

Предсказание повышения

Евгений Голованов

Задача

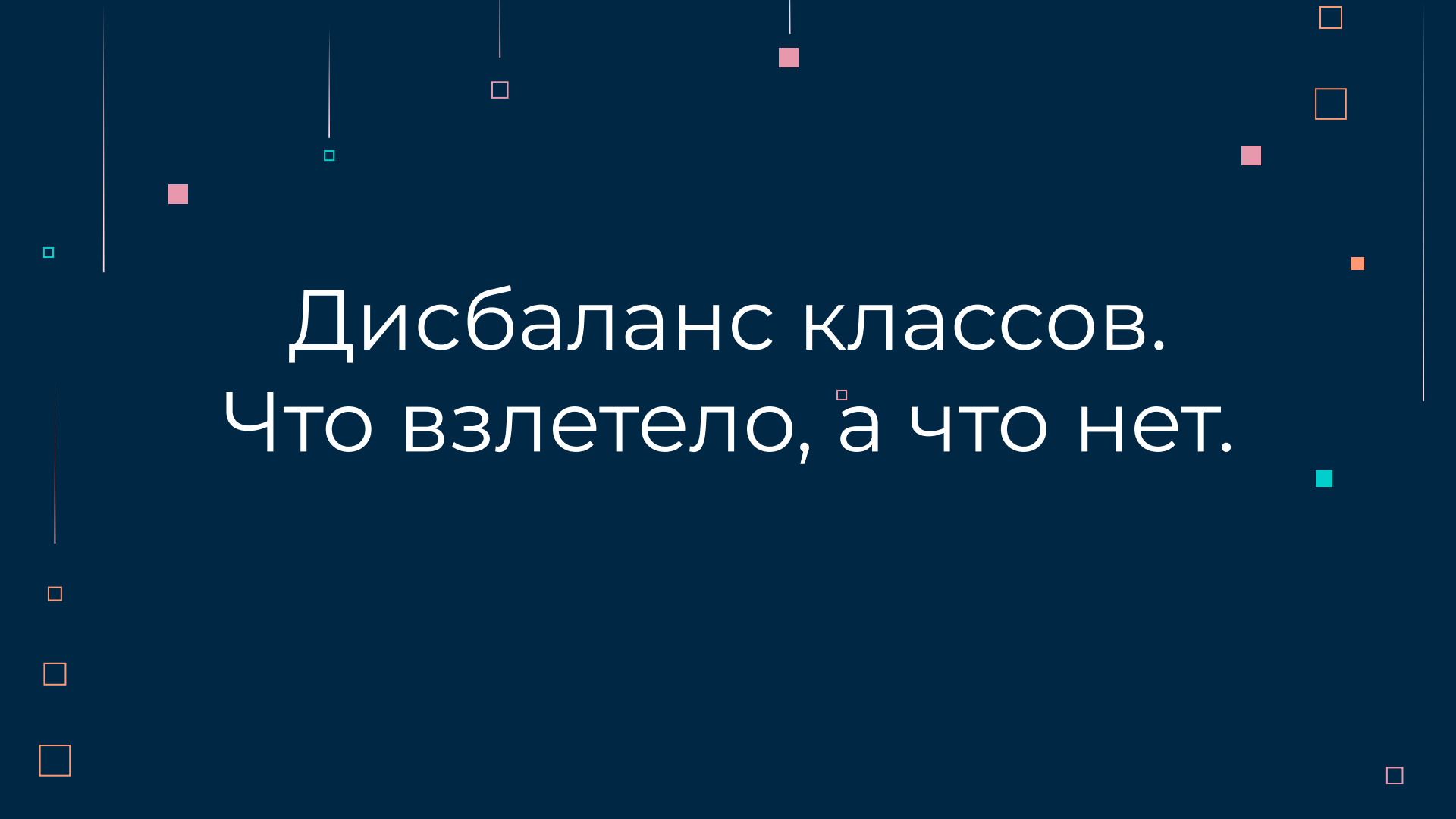
Предсказать повышение по должности и подразделению

