



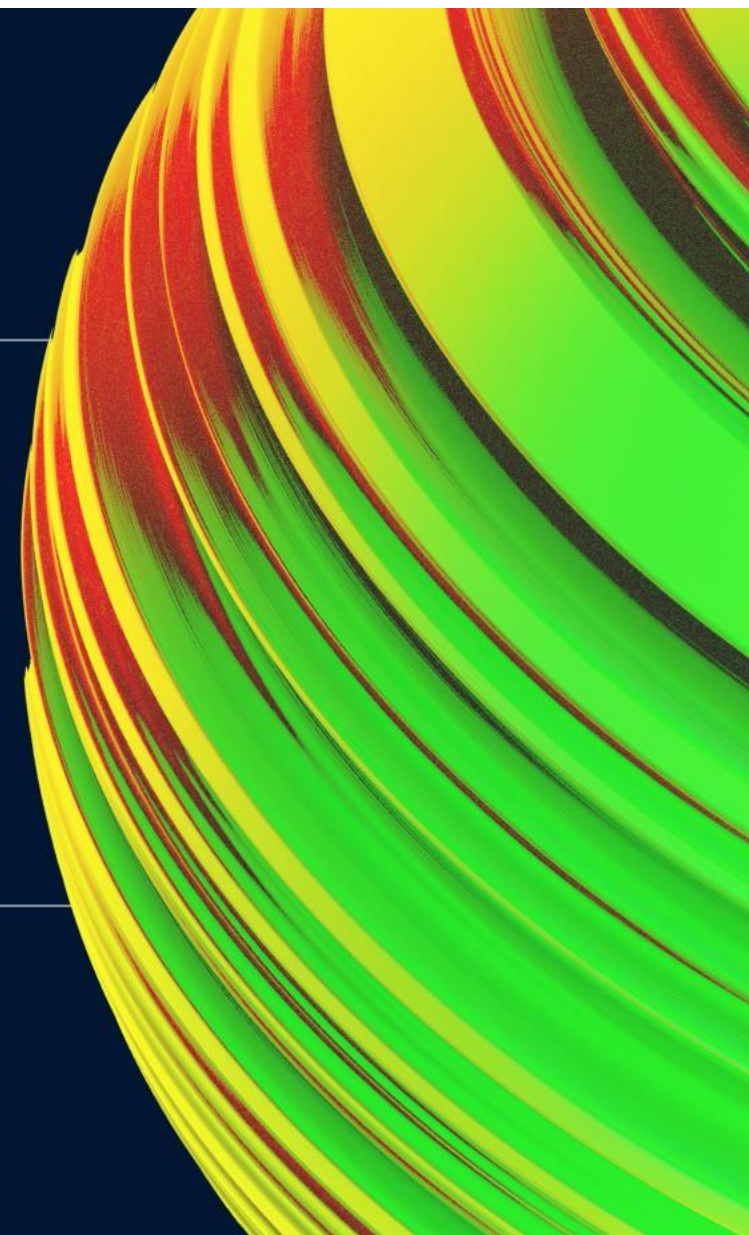
УНИВЕРСИТЕТ
ИННОПОЛИС

Введение в машинное обучение. KNN. Метрики качества. Матрица ошибок

Часть 2

Воробьёва Мария

- maria.vorobyova.ser@gmail.com
- [@SparrowMaria](#)



Важные понятия

Объект в обучающей выборке == строка

Целевой признак, target, label, метка, зависимая переменная, y

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

x, независимый признак, предиктор,

Классификация задач машинного обучения



Постановка задачи машинного обучения

Гиперпараметры модели — задаются перед обучением моделей

Параметры модели — то, что находим во время обучения

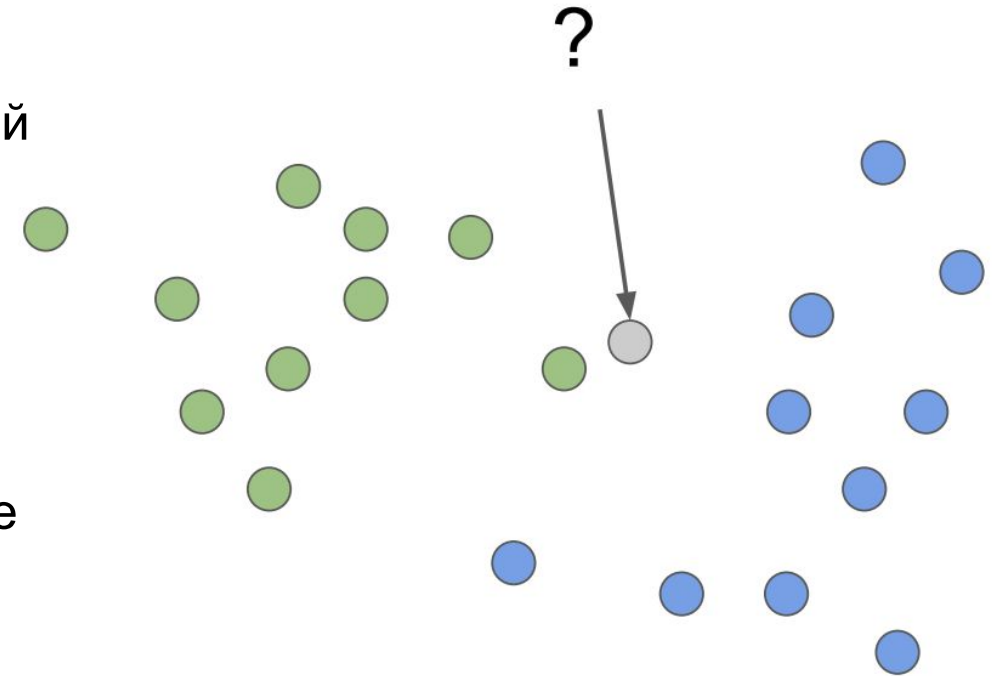
Loss function — функция, которая характеризует потери при неправильном принятии решений на основе наблюдений, то есть эта функция оценивает на сколько модель ошибается на данных

Метрики качества — это числовые показатели или статистики, используемые для оценки качества или производительности системы, модели или алгоритма. В контексте машинного обучения и анализа данных метрики часто используются для измерения точности, эффективности или других аспектов работы модели или алгоритма.

Алгоритм kNN - k Nearest Neighbours

Для **классификации** каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать объекты обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди ближайших соседей



Для **задачи регрессии** возвращается не метка, а число — среднее (или медианное) значение целевого признака среди соседей.

**А как понять, что мы построили
хорошую модель?**

Как измерить качество моделей?

Метрики качества зависят от типа целевой переменной:
если целевая числовая: вещественная

ID магазина	1	2	3	4	5	6	7	...	n	доход от магазина	прогноз		
1										1000000	999900	100	10000
2										200000	200100	100	10000
3										300000	299870	130	16900
4										500000	499800	200	40000
5										600000	600200	200	40000
6										1000000	1000150	150	22500
7										200000	199900	100	10000
8										300000	299900	100	10000
9										500000	499800	200	40000
10										600000	600200	200	40000
...										...			
k										600000	600300	300	90000
												161.8	29945.5

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - Y_i^p)^2,$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{N} abs(Y_i - Y_i^p)$$

где

Y_i^p - прогнозное значение ,

Y_i - фактическое значение,

N - количество объектов

Как измерить качество моделей?

Метрики качества зависят от типа целевой переменной:
если целевая числовая: вещественная

ID магазина	1	2	3	4	5	6	7	...	n	доход от магазина	прогноз		
1										1000000	999900	100	10000
2										200000	200100	100	10000
3										300000	299870	130	16900
4										500000	499800	200	40000
5										600000	600200	200	40000
6										1000000	1000150	150	22500
7										200000	199900	100	10000
8										300000	299900	100	10000
9										500000	499800	200	40000
10										600000	600200	200	40000
...										...			
k										600000	600300	300	90000
												161.8	29945.5

$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_i - Y_i^p)^2}{\sum_{i=1}^N (Y_i - Y_{avg})^2}$$

где

Y_i^p - прогнозное значение ,

Y_i - фактическое значение,

N - количество объектов

Коэффициент детерминации измеряет долю дисперсии, объяснённую моделью, в общей дисперсии целевой переменной.

Фактически, данная мера качества — это нормированная среднеквадратичная ошибка. Если она близка к единице, то модель хорошо объясняет данные, если же она близка к нулю, то прогнозы сопоставимы по качеству с константным предсказанием.

Как измерить качество моделей?

Не учитывает качество модели:

- Высокий R^2 не всегда означает, что модель хорошая. Модель может объяснять значительную часть дисперсии, но при этом иметь значительные ошибки предсказания.

Зависимость от числа предикторов:

- R^2 всегда увеличивается с добавлением новых предикторов, даже если они незначительно улучшают модель. Это может привести к избыточной подгонке модели (overfitting).

Не измеряет предсказательную силу:

- R^2 показывает, насколько хорошо модель объясняет дисперсию в обучающих данных, но не говорит о её предсказательной способности на новых данных. Для оценки предсказательной силы используются другие метрики, такие как RMSE, MAE и т.д.

$$R_{adj}^2 = 1 - \frac{s^2}{s_y^2} = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2) \frac{(n-1)}{(n-k)} \leq R^2,$$

где

$$RSS = \sum_n e_t^2 = \sum_n (y_t - \hat{y}_t)^2 \text{ — сумма квадратов остатков регрессии,}$$

$$TSS = \sum_n (y_t - \bar{y})^2 = n\hat{\sigma}_y^2 \text{ — общая дисперсия,}$$

n — количество наблюдений в наборе данных,

k — количество параметров модели.

Как измерить качество моделей?

MSE (или RMSE)

- сильное влияние оказывают выбросы
- интерпретируемая метрика

MAE

- интерпретируемая метрика
- не учитывает масштаб ошибки, 10 будет ошибкой и для 1000 и 1010 и для ситуации 10 и 20

Absolute Total Difference (разность между суммарным прогнозом и суммарным фактом) или **Bias (Relative Total Difference)** - отношение суммарного прогноза к суммарному факту минус 1)

- подсвечивает есть ли в целом сдвиг модели вверх или вниз
- понятна бизнесу
- метрика отвечает только за глобальную точность

Как измерить качество моделей?

