ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

https://doi.org/10.31649/1997-9266-2023-170-5-25-31

УДК 004.89:159.944

С. С. Гладіголов[1]
О. Б. Мокін[1]

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ В ЗАДАЧІ ПЕРЕДБАЧЕННЯ ВИГОРАННЯ СПІВРОБІТНИКІВ

[1]Вінницький національний технічний університет

Розглянуто задачу передбачення синдрому емоційного вигорання співробітників, актуальність якої пов'язана з високим рівнем стресу в сучасному світі. У дослідженні використано публічний набір даних "Are your employees burning out" зі змагання на платформі HackerEarth. Проведено порівняльний аналіз трьох традиційних моделей машинного навчання, основаних на класичних підходах машинного навчання (лінійна регресія, Random Forest, XGBoost) та трьох баєсових моделей (баєсова лінійна регресія, модель регресії зі змінним вільним членом, модель регресії зі змінним вільним членом та кутовим коефіцієнтом). Досліджено зміну якості моделей на різних розмірах наборів даних, починаючи від 13000 (тобто від повної тренувальної вибірки, яка склала 70 % від всіх даних) до 25 спостережень включно з перевіркою на повному наборі даних. Продемонстровано, що за великих обсягів даних найкращою моделлю є XGBoost. Однак зі зменшенням розміру тренувальної вибірки до менше ніж 5000 спостережень валідаційні показники XGBoost моделі суттєво погіршилися та стали нижчими ніж відповідні значення метрик для баєсових моделей. Після оптимізації таких гіперпараметрів, як глибина дерев, кількість дерев, швидкість навчання та інші, якість XGBoost суттєво покращилась, але не зробила її достатньо стійкою, щоб продемонструвати кращі результати, ніж баєсові моделі на вибірках менше 600 спостережень. Баєсові ж моделі окрім кращої якості на малих вибірках також дозволяють оцінювати «впевненість» у прогнозованих значеннях, що є важливою особливістю для низки задач. Проте, вони мають і значний недолік у вигляді набагато більшої обчислювальної складності, що призводить до збільшення часу навчання. У висновку підкреслено важливість ретельного вибору моделі, яка враховує особливості обсягу та якості наявних даних. Баєсові моделі проявили високу ефективність у разі невеликого обсягу даних, завдяки їхньої здатності враховувати невизначеність та недостатність інформації.

Ключові слова: машинне навчання, баєсові моделі, синдром вигорання, малі набори даних.

Вступ

Сучасний світ стикається зі значним зростанням рівня стресу, що визначається різноманітними факторами, які включають в себе надмірну робочу завантаженість, невизначеність в професійному майбутньому, соціальні та особисті проблеми. Цей стрес, з яким не завжди вдається ефективно справлятися, може призвести до вигорання — стану, коли працівник втрачає інтерес та енергію для виконання своїх робочих обов'язків. Вигорання впливає не лише на ефективність роботи, але й загрожує здоров'ю та загальному самопочуттю [1], [2].

Розуміючи важливість цієї проблеми, деякі прогресивні компанії вже активно інвестують в системи відстежування емоційного стану своїх співробітників [3]. Однак, незважаючи на цей позитивний крок, наукових досліджень та розробок в галузі передбачення вигорання за допомогою методів машинного навчання все ще не вистачає для того, щоб розробити систему передбачення та попередження вигорання.

Однією з можливих причин цього є обмежена доступність відкритих даних та складність їхнього збору. Тому метою статті є проведення аналізу якості різних моделей машинного навчання, використовуючи різні розміри тренувальної вибірки, на основі відкритого набору даних «Are your

employees burning out» [4].

Це дослідження допоможе визначити, які моделі краще справляються з поставленою задачею в умовах роботи з наборами даних різного розміру та може бути використана для поліпшення прогнозування виникнення професійного вигорання, допомагаючи компаніям та організаціям попереджувати цей стан та зберігати здоров'я та продуктивність своїх працівників.

