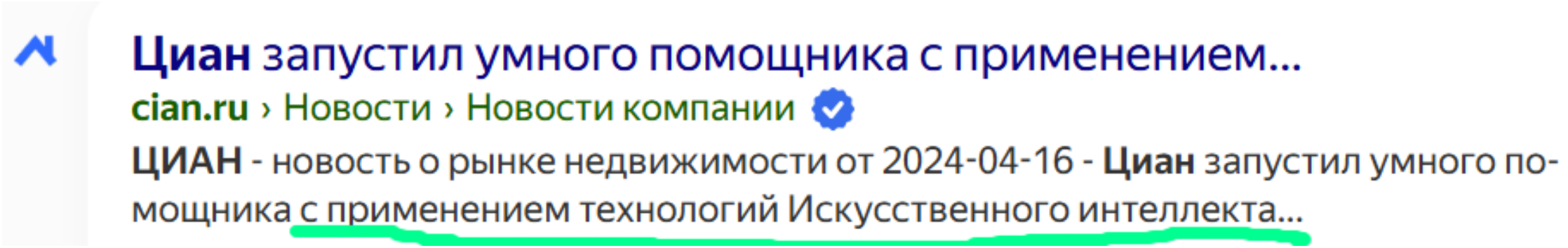
User cases при использовании моделей

Для разработки моделей была выбрана “категория” акций: компании, которые вкладываются в технологии ИИ и/или генетические. Был разбит набор данных на train/test субъективно.

1. Применение модели дало “открытие”

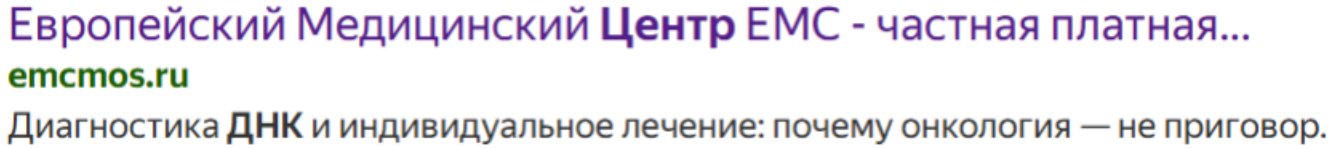
Пользователь не обратил внимание, но модель нашла (в данном случае компанию CIAN — неочевидно, но компания занимается разработками в ИИ области, скриншот из Интернет прилагается).



2. Расширение активов

За счет высокого recall модель удачно включила в результаты акции соответствующих категории компаний, которых не было в тренировке:

- GEMC ЕвроМедЦентр (скриншот прилагается: “ДНК” — генетика);
- МФК Займер: очевидно, применяет ML/DL.
- HHRU HeadHunter: очевидно, применяет ML/AI в работе.



Доля верных ответов

Приблизительно оценим вручную точность ответов в inference модели:
в результатах inference выдал 23 объекта:
4 явно неподходящих (авиакомпания, Русолово, аллюминий, Мостотрест);
4 под вопросом (непонятна принадлежность).
Получается грубо precision ≈ в пределах [0,65; 0,83].

Обновление модели (Модель 1Б)

После изучения полученных результатов набор помеченных данных train/test был дополнен (компанией CIAN). Поскольку размеченный набор данных очень маленький (особенно по положительному классу), то каждый объект имеет значение.

Доля верных ответов

Приблизительно оценим вручную точность ответов в inference модели Модель 1Б:
в результатах inference выдал 18 объектов:
1 явно неподходящая (Мостотрест);
1 под вопросом (непонятна принадлежность).
Получается грубо precision ≈ в пределах [0,89; 0,94].

При этом из результатов не пропала ни одна компания, в которой есть высокая уверенность принадлежности к категории, исчезли некоторые компании “под вопросом” (мало информации, чтобы понять их принадлежность).

Модель catboost

Проверим альтернативную модель на основе градиентного бустинга (catboost).

Доля верных ответов

Приблизительно оценим вручную точность ответов в inference модели catboost:
в результатах inference выдал 10 объектов.
Все объекты подходят.
Получается precision: 1.

Получается, что результаты более точные, чем при использовании Random Forest, но результатов меньше: Модель 1Б выдала, как минимум, 16 подходящих акций, а модель catboost только 10.

Сравнение и выводы

Результаты: “открытий” больше не стало, но субъективно улучшилась доля верных ответов, Модель 1Б стала более прицельной, а модель catboost показала самую высокую точность (за счет снижения покрытия).

Практическое использование моделей в течение нек. времени показало, что 100 % качество не достигается (и пользователь может отметить единицы объектов для тренировки модели), но модель ML полезна. Можно сравнить с моделями распознавания пальцев или лиц: работает далеко не 100 %, но в работе помогает. Если пользователю нужно, чтобы результаты были без шума, то лучше catboost. Если же нужно больше информации “на подумать”, то лучше random forest.

Использование модели позволяет

- расширить число акций в активе, потому что модель фильтрует акции беспристрастно и внимательно с высоким покрытием, исключается человеческий фактор;
- пользователь автоматически получит информацию о новых появляющихся на бирже компаниях, которые будут захвачены классификатором;
- исследовать рынок по выбранному категориальному критерию, потому что модель может «заметить» акции, на которые не обратил внимание человек (например, в случаях, когда неочевидно, что акция на самом деле подходит под критерии, а модель такие случаи может обнаружить) и привести к некоторым «открытиям»;
- также, модель экономит время при многократном/периодическом использовании inference, потому что человеку пришлось бы неоднократно просматривать и анализировать информацию по всем акциям, на такую работу ушло бы больше времени.