МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное
учреждение высшего образования
«Южный федеральный университет»

Институт математики, механики
и компьютерных наук им. И. И. Воровича

Кафедра информатики и вычислительного эксперимента

Волнобой Ирина Леонидовна

РАЗРАБОТКА КОМПОНЕНТОВ ПРИЛОЖЕНИЯ ДЛЯ АНАЛИЗА ОНЛАЙН-ПРОФИЛЯ ЖИВОТНОГО ИЗ ПРИЮТА

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по направлению подготовки
02.03.02 – Фундаментальная информатика и информационные технологии

Научный руководитель –
доцент., к. ф.-м. н. Абрамян Анна Владимировна

Допущено к защите:
заведующий кафедрой _____ Михалкович С. С.

Ростов-на-Дону – 2021

Оглавление

Постановка задачи

В постановке задачи коротко (по пунктам) указывается, что необходимо сделать в рамках работы. Раздел «Постановка задачи» должен соответствовать заданию на курсовую или выпускную квалификационную работу, подписанному научным руководителем.

Введение

Машинное обучение является одним из подразделов искусственного интеллекта. Оно нашло применение во многих сферах жизни: маркетинге, бизнесе, медицине, в банковской сфере, в различных научных исследованиях. Машинное обучение помогает в решении каких-либо вопросов, помогает принять решение, что нужно улучшить, чтобы увеличить или уменьшить какие-либо показатели и достигнуть цели.

Например, в данной работе требуется на основе данных о питомцах предсказать, с какой скоростью животное будет принято в семью, а также какие признаки влияют на принятие решения в большей степени. Данная задача предложена к решению малайзийским сайтом petfinder.my. Сама задача, а также все необходимые данные представлены на ресурсе Kaggle [1]. Скорость в данной задаче является категориальной переменной, поэтому необходимо решить задачу классификации. Задача классификации является разновидностью задачи обучения с учителем и решается с помощью методов машинного обучения.

Для решения данной задачи необходимо проанализировать исходные данные, выбрать наиболее подходящие для задачи методы обработки, применить различные алгоритмы машинного обучения, подобрать параметры, а также сравнить их качество и выбрать наиболее подходящий алгоритм, который показывает наилучшее качество на выбранной метрике.

В данной работе использовался язык программирования Python версии 3.8.2, библиотеки для визуализации matplotlib, seaborn, graphviz, библиотеки для обработки и анализа данных numpy и pandas, а также бибилотеки предоставляющие функционал для предварительной обработки данных и тренировки алгоритмов машинного обучения sklearn, XGBoost, LightGBM. В качестве инструмента для разработки использовался JupyterNotebook.

В качестве предварительной обработки данных использовались следующие методы:

- пропуски в данных заменены выборочным значением
- выбросы заменены выборочным значением или введена новая пе-

ременная, оценивающая исходную

- созданы новые признаки на основе имеющихся
- извлечены и обработаны признаки из данных, полученных от Google's Natural Language API и от Google's Vision API
- применены методы кодирования категориальных переменных: прямое кодирование, One Hot Encoding и Label Encoding
- данные приведены к одной шкале с использованием StandardScaler из sklearn

Создатели задачи рекомендуют к использованию метрику Quadratic Weighted Kappa, поэтому именно она используется для оценки качества моделей.

Для предсказания класса принятия на основе признаков были использованы алгоритмы логистической регрессии, дерева решений, случайного леса, а также три алгоритма градиентного бустинга из библиотек sklearn, XGBoost и LightGBM. Все модели и способы обработки были оценены и выбран наиболее оптимальный для данной задачи.

Также был использован алгоритм логистической регрессии с алгоритмом оптимизации стохастического градиентного спуска для предсказания целевой переменной на основе только текстовых признаков (описаний животных). Для этого описания животных были предварительно обработаны. Выполнена токенизация, нормализация, стемминг и лемматизация, векторизация методами Bags of Words и TF-IDF. Выполнен анализ эффективности модели на основе данных методов обработки текстов.

Настройка гиперпараметров моделей происходила по сетке с использованием GridSearchCV и RandomizedSearchCV из библиотеки sklearn.

Обзор

1. Анализ исходных данных и их предварительная обработка

1.1. Описание поставленной задачи и исходных данных

В данной работе использовались данные представленные на ресурсе Kaggle. Это два основных датасета, представленных в формате csv:

- train.csv — содержит данные для тренировки модели размерностью 14993 записи на 25 столбцов.
- test.csv — содержит данные для тестирования размерностью 3972 записи на 24 столбца. Данный датасет предназначен для тестирования модели на ресурсе Kaggle.

В train.csv на один столбец больше, чем в test.csv, так как тренировочный датасет содержит столбец AdoptionSpeed, который необходимо предсказать в тестовой выборке.

Так же имеется 3 csv файла, являющиеся словарями:

- breed_labels.csv — словарь пород
- color_labels.csv — словарь окрасов шерсти
- state_labels.csv — словарь штатов

В словарях содержатся 2 колонки: id и расшифровка. Эти словари необходимы для лучшего понимания данных, так как в тренировочном и тестовом датасетах в колонках содержатся именно id пород, окрасов и штатов, а сами названия находятся в словарях.

Ещё имеются папка, содержащая метаданные, полученные с помощью Google's Vision API с изображений, и папка, содержащая анализ тональности описаний, представленных в виде текста, который был получен с использованием Google's Natural Language API. Эти данные представлены в формате PetID.json.

Вся информация, представленная в датасетах была взята создателями соревнования Kaggle с малайзийского сайта petfinder.my (рис. 1). Это

ресурс, главная задача которого состоит в том, чтоб найти новый дом животным. Анкеты животных на сайт выкладывают либо приюты, либо отдельные люди, которые нашли животное на улице. Несмотря на то, что на самом сайте представлены профили различных видов животных (попугаев, кошек, собак, хомячков, кроликов и так далее), в датасетах содержится информация только о профилях кошек и собак. На этом сайте люди имеют возможность отдать животное в хорошие руки бесплатно или же за определенную плату. Для этого создаётся профиль животного и добавляется туда информация о животном: фотография, кличка, возраст, пол, описание и другое.

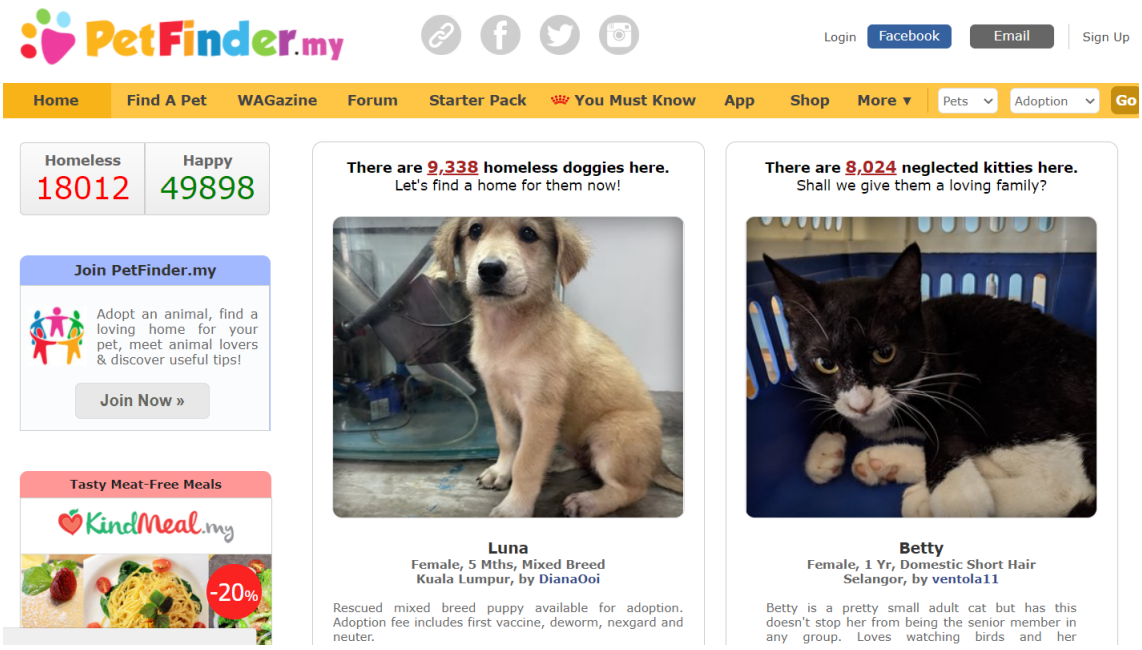

Рис. 1. Скриншот сайта petfinder.my

1.2. Описание признаков

Всего в датасете присутствует 24 признака. Посмотрим на матрицу корреляции (рис. 2). Наибольший коэффициент корреляции имеют переменные Dewormed и Vaccinated — 0,72. Но этот коэффициент недостаточно высок для того, чтобы утверждать, что данные переменные линейно зависимы. Поэтому нельзя выбрасывать из рассмотрения ни одну из них. Остальные переменные не коррелируют между собой.

