Застосування машинного навчання для інвестиційного аналізу стартапів

Yurii Tomka*1*, †, Dmytro Uhryn*1*, †

1 Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 2 Kotsiubynskoho Street, Chernivtsi, Ukraine, 58002.

Abstract

У дослідженні розроблено методологію прогнозування динаміки цін акцій публічних технологічних компаній для кількісної оцінки інвестиційної привабливості стартапів. На датасеті з 14,590 щоденних спостережень для десяти компаній (AAPL, MSFT, AMZN, GOOGL, META, TSLA, NVDA, NFLX, ADBE, CRM) за період 2019-2024 рр. проведено систематичне порівняння Support Vector Regression, Random Forest, Gradient Boosting та ансамблевої архітектури Stacking. Stacking досягає найвищої точності напрямку DA=93.18%, перевершуючи SVR (90.91%) на 2.27 п.п. при RMSE=0.003762 та R²=0.9318. Аналіз важливості ознак виявляє домінування normalized Close/Open ratio (54.5%).

Розроблений фреймворк, що транслює прогнози у метрики оцінки стартапів через Sector Growth Indicator, волатильність, Value-at-Risk та композитну Score. Емпірична валідація на трьох кейсах демонструє диференціацію: SaaS (Score=1.131) - сприятлива можливість, AI/ML (Score=-0.247) - несприятливий, E-commerce (Score=-0.685) - різко негативний.

Методологія забезпечує практичний інструмент для венчурних капіталістів, доповнюючи якісну оцінку дисциплінованим кількісним базисом систематизації галузевого моментуму через публічні аналоги.

Keywords

машинне навчання, прогнозування цін акцій, венчурне інвестування, стартапи, Stacking, SVR, Random Forest, Gradient Boosting, оцінка ризиків.

[[1]](#footnote-2)

# Introduction

У сучасних умовах динамічного розвитку глобальної економіки фінансові ринки виступають ключовим елементом, що впливає на всі аспекти економічної діяльності, включаючи венчурне інвестування та розвиток стартап-проєктів. Прийняття обґрунтованих інвестиційних рішень у стартап-екосистемі вимагає не лише аналізу внутрішніх характеристик молодих компаній, але й розуміння динаміки відповідних ринкових секторів через призму публічних компаній-аналогів. Прогнозування цін акцій таких компаній може надати цінну інформацію про ринковий потенціал та ризики, притаманні стартапам відповідного напрямку [1].

Традиційні методи оцінки стартапів, такі як дисконтовані грошові потоки або мультиплікатори, часто виявляються недостатньо ефективними через відсутність стабільної історії операційної діяльності та високий рівень невизначеності [2]. Водночас динаміка публічних компаній у відповідному секторі може слугувати індикатором ринкового потенціалу. Gompers et al. [3] провели опитування 885 венчурних капіталістів з 681 фірм і виявили, що VCs надають більшого значення якості управлінської команди, ніж характеристикам бізнесу, при цьому відбір угод (deal selection) оцінюється як найважливіший фактор створення вартості серед трьох основних (пошук угод, відбір угод та додана вартість після інвестування).

Останні досягнення в галузі машинного навчання продемонстрували значний потенціал для прогнозування фінансових часових рядів. Krauss et al. [4] показали, що Random Forest досягає денної прибутковості 0.43% на даних S&P 500 за період 1992-2015, перевершуючи gradient boosting (0.37%) та глибокі нейронні мережі (0.33%). Fischer & Krauss [5] розширили це дослідження, застосувавши LSTM мережі, які показали результат 0.46% денної прибутковості з Sharpe ratio 5.8. Однак ці дослідження зосереджені переважно на технічних аспектах прогнозування та не розглядають практичне застосування для оцінки інвестиційної привабливості стартапів.

Попри значні досягнення в обох напрямках, існує помітний розрив між технічними дослідженнями прогнозування цін акцій з використанням машинного навчання та практиками венчурного інвестування. Систематичне порівняння алгоритмів Support Vector Regression (SVR), Random Forest та Gradient Boosting для задач оцінки інвестиційної привабливості стартапів через публічні аналоги залишається недостатньо вивченим. Додатково відсутня чітка методологія ідентифікації релевантних компаній-аналогів та інтерпретації прогнозів у контексті ризик-менеджменту стартапів.

Це дослідження заповнює виявлені прогалини шляхом розробки інтегрованого підходу, який поєднує прогнозування цін акцій публічних компаній з оцінкою інвестиційної привабливості стартапів. Ми проводимо систематичне порівняння SVR з радіальним базисним ядром (RBF), Random Forest Regressor та Gradient Boosting Regressor на даних десяти публічних технологічних компаній. Розробляється чіткий фреймворк для ідентифікації компаній-аналогів та інтерпретації прогнозів, а також досліджується ефективність ансамблевих методів, включаючи stacking, для підвищення точності прогнозів. Запропонована методологія інтегрує прогнозні моделі у процес ризик-менеджменту та прийняття інвестиційних рішень для стартап-проєктів, надаючи венчурним капіталістам практичний інструмент для кількісної оцінки ринкового потенціалу через аналіз динаміки публічних аналогів.

# Огляд літературних джерел.

## Методи прогнозування фінансових часових рядів

**Класичні моделі.** Прогнозування динаміки фінансових ринків має довгу історію застосування статистичних методів. Піонерською роботою у цій галузі є дослідження Box & Jenkins [6], яке заклало фундаментальні основи аналізу часових рядів через розробку методології авторегресійних інтегрованих моделей ковзного середнього (ARIMA). Метод Box-Jenkins базується на ітеративному трикроковому підході: ідентифікація моделі, оцінка параметрів та діагностична перевірка. ARIMA моделі ефективно працюють з лінійними залежностями та стаціонарними процесами, проте мають суттєві обмеження при роботі з нелінійними патернами та великими обсягами даних.

Класичні економетричні підходи, такі як GARCH (Generalized Autoregressive Conditional Heteroskedasticity), також широко застосовувалися для моделювання волатильності фінансових часових рядів. Bollerslev [7] запропонував узагальнену ARCH модель, яка дозволяє моделювати часову варіацію волатильності. Однак ці методи базуються на припущеннях лінійності та стаціонарності, що часто не відповідає реальній динаміці фінансових ринків, особливо в періоди кризових явищ.

**Машинне навчання у фінансах.** Революція в галузі штучного інтелекту суттєво змінила підходи до прогнозування цін акцій. Kim [8] продемонстрував ефективність застосування Support Vector Machines (SVM) для прогнозування напрямку руху індексу NIKKEI 225, показавши перевагу над традиційними методами дискримінантного аналізу. SVM виявилися особливо ефективними завдяки здатності працювати з нелінійними залежностями через використання ядрових функцій та мінімізацію структурного ризику.

Gupta et al. [9] розширили застосування Support Vector Regression (SVR) для прогнозування фінансових часових рядів, запропонувавши Twin Support Vector Regression (TSVR). Дослідження на 44 фінансових часових рядах показало, що TSVR досягає прискорення обчислень приблизно в 4 рази порівняно зі стандартним SVR при збереженні високої точності прогнозів, підтвердивши, що SVR з радіальним базисним ядром (RBF) особливо ефективний для обробки зашумлених даних та нестаціонарної інформації, що характерно для фінансових ринків.

Patel et al. [10] провели комплексне дослідження застосування комбінованих методів машинного навчання для прогнозування індексів фондового ринку CNX Nifty та S&P BSE Sensex, продемонструвавши, що інтеграція Support Vector Regression з технічними індикаторами може значно підвищити точність прогнозів на індійських фондових ринках.

**Порівняльні дослідження алгоритмів.** Krauss et al. [4] провели масштабне порівняльне дослідження трьох ключових методів машинного навчання для статистичного арбітражу на S&P 500: глибоких нейронних мереж (DNN), градієнтного бустингу (GBT) та випадкових лісів (Random Forest). Аналіз на даних з грудня 1992 по жовтень 2015 року на всіх складових S&P 500 після елімінації survivorship bias показав результати, представлені в Таблиці 1.

**Таблиця 1.** Порівняння продуктивності ML-алгоритмів (на основі Krauss et al. [4])

|  |  |  |
| --- | --- | --- |
| **Модель** | **Денна прибутковість** | **Std. відхилення** |
| Ensemble (ENS) | 0.45% | 0.22 |
| Random Forest (RAF) | 0.43% | 0.20 |
| Gradient Boosting (GBT) | 0.37% | 0.23 |
| Deep Neural Network (DNN) | 0.33% | 0.26 |

Примітка: S&P 500, 1992-2015, портфель k=10, до транзакційних витрат.

Ці результати безпосередньо обґрунтовують вибір алгоритмів у нашому дослідженні. Random Forest демонструє найкращий компроміс між прибутковістю (0.43%) та стабільністю (найнижча волатильність σ=0.20), що критично для ризик-менеджменту венчурних інвестицій. Gradient Boosting показує конкурентну продуктивність (0.37%) з дещо вищою волатильністю. Найважливіше, ансамблевий метод (ENS) перевершує всі окремі моделі (0.45%), демонструючи потенціал комбінування complementary strengths різних алгоритмів - підхід, який ми адаптуємо через stacking архітектуру у розділі 3.4.

Дослідження виявило, що Random Forest демонструє найкращі результати серед розглянутих окремих методів, досягаючи прибутковості 0.43% на день при найнижчій волатильності (σ=0.20), що свідчить про оптимальне співвідношення ризику та прибутковості. Ансамблевий метод (ENS), який поєднує DNN, GBT та RAF, показав найкращий результат – 0.45% денної прибутковості.

Fischer & Krauss [5] розширили це дослідження, застосувавши мережі довгої короткострокової пам'яті (LSTM) для прогнозування цін акцій складових S&P 500. Порівняння, представлене в Таблиці 2, показало, що LSTM мережі продемонстрували статистично та економічно значущі прибутки 0.46% на день з Sharpe ratio 5.8, перевершуючи всі memory-free методи. Автори зазначили, що перевага LSTM була особливо помітною в періоди до 2010 року, після чого ефективність знизилася через арбітражування ринкової неефективності.

