

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 _____ перший (бакалаврський)

Інтелектуальна система оцінювання ринкової вартості футболістів на основі
аналізу ігрових та статистичних характеристик методами машинного навчання
(тема)

Виконав:
здобувач _____ четвертого _____ року навчання,
групи _____ ІТШ-22-3
_____ Максим ГРЕЧКА
(власне ім'я, прізвище)

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

Тип програми _____ освітньо-професійна
Освітня програма _____ Штучний інтелект

(повна назва освітньої програми)
Керівник _____ ас. Максим ПОЛІТ
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ

(підпис)

_____ Лариса ЧАЛА
(власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____

Кафедра _____ Штучного інтелекту _____

Рівень вищої освіти _____ перший (бакалаврський) _____

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

Тип програми _____ освітньо-професійна _____

Освітня програма _____ Штучний інтелект _____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

« _____ » _____ 20 ____ р.

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві _____ Гречці Максиму Сергійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _____ Інтелектуальна система оцінювання ринкової вартості футболістів на основі аналізу ігрових та статистичних характеристик методами машинного навчання _____

затверджена наказом університету від 5 червня 2026 р. № 488Ст _____

2. Термін подання здобувачем роботи до екзаменаційної комісії 10 липня 2026 р. _____

3. Вихідні дані до роботи _____ Науково-технічні публікації, інтернет-джерела з футбольної аналітики, набір даних Football Data from Transfermarkt, документація Python, pandas, scikit-learn, XGBoost, SHAP та Streamlit. _____

4. Перелік питань, що потрібно опрацювати в роботі _____

1) Аналіз предметної галузі _____

2) Опис методів машинного навчання _____

3) Розробка інтелектуальної системи _____

4) Експериментальні дослідження та аналіз результатів _____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / термін виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	08.06.2026	виконано
2	Аналіз предметної галузі	10.06.2026	виконано
3	Огляд існуючих підходів до оцінювання ринкової вартості футболістів	12.06.2026	виконано
4	Формування та попередня обробка набору даних	16.06.2026	виконано
5	Реалізація і навчання моделей машинного навчання	20.06.2026	виконано
6	Експериментальне дослідження та аналіз результатів	23.06.2026	виконано
7	Реалізація вебзастосунку системи	25.06.2026	виконано
8	Написання пояснювальної записки	30.06.2026	виконано
9	Нормоконтроль	03.07.2026	виконано
10	Захист перед ЕК	10.07.2026	

Дата видачі завдання 8 червня 20 26 р.

Здобувач 
(підпис)

Керівник роботи 
(підпис)

ас. Максим ПОЛІТ
(посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 81 с., 15 рис., 2 табл., 1 дод., 20 джерел.

ГРАДІЄНТНИЙ БУСТИНГ, ІНТЕЛЕКТУАЛЬНА СИСТЕМА, ІНТЕРПРЕТАЦІЯ МОДЕЛІ, МАШИННЕ НАВЧАННЯ, РЕГРЕСІЙНА МОДЕЛЬ, РИНКОВА ВАРТІСТЬ ФУТБОЛІСТА, СПОРТИВНА АНАЛІТИКА, ФУТБОЛІСТ.

Об'єкт дослідження – процес оцінювання ринкової вартості футболістів на основі статистичних та контекстних характеристик.

Предмет дослідження – методи машинного навчання для прогнозування ринкової вартості футболістів за ігровими та статистичними ознаками.

Мета роботи – розроблення інтелектуальної системи оцінювання орієнтовної ринкової вартості футболістів із використанням методів машинного навчання.

Методи дослідження – системний аналіз, методи попередньої обробки даних, регресійний аналіз, ансамблеві методи машинного навчання, градієнтний бустинг, методи оцінювання якості моделей, SHAP-інтерпретація, програмна реалізація засобами Python.

У роботі досліджено задачу оцінювання ринкової вартості футболістів на основі відкритих футбольних даних. Сформовано набір даних рівня «гравець-сезон», виконано попередню обробку даних, побудовано ознаки та навчено регресійні моделі Ridge Regression, Random Forest і XGBoost. Якість моделей оцінено за метриками MAE, RMSE та R^2 . Для коректного оцінювання використано групове розбиття даних за ідентифікатором футболіста. Практичним результатом є вебзастосунок, який прогнозує орієнтовну ринкову вартість футболіста та пояснює результат за допомогою SHAP-інтерпретації.

ABSTRACT

Bachelor's thesis contains: 81 pp., 15 fig., 2 tabl., 1 ann., 20 references.

FOOTBALL ANALYTICS, FOOTBALL PLAYER, GRADIENT BOOSTING, INTELLIGENT SYSTEM, MACHINE LEARNING, MARKET VALUE OF A FOOTBALL PLAYER, MODEL INTERPRETATION, REGRESSION MODEL, SPORTS ANALYTICS.

The object of research is the process of assessing the market value of football players based on statistical and contextual characteristics.

The subject of research is machine learning methods for predicting the market value of football players using playing, statistical and market-related features.

The aim of the work is to develop an intelligent system for estimating the approximate market value of football players using machine learning methods and prediction interpretation.

Research methods include system analysis, data preprocessing methods, regression analysis, ensemble machine learning methods, gradient boosting, model quality evaluation methods, SHAP interpretation, and software implementation using Python.

The work investigates the task of estimating the market value of football players based on open football data. A player-season dataset was formed, data preprocessing was performed, features were constructed, and regression models such as Ridge Regression, Random Forest and XGBoost were trained.

The quality of the models was evaluated using MAE, RMSE and R^2 metrics. For correct evaluation, group splitting by player identifier was used. The practical result is a web application that predicts the approximate market value of a football player and explains the result using SHAP interpretation.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів	7
Вступ.....	8
1 Аналіз предметної галузі	10
1.1 Ринкова вартість футболіста як об'єкт дослідження.....	10
1.2 Фактори та проблеми оцінювання ринкової вартості футболістів	12
1.3 Огляд існуючих платформ і методів оцінювання.....	16
1.4 Інноваційність та переваги розроблюваної системи	20
1.5 Постановка задачі дослідження.....	23
2 Опис методів машинного навчання.....	26
2.1 Задача регресійного прогнозування ринкової вартості футболістів .	26
2.2 Регресійні моделі машинного навчання	29
2.3 Метрики оцінювання якості моделей	34
2.4 Квантильна регресія для оцінювання невизначеності прогнозу	37
2.5 Методи інтерпретації результатів прогнозування	40
3 Розробка інтелектуальної системи оцінювання ринкової вартості футболістів	43
3.1 Загальна архітектура інтелектуальної системи.....	43
3.2 Формування та попередня обробка набору даних	46
3.3 Реалізація моделей машинного навчання.....	51
3.4 Реалізація оцінювання невизначеності та інтерпретації прогнозів ...	55
4 Експериментальне дослідження та аналіз результатів.....	66
4.1 Умови проведення експериментального дослідження.....	66
4.2 Порівняння якості моделей машинного навчання.....	67
4.3 Інтерпретація результатів прогнозування	70
4.4 Обмеження системи та напрями подальшого розвитку	73
Висновки	76
Перелік джерел посилання	78
Додаток А Відомість кваліфікаційної роботи	81

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ІІІ – штучний інтелект;

AI – Artificial Intelligence – штучний інтелект;

CSV – Comma-Separated Values – формат файлів із табличними даними, розділеними комами;

MAE – Mean Absolute Error – середня абсолютна помилка;

ML – Machine Learning – машинне навчання;

RMSE – Root Mean Squared Error – корінь із середньоквадратичної помилки;

R^2 – коефіцієнт детермінації;

SHAP – SHapley Additive exPlanations – метод інтерпретації моделей машинного навчання;

XGBoost – Extreme Gradient Boosting – алгоритм екстремального градієнтного бустингу.

ВСТУП

Сучасний трансферний ринок у футболі вирізняється великими фінансовими обсягами, гострим суперництвом між клубами та дедалі ширшим переходом до ухвалення рішень, що спираються на дані [1]. У професійному футболі гравця сприймають не тільки як учасника спортивного процесу, а й як вагомий економічний актив команди. Саме тому обґрунтоване визначення ринкової вартості футболіста постає важливим завданням для планування трансферної стратегії, керування бюджетом та зменшення фінансових ризиків.

Незважаючи на розвиток спортивної аналітики, процедура оцінювання вартості гравців лишається складною та частково суб'єктивною. На підсумкову оцінку здатні впливати не лише ігрові показники, а й вік, позиція, рівень чемпіонату та інші зовнішні фактори. З цієї причини традиційні експертні методи не завжди дають змогу достатньо об'єктивно врахувати справжній внесок футболіста, сталість його виступів і статистичні характеристики.

Завдання оцінювання ринкової вартості ускладнюється багатовимірністю вхідних даних та нелінійним характером залежностей між ознаками. Ті самі статистичні показники можуть мати неоднакову вагу залежно від віку футболіста, його позиції, рівня турніру й ігрової ролі в команді. Через це для опрацювання таких даних доцільно застосовувати методи машинного навчання, спроможні виявляти приховані закономірності у великих масивах статистичної інформації.

Нині наявні різноманітні інформаційні платформи та аналітичні сервіси, що надають відомості про футболістів і їхню ринкову вартість. Проте частина з них спирається на експертні або краудсорсингові оцінки, а інші послуговуються закритими алгоритмами, що утруднює розуміння логіки формування підсумку. Для практичного застосування у спортивній

аналітиці істотно не тільки одержати прогнозну оцінку вартості, а й роз'яснити, які саме характеристики найбільше позначилися на результаті.

Актуальність роботи зумовлена потребою у створенні інтелектуальної системи, що поєднує аналіз футбольних статистичних даних, методи машинного навчання та засоби інтерпретації прогнозів. Такий підхід дає змогу підвищити прозорість оцінювання, послабити залежність від суб'єктивних суджень і сформувати додаткову аналітичну підтримку для ухвалення рішень у сфері скаутингу та трансферної політики.

Мета роботи – створення інтелектуальної системи оцінювання орієнтовної ринкової вартості футболістів на основі аналізу ігрових та статистичних характеристик засобами машинного навчання.

Для досягнення зазначеної мети потрібно розв'язати такі задачі: проаналізувати предметну галузь оцінювання ринкової вартості футболістів; розглянути наявні підходи та аналітичні платформи; сформувати набір даних рівня «гравець-сезон»; здійснити попереднє опрацювання даних і конструювання ознак; навчити та зіставити регресійні моделі машинного навчання; оцінити якість моделей за метриками MAE, RMSE та R^2 ; реалізувати інтерпретацію прогнозів за допомогою SHAP; розробити програмний прототип системи.

Об'єктом дослідження є процес оцінювання ринкової вартості футболістів на підставі статистичних та контекстних характеристик.

Практичним підсумком роботи є програмний прототип інтелектуальної системи, що уможливорює введення характеристик футболіста, одержання прогнозу його орієнтовної ринкової вартості та пояснення впливу окремих ознак на підсумок. Отримані результати можуть слугувати допоміжним інструментом для спортивної аналітики, скаутингу та подальшого вдосконалення систем підтримки ухвалення рішень у футбольному менеджменті.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Ринкова вартість футболіста як об'єкт дослідження

Сучасна професійна футбольна індустрія – це складна комерційна екосистема, в якій футбольні клуби функціонують не лише як спортивні організації, а й як учасники економічної діяльності. У такій системі футболіст вважається одним із головних активів клубу, оскільки його спортивні досягнення, потенціал зростання, медійна привабливість та майбутні трансферні перспективи безпосередньо впливають на фінансову стратегію команди.

Ринкова вартість футболіста – це приблизна оцінка економічної вартості гравця на трансферному ринку. Вона не завжди збігається з фактичною сумою трансферу, оскільки реальна трансферна ціна може визначатися багатьма додатковими обставинами: терміном дії контракту, фінансовими можливостями клубу, що купує, суперництвом між командами, терміновістю угоди, агентськими угодами та загальним станом ринку. Незважаючи на це, ринкова вартість залишається значним аналітичним показником, який використовується для порівняння футболістів, оцінки складу команди та підтримки управлінських рішень у трансферній політиці [2].

Збільшення фінансового обороту у професійному футболі посилює роль об'єктивної оцінки гравців. Для клубів неточність у визначенні вартості футболіста може призвести до значних фінансових ризиків: переплати за переоціненого гравця або втрати можливості придбати перспективного футболіста за нижчою ціною. Обидва сценарії безпосередньо впливають на трансферний бюджет і довгострокову стратегію команди. Тому завдання оцінки ринкової вартості є актуальним як для спортивних директорів, так і для аналітичних відділів футбольних клубів, які прагнуть ухвалювати обґрунтовані та фінансово виважені рішення.

