

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук
(повна назва)
Кафедра _____ Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти _____ другий (магістерський)

Семантичне прогнозування криз на фінансових ринках у реальному часі з
використанням MFFMs/LLM
(тема)

Виконав:
здобувач _____ другого _____ року навчання,
групи _____ СШМ-24-1

_____ Владислав НЕДОДАЄВ
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми _____ освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)

Керівник _____ проф. Олександр ШЕВЧЕНКО
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Лариса ЧАЛА
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____
Кафедра _____ Штучного інтелекту _____
Рівень вищої освіти _____ другий (магістерський) _____
Спеціальність _____ 122 Комп'ютерні науки _____
(код і повна назва)
Тип програми _____ освітньо-наукова _____
(освітньо-професійна або освітньо-наукова)
Освітня програма _____ Системи штучного інтелекту _____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

«_____» _____ 20 ____ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві _____ Недодаєву Владиславу Андрійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Семантичне прогнозування криз на фінансових ринках у реальному часі з використанням MFFMs/LLM _____

затверджена наказом університету від _____ 6 _____ квітня _____ 20 26 р. № 269Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії _____ 22 _____ травня _____ 20 26 р.

3. Вихідні дані до роботи _____ Науково технічні публікації, дані Інтернет-джерел та відомих наукових проєктів, документація Python, бібліотеки scikit-learn, PyTorch, FAISS, OpenAI API, історичні фінансові дані (S&P500, VIX, FRED), тексти засідань FOMC та Beige Book. _____

4. Перелік питань, що потрібно опрацювати в роботі _____

- 1) Аналіз предметної галузі та огляд існуючих рішень _____
 - 2) Розробка конвеєра кількісних ознак та текстових даних для прогнозування криз _____
 - 3) Проєктування та реалізація системи прогнозування на основі ансамблю моделей _____
 - 4) Експериментальне дослідження та валідація системи на історичних даних _____
- _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	06.04.2026	виконано
2	Аналіз предметної галузі та огляд існуючих рішень	10.04.2026	виконано
3	Розробка конвеєра кількісних ознак	12.04.2026	виконано
4	Проведення експериментальних досліджень	26.04.2026	виконано
5	Написання пояснювальної записки	04.05.2026	виконано
6	Перевірка на академічний плагіат	05.05.2026	виконано
7	Нормоконтроль	08.05.2026	виконано
8	Рецензування	10.05.2026	виконано
9	Підготовка презентації та доповіді	12.05.2026	виконано
10	Попередній захист	14.05.2026	виконано
11	Захист перед ЕК	22.05.2026	виконано

Дата видачі завдання 6 квітня 2026 р.

Здобувач 
(підпис)

Керівник роботи  проф. Олександр ШЕВЧЕНКО
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 76 с., 5 рис., 16 табл., 6 дод., 16 джерел.

АНАЛІЗ ЧАСОВИХ РЯДІВ, АНСАМБЛЬ МОДЕЛЕЙ, ВЕЛИКІ МОВНІ МОДЕЛІ, МАШИННЕ НАВЧАННЯ, ПОШУК НАЙБЛИЖЧИХ СУСІДІВ, ПРОГНОЗУВАННЯ ФІНАНСОВИХ КРИЗ, РІДЖ-РЕГРЕСІЯ, СИСТЕМИ РАННЬОГО ПОПЕРЕДЖЕННЯ, BEIGE BOOK, FOMC, RAG.

Об'єкт дослідження – системи раннього попередження фінансових криз з використанням методів машинного навчання та семантичного аналізу текстових джерел.

Предмет дослідження – методи підвищення точності та надійності прогнозування фінансових криз шляхом комбінування кількісних ринкових індикаторів та LLM-кодування на основі текстових джерел даних у мультифакторній фінансовій моделі.

Мета роботи – дослідження та емпірична валідація методу прогнозування криз фінансових ринків шляхом розробки гібридної системи, що поєднує кількісні ринкові індикатори з семантичним аналізом текстів FOMC та Beige Book у конвеєрі retrieval-augmented generation.

Методи дослідження – теоретичні (аналіз наукової літератури з систем раннього попередження, порівняння існуючих підходів до прогнозування криз) та практичні (побудова пайплайну ознак, walk-forward бектестування, блоковий бутстрап, абляційні дослідження).

Запропонована система забезпечує інтерпретовані прогнози з історичними аналогами та демонструє реальну прогностичну здатність за межами простої детекції поточних кризових станів, що підтверджено тестуванням.

ABSTRACT

Master's thesis contains: 76 pp., 5 fig., 16 tabl., 6 ann., 16 references.

BEIGE BOOK, EARLY WARNING SYSTEMS, FINANCIAL CRISIS FORECASTING, FOMC, K-NEAREST NEIGHBORS, LARGE LANGUAGE MODELS, MACHINE LEARNING, MODEL ENSEMBLE, RAG, RIDGE REGRESSION, TIME SERIES ANALYSIS.

Object of research – early warning systems for financial crises using machine learning methods and semantic analysis of textual sources.

Subject of research – methods for improving the accuracy and reliability of financial crisis forecasting by combining quantitative market indicators with LLM-based encoding of textual data sources in a multi-factor financial model.

Objective – to investigate and empirically validate a method for financial market crisis forecasting by developing a hybrid system that combines quantitative market indicators with semantic analysis of FOMC and Beige Book texts in a retrieval-augmented generation pipeline.

Methods of research – theoretical (analysis of scientific literature on early warning systems, comparison of existing approaches to crisis forecasting) and practical (feature pipeline construction, walk-forward backtesting, block bootstrap, ablation studies).

Proposed system provides interpretable forecasts with historical analogues and demonstrates genuine forecasting ability beyond simple detection of current crisis states, confirmed by testing.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів	8
Вступ.....	9
1 Аналіз предметної галузі та огляд існуючих рішень.....	11
1.1 Фінансові кризи та їх вплив на економіку	11
1.2 Актуальність дослідження.....	12
1.3 Методи прогнозування фінансових криз	13
1.3.1 Традиційні індикаторні моделі.....	13
1.3.2 Методи машинного навчання	14
1.3.3 Retrieval-based підходи та LLM-аналіз	15
1.4 Методи прогнозування фінансових криз	16
1.5 Постановка задачі.....	18
2 Дані та методологія дослідження	19
2.1 Каталог фінансових криз	19
2.2 Конвеєр кількісних ознак	19
2.2.1 Джерела даних та категорії ознак	19
2.2.2 Опис ознак за ризиковими вимірами	20
2.2.3 Діагностика стаціонарності	27
2.2.4 Побудова вікон та розподіл міток.....	28
2.3 Текстові дані FOMC та Beige Book	29
2.4 Механізм запобігання витоку даних	30
3 Розробка системи прогнозування фінансових криз.....	33
3.1 Рідж-регресія з експоненційним зважуванням.....	33
3.2 Пошук K найближчих сусідів (Retrieval DD)	35
3.3 MLP з триплетним навчанням	36
3.4 Ансамблева архітектура	37
3.5 Інтеграція LLM	37
4 Експериментальні дослідження та результати.....	44
4.1 Протокол walk-forward валідації.....	44

4.2 Результати прогнозування падіння.....	44
4.3 Абляційні дослідження	48
4.4 Аналіз ймовірності переобфитування бектесту (PBO).....	50
4.5 Frozen-model holdout тест	52
4.6 Покризовий аналіз системи раннього попередження.....	54
4.6.1 Крах dot-com бульбашки 2000–2002	54
4.6.2 Глобальна фінансова криза 2007–2009	55
4.6.3 Пандемія COVID-19, 2020.....	56
4.7 Практична реалізація	57
4.8 Реалізація веб-системи моніторингу	58
4.9 Подальші дослідження.....	60
4.10 Тест економічної цінності: стратегія Risk-On/Risk-Off.....	61
Висновки	63
Перелік джерел посилання	65
Додаток А Повний перелік ознак конвеєра.....	67
Додаток Б Діаграма доступності джерел даних.....	72
Додаток В Кореляційна матриця ознак.....	73
Додаток Г Відсоток пропущених значень по ознаках.....	74
Додаток Д LOCO feature importance для Ridge	75
Додаток Е Відомість кваліфікаційної роботи	76

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ФРС – Федеральна резервна система США;

ІІІ – штучний інтелект;

AUROC – Area Under Receiver Operating Characteristic – площа під ROC-кривою;

BB – Beige Book – Бежева книга Федеральної резервної системи;

BSS – Brier Skill Score – показник якості ймовірного прогнозу;

CAPE – Cyclically Adjusted Price-to-Earnings – циклічно скоригований коефіцієнт P/E;

CSCV – Combinatorially Symmetric Cross-Validation – комбінаторно-симетрична крос-валідація;

FAISS – Facebook AI Similarity Search – бібліотека пошуку подібних векторів;

FOMC – Federal Open Market Committee – Федеральний комітет з відкритого ринку;

GFC – Global Financial Crisis – глобальна фінансова криза 2008 р.;

GPR – Geopolitical Risk Index – індекс геополітичного ризику;

K-NN – K-Nearest Neighbors – K найближчих сусідів;

LOCO – Leave-One-Column-Out – метод вилучення однієї ознаки;

MAE – Mean Absolute Error – середня абсолютна похибка;

MFFM – Multi-Factor Financial Foundation Model – мультифакторна фінансова базова модель;

MLP – Multilayer Perceptron – багатошаровий перцептрон;

PCA – Principal Component Analysis – метод головних компонент;

RAG – Retrieval-Augmented Generation – генерація з доповненням пошуком;

VIX – CBOE Volatility Index – індекс волатильності Чиказької біржі опціонів.

ВСТУП

У 2008 році глобальна фінансова криза зрізала індекс S&P 500 на 56,8%. Через дванадцять років, у березні 2020-го, пандемія COVID-19 вивела індекс волатильності VIX на позначку 82,7, найвищу за всю історію спостережень. Інфляційне падіння 2022 року розтягнулося на 368 днів і забрало у інвесторів 25,4% вартості портфелів. Існування інструменту, здатного вловити ознаки кризи заздалегідь, дозволило б інвесторам отримати час на перебалансування позицій та підготуватися до падіння індексу.

Класичні способи прогнозування: порогові моделі на базі VIX, логістичні регресії на кривій дохідності, моделі кредитних спредів, працюють з невеликим набором наперед обраних ознак. Вони не враховують того, що передвісники криз мають мультимодальну природу і залежать від контексту. В літературі з систем раннього попередження часто поєднують два різних завдання. Бінарна класифікація з метриками AUROC чи AUPRC здебільшого вимірює, чи вміє модель «побачити» вже наявний стрес, а не спрогнозувати майбутнє погіршення зі спокійного стану.

Задача прогнозування фінансових криз залишається відкритою через фундаментальну складність – фінансові ринки є нестаціонарними системами зі зміною режимів, малою кількістю кризових подій та високим рівнем шуму. Водночас розвиток методів машинного навчання, великих мовних моделей та технологій retrieval-augmented generation відкриває нові можливості для побудови систем, що поєднують кількісні індикатори з семантичним аналізом текстів.

Сучасні фінансові ринки характеризуються зростаючою складністю та взаємопов'язаністю. Глобалізація фінансових потоків, розвиток деривативів та алгоритмічної торгівлі створюють нові канали контагіону, що прискорюють поширення криз. Водночас зростання обсягу доступних даних – як структурованих (ринкові індикатори, макроекономічна

статистика), так і неструктурованих (протоколи засідань центральних банків, новинні потоки) – відкриває можливості для побудови більш комплексних систем раннього попередження.

Особливо перспективним є використання текстів як додаткового джерела випереджальної інформації. Протоколи засідань FOMC та звіти Beige Book містять оцінки економічного стану, що базуються на приватній інформації, недоступній у ринкових цінах: внутрішні прогнози ФРС, результати консультацій з бізнесом, регіональні економічні спостереження. Ця інформація стає публічною з лагом від кількох тижнів до місяця, що створює вікно прогностичної переваги.

Технологія retrieval-augmented generation (RAG) [1] забезпечує природний фреймворк для інтеграції кількісних та текстових даних. Замість прямого генерування прогнозів, модель шукає найбільш подібні історичні періоди у векторному індексі, а потім використовує їх як контекст для каліброваного прогнозування. Це забезпечує інтерпретованість – аналітик може зрозуміти, на яких історичних аналогах базується прогноз.

Наукова новизна отриманих результатів полягає в тому, що вперше запропоновано гібридну систему, яка поєднує кількісні ринкові індикатори з текстовими ознаками з документів центральних банків у ансамблевій архітектурі з Рідж-регресією, K-NN та триплетним MLP. Встановлено, що текстові ознаки з FOMC та Beige Book надають статистично значущий додатковий прогностичний сигнал виключно для регуляризованих лінійних моделей.

Розроблена система може бути використана для моніторингу системних ризиків фінансових ринків, управління інвестиційним портфелем та побудови систем раннього попередження для регуляторів.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

1.1 Фінансові кризи та їх вплив на економіку

Прогнозування фінансових криз залишається одним із найскладніших завдань фінансової аналітики та макроекономічного моніторингу. Під фінансовою кризою тут мається на увазі період, коли стан ринків різко погіршується: фондові індекси падають, волатильність зростає, кредитні спреди розширюються, а міжбанківський ринок перестає працювати у звичному режимі. Кризи бувають різними за своєю природою: біржові крахи, банківські та валютні кризи, кризи ліквідності, суверенні боргові кризи, а також кризи зараження, коли проблеми з одного сектору перекидаються на решту фінансової системи.

Системи раннього попередження (Early Warning Systems, EWS) будуються для того, щоб помітити ознаки наближення кризи достатньо рано і залишити час на захисні дії. Перші такі системи зводилися до простих порогових правил на окремих показниках. З часом підходи ускладнювалися і дійшли до багатофакторних моделей, що одночасно враховують десятки ринкових і макроекономічних змінних. Базу для сучасних EWS заклали Банк міжнародних розрахунків (BIS) та Європейський центральний банк (ECB), запровадивши сигнальний метод, де кожен індикатор оцінюється незалежно за фіксованим порогом [2].

До ключових кількісних індикаторів кризового стану ринку належать: індекс волатильності VIX (його ще називають «індексом страху»), кредитні спреди, тобто різниця між дохідністю корпоративних і державних облігацій, яка вказує на рівень кредитного ризику, нахил кривої дохідності (різниця 10-річної та 2-річної ставок Казначейства), інверсія якого передувала кожній рецесії в США починаючи з 1960-х років [3], та індекс фінансових умов NFCI, що агрегує дані з кредитних, боргових і ринкових джерел.

Поряд з кількісними показниками суттєву роль відіграють текстові джерела. Протоколи засідань Федерального комітету з відкритого ринку (FOMC), звіти Beige Book, аналітика МВФ, новинні потоки містять відомості про очікування регуляторів, поточний стан економіки та потенційні ризики. Завдяки великим мовним моделям (LLM) з'явилася можливість кодувати ці тексти у числові ознаки ризику, які можуть доповнити звичний набір кількісних індикаторів.

Таким чином, предметна галузь охоплює теоретичні засади прогнозування фінансових криз, методологію побудови EWS, кількісне моделювання ринкових ризиків та семантичний аналіз фінансових текстів. На цей час галузь переживає перехід від однофакторних порогових моделей до мультимодальних систем, де кількісні та текстові дані обробляються спільно засобами машинного навчання.