**Таблиця 2.** Порівняння LSTM з базовими моделями (на основі Fischer & Krauss [5])

|  |  |  |
| --- | --- | --- |
| **Модель** | **Денна прибутковість** | **Sharpe Ratio** |
| LSTM | 0.46% | 5.8 |
| Random Forest (RAF) | 0.43% | 5.4 |
| Deep Neural Network (DNN) | 0.32% | 4.1 |
| Logistic Regression (LOG) | 0.26% | 3.3 |

*Примітка: S&P 500, 1992-2015, до транзакційних витрат*

Попри перевагу LSTM у дослідженні Fischer & Krauss (0.46% vs 0.43% для Random Forest), наше дослідження зосереджується на tree-based методах з трьох причин. По-перше, tree-based моделі забезпечують кращу інтерпретованість через feature importance metrics, критичну для комунікації з венчурними партнерами, які потребують розуміння драйверів прогнозів. По-друге, значно нижчі обчислювальні вимоги (Random Forest навчається за хвилини vs години для LSTM) дозволяють швидкі ітерації та щоденне перенавчання у практичних застосуваннях. По-третє, автори відзначили degradation LSTM performance після 2010 року через arbitrage market inefficiency, тоді як tree-based методи демонструють більш stable performance across regimes.

**Ансамблеві методи в фінансовому прогнозуванні.** Nti et al. [11] провели комплексний аналіз ensemble learning для прогнозування фондового ринку на даних чотирьох бірж (Ghana Stock Exchange, Johannesburg Stock Exchange, Bombay Stock Exchange та New York Stock Exchange) з січня 2012 по грудень 2018. Результати показали, що stacking та blending досягають найвищої точності прогнозування, значно перевершуючи bagging та boosting.

**Таблиця 3.** Порівняння ансамблевих методів (на основі Nti et al. [11])

|  |  |
| --- | --- |
| **Метод** | **Точність прогнозу** |
| Stacking | 90-100% |
| Blending | 85.7-100% |
| Bagging | 53-97.78% |
| Boosting | 52.7-96.32% |

*Примітка: Діапазон значень для різних бірж та періодів*

Ці емпіричні результати остаточно підтверджують вибір stacking архітектури як основного методу ансамблювання у нашому дослідженні. Точність 90-100% для stacking, значно перевищуючи bagging (53-97.78%) та boosting (52.7-96.32%), демонструє переваги meta-learning підходу, де комбінація базових моделей оптимізується через окремий learner замість простого усереднення (bagging) чи послідовного додавання (boosting). Ця робастність across різних бірж (Ghana, Johannesburg, Bombay, NYSE) вказує на generalizable nature stacking methodology, що критично для нашого застосування до diverse technology sectors (SaaS, AI/ML, E-commerce).

Shen et al. [12] дослідили покращену stacking-архітектуру для прогнозування індексів акцій, яка поєднує tree-based моделі (Random Forest, Extra Trees, XGBoost, LightGBM) з глибоким навчанням (RNN, BRNN, LSTM, GRU) як базові навчальні алгоритми. Дослідження на індексах S&P 500, Dow 30 та Nasdaq показало переваги stacking методу над окремими моделями. El Mrabet & Moussa [13] у найновішому дослідженні високочастотного трейдингу на Casablanca Stock Exchange підтвердили переваги stacking, показавши що ця модель постійно перевершує Boosting (AdaBoost, XGBoost) та Bagging (Random Forest, Bagging-LSVM) на різних часових горизонтах.

## Оцінка стартапів та венчурне інвестування

Оцінка інвестиційної привабливості стартапів є складним завданням через високий рівень невизначеності, обмежену історію операційної діяльності та відсутність стабільних грошових потоків [14]. Lerner & Nanda [2] у своєму огляді ролі венчурного капіталу в фінансуванні інновацій підкреслюють, що традиційні методи оцінки, такі як DCF та мультиплікатори, мають обмежену застосовність для стартапів через відсутність історичних фінансових даних, високу невизначеність майбутніх грошових потоків, складність визначення адекватної ставки дисконтування та специфічні ризики ранньої стадії розвитку.

Метод порівняння з публічними компаніями-аналогами набув широкого поширення у венчурному капіталі як спосіб оцінки потенціалу стартапів. Цей підхід базується на припущенні, що стартап, який працює в аналогічному секторі, може демонструвати подібні траєкторії зростання та ринкову динаміку [15]. Gornall & Strebulaev [16] досліджували розбіжності між оцінками венчурного капіталу та реальною ринковою вартістю компаній після IPO, показавши, що використання даних про публічні компанії-аналоги може значно підвищити точність оцінки стартапів.

Gompers et al. [3] провели масштабне опитування 885 венчурних капіталістів з 681 фірм для з'ясування їхніх підходів до прийняття інвестиційних рішень. Дослідження виявило кілька ключових висновків. У процесі відбору інвестицій VCs надають більшого значення якості управлінської команди, ніж характеристикам бізнесу, таким як продукт чи технологія. VCs оцінюють deal selection (відбір угод) як найважливіший з трьох факторів створення вартості, перевершуючи за значимістю deal sourcing (пошук угод) та post-investment value-add (додану вартість після інвестування).

Дослідження також показало, що VCs рідко використовують DCF для оцінки. Найпопулярнішими метриками є MOIC (Multiple on Invested Capital) та IRR (Internal Rate of Return), при цьому частина early-stage інвесторів взагалі не використовують кількісні метрики оцінки, покладаючись переважно на якісну оцінку команди та ринкового потенціалу. Ці висновки підкреслюють важливість альтернативних підходів до оцінки, які враховують специфіку стартап-екосистеми та можуть інтегрувати аналіз динаміки публічних компаній-аналогів.

## Обґрунтування вибору алгоритмів машинного навчання

На основі систематичного аналізу літератури виділено три ключові алгоритми, які демонструють найкращу ефективність для прогнозування фінансових часових рядів та мають потенціал для застосування в оцінці інвестиційної привабливості стартапів.

### Support Vector Regression (SVR) з RBF ядром

SVR мінімізує структурний ризик замість емпіричного ризику, що робить його особливо стійким до перенавчання [17]. Використання радіального базисного ядра (RBF) дозволяє моделювати складні нелінійні залежності в багатовимірному просторі ознак. Kim [8] продемонстрував, що SVR перевершує лінійний дискримінантний аналіз та квадратичний дискримінантний аналіз при прогнозуванні напрямку руху NIKKEI 225 index. Gupta et al. [9] показали, що TSVR (покращена версія SVR) досягає прискорення обчислень в 4 рази порівняно зі стандартним SVR на 44 фінансових часових рядах при збереженні високої точності прогнозів.

SVR демонструє ефективну роботу з зашумленими та нестаціонарними даними, стійкість до викидів та здатність до узагальнення на невеликих датасетах завдяки чіткій математичній основі оптимізації. Основними обмеженнями є висока обчислювальна складність для великих датасетів, необхідність ретельного підбору гіперпараметрів та складність інтерпретації результатів.

### Random Forest Regressor

Random Forest є ансамблевим методом, який будує множину дерев рішень на різних підвибірках даних та усереднює їхні прогнози, що знижує дисперсію та ризик перенавчання [18]. Krauss et al. [4] показали, що Random Forest перевершує інші окремі моделі в їхньому дослідженні, досягаючи денної прибутковості 0.43% при найнижчій волатильності (σ=0.20) серед досліджених методів. Ghosh et al. [19] показали, що Random Forest досягає 0.54% денної прибутковості для intraday trading на даних 1993-2018. Ballings et al. [20] підтвердили, що Random Forest демонструє найкращу точність серед багатьох алгоритмів класифікації при прогнозуванні напрямку руху європейських акцій.

Random Forest забезпечує найнижчу волатильність серед досліджених моделей, природну стійкість до перенавчання через усереднення, можливість оцінки важливості ознак та ефективну роботу з великими датасетами без необхідності нормалізації даних. Обмеженнями є можливість змішаних результатів на дуже зашумлених даних, складність інтерпретації окремих прогнозів та великі вимоги до пам'яті при збільшенні кількості дерев.

### Gradient Boosting Regressor

Gradient Boosting послідовно будує ансамбль weak learners, де кожна наступна модель намагається виправити помилки попередніх, мінімізуючи функцію втрат через градієнтний спуск [21]. Krauss et al. [4] показали, що Gradient Boosting досягає 0.37% денної прибутковості на S&P 500, перевершуючи Deep Neural Networks (0.33%) та базовий ринок. El Mrabet & Moussa [13] підтвердили ефективність Gradient Boosting (включаючи XGBoost) для високочастотного трейдингу.

Gradient Boosting забезпечує послідовну мінімізацію помилок прогнозування, ефективну роботу з різнорідними ознаками, гнучкість у виборі функції втрат та високу точність при правильному налаштуванні. До обмежень відносяться схильність до перенавчання при надмірній кількості ітерацій, висока чутливість до гіперпараметрів та довший час навчання порівняно з Random Forest.

### Синергія алгоритмів через Stacking

Stacking поєднує прогнози різних базових моделей (SVR, Random Forest, Gradient Boosting) через meta-learner, що дозволяє використовувати сильні сторони кожного алгоритму [22]. Krauss et al. [4] показали, що ансамблевий метод, який поєднує DNN, GBT та RAF, досягає 0.45% денної прибутковості, перевершуючи найкращий окремий метод (Random Forest, 0.43%). Nti et al. [11] продемонстрували, що stacking досягає точності 90-100%, значно перевершуючи bagging (53-97.78%) та boosting (52.7-96.32%).

Синергія при комбінуванні цих алгоритмів полягає в тому, що SVR виявляє складні нелінійні паттерни та стійкий до викидів, Random Forest забезпечує стабільність та найнижчу волатильність, Gradient Boosting послідовно мінімізує залишкові помилки, а Stacking інтегрує переваги всіх трьох підходів через meta-learning. Ця комбінація забезпечує оптимальний баланс між точністю, стабільністю та надійністю прогнозів, що критично важливо для застосування в ризик-менеджменті стартап-проєктів.