На рисунку 1.1 наведено приклад динаміки міжнародних трансферів у чоловічому професійному футболі, що демонструє масштабність сучасного трансферного ринку.



Рисунок 1.1 – Динаміка міжнародних трансферів у чоловічому професійному футболі

Формування ринкової вартості футболіста – це багатофакторний процес. На остаточну оцінку впливають не лише основні ігрові показники, а й особистісні характеристики, позиція на полі, рівень чемпіонату, стабільність виступів та загальний ринковий контекст [3], [4]. Водночас вплив індивідуальних характеристик не однаковий для всіх гравців. Наприклад, для нападників важливіша продуктивність, тоді як для захисників та воротарів важливішими можуть бути інші характеристики, пов'язані з ігровою надійністю та роллю в команді.

У межах цієї роботи основну увагу зосереджено на тих характеристиках, які можна одержати з відкритих структурованих футбольних даних і застосувати для побудови моделей машинного навчання. Ці характеристики включають сезон, вік гравця, основну позицію, підпозицію, зріст, громадянство, лігу, ігровий час, середню кількість хвилин

за матч, голи, результативні передачі, дисциплінарні показники та кількість турнірів, у яких гравець грав протягом сезону.

Варто наголосити, що в роботі прогнозується не фактична сума трансферної угоди, а приблизний показник ринкової вартості футболіста. Такий підхід є виправданим, оскільки реальні суми трансферів є нерегулярними, залежать від ходу переговорів і не завжди повністю відображають спортивну цінність гравця. Натомість, історичні оцінки ринкової вартості дозволяють сформувати більш повний набір прикладів для навчання моделей машинного навчання.

Отже, ринкова вартість футболіста є складним об'єктом дослідження, що формується під впливом багатьох взаємопов'язаних факторів. Її аналіз вимагає використання методів, здатних обробляти різноманітні дані, враховуючи нелінійні залежності між ознаками та забезпечуючи подальшу інтерпретацію результатів. Саме тому використання методів машинного навчання є розумним підходом до вирішення проблеми оцінки ринкової вартості футболістів [5].

1.2 Фактори та проблеми оцінювання ринкової вартості футболістів

Ринкова вартість футболіста визначається великою кількістю взаємопов'язаних факторів. На відміну від звичайних товарів чи фінансових активів, вартість професійного спортсмена не може бути визначена одним показником. Вона залежить від спортивних результатів, віку, позиції на полі, стабільності виступів, рівня чемпіонату, попиту на гравців певного амплуа та загального стану трансферного ринку.

Однією з ключових груп факторів є особисті характеристики гравця. До них належать вік, зріст, національність, основна позиція та більш детальна ігрова роль. Вік має особливе значення, оскільки він пов'язаний не лише з поточною фізичною формою гравця, але й з потенціалом подальшого зростання. Молодші гравці часто мають вищу ринкову вартість через

очікування прогресу та можливість майбутнього перепродажу, тоді як вартість старших гравців зазвичай поступово знижується через коротший період їхньої кар'єри, що залишився.

Ігрові та статистичні характеристики мають велике значення. До них належать загальний ігровий час, голи, результативні передачі та дисциплінарні показники. Ці характеристики дозволяють оцінити не лише результативність футболіста, але й ступінь його залученості до командної гри. Наприклад, значна кількість хвилин, проведених на полі, може свідчити про довіру тренерського штабу та стабільну роль гравця в складі.

Окрему роль відіграє позиція футболіста на полі. Одні й ті ж статистичні показники можуть мати різне значення для гравців різних амплуа. Для нападників важливіші голи та результативні дії, для півзахисників – участь у розвитку гри, а для захисників – стабільність, регулярність виступів та надійність виконання ігрових завдань. Тому під час побудови моделі машинного навчання позиційні характеристики слід розглядати як окрему групу ознак.

Контекстуальні фактори також впливають на оцінку гравця. До них належать сезон, ліга, кількість турнірів, у яких грав гравець, та загальний ринковий попит на певний тип гравців. Статистика гравця може сприйматися по-різному залежно від рівня ліги. Наприклад, результати у вищій європейській лізі оцінюються інакше, ніж аналогічні показники в менш конкурентному турнірі.

Загалом фактори, що визначають ринкову вартість футболіста, можна поділити на кілька груп відповідно до способу їх використання в роботі. У моделі враховуються особисті характеристики гравця, зокрема вік, зріст і громадянство. Також використовуються позиційні характеристики, ігровий час, результативність, дисциплінарні показники та контекст виступів, до якого належать сезон, ліга і кількість турнірів. На противагу цьому, контрактні та нестатистичні фактори – термін дії контракту, зарплата, схильність до травм, популярність у ЗМІ, психологічні якості та вплив

агентів – можуть суттєво впливати на реальну вартість гравця, але у відкритих джерелах вони часто є неповними, закритими або важко кількісно оцінюваними. Тому вони не використовувалися в рамках цієї роботи, а розглядалися як обмеження предметної галузі та можливий напрямок подальшого розвитку системи.

Ще однією суттєвою проблемою є суб'єктивність історичних оцінок та наявність інформаційного шуму в даних. Ринкова вартість футболіста формується не лише на основі спортивної статистики, а й під впливом зовнішніх обставин, які не завжди піддаються точному вимірюванню. Через це модель машинного навчання оперує з наближеним відображенням реального трансферного ринку, і отриманий прогноз слід сприймати як приблизну аналітичну оцінку, а не як точну суму майбутнього трансферу.

Окремою складністю є проблема інтерпретованості моделей машинного навчання. Сучасні алгоритми, зокрема ансамблеві методи та градієнтний бустинг, здатні виявляти складні нелінійні залежності між ознаками та забезпечувати високоякісне прогнозування. Однак такі моделі можуть працювати за принципом «чорної скриньки», коли користувач отримує кінцевий результат без достатнього пояснення логіки його формування.

Проблема чорної скриньки особливо важлива для завдання оцінки ринкової вартості футболістів, оскільки кінцевий результат прогнозування може вплинути на подальші аналітичні висновки гравця щодо нього. Якщо система генерує лише числову оцінку, але не пояснює причини її отримання, користувачеві складно зрозуміти, чи є прогноз обґрунтованим. Наприклад, високе прогнозоване значення може бути пов'язане з молодим віком футболіста, стабільним ігровим часом, результативністю або виступами в сильнішій лізі. Без пояснення внеску цих ознак модель залишається менш корисною для практичного застосування. Наприклад, високе прогнозоване значення може бути пов'язане з молодим віком футболіста, стабільним ігровим часом, результативністю або виступами в сильнішій лізі. Без

пояснення внеску ці особливості моделі залишаються менш корисними для практичного застосування.

На рисунку 1.2 наведено узагальнену ілюстрацію проблеми «чорного ящика» у системах машинного навчання.

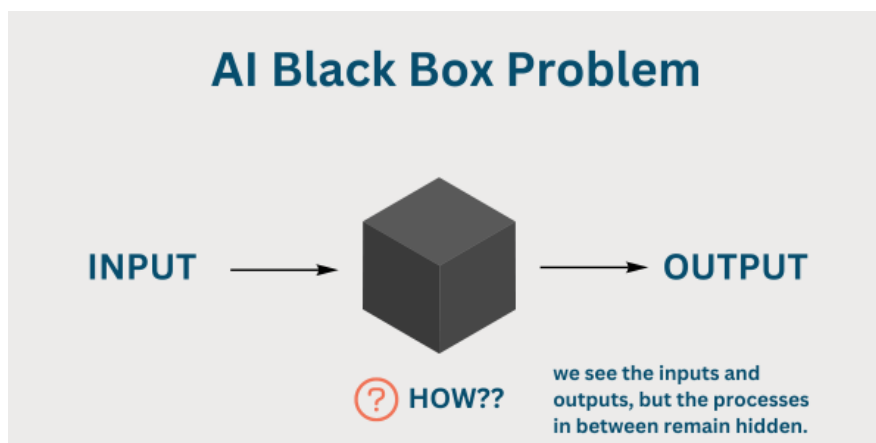


Рисунок 1.2 – Ілюстрація проблеми «чорного ящика» у системах машинного навчання

Для спортивного аналітика, скаута чи менеджера футбольного клубу недостатньо просто отримати прогнозовану ринкову вартість футболіста. Важливо також зрозуміти, які фактори найбільше вплинули на результат: вік, позиція, ігровий час, результативність, ліга чи інші характеристики. Без такого пояснення практична цінність системи знижується, оскільки користувач не може оцінити достовірність прогнозу.

Саме тому, в рамках цієї роботи, важливим компонентом системи є інтерпретація результатів прогнозування. Для пояснення впливу окремих характеристик доцільно використовувати пояснювальні методи штучного інтелекту, зокрема SHAP. Це дозволяє не тільки отримати прогноз ринкової вартості футболіста, але й показати, які характеристики збільшують або зменшують отриману оцінку.

Отже, оцінка ринкової вартості футболістів є складним завданням через багатофакторність, неповноту відкритих даних, наявність

суб'єктивних факторів та проблему інтерпретованості моделей. Для її вирішення доцільно використовувати методи машинного навчання в поєднанні з інструментами пояснення прогнозів, що дозволяє підвищити прозорість та практичну корисність інтелектуальної системи.

1.3 Огляд існуючих платформ і методів оцінювання

У сфері футбольної аналітики існує низка інформаційних ресурсів, комерційних сервісів та наукових підходів, що використовуються для аналізу футболістів та визначення їхньої ринкової вартості. Такі системи відрізняються джерелами даних, методологією оцінки, глибиною статистичної обробки, ступенем автоматизації та прозорістю отриманих результатів. Огляд існуючих рішень є важливим етапом роботи, оскільки дозволяє окреслити їх сильні сторони, обмеження та обґрунтувати доцільність створення власної інтелектуальної системи.

Одним із найвідоміших ресурсів у цій галузі є вебпортал Transfermarkt [6]. Платформа містить велику базу даних футболістів, клубів, трансферних транзакцій, статистики ігор та змін ринкової вартості. Оцінки вартості гравців, представлені тут, широко використовуються в інформаційному просторі футболу як орієнтир для порівняння футболістів та аналізу трансферного ринку. Однак методологія Transfermarkt значною мірою базується на експертних та краудсорсингових оцінках, що може вносити певну частку суб'єктивності.

Інтерфейс платформи Transfermarkt побудований навколо індивідуального профілю футболіста, який містить основну інформацію про гравця, його клубну приналежність, ігрову статистику та поточну ринкову вартість. Особлива увага приділяється візуалізації зміни вартості гравця з часом, що дозволяє відстежувати динаміку його оцінки протягом усієї кар'єри залежно від результатів, віку та трансферів між клубами. Саме така структура подання інформації зробила Transfermarkt одним із

найпопулярніших джерел для аналітиків, скаутів та дослідників футбольного ринку.

Приклад інтерфейсу профілю футболіста на платформі Transfermarkt подано на рисунку 1.3.



Рисунок 1.3 – Інтерфейс профілю гравця та динаміка його ринкової вартості на платформі Transfermarkt

Сильними сторонами Transfermarkt є широке охоплення футбольних ліг, доступність інформації та наявність історичних даних про зміни вартості гравців. Саме тому такі дані часто слугують основою для досліджень, пов'язаних з прогнозуванням ринкової вартості футболістів. Однак повна логіка формування остаточної оцінки залишається недостатньо прозорою для користувачів, а сама процедура оцінки може залежати від активності спільноти, експертної думки та загального інформаційного фону навколо гравця [7].

Інший підхід до оцінки вартості футболістів пропонує методологія CIES Football Observatory [8]. Ця платформа використовує статистичні

моделі для розрахунку трансферної вартості гравців, враховуючи історичні дані про трансфери, вік гравця, рівень клубу, чемпіонат та інші фактори [9]. На відміну від краудсорсингового підходу, CIES Football Observatory більше спирається на формалізоване моделювання та аналітичну обробку даних.

Приклад інтерфейсу оцінювання вартості футболіста на платформі CIES Football Observatory подано на рисунку 1.4.