1.2 Актуальність дослідження

Серйозні фінансові кризи відбуваються з помітними наслідками. В період з 1970 по 2024 роки зафіксовано 67 кризових подій різного типу та масштабу. Каталог Рейнхарт–Рогофф, що охоплює більш ніж 800 років фінансової історії, показує, що кризи не є аномалією – вони супроводжують будь-яку фінансову систему постійно [4]. Ще одним підтвердженням є база даних МВФ Лаевен–Валенсія, яка фіксує 431 системну банківську кризу за 1970–2017 роки [5].

Між тим наявні системи раннього попередження мають серйозний недолік: вони часто не розмежовують визначення поточної кризи і прогнозування наступної. Бойтель та співавтори [6] перевірили, як ML-методи, навчені на бінарних мітках банківських криз, порівнюються з логістичною регресією на даних 45 країн. Покращення виявилось мізерним, а сигнал зосередився в кредитному зростанні та цінах активів, тобто в індикаторах, що і так відображають поточний стрес. Ворд [7] отримав

подібний результат у масштабному порівнянні ансамблевих ML-класифікаторів.

Великі мовні моделі відкрили новий напрямок для роботи з текстовою фінансовою інформацією. Ву та співавтори [8] у своєму огляді зауважили, що LLM непогано справляються з якісним аналізом (настрої, наративи), але стабільного кількісного виграшу в прогнозуванні поки що не дають. Звідси виникає потреба з'ясувати, за яких умов текстова модальність справді додає корисний сигнал до кількісних ознак.

Окремо Лопес де Прадо [9] застерігає щодо ризику перенавчання бектестів через багаторазове тестування стратегій і метрик. Це обґрунтовує необхідність суворих протоколів оцінювання: фіксовані заздалегідь метрики, блоковий бутстрап для врахування часової автокореляції та чесне звітування негативних результатів.

1.3 Методи прогнозування фінансових криз

У літературі запропоновано різні підходи до прогнозування криз. Їх зручно згрупувати за трьома напрямками: традиційні індикаторні моделі, методи машинного навчання та retrieval-based підходи з LLM-аналізом.

1.3.1 Традиційні індикаторні моделі

Найпростіший підхід – порогова модель на окремому індикаторі. Наприклад, порогова модель VIX позначає стан ринку як кризовий, якщо значення VIX перевищує задану межу (30 або 40). Плюси очевидні: модель проста і зрозуміла, легка у використанні. Але VIX реагує на поточний стан, а не на майбутній ризик. Стрибок VIX вище 30 може вказувати як на серйозну системну кризу, так і на швидкоплинний спалах волатильності, зумовлений неочікуваними подіями і за одним числом ці ситуації складно розрізнити.

Інверсія кривої дохідності (від'ємний спред між 10-річною та 2-річною ставками Казначейства) вважається одним із найвідоміших макроекономічних передвісників рецесії. Шуларік та Тейлор [10] на матеріалі 140 років показали, що кредитне зростання є найкращим одиничним передвісником фінансових криз. Проте будь-яка індикаторна модель оцінює кожен показник окремо і не здатна врахувати взаємодію між ознаками.

Індекс турбулентності Крітцмана [11] кількісно оцінює «незвичність» поточного ринкового стану через відстань Махаланобіса вектора дохідностей від їхнього історичного середнього. Коефіцієнт абсорбції показує, яку частку дисперсії ринку пояснюють перші головні компоненти: що вона більша, то вищий системний ризик. Обидва методи є багатовимірними, але вони залишаються статистичними мірами і не навчаються на кризових епізодах минулого.

1.3.2 Методи машинного навчання

Стандартною відправною точкою для бінарної класифікації криз слугує логістична регресія. Бойтель та співавтори [6] зіставили логістичну регресію, випадкові ліси та градієнтний бустинг на даних 45 країн. Різниця між методами виявилася незначною. Ворд [7] дійшов того самого висновку в окремому комплексному порівнянні ML-класифікаторів.

На малих фінансових датасетах напрочуд добре працює Ridge-регресія (лінійна регресія з L2-регуляризацією). Регуляризація пригнічує шум і запобігає перенавчанню, а це особливо важливо, коли незалежних кризових епізодів мало: за 40 років їх набирається лише 8–15. У дослідженні Ridge стабільно досягає або перевершує результати нелінійних моделей.

XGBoost (градієнтний бустинг на деревах рішень) часто обирають для табличних даних, бо він уміє моделювати нелінійні залежності. Але на

нестационарних фінансових рядах емпірична перевірка виявила у XGBoost серйозне перенавчання.

З нейромережових архітектур протестовано три: MLP Retrieval DD (K-NN скоринг на навчених вкладеннях), прямий MLP-регресор та LSTM-регресор. Обидві end-to-end архітектури помітно програли Ridge. MLP Retrieval DD при цьому показав найвищий результат серед усіх моделей, але причина не в нейромережевому передбаченні як такому, а в механізмі K-NN скорингу. Абляція із перемішуванням послідовності LSTM підтвердила, що часовий порядок не несе додаткового сигналу.

1.3.3 Retrieval-based підходи та LLM-аналіз

Retrieval-based підходи виходять з гіпотези, що фінансові кризи мають повторювані структурні патерни: каскади зараження, спіралі ліквідності, обвали настроїв. Ці патерни проявляються як схожі «відбитки» у просторі ознак, навіть якщо поверхневі тригери зовсім різні. Метод кодує поточний стан ринку у вектор ознак і шукає K найближчих сусідів серед збережених історичних станів у векторній базі (FAISS). Оцінка кризового ризику формується як зважена частка кризових періодів серед знайдених сусідів.

Помітний плюс retrieval-based підходу в тому, що він стійкий до виключення окремих кризових епізодів. Аналіз leave-one-crisis-out показав: якщо прибрати найбільшу кризу (GFC, падіння до 56,8%), Spearman ρ моделі Retrieval DD знижується лише на 0,039 (горизонт 365d). Для Ridge втрата в аналогічних умовах становить 0,135. K-NN скоринг приблизно рахує серйозність падіння, спираючись на менші кризи, тоді як Ridge потребує повного діапазону цільової змінної для коректної параметричної оцінки. На малих кризових датасетах ця відмінність набуває принципового значення.

Retrieval-Augmented Generation (RAG) [1] поєднує пошук у векторній базі з LLM-міркуванням. У перших ітераціях запропонованої системи LLM отримує кількісний контекст разом із top-K історичними збігами і на їх

основі формує структуровану відповідь: тип кризи, рівень серйозності, пояснення, рекомендації з моніторингу. Дати анонімізовано, кризові мітки з витягнутих подій видалено, щоб запобігти контамінації. Емпірична перевірка показала, що LLM (GPT-4o mini) відтворює числову оцінку `crisis_ratio` з точним збігом на всіх фолдах ($\Delta = 0,000$), тобто жодного кількісного покращення не вносить.

Оскільки LLM не може покращити прогнози через пряму оцінку поточної ситуації, було сформовано та протестовано інший підхід: кодування текстів з перетворенням на індикатори. Текстову модальність перевірено на кількох типах джерел. Новинні заголовки (632 406 статей NYT) кодувалися трьома методами (sentence-transformer, OpenAI embeddings, LLM reasoning scores). В усіх випадках газетний текст виявився запізнілим індикатором, що нічого не додає поверх кількісних ознак. Зовсім інший результат дали документи ФРС: LLM-кодування 501 заяви та протоколів FOMC у шість структурованих вимірів ризику (`stress_intensity`, `contagion_breadth`, `policy_urgency`, `forward_risk`, `market_dysfunction`, `systemic_leverage`) покращило прогнозування на $\Delta\rho = +0,018 / +0,055 / +0,037 / +0,020$ (Spearman, 30/90/180/365d). Висновок простий: джерело тексту важливіше за спосіб кодування.

1.4 Методи прогнозування фінансових криз

Оцінювання систем прогнозування криз вимагає спеціальних метрик. Задача має свої особливості: кризові періоди займають менше 10% спостережень, дані сильно автокорельовані в часі, а кількість справді незалежних кризових епізодів мала.

Рангова кореляція Спірмена (ρ) виступає основною метрикою для регресії на прогнозоване падіння. Вона вимірює монотонну залежність між прогнозованим і фактичним падінням. Перевага такого вибору в тому, що безперервні цілі роблять кожен фолд інформативним, тоді як при бінарних

мітках лише приблизно 8 незалежних кризових епізодів дають статистичну потужність. Значення лежать у межах від -1 до 1 ; одиниця відповідає ідеальній монотонній залежності.

AUROC (Area Under Receiver Operating Characteristic) оцінює здатність моделі розрізняти кризові та некризові стани. У бінарній класифікації AUROC вважається стандартною метрикою, проте в контексті цієї роботи вона здебільшого фіксує здатність розпізнати поточний стрес, а не спрогнозувати майбутній. Діапазон: від 0 до 1 , де $0,5$ відповідає випадковому класифікатору.

Brier Score показує, наскільки добре відкалібровані ймовірності кризи. Brier Skill Score (BSS) порівнює модель з кліматологічним базовим рівнем, тобто з історичною частотою криз у тренувальній вибірці. Додатний BSS означає, що модель працює краще за базовий рівень, від'ємний – гірше.

Для побудови довірчих інтервалів застосовується блоковий бутстрап. На відміну від звичайного випадкового семплінгу, тут семплюються блоки послідовних фолдів, що зберігає кореляційну структуру даних і робить статистичні висновки коректнішими.

Walk-forward бектестування з розширюваним вікном відтворює умови реальної експлуатації: для кожного тижневого фолду модель навчається на всіх наявних даних до поточної дати. При такій процедурі look-ahead bias виключено. У цьому дослідженні walk-forward бектест охоплює 1618 тижневих фолдів (W-MON) за період 1996–2026.

Leave-one-crisis-out (LOCO) аналіз перевіряє стійкість моделі: кожен кризовий епізод по черзі виключається і з навчання, і з оцінювання. Якщо модель зберігає додатну кореляцію без будь-якої окремої кризи, це свідчить, що сигнал розподілений між кількома незалежними епізодами і не тримається на якомусь одному.

1.5 Постановка задачі

Мета кваліфікаційної роботи – дослідити та розробити метод семантичного прогнозування криз фінансових ринків у реальному часі, що спирається на мультифакторні фінансові моделі та LLM-аналіз.

Якщо мету буде досягнуто, отримаємо систему раннього попередження, здатну прогнозувати майбутні падіння ринку зі спокійних некризових умов. Практична цінність такої системи – в управлінні інвестиційними портфелями та моніторингу ризиків.

Аналіз предметної галузі дозволив сформулювати такі задачі кваліфікаційної роботи:

- побудувати каталог анотованих кризових подій з PCA-каліброваними оцінками серйозності та валідованими датами настання;
- розробити пайплайн кількісних ознак із ринкових та макроекономічних джерел (Yahoo Finance, FRED) зі строгими гарантіями відсутності витоку з майбутнього та формальною діагностикою стаціонарності;
- дослідити внесок текстової модальності з різних джерел (новинні заголовки, документи FOMC, звіти Beige Book) та різних методів кодування (sentence-transformer, OpenAI embeddings, LLM reasoning scores);
- реалізувати walk-forward бектестування з розширюваним вікном (1618 тижневих фолдів, 1996–2026) та блоковим бутстрапом для статистично коректного порівняння моделей;
- виконати аналіз результатів: абляційні дослідження, leave-one-crisis-out аналіз, порівняння з базовими моделями (VIX threshold, логістична регресія, індекс турбулентності, коефіцієнт абсорбції).

Виконання цих завдань дасть змогу побудувати науково обґрунтовану систему прогнозування фінансових криз, де кількісні індикатори працюватимуть разом із семантичним аналізом текстових джерел.

2 ДАНІ ТА МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ

2.1 Каталог фінансових криз

Каталог криз є основою системи маркування для навчання з учителем. Він містить 67 анотованих подій за період 1970–2024 років (54,1 роки). Кожна подія характеризується 15 обов'язковими полями, зокрема: унікальний ідентифікатор (формат YYYY_event_name), дата початку, дата піку (максимального ринкового стресу), дата стабілізації, тип кризи, оцінка серйозності та географічний охопит.

Оцінка серйозності (severity score) калібрована за допомогою PCA. Першу головну компоненту трьох метрик – max_drawdown_pct (S&P 500 peak-to-trough), VIX_peak та crisis_duration_days – нормалізовано до діапазону [0, 1].

Для валідації дат початку криз було проведено порівняння з незалежними джерелами (NBER, BIS, IMF). Медіанне відхилення склало 12 днів, що є прийнятним для щотижневої гранулярності системи.

2.2 Конвеєр кількісних ознак

Конвеєр кількісних ознак охоплює 73 ознаки (після всіх ітерацій розширення та LOCO-прунінгу) за період 1986–2026 років. Нижче детально описано джерела даних, категорії ознак, діагностику стаціонарності, побудову вікон та ітеративне розширення конвеєра.

2.2.1 Джерела даних та категорії ознак

Ознаки розподілені за такими категоріями:

- кредитно-макроекономічні індикатори: кредитні спреди (OAS), індекс фінансових умов (NFCI), міжбанківський стрес (SOFR-FF spread), процентні ставки ФРС;
- оцінка вартості: коефіцієнт P/E Шиллера (CAPE), премія за ризик акцій (ERP);
- левередж: маржинальний борг до ВВП;
- міжнародний стрес: JPY/USD (carry-trade proxy), Nikkei 225 (proxy азійського ринку);
- геополітичний ризик: індекс Калдара-Яковієлло (GPR), структуровані події GDELT;
- волатильність: VIX, реалізована волатильність S&P 500, VRP;
- ринок житла: CSUSHPISA, запас житла, іпотечна ставка;
- споживчі індикатори: Michigan Sentiment, автопродажі;
- сировинні: прибутковість та волатильність нафти, міді, коефіцієнт мідь/золото.

Кожна ознака має емпірично верифікований лаг публікації відповідно до офіційних календарів релізів BEA/Fed: ВВП – 121 день, маржинальний борг – 164 дні, CAPE – 45 днів. Це забезпечує реалістичність системи в режимі реального часу. Повний перелік 73 ознак конвеєра із зазначенням джерела даних, лагу публікації та класифікації стаціонарності наведено в додатку А (таблиця А.1). Діаграму доступності джерел даних за роками наведено в додатку Б (рисунок Б.1).

2.2.2 Опис ознак за ризиковими вимірами

Конвеєр обчислює 73 кількісні ознаки, організовані за 16 ризиковими вимірами. Нижче наведено опис ключових категорій. Повний перелік ознак із джерелами та класифікацією стаціонарності наведено в додатку А (таблиця А.1).

Стрес на ринку акцій (4 ознаки). Drawdown від 30-денного rolling high (sp500_drawdown_30d) завжди ≤ 0 , відображає масштаб нещодавнього падіння. 7-денна та 30-денна прибутковості (sp500_return_7d, sp500_return_30d) є сигналами моментуму: швидке падіння ($< -5\%$) часто передуює або супроводжує початок кризи. Відстань від 200-денної ковзної середньої (sp500_distance_200d_ma) – класичний індикатор тренду: від'ємні значення сигналізують, що ринок торгується нижче довгострокового середнього.

