ANIMAL DELAY PERIOD PREDICTION IN THE SHELTER USING NLP METHODS

Marchenkov I.
6-th year student, Department of Information Systems and Technologies,
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv

ПРОГНОЗУВАННЯ ПЕРІОДУ ЗАТРИМКИ ТВАРИНИ В ПРИТУЛКУ З ВИКОРИСТАННЯМ МЕТОДІВ NLP

Марченков І.Д.
студент 6-го курсу кафедри інформаційних систем і технологій,
НТУУ «КПІ ім. Ігоря Сікорського», м. Київ
DOI: 10.24412/3453-9875-2021-73-1-65-69

Abstract

The problem of homeless animals is considered in the article. Among the 7-8 million animals that end up in rescue shelters, about 45% of animals are euthanized, 15-25% of animals die due to overcrowding in shelters each year. The aim of the study is to improve the rate of appropriation of animals in shelters, using machine learning methods: Logistic regression, Naive Bayes classifier, Support vector machine, Decision trees, Random forest and Extremely randomized trees, as well as Natural language processing methods.

Анотація

У статті розглянуто можливі шляхи вирішення однієї з проблем безпритульних тварин. Серед 7-8 мільйонів тварин, які потрапляють до рятувальних притулків, близько 45% тварин піддаються евтаназії, 15-25% тварин помирають через переповнення притулків щороку. Метою дослідження є оцінка періоду перебування тварини в притулку та зменшення цього періоду, за рахунок використання методів машинного навчання: логістичної регресію (Logistic regression), наївного баєсевого класифікатора (Naive Bayes classifier), методу опорних векторів (Support vector machine), дерева рішень (Decision tree), випадкового лісу (Random forest) та гранично випадкових лісів (Extremely randomized trees), а також методів обробки природньої мови (Natural language processing).

Keywords: Pet adoption, euthanasia problem, machine learning methods, natural language processing, decision support tool.

Ключові слова: привласнення тварин, проблема евтаназії, методи машинного навчання, обробка природної мови, інструмент підтримки прийняття рішень.

Вступ. Проблема безпритульних тварин в Україні стоїть досить гостро і є актуальною як у великих містах, так і в маленьких селах. Головними інструментами у подоланні безпритульності тварин є системна якісна стерилізація тварин-безхатьків, активне прилаштування їх у сім'ї, а також відповідальне ставлення господарів до своїх домашніх улюбленців – обов'язкові реєстрація та чіпування. У даному дослідженні розроблено алгоритм, що буде прогнозувати період, який тварина проведе у притулку до її привласнення господарем. Отриманий результат допоможе розробити стратегії для зменшення проведеного періоду в притулку та зниження показнику евтаназії.

Огляд літератури. Клевенджер Дж. провів дослідження [1] щодо позитивного впливу стерилізації на ймовірність привласнення домашніх тварин. Автор дослідив вплив співпраці ветеринарних медичних шкіл у збільшенні кількості привласнення домашніх тварин, пропонуючи безкоштовну стерилізацію. Результати продемонстрували, що співпраця між ветеринарними клініками та місцевими притулками для тварин зменшила евтаназію на 26 % домашніх тварин. Браун та ін. провели дослідження [2], в якому оцінювали вплив віку, породи, кольору та візерунку шерсті на тривалість перебування у притулку без евтаназії. Автори дійшли висновку, що колір не впливав на термін перебування, тоді як стать, візерунок шерсті та порода були суттєвими чинниками для періоду перебування котів у притулку.

У наборі даних, що використовується для вирішення задачі прогнозування періоду затримки, залежна змінна складається з п'яти класів, чотири з яких мають порядкову властивість, тобто чим більше значення змінної, тим пізніший період привласнення. Френк і Хол стверджують, що стандартні підходи вирішення можуть не враховувати порядкову властивість вихідної змінної [3], тому вони пропонують використовувати метод дерев рішень.

Опис набору даних. Для розуміння логіки побудови математичної моделі розглянутої задачі, спочатку було проаналізовано структуру набору даних, який використовуватиметься для дослідження проблеми прогнозування періоду привласнення тварин із притулку.