Вибір набору даних

Для проведення порівняльного аналізу моделей машинного навчання в задачі передбачення вигорання співробітників використано набір даних під назвою "Are Your Employees Burning Out", який був представлений як змагання на платформі HackerEarth [4]. Цей набір даних став основою для цього дослідження та містить інформацію про співробітників різних компаній з погляду їхнього стану щодо вигорання на робочому місці. Він складається з 22750 записів та включає такі ознаки:

– Employee ID — унікальний ідентифікатор кожного співробітника;
– Date of Joining ID — дата, коли співробітник почав працювати в компанії;
– Gender ID — стать співробітника;
– Company Type ID — тип компанії, у якій працює співробітник (продуктова чи сервісна);
– WFH Setup Available ID — бінарна ознака, яка показує, чи є можливість працювати віддалено;
– Designation ID — значення від 0 до 5, де більше число означає вищу посаду в компанії;
– Resource Allocation ID — кількість робочих годин;
– Mental Fatigue Score ID — оцінка рівня психологічної виснаженості співробітника;
– Burn Rate ID — цільова ознака, яку слід передбачити — рівень вигорання співробітника.

Для підготовки даних до подальшого аналізу бінарні змінні, такі як стать та наявність віддаленої роботи, перетворено до формату 0 і 1 для полегшення подальшої обробки. Інформацію про день початку роботи в компанії перетворено в роки досвіду співробітника, щоб отримати числову ознаку, яка може бути використана в моделях.

Для оцінки якості моделей машинного навчання набір доступних даних розділено на навчальний та тестовий набори. Тестовий набір становив 30 % від загальної кількості даних і залишався незмінним протягом усього дослідження. Для порівняння ефективності різних моделей, використано коефіцієнт детермінації, оскільки цей показник визначено як цільову метрику в оригінальному змаганні.

Для повноти аналізу ефективності вибраних моделей для заданого класу задач, а також беручи до уваги те, що набори даних у галузі психології часто є обмеженими за обсягом (через приватний характер інформації, складність створення опитувань тощо), проведемо порівняння на різних розмірах тренувальних вибірок: почнемо з повного набору даних і будемо поступово зменшувати тренувальний набір даних, залишаючи тестовий без змін для об'єктивності порівняння.

Вибір моделей машинного навчання

Подамо результати порівняльного аналізу шести моделей машинного навчання, які використано для передбачення вигорання співробітників. Цей аналіз включає в себе як традиційні, так і баєсові моделі, опис параметрів та результати експерименту на різних розмірів тренувальних наборів даних.

Частина моделей названі традиційними, через заснованість на частотному підході до теорії ймовірностей (що є традиційнішим, оскільки більшість книжок та університетських курсів починаються саме з цього підходу) та набагато поширенішим, ніж баєсові моделі машинного навчання.

Баєсові моделі машинного навчання відрізняються тим, що засновані на баєсовому підході до теорії ймовірностей. Замість невідомих коефіцієнтів у моделях розподіли з невідомими параметрами. На основі апріорної інформації про ці розподіли та наявних спостережень вираховуються апостеріорні параметри. Такий підхід дозволяє враховувати в самій структурі моделі специфічні знання предметної області та розуміти наскільки модель «впевнена» в своїх прогнозах, на відміну від традиційних моделей, які дають точкові оцінки. Але великим недоліком баєсових моделей є набагато більша обчислювальна складність.

Отже, розділимо вибрані шість моделей машинного навчання на дві групи: традиційні та баєсові.

До традиційних моделей віднесемо:

• Лінійна регресія (далі на графіках і в таблицях — *linear*): Модель базується на лінійному відношенні між вхідними змінними і виходом.

• Random Forest (далі на графіках і в таблицях — *random_forest*): Ансамбль рішень дерев для передбачення результатів [5].

• XGBoost (далі на графіках і в таблицях — *xgb*): Ансамбль градієнтного бустинга, який використовує дерева рішень для покращення точності передбачень [6].

Баєсові моделі написані на мові програмування Python з використанням бібліотеки PyMC [7]. Для порівняльного аналізу вибрано такі базові баєсові моделі:

• Баєсова лінійна регресія (далі на графіках і в таблицях — *bayes_linear*), яка використовує баєсовий підхід до лінійної регресії, що дозволяє враховувати невизначеність в параметрах моделі.

• Модель регресії зі змінним вільним членом (Varying intercept model) (далі на графіках і в таблицях — *bayes_var_int*), яка розширює баєсову лінійну регресію, додаючи ієрархічність параметрів і можливість варіювати параметр вільного члена [8].