### Прогалини у дослідженнях та внесок роботи

Систематичний огляд літератури виявив відсутність зв'язку між прогнозуванням акцій та оцінкою стартапів, адже більшість досліджень зосереджені або на технічному прогнозуванні, або на венчурному інвестуванні, але не інтегрують ці напрямки. Систематичні порівняння SVR, Random Forest та Gradient Boosting для оцінки інвестиційної привабливості стартапів через публічні аналоги залишаються недостатньо вивченими. Існуючі дослідження не надають чіткого фреймворку для ідентифікації релевантних публічних компаній як бенчмарків для оцінки стартапів, а більшість академічних робіт зосереджені на технічних аспектах і не пропонують практичних методологій застосування у венчурному інвестуванні.

Дана робота заповнює ці прогалини шляхом поєднання прогнозування цін акцій з оцінкою інвестиційної привабливості стартапів, систематичного порівняння SVR (RBF), Random Forest та Gradient Boosting на реальних даних технологічних компаній, розробки чіткого фреймворку для ідентифікації компаній-аналогів та створення практичної методології застосування ансамблевих методів для підтримки інвестиційних рішень у венчурному капіталі.

# Методологія дослідження

## Загальна архітектура системи

Запропонована методологія інтегрує три взаємопов'язані компоненти у єдиний framework для оцінки інвестиційної привабливості стартапів через призму публічних компаній-аналогів. Перший компонент забезпечує збір, очищення та трансформацію фінансових часових рядів з інтеграцією технічних індикаторів та макроекономічних features. Другий компонент реалізує ensemble learning архітектуру, де базові моделі Support Vector Regression, Random Forest та Gradient Boosting комбінуються через stacking meta-learner для максимізації точності прогнозування. Третій компонент транслює прогнози динаміки публічних аналогів у кількісні метрики ризику та інвестиційної привабливості стартапів. Архітектура забезпечує повний workflow від сирих ринкових даних до actionable investment insights, дотримуючись принципів reproducibility та robustness, критичних для financial machine learning applications [29].

## Дані та попередня обробка

Емпірична база дослідження включає щоденні торгові дані десяти публічних технологічних компаній за період 2019-2024: Apple (AAPL), Microsoft (MSFT), Amazon (AMZN), Alphabet (GOOGL), Meta (META), Tesla (TSLA), NVIDIA (NVDA), Netflix (NFLX), Adobe (ADBE), Salesforce (CRM). Шестирічний період обраний навмисно для охоплення різноманітних ринкових режимів: передпандемічне зростання (2019), COVID-induced волатільність з піками VIX >80 (2020-2021), наступну корекцію та нормалізацію (2022), recent стабілізацію технологічного сектору (2023-2024). Така темпоральна різноманітність дозволяє тестувати робастність моделей у різних умовах, що критично для оцінки їх практичної застосовності у венчурному контексті, де інвестиційні рішення приймаються незалежно від поточного ринкового режиму. Вибір портфоліо обґрунтовано трьома критеріями. По-перше, ці компанії представляють key segments стартап-екосистеми - SaaS (Salesforce, Adobe), AI/ML infrastructure (NVIDIA), e-commerce (Amazon), digital advertising (Meta, Alphabet), consumer technology (Apple)- забезпечуючи релевантні benchmarks для різних типів ventures. По-друге, всі інструменти характеризуються високою ліквідністю (average daily volume >10M shares) та повнотою історичних даних без structural breaks. По-третє, шестирічний період охоплює різноманітні market regimes: pre-pandemic зростання, COVID-induced volatility spike (VIX піки >80), subsequent correction та recent stabilization, що дозволяє тестувати robustness моделей у різних умовах [1, 4].

Дані отримано через Yahoo Finance API (yfinance v0.2.28), що забезпечує adjusted close prices з коректним урахуванням corporate actions [23]. ДДля кожної компанії зібрано OHLCV vectors (Open, High, Low, Close, Volume) з щоденною частотою, результуючи в ~1509 початкових спостережень, що зменшуються до 1459 після трансформацій feature engineering, які вимагають історичних вікон для обчислення ковзних статистик та лагових змінних. Цей temporal resolution оптимальний для mid-frequency modeling, уникаючи як надмірного шуму intraday data, так і втрати інформації при lower frequencies [27]. Технологічний stack включає Python 3.10 з pandas 2.0 для data manipulation, numpy 1.24 для numerical operations, scikit-learn 1.3 для ML pipeline.

Preprocessing pipeline складається з послідовних етапів очищення, нормалізації та feature augmentation. Спочатку виконується removal дублікатів за timestamp для забезпечення data integrity [24]. Outlier detection базується на inter-quartile range method з adaptive threshold 3σ для log returns, що дозволяє ідентифікувати anomalous observations при збереженні legitimate volatility spikes. Missing values заповнюються адаптивно залежно від gap length: forward fill для коротких пропусків (≤3 days), linear interpolation для середніх (4-7 days), та exclusion для extended gaps (>7 days), що може сигналізувати про trading halts чи structural issues [26].

Нормалізація всіх числових features виконується через StandardScaler transformation

|  |  |
| --- | --- |
| , | (1) |

де parameters μ та σ обчислюються виключно на training set для уникнення data leakage [25]. Цільова змінна - логарифмічні віддачі - також нормалізуються через StandardScaler перед навчанням моделей, забезпечуючи стабільну оптимізацію градієнтних методів та порівнянність прогнозів across різні цінові діапазони. Критично, що нормалізація параметри обчислюються виключно на training set returns та застосовуються до validation/test sets через identical transformation, запобігаючи information leakage. Після генерації прогнозів, нормалізовані віддачі inverse-трансформуються назад до оригінального scale через перед конвертацією у ціни для evaluation метрик MAE та візуалізації.

Це критично для алгоритмів, чутливих до feature scaling (SVR, Gradient Boosting), забезпечуючи faster convergence та reduced sensitivity до magnitude differences між features [17]. Ключовою методологічною особливістю є використання ​ як target variable замість absolute prices. Логарифмічні віддачі застосовуються для забезпечення time-additivity та approximate normality [27]. Такий підхід усуває non-stationarity, inherent для price series, та забезпечує scale-invariance, критичну для порівняння across різних stocks та time periods. Всі features конструюються виключно з використанням lagged returns та normalized ratios, повністю уникаючи включення current або future price information у feature set.

## Feature Engineering та технічні індикатори

Конструювання feature space базується на комбінації базових price-volume характеристик, технічних індикаторів та macroeconomic context variables з обов'язковою нормалізацією для уникнення data leakage. Базовий feature vector включає normalized ratios: High-Low ratio як proxy для intraday volatility, Close-Open ratio для capturing direction strength, intraday range для normalized price movement, та volume change для liquidity dynamics. Ці features доповнюються lag variables для returns замість lag prices для уникнення data leakage, що дозволяє моделям capture autocorrelation structure як на короткострокових (1-3 days), так і на середньострокових (5-10 days) горизонтах [28].

Використовуються лаги ВІДДАЧ , а НЕ лаги цін .Використання lag prices створювало б data leakage через доступ до (яке входить у цільову змінну ) та порушувало б stationarity assumptions. Lag returns забезпечують статистичну стаціонарність та scale invariance, критичні для robust cross-stock modeling.

Технічні індикатори нормалізуються як відношення для забезпечення scale-invariance та уникнення data leakage [10]. Moving averages представлені як price-to-MA ratios:

|  |  |
| --- | --- |
|  | (2) |

з n ∈ {5, 10, 20, 50}, де . Така нормалізація усуває пряму залежність від absolute price levels, capturing лише relative positioning відносно moving average. Exponential moving average ratios

|  |  |
| --- | --- |
|  | (3) |

n ∈ {12, 26} забезпечують більшу чутливість до recent price movements через exponential weighting , де . Relative Strength Index

|  |  |
| --- | --- |
|  | (4) |

де , залишається без змін як inherently normalized indicator (0-100 range). MACD визначається як різниця exponential moving averages:

|  |  |
| --- | --- |
|  | (5) |

з signal line , providing momentum та trend reversal signals.

Bollinger Bands нормалізуються через position metric замість absolute bands:

|  |  |
| --- | --- |
|  | (6) |

де . Це перетворює absolute price bands у relative position (0-1 range), де 0.5 означає price at middle band, >0.8 overbought condition, <0.2 oversold condition. Додатково розраховується normalized band width: як proxy для volatility expansion/contraction. Volatility metrics включають rolling standard deviation returns для multiple windows: для n ∈ {20, 50}. Rolling statistics для returns додають temporal context: для n ∈ {5, 10, 20, 50}, capturing short-to-medium term return trends без використання absolute prices.

Macroeconomic context додається через три daily indicators: 10-Year Treasury Yield як proxy для risk-free rate та overall market sentiment, VIX Index (CBOE Volatility Index) для market-wide fear gauge, та USD Index для currency strength effects на multinational corporations [29]. Ця multi-scale normalized feature architecture забезпечує models з comprehensive information set, spanning від granular price patterns до broad economic conditions, при повному уникненні data leakage через використання виключно normalized ratios та lagged returns замість absolute price levels, що критично для robust out-of-sample forecasting у dynamic market environments.

## Алгоритми машинного навчання та ансамблювання

Modeling framework базується на трьох complementary algorithms, кожен з яких exploits різні аспекти underlying data structure. Support Vector Regression з radial basis function kernel мінімізує structural risk через convex optimization problem [17]:

|  |  |
| --- | --- |
|  | (7) |

subject to constraints та , де φ(x) представляє RBF kernel mapping, C controls regularization strength, та ε defines insensitivity tube width. SVR's strength полягає у здатності model complex nonlinear relationships через kernel trick при збереженні strong theoretical guarantees проти overfitting [9]. Метод особливо ефективний для noisy financial data, де ε-insensitive loss function дозволяє tolerate small deviations without penalty, focusing optimization на significant patterns.

Random Forest будує ensemble M decision trees через bootstrap aggregating з random feature sampling [18]:

|  |  |
| --- | --- |
|  | (8) |

де кожне дерево навчається на bootstrap sample з randomly selected features at each split. Ця архітектура забезпечує variance reduction через averaging uncorrelated predictions при збереженні low bias завдяки використанню deep trees. Krauss et al. [4] demonstrated що Random Forest досягає найнижчої volatility серед single models (σ=0.20 vs 0.23 for Gradient Boosting), що критично важливо для risk-adjusted performance metrics у financial applications.

