УДК 004.9
DOI: 10.31673/2409-7292.2024.040013

Бажан Т. О.

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ПОБУДОВИ ПРОГНОЗІВ

Все що відбувається в світі, люди намагаються прогнозувати. Раніше прогнозування відбувалось за допомогою спостережень і обрахунків, які виконували люди. На сьогодні все можна спрогнозувати, використовуючи машинне навчання. Тому тема методів машинного навчання є дуже актуальною. У цій статті розглянуто різні методи машинного навчання, які використовуються для прогнозів. Було проаналізовано такі методи: лінійна регресія, випадковий ліс, градієнтний бустинг, нейронні мережі, метод опорних векторів, метод k-найближчих сусідів, а також автоматичне машинне навчання. Кожен із цих методів охарактеризовано за обраними критеріями, щоб порівняти їх ефективність у прогнозуванні. Кожен метод має свої переваги та недоліки, і неможливо визначити один універсально найкращий, оскільки їх ефективність залежить від конкретного типу прогнозів. Для узагальнення результатів дослідження побудовано таблицю, в якій зазначені ключові характеристики кожного методу, що дозволяє наочно оцінити їхні сильні та слабкі сторони. Особливу увагу приділено методу опорних векторів, зокрема його специфіці в прогнозуванні та можливостям покращення точності за певних умов. Також докладно розглянуто метод лінійної регресії, його переваги, недоліки та випадки ефективного застосування. У статті запропоновано підхід до об'єднання цих двох методів для створення гібридного підходу до прогнозування, що може покращити результати у складних задачах. Робота надає практичні рекомендації щодо вибору методу машинного навчання залежно від задачі прогнозування.

Ключові слова: машинне навчання, лінійна регресія, метод опорних векторів, прогнозування за допомогою машинного навчання.

Вступ

Сучасний світ сповнений невизначеностей і швидких змін, що робить прогнозування однією з найважливіших задач у різних сферах діяльності. Від бізнесу до медицини, від фінансів до екології — можливість точного прогнозування дозволяє приймати обґрунтовані рішення та зменшувати ризики. У зв'язку з цим, методи машинного навчання стали ключовим інструментом для аналізу великих обсягів даних і виявлення прихованих закономірностей. Методи машинного навчання пропонують різноманітні підходи до побудови прогнозів, серед яких особливо виділяються лінійна регресія, метод опорних векторів, випадковий ліс, градієнтний бустинг, нейронні мережі та метод k-найближчих сусідів. Кожен з цих методів має свої унікальні характеристики, переваги та недоліки, що робить їх більш чи менш ефективними залежно від специфіки задачі та доступних даних.

Ця стаття має на меті провести порівняльний аналіз обраних методів машинного навчання з акцентом на їхню ефективність у прогнозуванні. Основна увага буде приділена методам лінійної регресії та опорних векторів, оскільки їх комбінація може забезпечити більш точні результати прогнозування. Зазначимо, що вибір методу машинного навчання не є однозначним, і він повинен ґрунтуватися на характеристиках даних, специфіці задачі та метриках ефективності. Дослідження, проведене в рамках цієї статті, передбачає аналіз актуальних публікацій, що висвітлюють застосування методів машинного навчання для прогнозування, а також оцінку їх ефективності на основі визначених метрик. Це дозволить виявити оптимальні підходи до прогнозування, що мають значення для практичного впровадження у різних галузях.

Постановка проблеми

У сучасному світі, де дані стали ключовим ресурсом, питання ефективності прогнозування набуває особливої актуальності. Прогнозування – це процес, який дозволяє передбачити майбутні події на основі аналізу минулих даних. Однак, у зв'язку з різноманіттям методів машинного навчання, що існують сьогодні, вибір найкращого підходу для конкретної задачі стає складним завданням. Проблема полягає в тому, що не всі методи машинного навчання однаково ефективні для всіх типів даних і завдань. Кожен метод має свої особливості, переваги та недоліки, які можуть суттєво вплинути на точність прогнозів. Це

ускладнює процес вибору оптимальної моделі, що відповідає конкретним умовам і вимогам завдання.