Набір даних розміщено на провідній платформі захисту тварин Малайзії PetFinder та у соціальній мережі спеціалістів по обробці даних та машинному навчанню (англ. Machine Learning) Kaggle. Набір складається з трьох частин. Перший – це файл CSV із детальною інформацією, що включає тип тварини (собака чи кіт), породу, стать, забарвлення, довжину хутра, стерилізацію, стан здоров'я та опис. Опис інформації подано в таблиці 1.

Другий набір даних – це файли описів JSON з оцінкою привабливості, а третій набір – велика колекція файлів зображень. Загалом набір складається з

14993 записів домашніх тварин та 23 характеристик по кожній із них. Вихідна змінна AdoptionSpeed, що є показником періоду привласнення має п'ять класів (табл. 2). Привласнень у проміжку з 91 по 100 день не було. Проміжки часу у різних класах неоднакові. Клас «0» містить у собі інтервал розміром 1 день, клас «2» - інтервал 23 дні.

Таблиця 1

Опис набору даних

Назва незалежної змінної	**Опис**	**Вимір (для дискретних змінних)**
PetID	Унікальний ідентифікатор тварини	
Type	Тип тварини	1 – Собака, 2 – Кіт
Name	Ім`я тварини	
Age	Вік тварини у місяцях	
Breed1	Первинна порода тварини	306 можливих варіантів
Breed2	Вторинна порода тварини (у випадку якщо порода є змішаною)	306 можливих варіантів
Gender	Стать тварини	1 – Самець, 2 – Самиця, 3 – Змішана (якщо запис містить декілька тварин)
Color1	Колір 1	7 можливих варіантів
Color2	Колір 2	7 можливих варіантів
Color3	Колір 3	7 можливих варіантів
MaturitySize	Розмір тварини у зрілому віці	1 – Малий, 2 – Середній, 3 – Великий, 4 – Дуже великий, 0 – Не вказано
FurLength	Довжина хутра	1 – Коротке, 2 – Середнє, 3 – Довге, 0 – Не вказано
Vaccinated	Наявність вакцинації	1 – Так, 2 – Ні, 3 – Не впевнений
Dewormed	Дегельмінтизація	1 – Так, 2 – Ні, 3 – Не впевнений
Sterilized	Наявність стерилізації	1 – Так, 2 – Ні, 3 – Не впевнений
Health	Стан здоров'я	1 - Здоровий, 2 - Незначна травма, 3 - Серйозна травма, 0 - Пропустити
Quantity	Кількість тварин у профілі	
Fee	Вартість привласнення	0 – Безкоштовно
State	Розташування штату в Малайзії	15 можливих варіантів
RescuerID	Унікальний ідентифікатор рятівника	
VideoAmt	Загальна кількість завантажених відео для цієї тварини	
PhotoAmt	Усього завантажених фотографій для цієї тварини	
Description	Текстовий опис профілю тварини. Основна мова – англійська, в деяких профілях присутня малайська або китайська	
SentimentMagnitude	Величина емоційності текстового опису (від 0.0 до +inf)	
SentimentScore	Загальна емоційність текстового опису тварини (від -1.0 до 1.0)	

Змінні PetID, Name, RescuerID видалимо з навчальної вибірки, оскільки вони не впливатимуть на значення залежної змінної при побудові моделі. До ознак також було додано довжину текстового опису. Для обробки безпосередньо текстового опису було використана модель «торба слів» (bag-of-words), що представляє текст у вигляді мультимножини його слів, не беручи до уваги граматику і навіть порядок слів, підраховуючи частоту появи кожного слова [4]. Після цього було застосовано швидке кодування (One-Hot Encoding) – процес, за допомогою якого категорійні змінні були перетворені на відповідну алгоритмам машинного навчання форму [5].

Методи

Логістична Регресія. Багатокласова логістична регресія може бути застосована для класифікації 5 змінних. Ймовірнісний розподіл відповіді по заданим вхідним параметрам:

$$P(Y_i = c|X;\beta) = \frac{e^{\beta_c * x_i}}{\sum_{k=1}^{K} e^{\beta_k * x_i}},$$

де X – матриця ознак, що описують спостереження, β – набір коефіцієнтів регресії, c – клас, ймовірність належності до якого треба визначити, Y_i – елемент із множини класів, $e^{\beta_c * x_i}$ – оцінка віднесення спостереження i до класу c, $\sum_{k=1}^{K} e^{\beta_k * x_i}$ – сума оцінок віднесення спостереження i до кожного класу k із множини K.