• Модель регресії зі змінним вільним членом та кутовим коефіцієнтом (Varying intercept and slope model) (далі на графіках і в таблицях — *bayes_var_int_slp*), яка доповнює попередню баєсову лінійну регресію, додаючи ієрархічність параметрів і можливість варіювати параметри вільного члену та кутового коефіцієнта [8].

Баєсові моделі, використані у цьому дослідженні, застосовують баєсовий підхід до статистичного моделювання. Вони дозволяють моделі враховувати невизначеність у параметрах і апріорні розподіли цих параметрів. Завдяки своїм відмінностям від традиційних статистичних моделей баєсові моделі можуть бути особливо корисним в задачах з невеликими обсягами даних, де інформація має пропуски чи є неповною.

Результати аналізу

Натренувавши всі вибрані моделі на всіх вибірках даних, отримано такі результати. Модель XGBoost, що тренувалася на повному наборі даних, показала найкраще валідаційне значення коефіцієнта детермінації для передбачення вигорання співробітників. Однак зі зменшенням розміру тренувальної вибірки до 5000 навчальних даних, показники XGBoost моделі суттєво знизилися. В умовах навчання з меншим набором даних найкращою моделлю стала модель регресії зі змінним вільним членом, що є свідченням важливості врахування особливостей тренувальної вибірки в задачі передбачення вигорання співробітників.

Зауважимо, що на цьому етапі для забезпечення порівнюваності результатів для всіх традиційних моделей (лінійна регресія, Random Forest та XGBoost) використано параметри за замовчуванням, а для баєсових моделей використані нормальні розподіли без додаткових обмежень як апріорних.

На рис. 1—3 показана зміна валідаційного значення коефіцієнта детермінації для всіх вибраних та натренованих моделей для різних розмірів тренувальної вибірки.

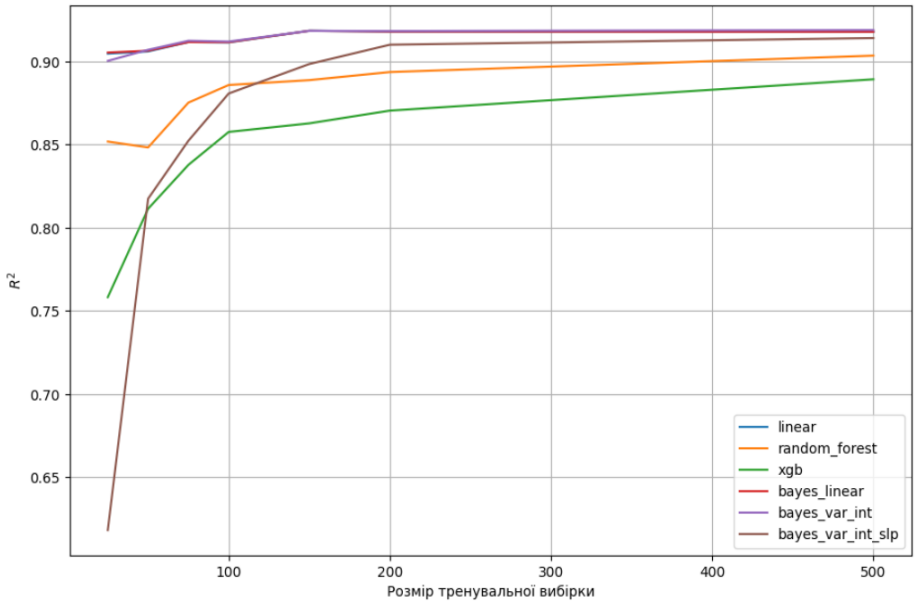

Рис. 1. Значення коефіцієнта детермінації для тренувальних вибірок з кількістю прикладів від 25 до 500