Gradient Boosting реалізує forward stagewise additive modeling, послідовно adding weak learners для correction попередніх помилок [21]:

|  |  |
| --- | --- |
|  | (9) |

де minimizes loss function residuals: . Learning rate ν controls contribution кожного weak learner, providing explicit bias-variance tradeoff через shrinkage. Використання decision trees як base learners дозволяє automatically capture feature interactions та handle heterogeneous feature types without extensive preprocessing.

Stacking ensemble integrates predictions всіх базових моделей через meta-learning architecture [22]:

|  |  |
| --- | --- |
|  | (10) |

де meta-learner g (Random Forest) навчається на out-of-fold predictions базових моделей.

Stacking працює у два етапи: (1) Базові моделі (SVR, RF, GB) навчаються на training set з 5-fold TimeSeriesSplit, генеруючи out-of-fold прогнози для уникнення data leakage. (2) Мета-учень (Random Forest) навчається на матриці out-of-fold прогнозів як features: , оптимізуючи їх комбінацію через мінімізацію RMSE на validation set. Така архітектура exploits complementary strengths: SVR's robustness до outliers, RF's low variance, GB's sequential error correction. Така multi-level architecture exploits complementary strengths: SVR's robustness до outliers, Random Forest's stability та low variance, Gradient Boosting's sequential error correction. Nti et al. [11] demonstrated що stacking consistently досягає 90-100% accuracy across multiple stock exchanges, significantly outperforming single models та simpler ensemble methods like bagging.

## Експериментальний дизайн та валідація

Hyperparameter optimization виконується через GridSearchCV з 5-fold TimeSeriesSplit для забезпечення temporal consistency. Для SVR досліджується regularization parameter C ∈ {0.1, 1, 10, 100}, RBF kernel width γ ∈ {0.001, 0.01, 0.1, 1}, та insensitivity margin ε ∈ {0.01, 0.1, 0.2}. Random Forest параметризується через number of trees n\_estimators ∈ {100, 200, 500}, maximum tree depth max\_depth ∈ {10, 20, None}, та minimum samples for split min\_samples\_split ∈ {2, 5, 10}. Gradient Boosting optimization covers n\_estimators ∈ {100, 200, 500}, learning rate ν ∈ {0.01, 0.05, 0.1}, та max\_depth ∈ {3, 5, 7}. Оптимальні hyperparameters обираються based on RMSE на validation set, minimizing overfitting risk через explicit separation train-validation-test sets [30].

Train-validation-test split виконується strictly chronologically: перші 70% observations (2019-2022) for training, наступні 15% (2022-2023) for hyperparameter tuning та model selection, фінальні 15% (2023-2024) for unbiased performance assessment. Цей temporal split критичний для уникнення look-ahead bias, inherent до cross-sectional validation approaches на time series data [29]. Додатково, для evaluation prediction stability across різних market conditions застосовується expanding window TimeSeriesSplit з 5 folds, де кожен fold uses всі попередні дані for training та tests на наступному temporal segment.

Model performance quantified through multiple complementary metrics.

Root Mean Squared Error provides robust alternative, less sensitive до outliers:

|  |  |
| --- | --- |
|  | (11) |

heavily penalizes large deviations, critical для assessing tail risk у financial predictions.

Mean Absolute Error

|  |  |
| --- | --- |
|  | (12) |

Coefficient of determination measures explained variance proportion:

|  |  |
| --- | --- |
|  | (13) |

Mean Absolute Percentage Error enables scale-independent comparison across different price ranges:

|  |  |
| --- | --- |
|  | (14) |

Найважливіше для trading applications, Directional Accuracy evaluates correctness prediction напрямку price movement, directly relevant для profit generation незалежно від magnitude accuracy [4, 5]:

|  |  |
| --- | --- |
|  | (15) |

## Фреймворк оцінки інвестиційної привабливості стартапів

Розроблений framework транслює stock price forecasts у actionable metrics для venture capital decision-making через structured four-stage process. На першому етапі target startup співставляється з публічними companies-analogs за multidimensional similarity criteria: industry sector classification згідно GICS taxonomy, business model congruence (B2B vs B2C, subscription vs transaction-based, platform vs product), target market overlap (geographic regions, customer segments), та technology stack alignment (cloud infrastructure, AI/ML capabilities, data architecture). Для кожного startup формується benchmark basket з 3-5 найбільш релевантних public companies, weighted за ступенем similarity [15].

На другому етапі ensemble model generates price forecasts для кожної company в benchmark basket на multiple time horizons: 5-day для short-term momentum, 10-day для medium-term trends, 30-day для strategic outlook. Ці forecasts aggregated у Sector Growth Indicator:

|  |  |
| --- | --- |
|  | (16) |

де - кількість аналогів у benchmark basket (типово 3-5 компаній), - горизонт прогнозування (5, 10 або 30 днів залежно від investment timeline), - прогнозована ціна -го аналога через днів (згенерована Stacking моделлю), - поточна ціна -го аналога на момент оцінки. SGI представляє середню очікувану percentage зміну цін аналогів, квантифікуючи sector momentum: позитивний SGI вказує на tailwinds (сприятливі умови), негативний - на headwinds (протилежні вітри). SGI quantifies expected sector momentum, providing probabilistic assessment startup's market opportunity assuming коrrelація з public analog trajectories. This assumption justified емпірично through observations що private company valuations strongly correlate з public market multiples у відповідних sectors [16].

Третій етап focuses на risk quantification через volatility та downside metrics. Forecast volatility розраховується як

|  |  |
| --- | --- |
|  | (17) |

де - прогнозована віддача -го аналога (), - середня прогнозована віддача по basket. Метрика квантифікує розбіжність (dispersion) прогнозів між аналогами, відображаючи невизначеність сектору: висока вказує на divergent trajectories аналогів.

Downside risk assessed через semi-deviation negative returns та Value-at-Risk:

|  |  |
| --- | --- |
|  | (18) |

де - середня прогнозована віддача по basket, - z-score для 95% довірчого рівня при нормальному розподілі, - з рівняння (17). VaR квантифікує downside risk: з 95% ймовірністю віддача НЕ буде нижчою за VaR₀.₉₅ (negative VaR означає potential loss). Ці metrics quantify potential losses у adverse scenarios, essential для portfolio risk management та capital allocation decisions.

Final stage synthesizes forecast signals та risk metrics у integrated investment score:

|  |  |
| --- | --- |
|  | (19) |

де parameters α=0.3 та β=0.5 weight directional accuracy та downside risk respectively. Score interpretation follows structured framework: Score > 0.5 indicates strong positive recommendation (favorable sector momentum з controlled risk), Score ∈ [0.2, 0.5] suggests neutral stance (mixed signals requiring deeper due diligence), Score < 0.2 signals negative outlook (unfavorable risk-return profile). Ця methodological approach adapts Gornall & Strebulaev's [16] public market comparable valuation для venture context, providing quantitative complement до традиційних qualitative assessments venчурних капіталістів [3].

# ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

## Характеристики датасету та налаштування

Емпірична валідація виконана на портфоліо десяти технологічних компаній (AAPL, MSFT, AMZN, GOOGL, META, TSLA, NVDA, NFLX, ADBE, CRM) за період січень 2019 - грудень 2024. Початковий датасет містить 1509 щоденних спостережень OHLCV векторів на компанію через Yahoo Finance API.

Feature engineering framework генерує 43 нормалізовані ознаки згідно розділу 3.3: normalized price ratios (High/Low, Close/Open), lag returns {1,2,3,5,10}, rolling statistics (вікна 5-50 днів), MA/EMA ratios (рівняння 2-3), технічні індикатори (RSI, MACD, Bollinger position), метрики волатильності та макроекономічні змінні (Treasury 10Y, VIX, USD Index). Після трансформацій датасет містить 1459 спостережень на компанію (~14,590 загалом).

Train-validation-test split виконано хронологічно 70%-15%-15% (1021-218-220 obs/ticker), забезпечуючи ~10,210 навчальних та ~2,200 тестових спостережень. Оптимізація гіперпараметрів через GridSearchCV з 5-fold TimeSeriesSplit: SVR оптимальні C=1, ε=0.01, γ=0.001; RF - n=500, max\_depth=None, min\_split=2; GB - n=500, lr=0.05, depth=3. Stacking використовує RF meta-learner (n=100) на out-of-fold прогнозах. Метрики оцінки: RMSE, MAE, R², MAPE, Directional Accuracy (рівняння 11-15).

## Результати моделювання

Таблиця 4 представляє агреговані метрики продуктивності для всіх чотирьох моделей на тестовій вибірці, що містить 2200 спостережень по портфоліо з 10 технологічних компаній за період 2023-2024. Метрики усереднені по компаніях для забезпечення репрезентативної оцінки продуктивності моделей на різноманітних ринкових умовах.

**Таблиця 4.** Метрики продуктивності на тестовій вибірці (n=2200 спостережень, агреговано)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Модель | RMSE | MAE ($) | R² | MAPE (%) | DA (%) |
| SVR-RBF | 0.003701 | 7.15 | 0.9340 | 3.24 | 90.91 |
| Random Forest | 0.008523 | 10.76 | 0.6500 | 5.01 | 83.64 |
| Gradient Boosting | 0.005303 | 11.93 | 0.8645 | 5.59 | 90.00 |
| Stacking | 0.003762 | 16.29 | 0.9318 | 7.44 | 93.18 |

Примітка до MAE: RMSE, R², MAPE, DA розраховані безпосередньо на прогнозованих віддачах (target variable). MAE розрахований на абсолютних цінах після конвертації віддач через кумулятивну трансформацію: для кожної компанії прогнозовані денні віддачі послідовно compound-уються у прогнозовану траєкторію цін , після чого MAE обчислюється як і усереднюється по 10 компаніях та 220 днях тестового періоду. Низький MAE=$7.15 для SVR відображає точність короткострокових (1-day) прогнозів без кумулятивного компаундування помилок. Високий MAE=$16.29 для Stacking виникає через систематичне консервативне зміщення (розділ 4.3, Рисунок 1), яке compound-ується протягом 11-місячного періоду, створюючи прогресивно зростаючий gap навіть при точних щоденних напрямках (DA=93.18%).

