



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

МАШИННОЕ ОБУЧЕНИЕ. ЗАДАЧА КЛАССИФИКАЦИИ. МЕТРИКИ КЛАССИФИКАЦИИ.

Теванян Элен
24.05.2019

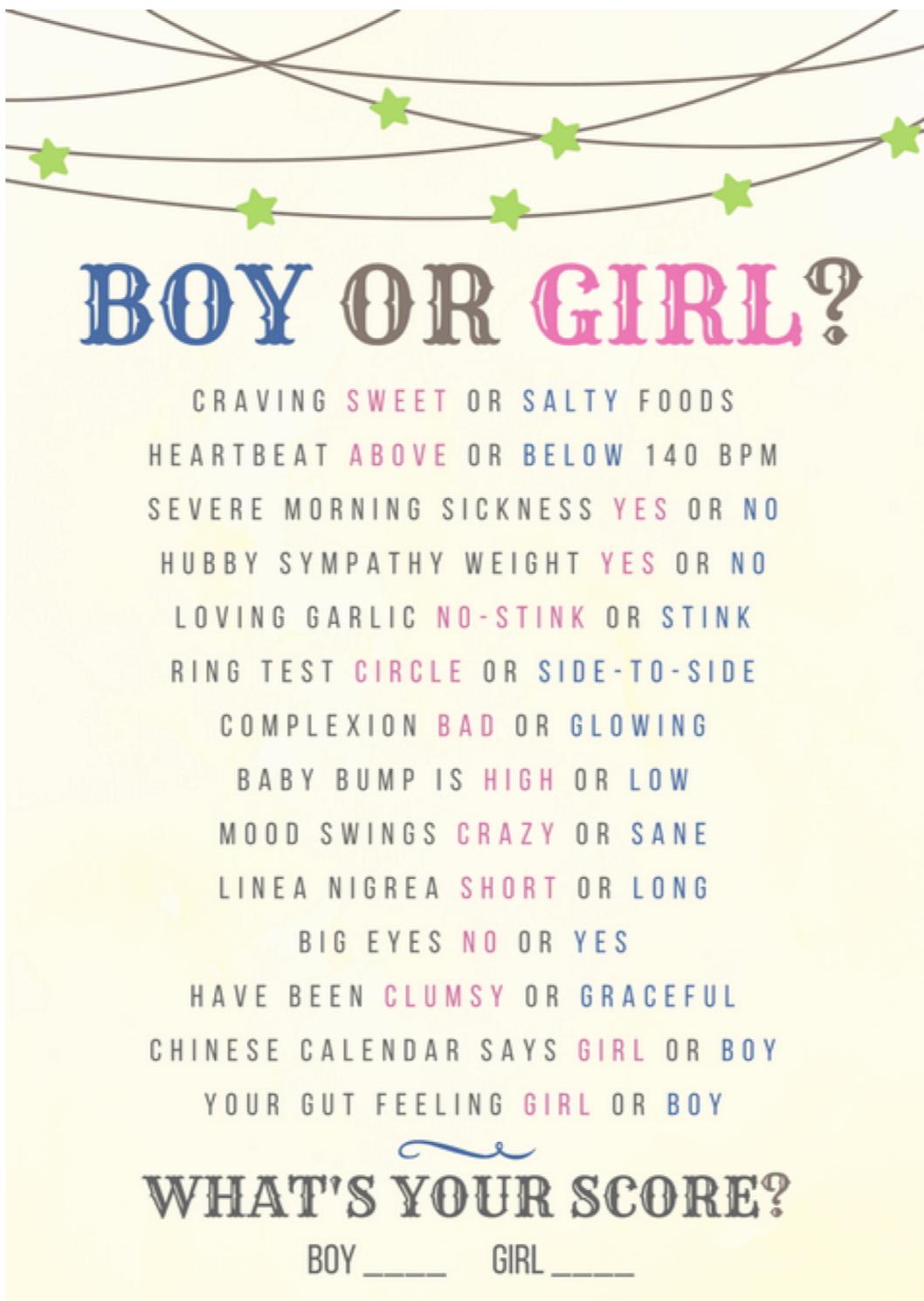
Москва 2019

КЛАССИФИКАЦИЯ.

Классификация – задача восстановления алгоритма, по которому объекту можно присваивать класс

- Y – это разметка объектов по классам.
- X_1, \dots, X_n – это признаковое описание объектов. Построение модели – выделение зависимостей между признаками и классом объектов.

ПРЕДСКАЗАНИЕ ПОЛА



ОПРЕДЕЛЕНИЕ ТИПА ДОКУМЕНТА



ОПРЕДЕЛЕНИЕ ЯЗЫКА ТЕКСТА

The screenshot shows a user interface for language detection. At the top, there are three tabs: "РУССКИЙ (ОПРЕДЕЛЕН АВТОМАТИЧЕСКИ)" (Russian (Automatically Detected)), "ИСПАНСКИЙ" (Spanish), and "РУССКИЙ" (Russian) with a dropdown arrow. Below the tabs, the text "день сегодня начался не очень" is displayed in Russian, with a green circular icon containing a white letter 'G' to its right. An "X" button is located to the right of the text. Underneath the Russian text, its Spanish translation "den' segodnya nachalsya ne ochen'" is shown. At the bottom left are microphone and speaker icons. On the bottom right, it says "29/5000" and has a small edit icon.