Багатокласова логістична регресія є розширенням бінарного випадку. Модель передбачає обчислення ймовірності належності об'єкту до кожного із класів.

Таблиця 2

Опис залежної змінної

Залежна змінна	Період привласнення
0	У той же день
1	Від 1 до 7 днів
2	Від 8 до 30 днів
3	Від 31 до 90 днів
4	Після 100 днів перебування у притулку привласнення відсутнє

Наївний баєсів класифікатор. Метод базується на теоремі Баєса із передбаченням класу на основі незалежних предикторів. Наївну баєсову модель легко побудувати та вона особливо підходить для великих масивів даних. Відомо, що наївний баєсів метод не тільки простий, але й у деяких випадках перевершує найскладніші методи класифікації [6]. Теорема Баєса представляє спосіб обчислення апостеріорної ймовірності $P(c|x)$ з $P(c)$, $P(x)$ та $P(x|c)$. Рівняння теореми Баєса:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)},$$

де $P(c|x)$ – апостеріорна ймовірність класу c для предиктора x, $P(c)$ – апріорна ймовірність класу c, $P(x|c)$ – умовна ймовірність предиктора x при класі c, $P(x)$ – апріорна ймовірність предиктора.

Метод Опорних Векторів. Основна ідея методу – переведення вихідних векторів у простір більш високої розмірності і пошуку роздільної гіперплощини з максимальним зазором у цьому просторі. Дві паралельні гіперплощини будуються по обидва боки гіперплощини, що розділяє класи. Роздільною гіперплощиною буде гіперплощина, що максимізує відстань до двох паралельних гіперплощин. Алгоритм працює у припущенні, що чим більша різниця або відстань між цими паралельними гіперплощинами, тим меншою буде середня помилка класифікатора.

Дерева Рішень (Decision Trees). Дерево рішень, або метод дерева класифікації, на кожному кроці обирає змінну, яка є найбільш відповідною для поділу, а потім знаходить точку поділу. З точки зору найкращого поділу, інтерпретовано це як поділ, який найкращим чином зменшує похибку та покращує чистоту листкових вузлів [7].

Випадковий ліс. Випадковий ліс (random forest) ефективно оброблює дані з великою кількістю ознак і класів [8], тому вони можуть стати у нагоді, особливо при побудові моделі із закодованим текстом при обробці природньої мови. Добре оброблюють і дискретні, і безперервні ознаки, а також надають можливість оцінити рівень впливовості кожної змінної на модель (тест out-of-bag). Гранично випадкові ліси (extremely random forests) ще більше посилюють роль випадковості [9]. Поряд із випадковим вибором ознак випадково вибираються також порогові значення. Ці випадково генеровані значення стають правилами розбиття, що додатково зменшують варіативність моделі. Тому використання гранично випадкових лісів зазвичай призводить до більш гладких кордонів прийняття рішень, порівняно з тими, які вдається отримати за допомогою випадкових лісів.

Результат. Були побудовані моделі, що описані вище, з урахуванням та без урахування текстового опису тварини. Результати наведені в таблиці 3. Точність моделі логістичної регресії отримала незначну перевагу із урахуванням опису на тестових даних. Модель наївного Баєсу та метод опорних векторів відчутно погіршили результат з використанням NLP. Дерева прийняття рішень показали незначне погіршення. Випадковий ліс та гранично випадкові ліси покращили свій результат при використанні природньої обробки мови. Результат точності при використанні випадкового лісу покращився на 1,16 %, гранично випадкових лісів – 8,52 %. Найкращі результати мають випадковий ліс та гранично випадкові ліси, з невеликою різницею точності.

Для моделі випадкового лісу, що є найкращою за точністю, було побудовано гістограму впливовості ознак на модель (рис. 1). Найвпливовішими ознаками є вік тварини, кількість фотографій, показники величини та загальної емоційності текстового опису, а також розмір опису, що сумарно мають впливовість понад 40 %. Тобто, опис тварини та його наявність є значними чинниками, що впливають на вибір тварини для привласнення новим господарем.