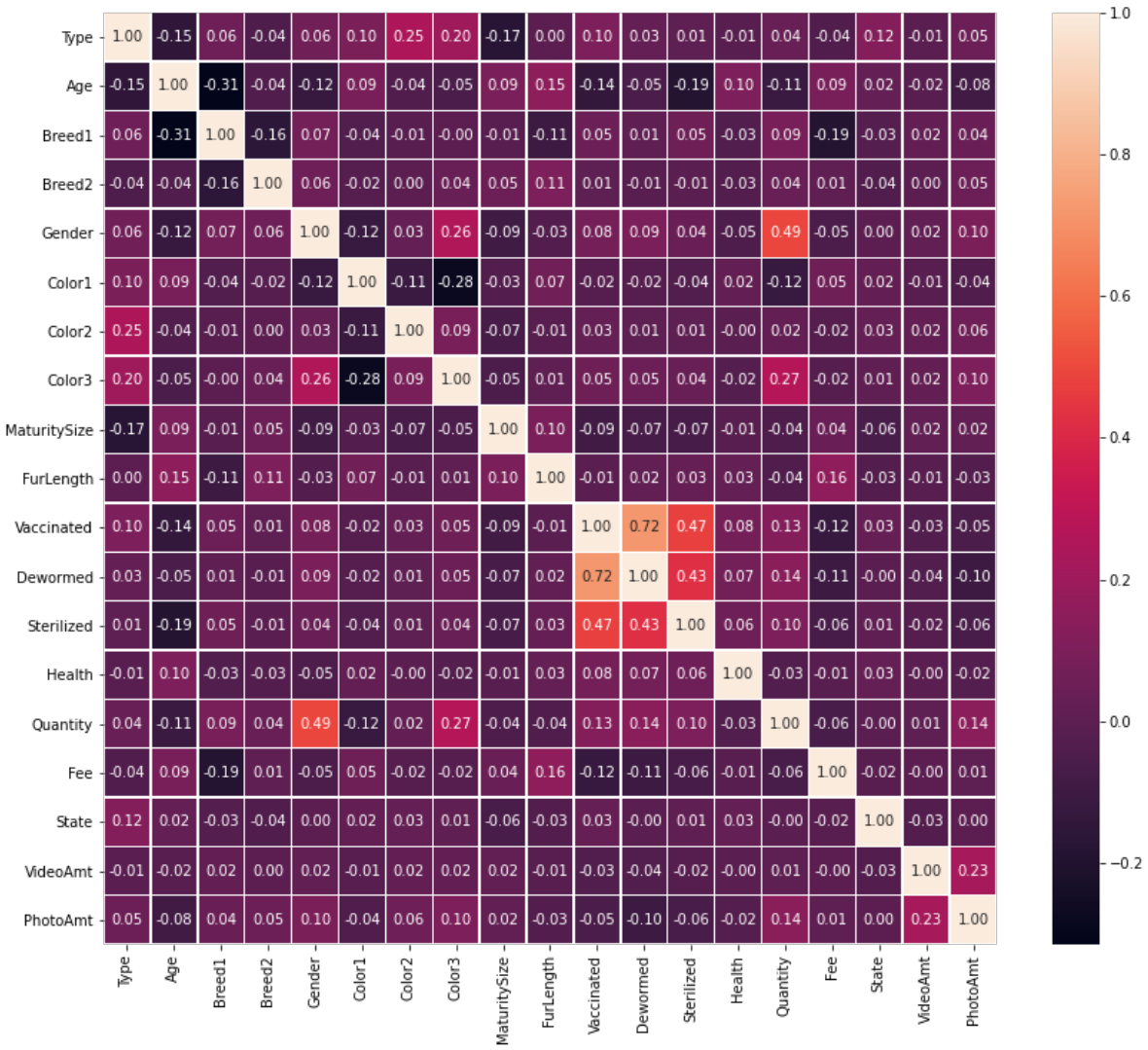

Рис. 2. Матрица корреляции переменных

Рассмотрим более подробно каждый признак.

PetID — это уникальный идентификатор питомца. Так как каждый питомец имеет свой идентификатор, то не имеет смысла использовать данный признак для обучения модели.

Type — тип животного. Это категориальная переменная, принимающая два значения: кошка или собака. Количество собак составляет 8132 особи, а кошек — 6861 (рис. 3).

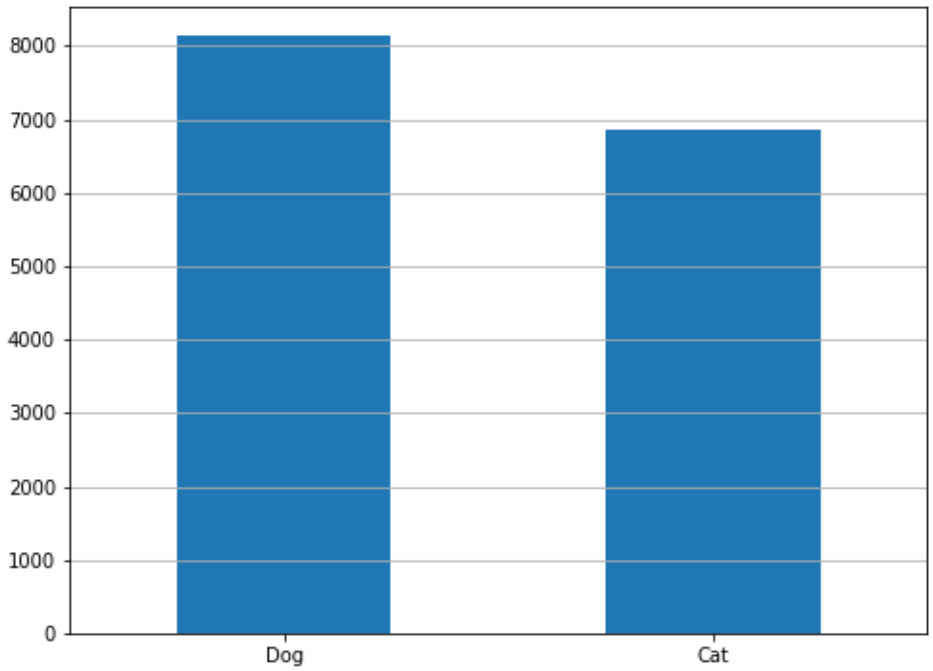

Рис. 3. Распределение кошек и собак в тренировочной выборке

Name — имя питомца. На рисунке 4 изображены наиболее популярные клички среди кошек и собак. Можно заметить, что большинство людей не придумывали животным имена, а просто называли их “Kitten”, “Puppy” или “No name”. Также была часть животных, имена которых состояли из 1, 2 или 3 символов и не имели смысла. Например, “BB”, “C7C”, “Z4”.

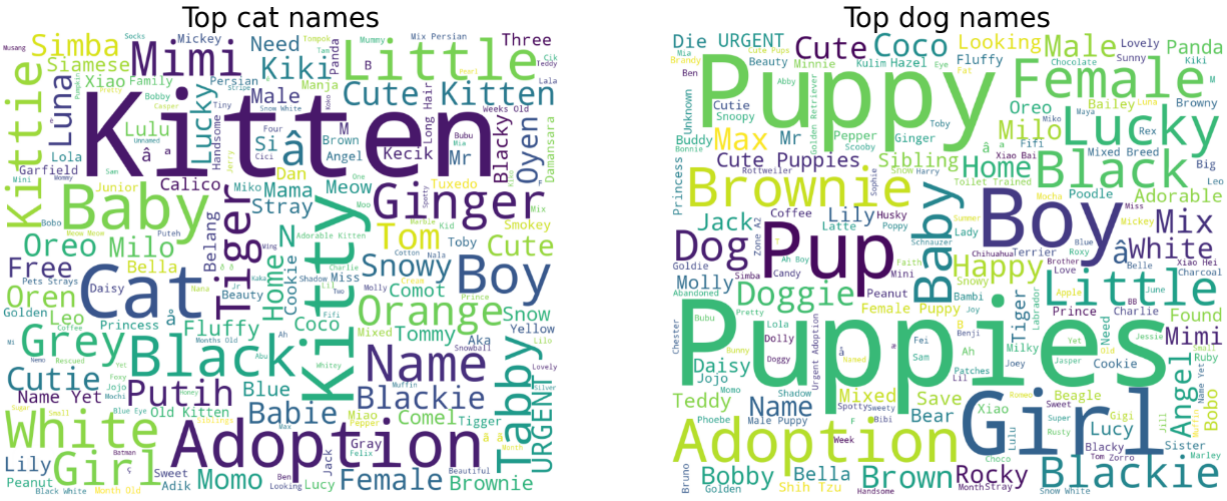

Рис. 4. Наиболее распространённые имена среди кошек и собак

Age — возраст питомца в месяцах.

Breed1 и Breed2 — переменные, содержащие идентификатор породы, который ссылается на словарь пород breed_labels.csv. Если питомец чистокровной породы, то в Breed 2 стоит идентификатор 0.

