УДК 004.82:004:85

Бушменьов.В.Є., студент 4 курсу СО Бакалавр
Зелінська О.В., к.т.н., доцент, доцент кафедри інформаційних технологій

ПРИКЛАДНЕ ВИКОРИСТАННЯ ДЕРЕВ РІШЕНЬ НА ПРИКЛАДІ АНАЛІЗУ ДАТАСЕТУ ЗА ДОПОМОГОЮ МОВИ ПРОГРАМУВАННЯ PYTHON

Донецький національний університет імені Василя Стуса, м. Вінниця

Аналіз даних - це розділ математики, що займається розробкою різних методів для обробки даних незалежно від того, звідки ці дані потрапляють до нас. Дерева рішень - використовується в галузі статистики та аналізу даних для прогнозних моделей. Структура дерева містить такі елементи: «листя» і «гілки». На ребрах («гілках») дерева ухвалення рішення записані атрибути, від яких залежить цільова функція, в «листі» записані значення цільової функції, а в інших вузлах — атрибути, за якими розрізняються випадки [1,3].

Актуальність полягає в тому, що цей статистичний метод підтримує різні набори параметрів, які можуть набувати різну ефективність залежно від вхідного набору даних. Отже, правильний вибір параметрів наблизить нас на найкращої оцінки відгуку на тестових даних, що і є головною ціллю будь-якого аналізу даних [4]. В нашому випадку для дослідження ми будемо використовувати мову програмування “Python”, яка є програмним середовищем для математичних обчислень.

Розглянемо приклад такого аналізу на основі даних представлених на платформі для змагань з аналітики та передбачуваного моделювання, в рамках якого, статистики та добувачі даних конкурують у створенні найкращих моделей для прогнозування та опису даних, запропонованих компаніями або користувачами “Kaggle”. Ми будемо використовувати датасет “Kaggle” [2].

Опис датасету: в тренувальній виборці міститься 891 спостереження, а в тестовій - 418. Відгуком в даному датасеті є змінна Survived, яка визначає вижив пасажир внаслідок корабельної аварії чи ні. Поля Pclass, Age, Sibsp, Parch, Fare є числовими і вимагають перетворення. Візуальне вивчення даних показало, що поля Age та Fare можуть набувати порожніх значень. Для пасажирів, значення яких немає, передбачається використання медіани серед ненульових значень цього параметра.

Під час аналізу ми будемо використовувати такі методи дерев рішень: змішані дерева та градієнтний бустинг. Для підбору найефективніших параметрів для змішаних дерев, будемо використовувати пошук по сітці з 5 кросс-валідаційними фолдами та з параметрами, які відображені на Рисунку 1.

```
param_grid = {
    'classifier__class_weight': [None,"balanced", 'balanced_subsample'],
    'classifier__min_samples_split': [6,10,14,18,20],
    'classifier__max_depth' : [150,175,200,225,None],
    'classifier__min_samples_leaf': [1,2,3],
    'classifier__max_features':["log2","auto","sqrt"],
```

Рисунок 1 - пошук по сітці

Після цього, скориставшись найкращими параметрами підсумуємо результати алгоритму на тестовій та тренувальній виборці [Рисунок 2].

```
Training Prefomence : 0.9434878735093126
Test Perfomence : 0.92072192513369
Accuracy metric: 0.848314606741573
F1 metric : 0.8085106382978724
```

Рисунок 2 - результати пошуку по сітці

Отже, ми бачимо, що найкращого результату ми досягли на тестовій виборці з точністю 0.848. Тепер перейдемо до методу градієнтного бустингу. Спробуємо віднайти оптимальну кількість дерев для отримання найкращого результату та найкращий параметр звуження відповідно двох метрик : accuracy та AUG [Рисунок 3].

