Отчёт

Stardust Crusaders

18 сентября 2022 г.

Содержание

Используемый датасет и организация проекта

Предварительная обработка данных

Основной трудностью обработки данных было аккуратно категоризовать события и агрегировать историю болезни. Этому препятствует ряд проблем:

- Некоторые поля "примечания к событию" заполнены некорректно (в несоответствии с форматом, указанным в таблице)
- К некоторым событиям поле "примечание к событию" подразумевает свободное заполнение. Из-за частого использования аббревиатур и сокращений, некоторые одинаковые события при наивной категоризации будут помечены как разные (например событие 'НЕОСЕМ' с примечанием 'МАСТ' и примечанием 'МАСТИТ' будут размечены как разные).
- Организация таблицы как поочередной последовательности событий создаёт трудность при агрегировании истории болезни одной коровы: так, чтобы понять, помог ли определенный протокол лечения, надо посмотреть все события относящиеся к данной корове и проверить что среди событий произошедших после назначения лечения, идет выписка или наоборот, очередное назначение протокола.
- Ряд других, мелких проблем: при заполнении поля "примечания к событию" довольно часто встречаются опечатки (например к событию 'ЗДОРОВА' поле примечания заполнено как 'ГЕНЕКОЛ'); при описании применяемого протокола лечения к событиям 'МАСТИТ' и 'ХРОМОТА' указание пораженных долей/конечностей происходит в произвольном формате: например '1-3' или '12,3'; часть событий указана на английском, часть на русском;

Схема парсинга:
Сначала данные приводятся к относительно единому формату: исправляются опечатки, значения поля 'событие' конвертируются на английский язык, поля 'примечания к событию' где указаны протоколы лечения также приводятся к единому шаблону. После этого в соответствии с определенными в скрипте правилами, события "категоризуются" — для каждой категории событий появляется свой столбец со значениями 0 и 1, где значение 1 указывает что событие относится к данной категории, значение 0 означает иное. Такая организация данных удобна для подсчета статистики и алгоритмов машинного обучения.

Для того чтобы агрегировать историю коровы, события для каждой коровы упорядочиваются по дате и после этого для каждого события 'МАСТИТ' просто поиском по событиям проверяется помог ли данный протокол или нет. После этого создается новая таблица в которой строка отвечают одному заболеванию одной коровы. Столбцы в этой таблице уже содержат информацию об истории коровы: сколько раз она болела и чем, чем её лечили и прочее. Эта таблица содержит уже только коров, так как предназначена для анализа заболеваемости маститом.

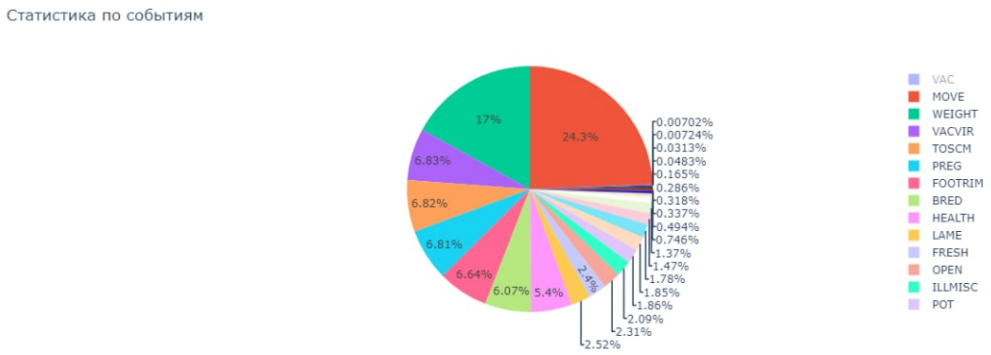
Отдельно парсинг проходит для алгоритма на основе трансформера (см. секцию ниже). Отличия TODO

Организация предварительно обработанных данных:
В результате предобработки исходный файл `raw.csv` преобразовывается в два файла:

- `counters.csv` — csv-файл строки которого это некоторые события.
- `target.csv` — csv-файл с агрегированной историей болезни.
- `events.csv` — csv-файл для нейросети.