Плюсы:

- Прост в интерпретации и расчете.
- Учитывает масштаб ошибки, отклонение на 10 ед при факте 2000 менее будет критично, чем при факте 20

Минусы:

- Неустойчив к нулевым значениям в базовых данных (когда фактические значения равны нулю), что может вызвать проблемы при расчете. Есть доработка, например, $\max(y, \epsilon)$ или y заменяем на $y + \epsilon$, но нет алгоритма как правильно выбрать ϵ
- Не симметричен и может давать разные результаты в зависимости от порядка фактических и прогнозных значений. За перепрогноз штрафует больше, чем за недопрогноз

$$MAPE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - f(x_i)|}{|y_i|}$$

Month Year	Actual Spend	Forecasted Spend	Absolute Percentage Error
Jan-22	0.2	1	400.00

Month Year	Actual Spend	Forecasted Spend	Absolute Percentage Error
Jan-22	500	600	20.00
Feb-22	600	500	16.67

Как измерить качество моделей?

$$SMAPE(y^{true}, y^{pred}) = \frac{1}{N} \sum_{i=1}^N \frac{2 |y_i - f(x_i)|}{y_i + f(x_i)}$$

SMAPE (Symmetric Mean Absolute Percentage Error):

Плюсы:

Симметричен и не зависит от порядка фактических и прогнозных значений.

Обрабатывает нулевые значения более устойчиво, чем MAPE.

Минусы:

Может быть бесконечным, если фактическое значение равно нулю, и прогноз также равен нулю.

Менее интуитивен в интерпретации

Как измерить качество моделей?

Метрики качества зависят от типа целевой переменной: если целевая категориальная:

- матрица ошибок (Confusion matrix)
- Accuracy
- точность и полнота (Precision и Recall)
- F1-мера
- ROC-кривая (ROC Curve)
- ROC AUC (площадь под ROC Curve)

Как измерить качество моделей?

Необходимо снова сравнить факт с прогнозом, задаем cutoff для прогноза и от оценки вероятностей переходим к прогнозу 0

↓

id клиента	fact	прогноз модели	бинарный прогноз модели
1	1	0.2	0
2	0	0.3	0
3	1	0.5	1
4	0	0.1	0
5	0	0.2	0
6	1	0.55	1
7	1	0.7	1
8	1	0.8	1
9	0	0.55	1



COUNT of id прогноз		
fact	0	1
0	3	1
1	1	4

Матрица ошибок

True Positives	False Positives	Число наблюдений классифицированных как P
False Negatives	True Negatives	Число наблюдений классифицированных как N
Число наблюдений из P (TP + FN)	Число наблюдений из N (FP + TN)	

Если результат классификации положительный (или 1) и фактическое значение тоже положительное (то есть тоже 1), то **TRUE POSITIVE (TP)**

Если результат классификации положительный (или 1) и фактическое значение отрицательное (то есть -1 или 0), то **FALSE POSITIVE (FP)**

Если результат классификации отрицательный (-1, или 0) и фактическое значение положительное (то есть тоже 1), то **TRUE NEGATIVE (TN)**

Если результат классификации отрицательный (-1 или 0) и фактическое значение тоже отрицательное (-1 или 0), то **FALSE NEGATIVE (FN)**

Accuracy

True Positives	False Positives	Число наблюдений классифицированных как P
False Negatives	True Negatives	Число наблюдений классифицированных как N
Число наблюдений из P (TP + FN)	Число наблюдений из N (FP + TN)	

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy

True Positives	False Positives	Число наблюдений классифицированных как P
False Negatives	True Negatives	Число наблюдений классифицированных как N
Число наблюдений из P (TP + FN)	Число наблюдений из N (FP + TN)	

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Пример:

пусть в выборке

1. 10000 выданных кредитов
2. 100 кредитов достигли просрочки 90+ за 1-ый год, то есть доля “1” = 0.01

Теперь возьмем “самый глупый” алгоритм, который всем объектам прогнозирует 0, то accuracy = 0.99.

Это хороший алгоритм?

Матрица ошибок

True Positives	False Positives	Число наблюдений классифицированных как P
False Negatives	True Negatives	Число наблюдений классифицированных как N
Число наблюдений из P (TP + FN)	Число наблюдений из N (FP + TN)	

$$\text{Precision} = TP / (TP + FP)$$

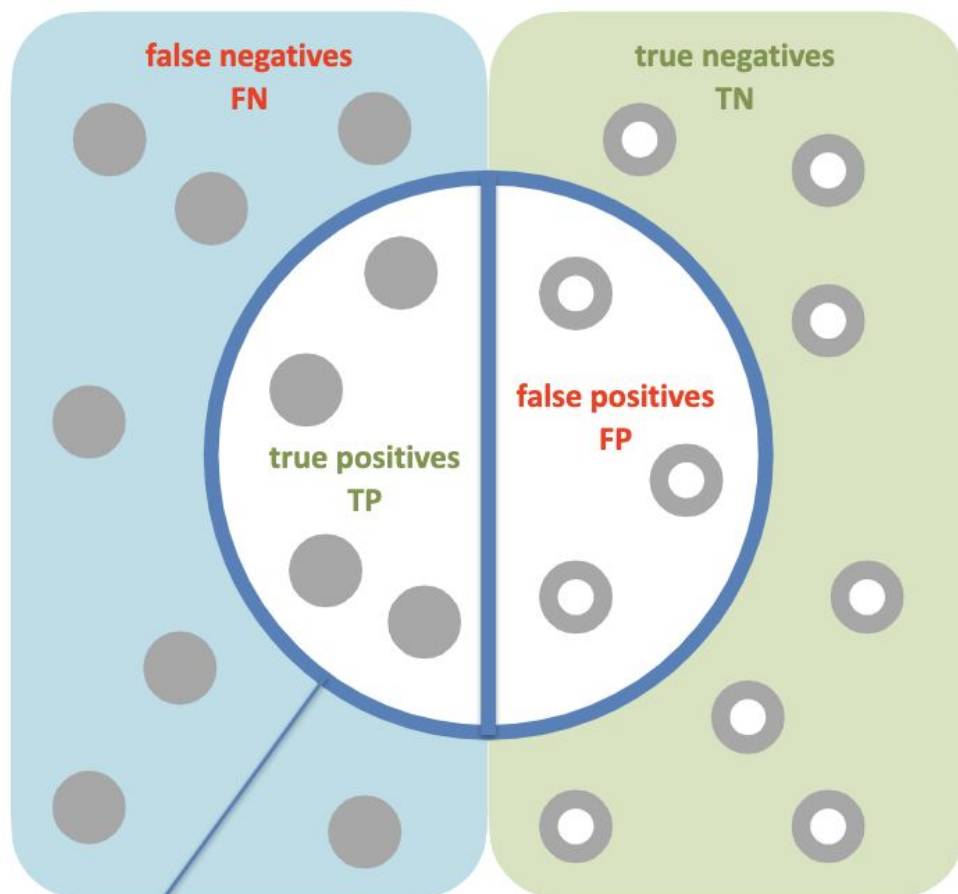
Чувствительность алгоритма
(sensitivity) = $TP / (TP + FN)$

$$\text{Recall} = TP / (TP + FN)$$

$$\text{TPR (True Positive Rate)} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP) = 1 - \text{FPR} = 1 - FP / (TN + FP)$$

Precision и recall



Выбранные объекты

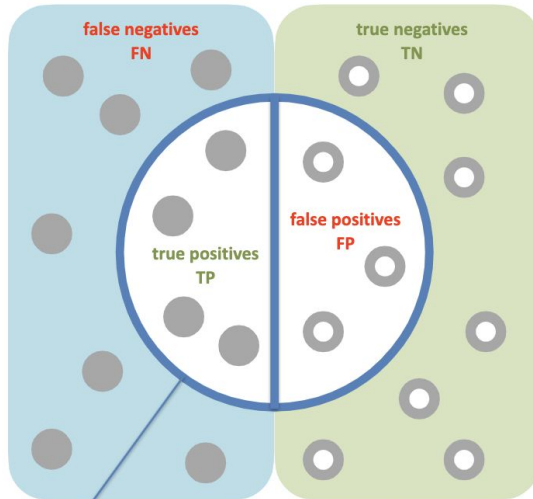
Сколько выбранных
объектов **корректны**?

$$precision = \frac{TP}{TP + FP}$$

Как много корректных
объектов выбрано?

$$recall = \frac{TP}{TP + FN}$$

Precision и recall



Сколько выбранных
объектов **корректны**?

$$precision = \frac{TP}{TP + FP}$$

Как **много** корректных
объектов выбрано?

$$recall = \frac{TP}{TP + FN}$$

Выбранные объекты

Пример:

пусть в выборке

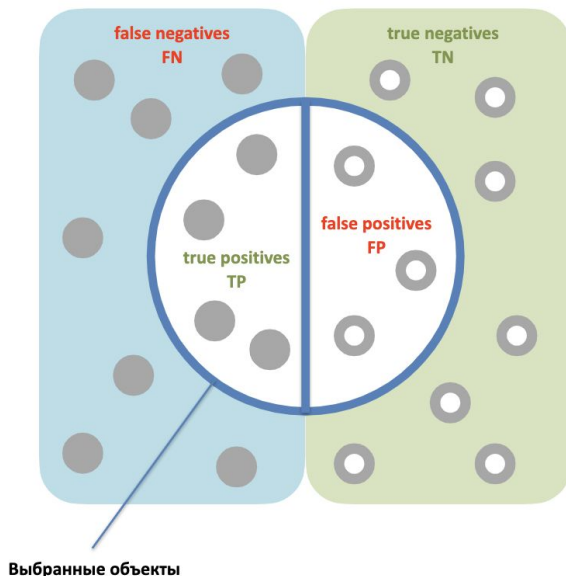
1. 10000 выданных кредитов
2. 100 кредитов достигли просрочки 90+ за 1-ый год, то есть доля "1" = 0.01

Теперь возьмем "самый глупый" алгоритм, который всем объектам прогнозирует 0

Precision = 0/0

Recall = 0/100 = 0

Precision и recall



Сколько выбранных
объектов **корректны**?

$$precision = \frac{TP}{TP + FP}$$

Как много корректных
объектов выбрано?

$$recall = \frac{TP}{TP + FN}$$

Выбранные объекты



Невозможно уменьшить и precision, и recall

Что же делать?

