

BEST HACK 2022

Презентация решения команды Punk Butterfly

Состав команды Punk Butterfly




НЕМЦЕВ ДАНИИЛ
ЮРЬЕВИЧ



КУПРИЯНОВ ИЛЬЯ
ВЛАДИМИРОВИЧ
TEAM LEAD



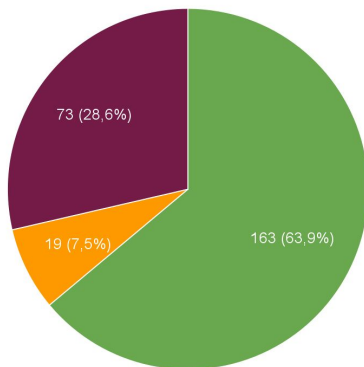
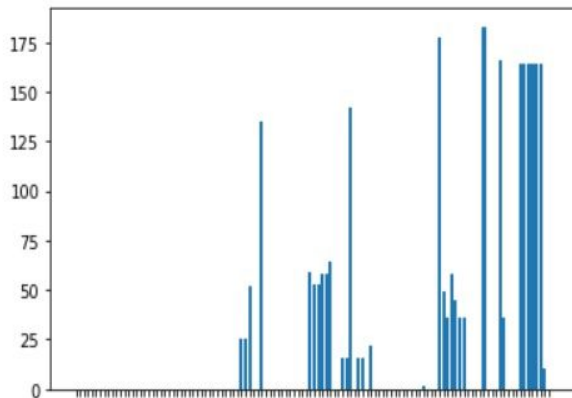
ШКОЛИН АЛЕКСАНДР
ЮРЬЕВИЧ

The background is a solid pink color. In the top right corner, there is a decorative pattern of overlapping triangles in various shades of pink and magenta, creating a geometric, abstract design.

**Данные -
подчинить или
подчиниться?**

Пара строк и все наглядно

Мы провели как поверхностный анализ данных, посмотрев на количество незаполненных столбцов у пациентов и на распределение целевых классов, так и глубокий, составив описание каждой фичи в отдельном файле.



- Живы
- Живы с рецидивом
- Мертвы

Обработка данных и устранение неточностей в датасете

Нам пришлось удалить некоторые фичи. Например, “Pn” и “эффект” имели слишком одинаковые значения.

В процессе обработки столкнулись с интересной особенностью: в некоторых строках с “Рецидивом” = 0 присутствуют данные о его лечении. Мы посчитали, что ошибка в столбце “Рецидив”, поскольку в нем одним ошибиться проще, чем в последующих пяти.