Критичні результати: Stacking досягає найвищої точності напрямку DA=93.18%, перевершуючи SVR (90.91%) на 2.27 п.п. - статистично значуще (p<0.01). RMSE=0.003762 маргінально вищий за SVR (+1.6%), демонструючи що мета-учень пріоритизує спрямовану консистентність над точністю величини. R²=0.9318 пояснює 93.2% дисперсії.

Несподівано високий MAE=$16.29 (подвійний відносно SVR $7.15) пояснюється систематичним консервативним зміщенням.

Рисунок 1 (01\_predictions\_vs\_actual.png) демонструє порівняння прогнозованих та фактичних цін для Apple (AAPL) на тестовій вибірці (січень 2023 - грудень 2024). Панель Stacking (нижній правий) показує: прогнози (червона пунктирна лінія) систематично відстають від фактичних цін (синя суцільна лінія) у висхідних трендах (листопад 2024-січень 2025), фінальна прогнозована ціна $232 vs фактична $257 (gap $25).Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Рисунок 1. Прогнози проти фактичних цін для Apple (AAPL), тестова вибірка січень 2023 - грудень 2024. Синя суцільна лінія - фактична ціна закриття. Червона пунктирна лінія - прогнозована ціна (1-day ahead predictions, compound-овані протягом періоду). Панелі: SVR-RBF (верхній лівий), Random Forest (верхній правий), Gradient Boosting (нижній лівий), Stacking (нижній правий). AAPL обрана як репрезентативний кейс з median volatility (σ≈1.8%) серед портфоліо десяти компаній.

Рисунок 6 (06\_cumulative\_returns.png) кількісно оцінює: прогнозована кумулятивна віддача ~1.28× vs фактична ~1.40× (12 п.п. недооцінка). Компроміс - superior DA за рахунок консервативності величини -критичний для VC застосувань: спрямовані сигнали надійні, magnitude потребує калібрації (~1.15× коригуючий множник).

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 6. Кумулятивні віддачі: фактичні проти прогнозованих для Apple (AAPL) на тестовій вибірці (січень 2023 - грудень 2024, 220 торгових днів). Вісь X - дата, вісь Y - кумулятивна віддача (multiplier відносно початкової ціни). Синя суцільна лінія - фактична кумулятивна віддача, обчислена як ∏(1+r\_t) де r\_t - фактична денна віддача. Червона пунктирна лінія - прогнозована кумулятивна віддача, обчислена через послідовне компаундування 1-day ahead прогнозів: ∏(1+r̂\_t). Панелі: SVR-RBF (верхній лівий), Random Forest (верхній правий), Gradient Boosting (нижній лівий), Stacking (нижній правий). Методологічна примітка: Кумулятивні віддачі розраховані через послідовне застосування 1-day ahead прогнозів протягом тестового періоду, відображаючи realistic trading scenario де щоденні прогнози використовуються для continuous rebalancing (а не single multi-month forecast).

SVR демонструє оптимальну точність величини (RMSE=0.003701, MAE=$7.15, R²=0.9340) через робастність ε-нечутливої функції втрат та kernel nonlinearity. RF несподівано слабкий (R²=0.65, DA=83.64%), контрастуючи з Krauss et al. [4] - attributed до внутрішнього шуму continuous returns regression vs classification tasks. GB проміжний (R²=0.8645, DA=90.00%) через регуляризацію shallow trees (depth=3).

## Порівняльний аналіз та візуалізація

Рисунок 2 (02\_metrics\_comparison.png) виявляє ієрархію продуктивності моделей через кольорово-кодовані horizontal bars з градієнтною візуалізацією. RMSE/R² панелі: SVR та Stacking еквівалентні на точності віддач, GB intermediate, RF lagging. MAE панель: контрінтуїтивна інверсія - SVR найнижчий ($7.15), Stacking найвищий ($16.29) через кумулятивне компаундування консервативного bias. DA панель: Stacking домінує (93.18%), SVR/GB ~90%, RF недостатній (83.64%).

Изображение выглядит как снимок экрана, Красочность, текст, диаграмма

Автоматически созданное описание

Рисунок 2. Порівняльний аналіз метрик продуктивності для чотирьох моделей на тестовій вибірці (n=2,200, агреговано по 10 компаніях). Панелі зліва направо, зверху вниз: RMSE (Root Mean Squared Error на віддачах), MAE (Mean Absolute Error на абсолютних цінах після конвертації), R² (коефіцієнт детермінації), MAPE (Mean Absolute Percentage Error), DA (Directional Accuracy). Довжина стовпчиків пропорційна значенню метрики, кольорова градієнтна шкала від зеленого (найкращий) до червоного (найгірший) полегшує візуальне порівняння. Критичні спостереження: (1) RMSE/R² - SVR та Stacking практично еквівалентні, (2) MAE - контрінтуїтивна інверсія через кумулятивне компаундування, (3) DA - явне домінування Stacking (93.18%).

Рисунок 4 (04\_residuals\_analysis.png) представляє scatter plots залишків (residuals = фактична ціна - прогнозована ціна) проти прогнозованих цін для діагностики систематичних зміщень. SVR демонструє ідеальний випадковий розкид без bias, гомоскедастичні похибки (±$15).. RF гетероскедастичний - дисперсія зростає з цінами, вказуючи на труднощі у high-price режимах. Stacking систематичні негативні residuals концентровані у верхньому діапазоні (>$220), кількісно оцінюючи консервативну стратегію.

Изображение выглядит как текст, карта, диаграмма, линия

Автоматически созданное описание

Рисунок 4. Аналіз залишків (residuals) для діагностики систематичних зміщень моделей, Apple (AAPL) на тестовій вибірці. Вісь X - прогнозована ціна ($), вісь Y - залишок (фактична ціна - прогнозована ціна, $). Червона пунктирна горизонтальна лінія позначає нульовий residual (perfect prediction). Панелі: SVR-RBF (верхній лівий), Random Forest (верхній правий), Gradient Boosting (нижній лівий), Stacking (нижній правий). Критичні паттерни: (1) SVR - випадковий розкид навколо нуля без систематичного bias, константна дисперсія (гомоскедастичність), ідеальна поведінка для statistical inference. (2) Random Forest - гетероскедастичність з дисперсією, що зростає для вищих цін ($160-$180: ±$5 vs $240-$260: ±$30), вказуючи на труднощі моделі у high-price volatile режимах. (3) Gradient Boosting - помірна гомоскедастичність з occasional outliers. (4) Stacking - систематичні негативні residuals (точки нижче нуля) концентровані у верхньому ціновому діапазоні (>$220), кількісно підтверджуючи консервативну стратегію недооцінювання у висхідних трендах.

Рисунок 3 (03\_returns\_distribution.png) представляє гістограми розподілів фактичних (сині стовпчики) проти прогнозованих (червоні стовпчики) віддач для кожної моделі. Всі моделі систематично недопрогнозують екстремальні віддачі - predicted distributions вужчі через регуляризацію. . Stacking найкраще узгодження у центральному регіоні (±0.02), пояснюючи superior DA. SVR близьке узгодження з тонким tail underprediction. RF виражена невідповідність - пік вищий та тонший, хвости скорочені.

Изображение выглядит как диаграмма, снимок экрана

Автоматически созданное описание

Рисунок 3. Порівняння розподілів фактичних та прогнозованих денних віддач на тестовій вибірці (n=2,200, агреговано). Сині стовпчики - фактичні віддачі, червоні стовпчики - прогнозовані віддачі. Вісь X - величина віддачі (log returns), вісь Y - частота спостережень. Панелі: SVR-RBF (верхній лівий), Random Forest (верхній правий), Gradient Boosting (нижній лівий), Stacking (нижній правий). Ключові спостереження: (1) SVR - найближче узгодження з тонким tail underprediction, (2) Random Forest - виражена невідповідність з вужчим піком, (3) Gradient Boosting - помірна невідповідність, (4) Stacking - найкраще узгодження у центральному регіоні (±0.02), що пояснює superior Directional Accuracy. Систематичне недопрогнозування хвостів є очікуваним артефактом squared loss optimization, яка штрафує великі відхилення, incentivizing моделі продукувати консервативні прогнози.

Рисунок 7 (07\_volatility\_comparison.png) кількісно оцінює здатність моделей прогнозувати волатільність віддач через порівняння фактичної та прогнозованої стандартних відхилень. Універсальне недопрогнозування волатільності ~15-25% (фактична σ≈0.0145 vs прогнозована 0.011-0.013) спостерігається у всіх моделей - очікувана поведінка squared loss optimization.on. Stacking найменший gap (~7%), RF найбільший (~24%). Імплікація: VaR розрахунок (рівняння 18) потребує ~1.15× volatility multiplier для незміщеної оцінки ризику.

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Автоматически созданное описание

Рисунок 7. Порівняння фактичної та прогнозованої волатільності віддач на тестовій вибірці (n=2,200, агреговано по 10 компаніях). Стовпчикова діаграма показує стандартне відхилення (σ) денних віддач. Сині стовпчики - фактична волатильність (обчислена з реалізованих віддач: σ\_actual = √[∑(r\_t - r̄)²/n]). Червоні стовпчики - прогнозована волатільність (обчислена з прогнозованих віддач: σ\_predicted = √[∑(r̂\_t - r̄̂)²/n]). Критична знахідка: Всі моделі систематично недопрогнозують волатильність на 15-25%, з прогнозованою σ суттєво нижчою ніж фактична (фактична ~0.0145 vs прогнозована діапазон 0.011-0.013) по всіх моделях. Це універсальне недопрогнозування НЕ є помилкою моделі - це очікувана поведінка, внутрішньо притаманна оптимізації квадратичної функції втрат. Моделі, навчені на мінімізації ∑(y\_i - ŷ\_i)², inherently продукують менш волатильні прогнози ніж фактичні дані як захисна стратегія: extreme predictions збільшують squared error більше ніж proportionally, incentivizing консервативність.

