Задача 2

Информация о нас

Команда: Насосные эксперты

Имя капитана команды: Рузманов Дмитрий Вячеславович

Состав команды:

- Рожков Павел Дмитриевич
- Мынко Семён Андреевич
- Рузманов Дмитрий Вячеславович

План

1) Проблема
2) Анализ данных
3) Очистка данных
4) Выбор модели
5) "Оу май, это же точность 1.0" или почему мы считаем, что модель не переобучена
6) Выводы

Проблема

Выявляем проблему

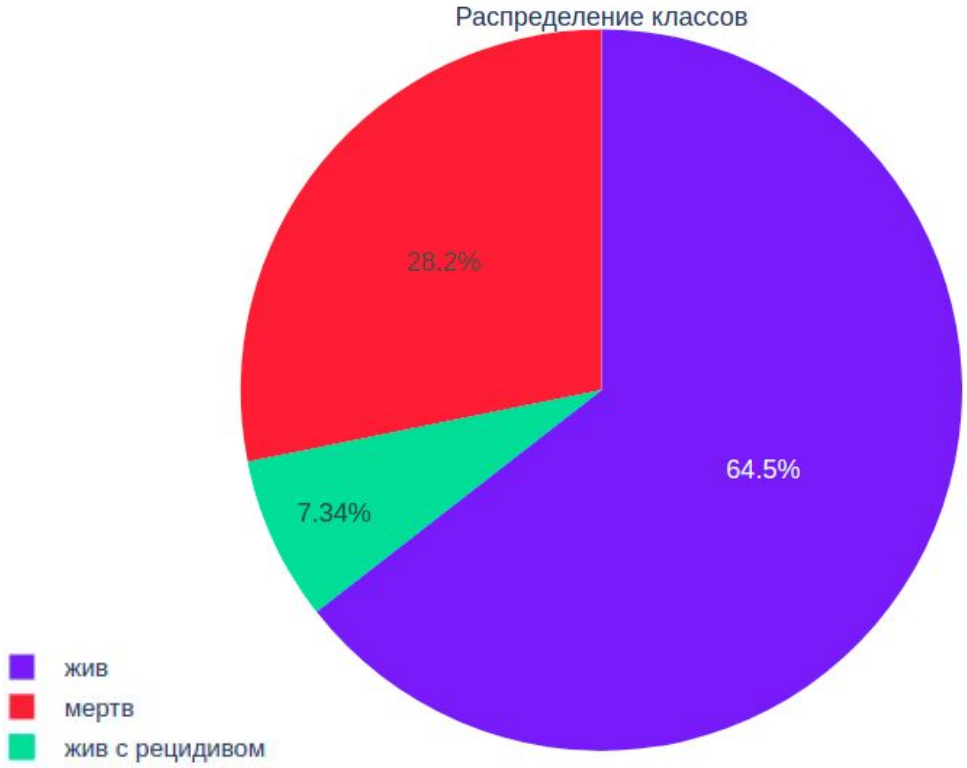
Цель нашей работы предсказать состояние человека:

- жив
- мертв
- жив с рецидивом

Анализ данных

Анализируем данные

```
Размер данных:
количество строк(259)
количество признаков(117)
Какие типы данных используются: int64 float64
Количество дубликатов строк: 0
Количество дубликатов колонок: 4
Процентн Nan в датасете: 8.67 %
```



Анализируем данные

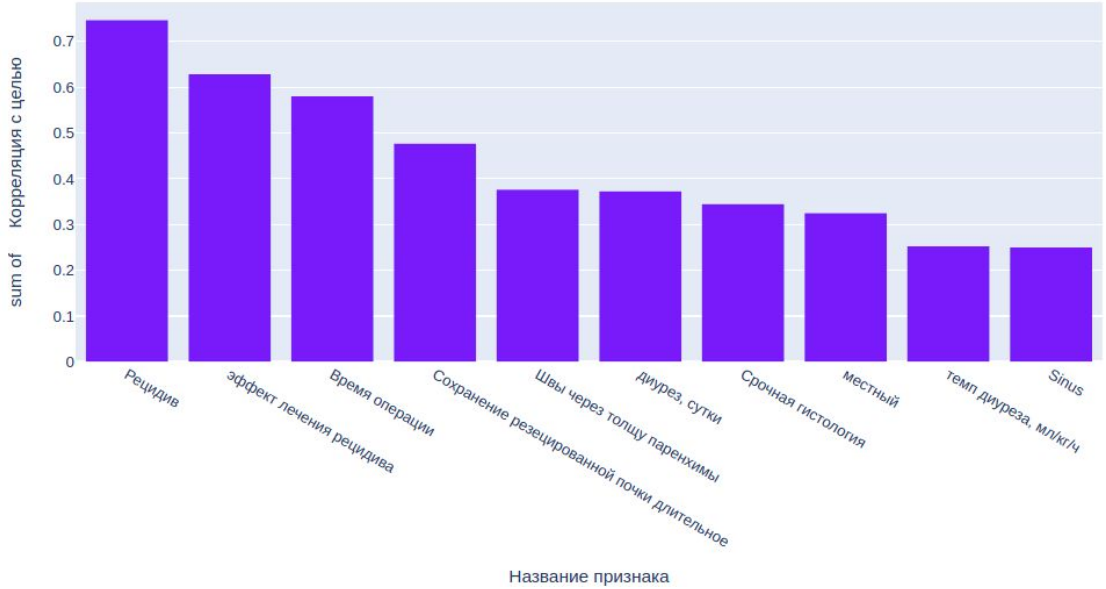
При анализе данных были выявлены:

- Неправильно рассчитанные признаки (на основе медицинских определений)
- Сильно коррелирующие с другими признаками (участвующие в зависимостях)
- Неинформативные колонки :
    - Большое количество Nan (>50%)
    - Сильное преобладание одного из значений (>70%)
    - Дубликатные колонки

к презентации прикреплен файл “Информация_о_признаках.pdf” с дополнительным обоснованием

Корреляция с таргетом

При анализе **необработанных данных** были выявлены признаки наиболее коррелирующие с классами предсказания



Очистка данных

Очистка данных

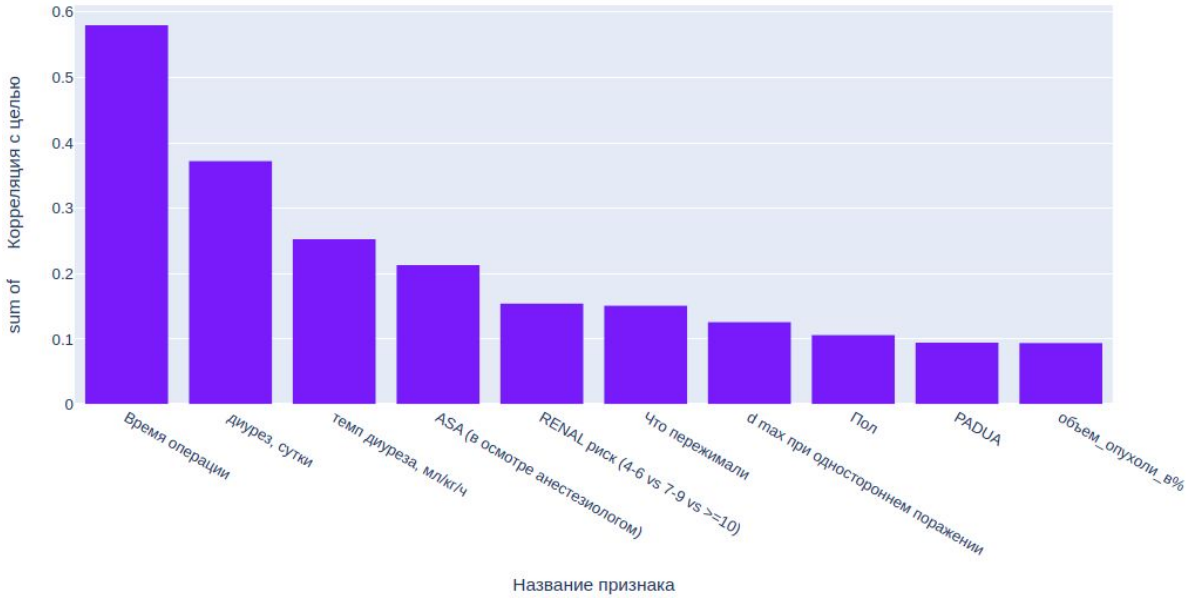
- пересчет зависимых признаков
- удаление признаков:
  - с помощью которых пересчитывали зависимые
  - дубликатных
  - с более 50% nan
  - с преобладанием одного значения более 70%

После очистки

```
Размер данных:
количество строк(259)
количество признаков(44)
Какие типы данных используются: int64 float64
Количество дубликатов строк: 0
Количество дубликатов колонок: 0
Процентн Nan в датасете: 5.16 %
```

Корреляция с таргетом

После обработки данных были выявлены признаки наиболее коррелирующие с классами предсказания



Выбор модели

Выбор модели

Была выбрана модель градиентного бустинга над решающими деревьями CatBoostClassifier.

