-AUC

## ROC-AUC

- Применяется для оценки «вероятностной» классификации\*
- «Качество» ранжирования объектов по вероятности принадлежности к целевому классу;
- Доля «правильно» отранжированных пар;
- Вероятность встретить объект целевого класса раньше, чем объект нецелевого класса.

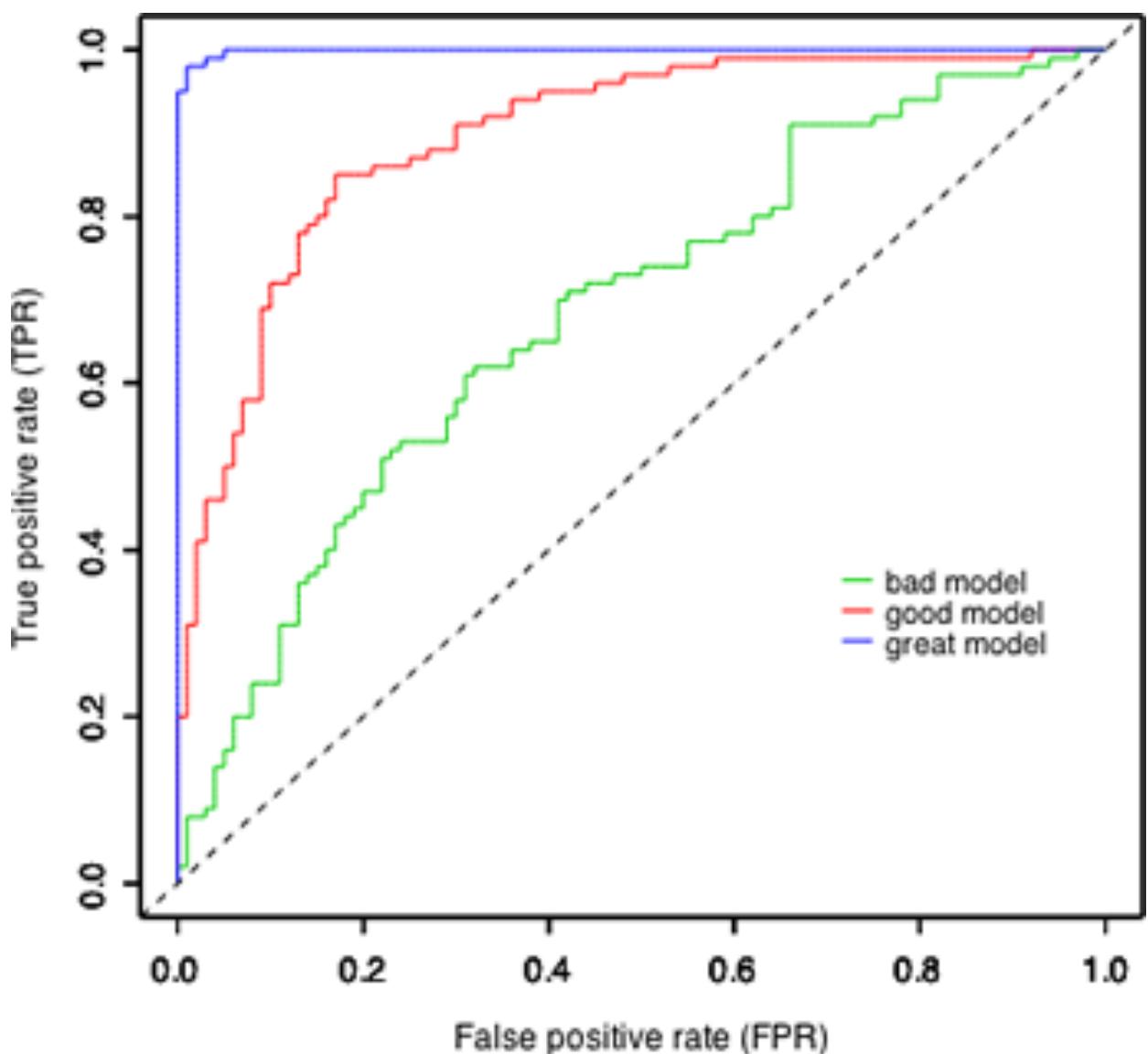
# ROC

		Actual Class	
		Yes	No
Predicted Class	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$TPR = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$FPR = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}.$$