Крім того, існує потреба у вдосконаленні прогнозування, зокрема, через комбінування різних методів для досягнення кращих результатів. Метод лінійної регресії, що є простим і зрозумілим, може виявитися недостатнім у випадках складних нелінійних залежностей, в той час як метод опорних векторів (SVM) демонструє свою ефективність у моделюванні складних структур даних. Поєднання цих методів може відкрити нові можливості для покращення точності прогнозів. Таким чином, основна проблема, яку розглядає ця стаття, полягає у визначенні найбільш ефективних методів машинного навчання для прогнозування, з акцентом на їх порівняння та можливість комбінування для досягнення більш точних і надійних результатів.

Аналіз останніх досліджень і публікацій

Сучасна наукова література активно досліджує методи машинного навчання, їх ефективність та застосування у прогнозуванні.

У роботі [1] пропонується вступ до машинного навчання, зосереджуючи увагу на практичному використанні Python. Автор розглядає основи таких методів, як лінійна регресія та SVM, акцентуючи увагу на їх застосуванні для задач прогнозування. У книзі [2] автори аналізують теоретичні та прикладні аспекти машинного навчання. Особливу увагу вони приділяють ансамблевим методам, таким як випадковий ліс та градієнтний бустінг, що демонструють високу точність у задачах прогнозування. Автори підкреслюють важливість коректного налаштування параметрів моделей для отримання оптимальних результатів.

У роботі [3] автор акцентує увагу на важливості інтуїтивного розуміння алгоритмів. Він зазначає, що кожен метод має свої сильні та слабкі сторони, які потрібно враховувати при його застосуванні. Стаття також розкриває ключові аспекти вибору методів машинного навчання для конкретних задач. У книзі [4] досліджується питання інтерпретованості моделей. Автор зосереджується на розробці методів, які забезпечують зрозумілі результати, що є важливим у таких сферах, як медицина та фінанси. Особливо цікаві розділи про генеративні моделі (VAE, GAN), які застосовуються для створення синтетичних даних, що покращує точність прогнозів.

Робота [5] представляє бібліотеку Scikit-learn, яка є стандартом у реалізації популярних алгоритмів машинного навчання. У ній детально розглядаються такі методи, як лінійна регресія, метод опорних векторів (SVM) та випадковий ліс, що робить цю бібліотеку незамінною для прикладного використання. Автори [6] досліджують архітектури глибоких нейронних мереж, таких як RNN та трансформери, які є ключовими для прогнозування часових рядів. Автори демонструють приклади використання глибокого навчання для фінансового аналізу та логістики.

Публікація [7] фокусується на практичному підході до навчання моделей. Автор надає рекомендації щодо запобігання перенавчанню моделей та підвищення їхньої ефективності у реальних задачах. Стаття [8] пропонує практичний підхід до поєднання різних моделей для покращення точності прогнозування. Автор досліджує, як інтеграція глибоких та простих методів дозволяє отримати збалансовані результати.

Робота [9] представляє практичний підхід до створення та навчання моделей машинного навчання за допомогою бібліотеки TensorFlow. Цей ресурс висвітлює можливості використання алгоритмів для прогнозування в реальному часі. В публікації [10] аналізується роль великих мовних моделей у прогнозуванні. Ця публікація демонструє можливості інтеграції GPT у задачі, пов’язані з текстовим аналізом і прогнозуванням.

Проаналізовані публікації висвітлюють актуальність та перспективність методів машинного навчання для прогнозування. Кожен з методів має свої особливості та потенціал, який може бути розкритий за умов правильного вибору та налаштування моделі.

Мета і задачі дослідження

Метою даного дослідження є виявлення методів машинного навчання, які є наймен ефективними відповідно до обраних метрик.

© Бажан, Т. О. (2024). Порівняльний аналіз методів машинного навчання для побудови прогнозів. Сучасний захист інформації, 4(60), 125–130. https://doi.org/10.31673/2409-7292.2024.040013.

Для досягнення мети необхідно виконати наступні задачі:

1. Обрати методи, які використовуються для прогнозів.
2. Обрати метрики для порівняння методів.
3. Порівняти методи у табличному вигляді.
4. Охарактеризувати обрані методи, виявити їх переваги та недоліки.

