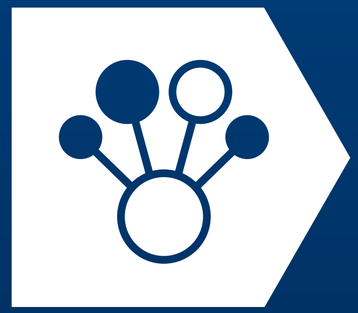


CatBoost

Библиотека Градиентного Бустинга

Градиентный бустинг

- Лучшее решение для разнородных данных
- Легко использовать
- Хорошо работает даже на малых объемах данных

Применения

Погода

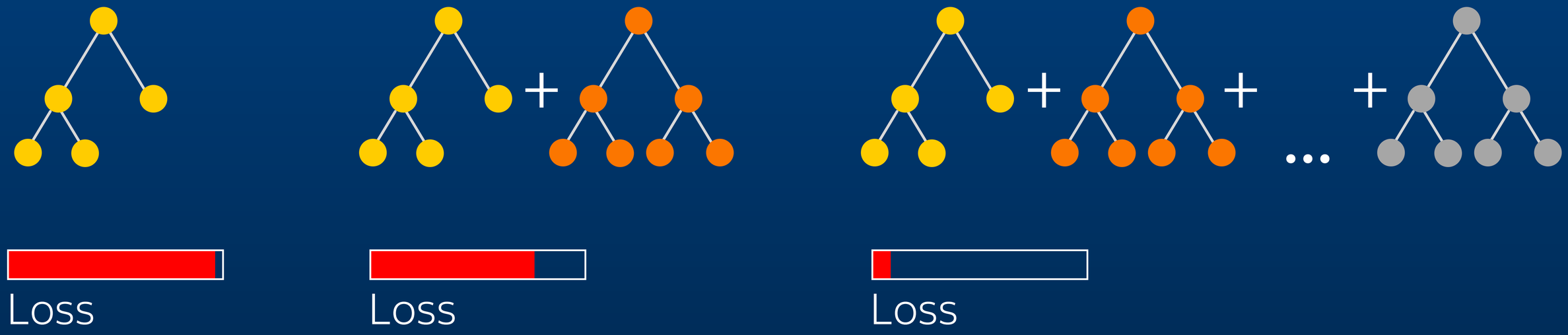
Такси

Музыка

Алиса

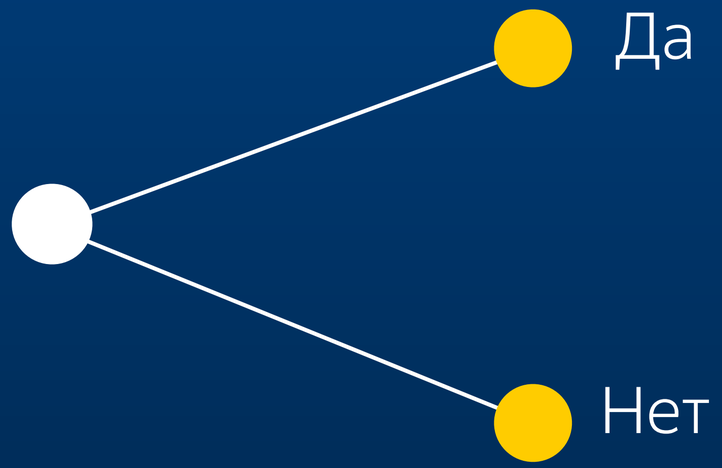
Поиск

Градиентный бустинг

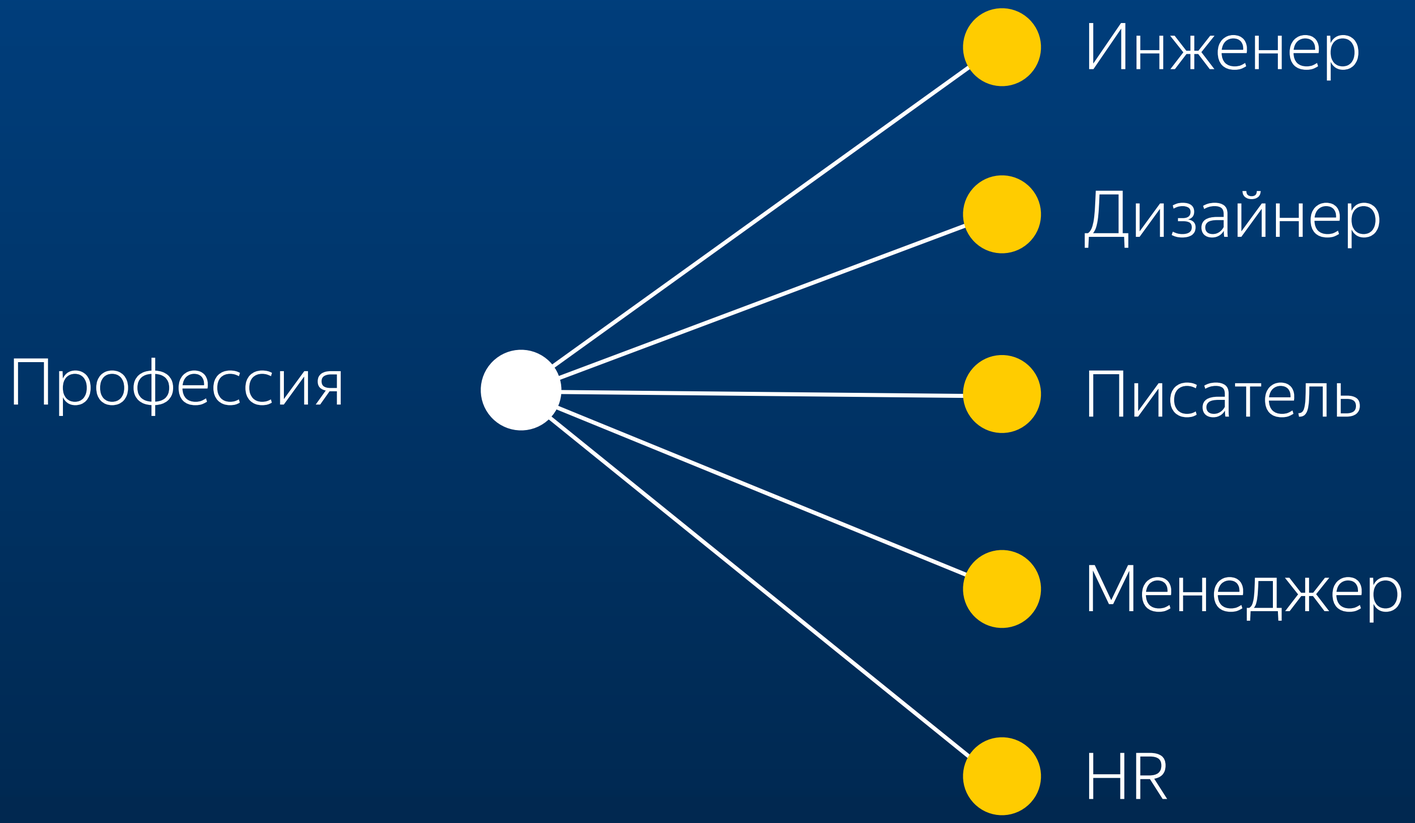


Числовые признаки

Рост > 170 см?



Категориальные признаки



Преимущества

- Поддержка категориальных признаков
- Хорошее качество с дефолтными параметрами
- Методы визуализации и анализа модели

Качество

	CatBoost	LightGBM		XGBoost		H2O	
Adult	0.269741	0.276018	+ 2.33 %	0.275423	+ 2.11%	0.275104	+ 1.99%
Amazon	0.137720	0.163600	+ 18.79 %	0.163271	+ 18.55%	0.162641	+ 18.09%
Appet	0.071511	0.071795	+ 0.40 %	0.071760	+ 0.35%	0.072457	+ 1.32%
Click	0.390902	0.396328	+ 1.39 %	0.396242	+ 1.37%	0.397595	+ 1.71%
Internet	0.208748	0.223154	+ 6.90 %	0.225323	+ 7.94%	0.222091	+ 6.39%
Kdd98	0.194668	0.195759	+ 0.56 %	0.195677	+ 0.52%	0.195395	+ 0.37%
Kddchurn	0.231289	0.232049	+ 0.33 %	0.233123	+ 0.79%	0.232752	+ 0.63%
Kick	0.284793	0.295660	+ 3.82 %	0.294647	+ 3.46%	0.294814	+ 3.52%