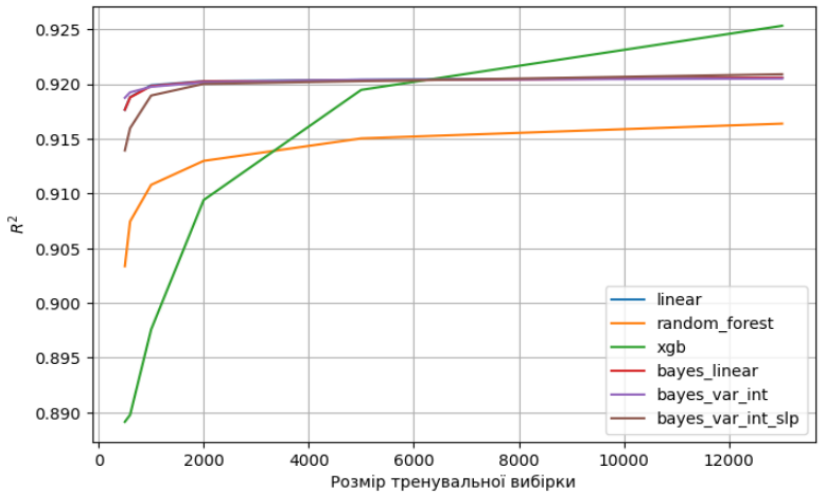

Рис. 2. Значення коефіцієнта детермінації для тренувальних вибірок з кількістю прикладів від 500 до 13000

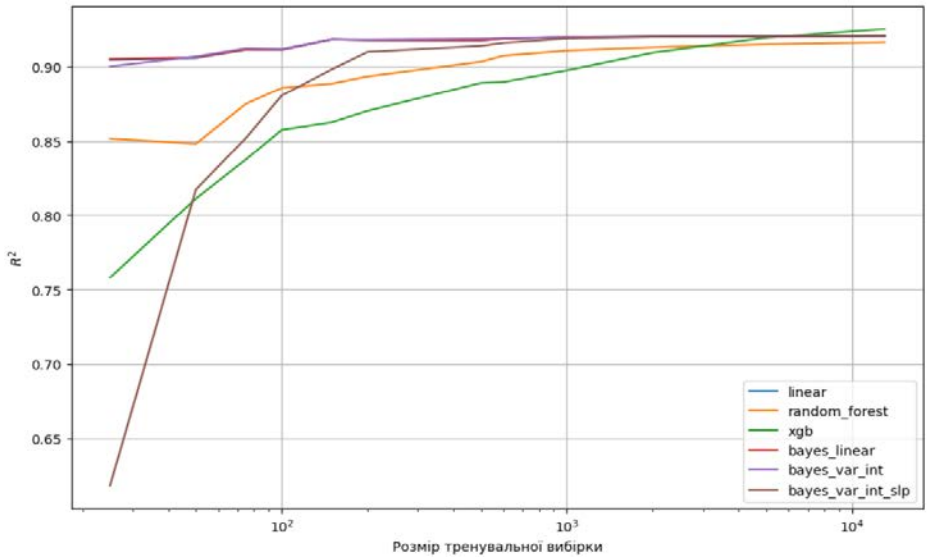

Рис. 3. Значення коефіцієнта детермінації для всього діапазону значень розмірів тренувальної вибірки (від 25 до 13000) у логарифмічному масштабі

Всі розміри тренувальних вибірок наведено у табл. 1.

Таблиця 1

Розміри тренувальних вибірок

Розмір вибірки	linear	random_forest	xgb	bayes_linear	bayes_var_int	bayes_var_int_slp
13013 (всі)	0,9206	0,9164	**0,9253**	0,9206	0,9205	0,9209
5000	**0,9204**	0,9150	0,9195	0,9203	**0,9204**	0,9203
2000	**0,9203**	0,9130	0,9094	0,9202	0,9202	0,9200
1000	**0,9199**	0,9108	0,8976	0,9198	0,9197	0,9189
600	0,9188	0,9074	0,8898	0,9188	**0,9192**	0,9160
500	0,9176	0,9034	0,8891	0,9176	**0,9187**	0,9139
200	0,9178	0,8935	0,8703	0,9178	**0,9181**	0,9099
150	**0,9183**	0,8886	0,8626	**0,9183**	0,9182	0,8983
100	0,9113	0,8857	0,8575	0,9114	**0,9118**	0,8806
75	0,9115	0,8751	0,8377	0,9115	**0,9123**	0,8522
50	0,9059	0,8482	0,8113	0,9063	**0,9069**	0,8173
25	0,9046	0,8517	0,7582	**0,9052**	0,9002	0,6182