CIES Football Observatory
Estimated transfer values
Last update: 05/05/2025

League: All | Club: Search club | Player: Search player | Clear filters

Top estimated value per club in 70 leagues worldwide, not including players on loan

Club	Player	Fair price [®]	Birth
FC Barcelona (ESP)	Lamine Yamal (w/2023)	€349-405	2007
Manchester City (ENG)	Erling Haaland (chw/2024)	€206-240	2000
Real Madrid (ESP)	Kylian Mbappé (chw/2025)	€158-184	1998
Bayern München (GER)	Michael Olise (w/2023)	€133-154	2001
Juventus FC (ITA)	Kenan Yildiz (w/2023)	€125-145	2005
Paris St-Germain (FRA)	Desiré Doué (w/2023)	€119-138	2005
Aston Villa (ENG)	Morgan Rogers (w/2023)	€118-137	2002
Liverpool FC (ENG)	Florian Wirtz (w/2023)	€117-136	2003
Chelsea FC (ENG)	Estêvão Willian (w/2023)	€110-128	2007
Atlético Madrid (ESP)	Julian Alvarez (chw/2023)	€104-121	2000
Arsenal FC (ENG)	Bukayo Saka (w/2023)	€104-120	2001
RB Leipzig (GER)	Yan Dabon (w/2023)	€94-109	2006
Newcastle United (ENG)	Nick Wolterbeek (chw/2023)	€90-105	2002
SSC Napoli (ITA)	Rasmus Højlund (chw/2023)	€90-105	2003
FC Inter (ITA)	Pio Esposito (chw/2023)	€89-103	2005

Estimates with actual contract duration available
for subscriptions
Multiplayer

cies FOOTBALL OBSERVATORY
www.cies-football-observatory.com

Рисунок 1.4 – Інтерфейс оцінювання ринкової вартості футболіста на платформі CIES Football Observatory

Перевагою підходу CIES є використання статистичного моделювання та зосередженість на кількісній оцінці цінності футболістів. Водночас, звітні особливості моделі не повністю розкриті широкому загалу, що забезпечує можливість незалежної перевірки результатів та знижує прозорість системи.

Подібну тенденцію представляє платформа Football Benchmark, яка пропонує інструменти для аналізу футболістів, клубів та футбольного ринку загалом [10]. Її Player Valuation Tool генерує оцінки ринкової вартості професійних гравців на основі власної методології та комбінації спортивних, фінансових та ринкових показників. Такий підхід корисний для комплексного аналізу футбольного ринку, але алгоритми платформи є

комерційними та закритими, тому користувач не має повного доступу до логіки формування остаточної оцінки.

Приклад аналітичного інтерфейсу платформи Football Benchmark подано на рисунку 1.5.

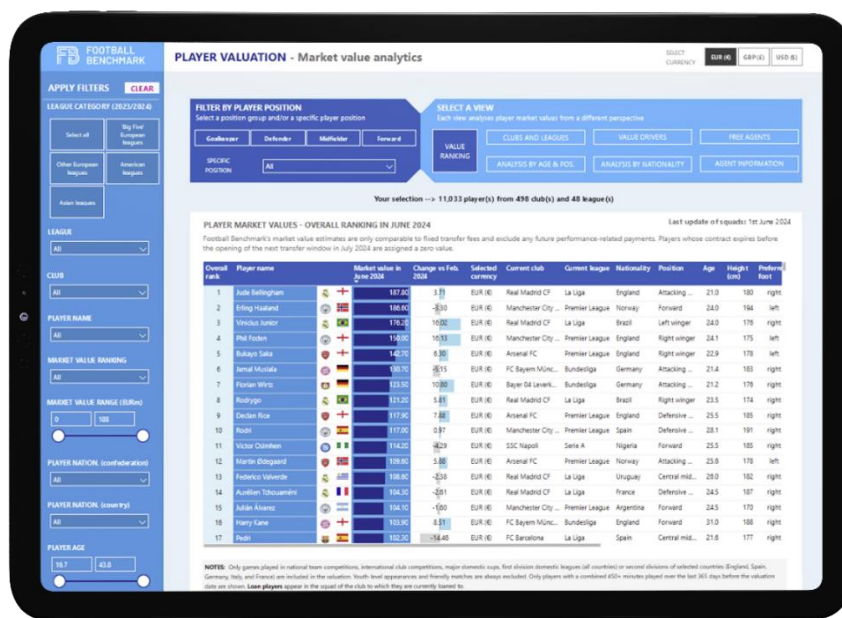


Рисунок 1.5 – Аналітичний інтерфейс оцінювання гравця на платформі Football Benchmark

Окрему групу утворюють платформи, які не завжди надають готову оцінку ринкової вартості футболіста, але забезпечують статистичне підґрунтя для подальшого аналізу. До таких ресурсів можна зарахувати FBref та StatsBomb, які містять докладну інформацію про матчі, команди, футболістів, ігрові дії та окремі статистичні показники [11], [12]. Такі дані мають важливе значення для побудови моделей машинного навчання, адже якість прогнозування великою мірою залежить від повноти, структурованості та достовірності вхідної інформації.

У наукових дослідженнях проблему оцінки ринкової вартості футболістів часто розглядають як задачу регресійного прогнозування. Для

її вирішення використовуються моделі лінійної регресії, ансамблеві методи, випадковий ліс, градієнтний бустинг та інші алгоритми машинного навчання. Такі підходи дозволяють аналізувати зв'язок між характеристиками футболіста та його ринковою вартістю, а також виявляти складні нелінійні залежності між ознаками.

Водночас, значна частина існуючих підходів зосереджена переважно на підвищенні точності прогнозу та менше уваги приділяє інтерпретації результатів. Для практичного застосування в галузі футбольного менеджменту важливо не лише отримати числову оцінку значення, а й пояснити, які характеристики гравця мали найбільший вплив на прогноз. Саме тому перспективним напрямком є поєднання моделей машинного навчання з пояснювальними методами штучного інтелекту.

Отже, аналіз існуючих платформ та методів показує, що поточні рішення мають як переваги, так і обмеження. Transfermarkt забезпечує широке охоплення та доступність історичних даних, але значною мірою покладається на експертні та краудсорсингові оцінки. Сервіси CIES Football Observatory та Football Benchmark пропонують більш формалізовані підходи, але їхні алгоритми залишаються недостатньо відкритими. Статистичні платформи надають цінні дані, але не завжди вирішують проблему оцінки ринкової вартості. Це вимагає розробки інтелектуальної системи, яка створює відкриті футбольні дані, методів машинного навчання та інтерпретації результатів прогнозування.

1.4 Інноваційність та переваги розроблюваної системи

Аналіз предметної галузі, факторів, що формують ринкову вартість футболістів, та існуючих платформ оцінювання показав, що сучасні рішення мають низку суттєвих обмежень. Деякі ресурси спираються на експертні або краудсорсингові оцінки, що може підвищити суб'єктивність результатів. Інші платформи використовують більш формалізовані аналітичні підходи,

але їхні алгоритми залишаються недостатньо відкритими для користувачів. Деякі статистичні ресурси надають цінні футбольні дані, але не завжди безпосередньо вирішують проблему прогнозування ринкової вартості.

Інноваційність розробленої системи полягає в об'єднанні трьох компонентів: автоматизованої обробки відкритих футбольних даних, застосування методів машинного навчання для прогнозування ринкової вартості та інтерпретації отриманих результатів. Такий підхід дозволяє не лише сформулювати приблизну оцінку вартості футболіста, але й пояснити, які характеристики були найбільше враховані в результаті.

На відміну від підходів, що повністю базуються на експертних оцінках або закритих комерційних алгоритмах, розроблена система використовує структуровані статистичні, особистісні, позиційні та контекстуальні характеристики футболістів. До них належать вік, зріст, громадянство, основна позиція, детальна позиція, ліга, сезон, загальний ігровий час, голи, результативні передачі, дисциплінарні показники та кількість турнірів. Використання таких характеристик дозволяє побудувати модель на основі фактичних даних про виступи футболістів, а не лише на суб'єктивних оцінках.

Значною перевагою системи є формування набору даних на рівні «гравець-сезон». Такий підхід дозволяє розглядати футболіста не як окремий статичний об'єкт, а як учасника певного сезону з відповідними статистичними показниками, ігровою активністю та ринковою оцінкою. Це робить дані більш придатними для навчання моделей машинного навчання, оскільки кожен запис містить узагальнену інформацію про виступи гравця протягом певного сезону.

Ще однією перевагою є використання кількох моделей регресійного машинного навчання. У роботі передбачено навчання та порівняння методів Ridge Regression, Random Forest та XGBoost. Це дозволяє оцінити ефективність як лінійної моделі, так і складніших ансамблевих методів, які можуть враховувати нелінійні залежності між ознаками. Доцільно звернути

особливу увагу на модель XGBoost, оскільки градієнтний бустинг ефективно працює з табличними даними та часто демонструє високу якість у завданнях регресійного прогнозування.

Значною перевагою розробленої системи також є коректний підхід до оцінки якості моделей. Для зменшення ризику витоку даних було використано групування за ідентифікатором футболіста. Це запобігає ситуації, коли записи одного й того ж гравця одночасно потрапляють до навчальної та тестової вибірок. Завдяки цьому експеримент стає більш виправданим та дозволяє більш реалістично оцінити здатність моделі працювати з новими, раніше небаченими футболістами.

Окремим інноваційним елементом є використання SHAP-інтерпретації для пояснення прогнозів. Для практичного використання в спортивній аналітиці недостатньо просто отримати числову оцінку ринкової вартості. Користувачеві важливо розуміти, які характеристики збільшують або зменшують прогнозовану вартість футболіста. SHAP дозволяє оцінити внесок окремих характеристик у результат моделі та зробити прогноз більш прозорим для користувача.

Додатковою перевагою є те, що результат представлений не лише у вигляді одного числа, а й у вигляді ймовірного діапазону значень. Для цього використовується квантильна регресія, яка формує інтервал, ширина якого відображає рівень невизначеності оцінки для конкретного гравця. Такий підхід відповідає реальному сценарію аналітика, який використовує оцінку як орієнтир для переговорів, а не як кінцеву суму, і вигідно відрізняє систему від ресурсів, що надають лише одну оцінку без міри її надійності.

Практичною перевагою системи є реалізація прототипу програмного забезпечення у вигляді вебзастосунку. Такий формат дозволяє користувачеві ввести характеристики футболіста, отримати прогноз його приблизної ринкової вартості, побачити можливий діапазон оцінки та проаналізувати пояснення результату. Завдяки цьому систему можна

використовувати не лише як дослідницьку модель, але й як демонстраційний інструмент для підтримки аналітичних рішень.

Водночас, розроблена система не замінює експертну оцінку скаута, спортивного директора чи футбольного аналітика. Її метою є надання додаткової аналітичної підтримки на основі відкритих даних та методів машинного навчання. Остаточне рішення щодо цінності футболіста також має враховувати контрактну ситуацію, фінансові можливості клубів, вплив ЗМІ, схильність до травм, психологічні характеристики та інші фактори, які не завжди доступні у структурованих наборах відкритих даних.

Отже, перевага розробленої системи полягає в поєднанні прогнозування, інтерпретації та програмної реалізації. Вона дозволяє автоматизувати первинний аналіз футболістів, зменшити залежність від суб'єктивних оцінок, пояснити вплив індивідуальних характеристик на прогноз та закласти основу для подальшого розвитку інтелектуальних систем у сфері спортивної аналітики.

1.5 Постановка задачі дослідження

Аналіз предметної галузі показав, що оцінка ринкової вартості футболістів є складним багатофакторним завданням. На цінність гравця впливають особистісні, позиційні, статистичні та контекстуальні характеристики, а також низка зовнішніх факторів, які не завжди доступні у відкритих джерелах. Крім того, існуючі платформи оцінювання мають певні обмеження, пов'язані з опорою на експертні судження, закритими алгоритмами або недостатньою інтерпретованістю результатів.

В рамках цієї кваліфікаційної роботи дослідницьким завданням є створення інтелектуальної системи оцінки приблизної ринкової вартості футболістів на основі відкритих структурованих футбольних даних з використанням методів машинного навчання. Система повинна

забезпечувати не лише формування прогностної оцінки, але й пояснення впливу індивідуальних характеристик футболіста на отриманий результат.

Вхідними даними для вирішення задачі є статистичні, особистісні, позиційні та контекстуальні характеристики футболістів. До них належать сезон, вік, зріст, громадянство, основна позиція, детальна позиція, ліга, загальний ігровий час, кількість голів, результативних передач, жовтих та червоних карток, а також кількість турнірів, у яких футболіст брав участь протягом сезону. Цільовою змінною є ринкова вартість футболіста в євро, яка розглядається як приблизна аналітична оцінка, а не як точна сума майбутнього трансферу.

Для досягнення мети роботи необхідно виконати такі завдання:

- проаналізувати предметну галузь оцінювання ринкової вартості футболістів і визначити основні чинники, що впливають на формування вартості гравця;
- розглянути наявні платформи та методи оцінювання ринкової вартості футболістів, окреслити їхні переваги й обмеження;
- сформувати набір даних рівня «гравець-сезон» на основі відкритих футбольних даних;
- здійснити попереднє опрацювання даних, зокрема очищення, узгодження таблиць, обробку пропущених значень і підготовку ознак для машинного навчання;
- побудувати регресійні моделі машинного навчання для прогнозування ринкової вартості футболістів;
- виконати навчання та зіставлення моделей Ridge Regression, Random Forest і XGBoost за метриками MAE, RMSE та R^2 ;
- застосувати групове розбиття даних за ідентифікатором футболіста для зниження ризику витоку даних між навчальною та тестовою вибірками;
- реалізувати оцінювання діапазону можливої ринкової вартості футболіста за допомогою квантильних моделей;

- застосувати SHAP-інтерпретацію для пояснення впливу окремих ознак на результат прогнозування;
- розробити програмний прототип інтелектуальної системи у вигляді вебзастосунку для введення характеристик футболіста, одержання прогнозу ринкової вартості, перегляду діапазону оцінки та пояснення результату.