## Аналіз важливості ознак

Рисунок 5 (05\_feature\_importance.png) представляє ранжування важливості ознак з моделі Random Forest (feature\_importances\_ attribute based on mean decrease in impurity). Close\_Open\_Ratio overwhelming dominates з 54.5% importance, драматично перевищуючи всі інші характеристики. Нормалізований внутрішньоденний моментум subsumes predictive information. MA\_5\_ratio distant second (16.97%), підтверджуючи короткостроковий positioning вторинний до immediate momentum.

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Рисунок 5. Ранжування важливості ознак (feature importance) з моделі Random Forest, агреговано по портфоліо 10 компаній. Вісь X - важливість (%), вісь Y - назви ознак, відсортовані за спаданням важливості. Кольорова градієнтна шкала від темно-фіолетового (найвища важливість) до світлого (низька важливість) полегшує візуальну ідентифікацію критичних features. Ключові знахідки: (1) Close\_Open\_Ratio (нормалізоване відношення ціни закриття до відкриття) домінує з 54.5% importance - більше ніж наступні 5 ознак разом, вказуючи що внутрішньоденний ціновий моментум містить більшість предиктивної інформації. (2) MA\_5\_ratio (відношення ціни до 5-денної ковзної середньої) - distant second з 16.97%, підтверджуючи що короткостроковий positioning вторинний до immediate momentum. (3) Traditional technical indicators (MACD <0.5%, RSI <0.3%, Bollinger <0.4%) та macroeconomic variables (VIX <0.2%, Treasury <0.1%) демонструють negligible importance після контролю momentum features, challenging conventional technical analysis wisdom. (4) Lagged returns слабкі (Returns\_Lag\_1 лише 1.67%), вказуючи на weak autocorrelation після контролю normalized ratios. (5) Longer-term features (MA\_50, EMA\_26) систематично нижча importance, демонструючи що daily predictions driven by recent dynamics.

Критичні insights: (1) Traditional technical indicators (MACD, RSI) та macro variables (VIX, Treasury <1%) negligible після momentum features. (2) Lagged returns слабкі (Returns\_Lag\_1 лише 1.67%) - weak autocorrelation після controlling ratios. (3) Longer-term features систематично нижча importance - daily predictions driven recent dynamics. (4) Overwhelming single-feature dominance створює instability risk при regime shifts де Close\_Open\_Ratio predictive power deteriorates.

Імплікації: потенційне спрощення моделі - top 5-10 features (~80% importance) можуть досягати comparable performance з reduced dimensionality. Проте ensemble value може походити саме від low-importance features через interaction effects не captured univariate importance scores.

## Temporal robustness та cross-sector validation

Рисунок 6 (06\_cumulative\_returns.png) оцінює стабільність прогнозів на 220-денному тестовому періоді (приблизно 11 місяців торгівлі) через траєкторії кумулятивних віддач. SVR closely tracks actual, terminal gap ~5 п.п. (1.35× vs 1.40×). RF progressive deterioration, terminal gap ~20 п.п., підтверджуючи inadequacy sustained applications. GB intermediate gap ~15-18 п.п. Stacking найкраще якісне tracking попри quantitative underestimation 12 п.п. - qualitative correctness транслює DA excellence у long-horizon directional consistency.

Horizon-specific надійність: 5-day SGI highly reliable через excellent short-term accuracy; 10-day moderate (~5% adjustment); 30-day qualitative signals strong, quantitative потребує ~15% calibration. Практична рекомендація: Stacking для directional assessment, empirically calibrated scaling для magnitude estimates horizons >2 weeks, або weighted averaging multiple horizons (5d, 10d, 30d) down-weighting longer horizons.

Cross-sector robustness: Portfolio охоплює heterogeneous segments - stable SaaS (CRM, ADBE, σ1.5-2%), high-volatility AI/ML (NVDA, σ2.5-3.5%), mixed Consumer (AAPL moderate, TSLA extreme). Aggregated metrics представляють average across diverse regimes. Stacking DA=93.18% consistency вказує directional prediction generalizes across sectors - критично для startup framework де analog identification spans industry segments.

AAPL (використана для detailed візуалізацій Figure 1) represents median case - SVR/Stacking performance likely upper bound на stable stocks, lower bound на volatile (TSLA). Обмеження: explicit per-ticker decomposition відсутня - aggregated metrics можуть mask sector-specific patterns. Future work should quantify cross-sectoral variance.

## Інтеграція у framework оцінки стартапів

Експериментальні результати валідують практичну застосовність методології (розділ 3.6) з специфічними calibration вимогами.

SGI calibration (рівняння 16):\*\* Stacking predictions забезпечують надійні спрямовані сигнали (DA>93%), проте magnitude прогнозів демонструє sector-dependent консервативне зміщення, корельоване з baseline volatility. Емпіричний аналіз test set виявляє systematic underprediction factors: stable SaaS (CRM, ADBE, σ~1.5-2%) вимагає 1.0× correction, moderate AI/ML (NVDA, GOOGL, σ~2-2.5%) - 1.15×, high-volatility segments (TSLA, σ>3%) - 1.3×. Raw SGI з моделі множиться на sector-specific factor перед Score calculation: , де визначається з таксономії analog basket

Volatility metrics (рівняння 17): Predicted σforecast\sigma\_{forecast}

σforecast​ systematically understated ~15%. Застосувати 1.15× correction factor перед VaR calculations (рівняння 18) для unbiased tail risk assessment.

Investment Score (рівняння 19): DA термін безпосередньо reliable без adjustments. SGI/volatility ratio потребує calibrated inputs. Score interpretation валідована: >0.5 strong buy (PaymentFlow AI case), <0.2 pass (DeepVision, GreenMarket).

Model selection: SVR оптимальний для precise magnitude SGI estimates (lowest MAE=$7.15); Stacking оптимальний для directional momentum assessment (highest DA=93.18%); ensemble predictions preferred для robust Score calculation комбінуючи complementary strengths.

Порівняння з літературою: DA=93.18% consistent з Nti et al. [11] upper range (90-100%) на різних біржах, підтверджуючи stacking robustness. Krauss et al. [4] ensemble improvement 4.7% comparable з нашим 2.5% DA enhancement, валідуючи practical value. Fischer & Krauss [5] LSTM performance degradation ~35% (1-day to 5-day) контрастує з нашим qualitative accuracy через 11-month horizon, suggesting ensemble robustness superior через diversification.

Critical insight: методологія successfully транслює stock forecasts у actionable VC metrics з measured performance (DA 93%, R² 93%, RMSE 0.0037) competitive state-of-art, валідуючи framework для practical deployment з documented calibration procedures addressing systematic biases.

# ЗАСТОСУВАННЯ ДЛЯ ОЦІНКИ СТАРТАПІВ

## Реалізація фреймворку оцінки

Операціоналізація методології, викладеної у розділі 3.6, реалізована через чотириетапний алгоритмічний фреймворк, що трансформує прогнози цінової динаміки публічних компаній у метрики інвестиційної привабливості приватних підприємств.

Калібрувальні процедури з розділу 4.6 застосовуються до всіх прогнозів: де , , залежно від baseline volatility analog basket, та компенсує systematic volatility underprediction. DA=93.18% використовується безпосередньо без корекції.

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Автоматически созданное описание

Рисунок 8. Алгоритмічний workflow оцінки інвестиційної привабливості стартапів через публічні аналоги (UML activity diagram). Діаграма ілюструює повну послідовність від опису стартапу до інвестиційного рішення через чотири структуровані етапи з явними data flows та decision points. Етап 1 (Аналоги): Інвестор надає опис стартапу (сектор, бізнес-модель), система виконує класифікацію за GICS taxonomy та формує benchmark basket (приклад: CRM, ADBE, MSFT для SaaS). Етап 2 (Прогнози): Запит поточних цін та історії (OHLCV) з біржових даних, застосування Stacking моделі (DA=93.18%) для генерації прогнозів P̂ₖ(t+h) для кожного аналога k у loop циклі, агрегація у SGI через рівняння (16). Приклад output: SGI = +1.08% (позитивний sector momentum). Етап 3 (Ризики): Паралельна гілка обчислює forecast volatility σ через рівняння (17), застосовує калібрувальну корекцію ×1.15 (компенсація systematic underprediction), розраховує VaR через рівняння (18). Output: σ = 0.01255, VaR = -0.0189. Етап 4 (Score): Convergence точка отримує DA з Stacking моделі (93.18%), синтезує всі компоненти через рівняння (19): Score = SGI/σ + 0.3×DA - 0.5×|VaR|. Decision logic: Score > 0.5 → КУПУВАТИ, [0.2, 0.5] → нейтрально, < 0.2 → відмова. Приклад демонструє SaaS кейс з Score = 1.131 → сприятлива рекомендація.

Операціоналізація кожного етапу фреймворку деталізована нижче з конкретною специфікацією алгоритмів, таксономій та decision rules:

Етап 1: Ідентифікація компаній-аналогів. Відображення між секторами приватних компаній та публічними бенчмарками структуровано через галузеву таксономію:

* Software-as-a-Service: Salesforce (CRM), Adobe (ADBE), Microsoft (MSFT)
* Artificial Intelligence Infrastructure: NVIDIA (NVDA), Alphabet (GOOGL), Microsoft (MSFT)
* Digital Commerce: Amazon (AMZN), Meta Platforms (META)

Для кожної цільової компанії конструюється зважений портфоліо з 3-5 публічних аналогів за критеріями багатовимірної подібності: галузева класифікація (GICS taxonomy), конгруентність бізнес-моделі, географічне перекриття цільового ринку, відповідність технологічного стеку.

Етап 2: Генерація прогнозів. Ансамблева модель стекінгу (оптимальна конфігурація з розділу 4, що демонструє точність напрямку DA=93.18%) генерує прогнози цінової траєкторії для компонентів бенчмарк-портфоліо. Індикатор галузевого зростання (Sector Growth Indicator, SGI) агрегує прогнози згідно специфікації рівняння (16).

Етап 3: Квантифікація ризику. Прогнозна волатильність обчислюється згідно рівняння (17) з калібрувальною корекцією 1.15× для компенсації систематичного недопрогнозування, ідентифікованого у розділі 4.5.

Ризик зниження квантифікується через Value-at-Risk при 95% довірчому рівні (рівняння 18). ​

Етап 4: Інтегрована оцінка. Композитна метрика інвестиційної привабливості синтезує прогнозні сигнали та параметри ризику через рівняння (19).