# ROC



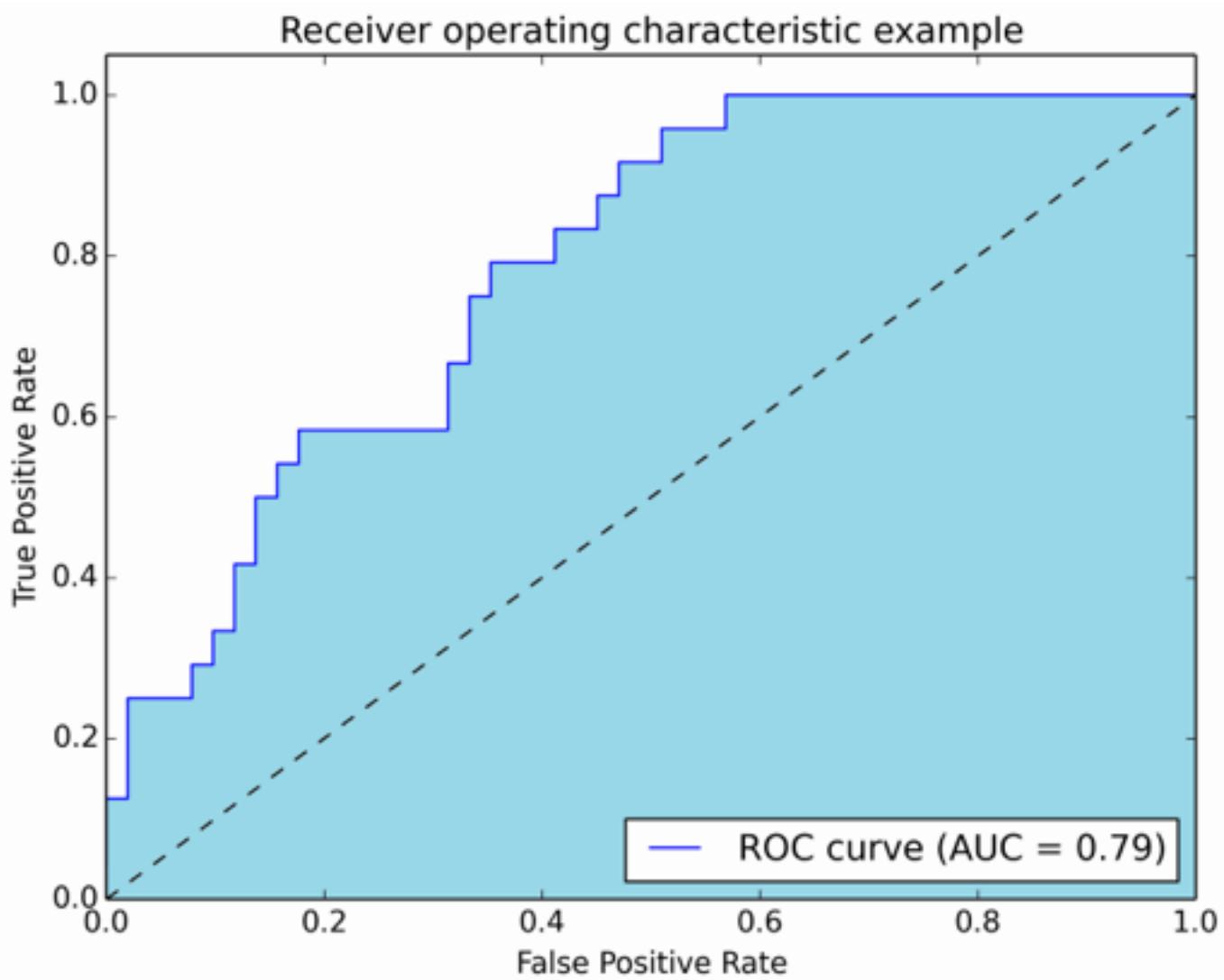
# ROC

- Как оценить кривую численно?

# ROC-AUC

- Как оценить кривую численно?
- Измерить площадь под кривой – area under the curve!

# ROC-AUC



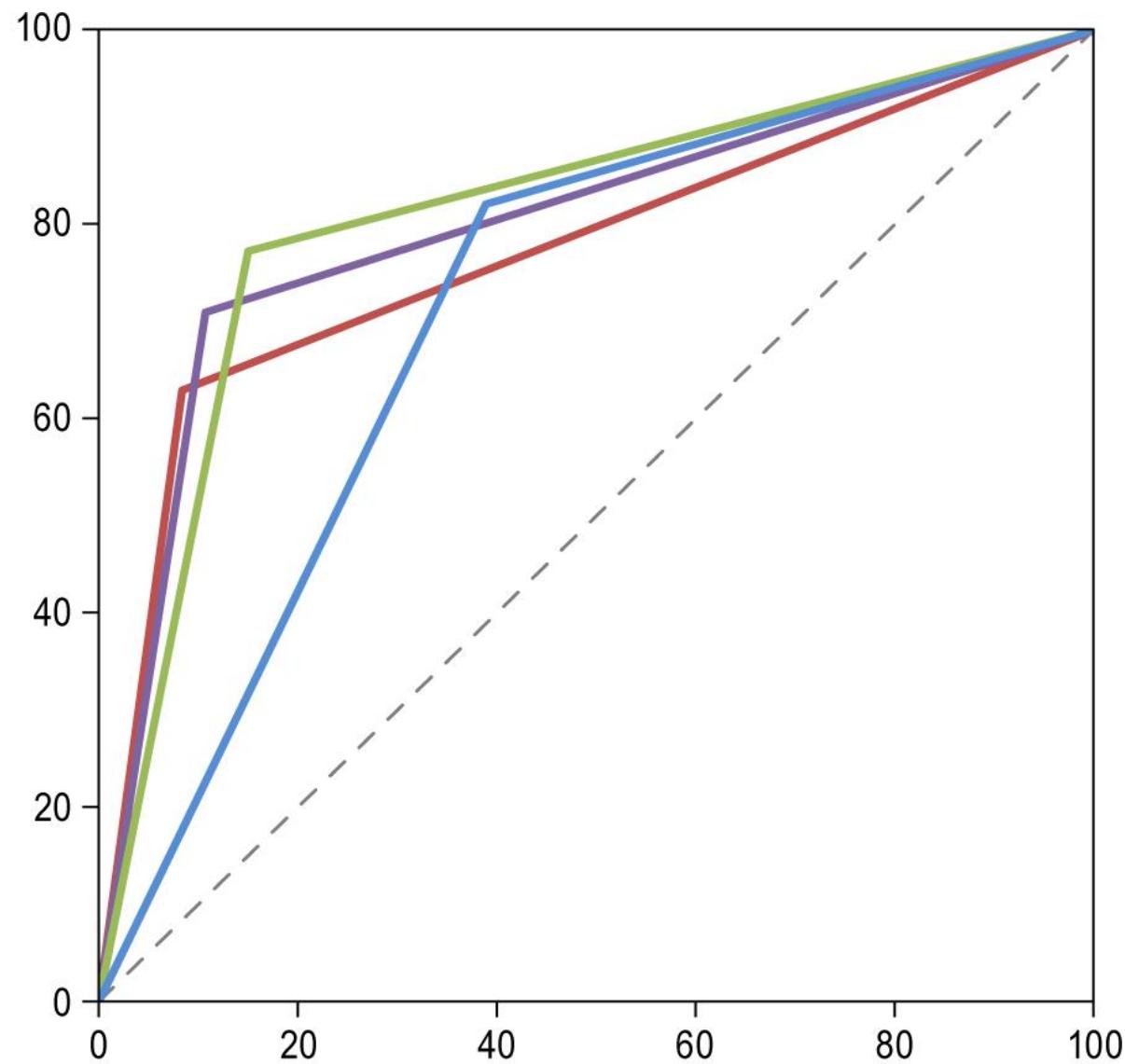
# ROC-AUC

- Что если классификация всё же не вероятностная?

# ROC-AUC

- Что если классификация всё же не вероятностная?
- Строим кривую по 3м точкам

# ROC-AUC



### III. Выбор метрик качества: пример

1. Выбираем, что оптимизировать (на примере рекомендаций)

## Что можем делать

- Прогнозировать, какие товары будут куплены
- Максимизировать прибыль

Остается вопрос: какие прогнозы нужны и как их использовать, чтобы денег стало больше?

# Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

# Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Вероятность:	$p_1$	$p_2$	$p_3$	$p_4$
--------------	-------	-------	-------	-------

# Максимизация дохода

	Товар 1	Товар 2	Товар 3	Товар 4
Вероятность:	$p_1$	$p_2$	$p_3$	$p_4$
Цена:	$c_1$	$c_2$	$c_3$	$c_4$

# Максимизация дохода



Puma  
Ветровка  
3 490 руб.



Crocs  
Сланцы  
1 990 руб.



Tony-p  
Слипоны  
~~1 999 руб.~~ 1 590 руб.



Champion  
Брюки спортивные  
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970

# Максимизация прибыли



Puma  
Ветровка  
3 490 руб.

Crocs  
Сланцы  
1 990 руб.

Tony-p  
Слипоны  
~~1 999 руб.~~ 1 590 руб.