АНАЛИЗ ТОНАЛЬНОСТИ ОТЗЫВОВ



koshka_Stasya
[рекомендации \(12\)](#) | [оценки](#) | [друзья](#) | [фильмы](#)

[Показать историю или заинтересовать ее](#)

26 января 2017 | 01:15

В последнее время выходит все больше и больше фильмов и сериалов на историческую тематику. Этот сериал, как и большинство других, можно бесконечно критиковать за неточности в истории, за то, что в кадре присутствуют детали, несоответствующие показанному времени и так далее.

Не стану спорить с теми, кто негодует по поводу исторической составляющей данного сериала. Впрочем, если сравнить историю, скажем, с теми же Тюдорами, в данном сценарии история пострадала явно в разы меньше.

Сериал **Медичи** — это не документальный фильм о великой семье, а художественный. Нужно признать, что приоритет здесь все-таки не 100% историческая достоверность, а хороший сценарий, актеры и красавая картинка, а с этой точки зрения мы получили абсолютно все.

Единственный актер, который у меня почему-то вызывает некоторый резонанс — это **Ричард Мэдден**. Нет, я вовсе не хочу сказать, что он плохо играет. Я хочу сказать, что во время просмотра сериала не могла отделаться от мысли, что эту роль должен был исполнить более зрелый, взрослый актер, потому что даже состаренный Мэдден кажется слишком юным для сурового правителя Флоренции.

P.S.: Считаю, что исторический сериал можно считать хорошим не тогда, когда он полностью передал нам всю историю, а тогда, когда он смог заинтересовать нас и заставить перерыть весь интернет, чтобы узнать о том, как было на самом деле и что создателям сериала пришлось выпустить, чтобы уложиться в формат. **Медичи** вызвали у меня желание прочитать все, что только можно найти на просторах интернета об этой, без сомнения, великой семье.

[прямая ссылка](#)  + [комментарий](#) [спойлер?](#)  [Полезная рецензия?](#) Да / Нет 11 / 1



Alexander Dokiyen
[рекомендации \(46\)](#) | [оценки](#) | [друзья](#) | [фильмы](#) | [звезды](#)

7 ноября 2016 | 00:13

Рецензия пилотной серии

Пилотная серия не богата на события, с которых начинается повествование. **Глава семьи Медичи** отравлен у себя в загородной усадьбе. Двое его сыновей обещают восстановить справедливость и отомстить за отца. Первый подозреваемый — глава конкурирующего семейства. Но **остальная хронометраж — о событиях двадцатилетней давности**, которые немного объясняют отношения между братьями и отцом Медичи и рассказывают об общем путешествии семьи в Рим.

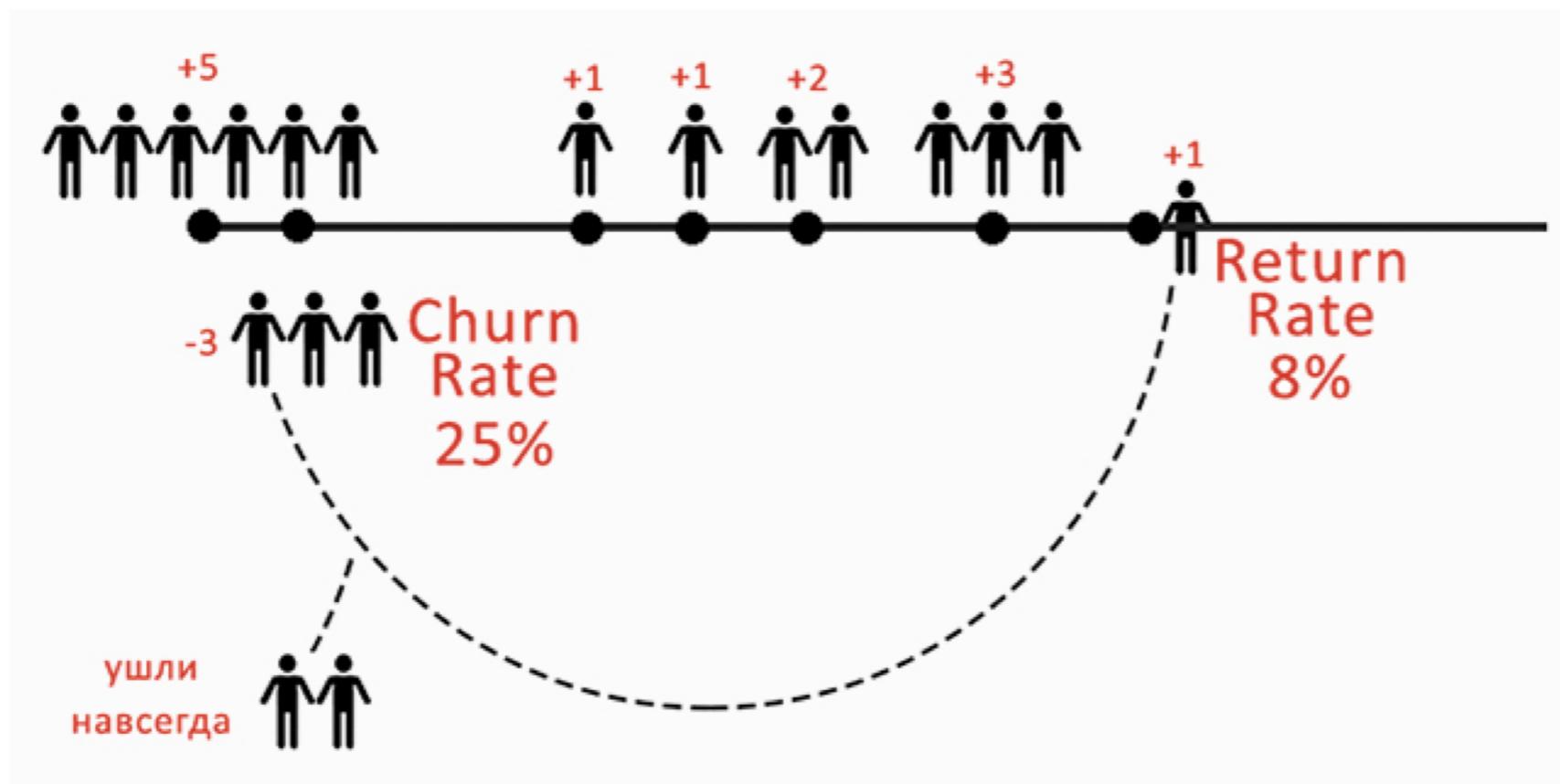
Своим антуражем сериал напоминает «Демоны да Винчи». Примерно та же эпоха, та же Италия. Но, увы, у маститого режиссера **Серджо Мимики-Газзана** получилась **история слишком неуверенная**. От просмотра создается впечатление, что автор собрал много разных клише, свел вместе и сказал: вот вам и сериал о Италии в период Возрождения. Тут вам и **полно эротических сцен** в эллинском стиле, художник, который по стечению обстоятельств должен стать меркантильным купцом, ну и куча **стандартных пафосных фраз** вложенных в уста актерам, типа «сын мой, после меня кто-то должен возглавить нашу семью».

