Ликбез по ML метрикам и их связи с бизнес-метриками

Pavel.Filonov@kaspersky.com
Data Science Manager

kaspersky

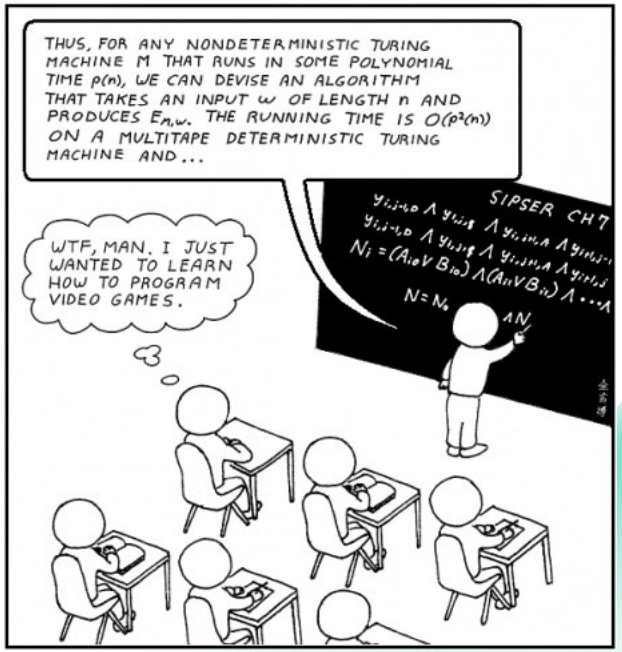
Образование $(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$