Очікуваним результатом роботи є інтелектуальна система, яка поєднує підготовлений набір футбольних даних, моделі машинного навчання, інструменти оцінки невизначеності та інструменти інтерпретації прогнозів. Така система може служити допоміжним аналітичним інструментом для оцінки футболістів та підтримки прийняття рішень у сфері спортивної аналітики, а також створювати основу для подальшого порівняння гравців за статистичними характеристиками.

Важливою особливістю завдання є те, що система не претендує на точне визначення фактичної суми трансферу. Прогнозована ринкова вартість інтерпретується як приблизна оцінка, сформована на основі доступних статистичних та контекстуальних ознак. Такий підхід дозволяє зменшити суб'єктивність первинного аналізу, але не усуває необхідності експертної оцінки з урахуванням контрактних, медійних, фінансових та інших нестатистичних факторів.

Отже, завдання полягає у створенні практично застосовного прототипу інтелектуальної системи, яка поєднує прогнозування ринкової вартості футболістів за допомогою методів машинного навчання, оцінку можливого діапазону значень та пояснення отриманих результатів. Це забезпечує зв'язок між теоретичним аналізом предметної галузі та подальшою практичною реалізацією системи.

2 ОПИС МЕТОДІВ МАШИННОГО НАВЧАННЯ

2.1 Задача регресійного прогнозування ринкової вартості футболістів

Оцінка ринкової вартості футболістів на основі статистичних, особистісних та контекстуальних характеристик є задачею машинного навчання з учителем [13], [14]. У таких задачах модель навчається на прикладах, для яких відомі як вхідні ознаки, так і правильне цільове значення. У цій роботі вхідними даними є характеристики футболіста, а цільовою змінною – його оцінена ринкова вартість.

За типом вихідної змінної задача є регресійною. На відміну від класифікації, де результатом є віднесення об'єкта до певного класу, в регресійному прогнозуванні результатом моделі є числове значення. У цій роботі таким числовим значенням є прогнозована ринкова вартість футболіста в євро.

Навчальний набір даних у задачі регресії подають як множину пар:

$$D = \{(x_i, y_i)\}_{i=1}^n, \quad (2.1)$$

де D – навчальний набір даних;

x_i – вектор вхідних ознак i -го об'єкта;

y_i – відоме значення цільової змінної для i -го об'єкта;

n – кількість об'єктів у наборі даних.

У завданнях машинного навчання важливим є правильне представлення об'єкта через набір ознак, оскільки саме ознаки є основою для пошуку закономірностей у даних. Якщо вхідні характеристики вибрані недостатньо повно або представлені у незручному для моделі вигляді, якість прогнозування може знизитися. Тому перед навчанням регресійної моделі кожен об'єкт необхідно формально представити як упорядкований вектор ознак.

Вектор x_i охоплює числові та категоріальні характеристики об'єкта, які модель використовує для формування прогнозу. У загальному випадку кожен такий вектор може складатися з m ознак:

$$x_i = (x_{i1}, x_{i2}, \dots, x_{im}), \quad (2.2)$$

де x_i – вектор ознак i -го об'єкта;

$x_{i1}, x_{i2}, \dots, x_{im}$ – окремі ознаки об'єкта;

m – кількість ознак, задіяних для прогнозування.

Метою навчання регресійної моделі є побудова функції, яка встановлює зв'язок між вхідними ознаками та цільовою змінною. Загалом, прогноз моделі представлений наступним чином:

$$\hat{y}_i = f(x_i), \quad (2.3)$$

де $\hat{y}_i$ – прогнозоване значення ринкової вартості для i -го об'єкта;

f – функція, що визначається моделлю машинного навчання;

x_i – вектор вхідних ознак i -го об'єкта.

Функція f не задається вручну – вона формується під час навчання моделі на історичних даних. У процесі навчання алгоритм прагне виявити закономірності між характеристиками футболістів та значеннями їхньої ринкової вартості. Згодом отриману модель можна використовувати для прогнозування вартості нових об'єктів, для яких відомі лише вхідні ознаки.

Особливістю завдання оцінки ринкової вартості футболістів є нерівномірний розподіл цільової змінної. Більшість футболістів мають відносно низьку або середню ринкову вартість, тоді як лише невелика частина гравців оцінюється в десятки мільйонів євро. Через це розподіл цільової змінної є асиметричним і зміщеним у бік менших значень, а окремі дуже великі значення можуть суттєво впливати на процес навчання моделі та зміщувати її оцінки.

Щоб послабити вплив великих значень і стабілізувати розподіл цільової змінної, доцільно застосовувати логарифмічне перетворення. У роботі використано перетворення $\log 1p$, що має такий вигляд:

$$y' = \ln(1 + y), \quad (2.4)$$

де y' – перетворене значення цільової змінної;

y – початкове значення ринкової вартості футболіста;

$\ln$ – натуральний логарифм.

Застосування перетворення $\ln(1 + y)$ зручне для практичної обробки даних, оскільки воно коректно працює з невід'ємними значеннями та зменшує масштаб великих чисел. Після цього модель вчиться прогнозувати не саму початкову ринкову вартість, а її логарифмічно перетворене значення. Це робить навчання більш стабільним та зменшує вплив окремих дорогих гравців на модель.

Одержавши прогноз у логарифмічному масштабі, необхідно виконати зворотне перетворення, щоб повернути результат у початкову одиницю вимірювання. Зворотне перетворення має вигляд:

$$\hat{y} = \exp(\hat{y}') - 1, \quad (2.5)$$

де $\hat{y}$ – прогнозоване значення ринкової вартості у початковому масштабі;

$\hat{y}'$ – прогнозоване значення у логарифмічному масштабі;

$\exp$ – експоненціальна функція.

Таким чином, під час навчання модель працює зі стабілізованою зміною цільового показника, але кінцевий результат видається користувачам у зрозумілій формі – як приблизна ринкова вартість футболіста в євро. Це важливо для практичного застосування системи, оскільки користувач повинен підтримувати результат у форматі, загальноприйнятому для футбольного ринку.

Слід наголосити, що прогнозована ринкова вартість не є точною сумою майбутнього трансферу. Реальна ціна трансферу футболіста може залежати від контрактної ситуації, фінансових досягнень клубів, ходу переговорів, висвітлення гравця в ЗМІ, попиту на певну позицію та інших зовнішніх факторів. Тому регресійна модель в рамках цієї роботи виступає інструментом для формування приблизної аналітичної оцінки на основі наявних даних.

Отже, задача оцінки ринкової вартості футболістів формалізується як задача регресійного прогнозування, в якій модель навчається встановлювати зв'язок між вхідними характеристиками гравця та числовим значенням його ринкової вартості. Використання логарифмічного перетворення цільової змінної дозволяє зменшити вплив великих значень та підготувати задачу до подальшого використання регресійних моделей машинного навчання.

2.2 Регресійні моделі машинного навчання

Для вирішення задачі прогнозування ринкової вартості футболістів доцільно використовувати моделі регресійного машинного навчання. Такі моделі встановлюють зв'язок між вхідними ознаками об'єкта та неперервною числовою цільовою змінною. У цій роботі цільовою змінною є оціночна ринкова вартість футболіста, а вхідними даними – набір характеристик, що описують гравця та його результати.

Моделі регресії різняться за складністю. Найпростішими є лінійні моделі, які припускають наявність прямого лінійного зв'язку між ознаками та цільовою змінною. Складніші моделі, такі як ансамблеві методи, здатні враховувати нелінійні зв'язки, взаємодію між ознаками та складні приховані закономірності в даних, недоступні для лінійних підходів. У цій роботі для порівняння різних за складністю підходів розглядаються Ridge Regression, випадковий ліс та XGBoost.

Основний підхід до задачі регресії – це лінійна регресія. Вона дає прогнозоване значення як лінійну комбінацію вхідних ознак. Загалом, модель лінійної регресії можна записати так:

$$\hat{y}_i = w_0 + \sum_{j=1}^m w_j x_{ij}, \quad (2.6)$$

де $\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта;

w_0 – вільний член моделі;

w_j – ваговий коефіцієнт j -ї ознаки;

x_{ij} – значення j -ї ознаки для i -го об'єкта;

m – кількість ознак.

Лінійну регресію легко зрозуміти та інтерпретувати, але в задачах з великою кількістю ознак вона може бути вразливою до мультиколінеарності та перенавчання. Для пом'якшення цих проблем застосовується регуляризація, яка додає штраф до функції втрат за надмірно великі значення коефіцієнтів моделі.

Ridge Regression є різновидом лінійної регресії з L2-регуляризацією. Такий підхід дозволяє обмежити значення вагових коефіцієнтів та зробити модель більш стійкою до шуму в даних.

У задачах регресійного прогнозування регуляризація є важливим засобом підвищення стійкості моделі. Вона дозволяє обмежити надмірне зростання вагових коефіцієнтів, що особливо важливо за великої кількості ознак або зв'язків між ними. Завдяки цьому модель менше адаптується до окремих ознак навчального набору даних і краще узагальнює закономірності для нових об'єктів. Для задачі оцінки ринкової вартості футболістів це має практичне значення, оскільки окремі ознаки можуть бути взаємопов'язані. Наприклад, ігровий час, кількість матчів, результативність та участь у турнірах можуть одночасно характеризувати рівень залученості футболіста до гри. За таких умов звичайна лінійна модель може надавати

надмірну вагу окремим коефіцієнтам, тоді як гребенева регресія згладжує цей ефект і робить модель більш стабільною.

Функцію втрат Ridge Regression можна подати у вигляді:

$$L(w) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^m w_j^2, \quad (2.7)$$

де $L(w)$ – функція втрат моделі;

y_i – фактичне значення цільової змінної для i -го об'єкта;

$\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта;

λ – параметр регуляризації;

w_j – ваговий коефіцієнт j -ї ознаки;

n – кількість об'єктів у навчальному наборі;

m – кількість ознак.

Параметр λ визначає силу регуляризації. При малих значеннях модель наближається до звичайної лінійної регресії. Зі збільшенням λ вагові коефіцієнти обмежені більш жорстко, що може зменшити ризик перенавчання. У задачі оцінки ринкової вартості футболістів, регресія Ріджа доречна для використання як базова модель, з якою порівнюються складніші алгоритми.

Наступним підходом є Random Forest, або випадковий ліс. Це метод ансамблю машинного навчання, який складається з багатьох дерев рішень. Кожне дерево робить свій власний прогноз, а кінцевий результат моделі формується шляхом усереднення прогнозів усіх дерев.

Особливістю випадкового лісу є використання випадковості під час побудови окремих дерев. Для кожного дерева може використовуватися окрема підвибірка навчальних даних, сформована шляхом випадкового відбору прикладів, а під час розгалуження враховується лише частина доступних ознак, а не весь їх набір. Завдяки цьому дерева в ансамблі відрізняються одне від одного й не повторюють одні й ті самі помилки, а кінцеве усереднення їхніх прогнозів дозволяє зменшити ризик

перенавчання та підвищити робастність моделі. Такий підхід робить випадковий ліс менш чутливим до окремих нетипових спостережень і забезпечує стабільніші результати порівняно з одиночним деревом рішень. Для задачі регресії прогноз випадкового лісу подають так:

$$\hat{y} = 1/K \sum_{k=1}^K h_k(x), \quad (2.8)$$

де $\hat{y}$ – прогнозоване значення цільової змінної;

K – кількість дерев в ансамблі;

$h_k(x)$ – прогноз k -го дерева рішень;

x – вектор вхідних ознак об'єкта.

Перевагою Random Forest є здатність опрацьовувати нелінійні залежності та взаємодію ознак без потреби явно задавати ці зв'язки. Метод також доволі стійкий до шуму, оскільки підсумковий прогноз формує не одне дерево, а ансамбль багатьох дерев. Водночас випадковий ліс може поступатися сучаснішим методам градієнтного бустингу в задачах, де важливо максимально підвищити якість прогнозування на табличних даних.

Для покращення якості прогнозування в роботі також розглядається XGBoost. Цей алгоритм належить до методів градієнтного бустингу над деревами рішень. На відміну від Random Forest, де дерева будуються переважно незалежно одне від одного, при бустингу кожна наступна модель навчається з урахуванням помилок попередніх. Завдяки цьому ансамбль поступово покращує прогноз.