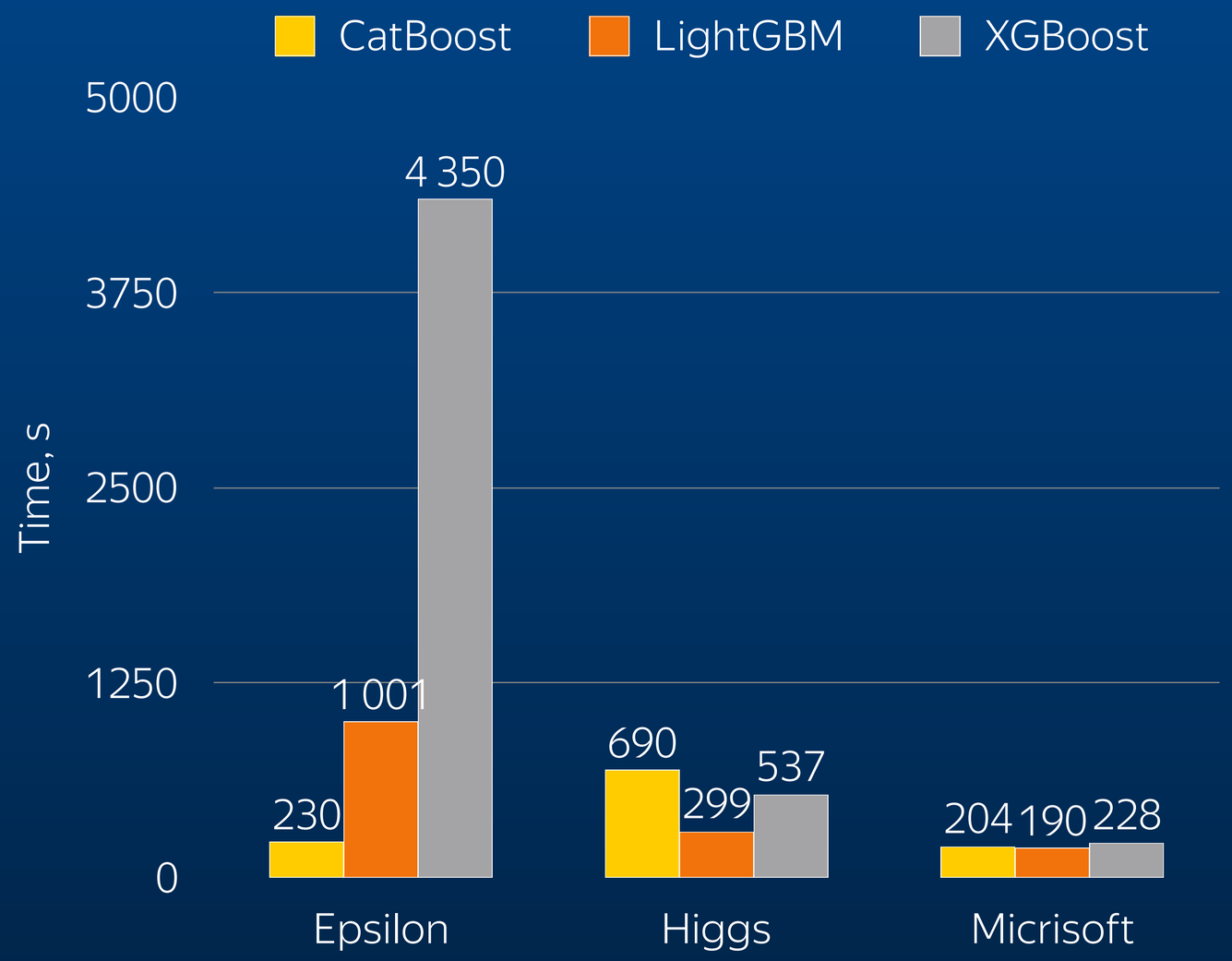
Metric: Logloss

Скорость: CPU Training

- Parameters: 1000 iterations

- Epsilon: 2k features, 400k objects
- Higgs: 28 features, 11kk objects
- Microsoft: 136 features, 1kk objects