Champion  
Брюки спортивные  
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970
Маржинальность:	0.1	0.4	0.4	0.2

## Мини-задача

Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

2. Выбираем метрику качества (на примере  
рекомендаций)

# Точность (Precision@k)

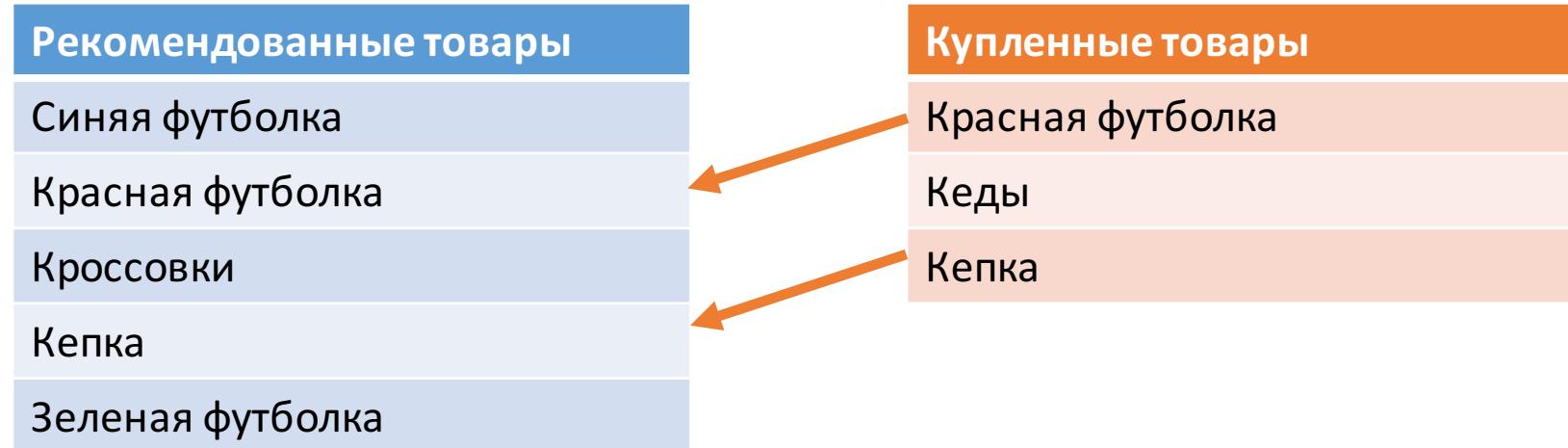
Рекомендованные товары	Купленные товары
Синяя футболка	Красная футболка
Красная футболка	Кеды
Кроссовки	
Кепка	Кепка
Зеленая футболка	

$k$  – количество  
рекомендаций

$$\text{Precision}@k = \frac{\text{купленное из рекомендованного}}{k}$$

AveragePrecision@k - усредненный по сессиям Precision@k

# Полнота (Recall@k)



k – количество  
рекомендаций

$$\text{Recall}@k = \frac{\text{купленное из рекомендованного}}{\text{количество покупок}}$$

AverageRecall@k - усредненный по сессиям Recall@k

# Взвешенный ценами recall@k

Рекомендованные товары	Купленные товары
Синяя футболка – 1000р	Красная футболка – 1200р
Красная футболка – 1200р	Кеды – 3000р
Кроссовки – 3500р	Кепка – 900р
Кепка – 900р	
Зеленая футболка – 800р	

Взвешенный ценами Recall@k =  $\frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$

AverageRecall@k - усредненный по сессиям Recall@k

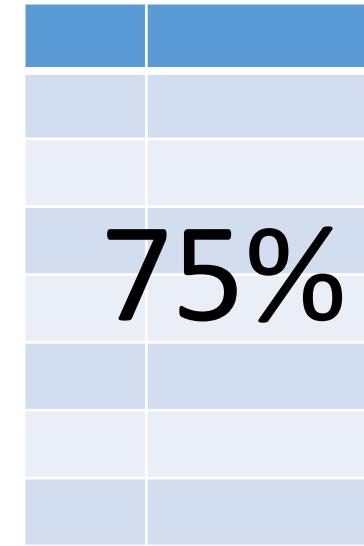
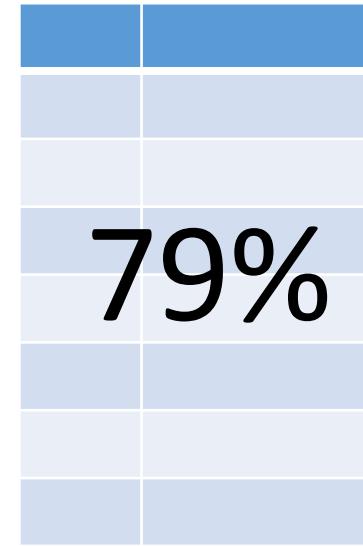
# Качество классификации против качества рекомендаций

Пример – 2 решения для прогноза купит/не купит товар

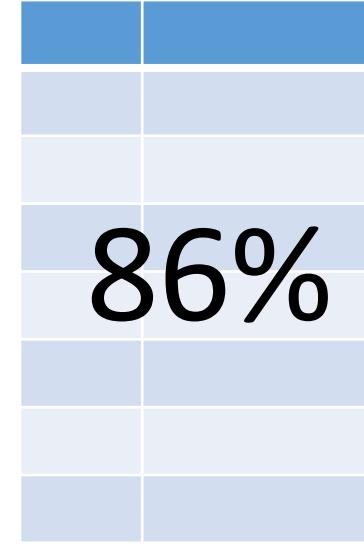
	Алгоритм 1	Алгоритм 2
AUC классификатора	0.52	0.85
Recall@5	0.72	0.71

## IV. Стабильность моделей

# Проблема: разброс качества на разных данных

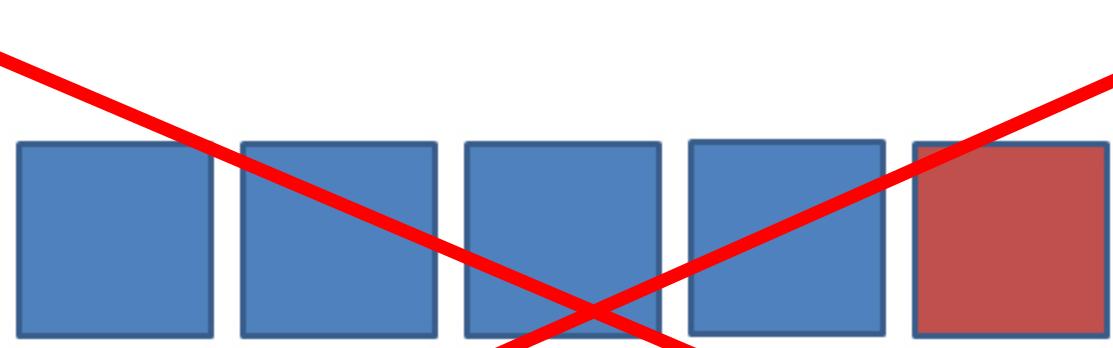


...