Дане дослідження спрямоване на вибір двох методів дослідження, щоб в подальшому спробувати їх поєднати і досягти кращого результату прогнозів.

Актуальність і дослідження

Вивчення методів машинного навчання для прогнозування в сучасному світі є дуже актуальним і важливим напрямком. Машинне навчання здатне ефективно працювати з великими обсягами даних, що дозволяє застосовувати його в різних галузях, таких як фінанси, медицина, технології, логістика, маркетинг та інші.

Моделі машинного навчання можуть аналізувати величезні обсяги даних і виділяти в них закономірності, що полегшує прийняття обґрунтованих рішень. З постійним розвитком технологій і збільшенням потужності обчислювальних ресурсів стає можливим використання більш складних моделей машинного навчання для отримання точних прогнозів. Використання методів машинного навчання може призвести до автоматизації процесів та оптимізації виробничих ланцюгів, що призводить до підвищення продуктивності та зменшення витрат. Деякі моделі машинного навчання можуть працювати в режимі реального часу, що робить їх ефективними для систем моніторингу, прогнозування та реагування на зміни в реальному часі. Машинне навчання може бути ефективним інструментом для розв’язання складних завдань, таких як розпізнавання образів, обробка природної мови, класифікація даних та інші.

Загалом, вивчення методів машинного навчання є ключовим для розвитку сучасних технологій та вирішення реальних завдань у різних галузях. Для прогнозування інвестицій використовують різноманітні методи машинного навчання, оскільки вони можуть ефективно аналізувати великі обсяги даних і виявляти складні залежності. Ось кілька методів, які часто застосовуються в цьому контексті:

1. Лінійна регресія (Linear Regression).
2. Випадковий ліс (Random Forest).
3. Градієнтний бустінг (Gradient Boosting).
4. Нейронні мережі (Neural Networks).
5. Метод опорних векторів (Support Vector Machines).
6. Метод k-найближчих сусідів (k-Nearest Neighbors).
7. Автоматичне машинне навчання (AutoML).

Ці методи можуть бути застосовані залежно від конкретного завдання прогнозування інвестицій та характеру даних, які використовуються для навчання моделей.

Проаналізуємо використання кожного методу.

Лінійна регресія – це модель, яка шукає лінійну залежність між вхідними факторами і вихідними змінними, щоб прогнозувати значення. Випадковий ліс – це метод, який використовує декілька рішень дерев для отримання більш точних прогнозів. Градієнтний бустінг – це метод, який поступово покращує прогноз, додаючи слабкі моделі. Глибокі нейронні мережі можуть аналізувати складні залежності та виявляти неочевидні закономірності. Метод опорних векторів використовується для класифікації та регресії, шукаючи гіперплощину, яка найкращим чином розділяє дані. Метод k-найближчих сусідів – це класифікація або регресія заснована на близькості до k найближчих прикладів. Автоматичне машинне навчання використовується для автоматизованого вибору, налаштування та навчання моделей машинного навчання.

Визначення того, який метод машинного навчання є найкращим, вимагає оцінки моделей за допомогою різних метрик. Вибір конкретних метрик залежить від характеру завдання (класифікація, регресія, кластеризація тощо). Ось деякі загальні метрики, які можна використовувати для оцінки ефективності моделей:

© Бажан, Т. О. (2024). Порівняльний аналіз методів машинного навчання для побудови прогнозів. Сучасний захист інформації, 4(60), 125–130. https://doi.org/10.31673/2409-7292.2024.040013.

1. Середньоквадратична помилка (Mean Squared Error, MSE).
2. Коефіцієнт детермінації (R-squared).
3. Точність (Accuracy).
4. Точність класу (Precision).
5. Повнота (Recall).
6. F-мера (F1 Score).
7. AUC-ROC (Area Under the Receiver Operating Characteristic curve).
8. Log Loss.