В начале карьеры

Недавно

Сейчас

План

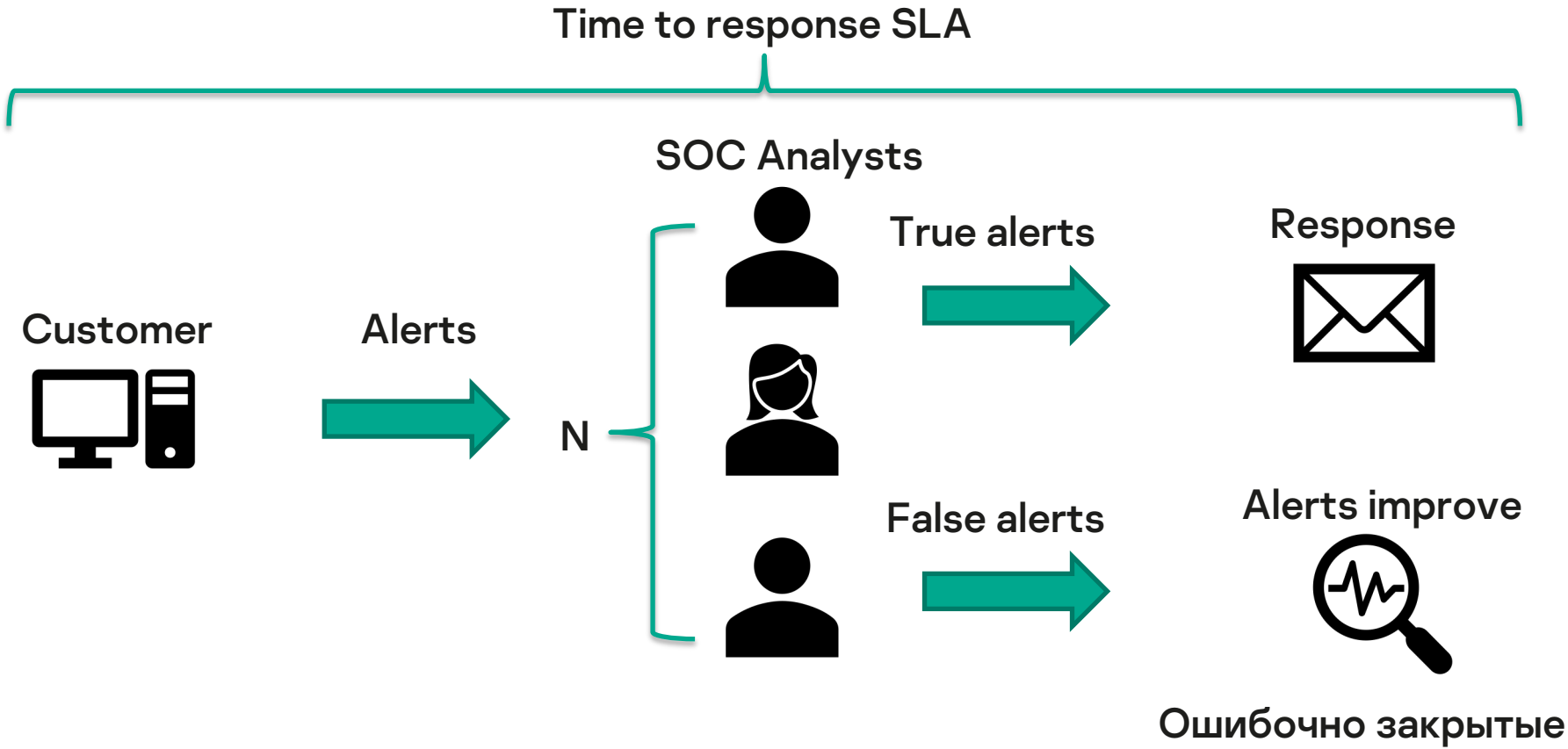


Пример из практики

Бизнес метрики

DS метрики

Как связать


Time to response SLA
SOC Analysts
Customer
Alerts
N
True alerts
Response
False alerts
Alerts improve
Ошибочно закрытые

Гипотеза Alpha

Мы предполагаем, что приоритизация оповещений позволит находить true alerts раньше и это уменьшит среднее время реакции на ? %.