## Шаг 1: усреднение качества в CV

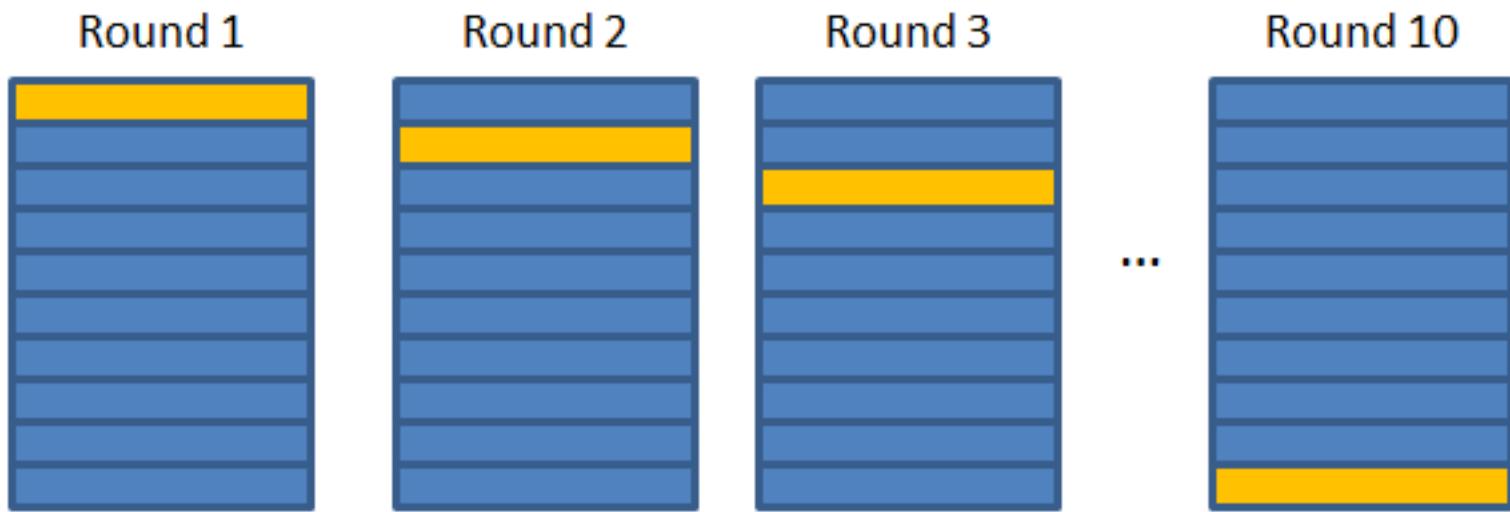
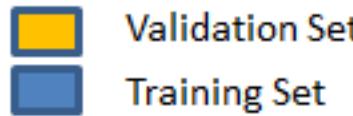
Если есть проблема со стабильностью модели, точно нужно избегать оценок на одном фиксированном датасете



Нужно использовать оценку качества в кросс-валидации

# Кросс-валидация

K-Fold cross validation:



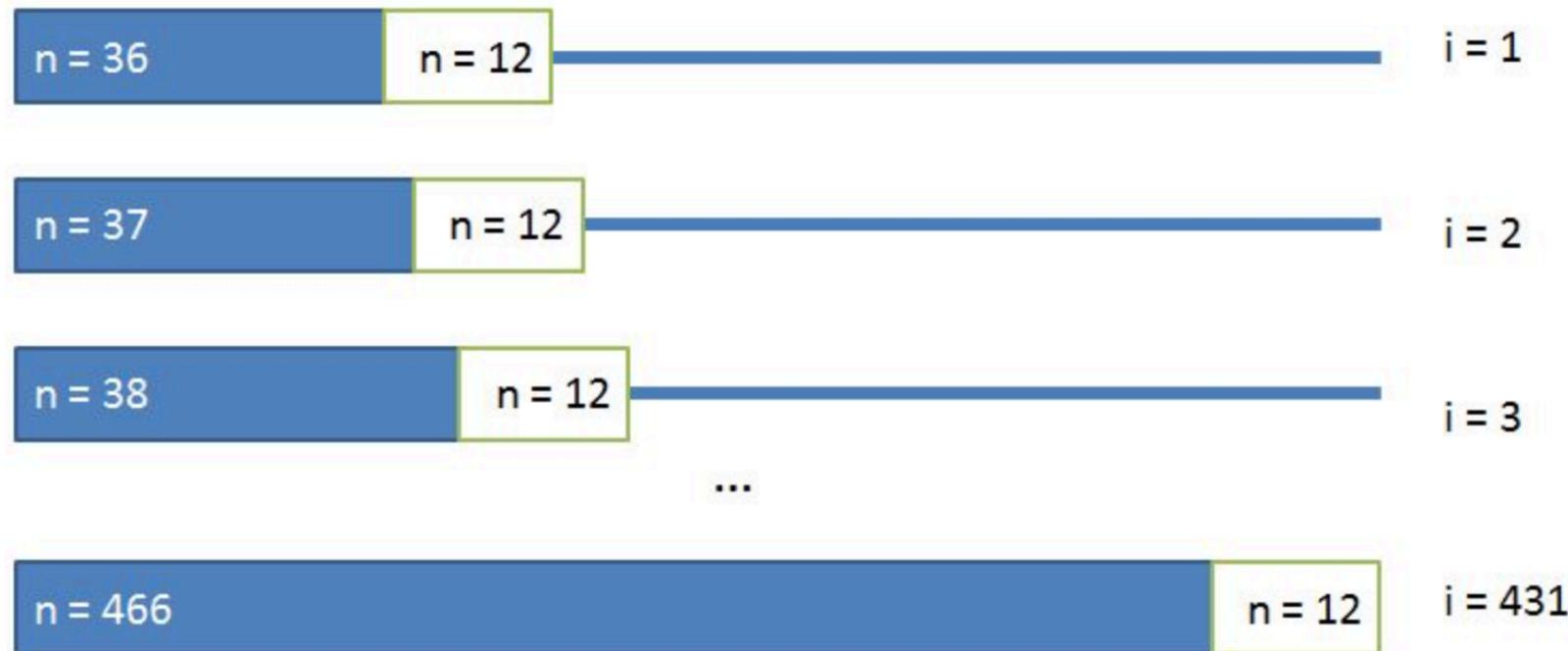
На картинке  $k = 10$ . Другие частые варианты – 3 и 5.