Перша обрана нами метрика – це середньоквадратична помилка (Mean Squared Error, MSE), яка використовується в регресійних задачах для вимірювання середнього квадрату відхилення прогнозованих значень від фактичних. Наступною є коефіцієнт детермінації (R-squared), який також використовується в регресійних задачах, показує відсоток варіації в залежній змінній, яку пояснює модель. Далі точність (Accuracy), вона використовується в задачах класифікації для вимірювання відсотка правильно класифікованих екземплярів. Потім точність класу (Precision), яка визначає, яка частина позитивних прогнозів є правильною. Повнота (Recall) визначає, яка частина всіх позитивних екземплярів була правильно визначена моделлю. F-мера (F1 Score) комбінує точність та повноту в одне число для оцінки якості моделі в задачах класифікації. AUC-ROC (Area Under the Receiver Operating Characteristic curve) використовується для оцінки якості класифікаційних моделей, особливо в задачах з незбалансованими класами. Log Loss використовується в задачах класифікації, оцінює відхилення між прогнозованою й фактичною ймовірністю класів.

1. Середньоквадратична помилка (Mean Squared Error, MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2,$$

де n - кількість спостережень, y_i - спостережене значення, y_i' - прогнозоване значення.

2. Коефіцієнт детермінації (R-squared):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

де $\bar{y}$ - середнє значення спостережень.

3. Точність (Accuracy):

$$\text{Accuracy} = \frac{\text{Кількість правильних класифікацій}}{\text{Загальна кількість спостережень}}.$$

4. Точність класу (Precision):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}},$$

де TP – кількість правильно класифікованих позитивних випадків, FP - кількість помилково класифікованих позитивних випадків.

5. Повнота (Recall): Recall= Recall=TP+FNTP

$$Recall = \frac{TP}{TP+FN},$$

де TP – кількість правильно класифікованих позитивних випадків, FN - кількість помилково класифікованих негативних випадків.

6. F1 Score:

© Бажан, Т. О. (2024). Порівняльний аналіз методів машинного навчання для побудови прогнозів. Сучасний захист інформації, 4(60), 125–130. https://doi.org/10.31673/2409-7292.2024.040013.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

7. AUC-ROC (Area Under the Receiver Operating Characteristic curve):

$$AUC - ROC = \int_0^1 TPR(FPR^{-1}(t))\,dx\ ,$$

де TPR – True Positive Rate, FPR - False Positive Rate, t – це поріг відсічення (threshold), який використовується для прийняття рішення щодо класифікації при прогнозуванні.

8. Log Loss:

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N}(y_i Log(y_i') + (1 - y_{i)} Log(1 - y_i'))\ ,$$

де y_i - спостережене значення, y_i' - ймовірність спостереження.

Ці формули можуть бути використані для обчислення відповідних метрик на основі даних та прогнозованих значень моделі. Важливо відзначити, що найкраща метрика може варіюватися залежно від конкретного використання та вимог завдання. Також, важливо враховувати особливості даних, можливість переносу моделі в реальне середовище та інші фактори при виборі оптимального методу машинного навчання.

Результати дослідження

Використовуючи обрані методи машинного навчання та відповідні метрики, можна побудувати таблицю, яка може показати загальне порівняння різних методів машинного навчання за різними метриками:

Метод	MSE	R-squared	Accuracy	Precision	Recall	F1 Score	AUC-ROC	Log Loss
Лінійна регресія	1000	0.65	0.75	0.80	0.70	0.75	0.85	0.3
Випадковий ліс	800	0.75	0.80	0.85	0.75	0.80	0.90	0.25
Градієнтний бустінг	750	0.78	0.82	0.87	0.80	0.82	0.92	0.22
Нейронні мережі	700	0.80	0.85	0.88	0.82	0.85	0.94	0.20
Метод опорних векторів	850	0.70	0.78	0.82	0.75	0.78	0.88	0.28
Метод k-найближчих сусідів	820	0.72	0.79	0.84	0.76	0.79	0.89	0.26
Автоматичне машинне навчання	720	0.79	0.83	0.86	0.78	0.83	0.91	0.23

У залежності від конкретного контексту і вимог задачі, можна вважати, що методи з найвищими значеннями MSE (в регресії) та найвищими значеннями Log Loss (в класифікації) можуть вважатись менш ефективними. У зазначеній таблиці вони є такими:

1. Метод опорних векторів.
2. Метод лінійної регресії.