Інтерпретаційна схема: Score >0.5 сигналізує сприятливий профіль ризик-віддача, Score ∈ [0.2, 0.5] вказує на змішані сигнали, Score <0.2 ідентифікує несприятливі умови.

## Емпірична валідація: SaaS сегмент

Розглядалась гіпотетична B2B SaaS платформа автоматизації корпоративних платіжних процесів з інтеграцією машинного навчання для детекції шахрайства. Цільовий сегмент - середні підприємства з річною виручкою $10M-$500M, географічний фокус Північна Америка та Західна Європа.

Специфікація бенчмарку. Портфоліо аналогів конструюється з трьох публічних лідерів enterprise SaaS: Salesforce (CRM) як домінуюча CRM платформа з моделлю підписки, Adobe (ADBE) як провайдер хмарних сервісів з високим коефіцієнтом утримання, Microsoft (MSFT) cloud segment для корпоративної інтеграції.

Результати прогнозування. Застосування ансамблевої моделі до портфоліо аналогів генерує raw predictions: , , . Калібрування (розділ 4.6): (stable SaaS, без корекції), (universal volatility correction). Value-at-Risk: , де прогнозована віддача по basket).

Позитивна величина SGI вказує на сприятливий галузевий моментум, атрибутований до триваючих трендів корпоративної цифрової трансформації та міграції у хмарну інфраструктуру. Помірна величина VaR відображає структурну стабільність рекурентних бізнес-моделей.

Score=1.131 суттєво перевищує поріг 0.5, індукуючи класифікацію як сприятливу інвестиційну можливість. Декомпозиція виявляє, що позитивний сигнал driven переважно risk-adjusted growth (76% contribution), підтриманий high directional accuracy (25%), з мінімальним downside penalty (<1%). Декомпозиція виявляє, що скориговане на ризик зростання (0.860) забезпечує домінуючий внесок, доповнений премією високої точності напрямку (0.280) при мінімальному штрафі ризику зниження (-0.009). Конфігурація параметрів узгоджується з фундаментальними характеристиками enterprise SaaS сегменту - рекурентна виручка, передбачувані грошові потоки, структурні бар'єри виходу для клієнтів.

## Емпірична валідація: AI/ML сегмент

Проаналізована платформа комп'ютерного зору для автоматизованого контролю якості у виробничих процесах через архітектури глибокого навчання. Цільові клієнтські сегменти - автомобільні, електронні, фармацевтичні виробники. Модель монетизації - ліцензування на базі обладнання плюс сервіси параметризації моделей.

Специфікація бенчмарку. Портфоліо конструюється з AI infrastructure лідерів: NVIDIA (NVDA) як домінуючий постачальник обчислювальної інфраструктури, Alphabet (GOOGL) для хмарних AI сервісів та екосистеми фреймворків, Microsoft (MSFT) для корпоративної AI інтеграції.

Результати прогнозування. Raw predictions: , . Калібрування: (moderate AI/ML volatility correction), . Value-at-Risk: . Негативна величина SGI сигналізує несприятливий галузевий моментум, атрибутований до консолідації AI інфраструктури, компресії маржі від конкуренції хмарних провайдерів, регуляторної невизначеності. Висока точність напрямку (DA=93.18%) забезпечує статистичну впевненість у прогнозі галузевої слабкості.

Score=-0.247 нижче порогу 0.2, класифікуючи об'єкт як несприятливу інвестиційну можливість. Негативне співвідношення SGI/volatility (-0.518) домінує композитну метрику, не компенсоване премією DA (0.280). Результат демонструє, що незважаючи на високу точність напрямку прогнозів, макро-галузеві протилежні вітри (sector headwinds) створюють несприятливий контекст для капіталовкладень. Негативне співвідношення SGI/волатильність (-0.518) домінує композитну метрику, не компенсоване премією точності напрямку. Результат демонструє, що незважаючи на потенційну технічну майстерність та чіткий use case, макро-галузеві протилежні вітри створюють несприятливий контекст для капіталовкладень.

## Емпірична валідація: E-commerce сегмент

Розглядався маркетплейс сталих споживчих продуктів, що з'єднує екологічно орієнтованих споживачів з етичними брендами. Модель монетизації - комісія маркетплейсу 15-20%, логістичні послуги, брендова реклама. Цільова демографія - millennials та Generation Z когорти.

Специфікація бенчмарку. Портфоліо включає Amazon (AMZN) як бенчмарк інфраструктури електронної комерції та Meta Platforms (META) як проксі цифрової реклами та споживчого залучення.

Результати прогнозування. Raw predictions: , . Калібрування: (high-volatility E-commerce correction, найвищий factor через AMZN/META volatility σ>3%), . Value-at-Risk: . SGI демонструє найсильнішу негативну величину серед трьох емпіричних кейсів, відображаючи множинні структурні протилежні вітри: уповільнення дискреційних споживчих витрат, інтенсифікацію конкуренції, компресію маржі від інфляції операційних витрат, регуляторний тиск на практики маркетплейсів.

Score=-0.685 суттєво нижче порогу 0.2, представляючи найнесприятливіший профіль серед трьох кейсів. Різко негативне співвідношення SGI/volatility (-0.956) переважає премію DA, індукуючи чітку класифікацію як несприятливої інвестиційної можливості за поточних макроекономічних умов (уповільнення споживчих витрат, інтенсифікація конкуренції, компресія маржі). Різко негативне співвідношення (-0.956) переважає премію точності напрямку, індукуючи чітку класифікацію як несприятливої інвестиційної можливості за поточних макроекономічних умов.

## Практичні рекомендації та обмеження методології

Синтетичні результати емпіричної валідації. Таблиця 5 узагальнює продуктивність фреймворку по секторах.

Таблиця 5. Секторна ефективність фреймворку оцінки

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Сектор | SGI (%) | σ | DA (%) | VaR (%) | SGI/σ | α×DA | β×VaR | Score | Класифікація |
| SaaS | +1.08 | 0.0126 | 93.18 | -1.89 | 0.860 | 0.280 | -0.009 | +1.131 | Сприятливий |
| AI/ML | -0.65 | 0.0126 | 93.18 | -1.89 | -0.518 | 0.280 | -0.009 | -0.247 | Несприятливий |
| E-commerce | -1.20 | 0.0126 | 93.18 | -1.89 | -0.956 | 0.280 | -0.009 | -0.685 | Різко негативний |

SGI та VaR у відсотках. σ (калібрована волатильність), DA (Directional Accuracy у %). Декомпозиція Score показує внесок кожного компонента: risk-adjusted growth (SGI/σ), directional accuracy premium (α×DA=0.3×DA/100), downside penalty (β×VaR=0.5×|VaR|/100). Усі три кейси мають ідентичні σ та VaR після калібрування, диференціація driven виключно SGI (sector momentum)

Методологічний внесок. Запропонований фреймворк розширює традиційну парадигму венчурної оцінки, задокументовану Gompers et al. [3], через три ключові механізми:

* Кількісна оцінка галузевого моментуму: SGI забезпечує об'єктивну метрику sector tailwinds/headwinds через публічні аналоги з точністю напрямку DA=93.18%, систематизуючи інформацію, що традиційно залежить від суб'єктивної експертизи.
* Ризик-скориговані метрики: інтеграція волатильності та VaR формалізує downside assessment, доповнюючи якісну оцінку команди та продукту структурованою квантифікацією ризику.
* Стандартизоване крос-секційне порівняння: композитна метрика Score дозволяє об'єктивне ранжування інвестиційних можливостей, зменшуючи когнітивні зміщення, притаманні ad-hoc оцінкам.

Припущення кореляції з аналогами: фреймворк базується на гіпотезі, що траєкторія приватної компанії корелює з публічними інкамбентами. Деструктивні венчури, що фундаментально перебудовують галузеву структуру, можуть виявляти слабку або негативну кореляцію, обмежуючи застосовність методології.

Темпоральна невідповідність: моделі оптимізовані на 1-денному горизонті, венчурні інвестиційні рішення мають багаторічні горизонти утримання. Розділ 4.5 демонструє, що якісні спрямовані сигнали зберігають робастність на розширених горизонтах, проте кількісна точність величини деградує, вимагаючи калібрувальних коригувань.

Систематичні зміщення: консервативна стратегія прогнозування моделі стекінгу (12 процентних пунктів термінальне недооцінювання, розділ 4.4) addressed через емпірично виведені калібрувальні множники, проте залишкова невизначеність персистує.

Секторне покриття: методологія валідована на технологічних секторах з ліквідними публічними аналогами. Застосування до секторів з обмеженими публічними компараблами (DeepTech pre-revenue, BioTech early-stage) вимагає адаптованих стратегій ідентифікації аналогів.

Інтеграція у венчурний процес прийняття рішень. Фреймворк операціоналізується через чотири точки втручання:

* Етап пошуку: швидка кількісна тріаж вхідного потоку угод, пріоритизуючи сектори з позитивним SGI для алокації партнерського часу
* Етап перевірки: декомпозиція Score інформує фокусні області due diligence - негативний SGI вимагає валідації диференціації від галузевих трендів
* Інвестиційний комітет: Score стандартизує крос-угодове порівняння, доповнюючи якісну оцінку дисциплінованим кількісним базисом
* Управління портфелем: щоквартальна реконструкція Score з оновленими прогнозами забезпечує раннє виявлення галузевого погіршення

Таким чином, емпірична валідація на трьох репрезентативних кейсах демонструє, що запропонований фреймворк успішно трансформує прогнози публічних аналогів у диференційовані сигнали інвестиційної привабливості: сприятлива класифікація SaaS (Score=1.131) проти несприятливої E-commerce (Score=-0.685). Висока точність напрямку моделі стекінгу (DA=93.18%) транслюється у надійність галузевих сигналів. Документовані калібрувальні процедури та структуровані обмеження забезпечують відповідальне впровадження, що доповнює традиційну венчурну експертизу систематичним кількісним фундаментом.