Все зависит от задачи:

Например, задача **детектировать болезнь клиента**:

Тогда нам важен recall, пропустить клиента с болезнью будет хуже, чем если мы кого-то лишний раз проверим

а вот в задаче **не пропустить безбилетника**, важна точность, то есть precision

Как объединить Precision и Recall

$$F_1 = 2 \cdot \frac{\textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{\textit{precision} \cdot \textit{recall}}{(\beta^2 \cdot \textit{precision}) + \textit{recall}}$$

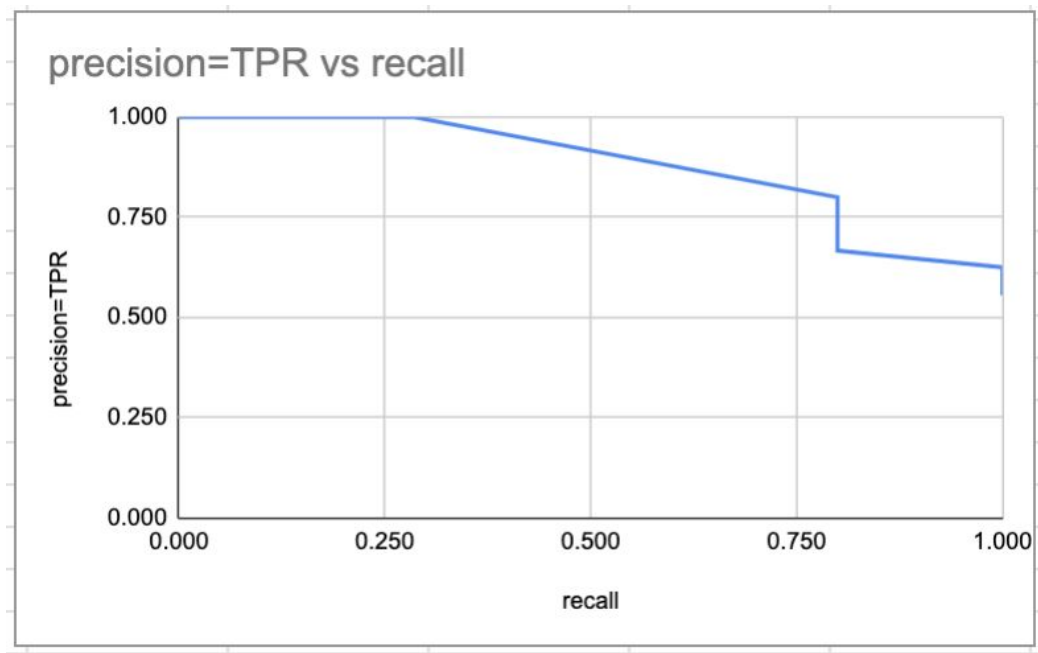
PR curve и ROC

id клиента	fact	прогноз модели	бинарный прогноз модели для cutoff 0.9	бинарный прогноз модели для cutoff 0.8	бинарный прогноз модели для cutoff 0.7	бинарный прогноз модели для cutoff 0.6	бинарный прогноз модели для cutoff 0.5	бинарный прогноз модели для cutoff 0.4	бинарный прогноз модели для cutoff 0.3	бинарный прогноз модели для cutoff 0.2	бинарный прогноз модели для cutoff 0.1
8	1	0,8	0	1	1	1	1	1	1	1	1
7	1	0,7	0	0	1	1	1	1	1	1	1
6	1	0,55	0	0	0	0	1	1	1	1	1
9	0	0,55	0	0	0	0	1	1	1	1	1
3	1	0,5	0	0	0	0	1	1	1	1	1
2	0	0,3	0	0	0	0	0	0	1	1	1
1	1	0,2	0	0	0	0	0	0	0	1	1
5	0	0,2	0	0	0	0	0	0	0	1	1
4	0	0,1	0	0	0	0	0	0	0	0	1

TP	0	1	2	2	4	4	4	5	5
FP	0	0	0	0	1	1	2	3	4
FN	5	5	5	5	1	1	1	0	0
TN	4	4	4	4	3	3	2	1	0
precision	1.000	1.000	1.000	1.000	0.800	0.800	0.667	0.625	0.556
recall=TPR	0.000	0.167	0.286	0.286	0.800	0.800	0.800	1.000	1.000
FPR	0	0	0	0	0.25	0.25	0.5	0.75	1

Precision-recall curve

TP	0	1	2	2	4	4	4	5	5
FP	0	0	0	0	1	1	2	3	4
FN	5	5	5	5	1	1	1	0	0
TN	4	4	4	4	3	3	2	1	0
precision	1.000	1.000	1.000	1.000	0.800	0.800	0.667	0.625	0.556
recall=TPR	0.000	0.167	0.286	0.286	0.800	0.800	0.800	1.000	1.000

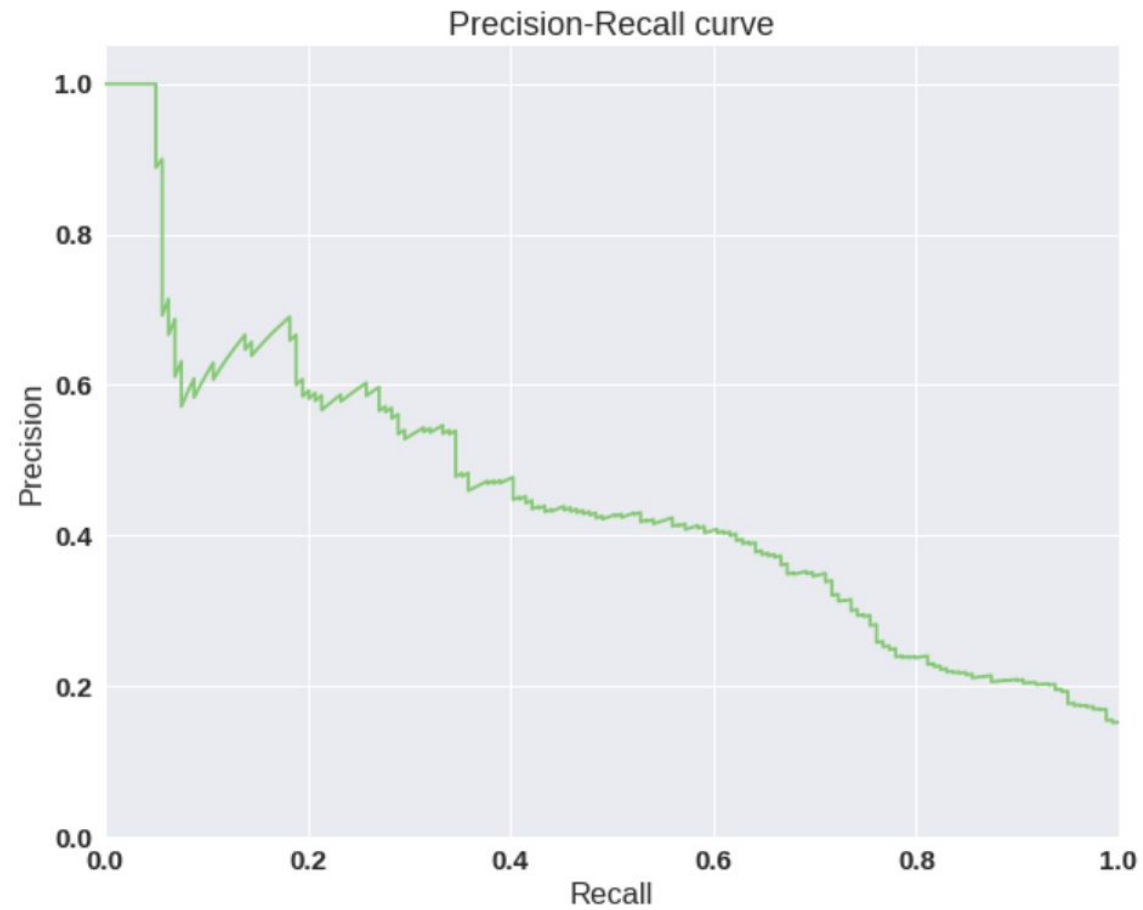


Кривая начинается в точке 0, 1. Договорились , что precision = 1

Завершается график точкой (1, доля 1 в выборке)

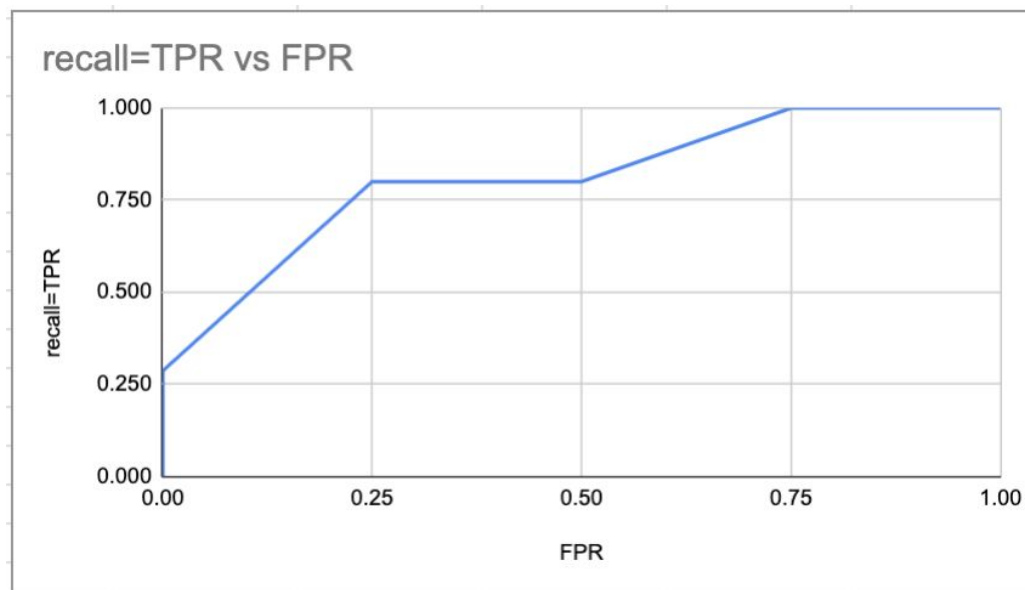
Можно посчитать площадь под кривой PRC: AUC PRC