Актеры. Ну конечно первое, что меня привлекло, это **Ричард Мэдден** на постере, тема Медичи мне тоже не безразлична. Кроме того, я сначала даже не поверил, что в сериале играет **Дастин Хоффман** (хороший грим). Сомневаться в актерских способностях последнего смысла нет, он играет великолепно. Но даже присутствие этих двух звезд не спасает картину от оценки ниже среднего.

5 из 10

[прямая ссылка](#)  + [комментарий](#) [спойлер?](#)  [Полезная рецензия?](#) Да / Нет 17 / 45

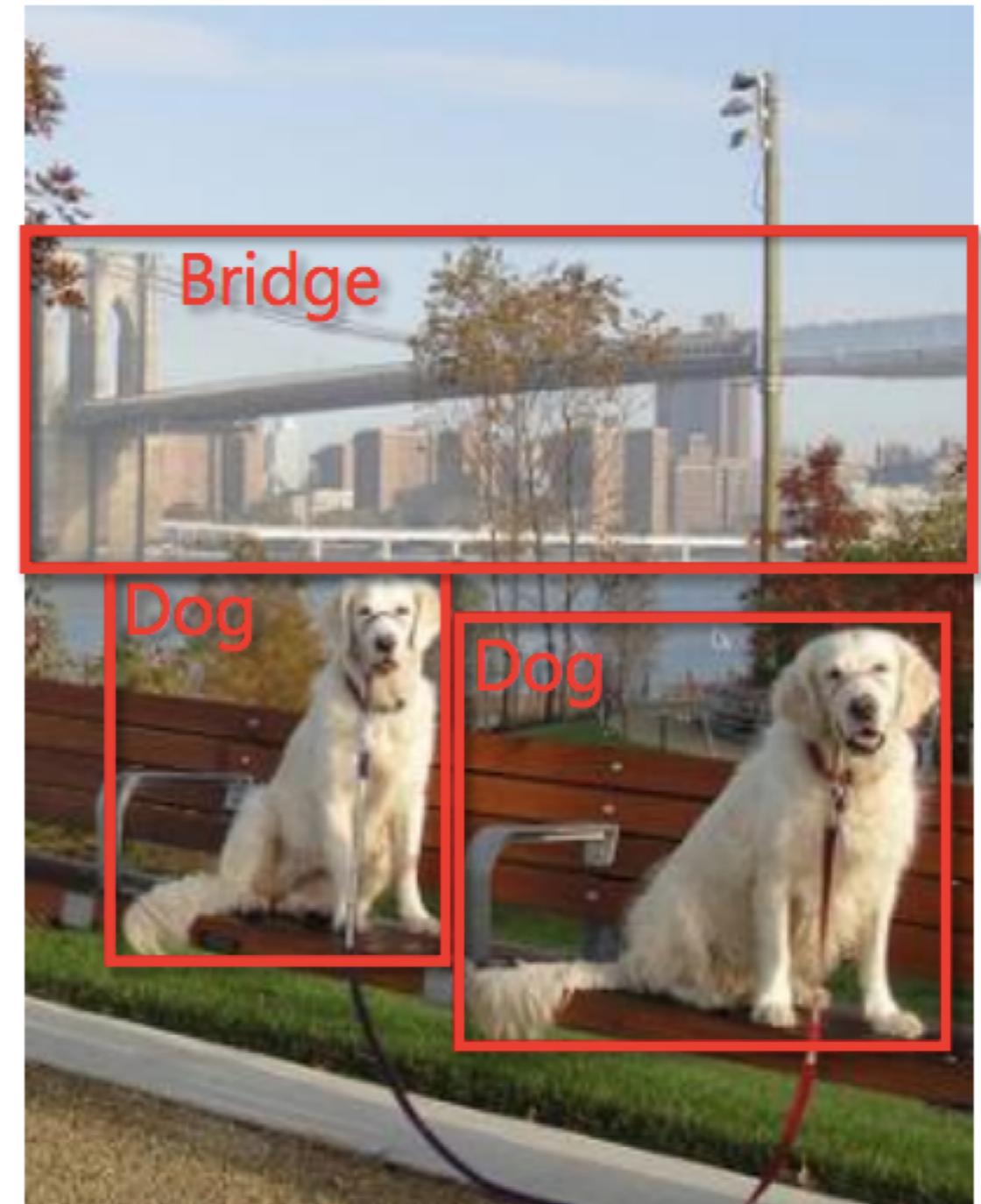
ОТТОК КЛИЕНТОВ



СПАМ/ НЕ СПАМ



РАСПОЗНАВАНИЕ ОБЪЕКТОВ



КЛАССИФИКАЦИЯ. ЗАЧЕМ?

1. Выделить характерные для класса признака
2. Определить вклад каждого признака
3. Объяснить значение признака отношением к классу

КЛАССИФИКАЦИЯ

- Есть обучающая выборка, в которой объекты представлены признаковым описанием и дана целевая переменная – метка класса.
- Метод обучения с учителем – требуются прецеденты (размеченная выборка)
- Задача: найти алгоритм, который каждому нового объекту будет присваивать метку класса.

КЛАССИФИКАЦИЯ

- Классов может быть много.
 - Бинарная классификация – 2 класса
 - Мноклассовая классификация – 3 и более.

ПРЕДЫСТОРИЯ ЗАДАЧИ

- В компании есть программа лояльности. Но не все ей пользуются честно.
- Хотим отлавливать нехороших и блокировать им карты! Иначе начисляем лишние баллы и теряем денежки

ПРИМЕР ЗАДАЧИ КЛАССИФИКАЦИИ

- Задача: клиент \rightarrow [плохой (1)/ хороший(0)]
 - x_i – объект, для которого строим предсказания (клиент)
 - y_i – целевая переменная: 1 или 0
 - (x_i, y_i) – прецедент
 - Обучающая выборка – все клиенты, кому предложили тестовую партию продукта.
-
- Обучили классификатор $a(x)$. Хороший? Плохой? Поможет принять решение?

ПРИМЕР ЗАДАЧИ КЛАССИФИКАЦИИ

- Учим модельку на 100 клиентах (для примера)