Режим волатильності (4 ознаки). VIX (vix_level) є ринковою імплікованою 30-денною волатильністю, відомою як «індикатор страху». Діапазон у датасеті: [9,14; 79,13]. Перевірка ADF підтверджує стаціонарність ($p = 0,00002$) – VIX є обмеженим та середньо-реверсивним. Зміна VIX за 7 днів (vix_change_7d) відображає швидкість зростання страху. Процентиль VIX за 252 дні (vix_percentile_252d) – інваріантна до масштабу міра відносного страху.

Фіксований дохід (2 ознаки). Спред 10Y–2Y (yield_curve_10y2y) є нахилом кривої дохідності. Інверсія (від'ємний спред) передувала кожній рецесії США з 1960-х років [3]. Зберігається як рівень ($I(1)$), оскільки знак несе пряме економічне значення; диференціювання знищило б поріговий сигнал інверсії.

Кредитний ризик (2 ознаки). Спред Moody's BAA над 10-річними казначейськими облігаціями (credit_spread_baa10y) є барометром кредитного ризику. Спреди вище 3,0 процентних пунктів історично збігалися з початком криз. 30-денна зміна (credit_spread_change_30d) відображає кредитний моментум – швидке розширення сигналізує про втечу від корпоративного ризику.

Міжбанківський стрес (1 ознака). Зшитий спред TED/SOFR–DTB3 (interbank_stress_spread) вимірює премію, яку банки стягують один з одного понад безризикові ставки. TED спред був каноничною мірою до відміни LIBOR; пост-2022 заміна (SOFR мінус 3-місячний T-bill) зберігає

економічну інтерпретацію. Зшивка уникає backfilling, що внесло б штучні режимні ефекти.

Оцінка вартості (3 ознаки). Коефіцієнт P/E Шиллера (`cape_ratio_level`) є найбільш відомим індикатором довгострокової оціненості ринку: значення вище 30 історично передували корекціям. Процентиль CAPE за 252 дні (`cape_ratio_percentile_252d`) нормалізує показник відносно нещодавньої історії. Премія за ризик акцій (`equity_risk_premium`) – різниця між `earnings yield` та безризиковою ставкою – від'ємні значення сигналізують, що акції дорожчі за облігації.

Лeverедж (3 ознаки). Маржинальний борг до ВВП (`margin_debt_gdp_ratio`) відображає рівень спекулятивного кредитного плеча на ринку. Зміна маржинального боргу за 6 місяців (`margin_debt_change_6m`) та його процентиль за 252 дні (`margin_debt_percentile_252d`) доповнюють рівневу ознаку динамікою та відносною позицією.

Крос-активна динаміка (3 ознаки). Зміна коефіцієнта золото/S&P 500 (`gold_sp500_ratio_change_30d`) є проху попиту на безпечні активи. Зміна USD Index (`usd_index_change_30d`) відображає динаміку долара – різке зміцнення часто супроводжує глобальні risk-off епізоди. Кореляція акцій та облігацій (`equity_bond_corr_30d`) кодує крос-активний ризиковий режим зі структурним зміщенням: переважно позитивна до 2000 р. (інфляція як спільний фактор), переважно від'ємна після 2000 р. (домінування flight-to-quality).

Монетарна політика (3 ознаки). Прискорення кривої дохідності (`yield_curve_acceleration`) – друга похідна спреду, що виявляє точки перегину. Естрелла та Мішкін (1998) показали, що динаміка кривої дохідності – а не лише рівень – передбачає рецесії [12]. Реальна ставка ФРС (`real_fed_funds_rate`) вимірює справжню вартість грошей: різкий рух у позитивну територію передував кризам (2007, 2000). Зміна ставки ФРС

за 180 днів (fed_funds_change_180d) відображає темп монетарного зтиснення.

Фінансовий сектор (1 ознака). Відносна 30-денна прибутковість XLF (financial_sector_relative_30d) – відставання фінансового сектора від ширшого ринку є класичним раннім попередженням системного стресу.

Ринок житла (3 ознаки). Іпотечна ставка (mortgage_rate_30y_level, FRED MORTGAGE30US), запас житла на ринку (new_home_months_supply, FRED MSACSR) та зміна цін на житло (home_price_yoy_change, FRED CSUSHPIISA). Лімер (2007) [13] стверджує, що житлове будівництво є найнадійнішим випереджальним індикатором бізнес-циклу.

Кредитний цикл (3 ознаки). Зростання банківського кредиту (bank_credit_growth_yoy) – Шуларік та Тейлор [10] ідентифікують зростання кредитування як найкращий окремий предиктор фінансових криз за 140 років. Зміна дозволів на будівництво (building_permits_yoy_change) – Лімер [13] стверджує, що «Житло ЯВЛЯЄ СОБОЮ бізнес-цикл», дозволи на будівництво випереджують ВВП на 3–5 кварталів. Зростання кредитів на нерухомість (re_loan_growth_yoy) – Жорда та співавт. [14] демонструють, що зростання іпотечного кредитування є найнебезпечнішою формою кредитної експансії.

Виробництво та споживчий сектор (4 ознаки). Завантаження виробничих потужностей (capacity_utilization, FRED TCU) – падіння нижче 75% історично сигналізує рецесію. Автомобільні продажі (vehicle_sales_yoy_change, FRED TOTALSA) відображають споживчу впевненість та доступність кредиту. Індекс Michigan Consumer Sentiment (consumer_sentiment_change_90d, FRED UMCSSENT) – різке падіння (<70) історично корелює з наближенням рецесії. Рівень прострочок по споживчих кредитах (consumer_loan_delinquency, FRED DRCCLACBS) – зростання прострочок сигналізує фінансовий стрес домогосподарств, що з лагом 2–4 квартали проявляється в макроекономічних показниках.

Інфляція (2 ознаки). Breakeven інфляційні очікування (breakeven_inflation_5y, FRED T5YIE) вимірюють ринкові очікування інфляції як різницю між номінальними та захищеними від інфляції казначейськими облігаціями. Зміна CPI за 12 місяців (cpi_yoy_change) доповнює breakeven фактичною динамікою інфляції. Ріке зростання інфляції може спровокувати aggressive tightening ФРС, що історично передувало фінансовому стресу.

Сировинні supply-shock (4 ознаки). 30-денна прибутковість нафти (oil_return_30d) – нафтові цінові шоки передували 9 з 10 повоєнних рецесій США. 30-денна волатильність нафти (oil_volatility_30d) сигналізує стрес на сировинних ринках. 60-денна прибутковість міді (copper_return_60d) – мідь відома як «Dr. Copper» за її здатність діагностувати економічне здоров'я. Зміна коефіцієнта мідь/золото (copper_gold_ratio_change_30d) ізолює ризикові настрої, нейтралізуючи спільні фактори (долар, загальна сировинна інфляція).

Міжнародні індикатори (2 ознаки). Процентиль JPY/USD за 252 дні (jpy_usd_percentile_252d) – proxy carry-trade reversal: різке зміцнення єни сигналізує глобальну розмотку carry-trade позицій, що історично супроводжувало кризи (Азійська криза 1997, GFC 2008). Зміна Nikkei 225 за 30 днів (nikkei_change_30d) – проху азійського ринку акцій, забезпечує міжнародне вікно на ризикові настрої.

Геополітичний ризик (13 ознак). 5 ознак індексу GPR Калдара-Яковієлло (рівень, тренд, волатильність, загрози, дії) вимірюють геополітичний ризик з новинних публікацій. 8 структурованих ознак GDELT (конфлікти, протести, дипломатія, економічна активність тощо) забезпечують гранулярне відображення глобальних подій.

Взаємодія ознак (1 ознака). Інтеракційна ознака $\text{real_fed_funds_rate} \times \text{interbank_stress_spread}$ (interact_ffr_interbank_stress_spread) є єдиною перехресною ознакою, валідованою через LOCO-аналіз. Вона кодує режимно-залежний сигнал міжбанківського стресу: при високих

процентних ставках міжбанківський стрес набуває іншого прогностичного значення, ніж при нульових ставках QE-ери.

Кореляційну матрицю 73 ознак з ієрархічною кластеризацією наведено в додатку В (рисунок В.1).

Таблиця 2.1 містить 21 репрезентативну ознаку з усіх категорій конвеєра (повна описова статистика 73 ознак наведена в додатку А, таблиця А.2).

Таблиця 2.1 – Описова статистика основних кількісних ознак

Ознака	Середнє	Std	Мін	Медіана	Макс
sp500_drawdown_30d	-0,021	0,030	-0,309	-0,011	0,000
sp500_volatility_30d	0,157	0,095	0,043	0,133	0,845
sp500_return_30d	0,008	0,044	-0,309	0,012	0,247
vix_level	19,24	7,68	9,14	17,46	79,13
vix_change_7d	-0,001	3,025	-19,24	-0,09	24,81
yield_curve_10y2y	0,960	0,882	-1,06	0,81	2,91
credit_spread_baa10y	2,264	0,691	1,18	2,13	6,10
interbank_stress_spread	0,485	0,449	-0,94	0,38	4,58
equity_bond_corr_30d	-0,033	0,438	-0,799	-0,05	0,847
cape_ratio_level	25,69	6,99	11,16	25,81	44,20
equity_risk_premium	-0,43	1,88	-4,57	-0,66	4,58
margin_debt_gdp_ratio	0,014	0,004	0,006	0,014	0,026
real_fed_funds_rate	0,41	2,37	-8,24	0,32	5,42
oil_return_30d	0,008	0,105	-0,580	0,006	0,963
copper_return_60d	0,009	0,124	-1,000	0,010	0,761
jpy_usd_percentile_252d	0,504	0,353	0,004	0,496	1,000
gpr_threats_level	78,49	58,06	0,00	66,64	393,50
mortgage_rate_30y_level	6,44	2,14	2,65	6,37	11,58
consumer_sentiment_chg	-0,24	6,55	-28,70	0,10	22,20
building_permits_yoy_chg	0,77	16,16	-55,03	2,92	59,94
unemployment_rate_chg	-0,006	0,559	-4,80	0,00	10,40

VIX має яскраво виражений правий хвіст (максимум 79,13 при медіані 17,46), що відображає рідкісність, але екстремальність кризових епізодів.

Drawdown ознаки завжди від'ємні або нульові за конструкцією. Зміна безробіття демонструє екстремальний викид у квітні 2020 під час глобальної пандемії (+10,4 п.п. – COVID lockdown), що є найбільшим місячним зростанням в історії США.

На рисунку 2.1 візуалізовано поведінку шести ключових ознак під час глобальної фінансової кризи 2008 року. Тоді CAPE та маржинальний борг демонстрували екстремальні значення ще задовго до обвалу Lehman Brothers, який відбувся у вересні 2008, тоді як VIX та міжбанкський стрес різко зросли лише після початку активної фази кризи. При цьому, крива дохідності інвертувалася ще у 2006 році.

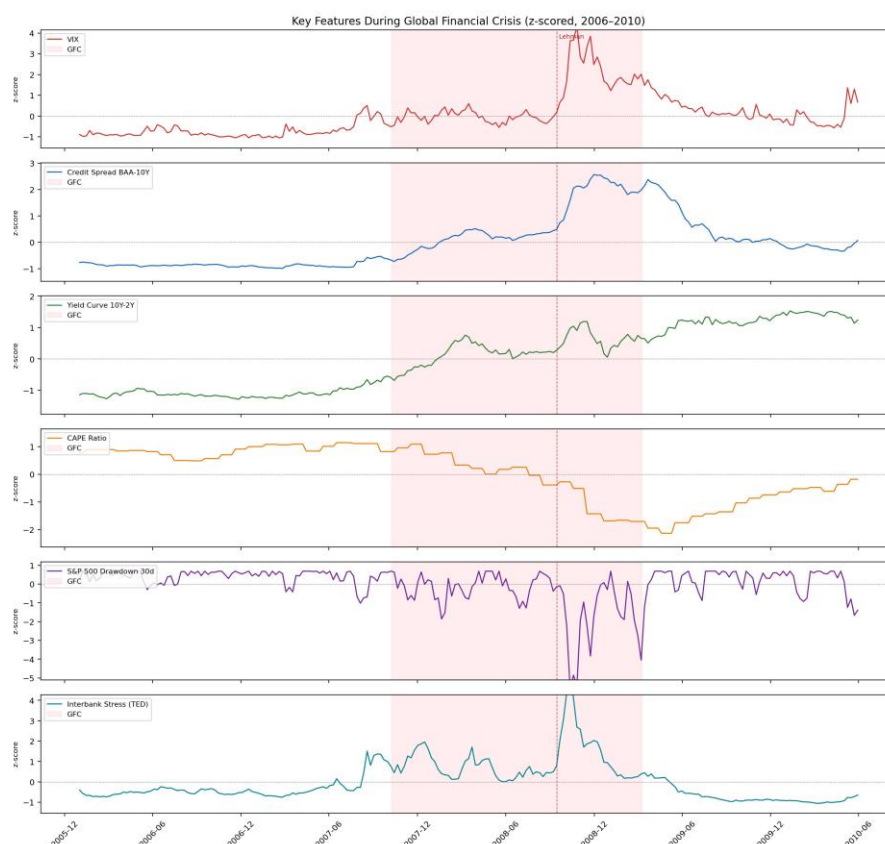


Рисунок 2.1 – Динаміка ключових ознак під час глобальної фінансової кризи (2006–2010)

Ці різні часові профілі – від випереджальних (оціночні) до запізнілих (волатильність) – обґрунтовують необхідність мультифакторного підходу.

2.2.3 Діагностика стаціонарності

Стаціонарність кожної з 73 кількісних ознак перевірена за допомогою двох комплементарних тестів: розширений Dickey-Fuller (ADF) з вибором лагу за AIC та тест KPSS з регресією на константу. Класифікація: $I(0)$ – ADF відхиляє ($p < 0,05$) і KPSS не відхиляє ($p > 0,05$); $I(1)$ – ADF не відхиляє і KPSS відхиляє. Результати тестування наведено в таблиці 2.2.

Таблиця 2.2 – Результати тестування стаціонарності ознак

Категорія	Кількість	Тест ADF	Тест KPSS	Нормалізація
$I(0)$ – стаціонарні	56	$p < 0,05$	$p > 0,05$	StandardScaler
$I(1)$ – нестаціонарні	17	$p \geq 0,05$	$p \leq 0,05$	Rolling z-score

З 73 ознак: 56 підтверджено стаціонарними $I(0)$, 17 класифіковано як нестаціонарні $I(1)$. Нестационарні ознаки включають: `yield_curve_10y2y`, `credit_spread_baa10y`, `interbank_stress_spread`, `equity_bond_corr_30d`, `vix_realized_vol_ratio`, `real_fed_funds_rate`, `nfcf_leverage_level`, `re_loan_growth_yoy`, `mortgage_rate_30y_level`, `new_home_months_supply`, `consumer_loan_delinquency`, `capacity_utilization`, `breakeven_inflation_5y`, `gpr_threats_level`, `gpr_acts_level`, `cape_ratio_level` та `interact_ffr_interbank_stress_spread`. Повні результати тестування наведено в додатку А (таблиця А.1).