# Предупреждение: будьте осторожны с CV

N = 478 (month-end data)

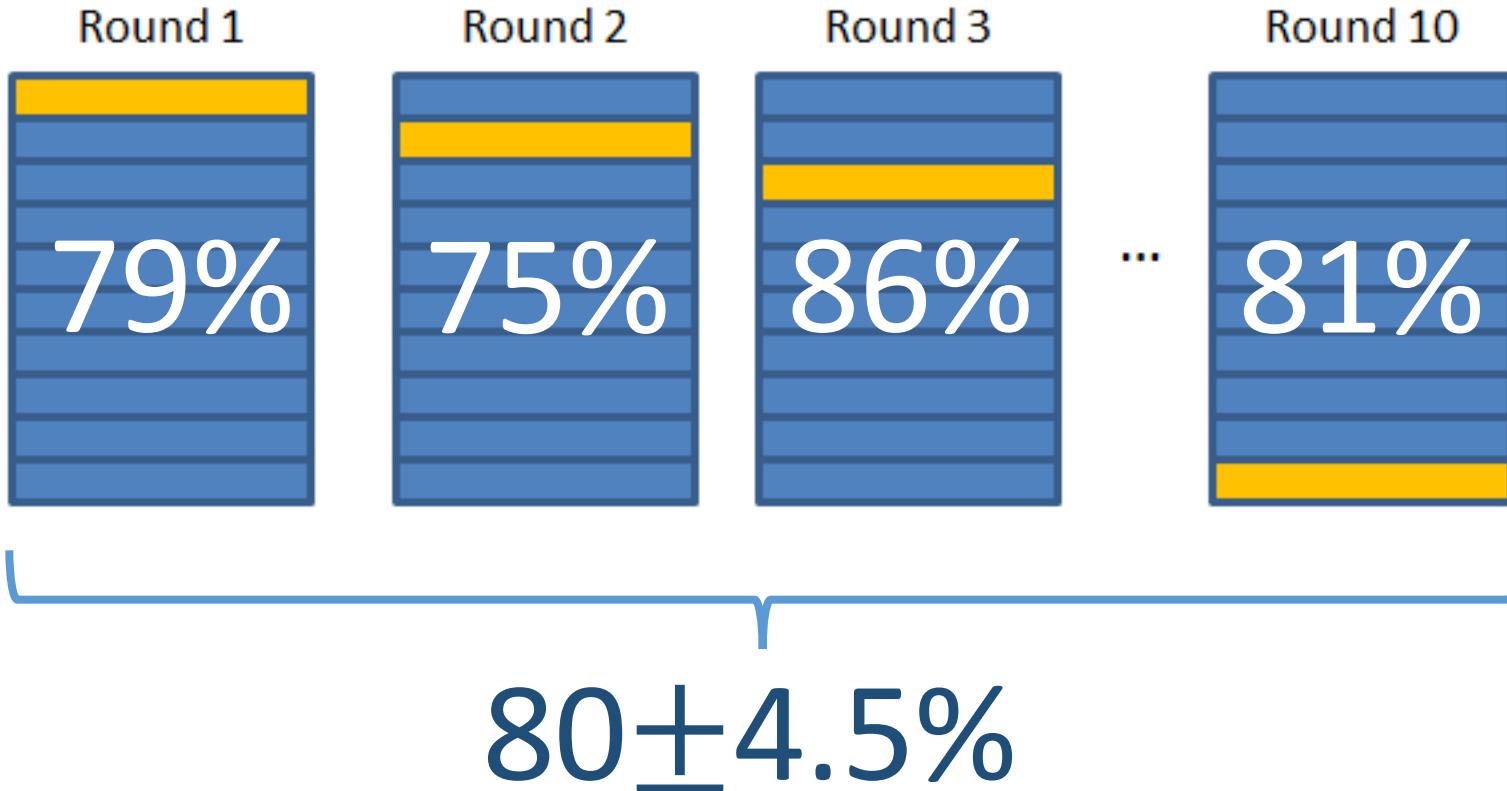
June 1967

March 2007

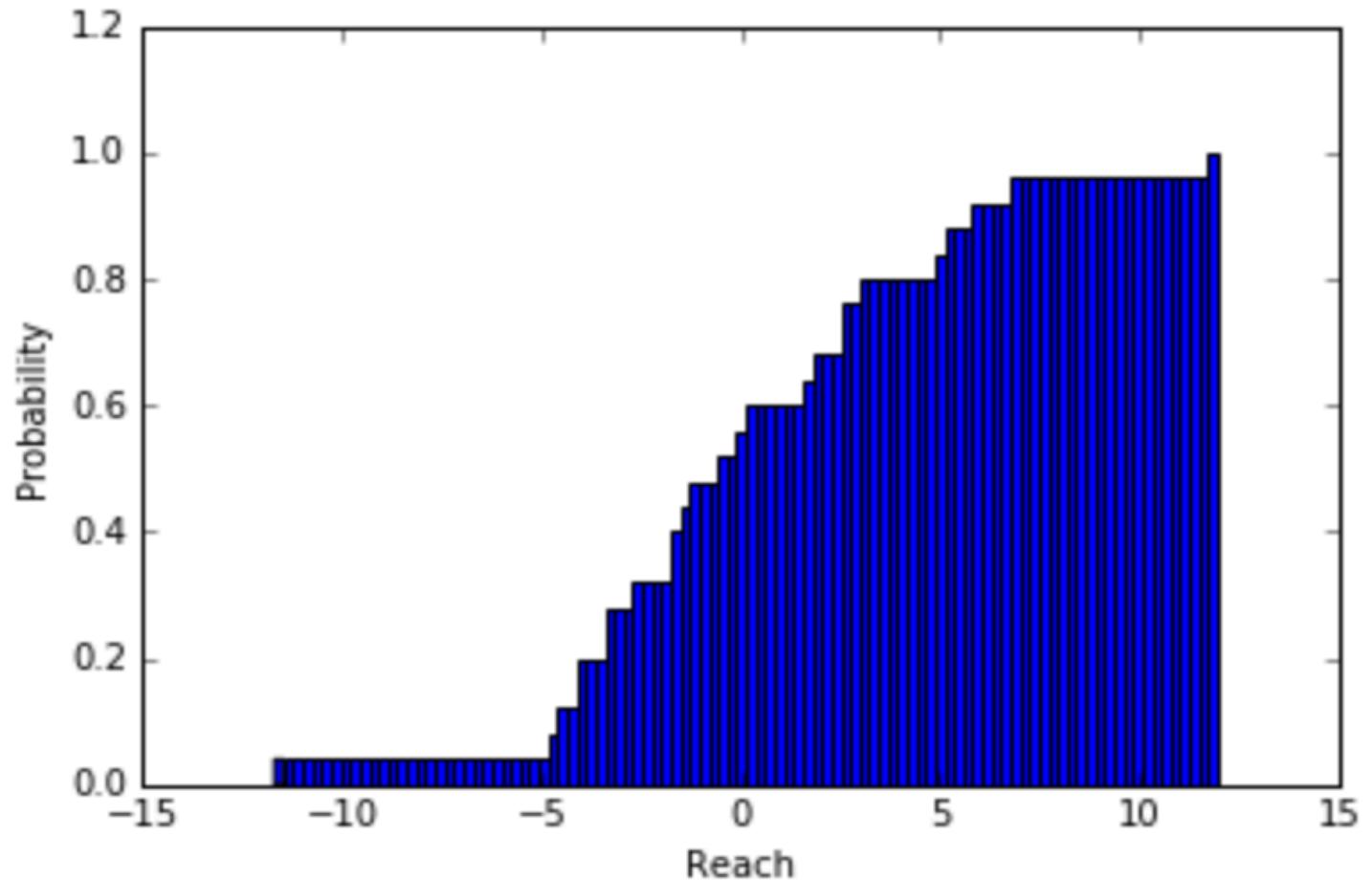


## Шаг 2: учет разброса и распределения в CV

 Validation Set  
 Training Set



## Шаг 2: учет разброса и распределения в CV



## Шаг 3: анализ топа важных признаков

На одном фолде:

0.211268 Номер  
0.147105 Ширина  
0.128326 Вес  
0.0954617 Параметр 1  
0.0688576 Высота  
0.057903 Параметр 2  
0.0438185 Параметр 3

...