- В таблице приведены результаты обучения модели машинного обучения

	Клиент плохой $y = 1$	Клиент хороший $y = 0$
Модель считает, что клиент плохой $a(x) = 1$	60 True Positive (TP)	5 False Positive(FP)
Модель считает, что клиент хороший $a(x) = 0$	25 False Negative (FN)	10 True Negative (TN)

ДОЛЯ ПРАВИЛЬНЫХ ОТВЕТОВ

- Доля правильных ответов (accuracy) показывает долю объектов в выборке, которым классификатор присвоил их истинный класс.

$$\text{Accuracy} = \frac{\sum_{i=1}^{\ell} [a(x_i) = y_i]}{n}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

ДОЛЯ ПРАВИЛЬНЫХ ОТВЕТОВ

- Доля правильных ответов (accuracy) показывает долю объектов в выборке, которым классификатор присвоил их истинный класс.

$$\text{Accuracy} = \frac{\sum_{i=1}^{\ell} [a(x_i) = y_i]}{n}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

$$\text{Accuracy} = \frac{60+10}{100} = 0.7$$

ТОЧНОСТЬ

- Точность (precision) показывает уровень доверия к классификатору при $a(x) = 1$

$$\text{Precision} = \frac{TP}{TP+FP}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

ТОЧНОСТЬ

- Точность (precision) показывает уровень доверия к классификатору при $a(x) = 1$

$$\text{Precision} = \frac{TP}{TP+FP}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

$$P = \frac{60}{60+5} = 0.92$$

- Отвечаем на вопрос: можно ли доверять классификатору при $a(x) = 1$

ПОЛНОТА

- Полнота (recall) показывает, какую долю правильно-положительных ответов находит классификатор

$$\text{Recall} = \frac{TP}{TP+FN}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

ПОЛНОТА

- Полнота (recall) показывает, какую долю правильно-положительных ответов находит классификатор

$$\text{Recall} = \frac{TP}{TP+FN}$$

	$y = 1$	$y = 0$
$a(x) = 1$	60	5
$a(x) = 0$	25	10

$$P = \frac{60}{60+25} = 0.71$$

- Отвечаю на вопрос: много ли положительных ответов находит классификатор?

ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ

- Будем жить в мире, где два класса: +1 и -1
 - t – порог
 - $b(x)$ – оценка принадлежности к классу +1
-
- Классифицируем так:

$$a(x) = [b(x) > t]$$

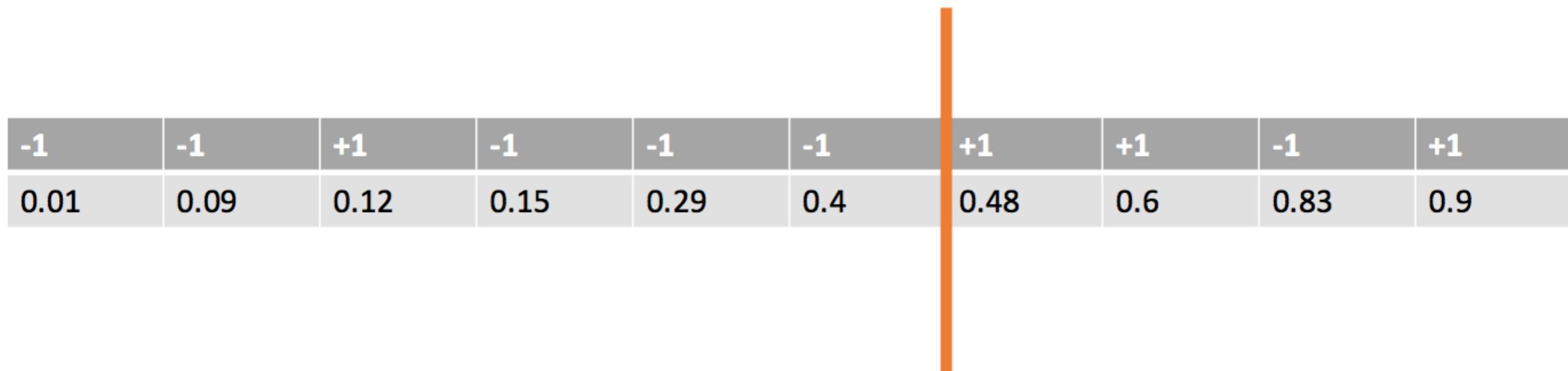
ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ

- Как оценить качество $b(x)$?
- Как выбрать порог?

ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ

- Как оценить качество $b(x)$?
- Как выбрать порог?
- Порог выбирается позже
 - Порог высокий:
 - Мало объектов к классу +1
 - Точность выше
 - Полнота ниже
 - Порог низкий:
 - Много объектов относим к +1
 - Точность ниже
 - Полнота выше

ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ



ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ

-1	-1	+1	-1	-1	-1	+1	+1	-1	+1
0.01	0.09	0.12	0.15	0.29	0.4	0.48	0.6	0.83	0.9

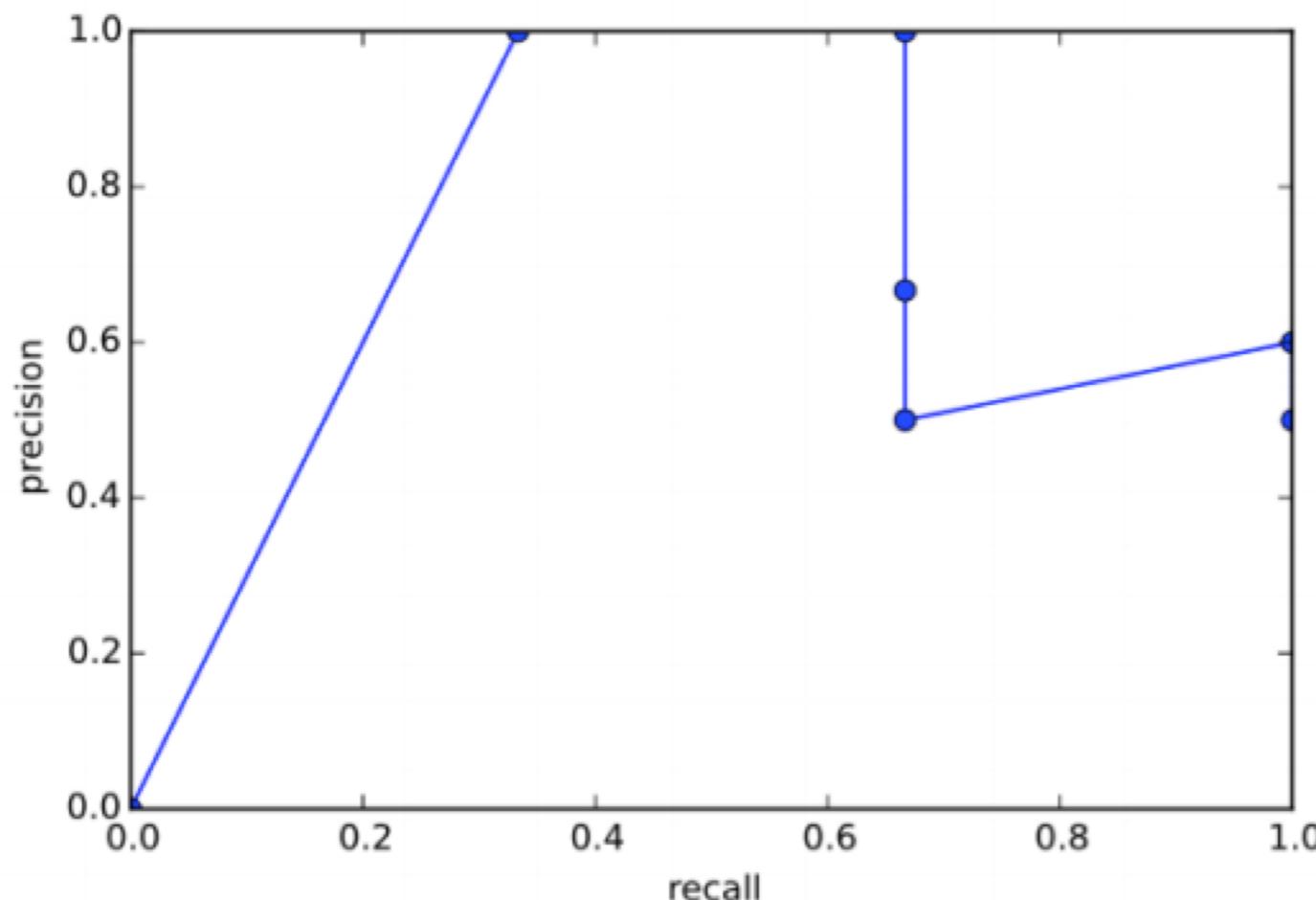


ОЦЕНКА ПРИНАДЛЕЖНОСТИ К КЛАССУ

- Пример: кредитный scoring
- $b(x)$ – оценка вероятности возврата кредита
- $a(x) = [b(x) > 0.5]$
- precision = 0.1
- recall = 0.7
- ??????