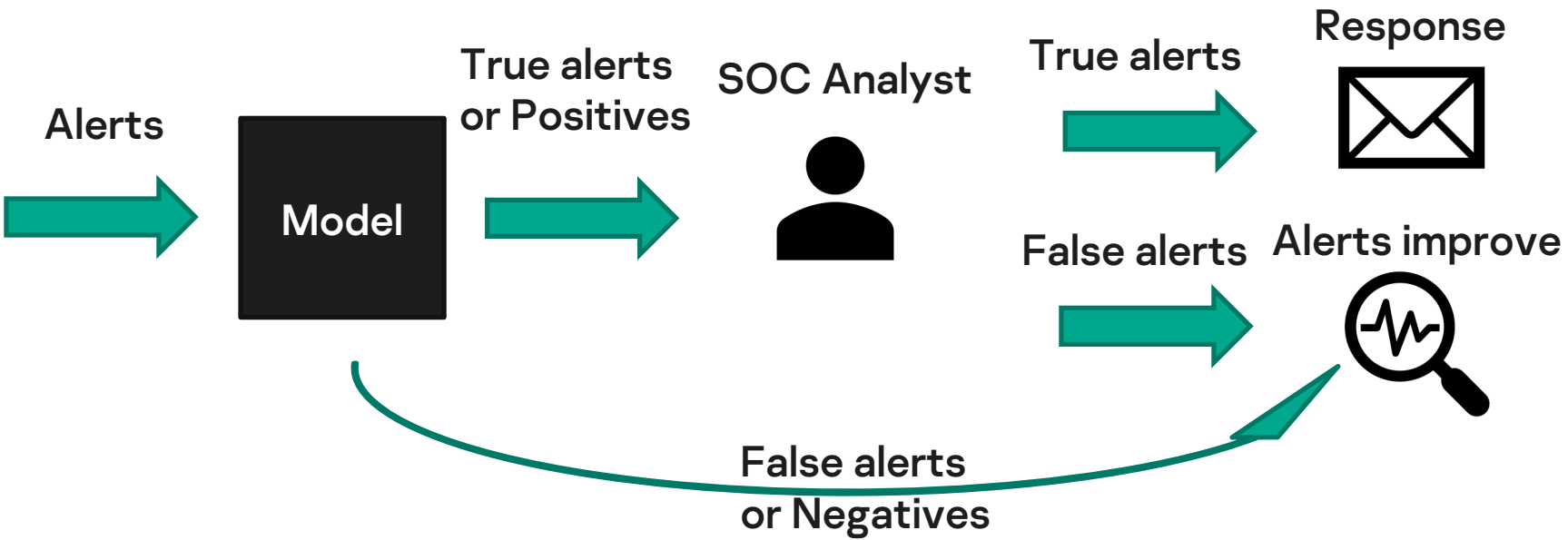
Гипотеза Beta

Подсказка в интерфейсе аналитика позволит снизить время на разбор ложных инцидентов, что повысит пропускную способность на 10%.

Гипотеза Charlie

Мы предполагаем, что автоматическая фильтрация false alerts снизит нагрузку на SOC аналитиков и повысит их пропускную способность на 20%. При этом доля ошибок не превысит 2%

