

ІНТЕЛЕКТУАЛЬНА СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ АНАЛІЗУ ЕКОНОМІЧНИХ РИЗИКІВ

Мельник М.С.¹, Левенчук Л.Б.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

¹ melnyk.mykola@lil.kpi.ua

У роботі досліджено розробку інтелектуальної системи підтримки прийняття рішень (ІСППР) для аналізу економічних ризиків та прогнозування банкрутства компаній. Метою є порівняння ефективності сучасних методів машинного навчання для цієї задачі. Дослідження включає застосування низки моделей, зокрема DecisionTree, RandomForest, SVM, GradientBoosting та XGBoost. Методи дослідження зосереджені на подоланні проблеми незбалансованості вхідних даних за допомогою техніки SMOTE та відборі найбільш значущих ознак з використанням RFE. Основні отримані результати демонструють, що хоча SVM показує найвищий F1 Score (~0.96), XGBoost є беззаперечним лідером за метрикою ROC AUC (~0.85), що вказує на його кращу розрізнявальну здатність. Наукова новизна полягає у комплексному порівнянні моделей в умовах дисбалансу класів та акценті на інтерпретованості результатів за допомогою методів, як-от SHAP. Практична значимість роботи полягає у проектуванні архітектури веб-орієнтованої СППР, яка надає ризик-менеджерам прозорий та ефективний інструмент для прийняття обґрунтованих рішень.

Ключові слова: інтелектуальна система підтримки прийняття рішень (ІСППР), аналіз економічних ризиків, прогнозування банкрутства, машинне навчання, XGBoost, SVM, незбалансовані дані, SMOTE.

1. ВСТУП

Сучасне економічне середовище є надзвичайно складною і динамічною системою, що піддається впливу безлічі факторів – від макроекономічних показників до поведінкових аспектів ринку [1]. Ефективний аналіз економічних ризиків у таких умовах вимагає використання аналітичних інструментів, здатних не лише швидко обробляти великі обсяги інформації, але й виявляти приховані закономірності та прогнозувати ймовірність настання кризових подій. У цьому контексті, інтелектуальні системи підтримки прийняття рішень (ІСППР) стають незамінним інструментом.

Такі системи є критично важливими для осіб, що приймають швидкі стратегічні рішення: фінансових аналітиків, ризик-менеджерів та керівників, які відповідають за стабільність компанії. ІСППР, що використовують потужні моделі машинного навчання, надають таким фахівцям змогу отримувати поглиблений аналіз та прогнози, що ґрунтуються на об'єктивних даних.

Застосування сучасних ансамблевих методів, таких як XGBoost та LightGBM, дозволяє досягти високої якості прогнозів, враховуючи навіть таку складну проблему, як незбалансованість вхідних даних [2].

2. ОПИС НАБОРУ ДАНИХ

В даній роботі було використано набір даних, який відображає історичні фінансові дані американських компаній. У вибірці містяться 18 фінансових показників (ознаки X1–X18, такі як: поточні активи X1, чистий прибуток X6, ринкова вартість X8, X12 та загальні активи X10). Цільова характеристика – status_label – відображає стан компанії («alive» або «failed»). Моделі машинного навчання мають певні вимоги до формату вхідних даних і, зокрема, більшість моделей потребує виключно чисельні значення. Оскільки цільова характеристика status_label в нашому наборі представляє категоріальні значення («alive»/«failed»), постала необхідність в її конвертуванні в чисельний формат (де «alive» = 0 та «failed» = 1). На наступному етапі було виконано масштабування всієї сукупності фінансових показників X1-X18 для покращення подальших результатів побудованих моделей.

Для аналізу вибірки була побудована кореляційна матриця (рис. 1). Аналіз матриці виявив наступні залежності.

