Описание набора данных

Набор данных представляет собой сведения о жизни животных, пребывавших в Austin Animal Center в последние три года.

Каждое животное снабжено уникальным идентификатором,а также для каждого животного известны следующие данные:

1. Имя
2. Дата и время, когда животное покинуло центр
3. Статус, в котором животное покинуло центр
4. Вид животного
5. Пол животного
6. Наличие или отсутствие кастрации на момент завершения пребывания животного в центре
7. Возраст животного на момент окончания пребывания животного в центре
8. Порода животного
9. Окрас животного

Некоторые из приведенных выше признаков требуют более детального описания, которое представлено ниже: статус животного на момент завершения пребывания в центре встречается нескольких типов, а именно:

1. **Adoption** - животное обрело новых хозяев.
2. **Died -** животное умерло.
3. **Euthanasia -** животное было подвергнуто эвтаназии.
4. **Return to Owner -** было возвращено владельцу.

5. **Transfer -** животное было переведено в другой центр содержания животных.

Анализ данных имеет целью установить статус животного на момент завершения пребывания в центре, имея в распоряжении все остальные данные.

Пол животного представлен в виде текста, описывающего как пол, так и наличие или отсутствие кастрации у животного.

Целью исследования данных является получить следующую информацию:

1. Тенденции в выборе животных для содержания
2. Какие параметры, влияющие на вероятность принятия животного на содержание возможно изменить искусственно во время нахождения животного в центре.
3. Животных с какими параметрами следует усиленно представлять людям для принятия на содержание, рекламировать.

Описание преобразований данных

Общие преобразования данных

Следующие данные были преобразованы для большей репрезентативной силы выборки:

1. Имя: из строкового типа данных было преобразовано в булевский, означающий наличие или отсутствие имени у животного на момент завершения пребывания животного в центре.
2. Возраст был преобразован в количество дней, для введения естественного порядка на множестве значений признака.

Генерация новых признаков

1. Из возраста животного, был сгенерирован булевский признак говорящий о том, является ли животное “маленьким”, т.е. котенком или собачкой. Граница возраста, после которого животное уже нельзя считать “маленьким” была установлена в 8 месяцев, исходя из данных о взрослении собак и кошек.
2. Из данных об окрасе животного были получены признаки:
    a. Булевский признак: является ли животное одноцветным.
    b. Булевские признак: является ли животное полосатым (tabby)
    c. Булевские признаки: имеет ли животное цвет белый, черный, любой из “теплых” цветов, любой их “холодных” цветов.

3. Из пола животного выделены два новых признака: непосредственно пол и было ли животное кастрировано.

4. Было испробовано три способа преобразования признака **Color**:
1. Каждый из цветов был отображен в случайное число
2. Цвет не учитывался во все

5. Было испробовано два способа преобразования признака **Breed**:
1. Каждая из пород был отображен в случайное число
2. Порода не учитывалась во все

Преобразования данных для обработки пропусков в данных

У менее чем 25 животных из всей выборки отсутствовала информация о возрасте, ввиду незначительного количества животных с неизвестным возрастом, их было решено исключить из выборки.

Признак **DateTime** был исключен из рассмотрения, так как, репрезентуя лишь время и дату, когда животное покинуло центр содержания, не является значимым для нашей модели.

Пропуски в признаке **SexuponOutcome** были преобразованы в среднее цифровое значение цифр, характеризующих пол.

Преобразования данных для установления зависимости между признаками

Для подсчета коэффициента корреляции Пирсона, данные о статусе, в котором животное покинуло центр были преобразованы в числа, таким образом, что чем больше число, тем успешнее закончилось пребывание животного в центре:

Adoption, **Return_to_Owner** были преобразованы в 1.

Transfer (фактически судьба животного неизвестна) был преобразован в 0.

Died, **Euthanasia** самые плохие исходы были преобразованы в -1.

Установление зависимостей между признаками

Чтобы выявить влияние на наличие каждого из признаков на судьбу животного, решено было посчитать коэффициент корреляции Пирсона, для чего данные о судьбе животного были естественным образом преобразованы в числа, в порядке успешности.

Также для подсчета корреляции в число был преобразован признак **AnimalType**, в виде: **Cat** в 1, **Dog** в 0, таким образом положительная корреляция с признаком **OutcomeType** свидетельствовала бы о более успешной судьбе кошек, в то время как отрицательная - о более успешной судьбе собак. Результат корреляции **-0.22**, указывает на то, что пребывание собак в приюте, как правило, более успешно, чем кошек.

Корреляция признака **Age,** преобразованного в дни с признаком **OutcomeType** показала результат в **0.06**, что указывает на незначительную успешность взрослых животных по сравнению с детенышами.