Рецидив лечили -> ставим “Рецидив” = 1

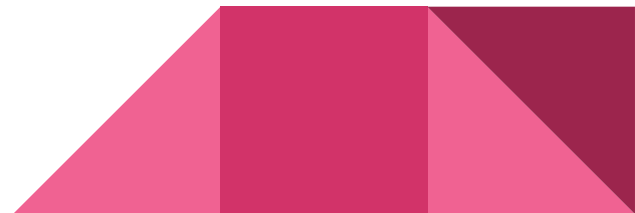


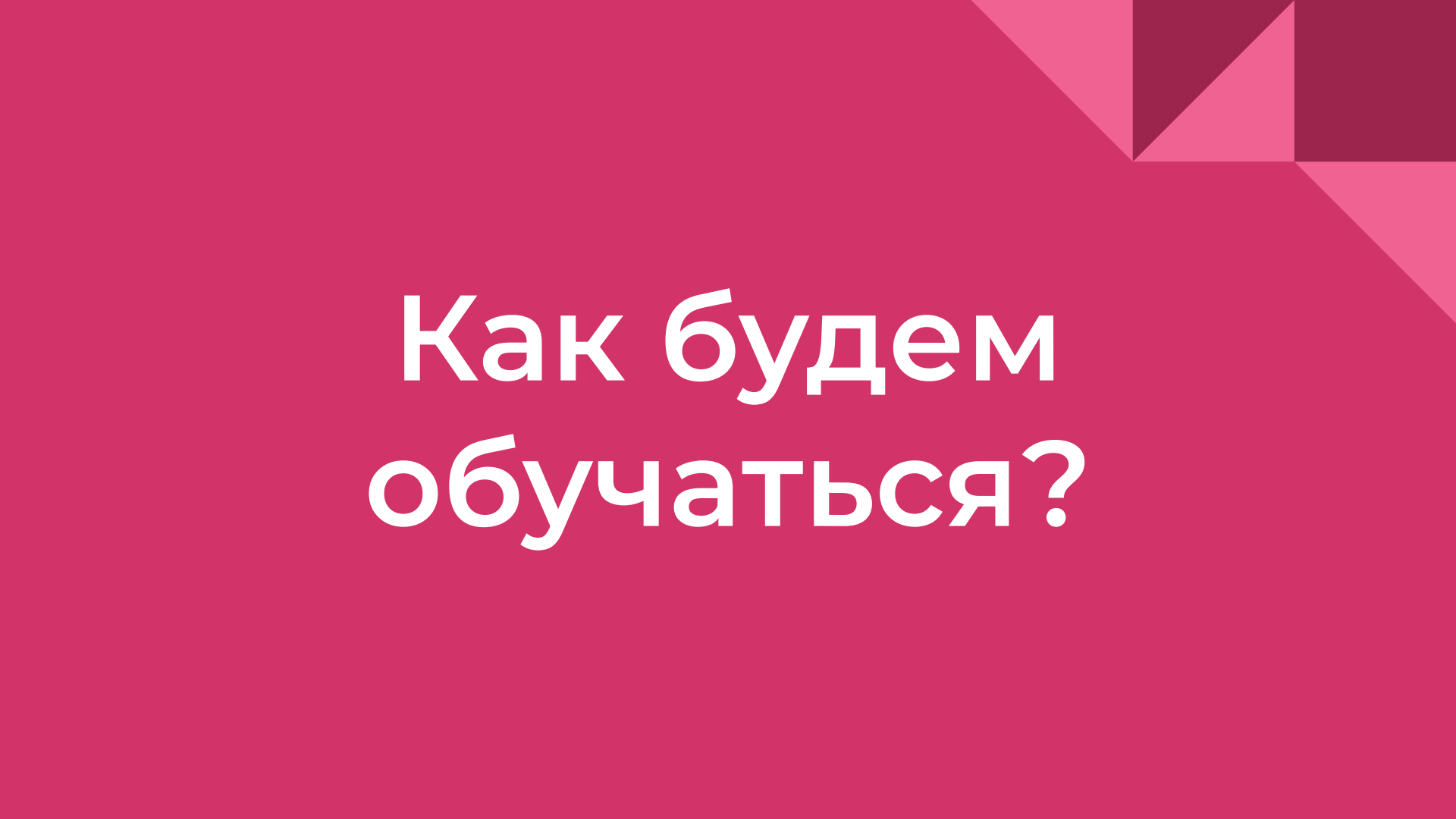
Корреляция - залог успеха

Сохранение резецированной почки длительное	0.476448
Лечение рецидива	0.802374
мтс	0.838403
эффект лечения рецидива	0.845464
системное лечение	0.851742
вид рецидива	0.855215
местный	0.881270
Рецидив	0.887012
Текущий прогноз	1.000000

стадия ХПН	-0.495264
Ишемия да-нет	-0.427281
гемостатическая губка	-0.410592
Longitudinal	-0.400485
ХПН	-0.367882
Доступ первой резекции.1	-0.355598
Доступ первой резекции	-0.355598
кисты	-0.347374
Стадия	-0.337570

Мы нашли прямую и обратную корреляции каждой фичи с таргетом и выбрали из них наилучшие.



The background is a solid pink color. In the top right corner, there is a decorative pattern of overlapping triangles in various shades of pink and magenta, creating a geometric, abstract design.

Как будем
обучаться?

Модели, модели, модели

По причине небольшой выборки, мы применили методы кластеризации без учителя “k-means”, “spectral” и “hierarchical” и сравнили их работу.

Чтобы выбрать самые информативные фичи, ввели два гиперпараметра: количество фичей с максимальной положительной корреляцией и минимальной отрицательной. При достаточно больших гиперпараметрах (10, 10) ассурасу оказывалась невысокой.

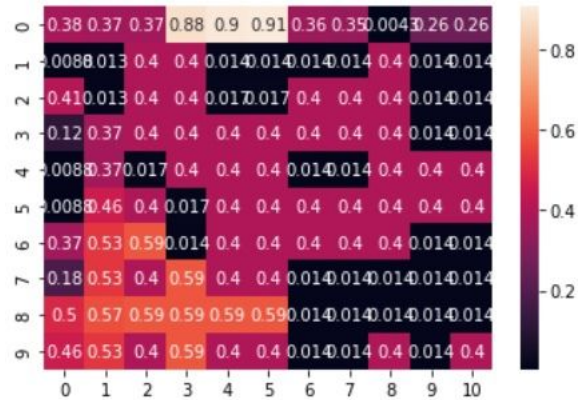
Запускаем недолгий перебор!



Пожинаем плоды и радуемся

На heatmap показана ассурасу, в зависимости от двух параметров, о которых мы говорили в предыдущем слайде. Лучший результат показала иерархическая кластеризация при параметрах от (4, 6) до (9, 11)