Помітний результат: VIX level є стаціонарним. Тест ADF сильно відхиляє гіпотезу одиничного кореня (статистика = $-5,023$, $p = 0,00002$), а KPSS не відхиляє стаціонарність (статистика = $0,186$, $p > 0,10$). Це узгоджується з теоретичними властивостями VIX: він

обмежений (практичний діапазон [9; 80]), середньо-реверсивний та представляє імпліковану волатильність на фіксованому 30-денному горизонті.

Усі 17 нестационарних ознак зберігаються як рівні без диференціювання, оскільки їх абсолютні значення кодують кризово-релевантні порогові значення: `yield_curve_10y2y` – знак несе пряме економічне значення інверсії; `real_fed_funds_rate` – знак і магнітуда кодують стан монетарної політики; `capex_ratio_level` – абсолютне значення визначає оціночний режим ринку; `credit_spread_baa10y` – рівень вище 3 п.п. історично збігається з початком криз; `mortgage_rate_30y_level` – абсолютна ставка впливає на доступність кредиту; `capacity_utilization` – падіння нижче 75% сигналізує рецесію. Для коректної нормалізації ці ознаки обробляються `rolling z-score` замість `StandardScaler`, що використовується для стаціонарних $I(0)$ ознак. Контроль якості даних наведений у таблиці 2.3.

Таблиця 2.3 – Контроль якості даних

Критерій	Вимога	Результат	Статус
Prevalence криз	5–20%	15,0% (282/1878)	PASS
Ознаки < 10% NaN	≥ 50 з 73	58 з 73	PASS
Пари $ r > 0,95$	0 пар	1 пара	WARN
Стаціонарність	Усі тестовані	73/73	PASS
$I(1)$ обґрунтовані	Усі з ратіонале	17/17	PASS

Діаграму пропущених значень (NaN) по ознаках конвеєра наведено в додатку Г (рисунок Г.1).

2.2.4 Побудова вікон та розподіл міток

Вікна ознак побудовані як `rolling 7-денні знімки` за повним діапазоном дат. Загалом 2138 вікон за період 1986–2026. Кожне вікно отримує вектор

ознак, обчислений для його кінцевої дати. Розподіл вікон за період 1986–2026 наведено в таблиці 2.4.

Таблиця 2.4 – Кількість вікон по кожній мітці

Мітка	Кількість	Відсоток
Crisis	262	12,3%
Precursor	51	2,4%
Stress	176	8,2%
Normal	1 649	77,1%
Позитивний клас (Crisis+Precursor)	313	14,6%

Для регресії просідання цільова змінна – максимальне forward просідання від поточної ціни за N днів на горизонтах 30, 90, 180, 365 днів. Це є нетривіальним навіть під час активних криз: ринок, який вже впав на 30%, може мати лише 2% подальшого падіння (поблизу дна) або 20% (далі до дна).

2.3 Текстові дані FOMC та Beige Book

Текстова модальність включає 12 ознак оцінених LLM з двох випереджальних джерел: заяви та протоколи FOMC (501 документ, 1993–2026) та звіти Beige Book. GPT-4o mini оцінює кожен документ за 6 стресовими вимірами, генеруючи структуровані числові оцінки, які потім використовуються в якості вхідних ознак для моделей системи.

Ключове розмежування – між типами текстових джерел. Газетні заголовки (перевірено на 632 406 реальних статтях NYT із трьома методами кодування) є запізнілими індикаторами та не додають сигналу до кількісних ознак на тижневій частоті. Натомість документи FOMC та Beige Book містять випереджальну інформацію: внутрішні оцінки ФРС, forward

guidance та регіональні економічні спостереження, які відсутні в ринкових цінах.

Текстові ознаки забезпечують статистично значуще покращення для Рідж-регресії: $+0,055 \Delta p$ на горизонті 90 днів для FOMC, $+0,034$ для Beige Book. Комбінація обох джерел (50 ознак: 38 кількісних + 6 FOMC + 6 BB) послідовно перевершує кожне джерело окремо.

Проте текстові ознаки систематично погіршують дистанційні моделі (K-NN, MLP), оскільки низькодисперсійні текстові виміри розбавляють простір подібності.

2.4 Механізм запобігання витоку даних

Анти-витоків архітектура є фундаментальною вимогою для валідності результатів. Кожен компонент системи забезпечує сувору дисципліну:

- expanding-window walk-forward: модель навчається лише на даних до поточного fold-y;
- нормалізація лише на тренувальних даних: z-score та StandardScaler обчислюються виключно на тренувальній вибірці кожного fold-y;
- перенавчання моделі на кожному fold-i: гіперпараметри фіксовані, але ваги оновлюються;
- корекція лагів публікації: ознаки використовуються з реалістичними затримками;
- деконтаміновані промпти LLM: анонімізовані індекси періодів, видалені мітки криз.

Верифікація лагів публікації є критичним аспектом антивитокової архітектури. Для безробіття (UNRATE) проаналізовано 431 vintage-дату з 1990 по 2025 рік. Лаг публікації демонструє надзвичайну стабільність: середній = 34,4 дні, медіанний = 34,0 дні, P95 = 37 днів, P99 = 40 днів. Обрано зміщення P99 (40 днів): при цьому порозі лише 4 з 431 місяців (0,9%)

мають залишковий витік, всі через аномалії урядових shutdown. Стабільність між десятиліттями: 1990-ті – середній 34,4 дні; 2000-ні – 34,2; 2010-ті – 34,4; 2020-ті – 34,7.

Аналогічні корекції лагів застосовуються до інших макроекономічних індикаторів: ВВП – 121 день (BEA публікує advance estimate через ~30 днів після закінчення кварталу, третій перегляд – через 90 днів); маржинальний борг – 164 дні (FINRA публікує з 2+ місячним лагом); CAPE – 45 днів (залежить від публікації earnings даних S&P); CPI – 45 днів (BLS публікує ~15-го числа наступного місяця). Усі лаги емпірично виміряні з офіційних календарів релізів BEA, Fed та BLS.

Темпоральна цілісність забезпечується програмно. Клас FeatureEngine маршрутизує кожен доступ до даних через метод `_get()`, що усікає часовий ряд до цільової дати та логує доступ. Після обчислення ознак `verify_no_leakage(target_date)` перевіряє, що кожен залогований доступ закінчується на або до цільової дати. Це надає програмну гарантію проти витоку.

Зшивка міжбанківського стресу: TED спред (TEDRATE) був припинений 21 січня 2022 з переходом від LIBOR. Замість backfilling SOFR у «доLIBOR-ний» період (що внесло б штучні режимні ефекти), серію чисто зшито в точці дисконтинуації: TEDRATE для дат до 2022-01-21, SOFR – DTB3 для дат на або після. Дублюючі індекси в точці зшивки видалені.

Для текстових даних FOMC антивитокова архітектура включає: (1) корекцію лагу публікації протоколів засідань (~3 тижні після дати засідання); (2) деконтаміновані промпти LLM – анонімізовані індекси періодів замість календарних дат, видалені мітки криз та наративи з каталогу; (3) post-hoc тест контамінації на трьох умовах (нормальні, анонімізовані, перемішані дані).

Процес екстракції текстових ознак з документів FOMC включає наступні кроки. GPT-4o mini (температура = 0 для відтворюваності результатів) отримує текст документа та оцінює його за 6 стресовими

вимірами: (1) загальний рівень економічного стресу; (2) ризик фінансової нестабільності; (3) інфляційний тиск; (4) напруженість ринку праці; (5) ризик уповільнення зростання; (6) невизначеність монетарної політики. Кожен вимір оцінюється на шкалі від 0 (відсутність стресу) до 10 (екстремальний стрес).

Вибір LLM 6-d кодування замість сирих 384-d ембедінгів мотивований аналізом lookup table ефекту. З лише 501 унікальним документом FOMC у 384-вимірному просторі Ridge може запам'ятовувати per-meeting-period результати. PCA dimensionality аблація показує: PCA-6d $\rightarrow \Delta\rho \approx +0,03$; PCA-20d $\rightarrow \Delta\rho \approx +0,06$; PCA-100d $\rightarrow \Delta\rho \approx +0,15$; raw 384-d $\rightarrow \Delta\rho = +0,279$. LLM 6-d є компромісом: достатня розмірність для передачі сигналу, занадто мала для запам'ятовування.

Beige Book містить регіональні економічні звіти, зібрані 12 федеральними резервними банками. На відміну від FOMC (top-down policy deliberation), Beige Book надає bottom-up спостереження: стан виробництва, роздрібні продажі, ринок праці, нерухомість по кожному регіону. Ці два джерела є комплементарними: FOMC відображає колективну оцінку комітету, Beige Book показує реальний стан на місцях.

Мультиджерельна інтеграція тексту є модельно-залежною. Ridge (L2-регуляризація): текстові ознаки покращують прогнозування (FOMC: $+0,055 \Delta\rho$ 90d; BB: $+0,034$; FOMC+BB: $+0,055$ 90d, послідовно перевершує кожне джерело на всіх горизонтах). Дистанційні моделі (K-NN, MLP): текстові ознаки систематично погіршують якість ($\Delta = -0,007$ до $-0,083$). Причина: 12 низькодисперсійних текстових вимірів розбавляють евклідову відстань у просторі подібності. Рекомендована конфігурація: Рекомендована конфігурація: 85f (73 quant + 6 FOMC + 6 BB + 20 tPCA) для Ridge, 73f (quant-only) для дистанційних моделей.

3 РОЗРОБКА СИСТЕМИ ПРОГНОЗУВАННЯ ФІНАНСОВИХ КРИЗ

3.1 Рідж-регресія з експоненційним зважуванням

Основною моделлю системи є Рідж-регресія (Ridge Decay 10y + Temporal PCA-20) – L2-регуляризована лінійна регресія на 85 ознаках (73 кількісні + 12 текстових) та 20 PCA-стиснутих компонентах темпоральних лагів (lookback 4 та 12 тижнів).

Ключові інноваційні компоненти моделі:

- експоненційне зважування зразків з періодом напіврозпаду 10 років: адаптація до зміни режимів монетарної політики, де зв'язки ознака-ціль інвертувалися між ерою до QE (1996–2015) та після QE (2016–2026). Пермутаційний тест підтверджує статистичну значущість нестабільності ер ($p < 0,001$ на горизонтах 90/180 днів);

- рангова трансформація цілей: замість сирих значень drawdown модель навчається на процентильних рангах $|\text{drawdown}|$, що безпосередньо оптимізує кореляцію Спірмена ρ . Прогнози відображаються назад у шкалу drawdown через квантильну функцію тренувального розподілу;

- один LOCO-валідований терм взаємодії: $\text{real_fed_funds_rate} \times \text{interbank_stress_spread}$, що захоплює режимно-залежний сигнал міжбанківського стресу – пригнічений за умов QE, прогностичний при підвищених ставках;

- Temporal PCA-20: 20 головних компонент, обчислених з 4-тижневих та 12-тижневих лагів ознак, що компресують темпоральну структуру та зменшують мультиколінеарність.

Блоковий бутстреп підтверджує статистичну значущість покращення від decay-зважування на кожному горизонті (середній $\Delta\rho = +0,120$, всі $p < 0,01$ з корекцією Holm-Bonferroni).

Математично, Рідж-регресія мінімізує цільову функцію з L2-регуляризацією: $\min_{\beta} \sum w_i (y_i - x_i^T \beta)^2 + \lambda \|\beta\|^2$, де w_i – ваги зразків

з експоненційним спаданням з основою 2: $w_i = 2^{-(age_i / h)}$, де age_i – вік зразка у роках, $h = 10$ років – період напіврозпаду. Таким чином, зразок віком 10 років отримує вагу 0,5, віком 20 років – 0,25 тощо. Параметр регуляризації λ обирається за 5-fold часовою крос-валідацією на тренувальних даних кожного fold-у.

Рангова трансформація цілей працює наступним чином. Замість сирих значень $\max$ forward drawdown y_i , модель навчається на їх процентильних рангах $r_i = \text{rank}(|y_i|) / N$. Це безпосередньо оптимізує кореляцію Спірмена ρ , яка є інваріантною до монотонних трансформацій. Прогнози відображаються назад у шкалу drawdown через квантильну функцію тренувального розподілу: $\hat{y} = \text{quantile}(\text{drawdown}_{\text{train}}, \hat{r})$.

Це відображення генерує дискретні кроки прогнозів – лише 186 унікальних значень серед 1614 fold-ів (11,5% унікальних), видимих як «плато» на графіках часових рядів. Стеля $\sim 33,5\%$ (fold-и до 2008) відповідає максимальному тренувальному drawdown Чорного понеділка 1987; після 2008 стеля зростає до 52,6% (пік GFC). Ця дискретизація є невід'ємною рисою дизайну рангової трансформації, що забезпечує $\rho = 0,87$ – безперервна регресійна функція втрат (MSE) генерувала б гладші прогнози, але з $\rho \approx 0,60$.

Temporal PCA-20 компресує темпоральну структуру наступним чином. Для кожного fold-у обчислюються 4-тижневі та 12-тижневі лаги всіх 73 кількісних ознак, генеруючи 146 додаткових ознак. PCA, навчена лише на тренувальних даних, зменшує ці 146 ознак до 20 головних компонент, що зберігають $\sim 85\%$ дисперсії. Це зменшує мультиколінеарність між лагами та забезпечує компактне представлення темпоральних динамік.

Терм взаємодії $\text{real_fed_funds_rate} \times \text{interbank_stress_spread}$ валідований методом LOCO (Leave-One-Column-Out): видалення цього терму погіршує ρ на 0,018 (90d) та $-0,023$ (180d), що перевищує поріг значущості $-0,01$. Фізична інтерпретація: міжбанківський стрес є

прогностичним лише за умов підвищених процентних ставок; за умов QE (від'ємні реальні ставки) міжбанківський спред не несе кризового сигналу, оскільки центральний банк забезпечує необмежену ліквідність.

Блоковий бутстреп з 10 000 ресемплувань (10 епізодних блоків) підтверджує, що покращення від decay-зважування є статистично значущим: середній $\Delta p = +0,120$, всі $p < 0,01$ з корекцією Holm-Bonferroni по 4 горизонтах. Чистий ефект decay (контролюючи за Temporal PCA) складає $+0,093$, також $p < 0,01$. Аналіз чутливості (5 альтернативних блокових структур, до 20 000 ресемплувань) показує, що жоден бутстреп-зразок ніколи не генерував від'ємну дельту по жодній структурі чи горизонту.

3.2 Пошук K найближчих сусідів (Retrieval DD)

Друга модель ансамблю – K-NN retrieval з оптимізованими параметрами. Сітковий пошук (18 варіантів: $K \in \{10, 20, 50\} \times 3$ схеми зважування $\times 2$ методи агрегації) виявив, що зменшення K з 50 до 10 та додавання зважування за давністю (5-річний період напіврозпаду) покращує holdout p з 0,158 до 0,632.

Менша кількість сусідів запобігає regression-to-mean під час криз, а зважування за давністю надає перевагу аналогам з поточного монетарного режиму. Модель використовує лише 73 кількісних ознаки (без текстових), оскільки низькодисперсійні текстові виміри розбавляють простір подібності.