Gender — категориальная переменная, содержащая пол питомца. Может принимать 3 значения: male, female и mixed. Mixed ставится в том случае, когда в профиле питомцев больше одного. Наибольшее число питомцев имеет мужской пол, наименьшее — смешанный (рис. 5)

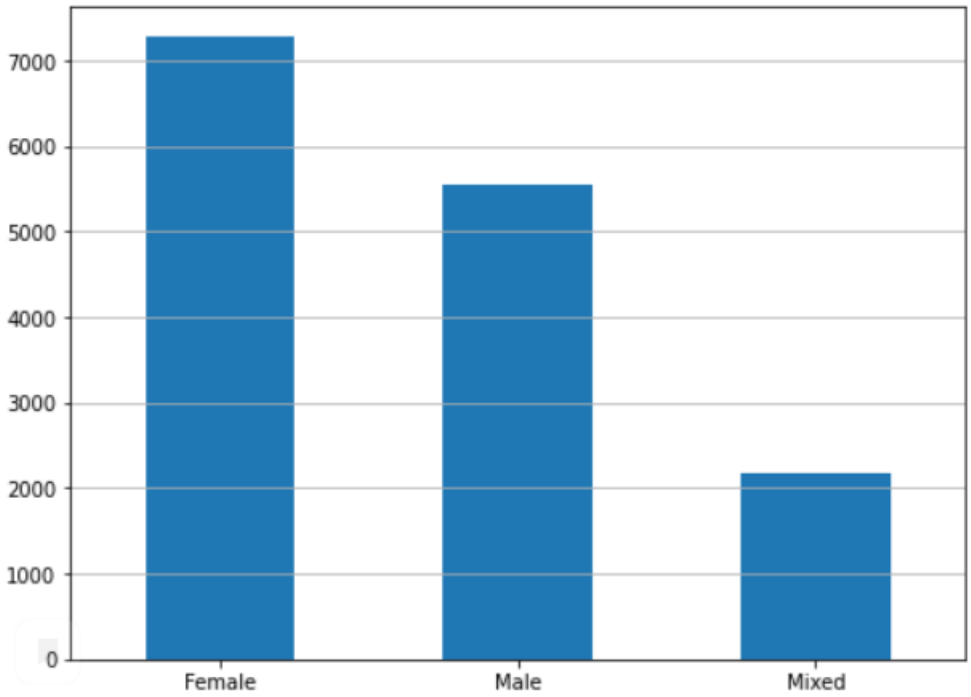

Рис. 5. Распределение питомцев по полу в train.csv

Переменные Color1, Color2 и Color3 содержат идентификаторы окраса, расшифровка которых находится в файле color_labels.csv. Если питомец имеет всего один цвет, то в переменных Color2 и Color3 стоит значение 0. Наиболее часто встречаются питомцы коричневого, черно-коричневого и черно-белого цветов (рис. 6).

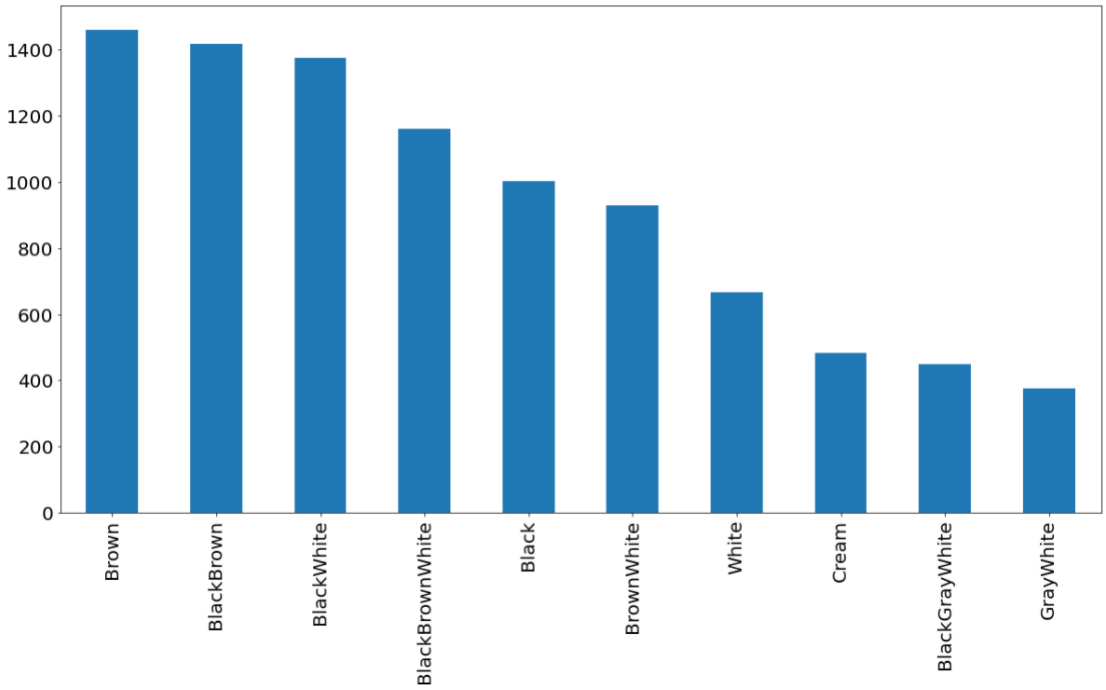

Рис. 6. Наиболее распространённые комбинации окрасов

MaturitySize и FurLength — категориальные переменные, обозначающие размер в зрелом возрасте и длину шерсти соответственно. MaturitySize принимает 4 значения: “small”, “medium”, “large” и “extra large”, а FurLength принимает 3 значения: “short”, “medium”, “long”. Наиболее распространенный размер животного в зрелом возрасте — средний, а наиболее популярная длина шерсти — короткая (рис. 7).

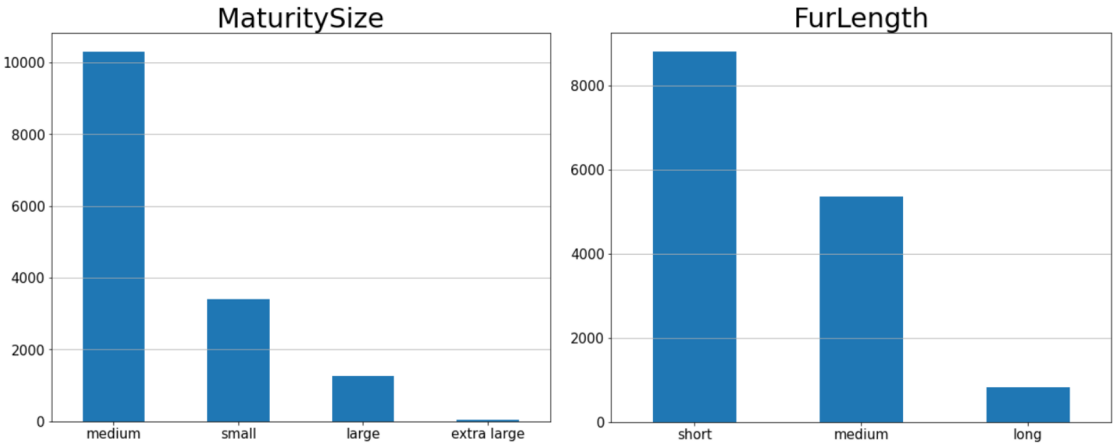

Рис. 7. Наиболее распространённые комбинации окрасов

Также есть 4 категориальных признака, относящихся к здоровью пи-

томца:

- Vaccinated (вакцинировано)
- Dewormed (избавлено от гельминтов)
- Sterilized (кастрировано)
- Health (общее состояние здоровья)

Vaccinated, Dewormed и Sterilized принимают 3 значения: “да”, “нет” и “неизвестно”. Переменная Health также принимает 3 значения: “Healthy”, “Minor Injury”, “Serious Injury”. Большинство животных здоровы, и лишь небольшая часть имеет травмы (рис. 8).