PR-КРИВАЯ

- Кривая точности-полноты
- Ось X – полнота
- Ось Y – точность
- Точки значения полноты и точности при последовательных порогах



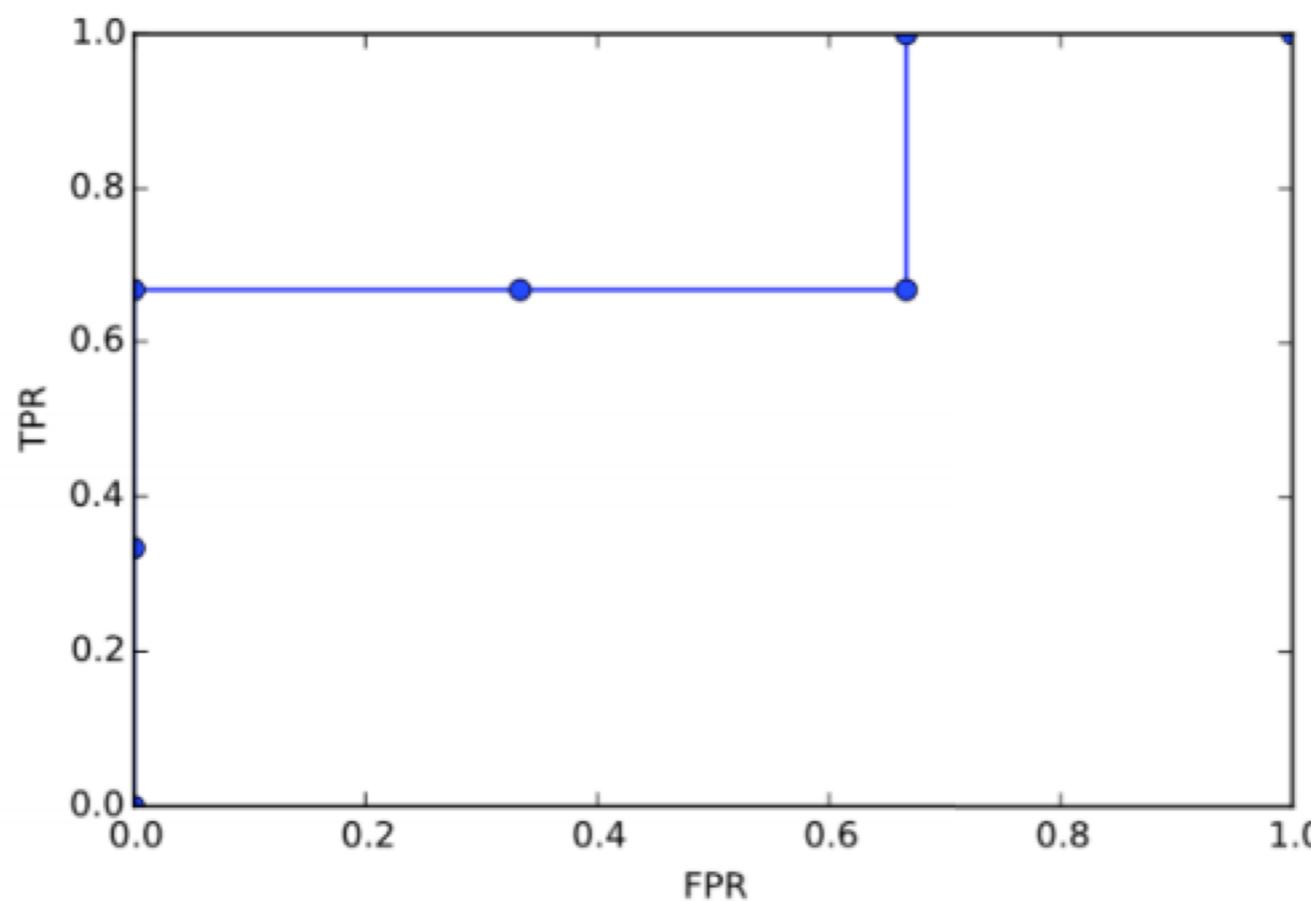
ROC- КРИВАЯ

- Receiver Operating Characteristic
- Ось X – False Positive Rate

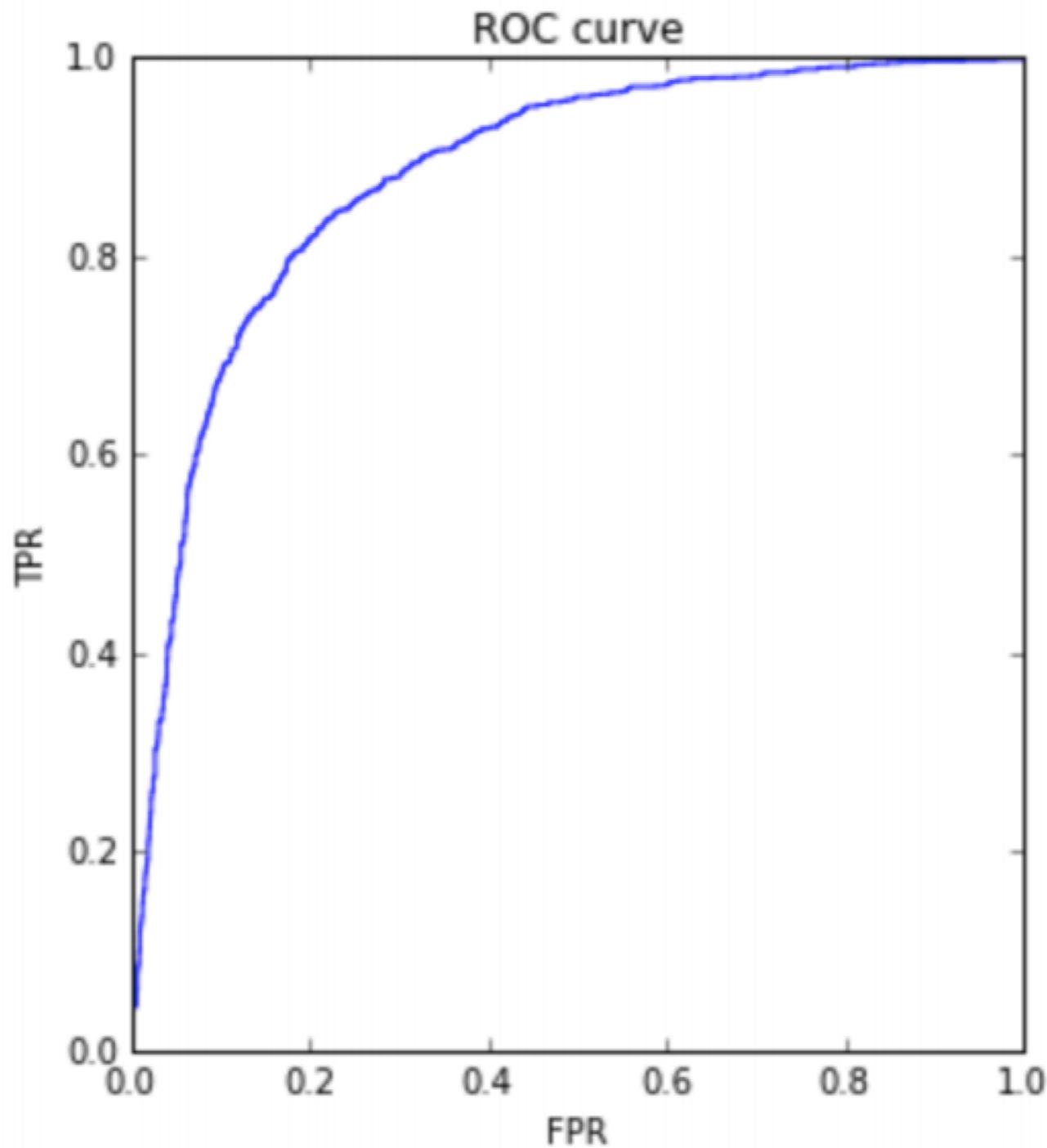