Номер класса	По должности	По подразделению
0	1	0
1	0	1
2	1	1
3	0	0

Особенности данных

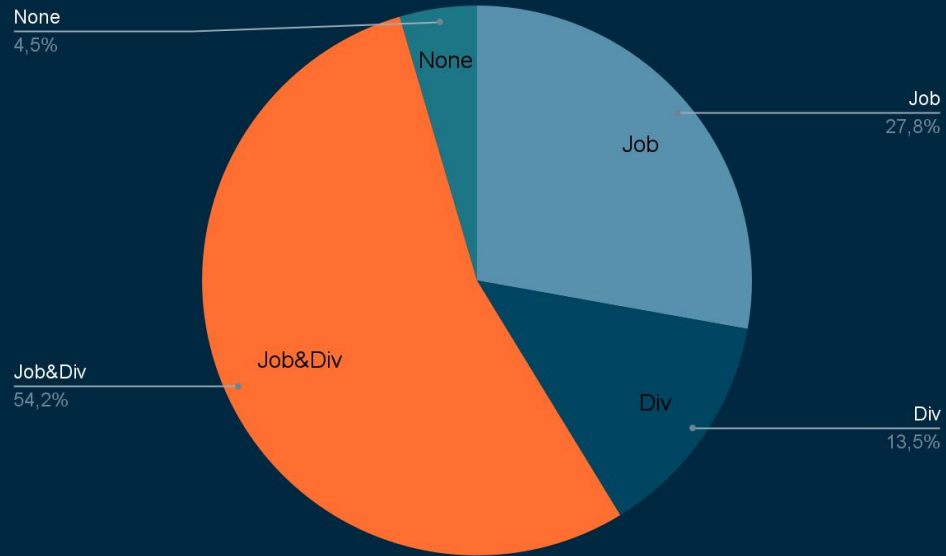
1. Данные избыточны и очень высоко разрежены
2. Очень сильный дисбаланс классов

ВЫВОД:
Нужна правильная
предобработка
данных и
постобработка
результатов
модели

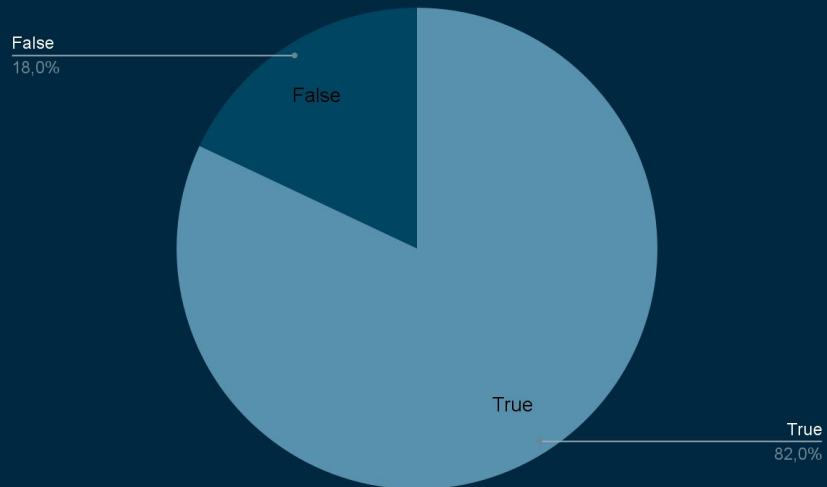


Дисбаланс классов.
Что взлетело, а что нет.