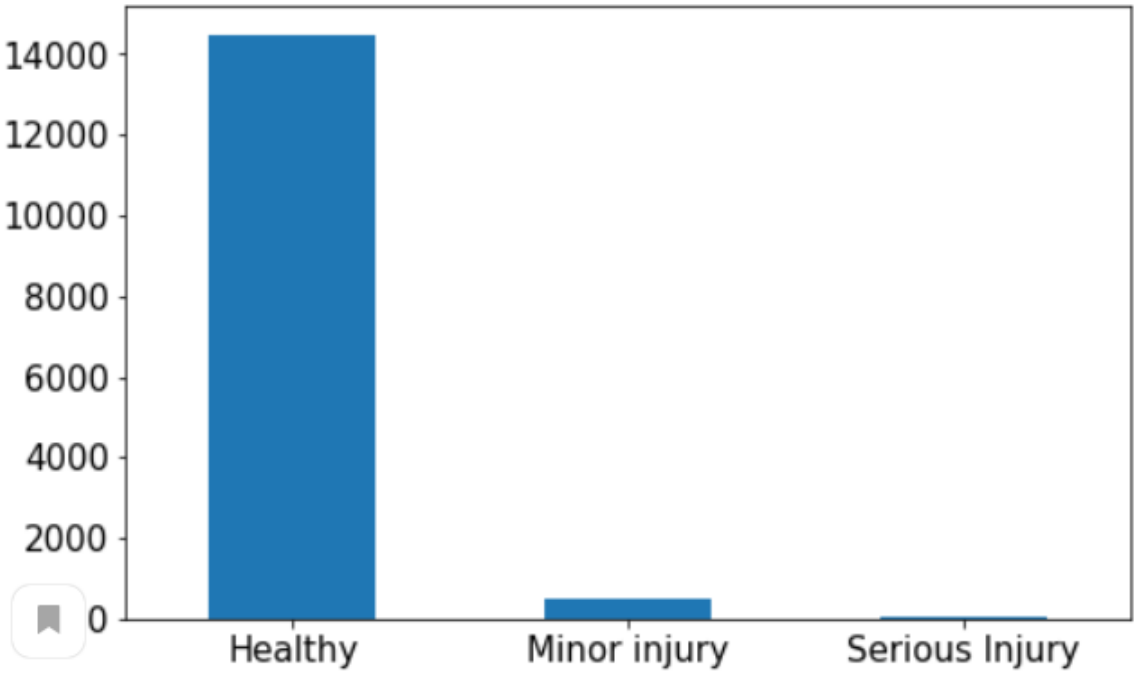

Рис. 8. Состояние здоровья питомцев

Quantity — количество животных в одном профиле. Наибольшее число профилей (11565) содержит одно животное, но также есть записи, где имеется и по 5, и по 10, и даже по 20 животных (рис. 9).

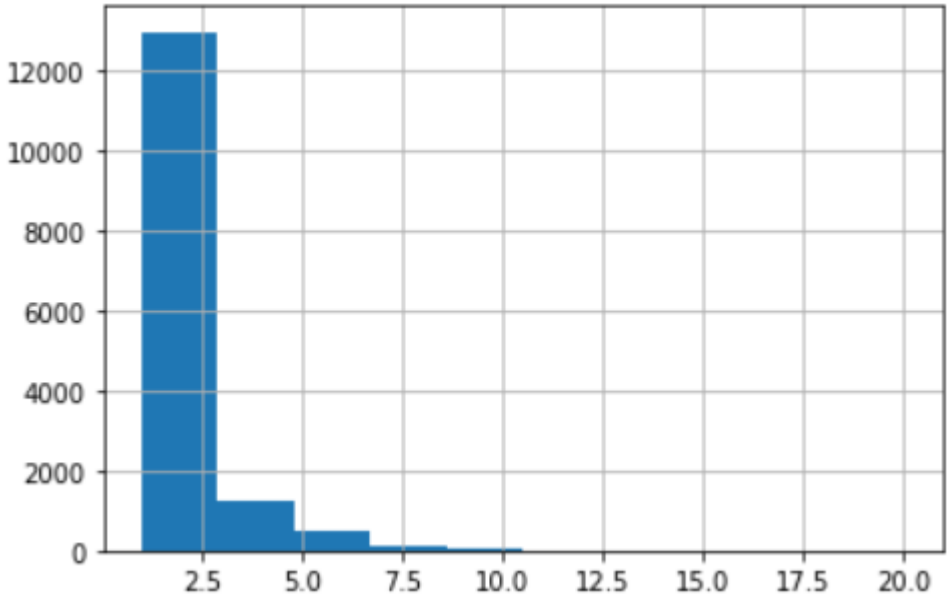

Рис. 9. Распределение количества животных в одном профиле

Fee — стоимость питомца. В тренировочном датасете 12663 животных отдают бесплатно и 2330 за определенную плату. Причем плата может быть как 10 малайзийских ринггитов, так и 3000 (рис. 10). За плату чаще отдают животных, которые являются чистопородными. Например, за 3000 отдают немецкую овчарку, за 2000 английского бульдога, за 750 персидскую кошку. Но также есть беспородные животные, которых отдают за плату, например, домашнюю длинношерстную кошку за 10 малайзийских ринггитов.

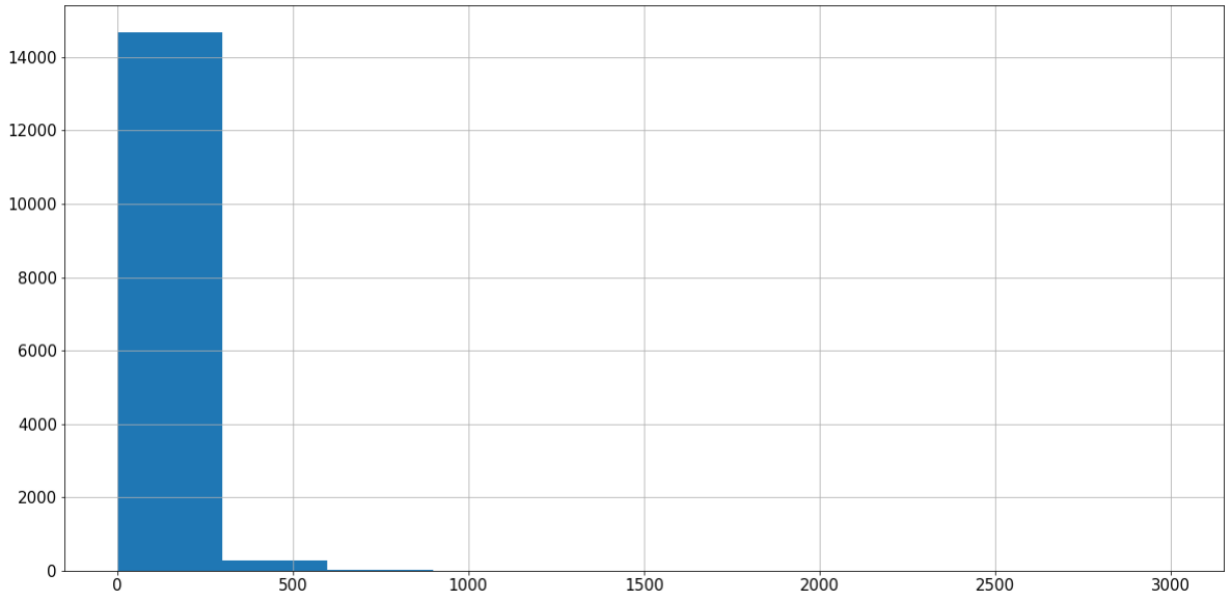

Рис. 10. Гистограмма стоимости питомца

State — штат в Малайзии, в котором находится питомец. Интересно, что 97% питомцев находится в 6 штатах, а именно в Selangor, Kuala Lumpur, Pulau Pinang, Johor и Perak (рис. 11).

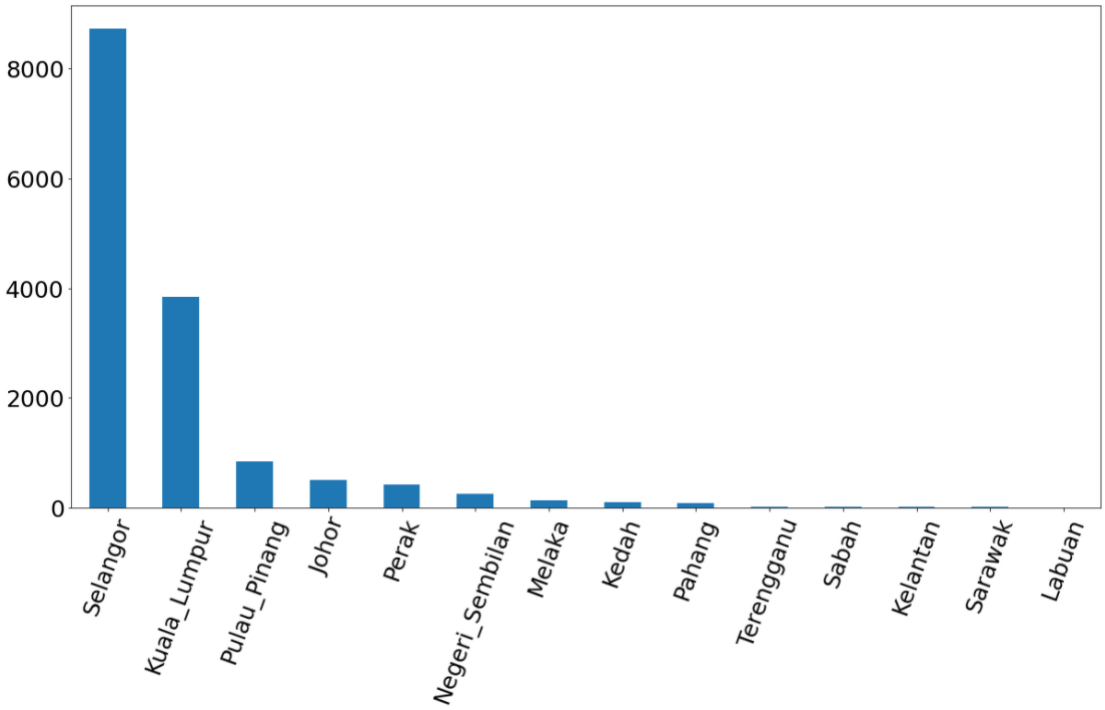

Рис. 11. Количество питомцев в штатах Малайзии

RescuerID — это идентификатор пользователя, который выкладывает

профиль животного на сайт. Есть несколько людей или организаций, которые выложили достаточно много объявлений (рис. 12). Наибольшее число профилей составило 459.