Схема реализации



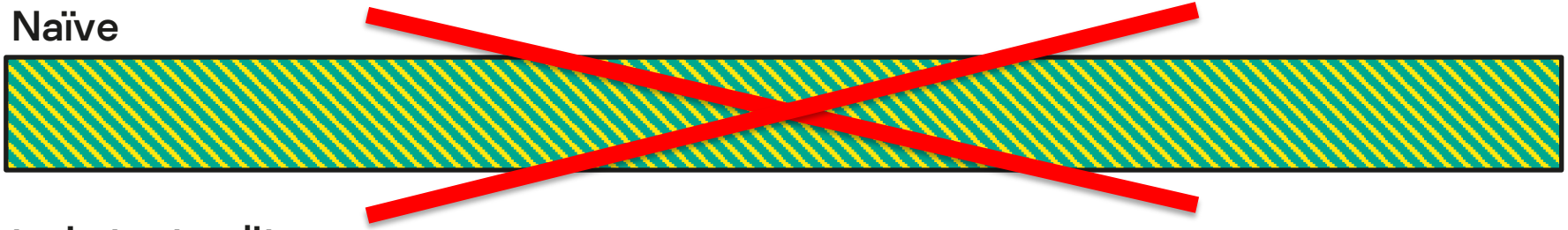
На чем считать метрики

train set

test set

validation set

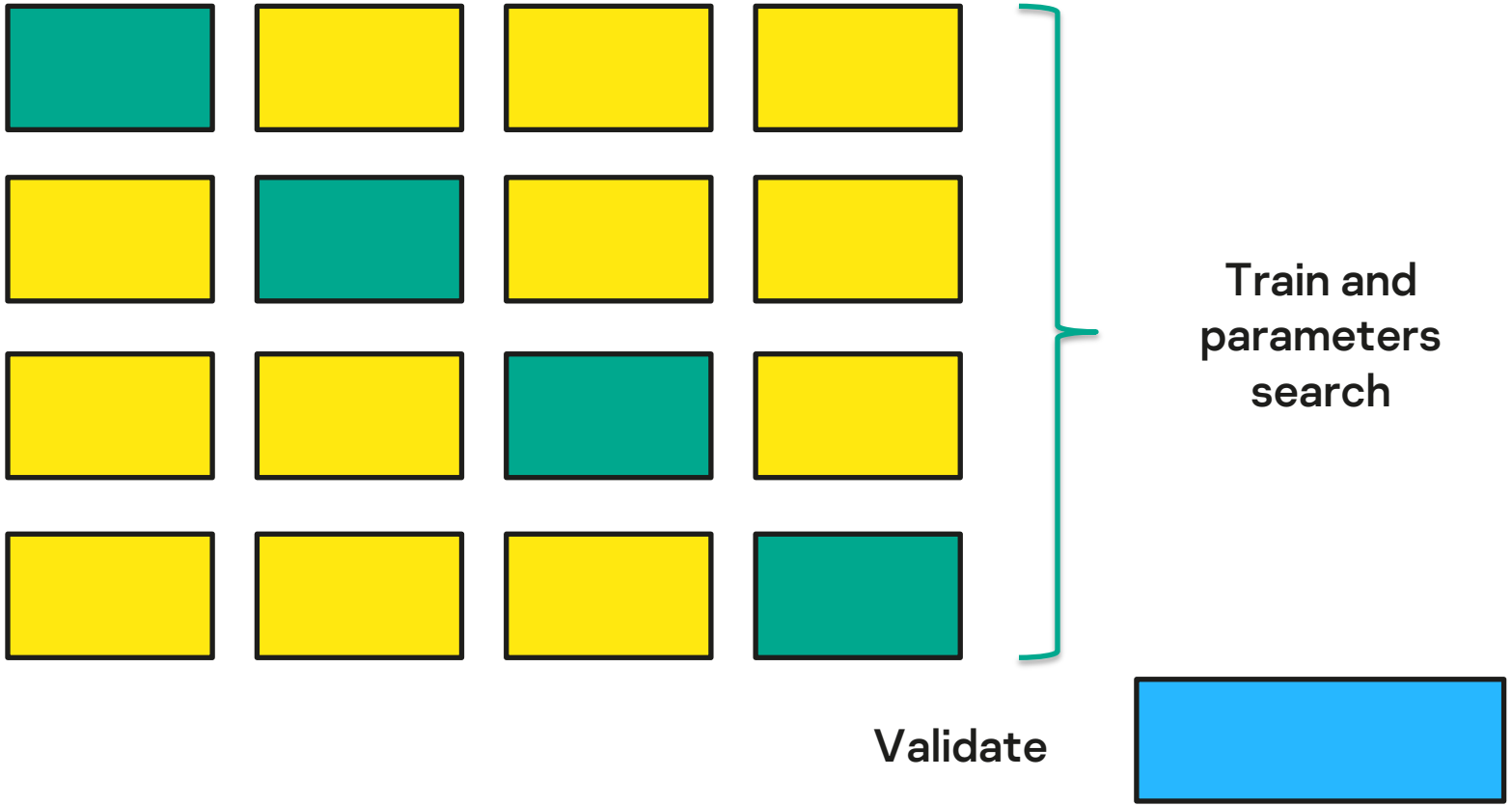
Naïve



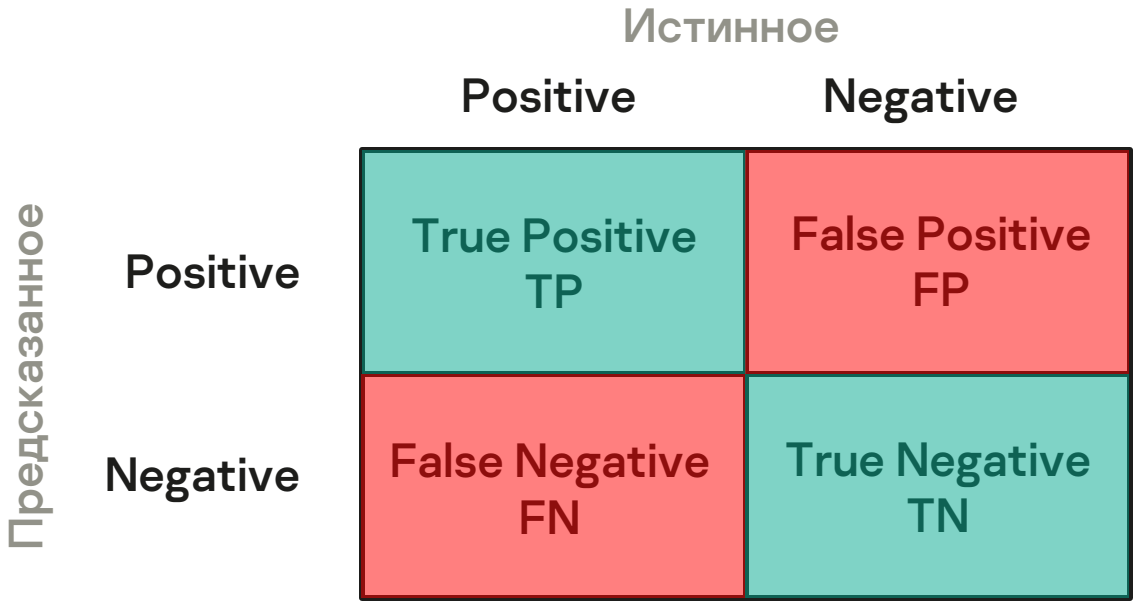
train test split

train test validate split

cross validation



Матрица ошибок



$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Предположим, что в данных 90% - Positive и 10% Negative