Precision-recall curve



ROC кривая

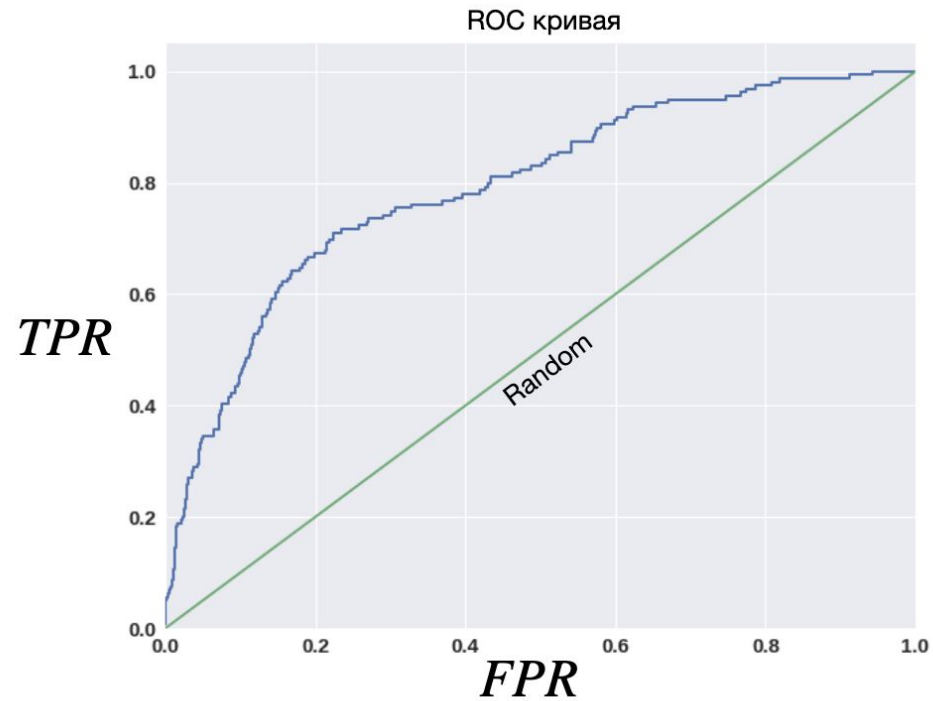
TP	0	1	2	2	4	4	4	5	5
FP	0	0	0	0	1	1	2	3	4
FN	5	5	5	5	1	1	1	0	0
TN	4	4	4	4	3	3	2	1	0
recall=TPR	0.000	0.167	0.286	0.286	0.800	0.800	0.800	1.000	1.000
FPR	0	0	0	0	0.25	0.25	0.5	0.75	1



Кривая начинается в точке 0, 0. И заканчивается в точке 1,1

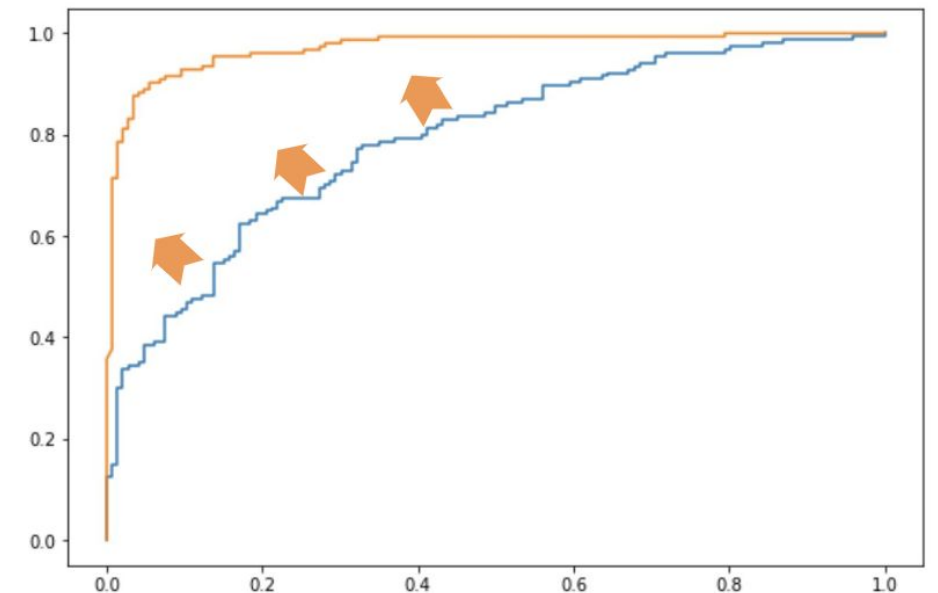
Можно посчитать площадь под кривой PRC: AUC ROC

ROC кривая

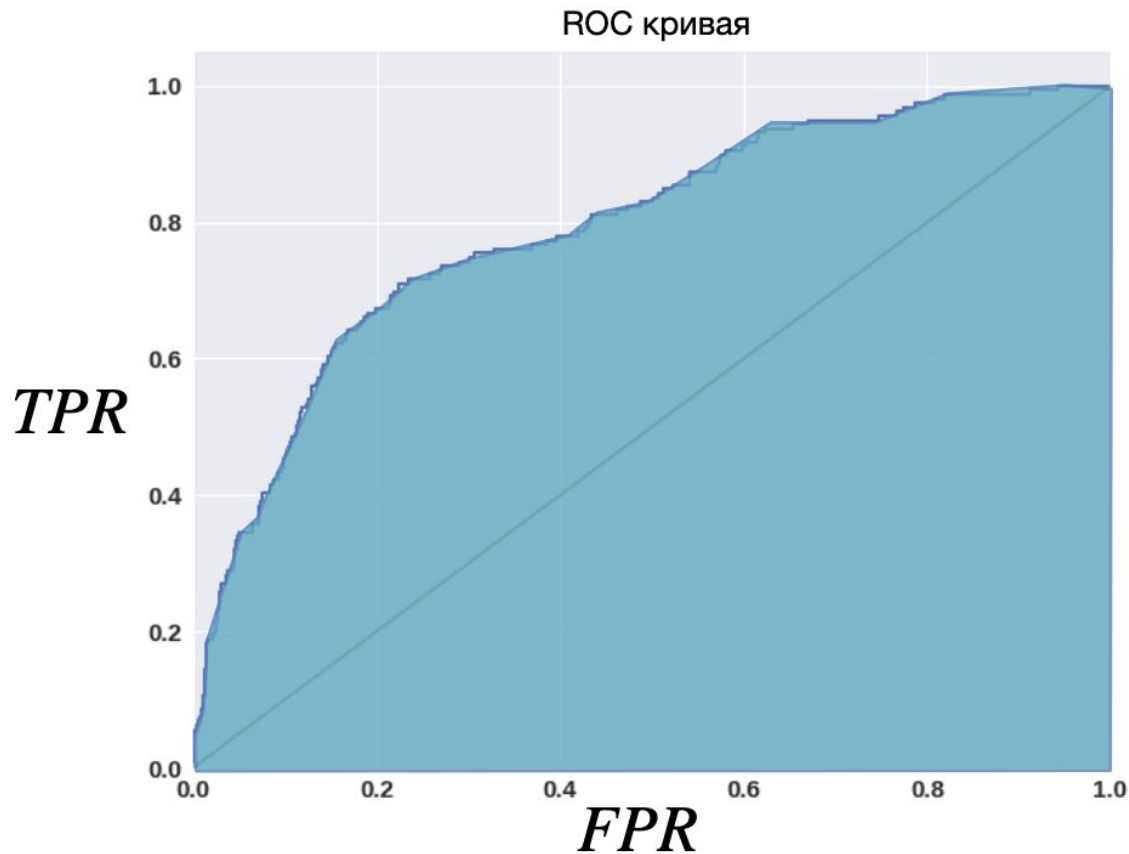


$$TPR = 1$$

$$FPR = 0$$



ROC кривая



AUC (Area Under ROC Curve) – это площадь под графиком ROC-кривой.

Эта величина используется для сравнения нескольких классификаторов:

- 0.5 – соответствует случайному классификатору
- 1.0 – соответствует идеальному классификатору

Когда ROC AUC плохо?

пусть дана выборка 1 - 100, 0 - 1 000 000

Мы построили алгоритм

$PR = 50000/1 \text{ млн} = 0.05$

факт	
0	первым 50000
1	для наших 100
0	
..	
0	
	остальное

TPR = 1

F

В итоге площадь AUC ROC = 0.95

а площадь AUC PRC = 0.05

Смысл ROC AUC

AUC-ROC интерпретация:

AUC-ROC равен **вероятности** того, что для случайно выбранных объекта good (x_+) и объекта bad (x_-) будет выполнено неравенство **прогноз(x_+) > прогноз(x_-)**

Смысл ROC AUC

AUC-ROC равен **вероятности** того, что для случайно выбранных объекта good (x_+) и объекта bad (x_-) будет выполнено неравенство **прогноз(x_+) > прогноз(x_-)**

Сравниваем всевозможные пары

To find concordant, discordant, and tied pairs, compare everyone who had the outcome of interest against everyone who did not.

good



bad



Смысл ROC AUC

AUC-ROC равен **вероятности** того, что для случайно выбранных объекта good (x_+) и объекта bad (x_-) будет выполнено неравенство **прогноз(x_+) > прогноз(x_-)**

Concordant Pair

Compare a 20-year-old who bad with a 30-year-old who good.

Good, Age 30



$P(\text{bad}) = .4077$

bad, Age 20



$P(\text{bad}) = .4272$

The actual sorting agrees with the model.
This is a **concordant pair**.

Смысл ROC AUC

AUC-ROC равен **вероятности** того, что для случайно выбранных объекта good ($x+$) и объекта bad ($x-$) будет выполнено неравенство **прогноз($x+$) > прогноз($x-$)**

Discordant Pair

Compare a 45-year-old who is bad with a 35-year-old who is good.

Good, Age 35



$P(\text{bad}) = .3981$

bad, Age 45



$P(\text{bad}) = .3791$



The actual sorting disagrees with the model.
This is a **discordant** pair.

Смысл ROC AUC

AUC-ROC равен **вероятности** того, что для случайно выбранных объекта good (x_+) и объекта bad (x_-) будет выполнено неравенство **прогноз(x_+) > прогноз(x_-)**

Tied Pair

Compare two 50-year-olds.