Используемые методы машинного обучения

Наивный Байесовский классификатор

Вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости данных.

В зависимости наивной вероятностной модели, наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практических приложениях для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью, не веря в байесовскую вероятность и не используя байесовские методы.

Несмотря на наивный вид и, несомненно, очень упрощенные условия, наивные байесовские классификаторы часто работают намного лучше во многих сложных жизненных ситуациях.

Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации.

В наших экспериментах используются 3 вида наивного Байесовского классификатора:

1. С функцией правдоподобия в виде функции Гаусса
2. Со сглаженной функцией правдоподобия
3. С оценкой параметров с помощью правила Бернулли

Решающие деревья

Решающие деревья или Дерево “принятия решений” Структура деревьев представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Одним из преимуществ решающих деревьев является тот факт, что принятые алгоритмом решения могут быть наглядно объяснены с помощью визуализации построенного дерева, что и было сделано в ходе выполнения работы.

Случайные леса

Алгоритм машинного обучения, предложенный Лео Брейманом и Адель Катлер, заключающийся в использовании комитета решающих деревьев. В ходе его выполнения несколько решающих деревьев строятся независимо друг от друга, основываясь на случайно выбранной подвыборке данных из исходного массива. При принятии окончательного решения к какому классу отнести объект построенные деревья голосуют, влияя на ответ решениями, полученными самостоятельно.

Многоклассовая логистическая регрессия

Используется для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой. Подбор параметров основан на максимизации функции правдоподобия методом градиентного спуска или методом Ньютона.

Оценки качества алгоритма

Оценка качества классификации

Для оценки качества используемых классификаторов, была использована метрика под названием **accuracy score,** вычислении которой происходит по формуле:

$$A = \frac{NP}{S}$$

Где **A -** значение метрики, **NP -** кол-во верно предсказанных исходов, **S** количество всех сделанных предсказаний.

Результаты метрики **accuracy score** для алгоритмов классификации точного предсказания типа:

	Gaussian Naive Bayes	Multinomial Naive Bayes	Bernulli Naive Bayes	Logistic Regression	Extra Trees	Random Forest	Decision Tree
Without **Color** and **Breed**	0.6	0.53	0.58	0.63	0.64	0.65	0.64

With **Color** and **Breed**	0.6	0.17	0.58	0.62	0.57	0.57	0.56
Cat/Dog	0.73/0.68	0.6/0.53	0.7/0.67	0.68/0.63	0.67/0.57	0.7/0.6	0.7/0.62

Оценка качества регрессии

Для оценки качества используемых регрессоров была использована метрика под названием **Multiclass Loss.** Логарифмическая функция потерь широко распространена для оценки качества работы алгоритмов, решающих задачу многоклассовой регрессии. Определяется она как отрицательное логарифмическое правдоподобие того, что объект принадлежит к предсказанному алгоритмом классу. В англоязычной литературе функция также известна как **Cross Entropy Loss.**

Вычисляется значение функции по следующей формуле:

$$V(f(\vec{x}), t) = -t \ln(f(\vec{x})) - (1 - t) \ln(1 - f(\vec{x}))$$

Где $f(\vec{x})$ является результатом предсказания алгоритма для некоторого вектора данных, характеризующего объект.

$\vec{x}$ - вектор данных.

t - достоверно известный класс объекта, который характеризуется вектором данных $\vec{x}$.

V - Значение функции потерь для конкретного объекта и его наблюдаемых признаков.

Приведенная выше формула дает выражение значения логарифмической функции потерь для одного объекта, применительно к множеству, значения этой функции складывают и берут среднее.

Результаты метрики **Multiclass Loss** для алгоритмов регрессии точного предсказания типа:

	Logistic Regression	Extra Trees	Random Forest	Decision Tree
Without **Color** and **Breed**	0.92	1.16	1.12	1.18
With **Color** and **Breed**	0.92	6	4	11
Cat/Dog	0.77/0.89	0.83/0.93	0.97/0.99	0.94/0.99

Вывод

Исходя из результатов, представленных выше, для максимально точного предсказания судьбы животного, попавшего в центр, наилучшим методом является алгоритм многоклассовой логистической регрессии с использованием L2 нормы штрафов, примененный без использования данных о цвете и породе животного, которые только уменьшают точность алгоритма, являясь причиной переобучения.

Значительно увеличивает оценку точности алгоритма отдельное обучение алгоритма для кошек и собак, что логично вытекает из разного подхода людей к принятию на содержание разных видов животных.

При использовании методов предложенных в данной работе можно точно предсказать около 73% исходов для котов и около 68% исходов для собак.