Общие наблюдения

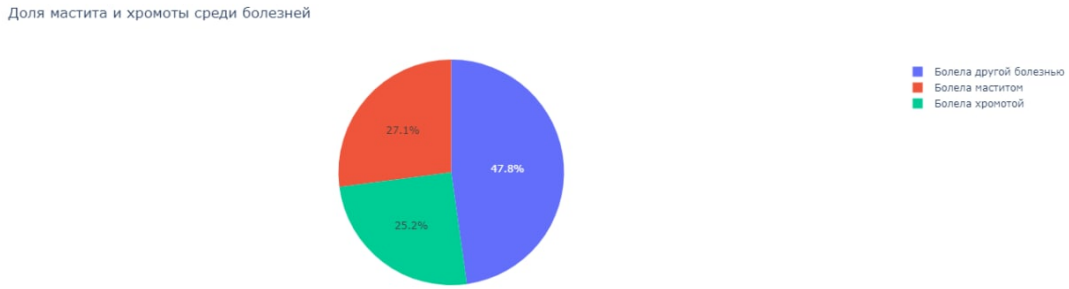
Распределение событий в исходном датасете:

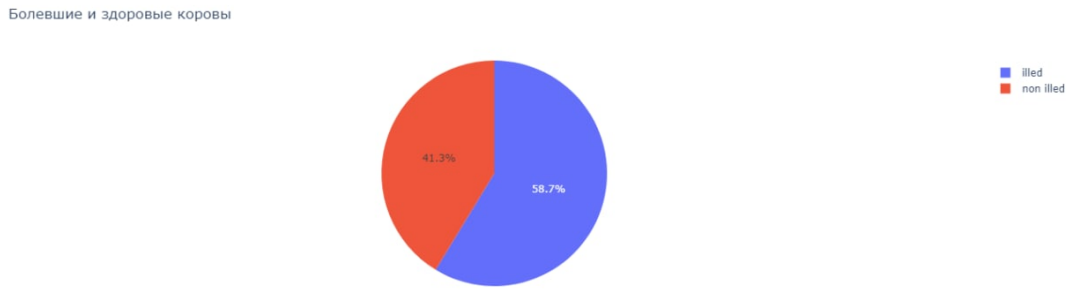


Некоторые числовые характеристики:

- Число коров: ~15300
- Число коров которые болели маститом: ~2700
- Число заболеваний маститом: ~5500
- Число коров которые болели хромотой ~2600

При этом видно что мастит и хромота представляют значительные доли среди всех заболеваний:



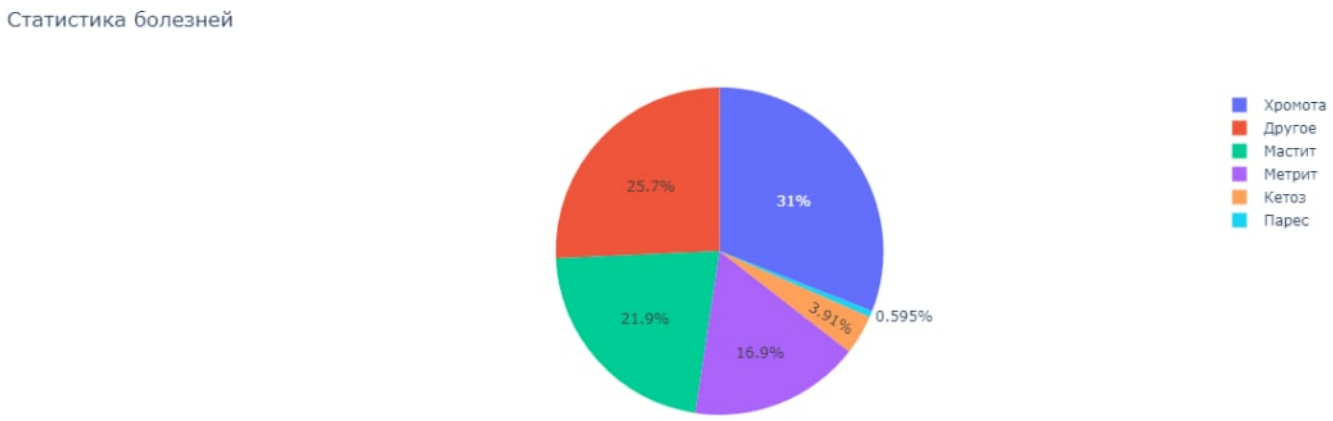


Из-за нехватки времени мы сфокусировались только на мастите.

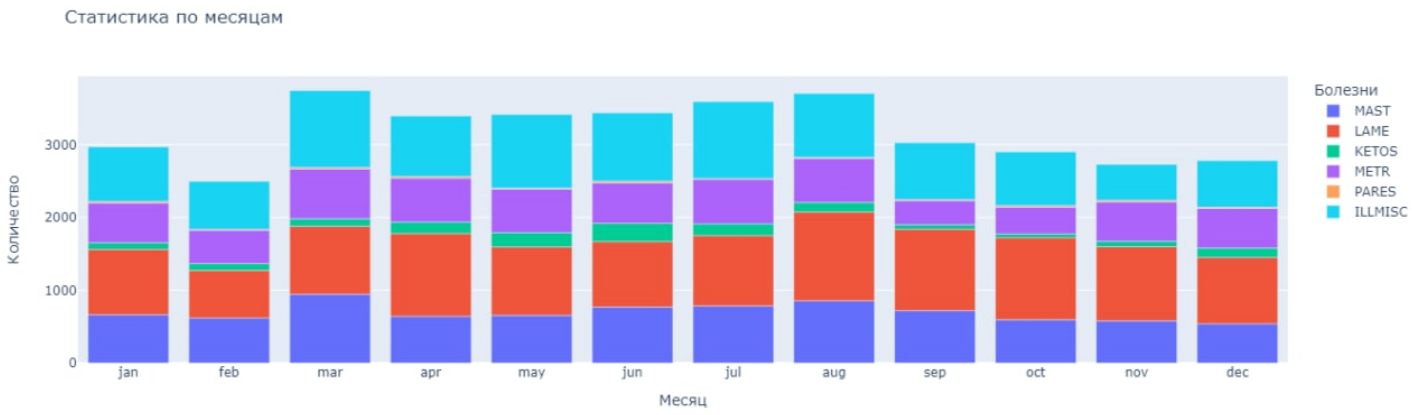
Анализ и поиск закономерностей

Анализ заболеваемости коров

Общее распределение болезней:



Гипотеза: число заболеваний зависит от месяца:



Визуально выбросов нет.

Анализ по группам

Несколько замечаний:

- В некоторых группах животные маститом не болели вообще, вероятно это может быть связано
- Из анализа убраны служебные группы: 1, 2, 3, 12, 17, 18, 21, 34, 41, 42, 43, 44

В силу малости выборки делать какие-либо заключения относительно статистики распределения заболеваемости маститом по группам сложно. Однако, в группах 4, 5, 6 где число заболеваний достаточно велико, возможно есть какие-то нарушения гигиены.

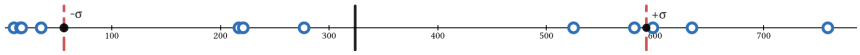


Рис. 1: Распределение числа заболеваний в группах

Закономерности и гипотезы

Связь заболеваемостью маститом и наличия атрофированных сосков: Все коровы, у которых были атрофированны соски также когда-либо болели маститом (правда, пораженные соски и атрофированные соски совпадают не всегда). При этом в половине случаев, атрофированию сосков предшествовал клинический мастит за одну две недели)

Замечание: за все время было зафиксировано очень малое количество случаев атрофирования сосков, поэтому невозможно сделать какой-то статистический вывод.