Важливою перевагою градієнтного бустингу є послідовний характер ансамблевого навчання. Кожна нова базова модель не просто додається до вже існуючих, а цілеспрямовано будується так, щоб зменшити помилки, які залишилися від попередніх кроків навчання. Завдяки цьому кінцевий ансамбль може поступово уточнювати прогноз, виправляти власні неточності та значно краще враховувати складні нелінійні залежності між ознаками.

У завданні оцінки ринкової вартості футболістів ця властивість є важливою, оскільки вплив індивідуальних характеристик гравця може проявлятися не ізольовано, а у взаємодії з іншими ознаками. Наприклад, значення віку, позиції чи результативності може по-різному впливати на прогноз залежно від рівня ліги та ігрової активності гравця.

У загальному вигляді прогноз градієнтного бустингу можна подати як суму прогнозів окремих базових моделей:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad (2.9)$$

де $\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта;

K – кількість базових моделей в ансамблі;

f_k – k -та базова модель;

x_i – вектор вхідних ознак i -го об'єкта.

XGBoost розширює класичний градієнтний бустинг завдяки регуляризації, ефективній оптимізації, обробці пропущених значень і контролю складності моделі. Тому алгоритм використовується для задач регресії на структурованих табличних даних [15]. У контексті оцінки ринкової вартості футболістів XGBoost є виправданим вибором, оскільки зв'язок між характеристиками гравця та його вартістю може бути нелінійним та залежати від взаємодії багатьох ознак.

Порівняння Ridge-регресії, випадкового лісу та XGBoost дозволяє нам оцінити ефективність різних класів регресійних моделей на одній і тій самій задачі. Ridge-регресія виступає як базова лінійна модель, випадковий ліс являє собою ансамблевий підхід, заснований на незалежних деревах рішень, а XGBoost дозволяє нам дослідити можливості градієнтного бустингу з послідовним навчанням. Такий набір моделей добре підходить для задачі прогнозування ринкової вартості футболістів, оскільки дозволяє нам порівнювати як простіші, так і складніші підходи до моделювання й обрати найефективніший із них.

Отже, для вирішення проблеми доцільно використовувати кілька регресійних моделей та порівняти їхню якість. Остаточний вибір моделі повинен ґрунтуватися не лише на теоретичних перевагах алгоритму, а й на результатах експериментальної оцінки.

2.3 Метрики оцінювання якості моделей

Для порівняння моделей регресійного машинного навчання необхідно використовувати кількісні метрики оцінки якості. Такі метрики дозволяють визначити, наскільки прогнозовані значення відрізняються від фактичних, а також порівнювати різні алгоритми за спільними критеріями. У задачі оцінки ринкової вартості футболістів це особливо важливо, оскільки навіть моделі одного типу можуть по-різному реагувати на великі значення, викиди та нерівномірний розподіл цільової змінної.

Якість регресійної моделі оцінюється порівнянням фактичних значень цільової змінної з прогнозованими: що ближчий прогноз до реального значення, то меншою є похибка, і навпаки. Тому вибір метрик повинен враховувати як середній розмір похибки, так і чутливість до великих відхилень.

Однією з найпоширеніших метрик у задачах регресії є середня абсолютна похибка, або MAE. Вона відображає середнє абсолютне відхилення прогнозованих значень від фактичних значень. Формула MAE має вигляд:

$$MAE = 1/n \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2.10)$$

де MAE – середня абсолютна похибка;

n – кількість об'єктів у вибірці;

y_i – фактичне значення цільової змінної для i -го об'єкта;

$\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта.

Перевагою MAE є простота інтерпретації. Значення цієї метрики вимірюється в тих самих одиницях, що й цільова змінна. У контексті цієї роботи MAE показує, наскільки в середньому прогнозована ринкова вартість футболіста відрізняється від фактичної оцінки. Водночас MAE однаково враховує всі похибки, незалежно від їх напрямку, оскільки використовує модуль різниці між фактичним та прогнозованим значенням.

Водночас, MAE не надає підвищеної ваги великим помилкам. Це означає, що великі відхилення впливають на кінцеве значення метрики пропорційно своєму розміру. У завданні оцінки ринкової вартості футболістів це може бути як перевагою, так і обмеженням. З одного боку, метрика більш стійка до окремих великих відхилень, з іншого боку, вона може недостатньо підкреслювати помилки для дуже дорогих гравців.

Ще однією важливою метрикою є середньоквадратична помилка, або RMSE. На відміну від MAE, ця метрика зводить помилки в квадрат, тому більші відхилення мають більший вплив на результат. Формула RMSE має вигляд:

$$\text{RMSE} = \sqrt{(1/n \sum_{i=1}^n (y_i - \hat{y}_i)^2)}, \quad (2.11)$$

де RMSE – корінь із середньоквадратичної похибки;

n – кількість об'єктів у вибірці;

y_i – фактичне значення цільової змінної для i -го об'єкта;

$\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта.

RMSE є корисною метрикою тоді, коли потрібно ретельніше враховувати саме великі помилки прогнозування. У задачі оцінки ринкової вартості футболістів це особливо важливо, оскільки помилка для гравця з високою ринковою вартістю може бути набагато більшою в абсолютному вираженні, ніж помилка для гравця з нижчою вартістю. Завдяки квадратичному штрафу за відхилення RMSE дозволяє краще виявляти ті моделі, які припускають великі відхилення в окремих прикладах.

Водночас, RMSE чутливіший до викидів, ніж MAE. Якщо у вибірці є окремі об'єкти з дуже високою ринковою вартістю або нетиповими характеристиками, вони можуть суттєво впливати на значення цієї метрики. Тому доцільно використовувати MAE та RMSE разом: перша метрика показує середній абсолютний рівень похибки, а друга дозволяє оцінити наявність великих відхилень у прогнозах.

Для оцінки частки поясненої дисперсії цільової змінної забезпечує визначення коефіцієнта R^2 . Ця метрика показує, наскільки добре модель пояснює варіативність фактичних значень порівняно з простим прогнозуванням середнього значення. Формула коефіцієнта детермінації має вигляд:

$$R^2 = 1 - (\sum_{i=1}^n (y_i - \hat{y}_i)^2) / (\sum_{i=1}^n (y_i - \bar{y})^2), \quad (2.12)$$

де R^2 – коефіцієнт детермінації;

y_i – фактичне значення цільової змінної для i -го об'єкта;

$\hat{y}_i$ – прогнозоване значення цільової змінної для i -го об'єкта;

$\bar{y}$ – середнє значення цільової змінної;

n – кількість об'єктів у вибірці.

Значення R^2 зазвичай інтерпретується як частка варіації цільової змінної, яка пояснюється моделлю. Чим ближче R^2 до 1, тим краще модель відтворює взаємозв'язки в даних. Значення, близьке до 0, означає, що модель не працює суттєво краще, ніж прогнозування середнього значення. У деяких випадках R^2 може бути від'ємним, якщо модель працює гірше, ніж проста базова оцінка.

У задачі прогнозування ринкової вартості футболістів коефіцієнт детермінації корисний для загального розуміння якості моделі, але його не слід розглядати ізольовано. Високе значення R^2 не завжди свідчить про те, що абсолютна похибка прогнозу мала з практичної точки зору. Тому для більш повної оцінки якості доцільно поєднувати R^2 з MAE та RMSE.

Використання кількох метрик дозволяє збалансовано оцінити ефективність моделі. MAE дозволяє оцінити середню похибку в одиницях зміни цільового показника, RMSE краще реагує на великі відхилення, а R^2 показує, яка частина мінливості даних пояснюється моделлю. Разом ці метрики дозволяють не лише порівнювати моделі, але й краще розуміти природу їхніх помилок.

Отже, для оцінки якості регресійних моделей у рамках цієї роботи доцільно використовувати MAE, RMSE та R^2 . Для практичного розрахунку цих метрик і подальшого порівняння моделей може застосовуватися інструментарій бібліотеки `scikit-learn` [16]. Цей набір метрик дозволяє провести комплексний аналіз точності прогнозування, чутливості моделей до великих помилок та можливості пояснення мінливості ринкової вартості футболістів. Саме на основі цих показників далі порівнюються Ridge Regression, Random Forest та XGBoost.

2.4 Квантильна регресія для оцінювання невизначеності прогнозу

У задачах регресійного прогнозування модель зазвичай видає одне числове значення, яке інтерпретується як найімовірніша оцінка цільової змінної. Однак у багатьох прикладних задачах цього результату недостатньо, оскільки користувачеві важливо знати не лише очікуване значення прогнозу, але й рівень його невизначеності. Це особливо актуально для задачі оцінки ринкової вартості футболістів, де кінцева вартість гравця залежить від багатьох факторів і не може бути визначена з повною точністю.

Ринкова вартість футболіста – це складний показник, на який впливають не лише статистичні характеристики, а й зовнішні ринкові умови, контрактна ситуація, попит на певну позицію, фінансові можливості клубів та інші фактори. Деякі з цих факторів можуть бути відсутніми у відкритих наборах даних або не мати точного кількісного представлення.

Тому доцільно сприймати прогноз моделі не як остаточну суму, а як приблизну аналітичну оцінку.

Квантильна регресія може бути використана для представлення невизначеності прогнозу. На відміну від звичайної регресії, яка переважно оцінює умовне середнє значення цільової змінної, квантильна регресія дозволяє оцінювати окремі квантілі її розподілу. Це дозволяє формувати не лише центральний прогноз, але й нижню та верхню межі можливого діапазону значень.

Квантиль рівня τ можна розглядати як значення, нижче якого знаходиться задана частка спостережень. Наприклад, квантиль рівня 0,1 відповідає нижній частині розподілу, квантиль рівня 0,5 – медіані, а квантиль рівня 0,9 – верхній частині розподілу. Конкретні рівні квантилів для нижньої та верхньої меж вибираються залежно від задачі та бажаної ширини інтервалу прогнозу (зокрема, квантілі 0,25 та 0,75 використовуються для формування 50-відсоткового інтервалу). У задачі оцінки ринкової вартості це дозволяє представити результат не одним числом, а як можливий інтервал, у межах якого може знаходитися оцінена вартість футболіста.

Основою квантильної регресії є асиметрична функція втрат, яка по-різному карає за переоцінку та недооцінку залежно від обраного рівня квантиля. Для квантиля τ функцію втрат можна записати як:

$$\rho_{\tau}(u) = u(\tau - I(u < 0)), \quad (2.13)$$

де $\rho_{\tau}(u)$ – функція втрат для квантіля рівня τ ;

u – похибка прогнозування;

τ – рівень квантіля, $0 < \tau < 1$;

$I(u < 0)$ – індикаторна функція, що дорівнює 1, якщо $u < 0$, і 0 в іншому разі.

Похибку прогнозування в цьому випадку визначають як різницю між фактичним і прогнозованим значенням:

$$u = y - \hat{y}, \quad (2.14)$$

де u – похибка прогнозування;

y – фактичне значення цільової змінної;

$\hat{y}$ – прогнозоване значення цільової змінної.

Особливість такої функції втрат полягає в тому, що для різних значень τ модель навчається прогнозувати різні частини умовного розподілу цільової змінної. Якщо τ менше 0,5, модель зосереджується на нижній межі можливих значень. Якщо τ більше 0,5, вона зосереджується на верхній межі. Для $\tau = 0,5$ квантильна регресія оцінює значення медіани.

Для формування інтервалу прогнозування можна застосувати дві квантильні моделі: одну – для нижньої межі, іншу – для верхньої. У загальному вигляді інтервал прогнозу записують так:

$$\hat{y} \in [q\alpha(x), q\beta(x)], \quad (2.15)$$

де $\hat{y}$ – прогнозоване значення цільової змінної;

$q\alpha(x)$ – нижній квантиль прогнозу для вхідного вектора ознак x ;

$q\beta(x)$ – верхній квантиль прогнозу для вхідного вектора ознак x ;

α, β – рівні квантілів, що задають межі інтервалу.

На практиці такий підхід дозволяє користувачеві бачити не лише одне прогнозоване значення, а й можливий діапазон ринкових значень. Наприклад, центральний прогноз може відображати найімовірнішу оцінку вартості гравця, тоді як нижня та верхня межі окреслюють прийнятний інтервал з урахуванням невизначеності моделі. Це робить результат значно інформативнішим для спортивного аналітика чи скаута, який ухвалює рішення в умовах неповної інформації.

Слід наголосити, що інтервал, згенерований квантильними моделями, не слід сприймати як гарантію фактичної трансфертної ціни. Він лише відображає діапазон можливих оцінок на основі закономірностей, які модель визначила в доступних даних. Якщо у вхідних даних відсутні важливі ринкові або договірні фактори, інтервал може не враховувати всі реальні умови майбутнього трансферу.