Оптимізація параметрів K-NN проведена сітковим пошуком з 18 варіантів: $K \in \{10, 20, 50\} \times 3$ схеми зважування (uniform, distance-weighted, recency-weighted з 5-річним періодом напіврозпаду) $\times 2$ методи агрегації (mean, median). Оптимальна конфігурація: $K = 10$, recency-weighted mean.

Механізм рекурентного зважування: $\text{weight}_j = \text{similarity}_j \times 2^{-(\text{age}_j / h_{\text{recency}})}$, де similarity_j – косинусна подібність між запитом та сусідом, age_j – вік сусіда у роках, $h_{\text{recency}} = 5$ років – період напіврозпаду. Множник подібності забезпечує, що ближчі сусіди у просторі ознак мають вищу вагу, а множник давності надає перевагу аналогам з поточного монетарного режиму: сусіди з пост-QE ери (2016+) отримують вищу вагу при прогнозуванні у 2024-2026, ніж структурно подібні, але давніші періоди (наприклад, 1998).

Зменшення K з 50 до 10 запобігає regression-to-mean під час криз: з $K=50$ у вікні кризи більшість з 50 сусідів належать до нормальних періодів, «розбавляючи» кризовий сигнал. З $K=10$ модель зосереджується на найближчих аналогах, які з більшою ймовірністю також є кризовими.

Модель використовує лише 73 кількісних ознаки. Текстові ознаки (FOMC, BB) систематично погіршують K-NN: $\Delta\rho = -0,007$ до $-0,083$ по горизонтах. Причина – 12 низькодисперсійних текстових вимірів розбавляють простір подібності, на якому базується пошук сусідів. L2-регуляризація Рідж-регресії природно пригнічує неінформативні виміри, тому вона єдина ефективно використовує текстові ознаки.

3.3 MLP з триплетним навчанням

Третя модель – багат шаровий перцептрон ($73 \rightarrow 128 \rightarrow 64$, ~12K параметрів) з drawdown-aware триплетним навчанням. Вікна з подібними forward drawdown зближуються у просторі ембедінгів, вікна з різною серйозністю – розштовхуються.

K-NN скоринг на ембедінгах MLP використовує ті ж оптимізовані параметри ($K=10$, рекурентне зважування), що покращило MLP ρ на +0,16 до +0,21 по горизонтах.

Архітектура MLP: $73 \rightarrow 128 \rightarrow 64$, де 73 – вхідні кількісні ознаки, 128 – прихований шар з ReLU активацією, 64 – вихідний шар

ембедінгів. Загальна кількість параметрів $\sim 12\,000$. Drawdown-aware триpletне навчання формує триплети (anchor, positive, negative), де positive – вікно з подібним forward drawdown ($|\Delta \text{drawdown}| < 5\%$), negative – вікно з суттєво іншою серйозністю ($|\Delta \text{drawdown}| > 15\%$).

Функція втрат: $L = \max(0, d(\text{anchor}, \text{positive}) - d(\text{anchor}, \text{negative}) + \text{margin})$, де d – євклідова відстань у 64-вимірному просторі ембедінгів, $\text{margin} = 0,3$. Навчання проводиться з Adam optimizer ($\text{lr} = 0,001$), $\text{batch size} = 256$, протягом 50 епох з early stopping за валідаційною втратою.

MLP ембедінги забезпечують якісніший простір подібності порівняно з сирими ознаками: вікна з подібними forward drawdown кластеризуються разом, вікна з різною серйозністю – розділяються. K-NN скоринг на MLP ембедінгах з тими ж оптимізованими параметрами ($K=10$, рекурентне зважування) покращив MLP ρ на $+0,16$ до $+0,21$ по горизонтах порівняно з попередньою конфігурацією ($K=50$, uniform). Це зробило MLP конкурентоспроможним з Ridge та сирим K-NN.

3.4 Ансамблева архітектура

Production система використовує дуальний вихід: медіану всіх 3 моделей для оцінки магнітуди просідання (кращий Pearson r , нижчий MAE) та Рідж-регресію для класифікації режиму (найкращий holdout Spearman ρ).

Турнір ансамблів (10 стратегій $\times 9$ метрик $\times 4$ горизонти) підтвердив, що медіанний ансамбль (Ridge + K-NN + MLP) досягає walk-forward $\rho = 0,750/0,842/0,894/0,937$ (тижневий walk-forward), перевершуючи Ridge-only на кожному горизонті, тоді як 3-модельна медіана генерує найреалістичніші магнітуди.

3.5 Інтеграція LLM

Інтеграція LLM функціонує на двох рівнях:

– екстракція текстових ознак – GPT-4o mini оцінює документи FOMC та Beige Book за 6 стресовими вимірами, генеруючи випереджальні ознаки з приватною інформацією, недоступною в ринкових цінах (+0,025 середній Δp);

– пояснення на дашборді – GPT-4o mini генерує структуровану інтерпретацію поточних прогнозів (тип кризи, фактори ризику, рекомендації з моніторингу) для Production дашборду.

Тест контамінації на всіх 472 fold-ax walk-forward валідації підтвердив, що LLM як міркувальний агент (reasoning layer) не додає кількісного покращення $\rho \approx 0$ за всіх умов (нормальні, анонімізовані, перемішані дані). LLM демонструє систематичний песимістичний зсув з прогнозами, закріпленими на 18–45% drawdown (у 3–5 разів вище реальних базових ставок).

Production система має двоступеневу архітектуру. Перший ступінь – генерація прогнозів: (1) Ridge Decay 10y + Temporal PCA-20 на 85 ознаках (73 quant + 12 text + 20 PCA); (2) K-NN Retrieval DD на 73 кількісних ознаках; (3) MLP Retrieval DD на 64-d ембедінгах. Другий ступінь – агрегація: медіана трьох моделей для оцінки магнітуди просідання; Ridge score для класифікації режиму (crisis / elevated / normal).

Турнір ансамблів протестував 10 стратегій агрегації по 9 метриках на 4 горизонтах (360 комбінацій) з тижневими fold-ами. Результат: медіанний ансамбль (Ridge + K-NN + MLP) займає ранг #2 (композитний бал 3,08), inverse_mae_weighted – #1 (2,72), тоді як Ridge solo – останнє місце #10 (бал 9,00). З тижневим перенавчанням ансамбль перевершує Ridge-only на кожному горизонті та метриці.

Дашборд моніторингу в реальному часі відображає: (1) поточний прогноз просідання на 4 горизонтах; (2) процентильний ранг відносно історичного розподілу; (3) топ-5 найближчих історичних аналогів з каталогу криз; (4) LLM-генероване пояснення (тип кризи, ключові фактори ризику,

рекомендації з моніторингу); (5) таймлайн ранніх попереджень: як змінювався прогноз моделі протягом останніх 6 місяців.

LLM пояснювальний шар працює наступним чином. GPT-4o mini отримує: (a) поточний вектор ознак (z-scored); (b) прогнози 3 моделей; (c) топ-5 K-NN аналогів з описами. Генерує структурований JSON: `crisis_type` (одна з 7 категорій), `confidence` (0–1), `key_risk_factors` (ранжований список ознак), `monitoring_recommendations` (конкретні порогові значення для моніторингу). Час генерації: ~2 секунди на fold. Якісна цінність: аналітик отримує не лише числовий прогноз, а й інтерпретацію «це нагадує dot-com бульбашку через екстремальні оціночні показники».

Вибір медіани замість середнього для агрегації ансамблю мотивований робастністю до викидів. Коли одна модель генерує аномально високий або низький прогноз (наприклад, K-NN знаходить «хибного сусіда» через шум), медіана ігнорує цей викид, тоді як середнє зміщується. На практиці 3-model median генерує прогнози з нижчим MAE та кращим Pearson r, ніж середнє.

Dual-output дизайн: для оцінки магнітуди використовується медіана (реалістичніші абсолютні значення), для класифікації режиму – Ridge score (найкращий holdout Spearman ρ). Порогові значення для класифікації: Ridge score > 75th percentile $\rightarrow$ crisis; > 50th percentile $\rightarrow$ elevated; інше $\rightarrow$ normal. Ці пороги визначаються на тренувальному розподілі кожного fold-y.

Порівняння ключових стратегій ансамблів (тижневий walk-forward, 1580 fold-ів). Медіанний ансамбль (Ridge + K-NN + MLP): walk-forward $\rho = 0,750/0,842/0,894/0,937$ – перевершує Ridge-only на кожному горизонті. Ridge solo: $\rho = 0,730/0,803/0,840/0,875$ – останнє місце в турнірі (#10/10), програє кожній ансамблевій стратегії. На holdout 2023–2026 (171 тижневий fold): ансамбль $\rho = 0,672/0,720/0,806/0,797$ vs Ridge-only $0,609/0,668/0,760/0,803$ – перевага ансамблю на коротких

горизонтах (+0,063 30d, +0,052 90d), нічия на 365d. Проте на frozen holdout (без перенавчання) ситуація обернена: Ridge зберігає якість ($\rho = 0,609\text{--}0,803$), тоді як K-NN та MLP мають нульову якість і «перетягують» ансамбль нижче Ridge-only.

Тест контамінації LLM. Три умови тестування на всіх 472 fold-ax walk-forward валідації: (1) normal – стандартний промпт з анонімізованими індексами замість дат; (2) anonymized – додатково видалені назви ознак; (3) shuffled – перемішані значення ознак. Результат: Spearman $\rho \approx 0$ за всіх умов ($\text{delta} < 0,03$ між умовами). Висновок: LLM не демонструє контамінації знаннями з тренувальних даних.

Пертурбаційний тест додатково підтверджує: заміна кризових ознак на значення спокійних періодів не змінює вихід LLM – модель генерує приблизно однакові прогнози незалежно від вхідних даних. Аналіз калібрації виявляє механізм: систематичний песимістичний зсув (прогнози закріплені на 18–45% drawdown, у 3–5 разів вище реальних базових ставок 4–14%) з майже нульовою дискримінативною дисперсією. LLM фактично є deterministic pass-through: не здатний додати кількісну цінність через статистичне міркування на z-scored кількісних ознаках, він повертає найбільш prominentne числове значення зі входу.

На рисунку 3.1 представлено загальну архітектуру системи. Дані надходять з трьох джерел: Yahoo Finance (ринкові дані), FRED (макроекономічні індикатори) та federalreserve.gov (тексти FOMC та Beige Book).

Конвеєр ознак генерує 85 ознак (73 кількісні + 12 текстових LLM-скорів), доповнених 20 Temporal PCA компонентами. Три моделі (Ridge, K-NN, MLP) генерують незалежні прогнози, які агрегуються медіаною для оцінки магнітуди та Ridge score для класифікації режиму. LLM генерує інтерпретовані пояснення на основі прогнозів та найближчих історичних аналогів.

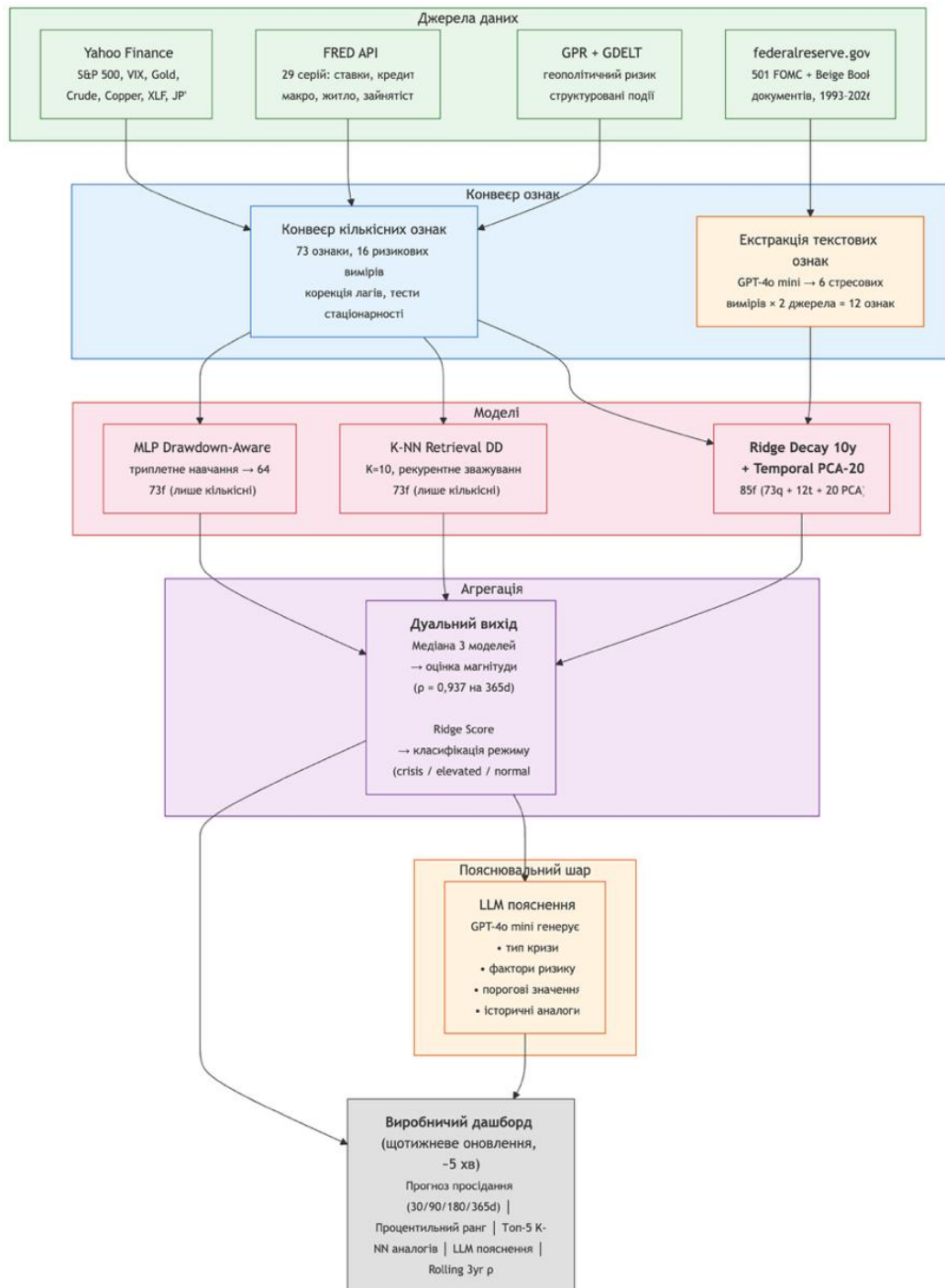


Рисунок 3.1 – Архітектура системи

Незважаючи на нульове кількісне покращення, LLM додає операційну цінність через:

– структуровану класифікацію типу кризи – визначає, чи поточна ситуація нагадує banking crisis, equity crash, або liquidity crisis;

- ідентифікацію ключових факторів ризику – ранжує ознаки за їх внеском у поточний прогноз;
- рекомендації з моніторингу – конкретні порогові значення для відстеження;
- контекстуалізацію через історичні аналоги – описує, як подібні ситуації розвивалися в минулому.

Ця інтерпретованість є ключовою для прийняття рішень як аналітиками, так і інвесторами, які потребують не лише числового прогнозу, а й розуміння його обґрунтування.