Таблиця 3

Результат тестування моделей класифікації

	Тренувальні дані		Тестові дані	
Назва методу	**Точність (без NLP)**	**Точність (з NLP)**	**Точність (без NLP)**	**Точність (з NLP)**
Логістична регресія	0.337	0.337	0.333	0.334
Наївний Баєс	0.338	0.608	0.338	0.231
Метод опорних векторів	0.307	0.278	0.311	0.283
Дерева прийняття рішень	0.442	0.445	0.383	0.378
Випадковий ліс	0.443	0.999	0.431	0.436
Гранично випадкові ліси	0.998	0.998	0.399	0.433

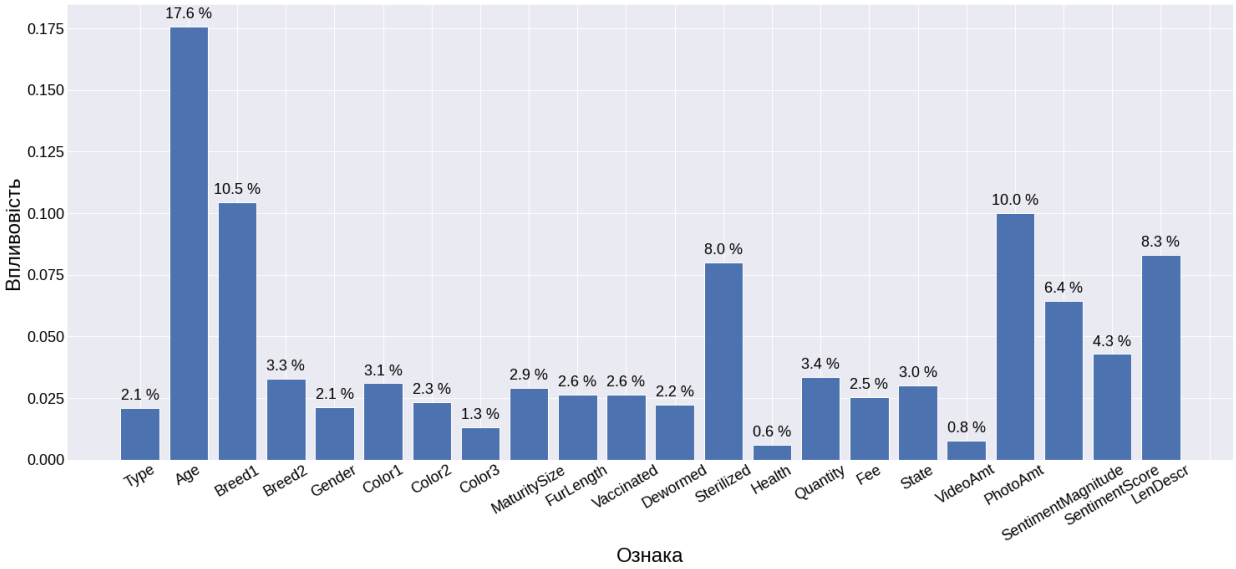

Рисунок 1

Впливовість незалежних змінних на результат при застосуванні класифікатору випадкового лісу

Побудовано матрицю невідповідностей (confusion matrix) для п'яти періодів затримки тварини в притулку, що характеризують відносну швидкість від найменшого значення до найбільшого (від 0 до 4). Дана матриця дає можливість унаочнити продуктивність алгоритму, зазвичай керованого навчання. Кожен з рядків цієї матриці представляє зразки прогнозованого класу, тоді як кожен зі стовпців представляє зразки справжнього класу [10]. Значення подані у відсотках відносно їх кількості по класу (рис. 2). Ми аналізуємо результати віднесення кожного класу та визначаємо частку невірно віднесених класів.

Тестова вибірка має 2,93% значень, що відносяться до класу «0». Це невелика кількість відносно інших класів, тому при побудові моделі ми зіштовхнулися з певними складнощами, оскільки вона віднесла до класу «0» лише 2 значення зі 132 можливих, причому зробила у цих випадках помилку. Тобто класифікатор у підсумку не зможе нам спрогнозувати за ознаками, чи зможуть привласнити тварину у той же день, що вона і потрапила до притулку.