Перевага квантильної регресії полягає в тому, що вона краще відповідає характеру задачі оцінки цінності футболістів. У реальній аналітичній практиці цінність гравця часто розглядається не як фіксоване число, а як певний діапазон, який може змінюватися залежно від переговорів, попиту, форми гравця та інших умов. Тому представлення прогнозу у вигляді інтервалу є більш гнучким та практично корисним.

Крім того, ширина прогнозного інтервалу може вказувати на рівень невизначеності в оцінці. Якщо інтервал вузький, це може свідчити про більшу впевненість моделі в прогнозі. Якщо інтервал широкий, модель має вищу невизначеність для конкретного об'єкта. У завданні оцінки ринкової вартості футболістів така інформація є важливою, оскільки дозволяє відрізнити більш стабільні прогнози від менш надійних.

Отже, квантильна регресія є корисним інструментом для оцінки невизначеності прогнозу. Вона дозволяє доповнити традиційне регресійне прогнозування інтервальною оцінкою, що підвищує практичну цінність інтелектуальної системи. Для задачі оцінки ринкової вартості футболістів це дозволяє представити результат не лише як одне числове значення, а й як діапазон можливих значень, що краще відповідає реальним умовам ринку футбольних трансферів.

2.5 Методи інтерпретації результатів прогнозування

Сучасні моделі машинного навчання здатні надавати високоякісні прогнози, але вони не завжди достатньо прозорі для користувача. Це

особливо стосується ансамблевих моделей та методів градієнтного бустингу, які можуть враховувати складні нелінійні залежності між ознаками. У таких випадках користувач отримує кінцеве прогнозоване значення, але не завжди розуміє, які ознаки мали найбільший вплив на результат.

Інтерпретованість має велике практичне значення для оцінки ринкової вартості футболістів. Спортивному аналітику чи скауту недостатньо знати лише прогнозовану вартість гравця. Важливо також розуміти, чому модель сформувала саме цю оцінку: чи вплинули на неї вік гравця, його позиція, рівень ліги, ігровий час, результативність чи інші характеристики. Без такого пояснення практична цінність системи знижується.

Одним із поширених підходів до пояснення результатів моделей машинного навчання є SHAP [17]. Цей метод базується на ідеї оцінки внеску кожної ознаки в прогноз моделі. SHAP дозволяє визначити, наскільки окрема ознака збільшила або зменшила прогнозоване значення порівняно з базовим рівнем.

У загальному вигляді адитивну модель пояснення SHAP можна подати так:

$$g(z') = \varphi_0 + \sum_{j=1}^M \varphi_j z'_j, \quad (2.16)$$

де $g(z')$ – пояснювальна модель прогнозу;

φ_0 – базове значення прогнозу;

φ_j – внесок j -ї ознаки у результат прогнозування;

z'_j – бінарний показник наявності j -ї ознаки у спрощеному поданні;

M – кількість ознак, що враховуються в поясненні.

Суть такого представлення полягає в тому, що кінцевий прогноз розглядається як сума базового значення та внеску окремих ознак. Якщо значення SHAP для певної ознаки додатне, ця ознака збільшує прогнозоване ринкове значення. Якщо значення від'ємне, то ознака зменшує прогноз.

Завдяки цьому користувач бачить не лише кінцеву оцінку, але й логіку її формування.

SHAP можна використовувати для локальної та глобальної інтерпретації. Локальна інтерпретація пояснює прогноз для конкретного футболіста, показуючи, які характеристики мають найбільший вплив на його індивідуальну оцінку. Глобальна інтерпретація дозволяє оцінити важливість характеристик для моделі в цілому, тобто з'ясувати, які характеристики найчастіше мають найбільший вплив на прогнозування ринкової вартості.

В інтелектуальній системі локальна інтерпретація особливо корисна, коли користувач аналізує конкретного гравця. Наприклад, система може показати, що прогнозоване значення збільшується через молодий вік, значний ігровий час або результативність, але зменшується через менш конкурентну лігу або інші характеристики. Це робить прогноз більш зрозумілим і придатним для аналітичного використання.

Глобальна інтерпретація, у свою чергу, дозволяє оцінити загальну поведінку моделі. Вона допомагає зрозуміти, які групи ознак є найважливішими для прогнозування: особистісні характеристики, ігрова активність, продуктивність, позиція чи контекстуальні фактори. Такий аналіз корисний для перевірки адекватності моделі та виявлення можливих упереджень у даних.

Слід зазначити, що інтерпретація не замінює оцінку якості моделі за допомогою метрик MAE, RMSE та R^2 . Метрики показують, наскільки точно модель прогнозує значення, тоді як SHAP пояснює, чому було отримано певне передбачення. Тому ці підходи слід використовувати разом: кількісні метрики дозволяють оцінити точність, а методи інтерпретації підвищують прозорість результатів.

Отже, застосування SHAP пом'якшує проблему «чорної скриньки» та робить прогнози зрозумілішими: система надає не лише числову оцінку, а й пояснення впливу окремих характеристик футболіста на результат.

3 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ОЦІНЮВАННЯ РИНКОВОЇ ВАРТОСТІ ФУТБОЛІСТІВ

3.1 Загальна архітектура інтелектуальної системи

Розроблена інтелектуальна система призначена для оцінки приблизної ринкової вартості футболістів на основі відкритих футбольних даних та методів машинного навчання. Її основним завданням є формування прогнозової оцінки вартості гравця, визначення можливого діапазону цієї оцінки та пояснення впливу індивідуальних характеристик футболіста на отриманий результат.

Загальна архітектура системи побудована за модульним принципом. Це означає, що процес її роботи розділений на окремі логічні етапи: завантаження та підготовка даних, формування ознак, навчання моделей машинного навчання, оцінка невизначеності прогнозу, інтерпретація результатів та представлення кінцевої інформації користувачеві через вебзастосунок. Такий підхід спрощує розробку, тестування та подальше вдосконалення системи.

Основними компонентами інтелектуальної системи є модуль обробки даних, модуль попередньої обробки, модуль машинного навчання, модуль оцінки невизначеності, модуль інтерпретації прогнозів та модуль користувацького інтерфейсу. Кожен з цих компонентів виконує свою функцію, але всі вони взаємопов'язані в єдиний процес прогнозування ринкової вартості футболіста.

Взаємодія між модулями системи є послідовною: вихідні дані поступово проходять через усі компоненти, від завантаження вхідної інформації до формування кінцевого результату для користувача. При цьому кожен наступний модуль використовує результати попереднього, що забезпечує цілісність процесу прогнозування. Така організація дозволяє чітко розмежувати функції окремих компонентів, спростити їх незалежне

налаштування та забезпечити прозорість загального процесу оцінки ринкової вартості футболіста. Узагальнена взаємодія основних модулів системи та послідовність обробки даних схематично представлені на рисунку 3.1.

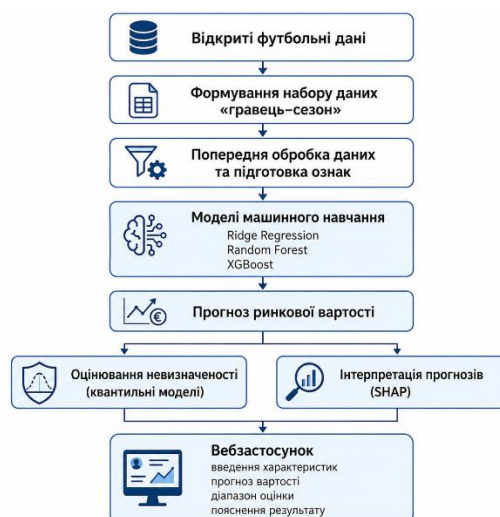


Рисунок 3.1 – Загальна архітектура інтелектуальної системи оцінювання ринкової вартості футболістів

Першим етапом роботи системи є формування набору вхідних даних. Для цього використовуються відкриті футбольні дані, які містять інформацію про футболістів, їхні особистісні характеристики, позиції, результати, статистичні показники та ринкову вартість. В рамках роботи дані організовані на рівні «гравець-сезон», що дозволяє аналізувати футболіста в контексті конкретного сезону та відповідних ігрових показників.

Другий етап – попередня обробка даних. На цьому етапі виконується очищення даних, зіставлення таблиць, обробка відсутніх значень, вибір ознак та підготовка цільових змінних. Оскільки ринкова вартість футболістів розподілена нерівномірно, для її стабілізації використовується логарифмічне перетворення. Це дозволяє зменшити вплив дуже великих значень та зробити навчання моделі більш надійним.

Після попередньої обробки дані передаються до модуля машинного навчання. Цей модуль реалізує навчання регресійних моделей, які прогнозують ринкову вартість футболістів на основі навчених ознак. У роботі розглядається кілька моделей різних типів: Ridge Regression, випадковий ліс та XGBoost. Це дозволяє порівняти лінійний підхід, ансамблеву модель на основі випадкового лісу та модель градієнтного бустингу.

Окремим компонентом системи є модуль оцінки невизначеності прогнозу. Його метою є формування не лише одного прогнозного значення, а й можливого діапазону ринкової вартості футболіста. Для цього використовується підхід, заснований на квантильних моделях, який дозволяє визначити нижню та верхню межі прогнозу. Такий результат є більш корисним для практичного аналізу, оскільки ринкова вартість у футболі не є абсолютно фіксованою величиною.

Для підвищення прозорості результатів система надає модуль інтерпретації прогнозу. Він служить для пояснення, які характеристики мали найбільший вплив на отриману оцінку ринкової вартості. Для цього використовується SHAP-інтерпретація, яка дозволяє визначити внесок окремих характеристик футболіста в прогноз моделі. Завдяки цьому користувач отримує не лише числовий результат, але й пояснення логіки його формування.

Заключним компонентом архітектури є вебдодаток, через який користувач взаємодіє із системою. У вебінтерфейсі можна ввести характеристики футболіста, виконати прогноз, отримати приблизну ринкову вартість, переглянути можливий діапазон оцінок та ознайомитися з поясненням впливу окремих функцій. Такий формат реалізації робить систему зручною для демонстрації та практичного застосування.

Загальний процес роботи системи можна описати як послідовний перехід від вхідних футбольних даних до аналітичних вихідних. Спочатку дані очищаються та перетворюються у форму, придатну для моделювання,

після чого на їх основі навчаються регресійні моделі. Далі система генерує прогноз, визначає можливий діапазон значень, пояснює результат за допомогою SHAP та надає користувачеві зведену інформацію у вебінтерфейсі.

Таким чином, запропонована архітектура поєднує етапи підготовки даних, машинного навчання, оцінки невизначеності, інтерпретації результатів та програмної реалізації. Така структура відповідає меті роботи, оскільки дозволяє не лише прогнозувати приблизну ринкову вартість футболістів, але й забезпечити прозорість та практичну застосовність отриманих результатів.

3.2 Формування та попередня обробка набору даних

Для побудови інтелектуальної системи оцінки ринкової вартості футболістів було використано відкриті футбольні дані, які містять інформацію про гравців, їхні особисті характеристики, ігрові показники, участь у змаганнях та ринкову вартість. Основною вимогою до набору даних була можливість представити кожного футболіста не лише як окремий об'єкт, а як учасника певного сезону з відповідними статистичними характеристиками.

В рамках роботи було сформовано набір даних рівня «гравець-сезон». Подібний підхід до подання футбольних даних є доцільним для задач машинного навчання, пов'язаних з оцінюванням вартості футболістів [18]. Такий підхід означає, що один запис у зведеній таблиці відповідає виступам конкретного футболіста протягом певного сезону. Це дає змогу враховувати зміни віку, ігрової активності, результативності, ліги та ринкової вартості гравця в різні періоди його кар'єри. На відміну від статичного представлення футболіста, рівень «гравець-сезон» краще підходить для задачі прогнозування, оскільки ринкова вартість гравця змінюється з часом.

Набір даних було створено за допомогою таблиць з інформацією про футболістів, їхні результати, клуби, ліги та ринкову вартість. Дані були узгоджені за ідентифікаторами гравців та сезонів, після чого були агреговані статистичні показники. В результаті для кожного запису було отримано набір ознак, що описує футболіста протягом певного сезону.

Цільовою змінною в роботі є ринкова вартість футболіста в євро. Щоб уникнути витoku інформації в часі, цільовим значенням для кожного сезону є остання оцінка ринкової вартості гравця в межах цього сезону – тобто оцінка, сформована після відповідних ігрових показників, а не до них. Вона використовується як приблизна аналітична оцінка, на основі якої навчаються регресійні моделі. Важливо зазначити, що така змінна не є фактичною сумою трансферної угоди, оскільки реальна трансферна ціна може залежати від договірних умов, переговорів між клубами, фінансової ситуації на ринку та інших зовнішніх факторів. Тому система прогнозує приблизну ринкову вартість, а не точну майбутню суму трансферу.