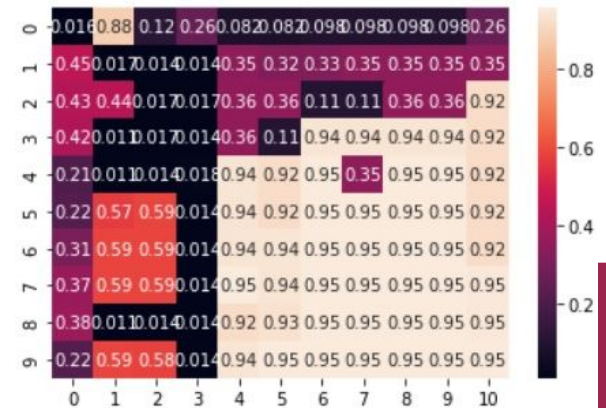
K-Means



Spectral

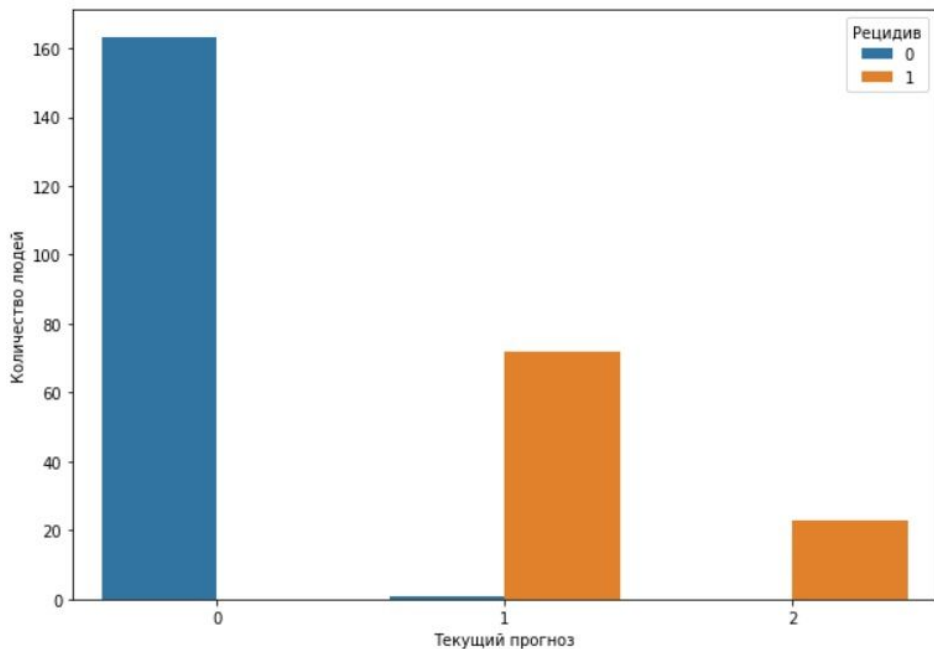


Hierarchical



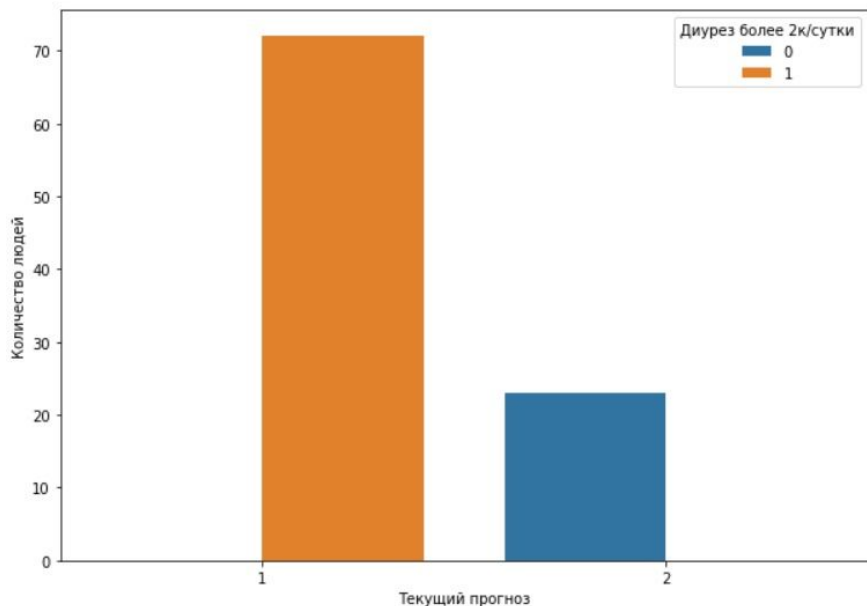
Accuracy > 0.95 -
мечта или цель?

Эвристика и немного работы “руками”



При учете неточности в данных нетрудно заметить, что признак “Рецидив” очень сильно влияет на выживаемость. В выборке Train всего лишь один человек, который умер без рецидива.

Тревожный звонок из приватной комнаты



Интересной фичей оказался “диурез в сутки”. Из 73 умерших людей у 72 значение этого показателя было равно или превышало 2000.

А из 186 выживших всего у 28 наблюдался повышенный диурез


В то же время у большинства выживших с рецидивом показатель “диурез в сутки” не замерялся или отсутствовал по другим причинам.

Решение в три строчки

Основываясь на зависимостях, полученных ранее, предсказываем исход для выборки Train вручную. На языке условий:

- Если случился рецидив и диурез в сутки ≥ 2000 , человек умер (Класс 1)
- Если случился рецидив и значение диуреза отсутствует, человек выжил с рецидивом (Класс 2)
- Все остальные люди выжили без рецидива (Класс 0)

Результаты этой теории, оказавшись лучше результатов иерархической кластеризации, нас впечатлили. Было решено использовать данный подход для итогового предикта.



BALANCED ACCURACY SCORE

0.9813373253493015

**Спасибо за
внимание!**

Ждем оценки нашей работы :)