Порівняння коефіцієнтів детермінації вибраних моделей на вибірках різного розміру. Порівняння з покращеною XGBoost моделлю

На повному наборі даних модель XGBoost виявилася найефективнішою. Однак зі зменшенням розміру тренувальної вибірки якість її результатів різко падає. Тому тепер ускладнимо задачу для баєсових моделей, виконавши підбір гіперпараметрів для моделі XGBoost для кожної з тренувальних вибірок. Тобто далі будуть розглянуті результати порівняння моделі XGBoost, налаштованої за допомогою пошуку значень її гіперпараметрів за сіткою (Grid Search) та перехресною валідацією за п'ятьма фолдами, з баєсовими моделями на різних розмірах тренувального набору даних.

Для кожної тренувальної вибірки виконувався пошук найкращих гіперпараметрів (див. рис. 4), таких як: глибина дерев (max_depth), кількість дерев (n_estimators), швидкість навчання (learning rate) та інші, і вимірювалася якість моделей за допомогою відповідних метрик. Після завершення пошуку ці оптимальні параметри використано для навчання найкращої моделі для відповідного тренувального набору.

```
param_grid = {
    'n_estimators': [20, 50, 100, 150, 200],
    'max_depth': list(range(3, 13)),
    'subsample': [1, 0.5],
    'learning_rate': [0.001, 0.005, 0.01, 0.03, 0.06, 0.1, 0.3, 0.6]
}
```

Рис. 4. Сітка гіперпараметрів для вибору кращої моделі XGBoost

Як видно з рис. 5, табл. 2 та рис. 6, результати оптимізації параметрів XGBoost показали очевидне підвищення якості моделі у разі зменшення розміру тренувального набору. Однак важливо відмітити, що навіть після ретельного підбору гіперпараметрів валідаційні показники XGBoost моделі для тренувальних вибірок менших 600 прикладів гірші за баєсові моделі.

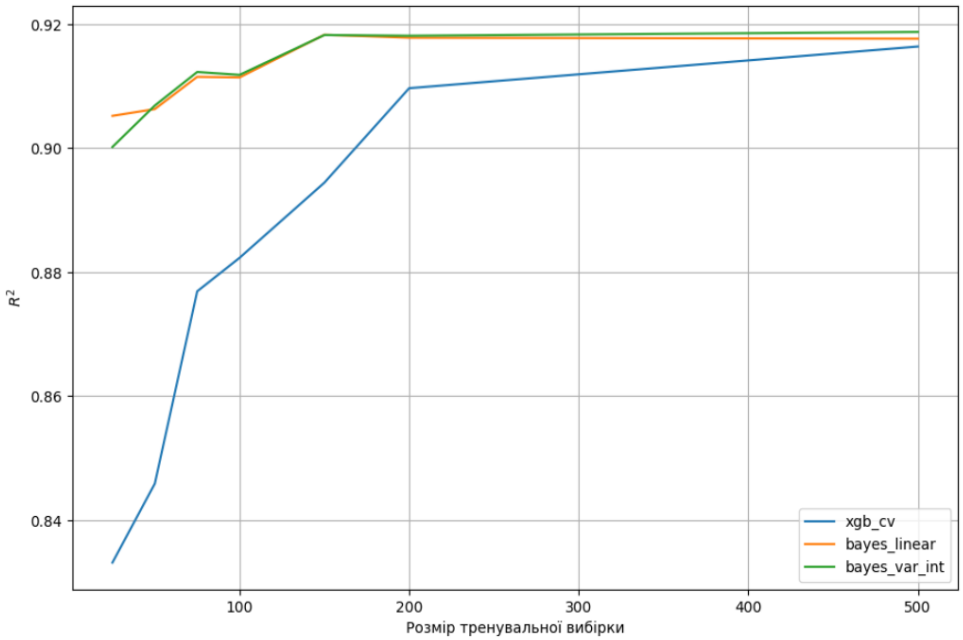

Рис. 5. Значення коефіцієнта детермінації для розмірів тренувальної вибірки від 25 до 500 для оптимізованої XGBoost моделі та двох кращих баєсових моделей