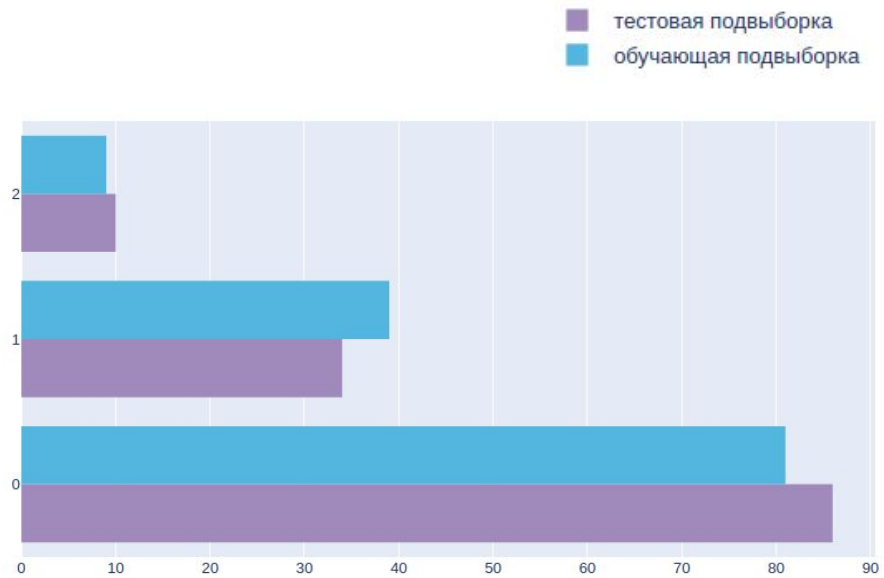
Причины выбора:

- Слабая чувствительность к выбросам
- Эффективно обрабатывает данные с большим числом признаков
- Одинаково хорошо обрабатывает дискретные и непрерывные признаки
- Редко переобучается
- Возможно оценивать важность признаков для модели
- Хорошо работает с пропущенными данными
- Результативно работает на небольших датасетах

Точность 100% или почему модель не переобучена

Работа модели

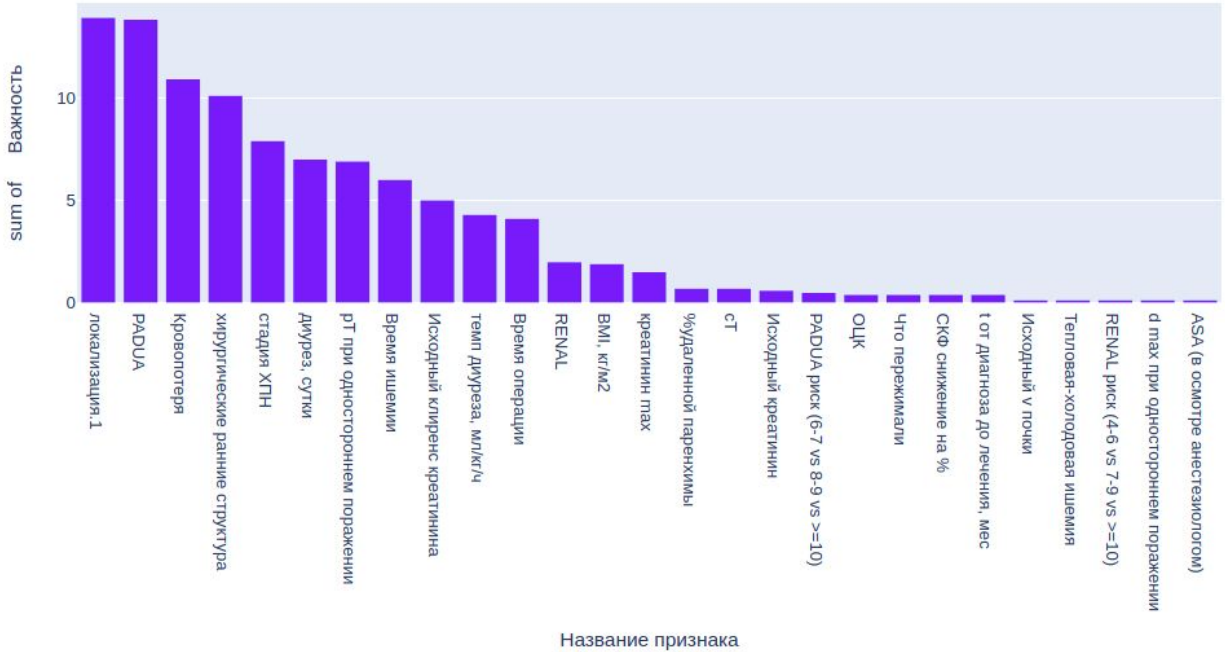
- Валидационная и тренировочная выборки 50%-50% для более точной проверки обобщающей способности модели
- Параметры модели:
  - количество деревьев = 100
  - максимальная глубина = 2
- Полученное качество на валидационной выборке: 100%



Почему модель не переобучена

Признаков с явно завышенной важностью нет. Из этого следует, что модель не переобучена и утечки данных не произошло.

Поскольку утечки данных не обнаружено, можно считать точность на валидационной выборке честной.



Почему модель не переобучена

Визуализировав некоторые важные для модели признаки видим, что по положению точек можно разделить 3 класса.

Из этого следует, что модель может давать точность 100%

Вывод

Вывод

Проведя исследование и очистку данных была обучена модель градиентного бустинга над решающими деревьями. Данная модель безошибочно предсказывает по важным для нее признакам состояния человека: жив, мертв, жив с рецидивом.

Конец