Good , Age 50	bad , Age 50
	
$P(\text{bad}) = .3697$	$P(\text{bad}) = .3697$

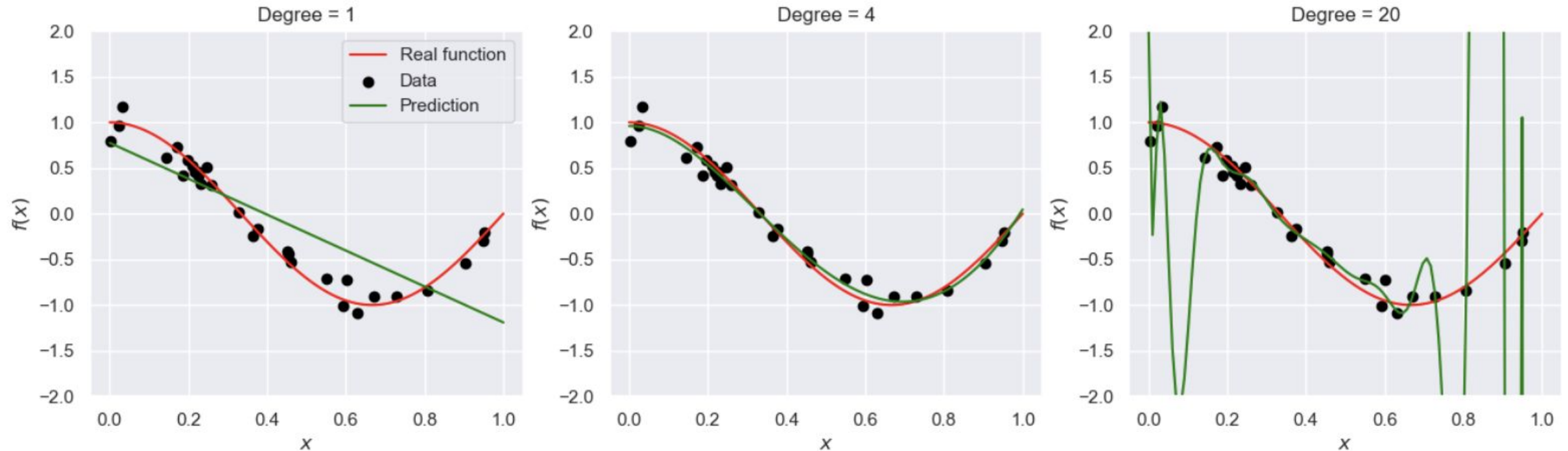
The model cannot distinguish between the two.
This is a **tied pair**.

Теперь все ясно!

**надо просто строить модель с идеальными
метриками качества!**

**Но тут мы встречаемся с такой проблемой как
ПЕРЕОБУЧЕНИЕ**

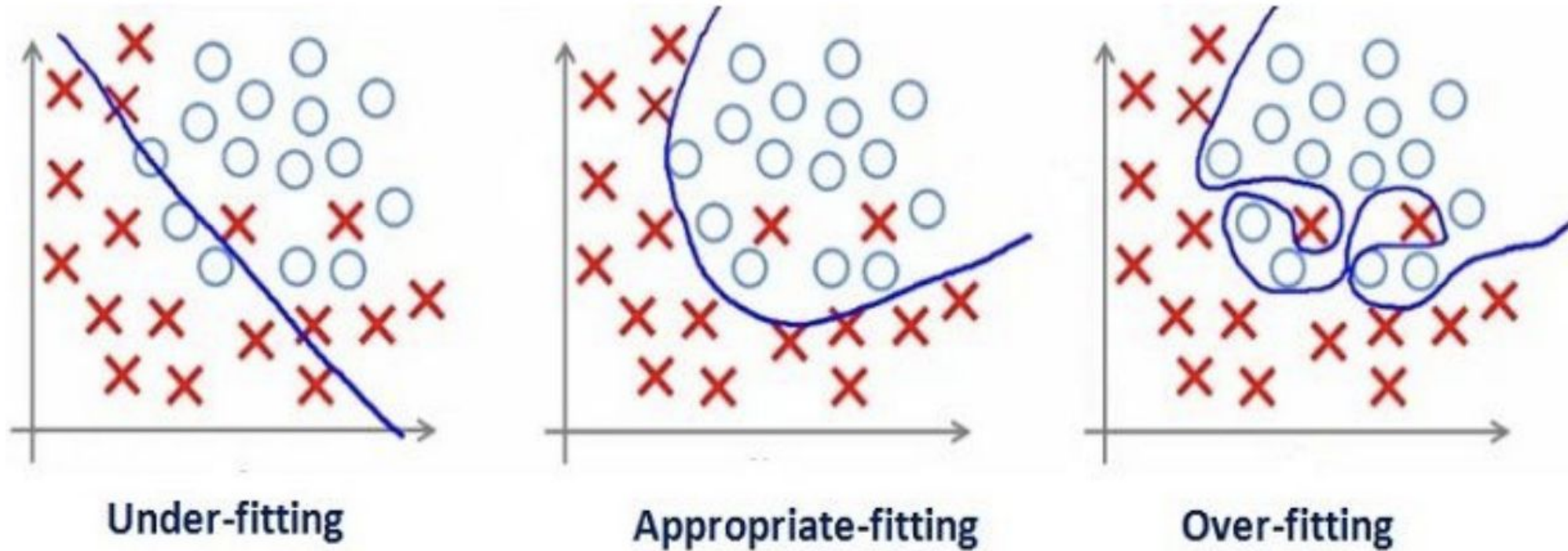
Проблема переобучения



Недообучение (underfitting) - модель слишком проста

Переобучение (overfitting) - модель слишком сложна

Проблема переобучения



Недообучение (underfitting) - модель слишком проста

Переобучение (overfitting) - модель слишком сложна

Проблема переобучения

Причина переобучения

- слишком сложная модель
- избыточные параметры в модели, то есть наблюдается мультиколлинеарность
- иногда переменные из “будущего”

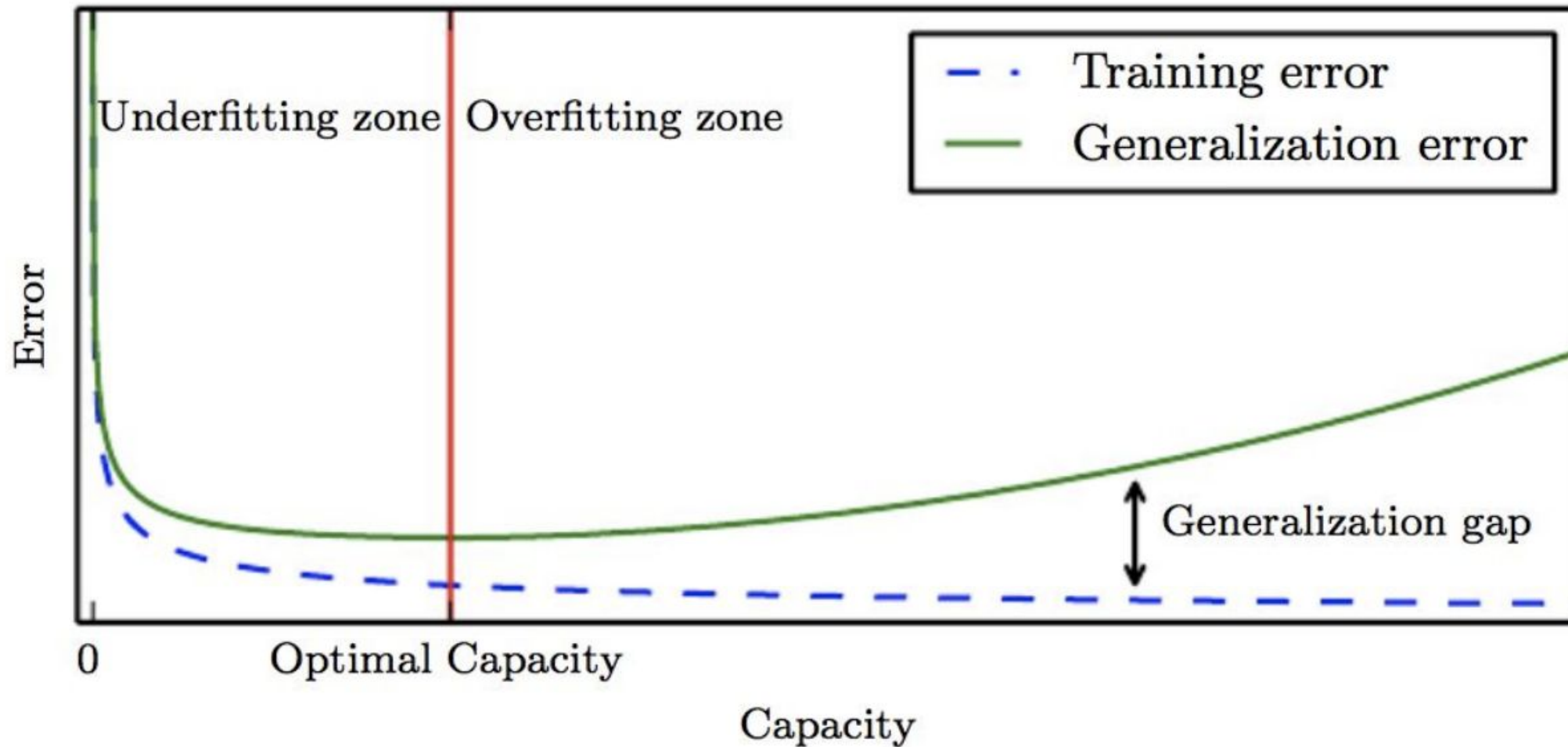
Как обнаружить переобучение

- эмпирически, путем разбиения на выборки и измерения качества моделей на отложенной выборке (тестовой)

Избавить от переобучения нельзя, его можно МИНИМИЗИРОВАТЬ

- накладывать ограничения на коэффициенты при независимых признаках
- выбрать модель по оценкам обобщающей способности

