УДК 004.04

Явгусішин Б.А., студент 3 курсу спеціальності «Комп'ютерні науки»
Нескородєва Т. В., д.т.н., завідувачка кафедри інформаційних технологій

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ДЕРЕВ РІШЕНЬ В СТАТИСТИЧНОМУ НАВЧАННІ

Донецький національний університет імені Василя Стуса, м. Вінниця

Дерева рішень є одним з ключових інструментів, що використовуються в статистичному навчанні для регресії та класифікації даних. Однією з переваг дерев є те, що їх легше трактувати, ніж інші регресійні моделі. Також їх можна дуже зручно представити графічно, але вони мають недоліки в передбачені даних на не навчальних виборках.

Головною метою дослідження є аналіз ефективності дерев рішень в статистичному навчанні, і порівняння результатів методів, які покращують результати передбачень.

Дослідження ефективності дерев рішень в статистичному навчанні зазвичай зосереджені на виявленні тих чинників, які можуть позитивно або негативно впливати на ефективність дерева рішень. Однією з основних проблем, які виникають при дослідженні ефективності дерев рішень, є вибір параметрів дерева, а саме такі як глибина дерева, кількість листків та критерії поділу.

Існують різні підходи для покращення точності передбачень, але в даній роботі будуть розглянуті такі, як беггінг, випадкові ліси та бустінг. Основою кожного з даних методів є побудова великої кількості дерев з подальшим їх об'єднанням, що призводить до покращення точності передбачення. Також будуть використані лінійна регресія та узагальнені адитивні моделі (GAM).

Розглянемо методи на прикладі датасета створеного на основі даних о народжуваності в штаті Північна Кароліна, США в 2001 році. Він містить містить 1450 унікальних рядків та 15 стовпців із характеристиками дитини та матері.

Атрибути:

1. ID Patient: ID код пацієнта
2. Plural: кількість народжених [1= одна дитина, 2= двійня, 3= трійня]
3. Sex: стать дитини [1=чоловік 2=жінка]
4. MomAge: вік матері [в роках]
5. Weeks: кількість тижнів вагітності [повних тижнів]
6. Marital: заміжня пацієнтка чи ні [1=заміжня, 2=незаміжня]
7. RaceMom: раса матері [1=біла, 2=чорношкіра, 3=американська індіанка, 4=китайка, 5=японка, 6=гавайка, 7=філіппінка або 8=уродженка інших азіатських або тихоокеанських островів]

8. HispMom: Латиноамериканське походження матері [C=кубинка, M=мексиканка, N=не латиноамериканка O=інша латиноамериканка, P=Пуерто-Ріко, S=Центральна/Південна Америка]
9. Gained: Вага, набрана під час вагітності [у фунтах]
10. Smoke: чи курить пацієнтка? [1=так або 0=ні]
11. BirthWeightOz: Вага при народженні [в унціях]
12. BirthWeightGm: Вага при народженні [в грамах]
13. Low: Показник низької ваги при народженні [1=2500 грамів або менше]
14. Premie: Показник передчасних пологів [1=36 тижнів або раніше]
15. MomRace: Раса матері [чорношкіра, латиноамериканська, інша або біла]

Будемо передбачувати показник Weeks і буде оцінювати продуктивність усіх моделей за допомогою повторної перехресної перевірки.

Досліджені моделі показали наступні результати середньоквадратичної похибки:

```
# A tibble: 5 × 2
  method               MSE
  <chr>              <dbl>
1 GAM (Smooth Terms)  2.5
2 GBM                 2.56
3 Bagged Trees        2.7
4 Random Forest       2.72
5 Linear Reg.         2.96
```

Рисунок 15 - Середньоквадратичні похибки

Беручи до уваги дані показники можемо сказати, що лінійна регресія має найгірший показник середньоквадратичної похибки. Тоді як методи бустінг та узагальнені адитивні моделі(GAM) мають найкращі показники, але інші методи є достатньо конкурентоспроможними.

За допомогою повторної перехресної перевірки кожна модель має 100 оцінок ефективності. Візуалізуємо результати дослідження, використовуючи модифіковану коробкову діаграму впорядковуючи найкращі найгірша продуктивність:

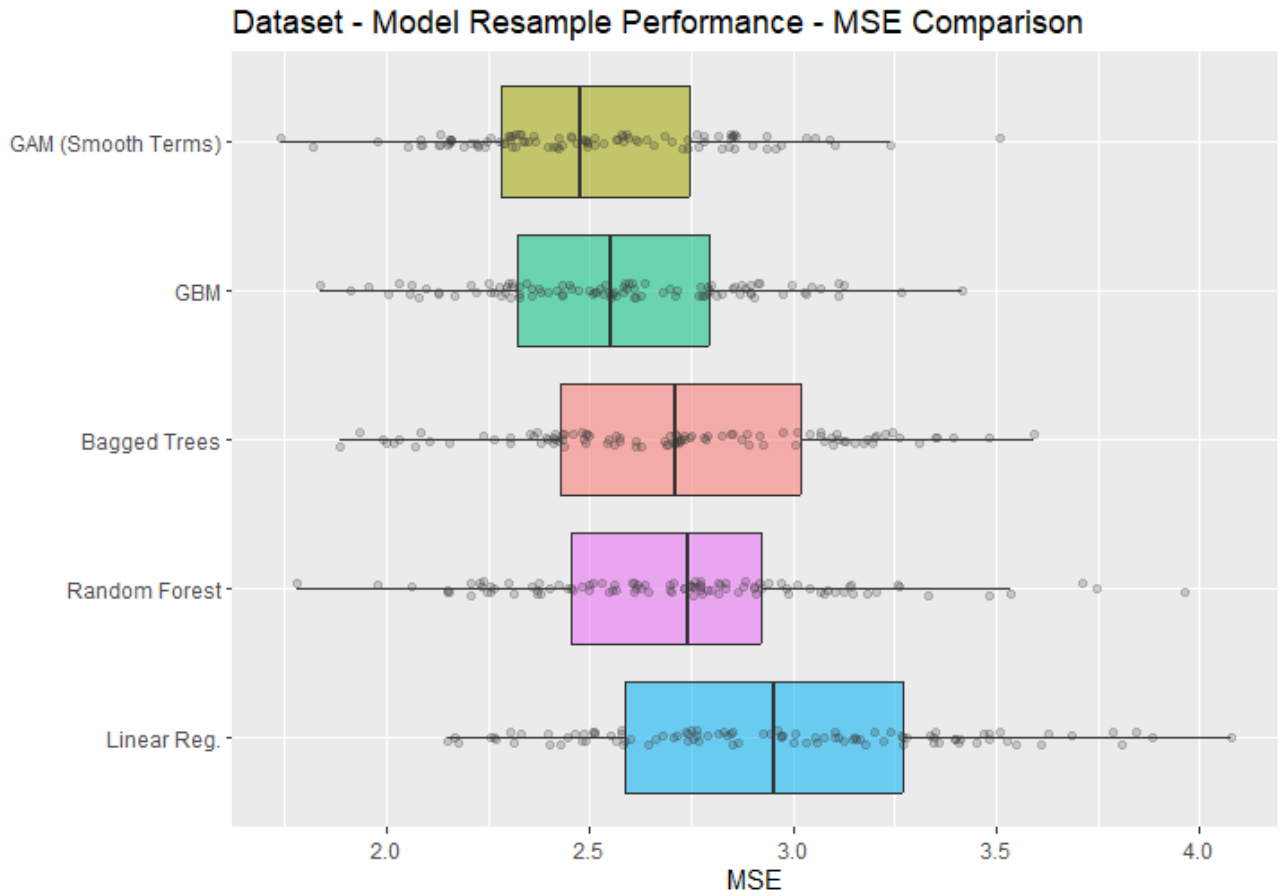

Рисунок 16 - Коробкова діаграма

Дослідження ефективності дерев рішень в статистичному навчанні вказують на те, що цей алгоритм може бути дуже ефективним інструментом для класифікації та прогнозування даних. Параметри дерева рішень, такі як глибина дерева, кількість листків та критерії поділу, можуть впливати на його ефективність, тому важливо обирати оптимальні параметри при застосуванні дерев рішень. Вони можуть допомогти в розв'язанні складних проблем, таких як класифікація та прогнозування даних, а також можуть забезпечити легку інтерпретацію результатів для додаткової аналітики та прийняття рішень. Однак, важливо враховувати, що дерева рішень не є універсальним рішенням для всіх проблем і що вони можуть бути вразливими до перенавчання, тому важливо ретельно відбирати параметри та контролювати якість моделі.

Список літератури

1. *Джеймс Г. Уіттон А Хасті Т. Тібшірані Р. Введення в статистичне навчання з прикладами на мові R. Пер. з англ. С. Е. Мастіцкого - М.: ДМК Пресс, 2017. - 456 с.*

2. *Що таке відеоняня та на що звертати увагу при її виборі. URL: https://vincentarelbundock.github.io/Rdatasets/datasets.html (дата звернення: 07.11.2022).*

3. *Датасети [Електронний ресурс] URL: https://gpsavto.com/articles/187-stezhennya-za-ditmy (дата звернення: 09.05.2023).*