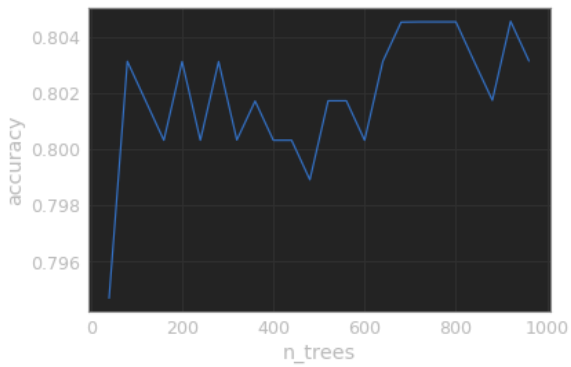
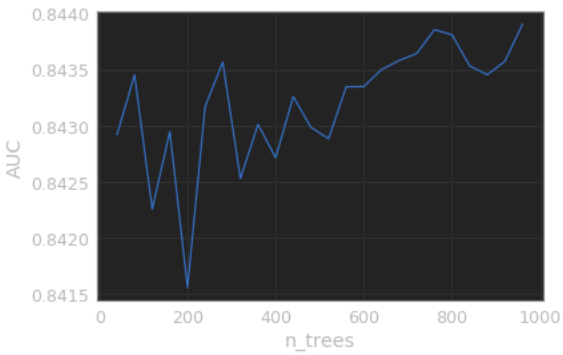
accuracy / n trees	AUG / n trees
	

Рисунок 3 - дослідження найефективнішої кількості дерев для алгоритму відносно метрик “accuracy” та “AUG”

З графіків бачимо, що найоптимальніший параметр кількості дерев для двох метрик рівний 760. Далі перейдемо до найоптимальнішого параметру звуження. Будемо досліджувати два методи звуження : L1 та L2 [Рисунок 4].

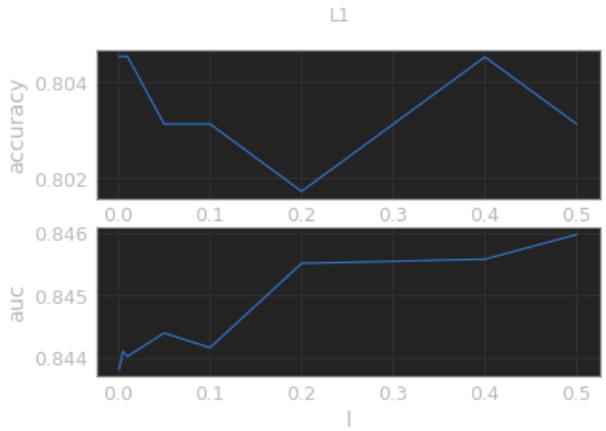
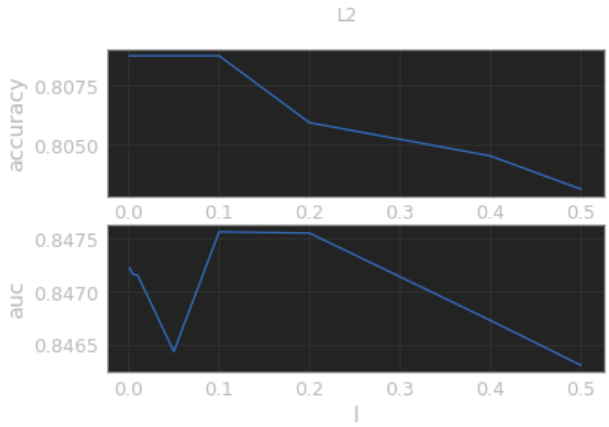
auc|accuracy / L1	auc|accuracy / L2
	

Рисунок 4 - дослідження найефективнішого параметру звужування для алгоритму відносно метрик “accuracy” та “AUG”

З графіків бачимо, що L2 при значенні 1.15 показує найкращі результати. Тепер оцінимо тестову виборку з отриманими найоптимальнішими значеннями для методу RGB. Точність методу на тестовій виборці рівна 87%.

Висновок. Отже, під час проведення дослідження ми виявили, як різні параметри можуть впливати на результати статистичних моделей та на скільки методи дерев рішень є ефективні в вирішенні актуальних проблем. Зробили дослідження ефективності двох моделей, які базуються на деревах рішень та порівняли їх. Метод градієнтного бустингу виявився кращим на датасеті “Титанік” та набрав 87% точності на тестовий виборці, що є на 3% більшим, чим метод змішаних дерев.

Список літературних джерел

1. *Дерева ухвалень рішень - [Електронний ресурс]. Режим доступу: https://uk.wikipedia.org/wiki/Дерево_ухвалення_рішень*
2. *Kaggle - [Електронний ресурс]. Режим доступу: https://uk.wikipedia.org/wiki/Kaggle*
3. *Practical statistics for Data Scientists: П. Брюс, Э. Брюс., O’Really Media Inc., 1005 Gravenstein Highway North, 2018. 304 с.*
4. *Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. An Introduction to Statistical Learning: with Applications in R. Springer New York Heidelberg Dordrecht London, 2017*