Влияние послеродовых заболеваний на заболеваемость маститом: Среди коров, которые перенесли одно из послеотельных заболеваний (парез, кетос, метрит или задержка последа), доля заболевших маститом *после* перенесенного заболевания:

- В течении 30 дней с постановки послеотельного заболевания: 9.35%
- В течении 60 дней с постановки послеотельного заболевания: 14.00%
- В течении 90 дней с постановки послеотельного заболевания: 17.9%
- За все время: 35.9%

При этом, коровы перенесшие мастит в течение 90 дней после послеотельного заболевания составляют 32% коров когда-либо болевших маститом, а если брать коров, перенесшие мастит за все время после послеотельного заболевания за все время после заболевания, то доля таких коров составит 65% .

Общая таблица:

	Кетоз	Парез	Метрит	Задержка	Всего
30 дней	9.0%	4.9%	7.5%	7.2%	9.3%
60 дней	13.2%	5.3%	12.2%	10.4%	14.00%
90 дней	16.7%	5.8%	16.4%	13.3%	17.9%
Все время	39.4%	14.2%	36.4%	28.5%	65.3%

Таблица 1: Доля коров заболевших маститом в течение некоторого времени от общего числа коров перенесших данное заболевание.

Если корова болела и парезом и метритом и оба раза после этого заболела маститом это считается и в столбец метрит и в столбец парез

На основании этого, можно сделать вывод что *коровы, перенесшие послеотельное заболевание находятся в группе повышенного риска заболевания маститом*

Склонность коров болеть в целом:

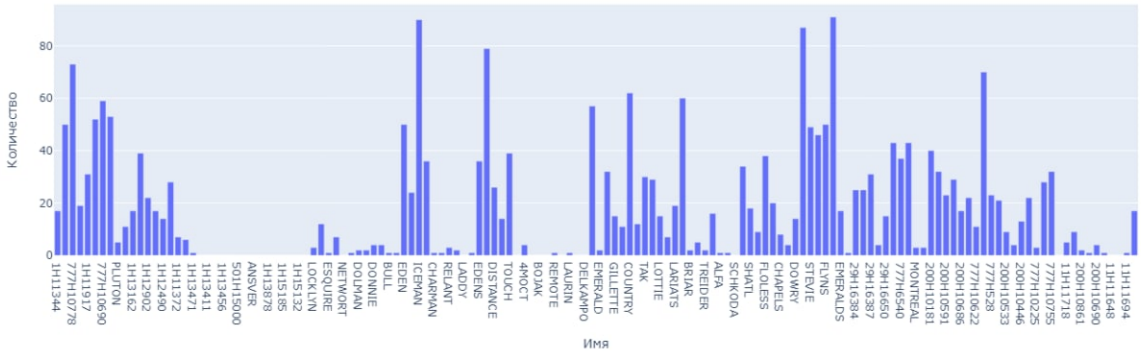
Эти данные позволяют нам оценить грубую вероятность того что корова *будет болеть* маститом при условии что она болела два раза другими болезнями:

$$\mathbb{P}(\text{Маститом} \,|\, \text{Две другие болезни}) = \frac{\mathbb{P}(\text{Две другие болезни} \,|\, \text{Мастит}) \cdot \mu}{\nu}$$