- CPU: Intel® Xeon® E5-2660 v4

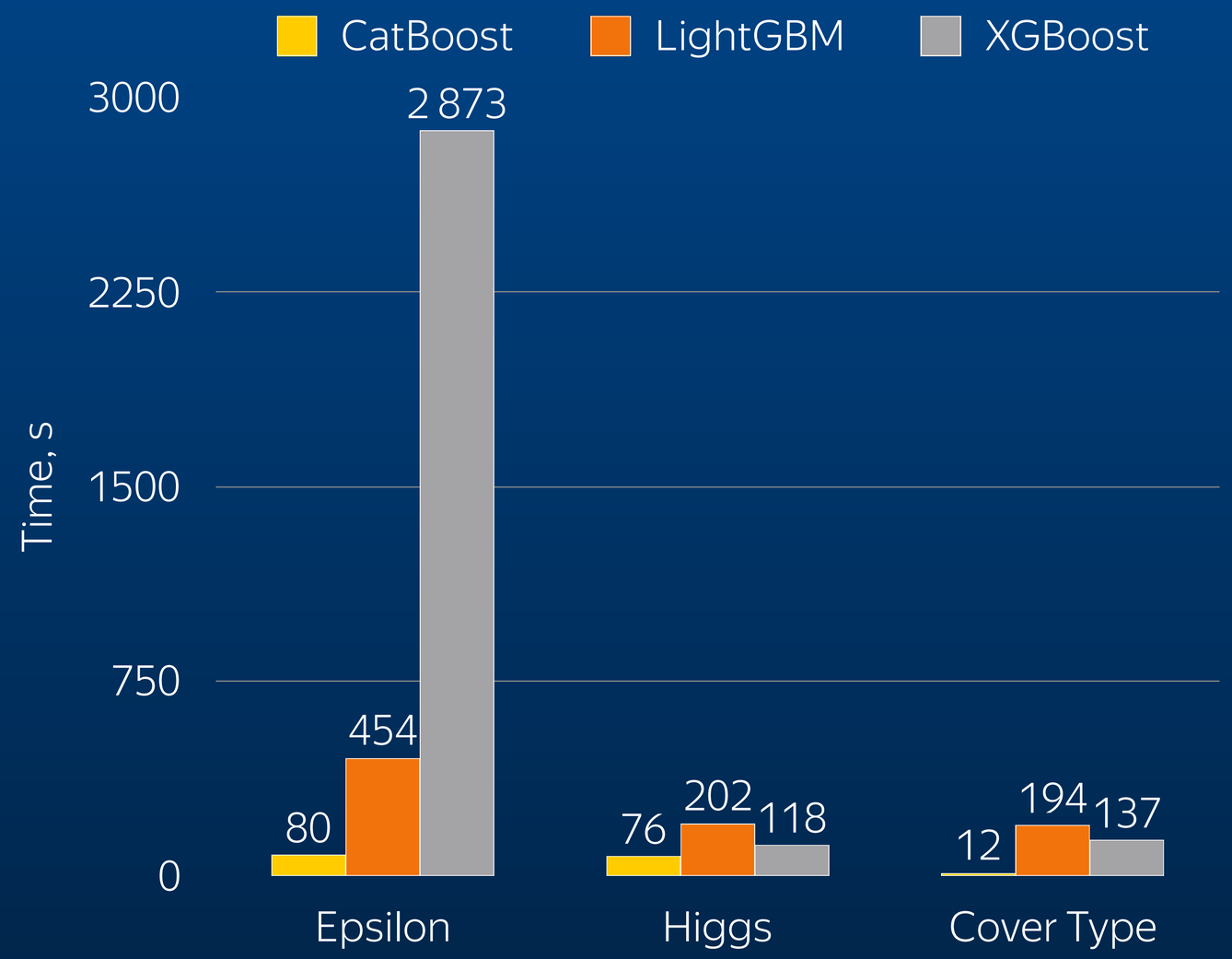


Скорость: GPU Training

- Parameters: 1000 iterations