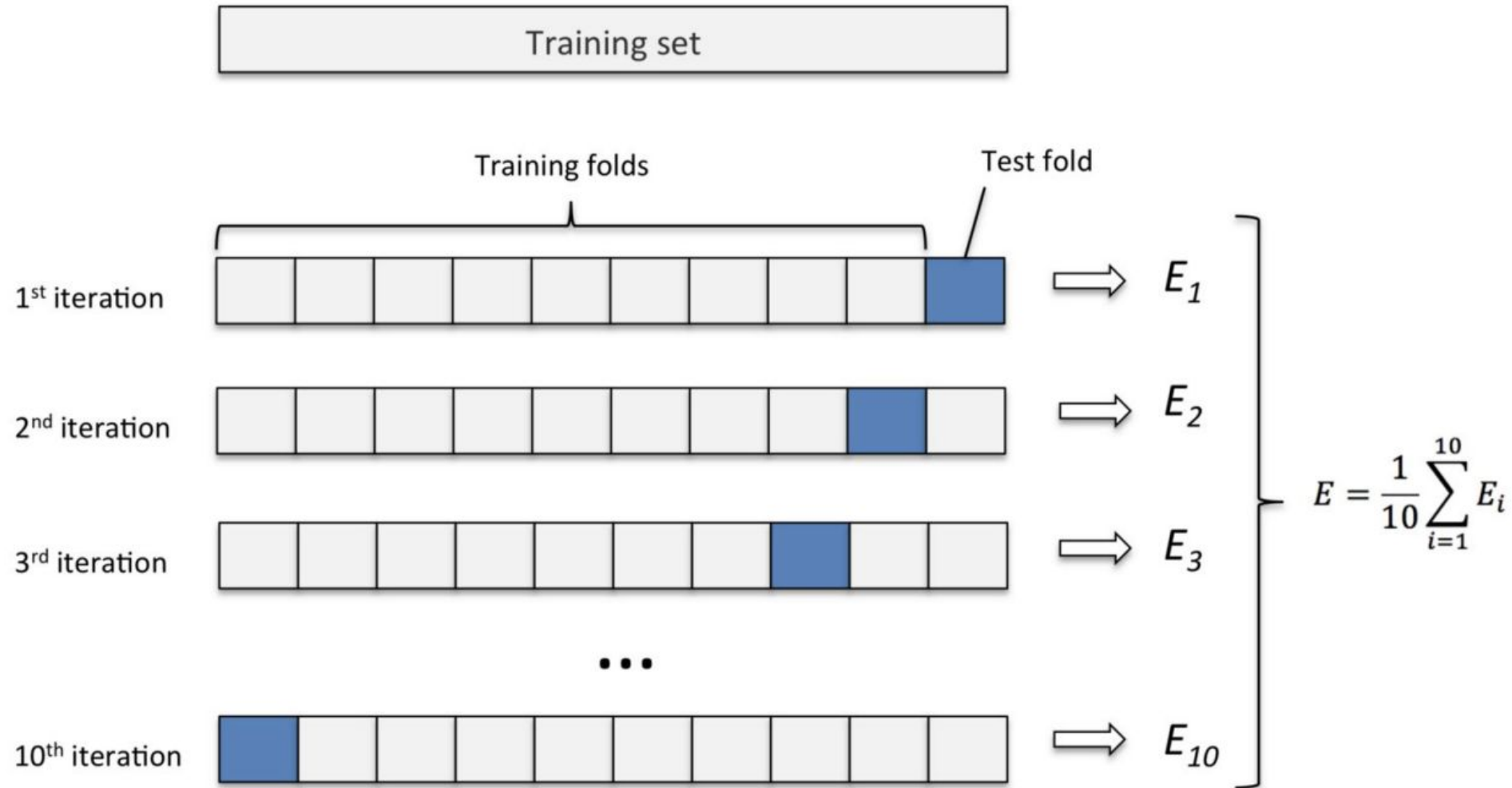
Проблема переобучения



Проблема переобучения



Проблема переобучения



Проблема переобучения

Утечка целевого признака

Утечка целевого признака (target leakage) — это проблема в моделировании машинного обучения, когда информация, которая не должна быть доступна модели при обучении, оказывается включенной в тренировочные данные.

Примеры утечки целевого признака:

1. **Использование будущих данных:** Если в тренировочном наборе данных имеются признаки, которые могут включать в себя информацию о будущем состоянии целевого признака, это может привести к утечке. Например, использование данных о продажах в следующем месяце для предсказания продаж в текущем месяце.
2. **Использование напрямую связанных признаков:** Если признаки напрямую связаны с целевым значением и могут помочь его предсказать без выполнения реальной задачи предсказания, это также является утечкой. Например, использование данных о завершенности транзакции для предсказания, завершится ли она.

**Хорошо! с чем же мы еще можем столкнуться,
когда будем строить модели?**

С чем можно столкнуться во время обучения модели?

- пропуски
- категориальные значения
- выбросы

Почему пропуск - это проблема?

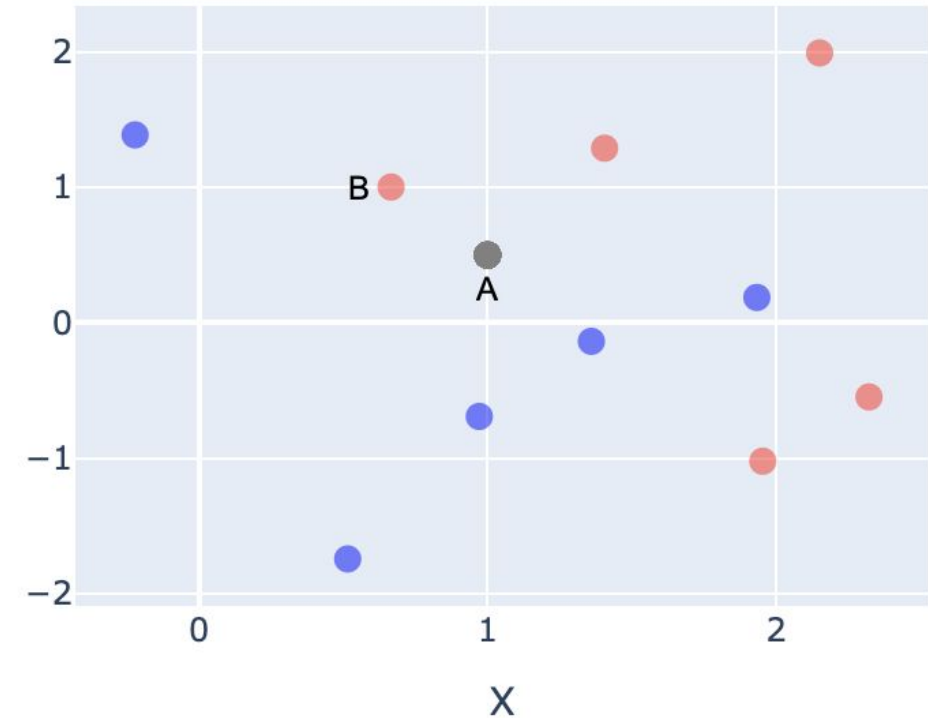
На примере алгоритма knn:

Для поиска ближайших соседей мы должны рассчитать расстояние между “серой точкой” и всеми остальными точками

Для примера: пусть координаты “серой точкой” $A = (1, 0.5)$, мы сможем рассчитать расстояние между всеми точками и дальше сделать вывод к какому классу принадлежит точка. например, расстояние между точками A $dist(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$

А если к нам приходит новая точка $C = (NaN, 1)$, то чему равно $dist(C, B) = ((NaN - 0.8)**2 + ((1-0.5)**2)**(1/2))$?

>



Что делаем с пропусками?

Решение:

- заполнить средним значением из распределения или другие статистики из распределения
- экстраполировать/интерполировать значения
- удалить наблюдения с пропусками (делать ОСТОРОЖНО)

например, мы можем рассчитать среднее = 1198 и пропуски заменить на 1198

ID магазина	площадь	количество этажей	в ТЦ?	доход от магазина
1	1000	1	1	1000000
2	1569	2	0	200000
3	870	1	0	300000
4	2000	2	0	500000
5	900	1	1	600000
6	850	1	1	1000000
7	1700	2	1	200000
8		2	1	300000
9		2	0	500000
10	700	1	0	600000

ID магазина	площадь	количество этажей	в ТЦ?	доход от магазина
1	1000	1	1	1000000
2	1569	2	0	200000
3	870	1	0	300000
4	2000	2	0	500000
5	900	1	1	600000
6	850	1	1	1000000
7	1700	2	1	200000
8	1198	2	1	300000
9	1198	2	0	500000
10	700	1	0	600000

Что делаем с пропусками?

Решение:

- построить модель, которая восстанавливает значение на основании других характеристик

То есть строим регрессию на наблюдениях с 1 по 7 и 10. Целевая переменная y = площадь магазина, независимые признаки: x_1 = “количество этажей” и x_2 = “в ТЦ”

$$y = 700 + 300 \cdot x_1 + 100 \cdot x_2$$

ID магазина	площадь	количество этажей	в ТЦ?	доход от магазина
1	1000	1	1	1000000
2	1569	2	0	200000
3	870	1	0	300000
4	2000	2	0	500000
5	900	1	1	600000
6	850	1	1	1000000
7	1700	2	1	200000
8		2	1	300000
9		2	0	500000
10	700	1	0	600000

ID магазина	площадь	количество этажей	в ТЦ?	доход от магазина
1	1000	1	1	1000000
2	1569	2	0	200000
3	870	1	0	300000
4	2000	2	0	500000
5	900	1	1	600000
6	850	1	1	1000000
7	1700	2	1	200000
8	1400	2	1	300000
9	1300	2	0	500000
10	700	1	0	600000

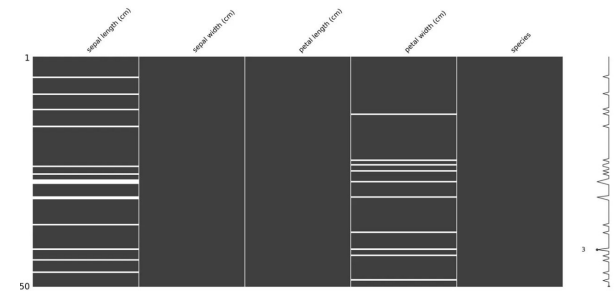
Что делаем с пропусками?