Таблиця 2

Порівняння коефіцієнтів детермінації баєсових та кращих XGBoost моделей

Розмір вибірки	bayes_linear	bayes_var_int	bayes_var_int_slp	xgb_cv
13013	0,9206	0,9205	0,9209	**0,9279**
5000	0,9203	0,9204	0,9203	**0,9257**
2000	0,9202	0,9202	0,9200	**0,9232**
1000	0,9198	0,9197	0,9189	**0,9213**

Продовження табл. 2

Розмір вибірки	bayes_linear	bayes_var_int	bayes_var_int_slp	xgb_cv
600	0,9188	**0,9192**	0,9160	0,9173
500	0,9176	**0,9187**	0,9139	0,9164
200	0,9178	**0,9181**	0,9099	0,9097
150	**0,9183**	0,9182	0,8983	0,8944
100	0,9114	**0,9118**	0,8806	0,8823
75	0,9115	**0,9123**	0,8522	0,8769
50	0,9063	**0,9069**	0,8173	0,8459
25	**0,9052**	0,9002	0,6182	0,8332

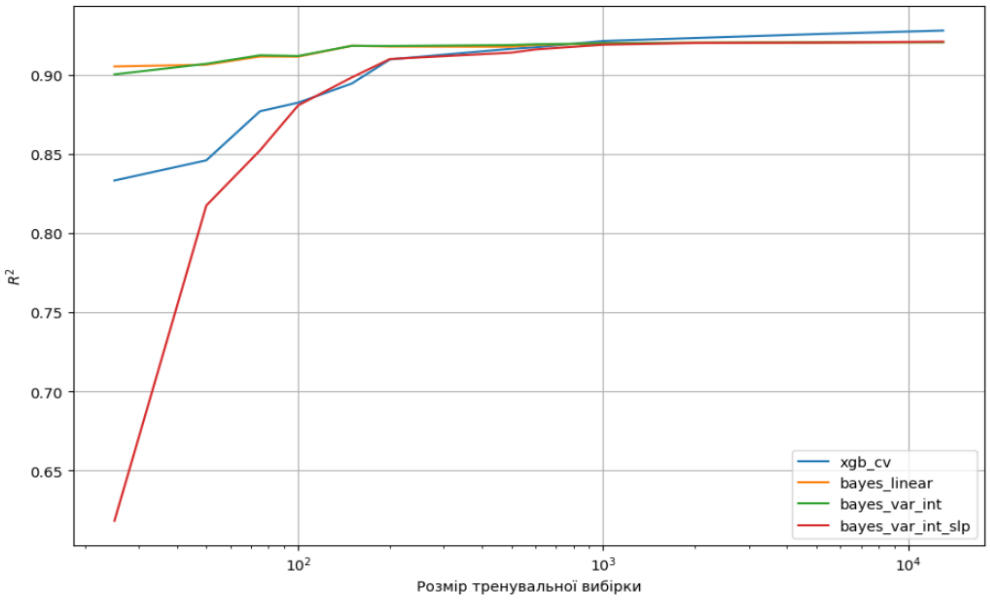

Рис. 6. Значення коефіцієнта детермінації для розмірів тренувальної вибірки від 25 до 13000 у логарифмічній шкалі для оптимізованої XGBoost моделі та двох кращих баєсових моделей

Висновки

Проведено порівняльний аналіз моделей машинного навчання для задачі передбачення вигорання співробітників. Розглянуто три традиційні моделі (лінійна регресія, Random Forest, XGBoost) та три баєсові моделі (баєсова лінійна регресія, модель регресії зі змінним вільним членом, модель регресії зі змінним вільним членом та кутовим коефіцієнтом).

Виявлено, що на повному тренувальному наборі даних модель XGBoost показала найкращу якість передбачення вигорання співробітників. Однак зі зменшенням розміру тренувальної вибірки до менше ніж 600 спостережень, навіть після ретельного підбору гіперпараметрів, валідаційні показники XGBoost моделі суттєво погіршилися і стали нижчими ніж відповідні значення метрик для баєсових моделей.

Цей висновок підкреслює важливість вибору моделі, яка враховує особливості обсягу та якості наявних даних. Баєсові моделі дозволяють враховувати невизначеність та недостатність інформації та можуть бути ефективнішими в умовах невеликого обсягу даних. Оптимізація параметрів XGBoost допомогла покращити його якість, але не зробила його достатньо стійким, щоб показати кращі результати, ніж баєсові моделі на малих вибірках.