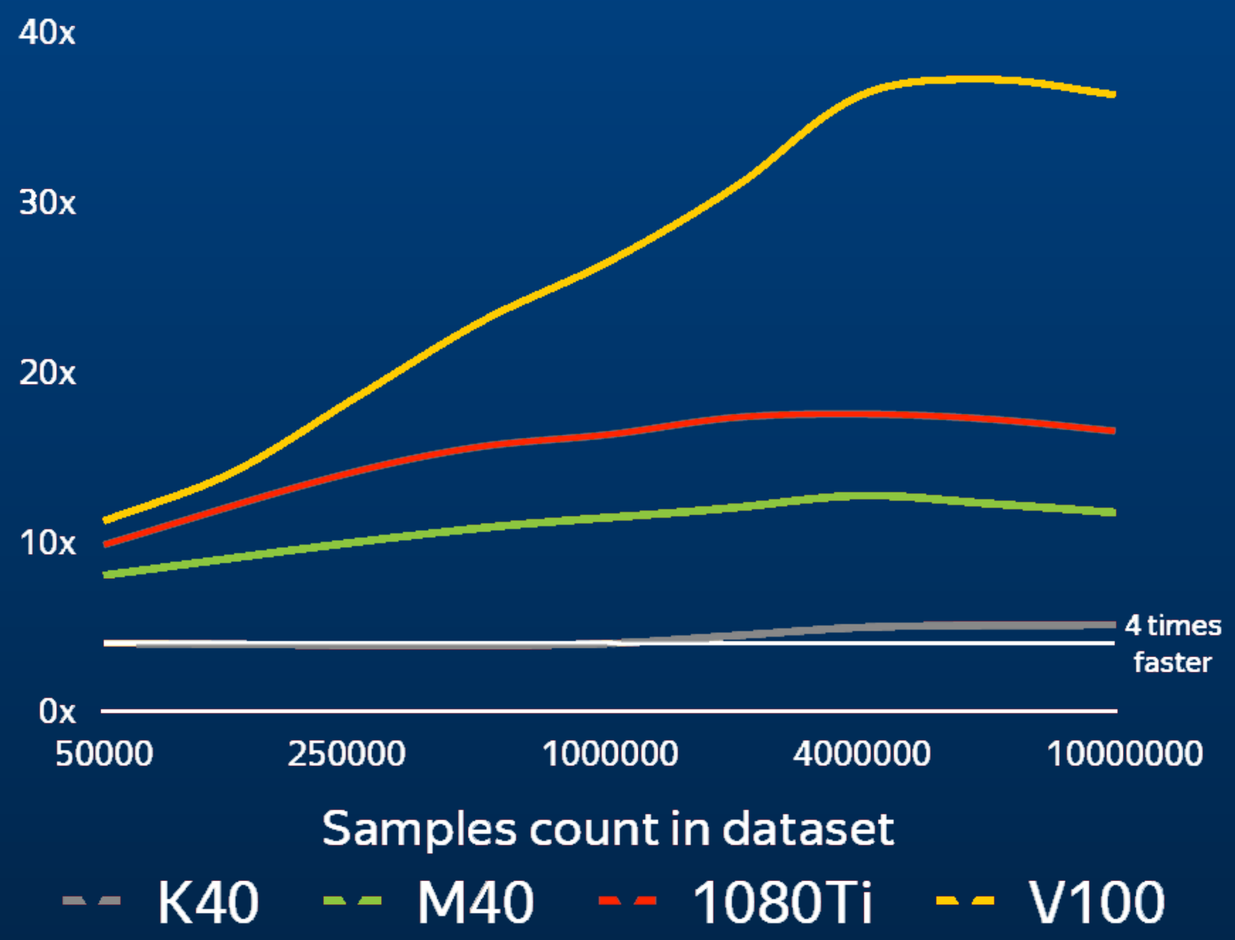
- Epsilon: 2k features, 400k objects
- Higgs: 28 features, 11kk objects
- Cover Type: 54 features,
  523k objects