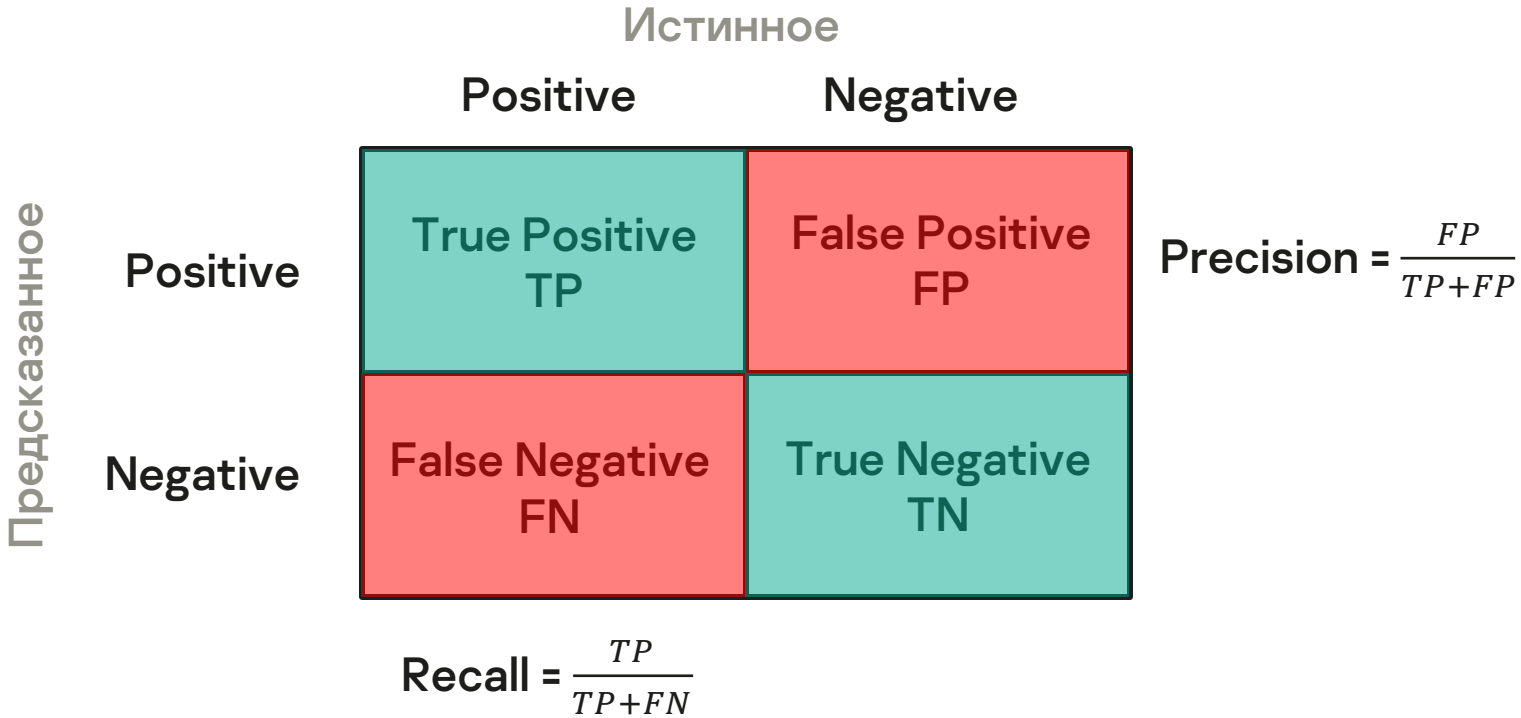
В качестве наивного классификатора выберем такой, который всегда предсказывает Positive

Предсказанное \ Истинное	Positive	Negative
Positive	90	10
Negative	0	0

Accuracy = $\frac{90}{100} = 0.9$

Вывод: accuracy не стоит использовать в случае дисбаланса классов

Матрица ошибок



Пусть в данных на 100 оповещений приходится 99 ложных и все 100 разбирались в ручную.