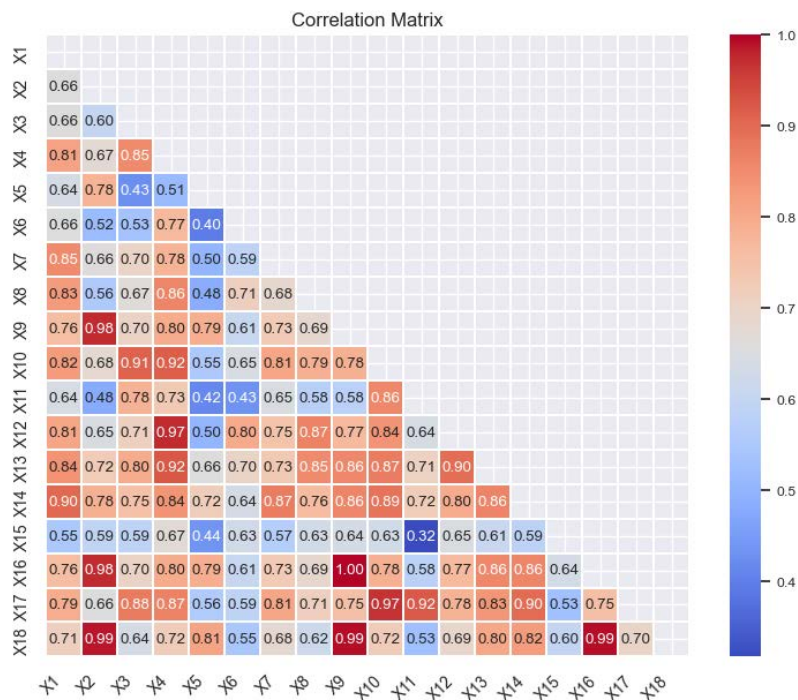


Рисунок 1. Кореляційна матриця

Сильні позитивні кореляції. Було виявлено зв'язки, що вказують на те, що змінні мають тенденцію зростати разом. Найсильніші з них спостерігаються між X3 та X16 (0.83), а також між X15 та X18 (0.78).

Помірні кореляції. Спостерігається значуща, але не абсолютна залежність, наприклад, між X7 та X14 (0.75).

Слабкі кореляції. Деякі пари, як-от X13 та X14 (0.01), демонструють практично повну відсутність лінійного зв'язку.

Негативні кореляції. Матриця також показала зворотні зв'язки, де збільшення однієї змінної відповідає зменшенню іншої, наприклад, між X1 та X15 (–0.67).

Ці результати вказують на потенційні проблеми мультиколінеарності, що може негативно вплинути на побудову прогностичних моделей. Для зменшення надмірності даних може знадобитися відбір ознак або зменшення розмірності.

3. ОПИС ВИКОРИСТАНИХ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

У роботі було досліджено та порівняно низку моделей машинного навчання для вирішення задачі прогнозування банкрутства. Оскільки дані для такого типу завдань зазвичай є сильно незбалансованими (кількість компаній-банкрутів значно менша за кількість фінансово здорових), ключовим етапом підготовки даних стало застосування методу SMOTE (Synthetic Minority Over-sampling Technique). Цей підхід дозволяє збалансувати класи шляхом генерації синтетичних прикладів для класу меншості, що є критично важливим для коректного навчання моделей та уникнення їх упередженості в бік мажоритарного класу [3].

Основні методи, що розглянуті в роботі, включають:

Дерево рішень (DecisionTree) є одним з базових методів керованого машинного навчання, що застосовується для задач класифікації та регресії. Модель являє собою ієрархічну деревоподібну структуру, де кожен внутрішній вузол відповідає перевірці певної ознаки (фінансового показника), кожна гілка – результату цієї перевірки, а кожен листковий вузол – кінцевому рішенню або класу (наприклад, «alive» або «failed»). Побудова дерева відбувається шляхом рекурсивного поділу вибірки на підмножини на основі критеріїв, що максимізують чистоту вузлів. Незважаючи на високу інтерпретованість, метод схильний до перенавчання [4].

Випадковий ліс (RandomForest) – це ансамблевий метод машинного навчання, що розширює концепцію дерев рішень для підвищення їхньої стабільності та точності. Алгоритм будує велику кількість (ансамбль) незалежних дерев рішень під час тренування, причому кожне дерево навчається на випадковій підмножині даних (бутстреп-вибірка) та з використанням випадкової підмножини ознак для кожного поділу. Кінцевий прогноз для задачі класифікації визначається шляхом голосування більшості дерев в ансамблі, що дозволяє нівелювати схильність окремих дерев до перенавчання та покращити узагальнюючу здатність моделі [4].

Метод опорних векторів (Support Vector Machine, SVM) є потужним алгоритмом керованого навчання, що використовується для лінійної та нелінійної класифікації. Основна ідея методу полягає у побудові оптимальної розділяючої гіперплощини у багатовимірному просторі ознак, яка б максимізувала зазор (margin) між найближчими точками різних класів (опорними векторами). Для нелінійно розділюваних даних SVM використовує «ядровий трюк» (kernel trick) для проектування даних у простір вищої розмірності, де лінійний поділ стає можливим, що робить його ефективним для складних задач класифікації [4].