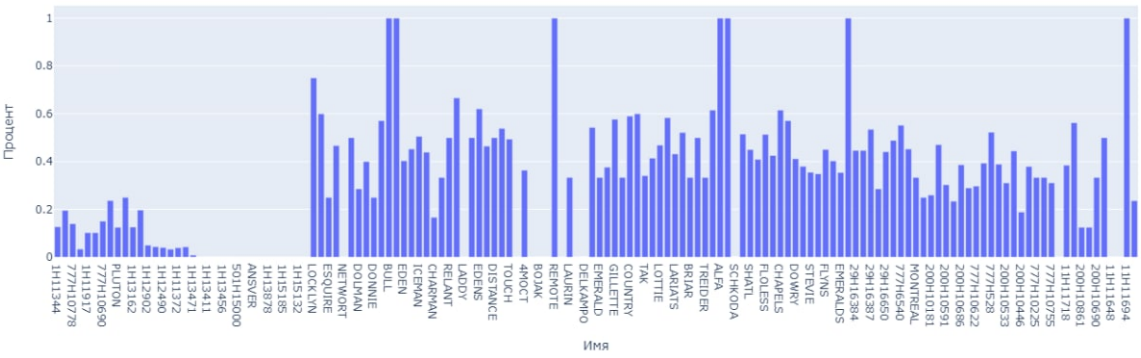
Где μ — доля всех коров, которые болели маститом от общего числа коров, ν — доля числа коров болевших двумя другими болезнями кроме мастита, а $\mathbb{P}(\text{Две другие болезни} \,|\, \text{Мастит})$ это доля коров болевших хотя бы болезнями, среди которых была мастит, просто к доле коров болевших хотя бы тремя болезнями. Искомая вероятность коровы заболеть при условии, что она болела двумя другими болезнями таким образом будет

равна 0.22. Аналогично можно посчитать и для другого любого числа болезней, однако вероятность всегда будет около 0.25

Генетическая предрасположенность к маститу: Мы также решили проверить, зависит ли предрасположенность к маститу от отца коровы. Для этого посчитали, сколько у каждого отца было потомков, которые заболели маститом:



Отнормировав, получаем процент заболевших детей:

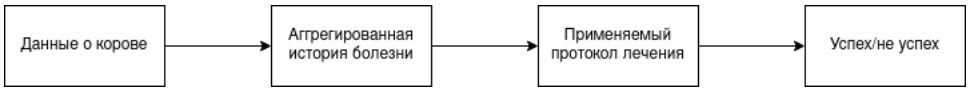


Вывод: неясно — есть некоторые выбросы, но их природа непонятна.

Предсказание эффективности лечения

Оценка эффективности протоколов

Чтобы оценить эффективность тех или иных протоколов, мы агрегировали историю болезни в следующий формат: Таким образом каждая строчка соответствует одному полному циклу лечения. Таким образом, задача



подбора протокола лечения свелась к задаче бинарной классификации (понять поможет протокол или нет). Для оценки эффективность протоколов мы выбрали следующий подход:

- Обучаем LightGBM для решения поставленной выше задачи бинарной классификации
- Выбираем случайное подмножество коров и запускаем на нем обученную модель, применяя к каждой корове поочередно каждый протокол лечения

Протокол	Успешность	Число применений
7	0.77	1742
9	0.79	436
8	0.82	910
2	0.82	626
3	0.83	1054
4	0.86	438
5	0.88	16
6	0.88	296
1	0.910	33

Таблица 2: Эффективность протоколов

- Сортируем протоколы по возрастанию успешных случаев.

Чем такой подход лучше чистого подсчета статистики? Статистический подход не очень чувствителен к неожиданным поведением групп и выбросам, в отличие от машинки. Семплеируя множество раз можно получить более приближенный к реальности результат.

Возникшие проблемы: из-за небольшого объема данных, точность модели после обучения невысока (balanced accuracy score составил 0.557). Поэтому мы решили остановиться на статистическом подходе:

Поиск нетривиальных закономерностей

Мы также придумали метод поиска нетривиальны закономерностей: для этого мы обучаем на *не агрегированной* истории болезни нейросеть-трансформер, однако на вход ей подавалась не прошлая история коровы а будущая. Таким образом, когда нейросеть предсказывает следующие события она выдаёт цепочку событий которые могут привести к заболеванию маститотм.

Однако, мы столкнулись с той же проблемой, что и раньше — нехваткой событий. Если же включать все события (в том числе переводы и взвешивания), то предсказания сети выгялдят примерно так: `ПЕРЕВОД` → `МАСТИТ ВЫДЕРЖКА` → `МАСТИТ`, т.е. каких-то нетривиальных закономерностей сеть не выдаёь