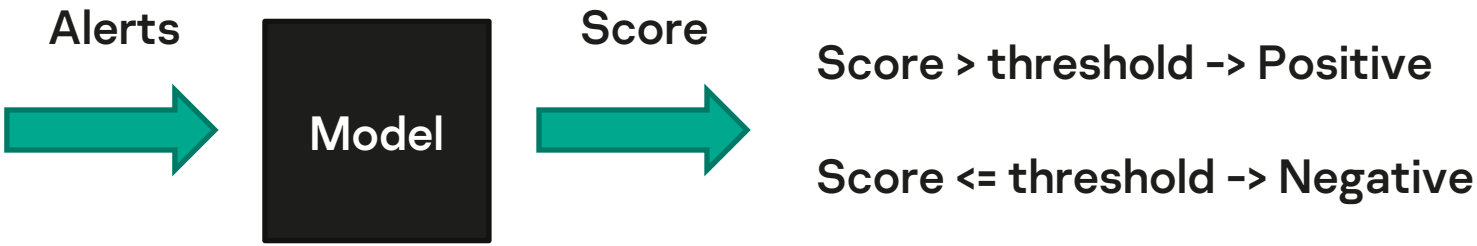
Точность модели – 70%.

Это означает, что из 100 оповещений, отправленных на разбор SOC аналитикам после фильтрации, 70 будут ими помечены как истинные, а 30 как ложные.

Полнота модели – 90%.

Это означает, что из 100 настоящих инцидентов 10 будут ошибочно отфильтрованы моделью.

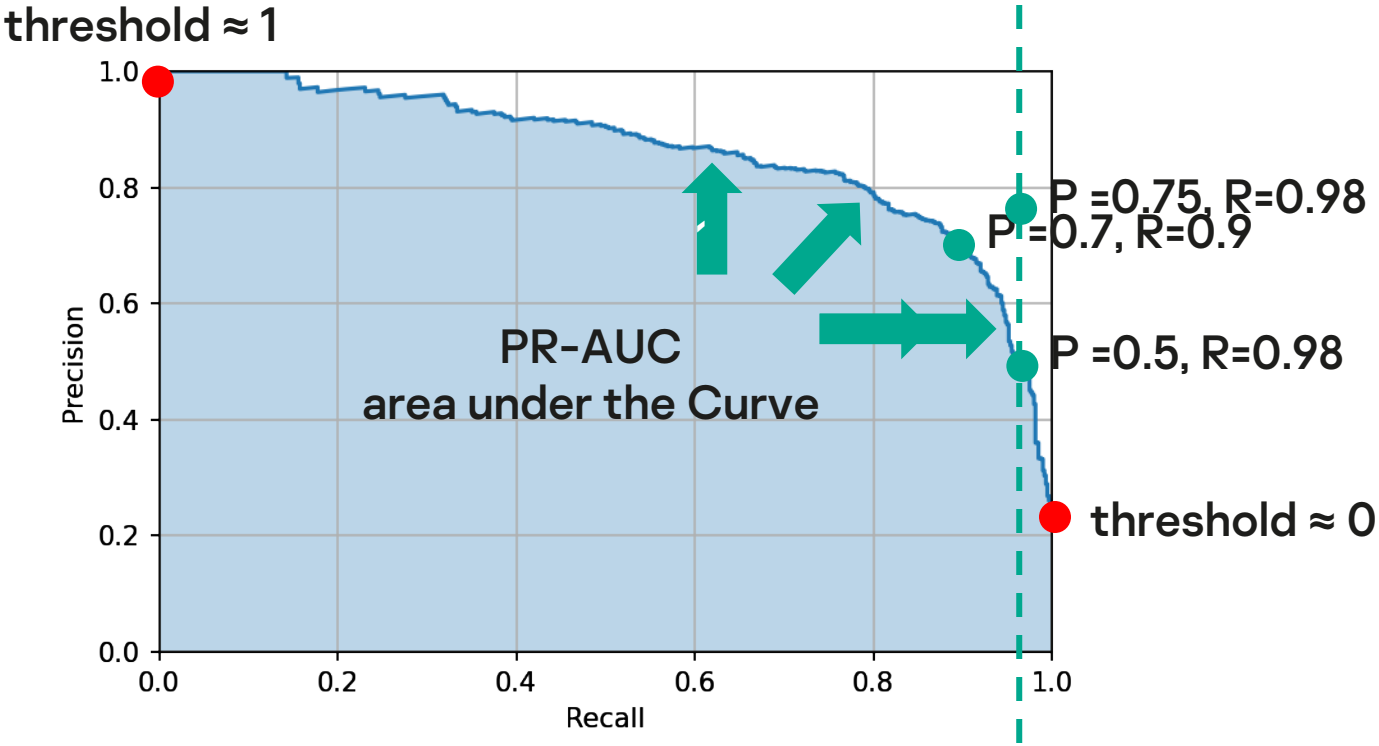
Score – обычно число от 0 до 1.
Можно трактовать как уверенность модели в позитивном предсказании