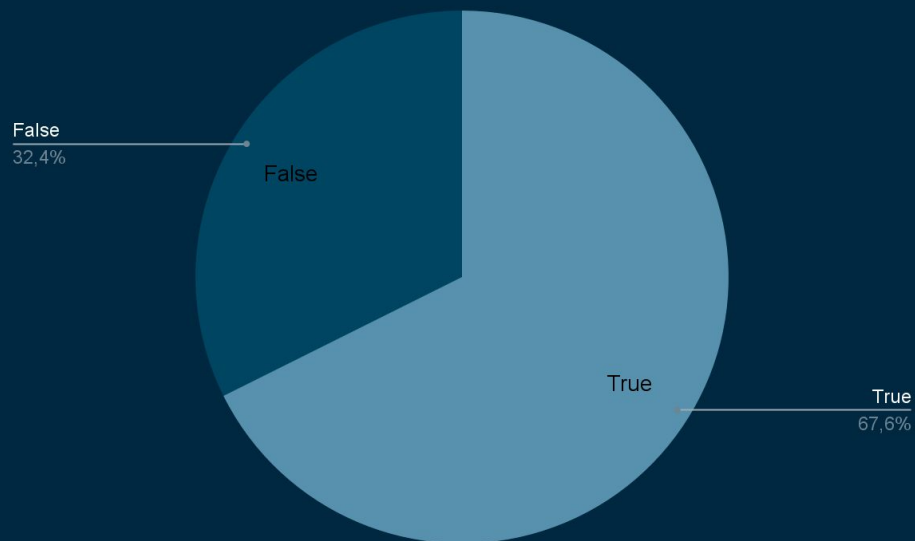
Распределение тренировочных данных



Job



Division

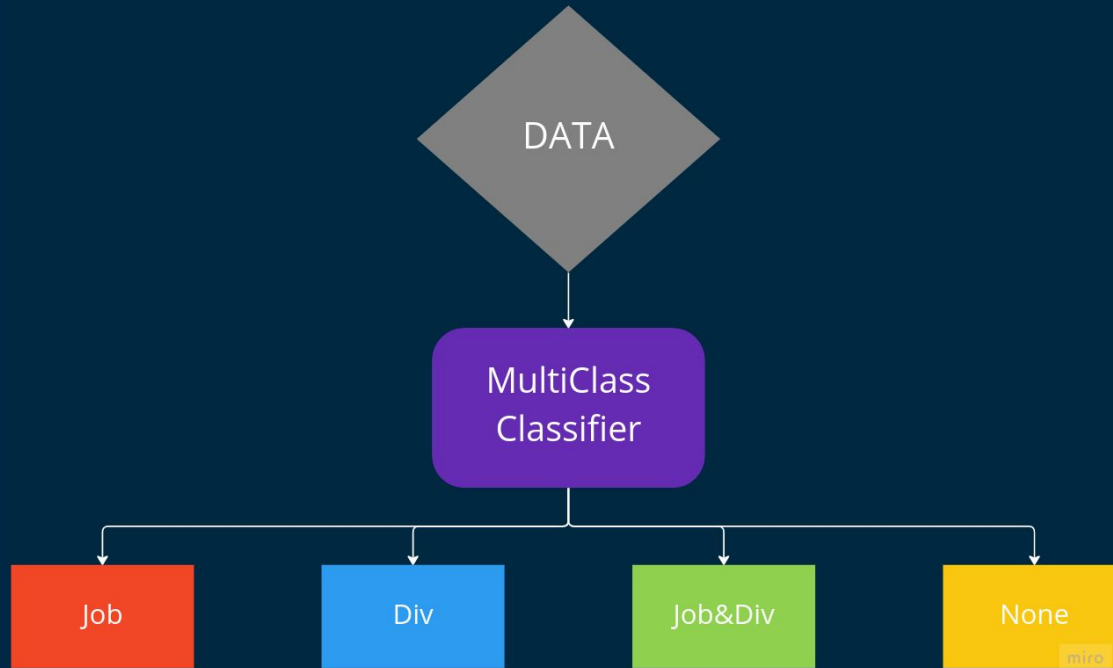


Модель

Почему catboost?