Гradientний бустинг (GradientBoosting) – це ансамблевий метод машинного навчання, що будує моделі у послідовний, поетапний спосіб. На відміну від випадкового лісу, де дерева будуються паралельно та незалежно, gradientний бустинг формує нову модель (зазвичай дерево рішень) на кожній ітерації таким чином, щоб вона виправляла помилки (залишки) попередньої моделі. Навчання відбувається шляхом оптимізації функції втрат за допомогою методу gradientного спуску, що дозволяє ансамблю поступово «посилувати» (boost) слабкі класифікатори, створюючи високоточну прогностичну модель [4].

Екстремальний gradientний бустинг (eXtreme Gradient Boosting) є не стільки окремим алгоритмом, скільки вискоєфективною та масштабованою реалізацією методу gradientного бустингу [2]. Його архітектура оптимізована для швидкодії та продуктивності, зокрема за рахунок паралелізації обчислень при побудові дерев та ефективної обробки пропущених значень. Ключовою відмінністю є включення L1 та L2 регуляризації (штрафів за

складність моделі) безпосередньо у функцію мети, що дозволяє ефективно контролювати перенавчання та часто забезпечує вищу точність порівняно зі стандартними реалізаціями градієнтного бустингу.

Також у роботі було застосовано метод рекурсивного усунення ознак (RFE – Recursive Feature Elimination). Цей підхід використовується для відбору найбільш значущих фінансових показників, дозволяючи зменшити розмірність даних і складність моделей, що в свою чергу може покращити їхню здатність до узагальнення та знизити ризик перенавчання.

4. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Результати порівняння продуктивності моделей, представлені на графіках (рис. 2–4), демонструють суттєві відмінності в їхній ефективності залежно від обраної метрики. Було використано наступні метрики: accuracy, precision, recall, F1, AUC-ROC і крива precision-recall [5].