На практике:

Сначала необходимо обнаружить пропуски:

- 1) либо методы describe, либо isna().sum()
- 2) смотрим комплексно

```
import missingno as msno
msno.matrix(df_miss, figsize=(10, 6))
```



Далее заменяем:

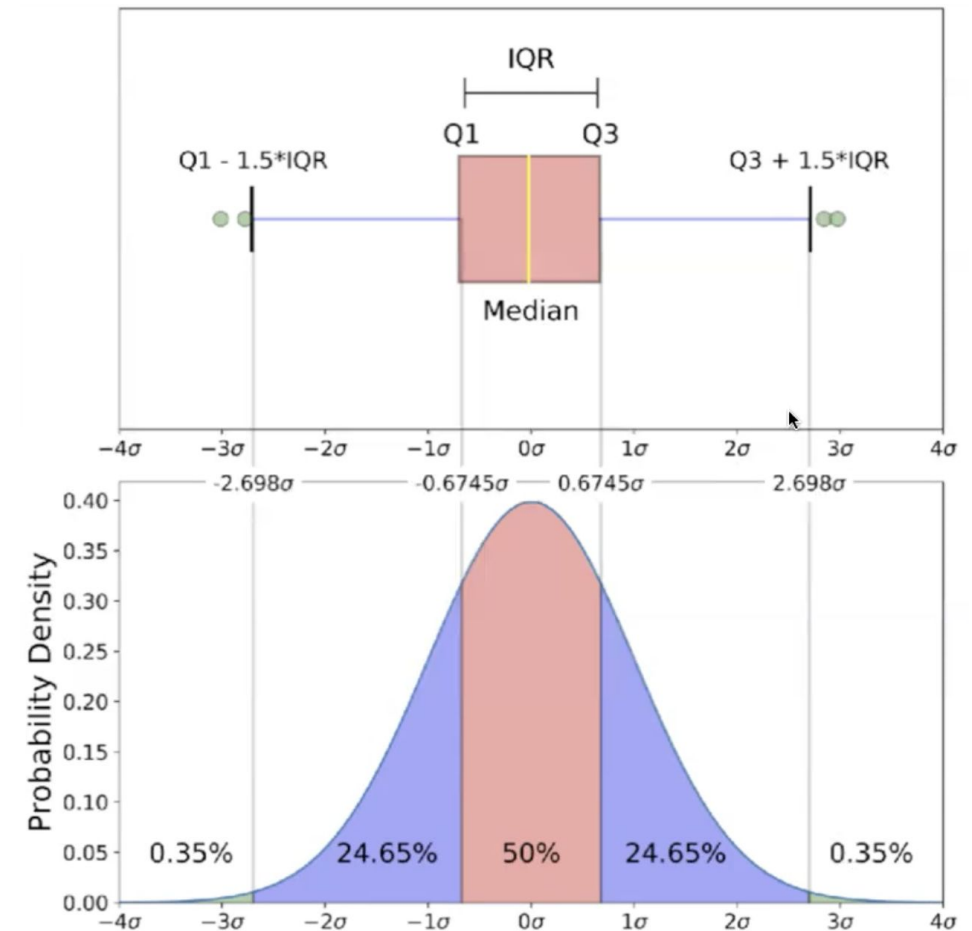
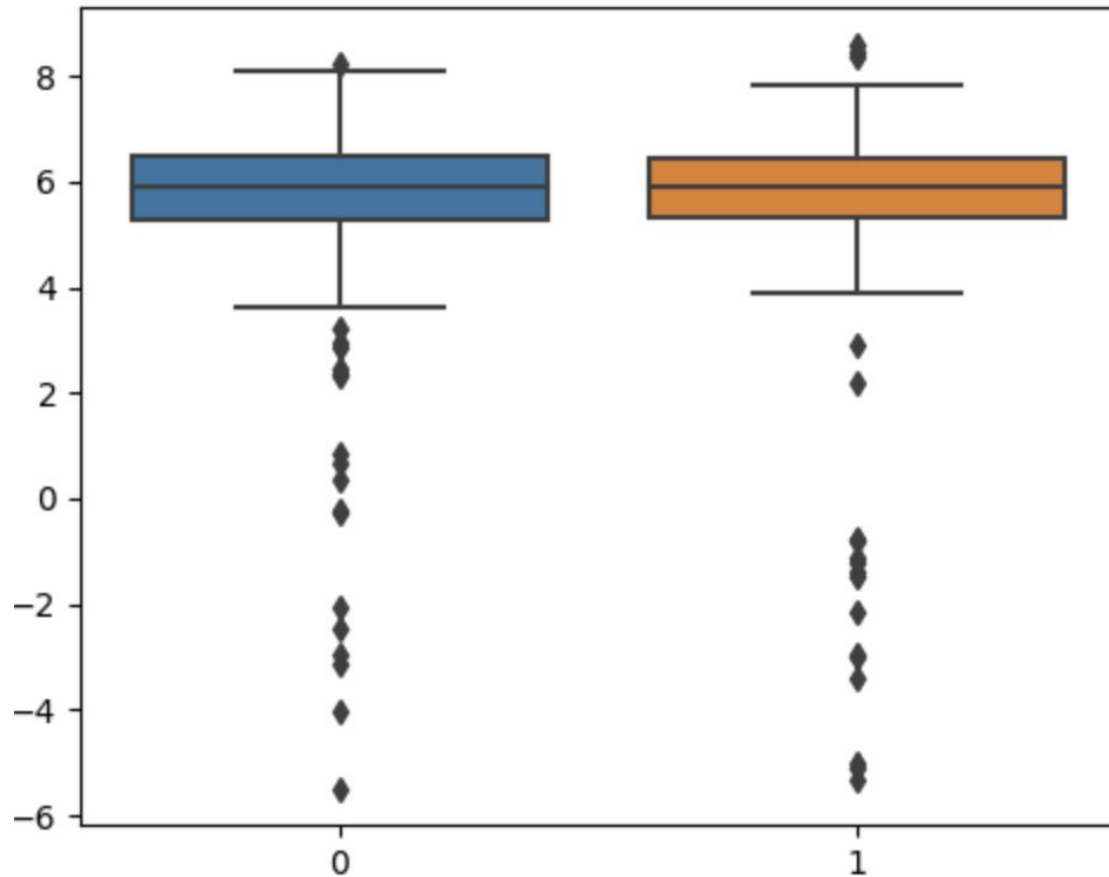
- 1) просто методы Pandas fillna()
- 2) методы sklearn.impute
<https://scikit-learn.org/stable/modules/impute.html#univariate-vs-multivariate-imputation>
- 3) Datawig, Fancyimpute и MissForest и так далее

Что делаем с выбросами?

На практике:

Сначала необходимо обнаружить выброс:

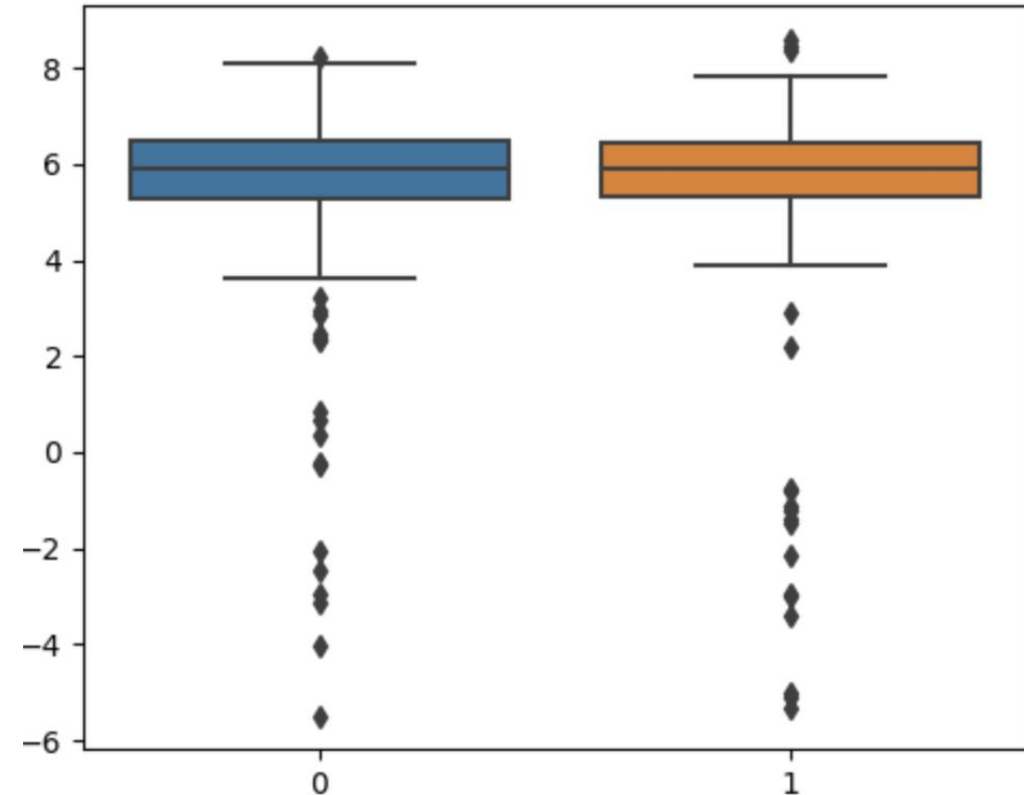
- Выбросы в признаках можно обнаружить, исследуя распределение признаков и в особенности хвосты распределений
- Метод Isolation Forest



Что делаем с выбросами?

Решение:

- заменить значение с выбросом на пропуск, а затем применить методы работы с пропусками
- удалить наблюдения с выбросами (делать ОСТОРОЖНО)
- выделить выбросы в отдельный сегмент и для них построить свою модель (если достаточно данных)



Что делаем с категориальными переменными?