- GPU: GTX1080Ti



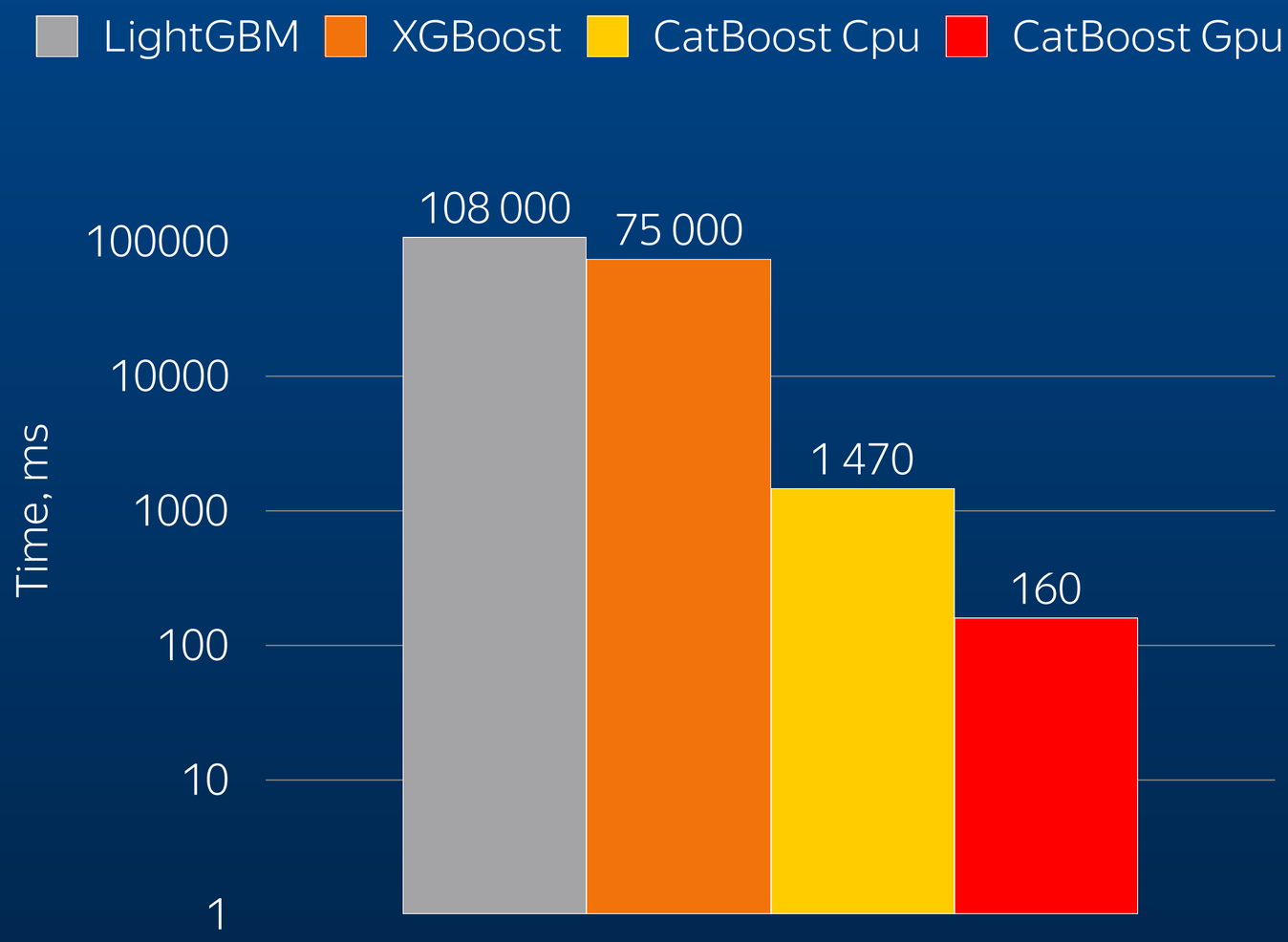
Скорость: CPU vs GPU Training

- Dual-Socket Intel Xeon E5-2660v4 as baseline
- Several modern GPU as competitors
- Dataset: 800 features



Скорость: Prediction

- Parameters: 1000 iterations
- Dataset: 2k features, 100k objects
- CPU: Intel® Xeon® E5-2660 v4



CatBoost Tutorial

- Зайдите на машинку и запустите Jupyter Notebook:

```
jupyter notebook --ip=0.0.0.0 --port=8888
```

- Откройте в браузере адрес **http://clx-#.boostcode.ru:8888**
  где # - номер вашей машинки и запустите туториал
  `catboost_tutorial/intel_hands_on_moscow_nov_07_2019.ipynb`

Планы

- Ускорение обучения на CPU
- Мульти-регрессия
- Полноценная поддержка текстовых признаков
- Новые обучающие материалы

https://twitter.com/CatBoostML

https://catboost.ai

https://github.com/catboost

https://ods.ai => slack => tool_catboost

https://t.me/catboost_ru

Вопросы?

Никита Дмитриев

Разработчик систем
машинного обучения