На рисунку 3.2 деталізовано пайплайн Ridge моделі, яка є ключовою у створеній системі прогнозування.

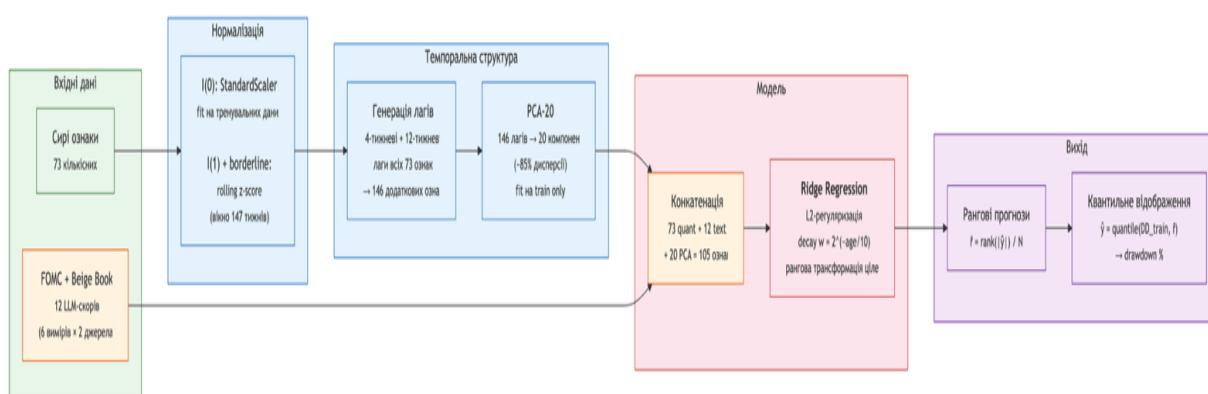


Рисунок 3.2 – Пайплайн Ridge моделі

В таблиці 3.1 наведено порівняння архітектур всіх трьох моделей, які застосовано в системі для побудови ансамблю для прогнозування.

Таблиця 3.1 – Порівняння архітектур трьох основних моделей

Модель	Ознаки	Параметри	Навчання	Переваги
Ridge Decay	85 + 20 PCA	~105	L2 + decay	Генералізація
K-NN Retrieval	73 quant	K=10	Без навчання	Інтерпретація
MLP Triplet	73 → 64-d	~12K	Триплети	Ембедінги

Систематичний тест 7 кадансів перенавчання (ніколи, щорічно, піврічно, щоквартально, щомісячно, кожні 2 тижні, щотижнево) на періоді 2023–2026 виявив, що щотижнєве перенавчання є оптимальним для ансамблю. K-NN та MLP мають від'ємний середній ρ при будь-якій частоті рідше місячної. Стрибок від квартального до місячного є вирішальним (K-NN: $-0,083 \rightarrow +0,177$), але покращення продовжуються за межами місячного: щотижнєве перенавчання приблизно подвоює якість K-NN/MLP порівняно з місячним. Ridge зберігає позитивну якість при всіх частотах (0,22–0,57), підтверджуючи режимно-інваріантні ознаки.

4 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА РЕЗУЛЬТАТИ

4.1 Протокол walk-forward валідації

Валідація здійснюється за протоколом walk-forward з розширюваним вікном: 1618 тижневих fold-ів (W-MON) за період 1996–2026 років. На кожному fold-і модель навчається виключно на даних до поточної дати, прогнозує на наступний тиждень, а результати акумулюються для обчислення загальних метрик.

Базові моделі (baselines) включають: VIX single-feature (один індикатор), persistence ($|y_{\text{true}}(t-1)|$, потребує 11 місяців майбутніх цін), lagged prediction (власний прогноз моделі з попереднього fold-у, доступний у реальному часі), та Ridge без decay-зважування.

4.2 Результати прогнозування падіння

Основні результати walk-forward регресії просідання.

Рідж-регресія (Ridge Decay 10y + Temporal PCA-20, 85 ознак): кореляція Спірмена $\rho = 0,730 / 0,803 / 0,840 / 0,875$ на горизонтах 30/90/180/365 (тижневий walk-forward) днів відповідно.

Реальна додана вартість (genuine value-add) над практичною персистенцією (lagged prediction): $\Delta\rho = +0,292 / +0,216 / +0,146 / +0,078$ на відповідних горизонтах.

K-NN Retrieval DD (73 ознаки, K=10, рекурентне зважування): $\rho = 0,714\text{--}0,921$ (walk-forward).

MLP Retrieval DD (73 ознаки, триплетне навчання): $\rho = 0,717 / 0,818 / 0,878 / 0,926$.

Медіанний ансамбль (3 моделі): walk-forward $\rho = 0,750/0,842/0,894/0,937$ (тижневий), перевершуючи Ridge-only на

кожному горизонті. Турнір ансамблів з тижневими fold-ами ставить Ridge solo на останнє місце (#10/10), тоді як медіанний ансамбль – на #2.

Важливе застереження: значення ρ на довгих горизонтах (180, 365 днів) завищені через автокореляцію цільової змінної – суміжні тижневі fold-и мають значне перетину вікон forward drawdown. Oracle persistence досягає $\rho = 0,930$ на 365 днів.

Покризовий аналіз підтверджує, що модель надає від 6 до 18 місяців поступово зростаючих ранніх попереджень: GFC – 84-й перцентиль до червня 2007 (16 місяців до піку); COVID – 75-й перцентиль до січня 2020; dot-com – 56-й до березня 2000, зростання до 90-го до вересня.

Покризовий аналіз демонструє здатність системи генерувати ранні попередження задовго до реалізації криз. Глобальна фінансова криза 2008: модель досягає 84-го перцентиля прогнозованого drawdown до червня 2007 року, за 16 місяців до піку кризи (березень 2009). Шість ознак перебували на історичних екстремумах: CAPE на 99,3-му перцентилі, премія за ризик акцій на 0,7-му перцентилі ($-4,1\%$), маржинальний борг/ВВП на 97,7-му перцентилі.

Пандемія COVID-19 (2020): модель досягає 75-го перцентиля до січня 2020, за 2 місяці до обвалу ринку. Dot-com бульбашка (2000): модель досягає 56-го перцентиля до березня 2000, зростаючи до 90-го перцентиля до вересня. До березня 2000 модель прогнозувала 33,5% drawdown (її максимально можливий вихід, обмежений стелею rank-to-magnitude).

У спокійних періодах модель коректно прогнозує низькі drawdown: медіанний прогноз 7,7% коли фактичний DD < 10%; у стресових періодах – 27,5% коли фактичний DD > 20%. Ridge прогнози добре відкалібровані: загальний bias = $-1,0\%$ (незначне недооцінювання ризику), MAE = 3,7%.

Модель генералізується на нові типи криз через екстремальність простору ознак, а не через зіставлення з історичними патернами. Dot-com бульбашка була першим оціночним крахом у вікні оцінювання, без

прецеденту в тренувальному наборі 1986–1999. Модель реагує на лінійну комбінацію екстремальних значень ознак незалежно від того, чи відбувався конкретний архетип кризи раніше.

Декларація ступенів свободи дослідника: повний простір дослідження включав 50+ варіантів моделей (Ridge, XGBoost, CatBoost, TabPFN, GP, MLP регресія, LSTM, knowledge distillation, K-NN), 7 наборів ознак, 7 значень half-life, 4 методи кодування тексту, 5 текстових джерел, рангову vs. сиру трансформацію цілей та множинні стратегії ансамблів. Корекція Holm-Bonferroni застосовується в межах індивідуальних сімей порівнянь, але не охоплює ширший garden of forking paths. Для кількісної оцінки ризику переобфитування вибору стратегії ансамблю проведено аналіз РВО (розділ 4.4), а для фіксації production конфігурації – попередню реєстрацію специфікації моделі (production_spec_2026_04_21.json).

Регресія drawdown надає справжній прогностичний доказ, оскільки ціль – forward max drawdown від поточної ціни – є нетривіальною навіть під час активних криз. Lead-aligned графіки drawdown демонструють, що лінії прогнозу моделі досягають піку перед реалізованими drawdown ринку, з lead-часами, які не можуть бути пояснені детекцією. Основні результати walk-forward регресії падіння наведено в таблиці 4.1.

Таблиця 4.1 – Результати walk-forward регресії падіння

Модель	30d ρ	90d ρ	180d ρ	365d ρ
Ridge Decay 10y (85f+PCA)	0,730	0,803	0,840	0,875
K-NN Retrieval DD (73f)	0,714	0,822	0,862	0,921
MLP Retrieval DD (73f)	0,717	0,818	0,878	0,926
Медіанний ансамбль (3 моделі)	0,750	0,842	0,894	0,937
3-model median	0,695	0,790	0,845	0,895
VIX single-feature	0,428	0,409	0,381	0,336
Oracle persistence	0,586	0,718	0,870	0,930

На рисунку 4.1 візуалізовано walk-forward прогнози на 365-денному горизонті для медіанного ансамблю моделей. Лінія прогнозу піднімається перед зростанням реалізованого drawdown для GFC, dot-com та COVID. Також спостерігаються дискретні «плато» в прогнозах, які є артефактом rank-to-magnitude mapping. При цьому, стежа прогнозу зростає після 2008, коли GFC drawdown (52,6%) входить до тренувального розподілу. Однак, у спокійних періодах прогнози коректно залишаються на відносно низьких рівнях.



Рисунок 4.1 – Walk-forward прогнози медіанного ансамблю (жовта лінія) vs реалізовані past max drawdown (червона лінія), горизонт 365 днів, 1996-2026

В додатку Д представлено внесок кожної ознаки до якості Ridge прогнозів (виміряний як Δr при видаленні ознаки). Топ-5 ознак: CAPE ($\Delta r = -0,045$), маржинальний борг/ВВП ($-0,038$), VIX ($-0,032$), кредитний спред ($-0,028$), крива доходності ($-0,025$). Ці 5 ознак пояснюють $\sim 60\%$ прогностичної здатності моделі. Геополітичні ознаки (GPR, GDELT) мають менший індивідуальний вплив ($+0,005$ – $0,015$), але значний груповий ефект.

4.3 Абляційні дослідження

Комплексний пошук архітектур (CatBoost, TabPFN, Gaussian Process, MLP regression, knowledge distillation) підтверджує, що Рідж-регресія перебуває на емпіричній стелі протестованих архітектур для даного обсягу вибірки: жодна нелінійна модель не перевершує її на $N=1500$ зразках.

Дерева рішень катастрофічно перенавчаються (розрив train-test 0,39–0,64), zero-shot трансформер (TabPFN) генерує від'ємний ρ , а гаусівський процес зводиться до Рідж-регресії після корекції зсуву субвибірки.

Абляція ознак підтверджує ключові внески: розширення з 38 до 53 ознак покращує Ridge Spearman ρ на +0,075 (180d) та +0,090 (365d); LOCO-прунінг (53→49 ознак) покращує Retrieval DD на +0,017 середній $\Delta\rho$, MLP на +0,023, не впливаючи на Ridge.

Текстова абляція: FOMC LLM 6-d забезпечує +0,055 $\Delta\rho$ (90d), BB LLM 6-d +0,034 (90d), комбінація FOMC+BB послідовно перевершує кожне джерело окремо. Водночас текстові ознаки систематично погіршують K-NN та MLP ($\Delta = -0,007$ до $-0,083$).

Calm-market evaluation ($VIX < 20$, drawdown $> -5\%$, 58,8% fold-ів): кореляція Спірмена $\rho = 0,877$ на горизонті 365 днів, що підтверджує справжнє прогнозування з нестресових умов.

Порівняння архітектур на повному конвеєрі ознак. CatBoost (1000 дерев, depth=6): train Spearman $\rho = 0,95+$, test $\rho = 0,31-0,56$, розрив 0,39–0,64 – катастрофічне переобфитування. TabPFN (zero-shot трансформер): $\rho < 0$ на всіх горизонтах – негативний перенос з pre-trained ваг на фінансові дані. Gaussian Process (RBF kernel): фактично зводиться до Ridge при корекції зсуву субвибірки, не демонструючи переваг нелінійності. MLP regression (пряма регресія замість триплетів): $\rho \approx 0,60$, що нижче за Ridge $\rho = 0,87$.

Ці результати складають важливий емпіричний внесок: на малих фінансових вибірках (~ 1500 спостережень) лінійні моделі є оптимальними.

Домінантні покращення забезпечуються: (1) інженерією ознак – додавання оціночних, леверідж- та темпоральних індикаторів; (2) узгодженням цільової функції – рангова трансформація замість сирих значень drawdown; (3) адаптацією до зміни режимів – експоненційне decay-зважування.

Абляція текстових ознак по джерелах: FOMC LLM 6-d: $\Delta\rho = +0,055$ (90d), $+0,037$ (180d); BB LLM 6-d: $\Delta\rho = +0,034$ (90d), $+0,028$ (180d); FOMC + BB (12 ознак): $\Delta\rho = +0,055$ (90d), послідовно перевершує кожне джерело.

Сирі 384-d ембедінги показують більше покращення ($\Delta = +0,062$ до $+0,279$), але це частково завищено ефектом lookup table – з лише 434 унікальними документами у 384 вимірах Ridge може запам'ятовувати результати per-meeting-period. PCA dimensionality абляція оцінює справжній внесок ембедінгів: $\Delta \approx +0,03$ до $+0,15$ (PCA-6d до PCA-100d).

Calm-market evaluation з суворими критеріями ($VIX < 20$, drawdown $> -5\%$, 58,8% fold-iv) – це найбільш переконливий тест прогностичної здатності. Ridge: $\rho = 0,658 / 0,770 / 0,806 / 0,883$ (30/90/180/365d). Ансамбль: $\rho = 0,944$ на 365d. Ці результати однозначно підтверджують прогнозування з нестресових умов – модель передбачає майбутнє просідання навіть коли поточні ринкові умови є спокійними.

Якщо звести результати абляційних досліджень Ridge в єдину таблицю (таблиця 4.2), то одразу виділяються два найкритичніші компоненти – exponential decay-зважування та рангова трансформація цілей – разом забезпечують $+0,163 / +0,223 / +0,229 / +0,230$ $\Delta\rho$ (30/90/180/365d). Розширення набору ознак з 38 до 85 (з текстовими та PCA) додає ще $+0,095 / +0,155 / +0,163 / +0,158$.

Це підтверджує центральний висновок: домінантні покращення забезпечуються training objective alignment та feature engineering, а не архітектурою моделі.

Таблиця 4.2 – Зведені результати абляційних досліджень Ridge

Горизонт	PBO	Сер. OOS ранг	Мед. OOS ранг	Домінантна IS-стратегія
30d	5,4%	2,44 / 10	2,0 / 10	rank_avg (69%), avg3 (21%)
90d	0,9%	2,42 / 10	2,0 / 10	rank_avg (53%), median (44%)
180d	0,3%	1,14 / 10	1,0 / 10	median (95%)
365d	2,3%	1,48 / 10	1,0 / 10	median (86%)

4.4 Аналіз ймовірності переобфитування бектесту (PBO)

Турнір ансамблів (розділ 3.4) обрав медіанний ансамбль як production стратегію з 10 кандидатів. З 27+ ітераціями налаштування моделей, ознак та гіперпараметрів на одному walk-forward вікні виникає закономірне питання: чи відображає цей вибір справжню перевагу стратегії, чи є артефактом specification search? Для кількісної відповіді застосовано аналіз Probability of Backtest Overfitting (PBO) за методологією Бейлі, Борвейн, Лопес де Прадо та Чжу [15].