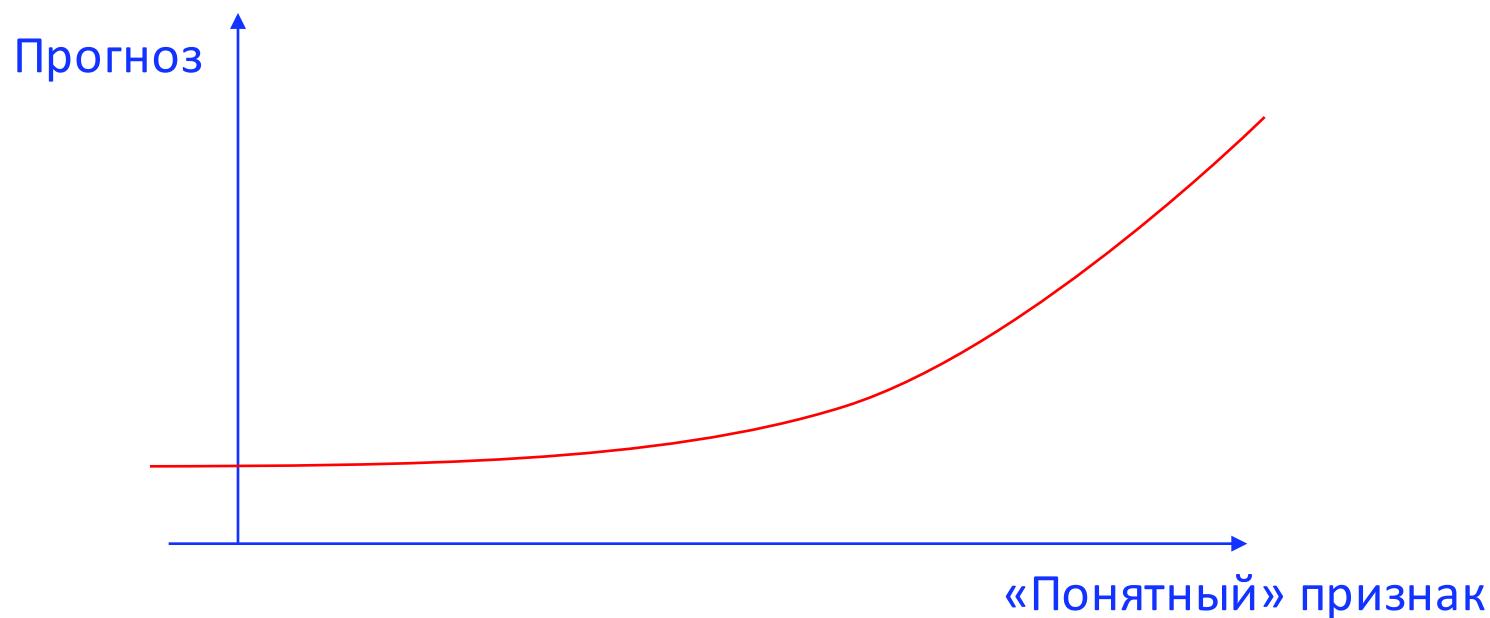
На другом:

0.285714 Номер  
0.163265 Параметр 1  
0.122449 Высота  
0.102041 Параметр 4  
0.0816327 Параметр 5  
0.0816327 Вес  
0.0612245 Параметр 2

...

## Шаг 4: Анализ зависимости от признаков

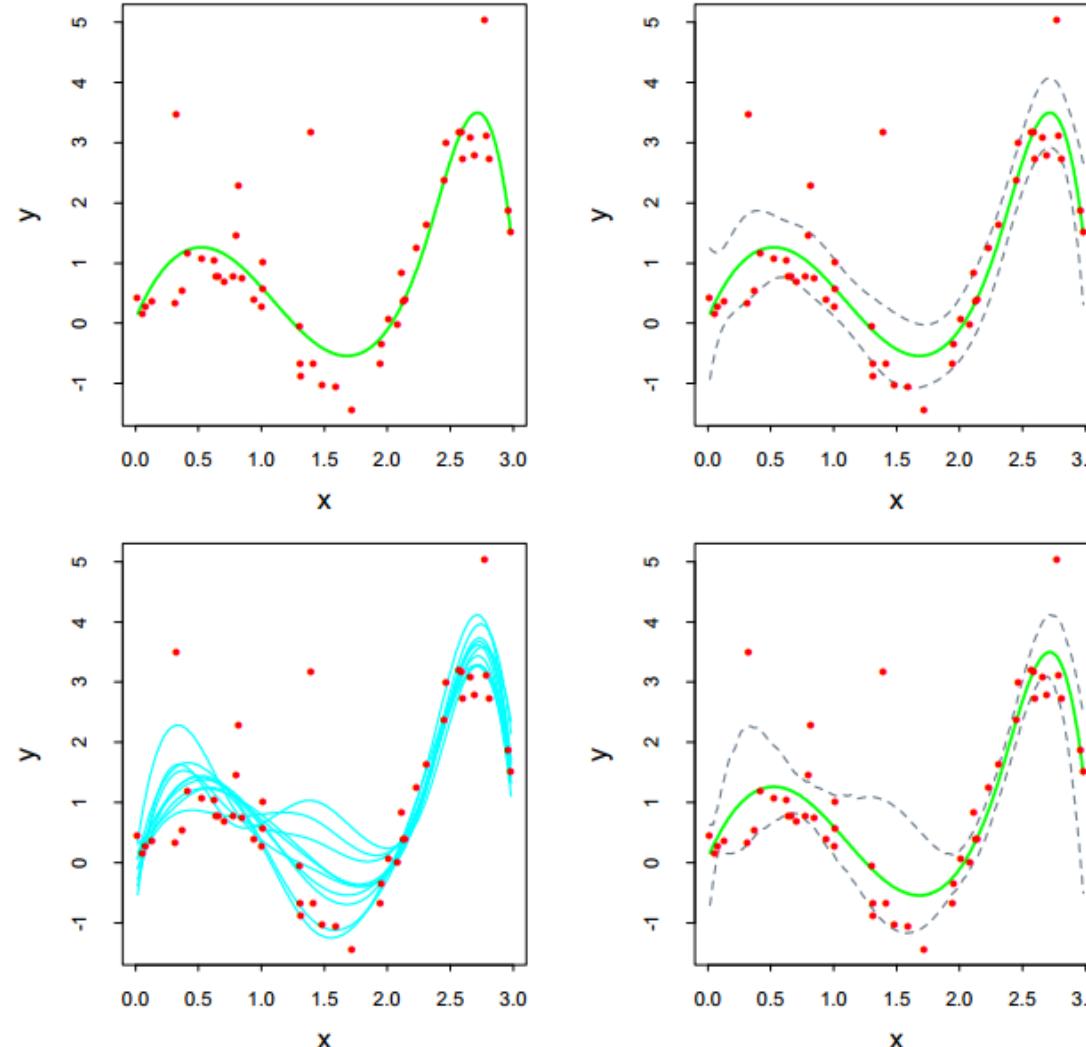
Если зависимость от каких-то признаков должна иметь понятный вид, можем поменять их (построить «искусственные» примеры) и посмотреть, как ведет себя прогноз