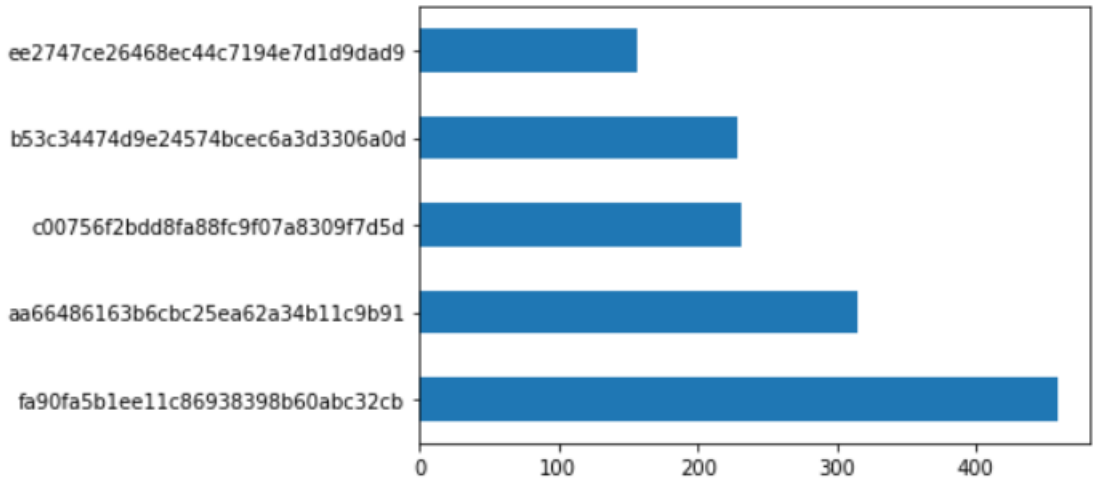

Рис. 12. Количество профилей, которые выложили пользователи

VideoAmt и PhotoAmt содержат количество видео и фото, которые содержатся в профиле питомца. Переменная Description хранит текстовое описание профиля. Основной используемый язык — английский.

AdoptionSpeed — это целевая категориальная переменная, которую необходимо предсказать. Всего имеется 5 классов:

- 0 — питомца забрали в первый же день, как профиль был создан
- 1 — питомца забрали в период от 1 до 7 дней после создания профиля
- 2 — питомца забрали в период от 8 до 30 дней после создания профиля
- 3 — питомца забрали в период от 31 до 90 дней после создания профиля
- 4 — питомец не был принят в семью после 90 дней ожидания

Для профилей, в которых несколько животных, скорость принятия в семью определяется как скорость, с которой все животные были приняты.

AdoptionSpeed имеет сильный дисбаланс классов (рис. 13). Количество животных, которых приняли в первый же день (0 класс), составляет

410 особей. А наибольшее число питомцев (4197 особей) находится в 4 классе. Таким образом, количество питомцев в наибольшем классе превышает количество питомцев в наименьшем классе в 10 раз.

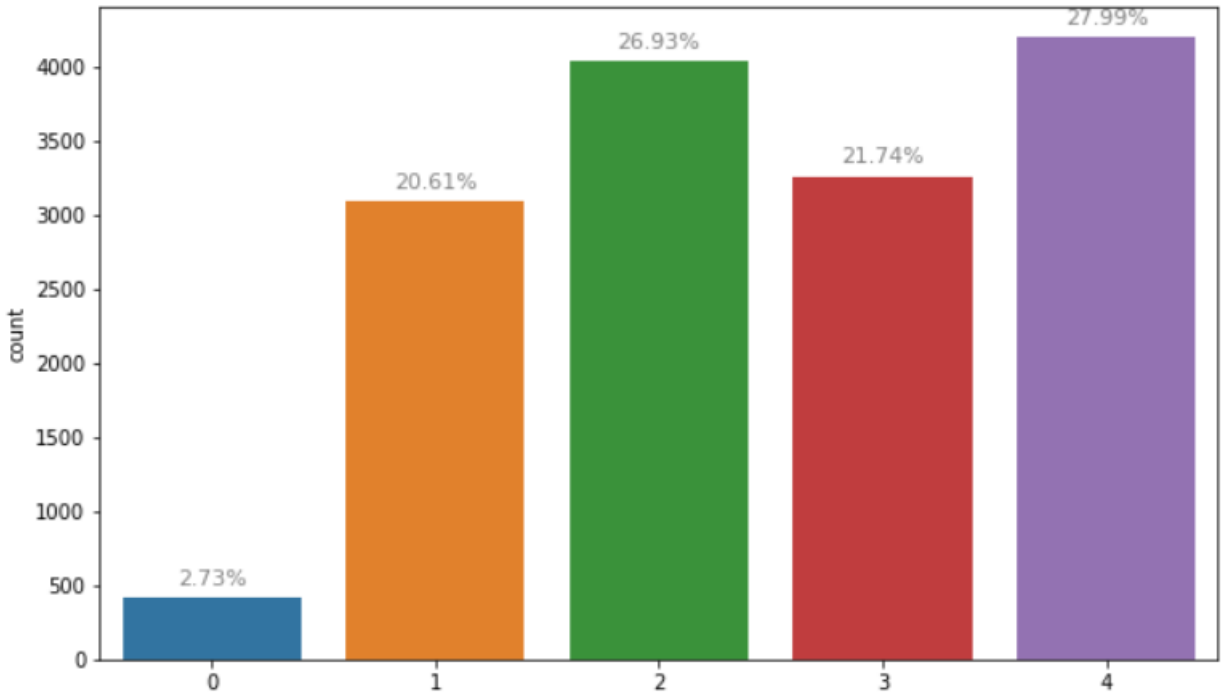

Рис. 13. Количество питомцев в каждом из классов

Для получения корректного качества модели все признаки необходимо предобработать, а именно обработать пропущенные значение, найти и обработать выбросы, извлечь признаки из имеющихся данных, закодировать категориальные переменные, привести данные к одной шкале.

1.3. Обработка пропущенных значений

В датасете всего 2 переменные содержат пропущенные значения: Name и Description (рис. 14).

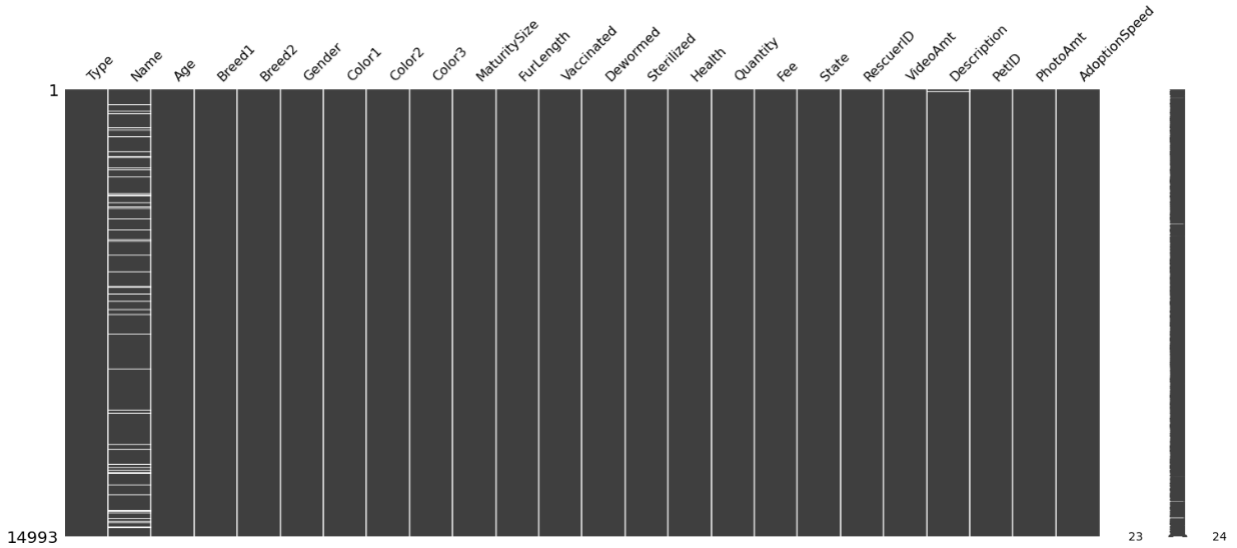

Рис. 14. Пропущенные значения в данных

Переменная Name содержит 1257 пропусков. Это составляет 8% от всего датасета. Переменная Description содержит 12 пропущенных значений, что составляет 0.08% от датасета. Пропущенные значения необходимо обрабатывать, так как не все модели способны работать с ними. Есть несколько способов борьбы с пропущенными значениями [2]:

- Удалить строки, содержащие пропущенные значения
- Заменить пропуски выборочным значением
- Заменить пропуски средней/медианой/модой
- Заполнить случайным значением

Удаление пропущенных строк не подходит, так как при использовании данного метода значительно сокращается объём датасета и, следовательно, теряется часть информации. Третий и четвертый метод также не подходят, потому что переменные Name и Description имеют строковый тип. Поэтому остаётся только заменить пропуски выборочным значением. Пропуски в Name были заменены значением ‘No_name’, а в Description пустой строкой. В дальнейшем данные переменные будут дополнительно преобразованы.