Самые популярные методы:

- one-hot-encoding (добавляем признаки бинарные признаки)
- mean target encoding (заменяем каждую категорию на среднее значение целевой переменной по всем объектам этой категории)
- биннинг - создание групп, а дальше one-hot-encoding или mean target encoding

ID магазина	месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
1	в центре	1000	1	1	1000000
2	на окраине	1569	2	0	200000
3	за пределами города	870	1	0	300000
4	в 10 км от центра города	2000	2	0	500000
5	за пределами города	900	1	1	600000
6	в 10 км от центра города	850	1	1	1000000
7	на окраине	1700	2	1	200000
8	в центре	1400	2	1	300000
9	в центре	1300	2	0	500000
10	в 10 км от центра города	700	1	0	600000

Что делаем с категориальными переменными?

one-hot-encoding

Сначала определяем сколько групп и что за группы в данных
(всего 4 значения): в 10 км от центра города, в центре, за
пределами города, на окраине

	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	на окраине
в 10 км от центра города	1	0	0	0
в центре	0	1	0	0
за пределами города	0	0	1	0
на окраине	0	0	0	1

Что делаем с категориальными переменными?

one-hot-encoding

Сначала определяем сколько групп и что за группы в данных
(всего 4 значения): в 10 км от центра города, в центре, за
пределами города, на окраине

	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	на окраине
в 10 км от центра города	1	0	0	0
в центре	0	1	0	0
за пределами города	0	0	1	0
на окраине	0	0	0	1

Что делаем с категориальными переменными?

one-hot-encoding

Сначала определяем сколько групп и что за группы в данных (всего 4 значения): в 10 км от центра города, в центре, за пределами города, на окраине

	bin_в 10 км от центра города	bin_в центре	bin_за пределами города
в 10 км от центра города	1	0	0
в центре	0	1	0
за пределами города	0	0	1
на окраине	0	0	0

Что делаем с категориальными переменными?

one-hot-encoding

И в итоге вместо 1 переменной получаем N-1 переменную, где N - количество значений переменной

ID магазина	месторасположение в городе	bin_в 10 км от центра города	bin_в центре	bin_за пределами города	площадь	количество этажей	в ТЦ?	доход от магазина
1	в центре	0	1	0	1000	1	1	1000000
2	на окраине	0	0	0	1569	2	0	200000
3	за пределами города	0	0	1	870	1	0	300000
4	в 10 км от центра города	1	0	0	2000	2	0	500000
5	за пределами города	0	0	1	900	1	1	600000
6	в 10 км от центра города	1	0	0	850	1	1	1000000
7	на окраине	0	0	0	1700	2	1	200000
8	в центре	0	1	0	1400	2	1	300000
9	в центре	0	1	0	1300	2	0	500000
10	в 10 км от центра города	1	0	0	700	1	0	600000

Что делаем с категориальными переменными?

one-hot-encoding

И в итоге вместо 1 переменной получаем $N-1$ переменную, где N - количество значений переменной

Что будет, если у переменной очень много значений, например 100? мы создадим вместо 1 переменной 99?

А если в наших данных таких категориальных переменных несколько?

Что делаем с категориальными переменными?

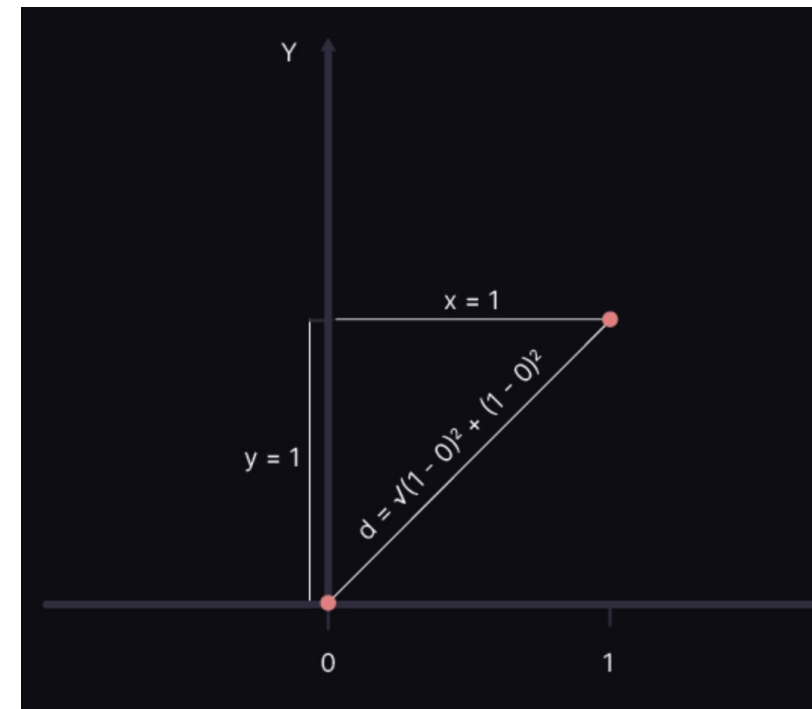
Термин «**проклятие размерности**» в 1961 году ввел американский математик Ричард Беллман.

Предположим, у нас есть две точки на прямой, 0 и 1. Эти две точки находятся на расстоянии друг от друга =1

Теперь мы вводим вторую ось Y – второе измерение. Положение точек определяется теперь списком из двух чисел – (0,0) и (1,1). Расстояние между точками теперь подсчитывается с помощью Евклидова расстояния и оно равно 1.44. В трехмерном пространстве будет 1.73

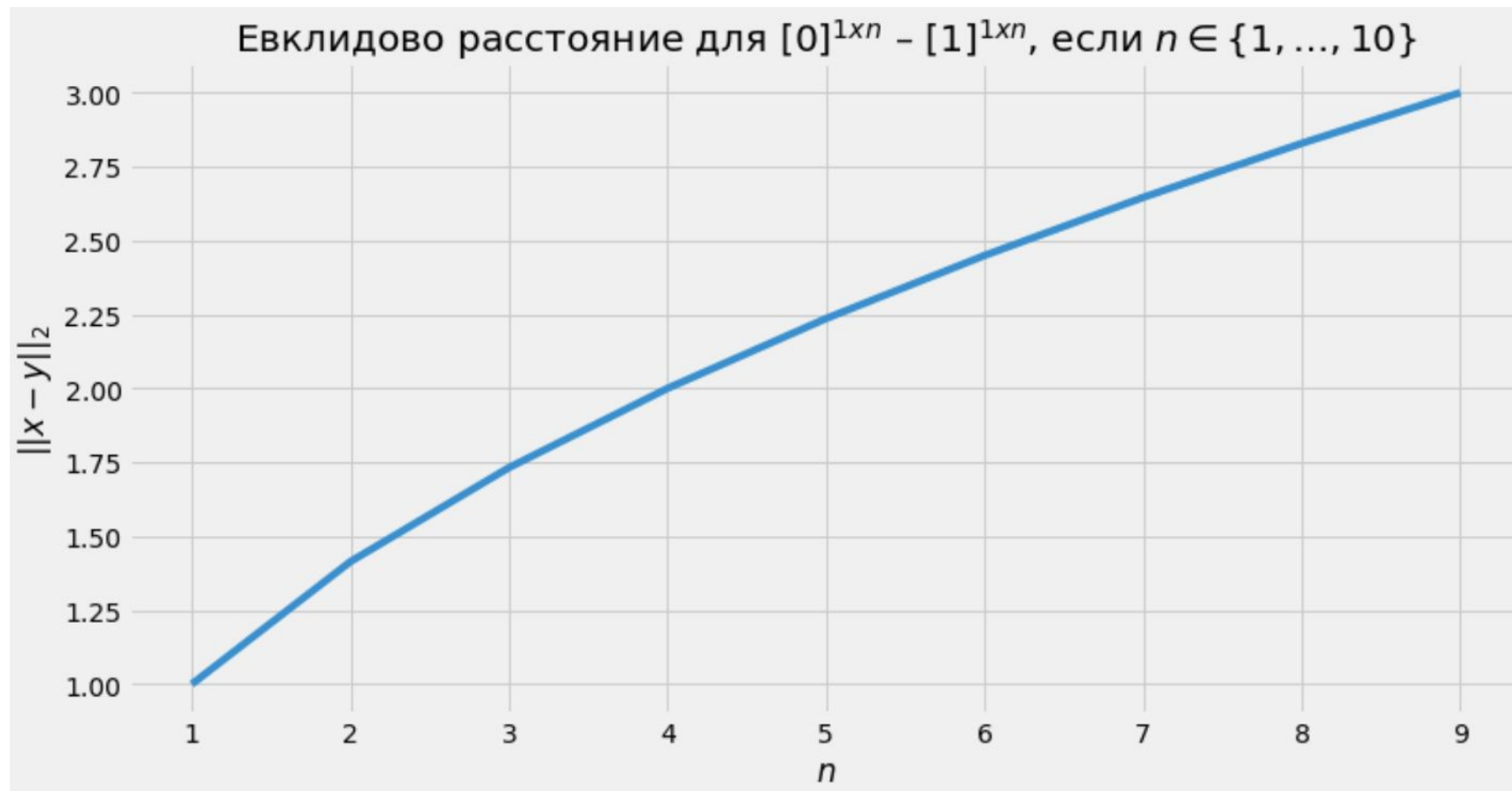


можно почитать - <https://www.helenkapatsa.ru/prokliatiie-razmiernostiei>



Что делаем с категориальными переменными?

Термин «**проклятие размерности**» в 1961 году ввел американский математик Ричард Беллман.



Что же делать?

Не использовать категориальные переменные? :((((

Что делаем с категориальными переменными?

mean target encoding (заменим каждую категорию на среднее значение целевой переменной по всем объектам этой категории)

месторасположение в городе	AVERAGE of доход от магазина
в 10 км от центра города	700000
в центре	600000
за пределами города	450000
на окраине	200000



месторасположение в городе	mean_месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
в центре	600000	1000	1	1	1000000
на окраине	200000	1569	2	0	200000
за пределами города	450000	870	1	0	300000
в 10 км от центра города	700000	2000	2	0	500000
за пределами города	450000	900	1	1	600000
в 10 км от центра города	700000	850	1	1	1000000
на окраине	200000	1700	2	1	200000
в центре	600000	1400	2	1	300000
в центре	600000	1300	2	0	500000
в 10 км от центра города	700000	700	1	0	600000

Что делаем с категориальными переменными?

месторасположение в городе	mean_месторасположение в городе	площадь	количество этажей	в ТЦ?	доход от магазина
в центре	600000	1000	1	1	1000000
на окраине	200000	1569	2	0	200000
за пределами города	450000	870	1	0	300000
в 10 км от центра города	700000	2000	2	0	500000
за пределами города	450000	900	1	1	600000
в 10 км от центра города	700000	850	1	1	1000000
на окраине	200000	1700	2	1	200000
в центре	600000	1400	2	1	300000
в центре	600000	1300	2	0	500000
в 10 км от центра города	700000	700	1	0	600000

Но и здесь нас ждут неожиданности :)

при обучении моделей mean target encoding необходимо рассчитывать на отложенной выборке, не на всей



УНИВЕРСИТЕТ
ИННОПОЛИС

ВОПРОСЫ И ОТВЕТЫ