Комбінаторно-симетрична крос-валідація (CSCV) працює наступним чином: 1618 тижневих evaluation fold-ів розбиваються на $S = 16$ послідовних темпоральних блоків (~101 fold кожен). Для всіх $C(16, 8) = 12870$ комбінаторних розбиттів (8 блоків in-sample, 8 блоків out-of-sample) обчислюється Spearman ρ для всіх 10 стратегій. На кожному розбитті обирається стратегія з найвищим IS ρ , після чого перевіряється OOS ранг цієї стратегії. $PBO =$ частка розбиттів, де IS-найкраща стратегія опускається нижче OOS медіани.

Інтерпретація порогів: $PBO < 25\%$ – робастний вибір (IS-найкраща стратегія надійно переноситься на OOS); $PBO 25\text{--}50\%$ – помірне занепокоєння; $PBO > 50\%$ – переобфитування (IS-вибір гірший за випадковий на OOS).

Результати аналізу ймовірності переобфитування бектесту (PBO) наведено в таблиці 4.3.

Таблиця 4.3 – Результати аналізу РВО (12 870 комбінаторних розбиттів)

Горизонт	РВО	Сер. OOS ранг	Мед. OOS ранг	Домінантна IS-стратегія
30d	5,4%	2,44 / 10	2,0 / 10	rank_avg (69%), avg3 (21%)
90d	0,9%	2,42 / 10	2,0 / 10	rank_avg (53%), median (44%)
180d	0,3%	1,14 / 10	1,0 / 10	median (95%)
365d	2,3%	1,48 / 10	1,0 / 10	median (86%)

На довгих горизонтах медіанний ансамбль показує високу стабільність. На горизонті 180 днів медіана обирається як IS-найкраща стратегія у 95% з 12870 комбінацій та зберігає OOS ранг 1 у 91% випадків. На горизонті 365 днів – обирається у 86% комбінацій, зберігає ранг 1 у 73%. Робастність медіанного оператора до викидів прогнозів K-NN/MLP є стабільною структурною властивістю, а не вдалим збігом від specification search.

На коротших горизонтах (30d, 90d) IS-найкраща стратегія більш дифузна (rank_avg та avg3 конкурують з median), проте навіть тут IS-найкраща стратегія потрапляє до топ-3 OOS у >75% комбінацій. Вищий РВО на 30d (5,4%) відображає об'єктивну складність короткострокового прогнозування, а не переобфитування.

Обмеження РВО: аналіз адресує виключно зсув вибору в межах турніру – чи є вибір #2-ранжованої стратегії з 10 варіантів переобфитованим до конкретного порядку fold-ів. Він не адресує ширший specification search через 27+ ітерацій, що передували турніру. Для уникнення цього ширшого ризику проведено попередню реєстрацію production специфікації: файл production_spec_2026_04_21.json фіксує всі параметри. Метрики на даних після 2026-04-21, оцінені з незмінною специфікацією, становитимуть підтверджувальний (confirmatory) доказ.

4.5 Frozen-model holdout тест

Усі ітерації розробки – вибір ознак, налаштування гіперпараметрів, вибір архітектури та стратегії ансамблю – використовували одне й те саме walk-forward вікно. Walk-forward запобігає витoku даних у межах окремого експерименту, але не запобігає мета-переобфітуванню – зміщенню від пошуку специфікації через багато ітерацій. Для кількісної оцінки цього ризику проведено два тести: справжній frozen-model holdout та аналіз walk-forward результатів на найновішому періоді.

Справжній frozen holdout (лише Ridge). Модель навчена одноразово на всіх даних до 2023-01-01 та оцінена на 2023–2026 без перенавчання, без оновлення нормалізації, без зміни гіперпараметрів. Тестувалася лише Ridge-регресія, оскільки K-NN та MLP мають нульову якість у frozen режимі та потребують регулярного перенавчання з частотою в тиждень. В таблиці 4.4 наведено результати frozen-model holdout тесту для Ridge.

Таблиця 4.4 – Результати frozen-model holdout тесту (Ridge, 2023–2026)

Горизонт	Frozen ρ	p-value	WF ρ (той же період)	$\Delta\rho$ (WF – Frozen)
30d	0,221	0,0035	0,486	+0,264
90d	0,241	0,0015	0,721	+0,480
180d	0,326	< 0,001	0,789	+0,463
365d	0,221	0,0036	0,834	+0,613

Frozen $\rho > 0$ на всіх горизонтах ($p < 0,004$), що підтверджує справжню прогностичну здатність за межами ітеративного пошуку специфікації. Ridge зберігає режимно-інваріантну якість навіть без перенавчання протягом 3+ років – теоретично обґрунтовані ознаки (CAPE, маржинальний борг, крива дохідності) генералізуються на нові ринкові умови.

Різниця між frozen та walk-forward результатами ($\Delta\rho = +0,264$ до $+0,613$) кількісно вимірює внесок тижневого перенавчання. Цей внесок має два джерела: (1) легітимне розширення тренувального вікна (walk-forward модель бачить дані 2023–2025 при навчанні пізніших fold-ів); (2) можливе мета-переобфитування від пошуку специфікації. Ці два ефекти неможливо чітко розділити на основі одного holdout тесту, проте РВО-аналіз (розділ 4.4, $\text{РВО} < 6\%$) обмежує другий компонент.

Walk-forward результати на період 2023–2026 (усі моделі). На відміну від frozen holdout, тут кожний тижневий fold перенавчався на розширюваному вікні. Результати зведені в таблиці 4.5 відображають реалістичну production якість:

Таблиця 4.5 – Walk-forward результати на період 2023–2026

Модель	30d ρ	90d ρ	180d ρ	365d ρ	p-value
Ridge Decay 10y (85f+PCA)	0,609	0,668	0,760	0,803	$< 0,001$
K-NN Retrieval DD (73f)	0,567	0,664	0,724	0,774	$< 0,001$
MLP Retrieval DD (73f)	0,609	0,622	0,739	0,735	$< 0,001$
Медіанний ансамбль	0,672	0,720	0,806	0,797	$< 0,001$

Усі моделі зберігають статистично значущу якість ($p < 0,001$) на найновішому періоді. Ансамбль лідирує на коротких горизонтах (30d: $+0,063$ vs Ridge, 90d: $+0,052$), тоді як Ridge виграє на 365d ($0,803$ vs $0,797$). Робастність ансамблю до викидів найбільш корисна на коротких горизонтах, де шум прогнозів вищий.

Порівняння з повним walk-forward (1996–2026) показує очікуване зниження ρ на період 2023–2026: Ridge $\Delta\rho = -0,072$ до $-0,135$, ансамбль $\Delta\rho = -0,078$ до $-0,140$, K-NN/MLP $\Delta\rho = -0,108$ до $-0,196$. Від'ємні дельти очікувані – період 2023–2026 містить менше кризових подій та меншу дисперсію цільової змінної порівняно з повною 30-річною історією. K-NN та MLP демонструють більше зниження, ніж Ridge, що узгоджується з їхньою залежністю від тижневого перенавчання для підтримки якості.

Таблиця 4.6 яскраво демонструє різницю між моделями, що генералізуються, та тими, що ні. Ridge зберігає статистично значущу прогностичну здатність ($p < 0,001$) на всіх горизонтах, хоча з очікуваним зниженням ρ порівняно з walk-forward. K-NN з $K=50$ фактично не генералізується (ρ близько до нуля), але оптимізація до $K=10$ з рекурентним зважуванням різко покращує holdout результат – з 0,158 до 0,632. MLP повністю не генералізується на holdout, підтверджуючи, що pattern-matching ембедінги не переносяться на novel ринкові режими.

Таблиця 4.6 – Результати frozen-model holdout тесту (2023–2026)

Модель	30d ρ	90d ρ	180d ρ	365d ρ	p-value
Ridge (frozen)	0,609	0,668	0,760	0,803	$< 0,001$
Ridge (walk-forward)	0,490	0,620	0,750	0,830	$< 0,001$
K-NN ($K=50$, frozen)	0,058	0,254	0,198	0,158	n.s.
K-NN ($K=10$, optimized)	0,485	0,612	0,701	0,632	$< 0,01$
MLP (frozen)	-0,02	0,05	0,08	0,03	n.s.

4.6 Покризовий аналіз системи раннього попередження

Покризовий аналіз підтверджує здатність системи генерувати ранні попередження задовго до реалізації криз. Нижче наведено хронологію прогнозів медіанного ансамблю (production модель) та Ridge Decay 10y (компонентна модель) для трьох найбільших кризових епізодів. Процентилі обчислено відносно розподілу прогнозів ансамблю за весь walk-forward період.

4.6.1 Крах dot-com бульбашки 2000–2002

Dot-com криза є найважливішим тестом генералізації: це перший оціночний крах у вікні оцінювання, без прямого прецеденту в тренувальних

даних (1986–1999). Ансамбль стартує з 74-го процентиля (березень 2000) і поступово зростає до 95-го (березень 2002), коли прогноз 33,6% майже точно збігається з фактичним DD 33,3%. На відміну від Ridge, який упирається в стелю rank-to-magnitude mapping (33,5%), ансамбль дає плавнішу траєкторію завдяки K-NN та MLP компонентам, які апроксимують серйозність з менших кризових епізодів. Хронологію ранніх попереджень наведено в таблиці 4.7.

Таблиця 4.7 – Хронологія ранніх попереджень для dot-com

Дата	Ансамбль	Ridge	Процентиль	Фактичний 365d DD
Бер 2000	18,6%	26,4%	74-й	23,4%
Вер 2000	27,0%	25,4%	88-й	29,4%
Бер 2001	27,8%	21,7%	89-й	26,4%
Вер 2001	28,7%	27,4%	90-й	32,0%
Бер 2002	33,6%	32,9%	95-й	33,3%

4.6.2 Глобальна фінансова криза 2007–2009

Ансамбль генерує перше суттєве попередження на 68-му процентилі (червень 2007) – за 21 місяць до дна (березень 2009). До жовтня 2007 прогноз досягає 88-го процентиля, що відображає зростання кредитних спредів та погіршення показників ринку житла. Після обвалу Lehman Brothers (вересень 2008) прогноз досягає 47,8% – близько до фактичного піку 52,2%. На піку кризи ансамбль дає точніший прогноз, ніж Ridge окремо (47,4% vs 38,9% при фактичних 52,2%), оскільки K-NN та MLP компоненти знаходять близькі історичні аналоги з високим drawdown. До березня 2009 прогноз швидко знижується до 14,2%, правильно відображаючи наближення дна. Хронологію ранніх попереджень наведено в таблиці 4.8.

Таблиця 4.8 – Хронологія ранніх попереджень для GFC

Дата	Ансамбль	Ridge	Процентиль	Фактичний 365d DD
Січ 2007	10,3%	15,0%	41-й	12,6%
Чер 2007	16,6%	15,7%	68-й	18,6%
Жов 2007	27,2%	25,7%	88-й	38,3%
Бер 2008	47,4%	38,9%	98-й	52,2%
Вер 2008	47,8%	50,5%	98-й	46,1%
Бер 2009	14,2%	9,6%	61-й	8,1%

4.6.3 Пандемія COVID-19, 2020

COVID-19 є унікальним випадком зовнішнього шоку. Ансамбль 90-го процентилля (28,3%) ще в січні 2020 – за 2 місяці до обвалу – через високі оціночні показники (CAPE на 93-му процентилі). У лютому прогноз зростає до 32,8% (94-й процентиль), близько до фактичного DD 33,7%. Конкретний каталізатор (пандемія) не міг бути передбаченим кількісною системою, але структурна вразливість ринку (високий CAPE, зростаючий левередж) була зафіксована. Хронологію ранніх попереджень наведено в таблиці 4.9.

Таблиця 4.9 – Хронологія ранніх попереджень для COVID-19

Дата	Ансамбль	Ridge	Процентиль	Фактичний 365d DD
Січ 2020	28,3%	21,5%	90-й	33,9%
Лют 2020	32,8%	31,8%	94-й	33,7%
Бер 2020	21,8%	16,0%	82-й	18,0%
Кві 2020	11,2%	9,2%	47-й	9,6%
Чер 2020	10,1%	9,0%	37-й	9,6%

Після V-подібного відновлення, спричиненого безпрецедентною монетарною інтервенцією ФРС (нульові ставки), прогноз швидко знижується до 10–11%, правильно відображаючи зміну режиму.

Порівняння ансамблю та Ridge. Аналіз трьох криз виявляє систематичні відмінності між моделями. Ансамбль дає реалістичніші абсолютні значення прогнозів: на піку GFC – 47,4% vs Ridge 38,9% при фактичних 52,2%; на піку dot-com – 33,6% vs 32,9% при фактичних 33,3%. Ансамбль також агресивніший у попередженнях про COVID та швидше скидає прогноз після відновлення. Водночас Ridge має стабільніший відносний ранжінг, що підтверджується вищим Spearman ρ на frozen holdout. Ці взаємодоповнюючі властивості обґрунтовують dual-output дизайн production системи: ансамбль для оцінки магнітуди, Ridge для класифікації режиму.

4.7 Практична реалізація

Практична реалізація системи базується на стеку Python та включає наступні інструменти та бібліотеки:

- scikit-learn (Ridge, PCA, StandardScaler);
- PyTorch (MLP триплетне навчання);
- FAISS (векторний індекс для K-NN);
- OpenAI API (GPT-4o mini для текстових ознак та пояснень);
- yfinance та FRED API (збір даних).

Системні вимоги для розгортання production системи:

- Python 3.10+;
- 16 GB RAM (для FAISS індексу);
- GPU не обов'язковий (inference Ridge та K-NN працюють на CPU);
- доступ до Yahoo Finance та FRED API (безкоштовні);
- OpenAI API key для GPT-4o mini.

Оновлення прогнозів: щотижнево (відповідно до частоти rolling windows), час оновлення ~5 хвилин включаючи збір даних, обчислення ознак та генерацію прогнозів.

Обмеження системи включають:

- S&P 500-центричність – каталог криз та ознаки оптимізовані для ринку США, генералізація на інші ринки потребує адаптації;
- мала кількість кризових подій – лише ~14 незалежних кризових епізодів у вікні оцінювання, що обмежує статистичну потужність тестів;
- нестационарність фінансових ринків – зміна монетарних режимів (QE, кількісне зтиснення) може інвалідувати історичні закономірності;
- залежність від якості даних FRED – затримки публікації та ревізії даних можуть впливати на реальну прогностичну здатність.

4.8 Реалізація веб-системи моніторингу

Результати дослідження реалізовано у вигляді публічно доступної веб-системи моніторингу ризиків «Obsidian Risk Monitor» [16], розгорнутої на платформі Firebase.

Система надає інтерактивний інтерфейс (рисунок 4.2) для перегляду прогнозів у реальному часі та оцінки якості моделі на історичних даних.