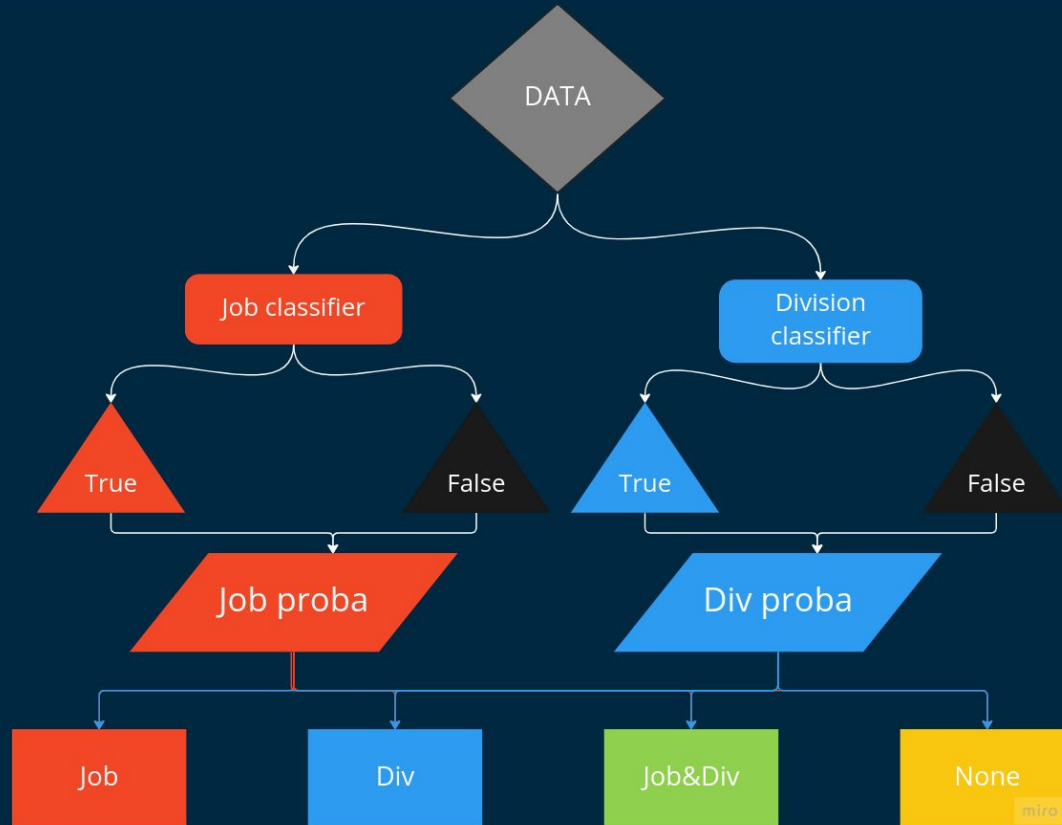
- Показывает самые лучшие результаты на большинстве современных бэнчмарков
- Имеет собственные адаптированные алгоритмы работы с категориальными и текстовыми данными и эмбедингами
- Удобна в эксплуатации и большая часть функций уже оптимизирована “из коробки”
- Применима к широкому спектру задач

Multiclass classification



Max score: 0.42

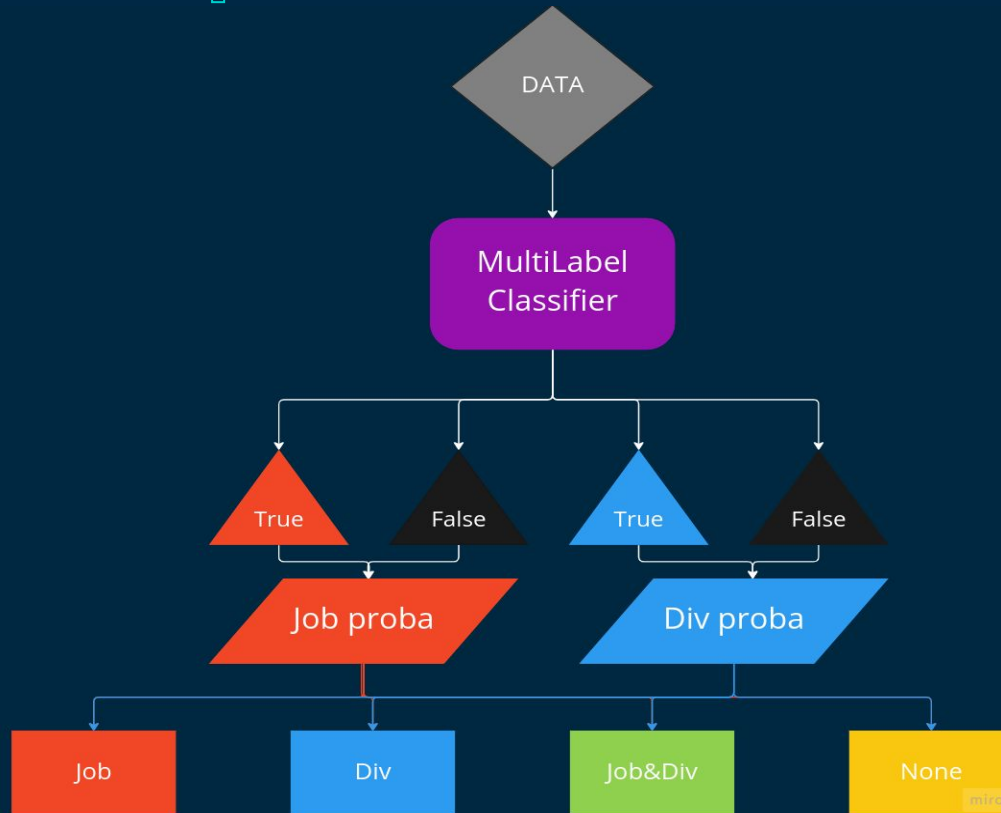
Binary classification



Max score < 0.42

Multilabel classification

Max score: 0.54



Метрика

Recall(“полнота” в пер. с англ.) - количество правильных ответов.