1.4. Детекция и обработка выбросов

Выбросы — это наблюдения, сильно отличающиеся от остальных наблюдений в выборке. Выбросы необходимо обрабатывать, так как алгоритмы машинного обучения чувствительны к диапазону и распределению переменных [3]. Наличие выбросов в данных может привести к увеличению времени обучения, а также к снижению точности.

Для автоматического обнаружения выбросов в данных был использован метод межквартильного расстояния [4].

Выбросами в данном случае считаются значения, которые не попадают в диапазон $[Q1 - 1,5 \times (Q3 - Q1), Q3 - 1,5 \times (Q3 - Q1)]$, где $Q1$ и $Q3$ — первый и третий квартиль соответственно.

Методы обработки выбросов аналогичны методам обработки пропущенных значений.

В переменной Age достаточно много выбросов (рис. 15). Максимальное значение возраста составило 255 месяцев. Это больше 21 года, что для кошки и собаки является достаточно большим возрастом. С помощью метода межквартильного расстояния нашли 1501 выброс и заменили их выборочным значением 27, то есть верхней границей полученного диапазона $[-13,\ 27]$.

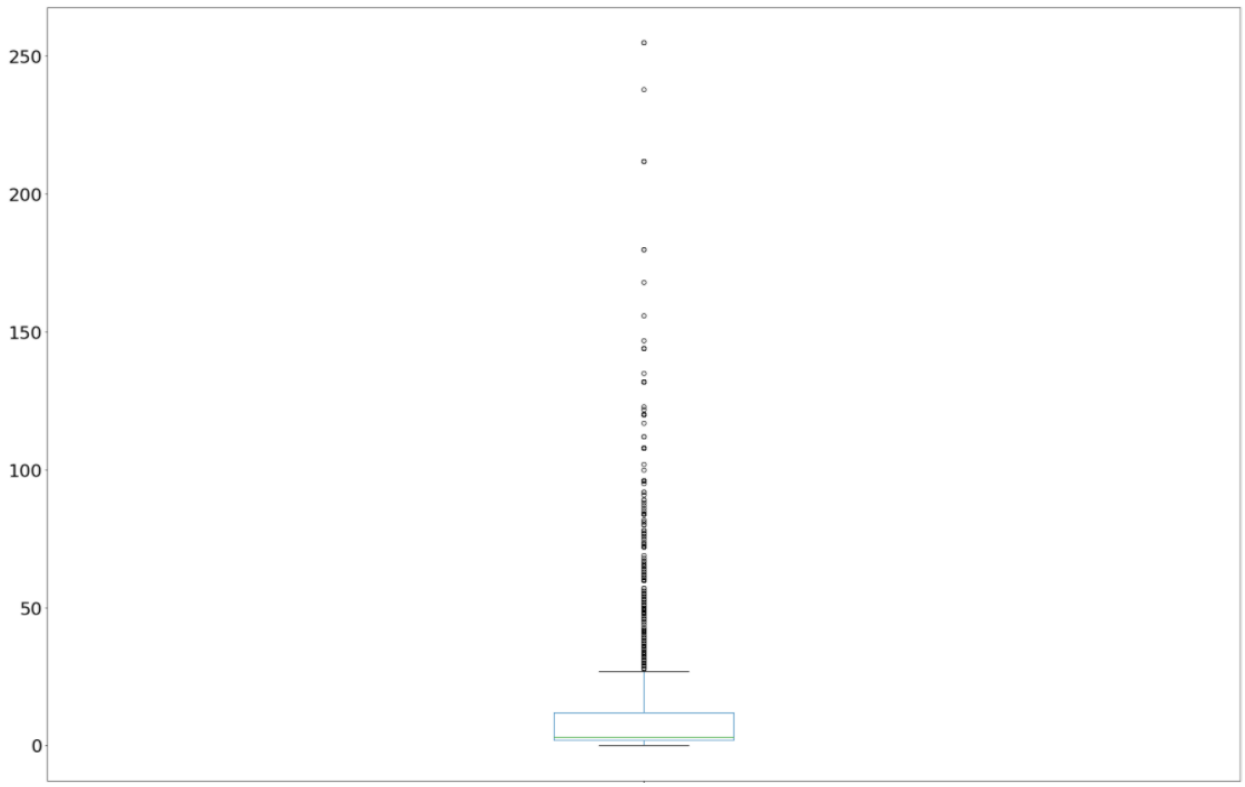

Рис. 15. Выбросы в переменной Age

Аналогично для переменной PhotoAmt было найдено 922 выброса, которые были заменены выборочным значением 9 (рис. 16).

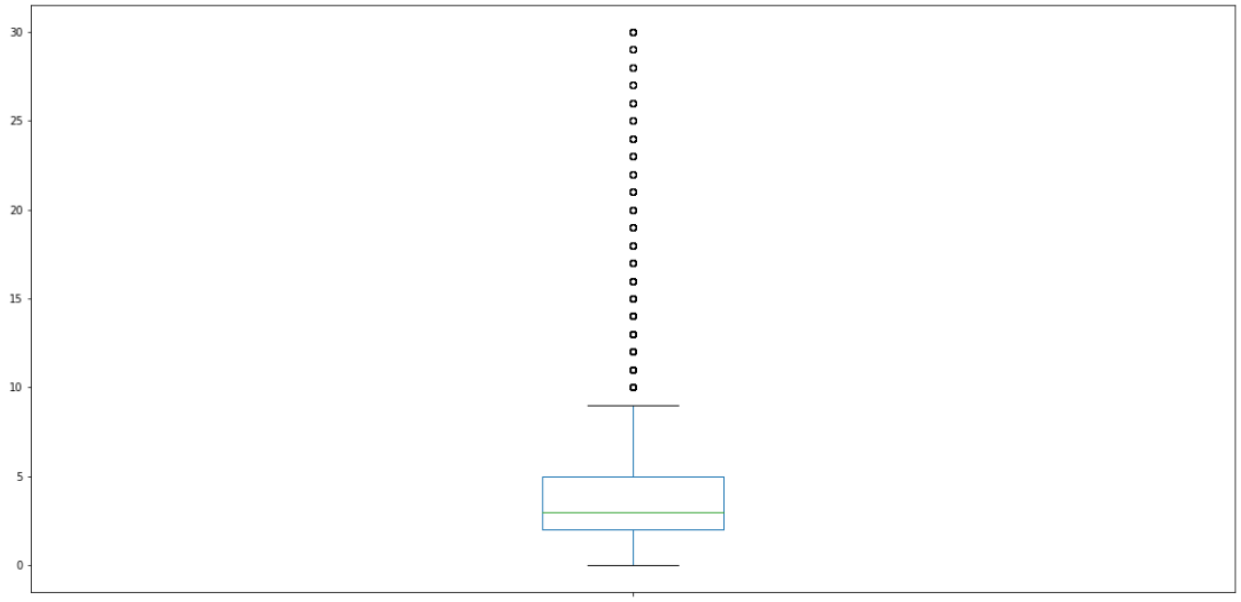

Рис. 16. Выбросы в переменной PhotoAmt

Также выбросы наблюдаются в переменных Quantity и Fee. Особен-

ностью данных переменных является то, что большая часть значений находится в 1 для Quantity и в 0 для Fee. Из-за этого на графике boxplot (рис. 17) среднее, медиана, нижняя и верхняя границы, первый и третий квартиль сливаются в одну линию. Если заменить все выбросы верхней границей диапазона (для Quantity — [1, 1], для Fee — [0, 0]), то переменная станет константной и не будет иметь значения для обучения модели. Поэтому для обработки данных переменных выбраны две стратегии:

- Замена части выбросов выборочным значением
- Создание новой переменной, которая оценивает исходную