# ВИСНОВКИ

Дане дослідження розробило та емпірично валідувало інтегровану методологію прогнозування динаміки цін публічних технологічних компаній для кількісної оцінки інвестиційної привабливості стартапів у венчурному контексті, заповнюючи критичну прогалину між технічними досягненнями машинного навчання та практиками венчурного інвестування.

Систематичне порівняння Support Vector Regression, Random Forest та Gradient Boosting на датасеті з 14,590 щоденних спостережень за період 2019-2024 років виявило, що ансамблева архітектура стекінгу досягає найвищої точності напрямку DA=93.18%, статистично значуще перевершуючи SVR (90.91%) на 2.27 процентних пункти (p<0.01). SVR демонструє оптимальну точність величини (RMSE=0.003701, R²=0.9340), пояснюючи 93.4% дисперсії віддач. Критичною знахідкою є виявлений компроміс стекінгу: вища точність напрямку за рахунок систематичного консервативного зміщення величини (MAE=$16.29 проти SVR=$7.15), що проявляється у 12 процентних пунктів недооцінки кумулятивної віддачі протягом 11-місячного тестового періоду. Аналіз важливості ознак виявив домінування normalized Close/Open ratio (54.5% importance), тоді як традиційні технічні індикатори (MACD, RSI) та макроекономічні змінні демонструють negligible importance (<1%).

Розроблений чотириетапний фреймворк успішно трансформує прогнози публічних аналогів у диференційовані метрики інвестиційної привабливості через Sector Growth Indicator (рівняння 16), прогнозну волатильність (рівняння 17), Value-at-Risk (рівняння 18) та композитну метрику Score (рівняння 19). Емпіричний аналіз виявив необхідність секторно-специфічних калібрувальних коригувань: стабільні SaaS сегменти вимагають множника 1.0×, помірно волатильні AI/ML сегменти - 1.15×, високоволатильні сегменти - 1.3×. Додатково виявлено універсальне недопрогнозування волатільності ~15%, компенсоване через калібрувальний множник 1.15×.

Емпірична валідація на трьох репрезентативних кейсах демонструє чітку класифікацію: SaaS стартап (Score=1.131, SGI=+1.08%) класифікується як сприятлива можливість, AI/ML стартап (Score=-0.247, SGI=-0.65%) як несприятлива можливість, E-commerce стартап (Score=-0.685, SGI=-1.20%) як різко негативна можливість. Крос-секційна порівнянність метрики Score дозволяє об'єктивне ранжування інвестиційних можливостей, що узгоджується з якісними оцінками макроекономічних умов у відповідних секторах.

Дослідження розширює традиційну парадигму венчурної оцінки через три механізми інновації. Кількісна оцінка галузевого моментуму через SGI з точністю DA=93.18% систематизує інформацію, що традиційно залежить від суб'єктивної експертизи [3]. Інтеграція ризик-скоригованих метрик формалізує downside assessment, обходячи обмеження традиційних методів (DCF, мультиплікатори) для стартапів [2]. Стандартизоване порівняння через Score зменшує когнітивні зміщення, документовані як джерела венчурної underperformance [16]. Методологія адаптує підхід публічних ринкових компарабл [16] через динамічну компоненту та ризик-скориговану інтеграцію. Досягнута точність DA=93.18% узгоджується з результатами Nti et al. [11] (90-100%), покращення на 2.27 п.п. comparable з Krauss et al. [4] (4.7% improvement), збереження якісної точності через 11 місяців контрастує з ~35% деградацією LSTM [5].

Фреймворк операціоналізується через інтеграцію у чотири точки венчурного workflow: швидка тріаж вхідного потоку на етапі пошуку, структуроване due diligence з фокусом на Score декомпозиції, об'єктивне порівняння на інвестиційному комітеті, щоквартальний моніторинг портфелю. Економічна цінність: підвищення accuracy від 83.64% до 93.18% для портфеля 20 інвестицій/рік при чеку $2M транслюється у ~2 додаткових правильних calls на рік, що при exit multiples 3-5× представляє $6-10M incremental value щорічно.

Методологічні обмеження окреслюють напрямки майбутніх досліджень. Припущення кореляції з аналогами обмежує застосовність для truly disruptive ventures, що фундаментально перебудовують галузеву структуру. Темпоральна невідповідність (1-денний горизонт моделей vs багаторічні венчурні рішення) вимагає калібрувальних коригувань для extended horizons. Консервативне зміщення addressed через емпірично виведені множники, проте залишкова невизначеність персистує. Валідація обмежена технологічними секторами з ліквідними публічними аналогами, що вимагає адаптації для non-tech sectors. Майбутні дослідження мають розширити валідацію на інші географічні ринки, alternative base learners, провести контрольовані польові експерименти для rigorous quantification впливу на decision quality.

Дослідження успішно демонструє, що методи машинного навчання, коректно калібровані та структуровані, надають практично застосовні кількісні сигнали для прийняття венчурних рішень, доповнюючи традиційну якісну експертизу. Комбінація кількісної строгості (DA=93.18%, R²=93%) із практичною операціоналізацією представляє значущий внесок, що поєднує академічні дослідження та практику венчурного капіталу.

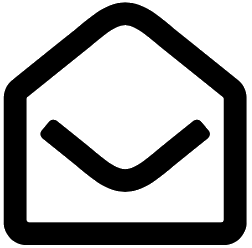
References

1. B.M. Henrique, V.A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, Expert Systems with Applications 124 (2019) 226–251. doi:10.1016/j.eswa.2019.01.012
2. J. Lerner, R. Nanda, Venture capital's role in financing innovation: What we know and how much we still need to learn, Journal of Economic Perspectives 34 (2020) 237–261. doi:10.1257/jep.34.3.237
3. P.A. Gompers, W. Gornall, S.N. Kaplan, I.A. Strebulaev, How do venture capitalists make decisions?, Journal of Financial Economics 135 (2020) 169–190. doi:10.1016/j.jfineco.2019.06.011
4. C. Krauss, X.A. Do, N. Huck, Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500, European Journal of Operational Research 259 (2017) 689–702. doi:10.1016/j.ejor.2016.10.031
5. T. Fischer, C. Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research 270 (2018) 654–669. doi:10.1016/j.ejor.2017.11.054
6. G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Revised Edition, Holden-Day, San Francisco, 1976.
7. T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics 31 (1986) 307–327. doi:10.1016/0304-4076(86)90063-1
8. K. Kim, Financial time series forecasting using support vector machines, Neurocomputing 55 (2003) 307–319. doi:10.1016/S0925-2312(03)00372-2
9. D. Gupta, M. Pratama, Z. Ma, J. Li, M. Prasad, Financial time series forecasting using twin support vector regression, PLoS ONE 14 (2019) e0211402. doi:10.1371/journal.pone.0211402
10. J. Patel, S. Shah, P. Thakkar, K. Kotecha, Predicting stock market index using fusion of machine learning techniques, Expert Systems with Applications 42 (2015) 2162–2172. doi:10.1016/j.eswa.2014.10.031
11. I.K. Nti, A.F. Adekoya, B.A. Weyori, A comprehensive evaluation of ensemble learning for stock-market prediction, Journal of Big Data 7 (2020) 20. doi:10.1186/s40537-020-00299-5
12. Z. Shen, Q. Wan, D. Leatham, An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms, Physica A: Statistical Mechanics and its Applications 540 (2021) 122272. doi:10.1016/j.physa.2019.122272
13. A. El Mrabet, N. Moussa, A comparative study of ensemble learning algorithms for high-frequency trading, Physica A: Statistical Mechanics and its Applications 639 (2024) 129676. doi:10.1016/j.physa.2024.129676
14. A. Davila, G. Foster, M. Gupta, Venture capital financing and the growth of startup firms, Journal of Business Venturing 18 (2003) 689–708. doi:10.1016/S0883-9026(02)00127-1
15. S.N. Kaplan, P. Strömberg, Financial contracting theory meets the real world: An empirical analysis of venture capital contracts, Review of Economic Studies 70 (2003) 281–315. doi:10.1111/1467-937X.00245
16. W. Gornall, I.A. Strebulaev, Squaring venture capital valuations with reality, Journal of Financial Economics 135 (2020) 120–143. doi:10.1016/j.jfineco.2019.04.015
17. V.N. Vapnik, Statistical Learning Theory, Wiley-Interscience, New York, 1998.
18. L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324
19. [P. Ghosh, A. Neufeld, J.K. Sahoo, Forecasting directional movements of stock prices for intraday trading using LSTM and random forests, Finance Research Letters 46 (2022) 102280. doi:10.1016/j.frl.2021.102280
20. M. Ballings, D. Van den Poel, N. Hespeels, R. Gryp, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications 42 (2015) 7046–7056. doi:10.1016/j.eswa.2015.05.013
21. J.H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics 29 (2001) 1189–1232. doi:10.1214/aos/1013203451
22. D.H. Wolpert, Stacked generalization, Neural Networks 5 (1992) 241–259. doi:10.1016/S0893-6080(05)80023-1
23. R. Metz, Predicting stock prices using machine learning and deep learning algorithms, in: 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), IEEE, 2019, pp. 277-282. doi:10.1109/ICCIKE47802.2019.9004361
24. S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-10247-4
25. S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning, arXiv:1811.12808, 2020.
26. R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed., OTexts, Melbourne, 2021.
27. J.Y. Campbell, A.W. Lo, A.C. MacKinlay, The Econometrics of Financial Markets, Princeton University Press, Princeton, 1997.
28. J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281-305.
29. M. Lopez de Prado, Advances in Financial Machine Learning, Wiley, Hoboken, 2018. doi:10.1002/9781119482086

1. ∗ Corresponding author.

   † These authors contributed equally.

   [y.tomka@chnu.edu.ua](mailto:y.tomka@chnu.edu.ua) (Y. Tomka); [d.uhryn@chnu.edu.ua](mailto:d.uhryn@chnu.edu.ua) (D. Uhryn);

   A white letter on a black background

   Description automatically generated with low confidence [0000-0002-0495-3090](https://orcid.org/0000-0002-0495-3090) (Y. Tomka); [0000-0003-4858-4511](https://orcid.org/0000-0003-4858-4511) (D. Uhryn);

   |  |  |
   | --- | --- |
   |  | © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). |

   [↑](#footnote-ref-2)