Вхідні ознаки включають особисті, позиційні, статистичні та контекстуальні характеристики футболістів. Особисті характеристики описують основну інформацію про гравця, включаючи вік, зріст та громадянство. Позиційні характеристики містять основну та детальну позицію футболіста на полі. Статистичні ознаки відображають ігрову активність та результативність: загальний ігровий час, кількість голів, результативних передач, жовті та червоні картки. Контекстуальні ознаки описують сезон, лігу та кількість турнірів, у яких брав участь футболіст.

Усі перелічені ознаки можна умовно розділити на кілька груп залежно від того, який аспект вони характеризують. Особистісні ознаки описують самого гравця, позиційні ознаки описують його роль на полі, статистичні ознаки описують його ігрову активність та результативність протягом сезону, а контекстуальні ознаки відображають умови, в яких завершилися виступи. Окремо виділено зміну цілі, яка є об'єктом прогнозування. Таке групування ознак спрощує розуміння структури набору даних і дозволяє

точно простежити, які характеристики футболіста використовує модель для формування оцінки його ринкової вартості.

Узагальнену структуру ознак, використаних для побудови моделей машинного навчання, наведено в таблиці 3.1.

Таблиця 3.1 – Групи ознак для прогнозування вартості футболістів

Група ознак	Приклади ознак	Призначення
Персональні	Вік, зріст, громадянство	Опис базових характеристик футболіста
Позиційні	Основна позиція, деталізована позиція	Урахування ігрової ролі
Статистичні	Ігровий час, голи, асисті, картки	Опис ігрової активності та результативності
Контекстні	Сезон, ліга, кількість турнірів	Урахування умов виступів
Цільова змінна	Ринкова вартість у євро	Значення, яке прогнозує модель

Після формування початкового набору даних було проведено попередню обробку. Цей етап є необхідним, оскільки відкриті футбольні дані можуть містити відсутні значення, різні формати представлення інформації, дублікати або ознаки, які не підходять для безпосереднього використання моделями машинного навчання. Якість попередньої обробки безпосередньо впливає на якість подальшого прогнозування.

Перший етап включав очищення даних. З набору видалялися записи, для яких була відсутня цільова змінна або неможливо було правильно визначити ринкову вартість футболіста. Також перевірялася сумісність типів даних, оскільки числові значення повинні бути представлені в

числовому форматі, а категоріальні характеристики повинні бути представлені окремим текстом або закодованими ознаками.

На другому етапі обробляються відсутні значення. Для числових ознак відсутні значення можуть бути замінені статистичними значеннями або оброблені залежно від характеру конкретної ознаки. Для категоріальних ознак доцільно замінити відсутні значення окремою категорією, що дозволяє зберегти запис у наборі даних та не втратити інформацію про інші характеристики футболіста.

На третьому етапі категоріальні ознаки готуються до використання в моделях машинного навчання. Оскільки більшість алгоритмів працюють з числовими значеннями, текстові характеристики, такі як посада, громадянство чи ліга, необхідно перетворити на числове представлення. Для цього використовуються методи кодування категоріальних змінних, які дозволяють зберігати інформацію про належність об'єкта до певної категорії.

Особлива увага приділяється цільовій змінній. Ринкова вартість футболістів має нерівномірний розподіл: більшість гравців мають відносно низьке або середнє значення, тоді як невелика частина футболістів оцінюється значно вище. Тому для стабілізації розподілу застосовується логарифмічне перетворення цільової змінної. Модель навчається прогнозувати перетворене значення, а після отримання результату виконується зворотнє перетворення для представлення прогнозу в євро.

Для правильної та об'єктивної оцінки якості моделей важливо коректно розділити наявні дані на навчальну та тестову вибірки. При звичайному випадковому розділенні може виникнути ситуація, коли записи одного й того ж футболіста з різних сезонів опиняються в навчальній та тестовій частинах одночасно. Це може призвести до витоку даних та завищеної переоцінки якості моделі, оскільки алгоритм фактично вже бачив інформацію про цього гравця під час тренування й запам'ятав пов'язані з ним закономірності, а отже, оцінка на тесті буде необ'єктивною.

Щоб зменшити ризик такого витоку, у роботі використовується груповий поділ за ID гравця. Це означає, що всі записи одного гравця потрапляють або до навчальної, або до тестової вибірки. Такий підхід дозволяє більш реалістично оцінити здатність моделі працювати з новими гравцями та краще відповідає практичному сценарію використання системи.

Процес формування та попередньої обробки набору даних можна представити як послідовність етапів: завантаження відкритих футбольних даних, зіставлення таблиць за ідентифікаторами, агрегування статистики до рівня «гравець-сезон», очищення даних, обробка відсутніх значень, підготовка числових та категоріальних ознак, перетворення цільової змінної та розділення даних на навчальну та тестову вибірки.

Фрагмент реалізації підготовки цільової змінної та розбиття даних на навчальну й тестову вибірки наведено в лістингу 3.1.

Лістинг 3.1 – Підготовка цільової змінної та розбиття даних на навчальну і тестову вибірки

```
# Логарифмічне перетворення цільової змінної

data["log_target_market_value"] =
np.log1p(data["target_market_value_eur"])
X = data[model_features]
y = data["log_target_market_value"]
groups = data["player_id"]

# Групове розбиття за гравцем (записи одного футболіста не
потрапляють

# одночасно до навчальної та тестової вибірок)
gss = GroupShuffleSplit(n_splits=1, test_size=0.2,
random_state=42)
train_idx, test_idx = next(gss.split(X, y, groups))
X_train, X_test = X.iloc[train_idx], X.iloc[test_idx]
y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]
```

Таким чином, в рамках розробки системи було сформовано структурований набір даних, придатний для навчання моделей регресійного машинного навчання. Попередня обробка дозволила привести дані до узгодженого формату, зменшити вплив пропусків та нерівномірного розподілу цільової змінної, а також підготувати ознаки для подальшого моделювання. Цей етап є основою для реалізації моделей прогнозування ринкової вартості футболістів.

3.3 Реалізація моделей машинного навчання

Після формування та попередньої обробки набору даних, наступним етапом розробки інтелектуальної системи стала реалізація моделей машинного навчання для прогнозування ринкової вартості футболістів. Основним завданням цього етапу було навчання кількох регресійних моделей, порівняння їхньої здатності відтворювати зв'язок між характеристиками футболіста та його ринковою вартістю, а також підготовка моделі, придатної для подальшого використання у вебзастосунку.

Моделі були реалізовані на Python з використанням бібліотек даних та машинного навчання. Pandas використовувався для обробки табличних даних, scikit-learn – для побудови та оцінки моделей, а XGBoost – для реалізації градієнтного бустингу. Цей набір інструментів підходить для роботи зі структурованими табличними даними, оскільки дозволяє поєднувати попередню обробку, навчання моделі, оцінку якості та подальше використання навчених моделей у прототипі програмного забезпечення.

У роботі реалізовано три основні регресійні моделі: Рідж-регресія, Випадковий ліс та XGBoost. Використання ансамблевих підходів у задачах прогнозування вартості футболістів також розглядається в сучасних дослідженнях футбольної аналітики [19]. Вибір цих моделей зумовлений

необхідністю порівняння різних підходів до прогнозування. Рідж-регресія використовується як базова лінійна модель з регуляризацією. Випадковий ліс дозволяє враховувати нелінійні залежності між ознаками завдяки ансамблю дерев рішень. XGBoost реалізує підхід градієнтного бустингу та здатний ефективно обробляти складні закономірності в табличних даних.

Моделі навчалися на підготовленому наборі ознак, отриманому після попередньої обробки даних. Вхідними даними для моделей є числові та кодовані категоріальні характеристики футболістів, а цільовою змінною – логарифмічно перетворена ринкова вартість. Використання перетвореної цільової змінної дозволяє послабити вплив дуже великих значень ринкової вартості та зробити процес тренування більш стабільним.

Для кожної моделі було проведено навчання на навчальній вибірці та згенеровано прогнози для тестової вибірки. Оскільки модель прогнозує значення в логарифмічному масштабі, після отримання прогнозу застосовується зворотне перетворення. Це дозволяє представити результат у вихідному масштабі, тобто в євро. Такий підхід важливий для практичного використання системи, оскільки користувач повинен отримати прогноз у зрозумілій формі.

Загальну послідовність реалізації моделі можна представити так: підготовка навчальної та тестової вибірок, навчання регресійної моделі, формування прогнозу в логарифмічній шкалі, зворотне перетворення прогнозу у вихідну шкалу, розрахунок показників якості та збереження результатів для подальшого порівняння. Ця послідовність застосовується до всіх моделей, що забезпечує коректність їх порівняння.

Першою реалізованою моделлю була Рідж-регресія. Вона використовується як базовий алгоритм, що дозволяє оцінити якість простішого лінійного підходу. Перевагою цієї моделі є її ясність та відносна стійкість до перенавчання завдяки L2-регуляризації. Водночас, через свою лінійну природу, така модель може не повністю враховувати складні

нелінійні зв'язки між характеристиками футболіста та його ринковою вартістю.

Друга – це модель випадкового лісу. Цей алгоритм використовує ансамбль дерев рішень і генерує остаточний прогноз шляхом усереднення прогнозів окремих дерев. У цій роботі ця модель служить для перевірки того, наскільки ансамблевий підхід без послідовного бустингу може покращити якість прогнозування порівняно з лінійною моделлю. Випадковий ліс – це корисний проміжний варіант між простою регуляризованою регресією та складнішою моделлю градієнтного бустингу.

Третя – це модель XGBoost. Цей алгоритм є одним із найефективніших підходів для задач регресії на табличних даних. Його використання виправдане тим, що ринкова вартість футболіста може залежати від складної взаємодії різних факторів: віку, позиції, ліги, ігрового часу, продуктивності та інших характеристик. XGBoost дозволяє враховувати такі залежності без їх явного визначення у вигляді правил.

Фрагмент реалізації навчання моделі XGBoost наведено в лістингу 3.2.

Лістинг 3.2 – Реалізація навчання моделі XGBoost

```
# Попередня обробка: масштабування числових і кодування
# категоріальних ознак
preprocessor = ColumnTransformer([
    ("num",
     Pipeline([("imputer",
                SimpleImputer(strategy="median")),
                ("scaler", StandardScaler())]),
     numeric_features),
    ("cat",
     Pipeline([("imputer",
                SimpleImputer(strategy="constant", fill_value="Unknown")),
                ("encoder",
                 OneHotEncoder(handle_unknown="ignore"))]),
     categorical_features),
])
```

```
# Модель градієнтного бустингу
xgb_model = Pipeline([
    ("preprocessor", preprocessor),
    ("model", XGBRegressor(
        n_estimators=300, learning_rate=0.05, max_depth=6,
        subsample=0.8, colsample_bytree=0.8,
        objective="reg:squarederror", random_state=42,
n_jobs=-1)),
])
xgb_model.fit(X_train, y_train)
```

Після навчання моделей було згенеровано прогнози для тестової вибірки, а результати було масштабовано назад. Далі для кожної моделі було розраховано метрики MAE, RMSE та R^2 . Отримані значення використовуються для подальшого експериментального порівняння моделей у четвертому розділі роботи. У цьому розділі основна увага приділяється реалізації моделей, тоді як детальний аналіз їхньої якості наведено в розділі експериментального дослідження.

Для подальшої інтеграції у вебзастосунок важливо, щоб модель могла працювати не лише в дослідницькому середовищі, але й як частина прототипу програмного забезпечення. Тому після навчання модель та необхідні елементи попередньої обробки слід зберегти у формі, придатній для повторного використання. Це дозволяє завантажити навчену модель у вебзастосунок, подавати їй нові вхідні дані та отримувати прогноз без перенавчання.

Таким чином, в рамках розробки системи було реалізовано кілька моделей регресійного машинного навчання, що дозволяє порівнювати різні підходи до прогнозування ринкової вартості футболістів. Рідж-регресія використовується як базова лінійна модель, Випадковий ліс – як ансамблевий метод на основі дерев рішень, а XGBoost – як градієнтна модель бустування для роботи зі складними нелінійними залежностями.

Реалізовані моделі є основою для подальшої оцінки якості прогнозування, інтерпретації результатів та використання системи у вебінтерфейсі.

3.4 Реалізація оцінювання невизначеності та інтерпретації прогнозів

Окрім формування прогнозу ринкової вартості футболіста за балами, розроблена система реалізує додаткові можливості для підвищення практичної цінності результату. До них належать оцінка невизначеності прогнозу та інтерпретація впливу окремих ознак на результуюче значення. Це дозволяє користувачеві отримати не лише єдину прогнозовану суму, а й ширше аналітичне пояснення результату.