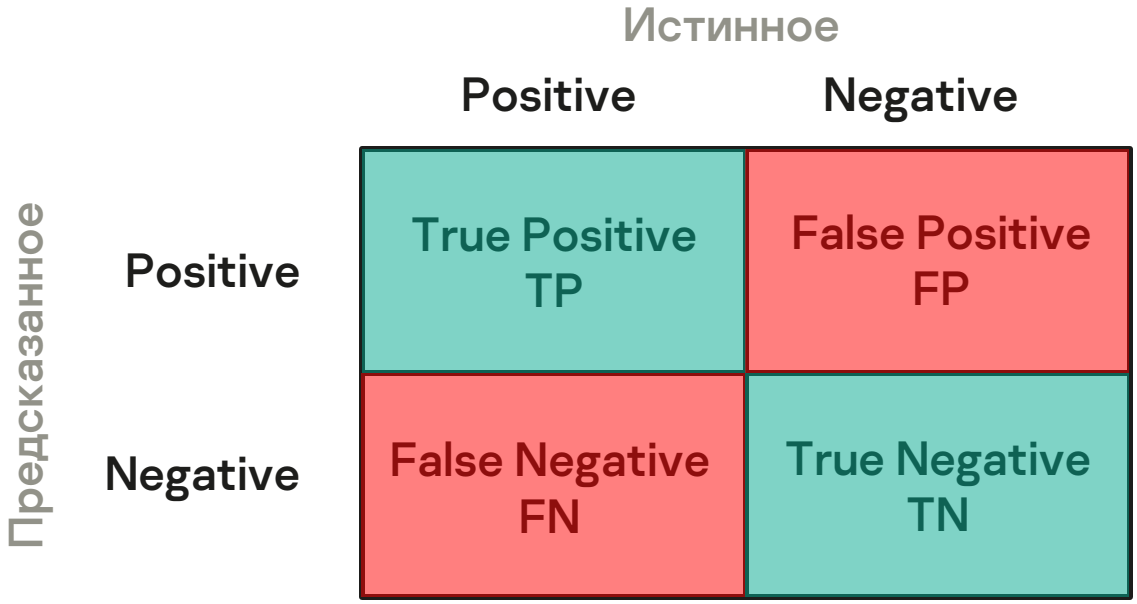


Если threshold -> 1, то Recall -> 0, а Precision -> 1

Если threshold -> 0, то Recall -> 1, а Precision -> min

PR-кривая


threshold ≈ 1
P =0.75, R=0.98
P =0.7, R=0.9
P =0.5, R=0.98
threshold ≈ 0
PR-AUC
area under the Curve
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0