## Шаг 5: Уменьшение разброса

- Вариант 1: нахождение допущенных ошибок
- Вариант 2: более устойчивые модели

# Bagging



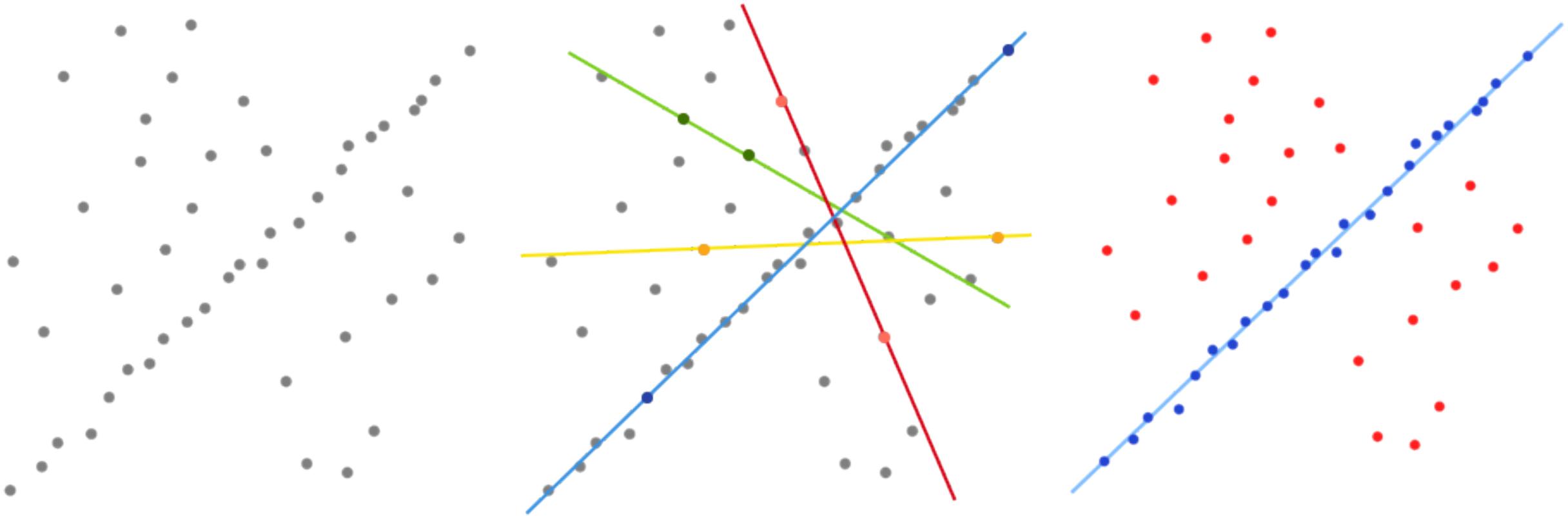
## Бэггинг в sklearn.ensembles

- BaggingRegressor
- BaggingClassifier

# Робастные модели в `sklearn.linear_model`

- RANSACRegressor
- HuberRegressor
- Theil-Sen Regressor

# RANSACRegressor



# HuberRegressor

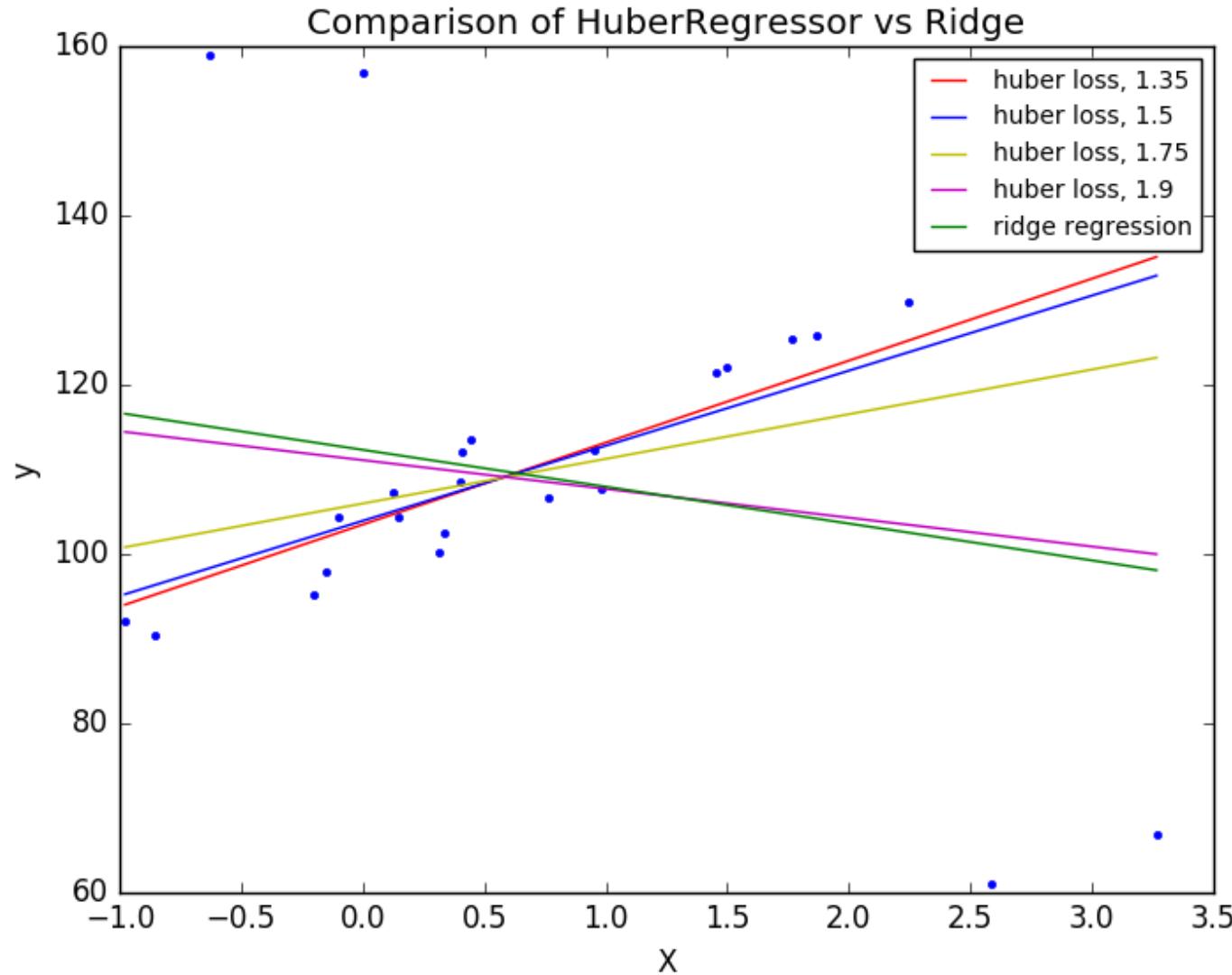
$$\min_{w, \sigma} \sum_{i=1}^n \left( \sigma + H_m \left( \frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha \|w\|_2^2$$