Отже, вибір моделі машинного навчання для задачі передбачення вигорання співробітників повинен бути обґрунтованим і враховувати конкретні умови дослідження, зокрема обсяг та якість наявних даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] D. A. J. Salvagioni, F. N. Melanda, A. E. Mesas, A. D. González, F. L. Gabani and S. M. de Andrade, “Physical, psychological and occupational consequences of job burnout: A systematic review of prospective studies,” *PLOS ONE*, no. 12, pp. e0185781, October 2017.

[2] M. C. & I. C., “The Role of the Stress in Development of the Diseases: Array,” *Precarpathian Bulletin of the Shevchenko Scientific Society Pulse*, pp. 25-32, October 2019.

[3] М. Гурська, «Я вигорів і боюсь звільнення — що робити? Топові IT-компанії відповіли, як вони реагують на вигоряння у працівників та кандидатів,» DOU.ua, 15.11.2022. [Електронний ресурс]. Режим доступу: https://dou.ua/lenta/articles/emotional-burnout-at-work . Дата звернення: 20.09.2023.

[4] “Hacker Earth Machine Learning Challenge: Are your employees burning out?” HackerEarth, 21.10.2021. [Online]. Available: https://www.hackerearth.com/challenges/new/competitive/hackerearth-machine-learning-challenge-predict-burnout-rate. Accessed on: 20.09.2023.

[5] L. Breiman, “Random Forests,” *Machine Learning*, no. 45, pp. 5-32, 2001.

[6] T. Chen, and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” в *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016.

[7] O. Abril-Pla, et al. “PyMC: a modern, and comprehensive probabilistic programming framework in Python,” *PeerJ Computer Science*, no. 9, pp. e1516, September 2023.

[8] A. Gelma, and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2006.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 26.09.2022

Гладіголов Сергій Сергійович — аспірант, кафедри системного аналізу та інформаційних технологій, e-mail: hladiholov.s@gmail.com ;

Мокін Олександр Борисович — д-р техн. наук, професор, професор кафедри системного аналізу та інформаційних технологій, e-mail: abmokin@gmail.com .

Вінницький національний технічний університет, Вінниця

S. S. Hladiholov[1]
O. B. Mokin[1]

Comparative Analysis of Machine Learning Models for Predicting Employee Burnout Problem

[1]Vinnytsia National Technical University

The article explores the problem of predicting the emotional burnout syndrome of employees , which is relevant due to the high level of stress in the modern world. The study uses the publicly available dataset "Are your employees burning out" from the competition on the HackerEarth platform. A comparative analysis of three traditional machine learning models based on classical machine learning approaches (linear regression, Random Forest, XGBoost) and three Bayesian models (Bayesian linear regression, varying intercept model, varying intercept and slope model) was carried out in the study. The change in the quality of the models is studied for different sizes of data sets, ranging from 13,000 (i.e., the full training set, which accounted for 70% of all data) to 25 observations, including testing on the full data set. It is demonstrated that XGBoost is the best model for large data sets. However, when the training sample size is reduced to less than 5000 observations, the validation performance of the XGBoost model becomes significantly less accurate and becomes lower than the corresponding metrics for Bayesian models. After optimizing such hyperparameters as tree depth, number of trees, learning rate, and others, the quality of XGBoost improved significantly, but did not make it stable enough to demonstrate better results than Bayesian models on samples of less than 600 observations. Bayesian models, on the other hand, in addition to being better on small samples, also allow estimating the "confidence" in the predicted values, which is an important feature for a specific tasks. However, they also have a significant disadvantage in the form of much greater computational complexity, which leads to an increase in training time. In conclusion, results of this study emphasize the importance of careful selection of a model that considers the peculiarities of the amount and quality of available data. Bayesian models have proven to be highly effective with a small amount of data, due to their ability to consider uncertainty and insufficient information.

Keywords: machine learning, Bayesian models, burnout syndrome, small data sets.

Hladiholov Serhii S. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: hladiholov.s@gmail.com ;

Mokin Oleksandr B. — Dr. Sc. (Eng.), Professor, Professor of the Chair of System Analysis and Information Technologies, e-mail: abmokin@gmail.com