True Positive Rate

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

False Positive Rate

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

True Negative Rate

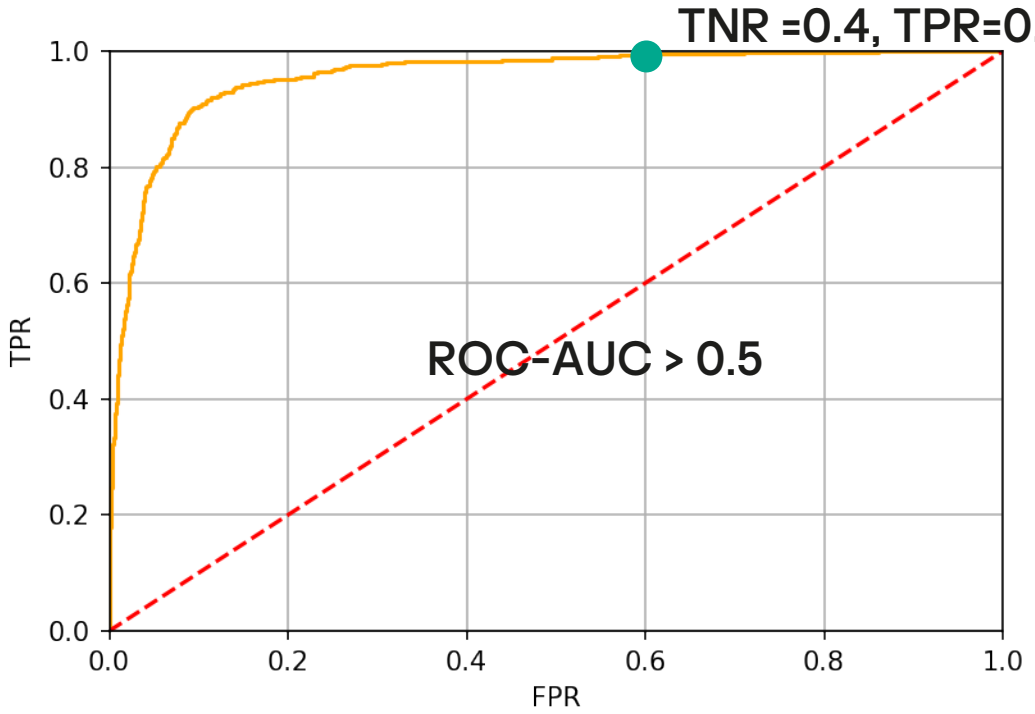
$$TNR = \frac{TN}{FP+TN} = 1\text{-}FPR$$

FPR модели 0.6:

60% всех негативных примеров будут распознаны некорректно

Вывод – можно автоматически отфильтровать 40% всех ложных оповещений, что повысит **пропускную способность**

ROC-кривая



Модель может автоматически отфильтровать **40%** ложных оповещений.

При этом доля ошибочно закрытых составит **2%**

Эксперимент показал, что пропускная способность аналитика растет с ростом TNR.

- Бизнес метрики зависят от гипотезы
- DS метрики также зависят от гипотезы
- Оптимизировать стоит DS метрики связанные с гипотезой
- Для гипотезы Charlie
    - TPR полностью ложиться на метрику по доле ошибок
    - TNR можно использовать как прокси-метрику для поиска корреляции с пропускной способностью аналитиков
- После подтверждения корреляции метрик, можно использовать в гипотезах целевые DS-метрики

1. What is a Confusion Matrix? – link
2. Метрики качества классификации – coursera
3. Classification metrics - scikit-learn docs
4. Multilabel ranking metrics - scikit-leran docs
5. Regression metrics - scikit-learn.org docs
6. gcunhase/NLPMetrics: The Natural Language Processing Metrics Python Repository - github.com
7. Biometrics performance - wikipedia.org
8. Evaluation Metrics for Recommender Systems - link

Спасибо за внимание!

 Pavel.Filonov@Kaspersky.com

 @pavel_filonov

If you are looking at this last slide, you are already a hero!

kaspersky