$$H_m(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$

# HuberRegressor

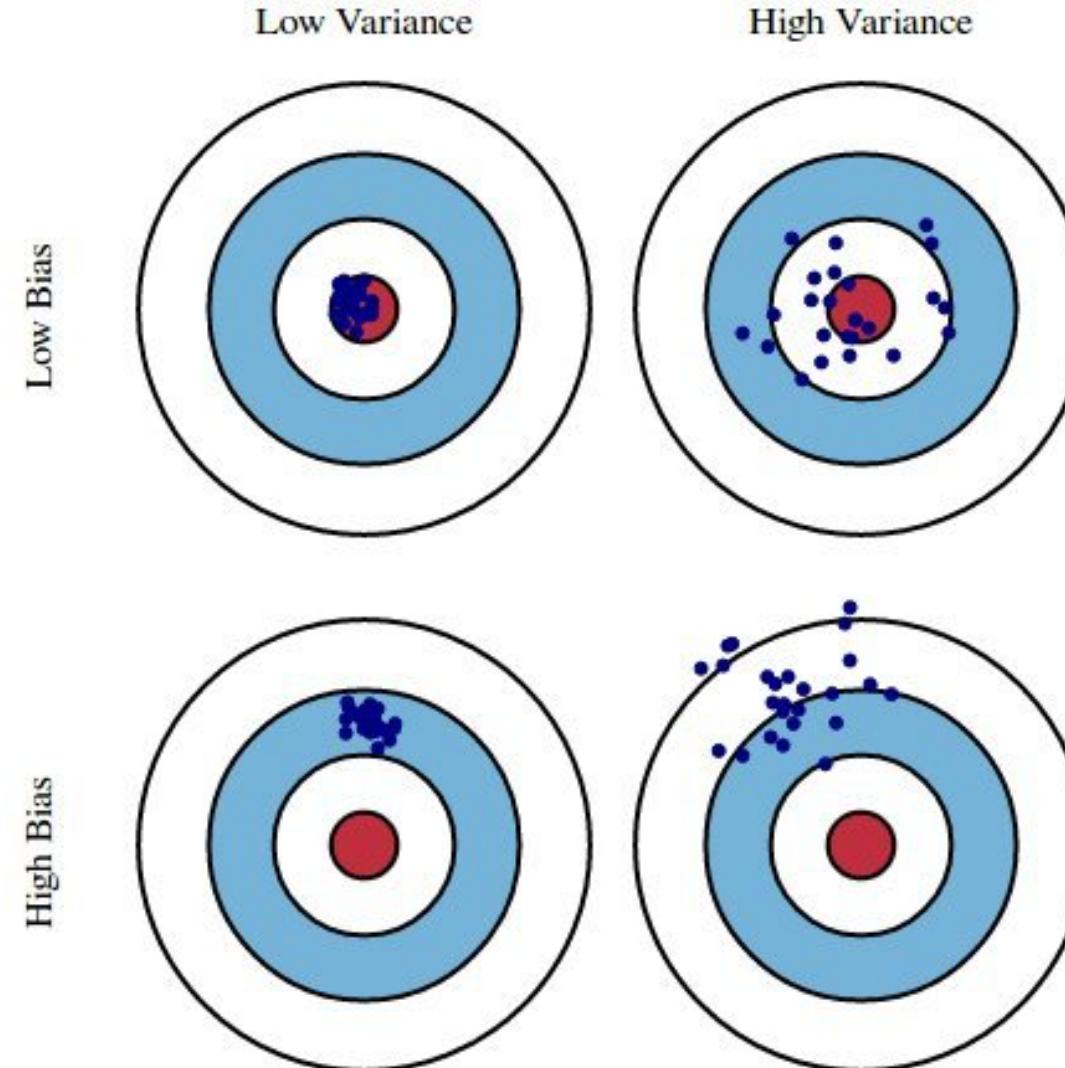
$$\min_{w, \sigma} \sum_{i=1}^n \left( \sigma + H_m \left( \frac{X_i w - y_i}{\sigma} \right) \sigma \right) + \alpha \|w\|_2^2$$

$$H_m(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases}$$



Дополнительная часть: bias-variance trade-off

# Bias-variance trade-off



# Bias-variance-noize decomposition

**Theorem.** For the squared error loss, the bias-variance decomposition of the expected generalization error at  $X = \mathbf{x}$  is

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = \text{noise}(\mathbf{x}) + \text{bias}^2(\mathbf{x}) + \text{var}(\mathbf{x})$$

where

$$\text{noise}(\mathbf{x}) = Err(\varphi_B(\mathbf{x})),$$

$$\text{bias}^2(\mathbf{x}) = (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2,$$

$$\text{var}(\mathbf{x}) = \mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}.$$

# Рассмотрели сегодня

- I. Метрики качества в задачах регрессии
- II. Метрики качества в задачах классификации
- III. Выбор метрик качества: пример
- IV. Устойчивость моделей

## Резюме

1. Существует множество различных метрик качества
2. Кроме того, можно придумывать их модификации
3. Важно выбрать такую, которая релевантна задаче
4. Полезно оценивать качество по нескольким метрикам
5. Важно оценивать стабильность модели

# Спасибо за внимание



[info@applieddatascience.ru](mailto:info@applieddatascience.ru)



[https://t.me/joinchat/B10lThC96v0BQCvs\\_joNew](https://t.me/joinchat/B10lThC96v0BQCvs_joNew)



[https://github.com/vkantor/ml2018jan\\_feb](https://github.com/vkantor/ml2018jan_feb)



<https://goo.gl/forms/jWxkcfcIEM5q4KUx1>