Особенности:

- *За $N - 1$ итераций можно узнать распределение классов*
- *За 2 итерации можно узнать приблизительный размер датасета*
- *Очень чувствительна к ошибкам*

Используя данные особенности, я узнал что классы в тестовой выборке распределены равномерно

Disbalance

- В тренировочном датасете ощутимый дисбаланс, классов, а в тестовом наоборот, таким образом предсказания обученной модели обесцениваются.

- Эту проблему решают трешхолдинг и синтетические данные

Thresholding

- Основная идея: выбрать трешхолды, при которых распределение классов будет наиболее близко к априорному.

$$Loss = abs(mean - median) + median + std + 2(max - min)$$

- В результате, трешхолд подбирается с точностью до 0.005

Synthetic Data

Основная идея: отобрать наиболее вероятных работников принадлежащих к самому непопулярному классу из неизвестных данных.

Алгоритм:

1. Находим трешхолды
2. Предсказываем на признаках всех сотрудников
3. Отсекаем по *трешхолду* - *const.* вероятные данные остальных классов
4. Объединяем с исходным трейном
5. Заново обучаем классификатор на новом

Feature engineering

The background is a solid dark blue. It is decorated with various geometric elements: thin white vertical lines of varying lengths, small solid squares in light blue, pink, and orange, and small hollow squares in orange. These elements are scattered across the slide, with some appearing to be connected by thin lines, creating a modern, abstract aesthetic.

Education

Данная таблица содержит информацию обо всех работниках и является самой информативной и наименее разреженной. Колонку с образованием упростил, так как некоторые классы были очень похожи(заменял похожие значения на ближайшие). Ключевые признаки из этой таблицы:

- До 6-и специализаций сотрудников в хронологическом порядке
- До 3-х образований в хронологическом порядке
- Продублировал все признаки(в т.ч. из других таблиц) на идентификаторе руководителя

Connection Time

■ Данная таблица очень важна с точки зрения повышения по должности.

□ Ключевыми признаками являются стандартные статистики(сумма, стандартное отклонение, медиана, количество - “count”)* по времени□опоздания.

□ * - те же статистики будут мною использоваться и
□ далее
□

Выбросы

Формулировка:

“Правильными” вещественные числовые данные считаются если входят в интервал

$$[med - mad; med + mad]$$

med - медиана, а

mad - медианное отклонение от медианы

Working Day

Ключевые признаки:

- Статистика по колонкам `ActiveTime` и `MonitorTime`
- Количество выбросов по будням и выходным по `ActiveTime` и `MonitorTime`
- Также добавил признак частоты ("frequency"), который представляет собой $\text{count}/(\text{max_count} - \text{min_count} + 1)$

Примечание: `max` и `min` - глобальные, т.е. среди всех сотрудников в данных

Network

Очень похожа на предыдущую, но имеет более релевантные значения и в ней нет классификации на будни/выходные, как и колонки ActiveTime.

Ключевые признаки:

- Статистика по Monitor Time + frequency
- Количество выбросов
- Статистика по всем id таких же специализации, образования(текущих) и руководителя(все по отдельности)

Tasks

Самая информативная таблица с точки зрения ценности информации, но она также отличается не очень большим количеством уникальных id(1272).

Ключевые признаки:

- Мода задачи
- Количество просроченных задач и сколько задач всего
- Распределение решаемых задач(количество данной задачи делить на общее количество задач)
- Статистика по количеству просроченных дней

Calls

Небольшая таблица необходимая для точного предсказания позиции людей работающих непосредственно с людьми. Ключевые признаки:

- Статистика по длительности звонков
- Количество звонков в будни, исходящих, всего
- Количество выбросов(пересечение выбросов в NumberOfCalls и CallTime)

SKUD

Самая маленькая по количеству уникальных id
таблица

Ключевые признаки:

- Количество выбросов общей длительности
- Усеченная статистика по общей длительности
рабочего дня(std и медиана)

Графики

Спасибо за внимание! Github