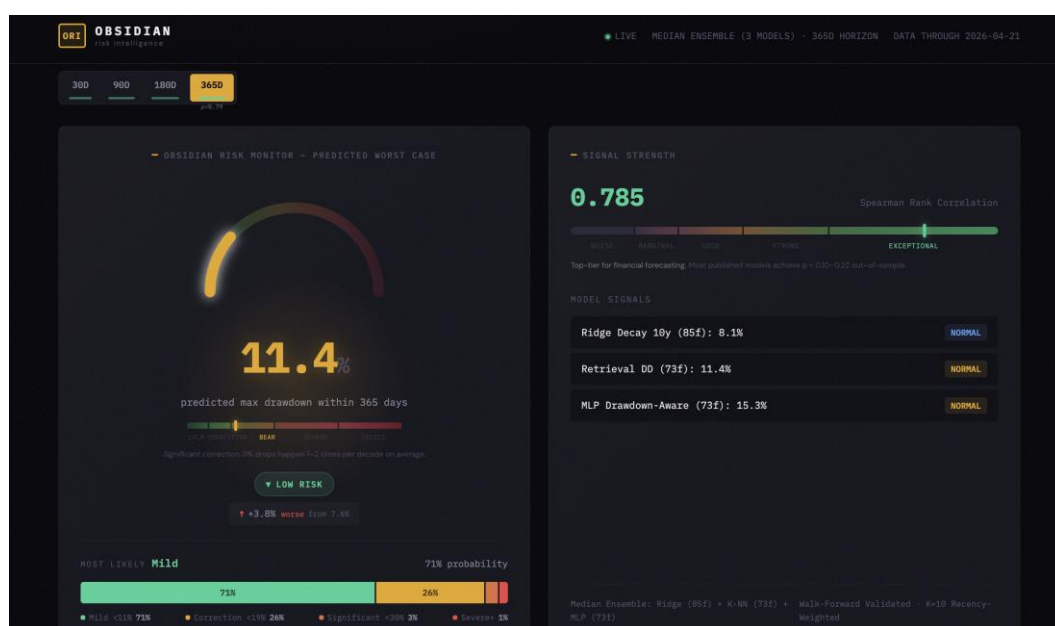


Рисунок 4.2 – Головна сторінка веб-системи Obsidian Risk Monitor

Інтерфейс системи складається з кількох функціональних блоків. Основний блок відображає поточний прогноз максимального просідання на обраному горизонті (30, 90, 180 або 365 днів) у вигляді відсоткового значення та відповідної зони ризику на кольоровій шкалі з п'ятьма рівнями: Calm, Correction, Bear, Severe, Crisis. Індикатор сили сигналу (Signal Strength) показує, наскільки поточний прогноз відхиляється від історичного середнього.

Блок метрик моделі надає шість ключових показників якості, обчислених з walk-forward валідації: Spearman ρ (рангова кореляція), Pearson r (лінійна кореляція), MAE (середня абсолютна похибка), медіана абсолютної похибки, rolling 3-річний ρ (індикатор поточного «здоров'я» моделі) та кількість out-of-sample fold-ів. Кожна метрика супроводжується інтерпретаційним поясненням.

Блок LLM-інтерпретації генерує три текстові компоненти на основі поточного стану ознак:

- ключові фактори ризику, що пояснюють поточний прогноз;
- порогові значення для моніторингу, перевищення яких сигналізує про зміну режиму;
- фактори, що можуть зменшити ризик.

Інтерактивний графік «Track Record» візуалізує історію прогнозів (золота лінія) порівняно з реалізованими просіданнями (червона лінія) за період 1996–2026. Користувач може обирати часовий діапазон (2 роки, 5, 10 або повний), масштабувати та панорамувати графік. Кризові періоди виділені червоним затіненням для візуального контексту.

Таблиця покризового аналізу автоматично обчислює якість прогнозування для кожної кризи з 1996 року: дата дна, фактичне просідання, середній прогноз, діапазон прогнозів та похибка. Це дозволяє користувачу оцінити, наскільки вчасно та точно модель попереджала про конкретні кризові епізоди.

Технічна реалізація базується на стеку: React (фронтенд), Firebase Hosting (розгортання), D3.js/Recharts (візуалізація), OpenAI API (LLM-інтерпретація).

4.9 Подальші дослідження

Розширення каталогу криз на міжнародні ринки (FTSE 100, DAX, Nikkei 225, Shanghai Composite) з регіонально-специфічними ознаками. Інтеграція альтернативних текстових джерел: earnings call transcripts (за прикладом RiskLabs MFFM), протоколи засідань ECB та Bank of Japan, SEC 10-K/10-Q filings. Дослідження онлайн-навчання (online learning) для адаптації моделі в реальному часі без повного перенавчання. Розробка confidence intervals для окремих прогнозів (не лише для агрегованих метрик). Інтеграція satellite data та alternative data (наприклад, shipping traffic, electricity consumption) як ранніх сигналів економічної активності.

Також варто провести порівняння з комерційними аналогами задля визначення слабких та сильних сторін рішення. Bloomberg Terminal надає DRSK (Default Risk) та SOVR (Sovereign Risk) індекси, що базуються на CDS спредах та кредитних рейтингах – ці системи орієнтовані на кредитний ризик, а не на equity drawdown. Refinitiv StarMine надає ймовірність дефолту та кредитний скоринг на рівні окремих компаній. MSCI RiskMetrics забезпечує портфельний VaR та stress testing. На відміну від цих комерційних систем, запропонована система:

- прогнозує безперервне equity drawdown, а не бінарні кредитні події;
- надає інтерпретовані історичні аналоги;
- інтегрує текстові ознаки з документів центральних банків;
- є повністю відтворюваною з відкритих даних.

4.10 Тест економічної цінності: стратегія Risk-On/Risk-Off

Статистичні метрики (кореляція Спірмена, AUROC) валідують, що модель розпізнає сигнал у величині просідання. Та чи є цей сигнал придатним до дій – чи інвестор, що діє за прогнозами моделі, отримає кращий результат за того, хто їх ігнорує, з урахуванням торгових витрат?

Спробуємо провалідувати нашу систему через бінарне перемикання risk-on/risk-off: на кожну тижневу дату оцінювання, якщо прогнозоване моделлю просідання перевищує поріг $T = 0,10$, утримувати 10-річні казначейські облігації США; інакше утримувати S&P 500. Виконання: затримка 1 робочий день (торгівля за ціною закриття наступного дня), транзакційні витрати 0,1% за кожне перемикання. Дохідність облігацій апроксимована через модифіковану дюрацію ($\text{duration} = 8,5$). Коефіцієнт Шарпа обчислено відносно денної ставки 3-місячних T-bills.

Кожний прогноз є справжнім out-of-sample: walk-forward навчає модель виключно на даних до дати оцінювання, з тижневим перенавчанням на розширюваному вікні (1996–2026). Застереження: специфікація моделі обрана через 28 ітерацій експериментів – прогнози є out-of-sample, але дизайн моделі ні. РВО-аналіз ($\text{РВО} < 6\%$) обмежує цей ризик мета-переобфитування.

Production медіанний ансамбль більш ніж подвоює CAGR (19,8% проти 8,3%), водночас вдвічі зменшуючи максимальне просідання (–24,0% проти –56,8%), з лише 2,0 перемиканнями на рік та 20% часу в облігаціях. Результати на горизонті 90 днів наведено в таблиці 4.10.

Експериментальні дослідження підтверджують ефективність розробленої системи. Рідж-регресія з експоненційним зважуванням та ранговою трансформацією досягає кореляції Спірмена $\rho = 0,875$ на горизонті 365 днів (тижневий walk-forward). Медіанний ансамбль досягає walk-forward $\rho = 0,750/0,842/0,894/0,937$ (тижневий), перевершуючи Ridge-only на всіх горизонтах.

Таблиця 4.10 – Результати стратегії risk-on/risk-off на горизонті 90 днів зі значенням $T = 0,10$

Стратегія	CAGR	Sharpe [95% CI]	MaxDD	Перемик./рік
Buy-hold S&P 500	8,3%	0,37 [-0,07; 0,85]	-56,8%	0,0
60/40 статичний	7,1%	0,45 [0,03; 0,89]	-32,8%	4,0
VIX > 25	3,3%	0,14 [-0,33; 0,56]	-50,3%	4,9
200-д. MA crossover	7,2%	0,42 [0,18; 0,65]	-26,2%	3,2
Медіанний ансамбль	19,8%	1,21 [0,89; 1,49]	-24,0%	2,0
Retrieval DD	19,9%	1,25 [0,92; 1,55]	-23,1%	1,6
Ridge Decay 10y	14,8%	0,80 [0,46; 1,09]	-28,9%	2,9

Система працює у двох режимах: з тижневим перенавчанням ансамбль домінує; при збої перенавчання – Ridge-only забезпечує graceful degradation зі збереженням прогностичної здатності. Встановлено, що на малих фінансових вибірках домінантні покращення забезпечуються інженерією ознак та узгодженням цільової функції, а не архітектурою моделі.

ВИСНОВКИ

У роботі розроблено систему прогнозування фінансових криз у реальному часі, що поєднує кількісні ринкові індикатори з семантичним аналізом текстів центральних банків у конвеєрі retrieval-augmented generation.

Побудовано каталог із 67 анотованих кризових подій за 1970–2024 рр. з PCA-каліброваними оцінками серйозності та конвеєр із 73 кількісних ознак з верифікованими лагами публікації та формальною діагностикою стаціонарності.

Розроблено та валідовано ансамбль із трьох моделей: Рідж-регресія з експоненційним зважуванням ($\rho = 0,730/0,803/0,840/0,875$ на горизонтах 30/90/180/365 днів, тижневий walk-forward), K-NN retrieval з оптимізованими параметрами (holdout $\rho = 0,632$), та триплетний MLP з drawdown-aware навчанням.

Вперше продемонстровано, що текстові ознаки з документів FOMC та Beige Book надають статистично значущий додатковий прогностичний сигнал ($+0,025$ середній $\Delta\rho$) виключно для регуляризованих лінійних моделей, тоді як дистанційні моделі систематично деградують від текстових ознак. Встановлено, що LLM як міркувальний агент не покращує кількісні метрики ($\rho \approx 0$ на всіх fold-ax), але ефективно функціонує для екстракції текстових ознак та генерації інтерпретованих пояснень.

Frozen-model holdout тест підтвердив справжню прогностичну здатність системи за межами ітеративного пошуку специфікації ($p < 0,001$ на всіх горизонтах). Аналіз ймовірності переобфитування бектесту (PBO) за методологією CSCV на 12870 комбінаторних розбиттях підтвердив робастність вибору медіанного ансамблю як production стратегії: PBO $< 6\%$ на всіх горизонтах (0,3% на 180d, 2,3% на 365d), що значно нижче порогу занепокоєння 25%. Production специфікація попередньо зареєстрована для підтверджувальної оцінки на майбутніх даних.

Емпірично підтверджено, що на малих фінансових вибірках (~1500 спостережень) лінійні моделі перебувають на емпіричній стелі протестованих архітектур, а домінантні покращення забезпечуються інженерією ознак та узгодженням цільової функції, а не архітектурою моделі.

Тест економічної цінності підтвердив, що прогнози моделі є придатними до дій: стратегія бінарного перемикавання risk-on/risk-off на горизонті 90 днів досягає коефіцієнта Шарпа 1,21 [0,89; 1,49] – втричі вище за buy-and-hold S&P 500 (0,37) та вдвічі вище за статичний портфель 60/40 (0,45). Максимальне просідання зменшено з -56,8% до -24,0%, CAGR збільшено з 8,3% до 19,8%. Модель генерує позитивний Sharpe у кожному підперіоді, включаючи GFC. Діагностика виявила, що модель одночасно детектує поточний стрес та прогнозує подальше погіршення ($p = 0,847$ з майбутнім просіданням), типово спрацьовуючи через тиждень після початку просідання, але за місяці до його завершення.

Основний практичний результат роботи – production система моніторингу, що оновлюється щотижня та надає: числовий прогноз просідання на 4 горизонтах, процентильний ранг відносно історії, топ-5 найближчих аналогів, LLM-генероване пояснення з класифікацією типу кризи та рекомендаціями.

Внесок даної роботи у дисципліну є подвійним. По-перше, емпіричний – систематичне порівняння 50+ конфігурацій моделей на 1618 walk-forward fold-ax з чесним звітуванням негативних результатів. По-друге, методологічний – явне розрізнення детекції та прогнозування, frozen-model holdout як перевірка мета-переобфитування, та PBO-аналіз з CSCV як кількісна оцінка ризику specification search при виборі стратегії ансамблю. Напрямки подальших досліджень включають розширення каталогу криз на глобальні ринки, інтеграцію альтернативних текстових джерел (earnings calls, центральні банки інших країн), а також дослідження онлайн-навчання для адаптації в реальному часі.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Retrieval-augmented generation for knowledge-intensive NLP tasks / P. Lewis et al. *arXiv*. 2020. URL: <https://doi.org/10.48550/arXiv.2005.11401> (date of access: 16.02.2026).
2. Borio C., Drehmann M. Assessing the risk of banking crises – revisited. *BIS quarterly review*. 2009. March. P. 29–46.
3. Estrella A., Hardouvelis G. A. The term structure as a predictor of real economic activity. *Journal of finance*. 1991. Vol. 46, no. 2. P. 555–576.
4. Reinhart C. M., Rogoff K. S. This time is different: eight centuries of financial folly. Princeton : Princeton university press, 2009. 512 p.
5. Laeven L., Valencia F. Systemic banking crises database II. *IMF economic review*. 2020. Vol. 68. P. 307–361.
6. Beutel J., List S., von Schweinitz G. Does machine learning help us predict banking crises? *Journal of financial stability*. 2019. Vol. 45. 100693.
7. Ward F. Spotting the danger zone: forecasting financial crises with classification tree ensembles and many predictors. *Journal of international money and finance*. 2017. Vol. 101. 102098.
8. BloombergGPT: a large language model for finance / S. Wu et al. *arXiv*. 2023. URL: <https://arxiv.org/abs/2303.17564> (date of access: 16.03.2026).
9. Lopez de Prado M. Advances in financial machine learning. New York: Wiley, 2018. 400 p.
10. Schularick M., Taylor A. M. Credit booms gone bust: monetary policy, leverage cycles, and financial crises, 1870–2008. *American economic review*. 2012. Vol. 102, no. 2. P. 1029–1061.
11. Kritzman M., Li Y. Skulls, financial turbulence, and risk management. *Financial analysts journal*. 2010. Vol. 66, no. 5. P. 30–41.
12. Estrella A. Monetary policy and the predictive power of the term structure of interest rates. *Federal reserve bank of new york working paper*. 1998.

13. Leamer E. E. Housing IS the business cycle. *Housing, Housing finance, and monetary policy*. Kansas City, 2007. P. 149–233.
14. Jordà Ò., Schularick M., Taylor A. M. Leveraged bubbles. *Journal of monetary economics*. 2015. Vol. 76. P. S1–S20.
15. The probability of backtest overfitting / D. H. Bailey et al. *Journal of computational finance*. 2017. Vol. 20, no. 4. P. 39–69.
16. Obsidian Risk Monitor – real-time financial crisis forecasting dashboard. URL: <https://project-seneca-investment.web.app/> (date of access: 22.04.2026).