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.

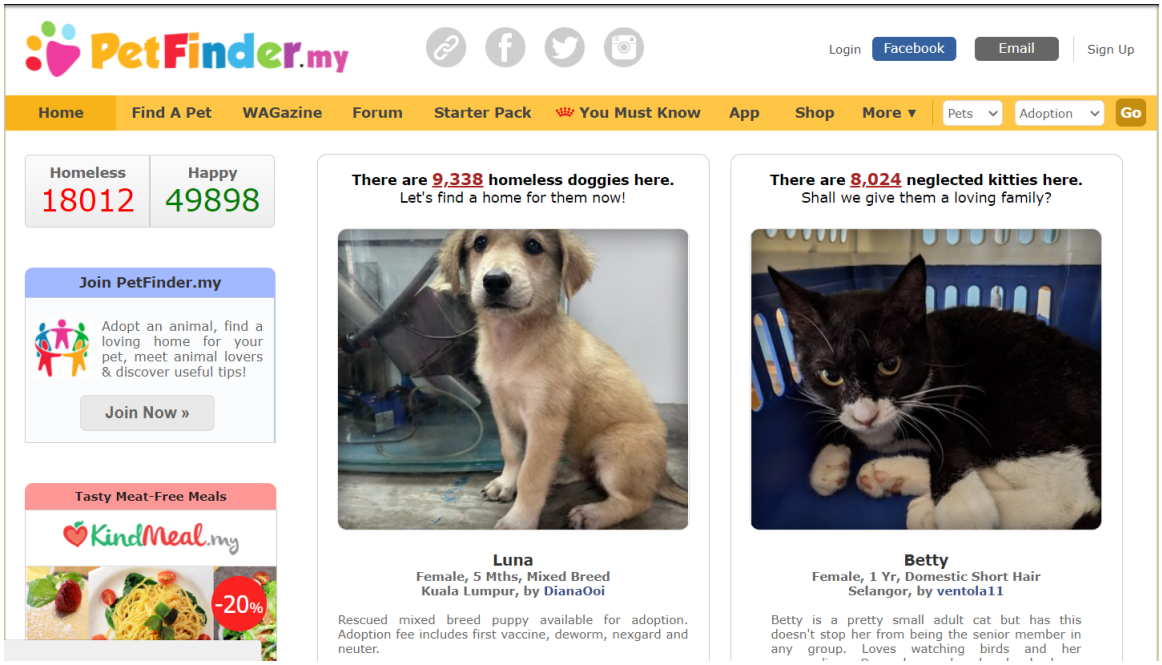

Рис. 1. Скриншот сайта petfinder.my

1.2. Описание признаков

1.3. Обработка пропущенных значений

1.4. Детекция и обработка выбросов

1.5. Создание новых признаков из имеющихся данных

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