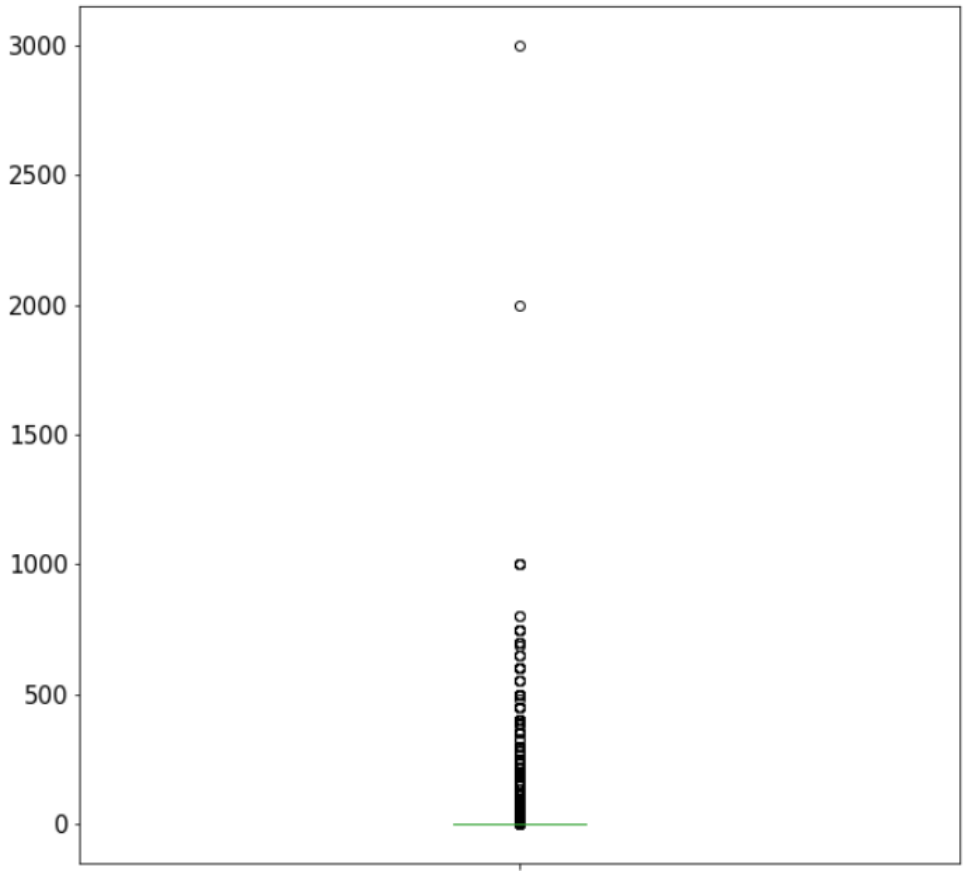

Рис. 17. Выбросы в переменной Fee

При использовании первой стратегии в переменной Quantity заменена часть выбросов, у которых значение больше 5, значением 5, а в переменной Fee заменены значения, превышающие 500, значением 500.

При использовании второй стратегии для Quantity создана новая переменная one_pet, которая имеет значение 1, если в профиле одно живот-

ное, и 0, если животных несколько. Аналогично для переменной Fee создана переменная Free, имеющая значение 1, если животное отдают бесплатно, иначе 0.

Таким образом, обучение будет происходить на двух датасетах, в одном из которых заменены переменные новыми, оценивающими значения, а в другом заменена часть выбросов выборочным значением.

1.5. Создание новых признаков из имеющихся данных

Переменная Name имеет строковый формат, и, так как модели машинного обучения не умеют работать со строковым типом, то необходимо эту переменную преобразовать. Для этого заменим признак Name новой переменной No_name, которая будет обозначать, есть ли у животного реальное имя. На этапе обработки пропущенных значений всем значениям равным NaN было поставлено значение 'No_name'. Также в поле Name есть значения, которые обозначают отсутствие имени. Например, "unnamed", "nameless", "no name yet". Ещё есть очень короткие имена, состоящие из 1–3 символов и не имеющие смысла. Все эти значения также были заменены на 'No_name'. Затем новой переменной No_name присваиваем значение 1, если в Name стоит значение 'No_name', иначе ставим 0. Таким образом, вместо строковой переменной Name получили булеву переменную No_name (рис. 18), которую и будем использовать в обучении.

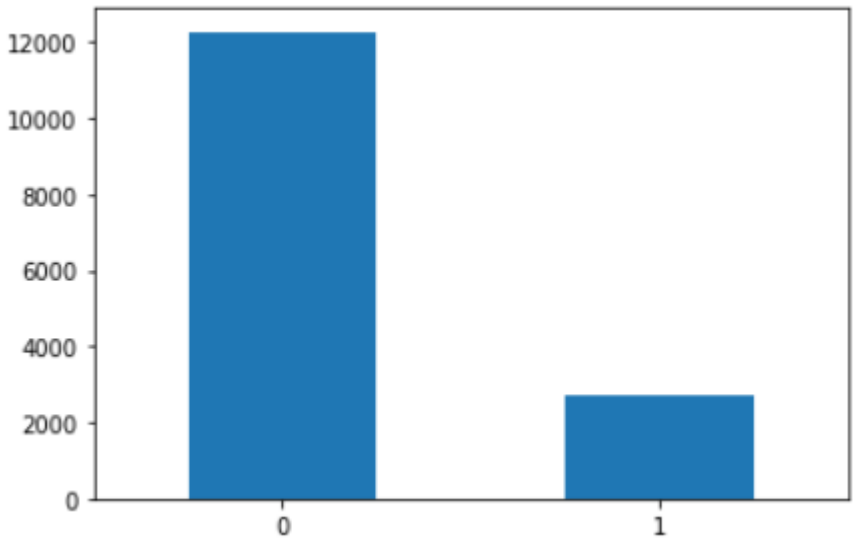

Рис. 18. Число питомцев, имеющих имя и не имеющих

Из переменных Breed1 и Breed2 была создана новая переменная Pure_breed (рис. 19), обозначающая, является ли животное породным или беспородным. Породным считалось животное, которое в Breed2 имеет значение 0 (то есть нет расшифровки в словаре) и в Breed1 не имеет значения ‘Mixed_Breed’, ‘Domestic_Long_Hair’, ‘Domestic_Medium_Hair’ или ‘Domestic_Short_Hair’, так как данные виды не считаются породистыми.

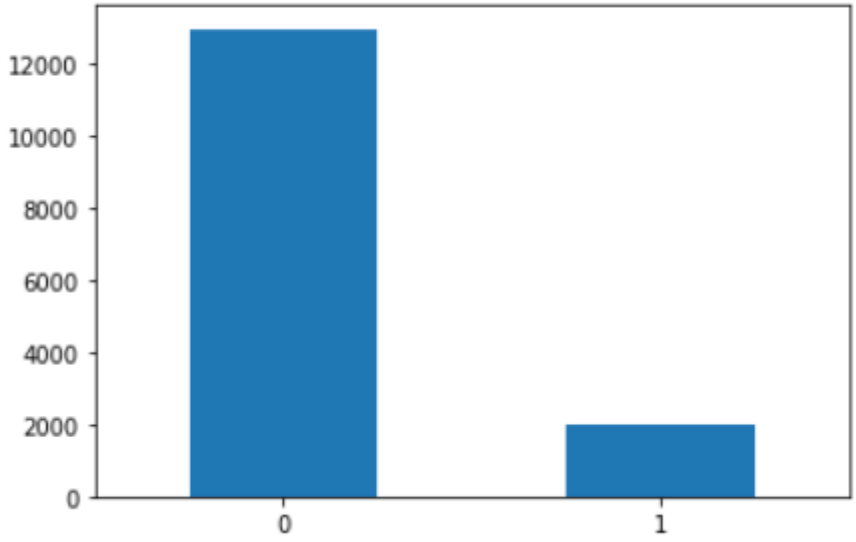

Рис. 19. Число породных и беспородных питомцев

Переменная RescuerID содержит идентификатор людей или органи-

заций, которые создают профиль питомца на сайте, а также отдают его. Есть идентификаторы, которые создали достаточно много профилей (рис. 20).