Якщо розглядати інші класи, то бачимо, що елементи по діагоналі мають найбільші значення відносно їх колонок. Чим значення елементу по діагоналі ближче до 100%, тим краще класифікує модель даний клас. У випадку класу «1» класифікатор правильно зробив прогноз для 37% записів, класу «2» - 41%, класу «3» - 18% та для класу «4» показав 55% точності.

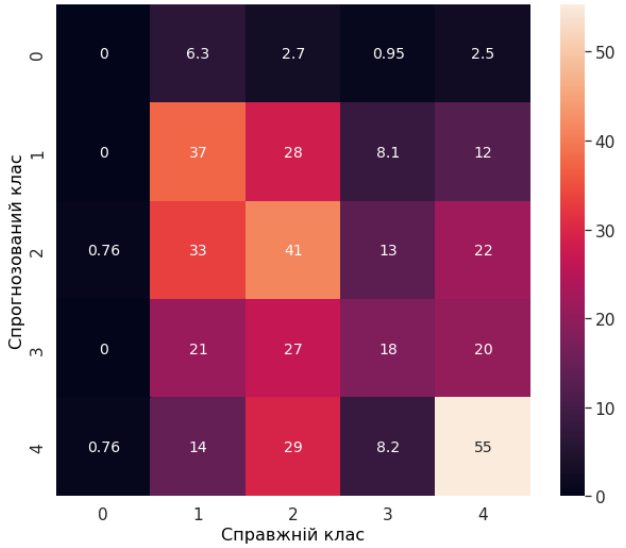

Рисунок 2 – Матриця невідповідностей моделі випадкового лісу

Висновки. У ході дослідження було побудовано базові моделі машинного навчання: логістичну регресію, наївний баєсів класифікатор, метод опорних векторів, дерева рішень, випадковий ліс та гранично випадкові ліси. Було зроблено порівняння ефективності роботи моделей: з та без врахування текстового опису. Текстовий опис було оброблено засобами NLP, за допомогою моделі «bag-of-words».

Серед усіх моделей машинного навчання найкраще себе показав випадковий ліс із точністю класифікації 0.436 на тестових даних. Методи обробки природньої мови покращили результат класифікатору випадкового лісу на 1.16 %. Найвпливовішими на результат ознаками виявилися: вік тварини, кількість фотографій, показники величини та загальної емоційності текстового опису, а також довжина опису.

СПИСОК ЛІТЕРАТУРИ:

1. Clevenger J, Kass PH. Determinants of adoption and euthanasia of shelter dogs spayed or neutered in the University of California veterinary student surgery program compared to other shelter dogs. J Veterinary Med Educs / 1. Clevenger J, Kass PH. – 2003. – №30. – С. 8.

2. Brown WP. Age, breed designation, coat color, and coat pattern influenced the length of stay of cats at a no-kill shelter / Brown WP, Morgan KT. // J Appl Anim Welf Sci. – 2015. – №18. – С. 80.

3. Frank E. A simple approach to ordinal classification. In European Conference on Machine Learning / E. Frank, M. Hall. // Springer. – 2001. – С. 145–156.

4. Zhang Y. Understanding bag-of-words model: a statistical framework / Y. Zhang, R. Jin, Z. Zhou. // International Journal of Machine Learning and Cybernetics volume. – 2010. – №1. – С. 43–52.

5. Potdar K. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers / K. Potdar, T. Pardawala, C. Pai. // International Journal of Computer Applications. – 2017. – №175. – С. 7–9.

6. Ray S. 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R / Sunil Ray. – 2015.

7. Classification And Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – Boca Raton, Florida, United States: Routledge, 2017. – 368 с.

8. Cutler A. Random Forests / A. Cutler, D. R. Cutler, J. R. Stevens // Ensemble Machine Learning: Methods and Applications / A. Cutler, D. R. Cutler, J. R. Stevens., 2011. – (Springer). – (45; кн. 1). – С. 157–176.

9. Chakrabarty N. Navo Minority Over-sampling Technique (NMOTe): A Consistent Performance Booster on Imbalanced Datasets / N. Chakrabarty, S. Biswas. // Journal of Electronics and Informatics. – 2020. – №2. – С. 96–136.

10. Powers D. M. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation / David Martin Ward Powers. // Journal of Machine Learning Technologies. – 2008. – №2. – С. 37–63.