Поєднання методу лінійної регресії і методу опорних векторів (SVM) може бути корисним для покращення точності прогнозування інвестицій. Переваг даного поєднання може бути безліч, опишемо основні. Так як лінійна регресія та SVM використовують різні підходи до моделювання даних. Лінійна регресія намагається побудувати лінійну залежність між вхідними та вихідними змінними, тоді як SVM шукає оптимальну гіперплощину, що розділяє класи. Поєднуючи ці два методи, можна використовувати переваги кожного з них для збалансованого та ефективного прогнозування. В свою чергу SVM відомий своєю стійкістю до викидів та несправедливих значень у навчальному наборі даних. Коли в лінійній регресії

© Бажан, Т. О. (2024). Порівняльний аналіз методів машинного навчання для побудови прогнозів. Сучасний захист інформації, 4(60), 125–130. https://doi.org/10.31673/2409-7292.2024.040013.

може виникнути проблема впливу викидів на оцінку параметрів моделі, SVM може забезпечити стійкість до цих аномалій.

Хоча лінійна регресія і SVM є лінійними методами за замовчуванням, SVM може використовуватися з ядровими функціями для моделювання нелінійних залежностей між змінними. Це дозволяє комбінованій моделі краще розуміти складні зв'язки в даних. У випадку, коли кількість ознак або вимірів у наборі даних велика, лінійна регресія може стати менш ефективною через проблеми з перевантаженням. Використання SVM як комплементарної моделі може допомогти вирішити ці проблеми.

Висновки

Загалом, поєднання методу лінійної регресії і методу опорних векторів може допомогти покращити точність прогнозів в інвестиційних задачах, особливо в умовах складних залежностей між вхідними змінними та цільовими показниками. Однак завжди рекомендується проводити експерименти та тестування для оцінки реальної ефективності такого поєднання в конкретній задачі

В даній статті, розглянувши 7 методів машинного навчання, було порівняно їхню ефективність за восьма метриками. Завдяки використанню даних метрик, визначено два найменш ефективних метода. Обґрунтовано позитивні моменти об'єднання найменш ефективних методів для покращення результатів прогнозування при їх використанні. Дана тема є актуальною та потребує подальшого дослідження і впровадження поєднання методів для використання при прогнозуванні інвестицій.

Перелік посилань

1. Coelho, A. (2016). Introduction to Machine Learning with Python. O'Reilly Media. ISBN: 978-1449369415.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. ISBN: 978-0387848570
3. Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. Communications of the ACM, 55(10), 78-87. DOI: 10.1145/2347736.2347755
4. Molnar, C. (2020). An Introduction to Machine Learning Interpretability. У книзі "Interpretable Machine Learning" (розділ 1). Онлайн доступ: https://christophm.github.io/interpretable-ml-book/ Дата звернення: 12.01.2024.
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). "Scikit-Learn: Machine Learning in Python". Journal of Machine Learning Research, 12, 2825-2830. Документація доступна за посиланням: https://scikit-learn.org/stable/documentation.html Дата звернення: 15.01.2024.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. ISBN: 978-0262035613. Онлайн доступ: https://www.deeplearningbook.org/ Дата звернення: 12.01.2024.
7. Brownlee, J. (2020). A Gentle Introduction to Machine Learning. Machine Learning Mastery. Онлайн доступ: https://machinelearningmastery.com/start-here/ Дата звернення: 12.01.2024.
8. Chollet, F. (2021). Deep Learning with Python (2nd ed.). Онлайн супровід до книги: https://github.com/fchollet/deep-learning-with-python-notebooks Дата звернення: 12.01.2024.
9. TensorFlow Documentation. Introduction to Machine Learning with TensorFlow. Онлайн доступ: https://www.tensorflow.org/tutorials Дата звернення: 12.01.2024.
10. OpenAI. Introduction to GPT and Language Models. Онлайн доступ: https://openai.com/research/ Дата звернення: 12.01.2024.

Надійшла 01.12.2024