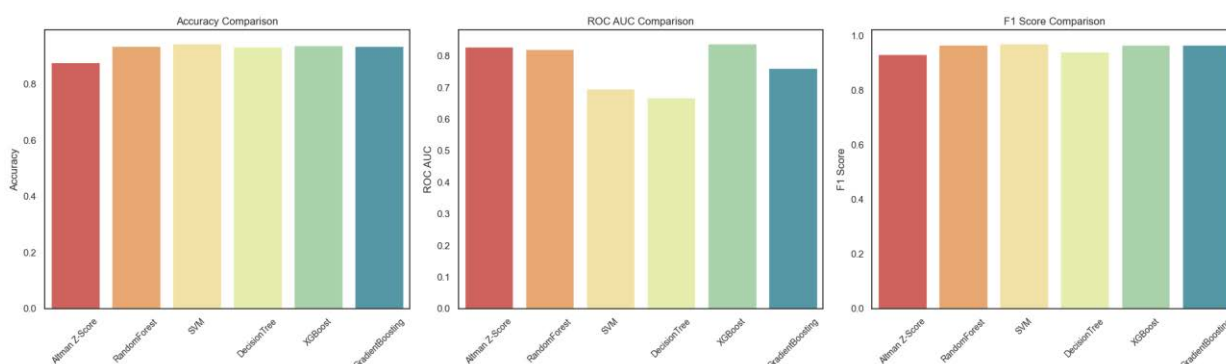


Рисунок 2. Метрики адекватності моделей на тестовій вибірці

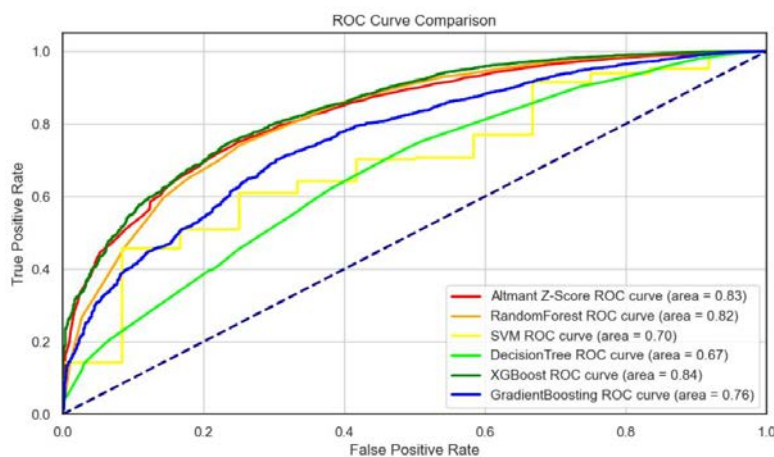


Рисунок 3. ROC-криві побудованих моделей

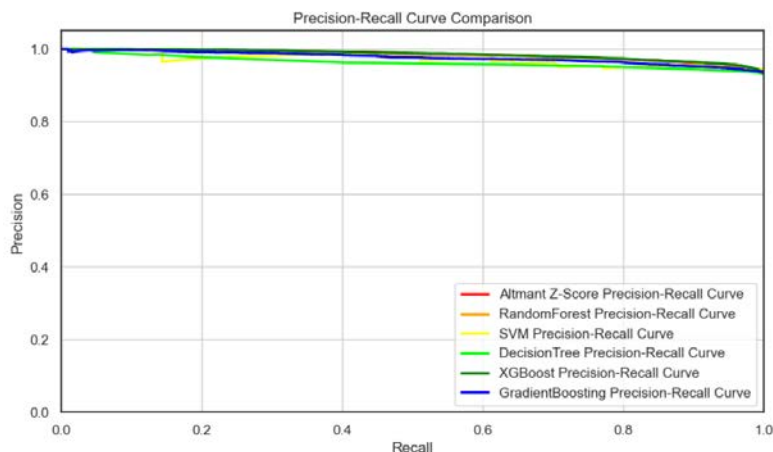


Рисунок 4. Precision-Recall криві побудованих моделей

За метрикою accuracy (рис. 2), всі моделі машинного навчання (RandomForest, SVM, DecisionTree, XGBoost, GradientBoosting) показали дуже високі та схожі результати, що коливаються в діапазоні ~ 0.93 – 0.94 . Це значно перевищує базовий показник Altman Z-Score (~ 0.88). Модель SVM демонструє незначну перевагу (~ 0.94).

При аналізі F1 Score (рис. 2), яка є більш показовою для даної задачі, SVM виразно лідує з показником приблизно 0.96. RandomForest, XGBoost та GradientBoosting також мають дуже високі результати (~ 0.95). Це свідчить про те, що SVM найкраще збалансувала виявлення компаній-банкрутів, не створюючи при цьому надмірної кількості помилкових тривог.

Однак, аналіз ROC AUC (рис. 2) показує кардинально іншу картину. XGBoost є беззаперечним лідером з показником AUC ~ 0.85 , що вказує на його потужну здатність до розрізнення класів. RandomForest (~ 0.82) та, що цікаво, традиційний Altman Z-Score (~ 0.83) також демонструють високу ефективність. Натомість, SVM (~ 0.69) та DecisionTree (~ 0.67) показують значно слабші результати, що лише трохи краще за випадкове вгадування.

5. АРХІТЕКТУРА СППР

Архітектуру СППР наведено на рисунку 5.

СППР складається з наступних модулів:

- **Веб-застосунок.** Це загальна оболонка системи, з якою взаємодіє користувач.
- **Вікно: Введення даних.** Це вікно має одну дію: «Вибір файлу вхідних даних», воно відповідає за завантаження датасету.
- **Вікно: Візуалізація даних.** Користувач може обрати для перегляду графік. Доступні графіки: співвідношення збанкрутілих компаній до живих, матриця кореляцій та викиди.
- **Вікно: Підготовка вхідних даних.** У цьому вікні користувач може обрати модель для якої підготувати дані, а саме обрати ключові ознаки та зберегти оброблені дані у окремому файлі.
- **Вікно: Навчання моделей.** Це основне вікно системи, де після обробки даних стає доступна можливість навчання моделей. Як результат виводяться ключові метрики та їх графічне відображення.
- **Вікно: Візуалізація кінцевих результатів.** Вікно для співставлення метрик навчених моделей у вигляді графіків на вибір користувача. Вигляд доступних графіків наведений на рисунках 2-4.
- **Вікно: Прогнозування результатів.** Надає користувачеві можливість обрання однієї з навчених моделей та введення вхідних даних для прогнозування ризику банкрутства.