$$FPR = \frac{FP}{FP+TN}$$

- Ось Y – True Positive Rate

$$TPR = \frac{TP}{TP+TN}$$

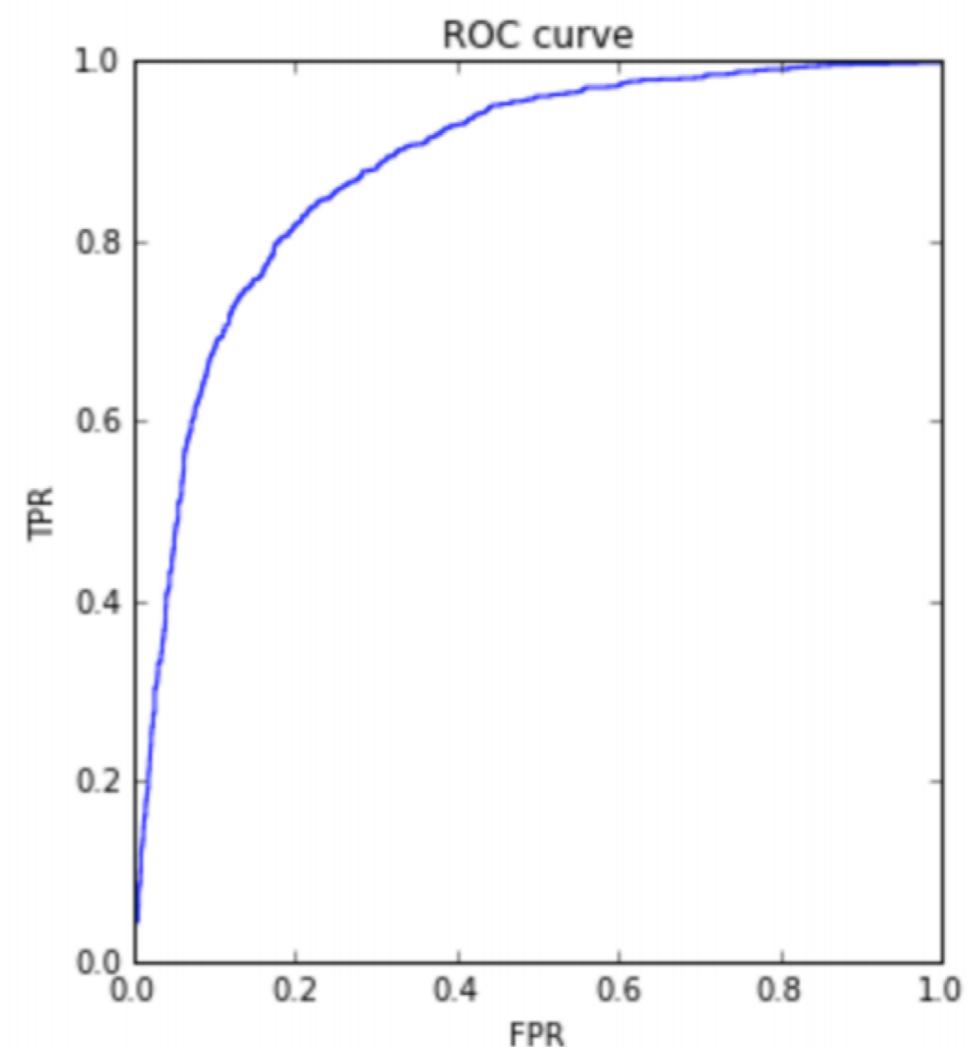


ROC- КРИВАЯ В РЕАЛЬНОСТИ



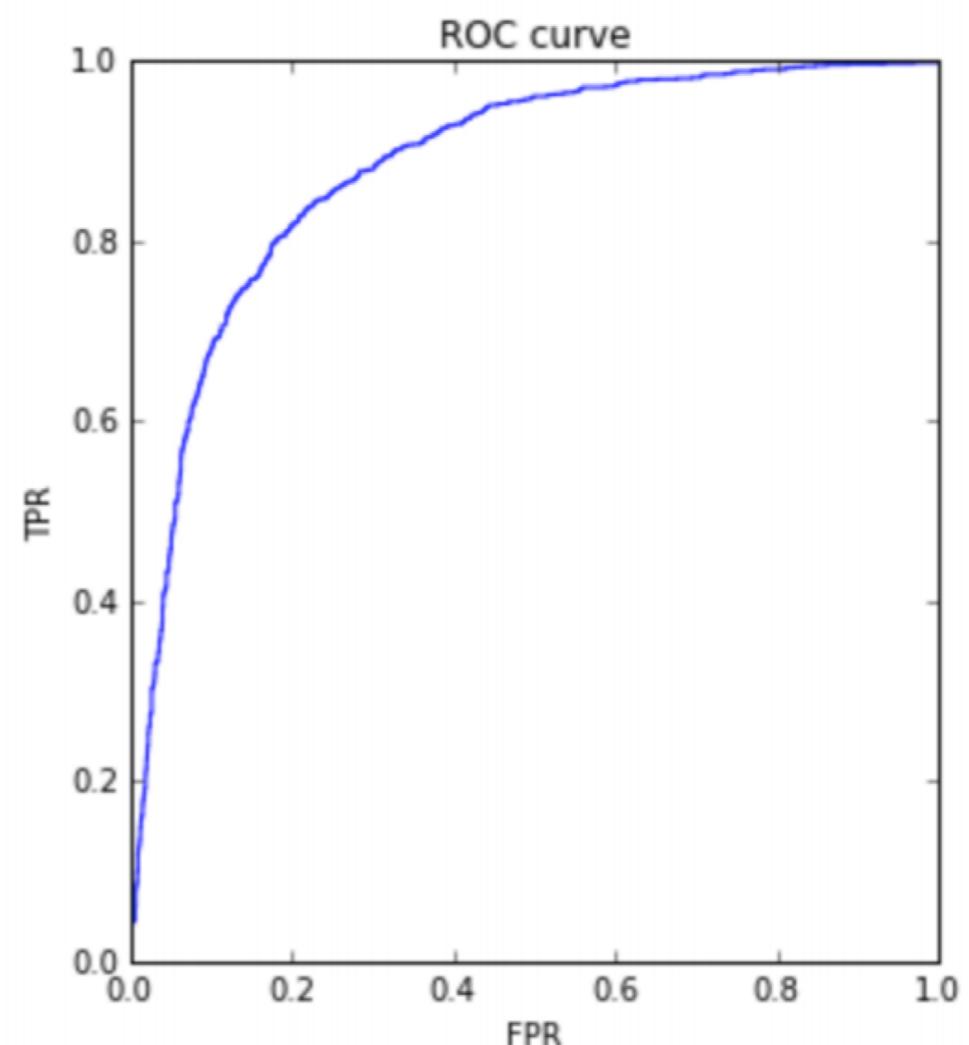
ROC-КРИВАЯ

- Левая точка: (0, 0)
 - Правая: (1,1)
 - Идеально:
захватили точку (0, 1)
-
- Считаем еще AUC-ROC –
площадь под кривой



ROC-КРИВАЯ

- Левая точка: (0, 0)
 - Правая: (1,1)
 - Идеально:
захватили точку (0, 1)
-
- Считаем еще AUC-ROC –
площадь под кривой
 - Идеально: AUC-ROC = 1
 - Плохо: AUC-ROC = 0.5





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