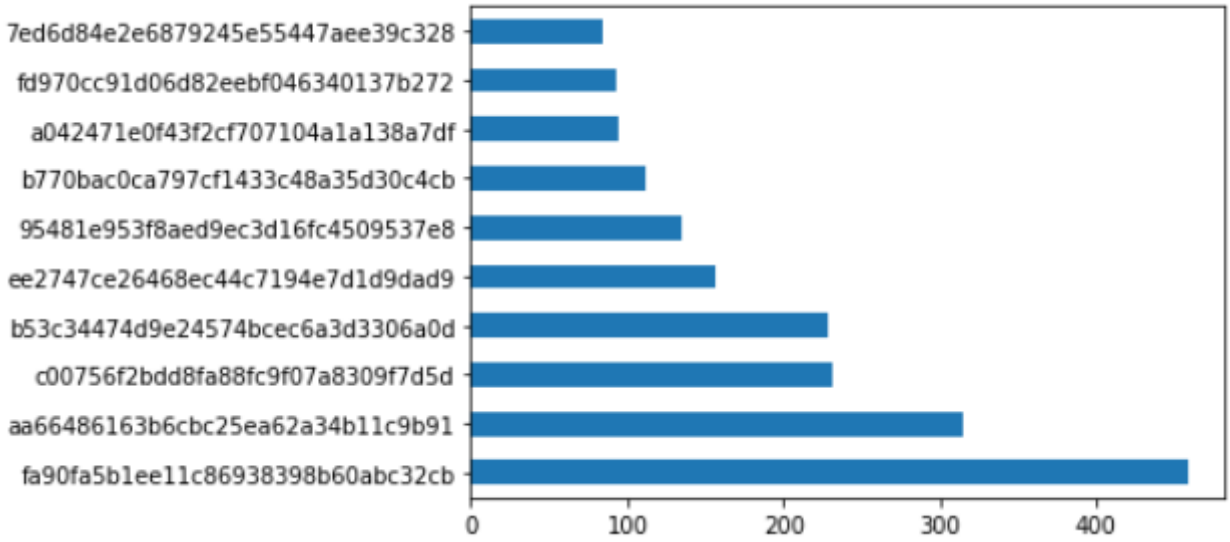

Рис. 20. Топ-10 пользователей, создавших наибольшее число профилей

Самое наибольшее — 459 профилей. Но так как данная переменная строкового типа, а всего уникальных значений 5415, то невозможно считать её категориальной, так как очень сильно расширится пространство признаков при использовании OneHotEncoding, что может негативно сказаться на времени обучения и качестве модели. Поэтому была создана новая переменная RankRescuer. Это переменная обозначает рейтинг пользователей, кто на каком месте по количеству объявлений. То есть пользователь с 459 объявлениями на 1 месте, с 315 — на 2 и так далее. Если у кого-то совпадает количество объявлений, то они делят одно место. Таким образом, в RankRescuer получилось 61 значение.

Из переменной VideoAmt (количество видео) была создана новая переменная has_video, которая обозначает наличие видео в профиле. Это было сделано из-за того, что VideoAmt в основном принимает значение 0, а остальные значения считаются выбросами (рис. 21).

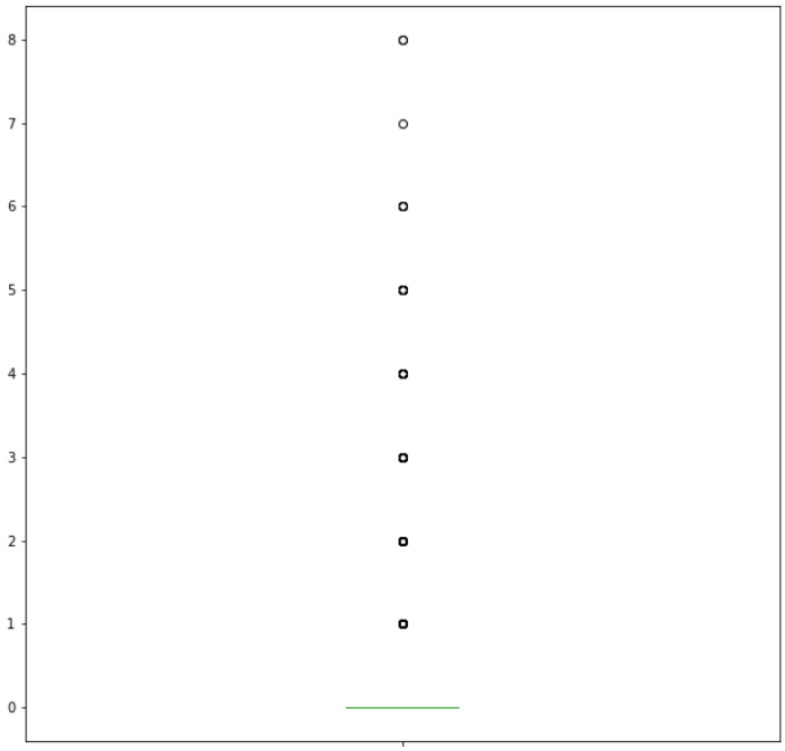

Рис. 21. Boxplot для VideoAmt

В переменной Description находится описание питомцев. Для каждого описания создатели задачи выполнили анализ эмоциональной окраски текста с помощью Google's Natural Language API и записали результаты в файлы формата JSON.

Из данных из этих файлов были созданы новые переменные lang, magnitude и score. В переменной lang хранится язык, на котором написаны описания (рис.22). Модель Google's Natural Language распознала английский (en), китайский упрощенный (zh), китайский традиционный (zh-Hant) и немецкий (de). Также есть часть наблюдений, где модель не смогла распознать, на каком языке написан текст. Этим наблюдениям присвоено значение 'no' в переменной lang. Наблюдений на немецком языке всего 2 штуки, поэтому они были удалены из датасета, чтоб не увеличивать количество категорий. С этой же целью китайский традиционный и китайский упрощенный были объеденены в один язык.

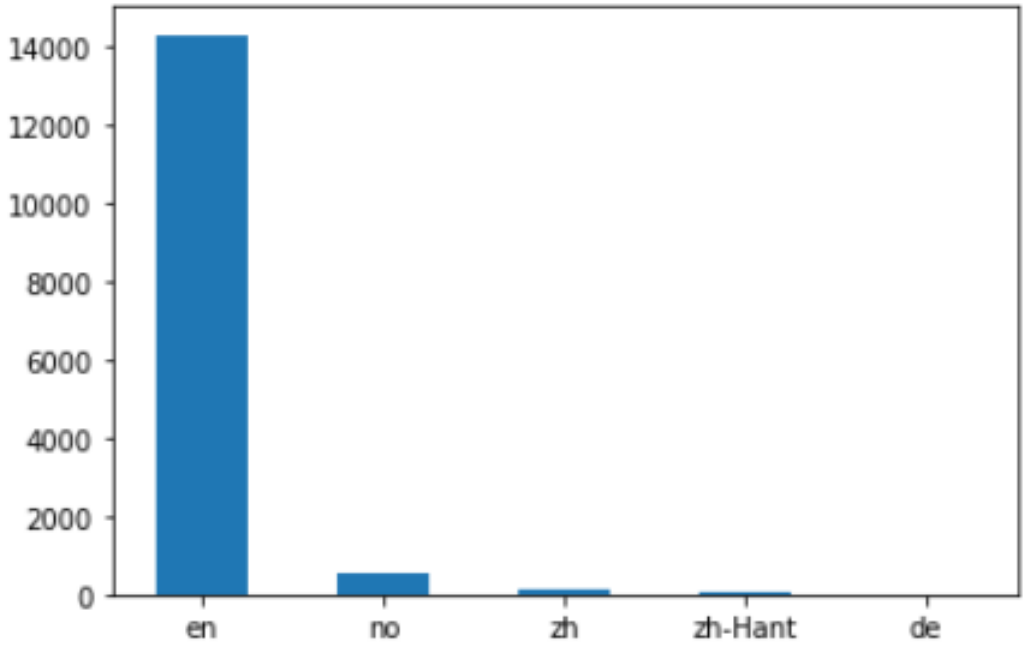

Рис. 22. Значения переменной lang

Score — это переменная со значениями из диапазона $[-1,\ 1]$. Отрицательное значение указывает на негативную окраску, положительное — на положительную. Чем ближе это значение к нулю, тем более текст нейтрален. Magnitude указывает, насколько эмоционален текст [5].

1.6. Кодирование категориальных переменных

1.7. Шкалирование переменных

2. Используемая метрика

3. Использованные модели

3.1. Baseline

3.2. Дерево решений

3.3. Логистическая регрессия

3.4. Случайный лес

3.5. Градиентный бустинг

4. Классификация только с использованием текстовых признаков

4.1. Предобработка текстов и выделение признаков

4.2. Обучение модели

4.3. Полученные результаты

5. Выбор датасета, модели и тестирование на Kaggle

Заключение

Заключение должно содержать информацию о проделанной работе и полученных результатах.

При написании текста работы следует иметь в виду, что её цель состоит в том, чтобы продемонстрировать квалификацию автора. Поэтому следует избегать общих и, тем более, тривиальных или нравоучительных высказываний. Мотивация выполняемой работы не должна носить слишком конкретный характер. Во время выступления на защите желательно избегать упоминаний об особенностях стандартных компонентов пользовательского интерфейса программ («нажимаем на правую кнопку», «перетаскиваем фрагмент мышью» и т. д.). Не следует комментировать задаваемые после защиты вопросы. Ответы на вопросы должны быть краткими.

Литература

1. Рекомендации по оформлению и представлению курсовых и выпускных квалификационных работ студентов института математики, механики и компьютерных наук. – Ростов н/Д, 2020.
2. Жуков М. Ю., Ширяева Е. В. LaTeX 2_ε: искусство набора и вёрстки текстов с формулами. – Ростов н/Д : Изд-во ЮФУ, 2009.

Приложение