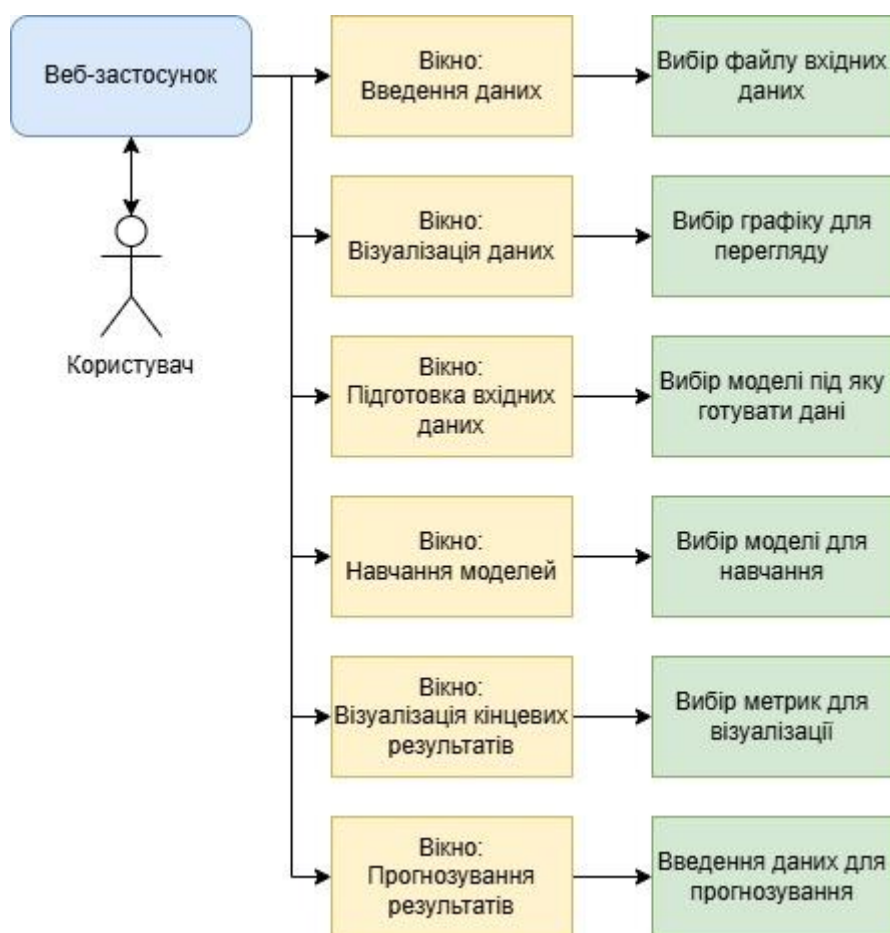


Рисунок 5. Архітектура СПДР

6. ВИСНОВКИ

У даній роботі розглянуто можливості застосування моделей машинного навчання для вирішення задачі аналізу економічних ризиків та прогнозування банкрутства компаній. Було проведено порівняльний аналіз низки алгоритмів, включаючи DecisionTree, RandomForest, SVM, GradientBoosting та XGBoost, із застосуванням методу SMOTE для вирішення проблеми незбалансованості вхідних даних. Проведений аналіз і отримані результати підтверджують доцільність використання цих моделей: зокрема, XGBoost продемонстрував найвищу розрізнявальну здатність (ROC AUC ~0.85), а SVM – найкращий баланс точності та повноти (F1 Score ~0.96). Крім того, було спроектовано архітектуру системи підтримки прийняття рішень (СПДР), що базується на даних моделях, у вигляді веб-застосунку з продуманою модульною структурою, яка забезпечує повний цикл роботи з даними та робить СПДР ефективним інструментом для ризик-менеджерів.

У рамках подальших досліджень доцільно розглянути інші методи класифікації, зокрема нейронні мережі, які можуть покращити точність прогнозів за рахунок виявлення складніших нелінійних залежностей у фінансових показниках. Перспективним напрямком є також поглиблене впровадження методів інтерпретації, як-от SHAP, безпосередньо в архітектуру СПДР, що дозволить надавати користувачам не лише прогноз, але й його миттєве пояснення. Такий підхід додасть новий рівень прозорості в оцінювання ризиків та допоможе фінансовим установам підвищити довіру до автоматизованих систем аналізу

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Python for Finance and Economics - Medium, 2025. URL: <https://medium.com/@quanticascience/python-for-finance-and-economics-baec27219da1> (дата звернення: 15.11.2025)
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
3. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5.
4. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (2009). Сполучені Штати Америки: Springer.
5. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Швейцарія: Springer New York.