Необхідність оцінки невизначеності зумовлена тим, що ринкова вартість футболіста не є абсолютно точним значенням. Вона залежить не лише від статистичних характеристик гравця, а й від зовнішніх факторів, які не завжди доступні у відкритих даних. До таких факторів належать контрактна ситуація, процес переговорів між клубами, попит на певну позицію, фінансові можливості команд, висвітлення футболіста в ЗМІ та інші умови трансферного ринку. Тому доцільно доповнити прогноз очок можливим діапазоном оцінки.

В рамках системи оцінки невизначеності реалізовано квантильні моделі. Основна модель формує центральний прогноз ринкової вартості, тоді як додаткові квантильні моделі служать для визначення нижньої та верхньої меж можливого діапазону. Такий підхід дозволяє представити результат у більш гнучкій формі, яка краще відповідає реальній практиці спортивної аналітики.

Для нижньої межі використовується модель, орієнтована на нижній квантиль розподілу прогнозу, а для верхньої межі – модель, орієнтована на верхній квантиль. В результаті, для кожного футболіста система може згенерувати три значення: центральний прогноз, нижню межу та верхню межу можливої ринкової вартості. Центральний прогноз відображає

основну оцінку моделі, а межі інтервалу окреслюють допустимий діапазон з урахуванням невизначеності.

Фрагмент реалізації формування прогнозного діапазону наведено в лістингу 3.3.

Лістинг 3.3 – Реалізація формування діапазону прогнозованої ринкової вартості

```
# Навчання квантильних моделей для нижньої (0.25) та
верхньої (0.75) меж
quantile_models = {}
for name, q in {"lower": 0.25, "upper": 0.75}.items():
    model_q = Pipeline([
        ("preprocessor", preprocessor),
        ("model", XGBRegressor(
            objective="reg:quantileerror",
            quantile_alpha=q,
            n_estimators=300,          learning_rate=0.05,
            max_depth=6,
            subsample=0.8,            colsample_bytree=0.8,
            random_state=42, n_jobs=-1)),
    ])
    model_q.fit(X_train, y_train)
    quantile_models[name] = model_q

# Формування діапазону для нового гравця (зі зворотним
перетворенням у євро)
low =
np.expml(quantile_models["lower"].predict(input_data)[0])
high =
np.expml(quantile_models["upper"].predict(input_data)[0])
```

Після отримання прогнозів у логарифмічному масштабі вони обернено перетворюються, щоб отримати всі значення в євро. Це стосується як центрального прогнозу, так і нижньої та верхньої меж. Це надає

користувачеві результат у зрозумілій формі: орієнтовну ринкову вартість гравця та можливий інтервал, у якому може знаходитися оцінка.

Важливою частиною системи також є інтерпретація результатів прогнозування. Для цього було використано SHAP, який дозволяє оцінити внесок окремих характеристик у кінцевий прогноз моделі. У контексті розробленої системи це особливо важливо, оскільки користувачеві потрібно розуміти, які характеристики футболіста збільшують або зменшують прогнозовану ринкову вартість.

Реалізація SHAP-інтерпретації виконується після навчання основної моделі машинного навчання. Для моделі обчислюються значення SHAP, які показують вплив кожної ознаки на результат прогнозування. Якщо значення SHAP для певної ознаки додатне, ця ознака збільшує прогнозоване значення футболіста. Якщо значення від'ємне, ознака зменшує прогноз. Такий підхід дозволяє перейти від простого числового прогнозування до поясненого результату.

Система може використовувати два рівні інтерпретації: глобальний та локальний. Глобальна інтерпретація показує, які характеристики є найважливішими для моделі в цілому. Вона дозволяє зрозуміти, які характеристики найчастіше впливають на прогноз ринкової вартості. Локальна інтерпретація пояснює прогноз для конкретного футболіста, тобто показує, чому система сформувала саме цю оцінку для вибраного набору характеристик.

Для практичного використання у вебзастосунку локальна інтерпретація має першорядне значення. Після введення характеристик футболіста користувач отримує не лише прогноз ринкової вартості, а й зрозуміле пояснення того, які саме характеристики мали найбільший вплив на результат і в якому напрямі вони його змінили. Наприклад, система може показати, що вік, ігровий час, позиція або результативність збільшують прогнозоване значення, тоді як інші характеристики, навпаки, можуть його

зменшувати, знижуючи підсумкову оцінку. Фрагмент SHAP-інтерпретації прогнозу наведено в лістингу 3.4.

Лістинг 3.4 – Реалізація SHAP-інтерпретації прогнозу

```
# Обчислення SHAP-значень для прогнозу конкретного
футболіста

explainer =
shap.TreeExplainer(xgb_model.named_steps["model"])

X_transformed =
xgb_model.named_steps["preprocessor"].transform(input_data)

shap_values = explainer(X_transformed)

# Внески one-hot ознак зводяться до вихідних характеристик
contributions = {}

for col, value in zip(feature_names,
shap_values.values[0]):
    original = map_to_original(col)
    contributions[original] = contributions.get(original,
0.0) + value
```

Результати інтерпретації можуть бути представлені у вигляді графіків або списку ознак з відповідним внеском у прогноз. Такий формат зручний для користувача, оскільки дозволяє швидко побачити, які ознаки мали найбільший вплив. Це підвищує прозорість системи та пом'якшує проблему «чорної скриньки», типову для складних моделей машинного навчання.

Поєднання оцінки невизначеності та SHAP-інтерпретації робить систему більш придатною для практичного аналізу. Користувач отримує не лише прогнозовану ринкову вартість, але й діапазон можливих оцінок та пояснення того, які характеристики вплинули на результат. Завдяки цьому систему можна використовувати як повноцінний допоміжний аналітичний інструмент, а не просто як модель, що повертає одне числове значення без жодного обґрунтування.

Таким чином, у рамках розробки інтелектуальної системи було реалізовано два важливі додаткові компоненти: оцінку невизначеності прогнозу за допомогою квантильних моделей та інтерпретацію результатів за допомогою методу SHAP. Перший компонент дозволяє представити результат не як фіксоване число, а у вигляді можливого діапазону ринкової вартості, що краще відповідає реальним умовам трансферного ринку. Другий компонент пояснює вплив окремих ознак на прогноз, показуючи, які характеристики підвищують або знижують оцінку. Разом ці компоненти суттєво підвищують інформативність, прозорість та практичну корисність розробленої системи, роблячи її придатною не лише для формування прогнозу, а й для його аналітичного обґрунтування.

3.5 Реалізація вебзастосунку та інтерфейсу користувача

Заключним етапом розробки інтелектуальної системи стала реалізація прототипу програмного забезпечення у вигляді вебзастосунку. Такий формат зручний для демонстрації роботи системи, оскільки користувач може взаємодіяти з моделлю через графічний інтерфейс без необхідності запуску коду або роботи безпосередньо з файлами даних.

Для реалізації вебзастосунку було використано Streamlit. Цей інструмент дозволяє швидко створювати інтерактивні вебінтерфейси для моделей машинного навчання на Python. Його перевагою є можливість поєднувати введення користувачем даних, запуск моделі прогнозування, виведення результатів та побудову пояснювальних елементів в межах одного застосунку.

Загальний дизайн вебзастосунку орієнтований на простоту та зрозумілість для користувача, який не обов'язково має технічну підготовку. Усі елементи керування згруповані за змістом, а результат прогнозу представлений у візуальному вигляді з акцентом на ключові показники. Робота з застосунком не вимагає знання мови програмування чи розуміння

внутрішньої структури моделі: користувач взаємодіє лише з графічним інтерфейсом, тоді як усі розрахунки, перетворення даних та формування прогнозу виконуються у фоновому режимі. Це робить систему придатною для практичного використання спортивними аналітиками, скаутами та іншими футбольними фахівцями.

Головна сторінка містить короткий опис призначення системи та елементи для введення характеристик футболіста. Інтерфейс розроблений таким чином, що користувач може послідовно заповнювати необхідні поля, після чого система формує прогноз приблизної ринкової вартості. Приклад головної сторінки вебзастосунку показано на рисунку 3.2.

СПОРТИВНА АНАЛІТИКА - MACHINE LEARNING
Оцінювання ринкової вартості футболістів

Введіть характеристики гравця, щоб отримати орієнтовну вартість, діапазон для переговорів та пояснення оцінки.

Профіль гравця

Вік	24	Основна позиція	Attack	Громадянство	Albania
Зріст, см	180,00	Підпозиція	Centre-Forward	Ліга / турнір	A League Men (Australia)
Кількість турнірів за сезон	2	Робоча нога	both		

Статистика сезону

Хвилини на полі	2200	Асисті	4	Червоні картки	0
Голи	5	Жовті картки	3		

Спрогнозувати ринкову вартість

Рисунок 3.2 – Головна сторінка вебзастосунку

Форма введення розділена на логічні групи відповідно до типів характеристик футболіста, що робить процес заповнення зручнішим, структурованішим і зрозумілішим для користувача, який не обов'язково має технічну підготовку. Перша група стосується профілю гравця та містить його особисті й позиційні параметри: вік, зріст, громадянство, основна та детальна позиція, робоча нога, ліга та кількість турнірів, у яких гравець брав участь.

Приклад цієї частини форми показано на рисунку 3.3.

Профіль гравця

Вік	Основна позиція	Громадянство
24	Attack	Albania
Зріст, см	Підпозиція	Ліга / турнір
180,00	Centre-Forward	A League Men (Australia)
Кількість турнірів за сезон	Робоча нога	
2	both	

Рисунок 3.3 – Введення персональних і позиційних характеристик футболіста

Друга група полів стосується ігрової статистики гравця за сезон. До неї належать загальний ігровий час, кількість голів, результативних передач та дисциплінарні показники – жовті та червоні картки. Такий поділ форми на профіль гравця та статистику результатів робить введення даних більш логічним та зручним для користувача. Приклад цієї частини форми показано на рисунку 3.4.

Статистика сезону

Хвилини на полі	Асисті	Червоні картки
2200	4	0
Голи	Жовті картки	
5	3	

Рисунок 3.4 – Введення ігрової статистики футболіста за сезон

Після введення ознак користувач запускає прогнозування. Вебзастосунок формує вхідний вектор, виконує необхідні перетворення та передає значення навченій моделі. Оскільки модель працює з логарифмічно перетвореною цільовою змінною, результат інвертується й подається у євро.

Окрім точкового прогнозу, вебзастосунок відображає можливий діапазон вартості футболіста, межі якого формують квантильні моделі, що дає змогу врахувати невизначеність оцінки. Приклад наведено на рисунку 3.5.

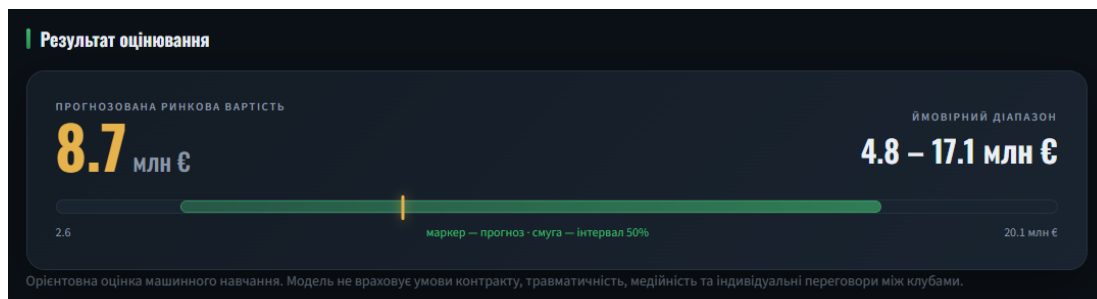


Рисунок 3.5 – Відображення прогнозу та діапазону ринкової вартості у вебзастосунку

Важливим елементом інтерфейсу є блок, що пояснює результат прогнозування. Для цього вебдодаток використовує SHAP-інтерпретацію, яка показує внесок окремих характеристик у підсумкову оцінку: користувач бачить, які ознаки збільшили прогнозовану ринкову вартість футболіста, а які знизили її. Пояснення подано в зрозумілій формі – окремо виділено основний фактор оцінки (зазвичай це ліга), а решту характеристик наведено з відповідним графічним представленням їхнього впливу.

Така подача дозволяє швидко зорієнтуватися в результаті: замість списку технічних коефіцієнтів користувач бачить у наочній формі, які характеристики працювали на користь оцінки, а які, навпаки, проти неї. Кольорове кодування, за яким позитивний вплив і негативний вплив позначено різними кольорами, а також компактний графік роблять пояснення інтуїтивно зрозумілим навіть без знання принципів роботи моделі та внутрішньої логіки її обчислень. Завдяки цьому сам користувач отримує не лише кінцевий прогноз, а й чітку аргументацію його формування, що дає змогу оцінити обґрунтованість результату та зіставити його з власними експертними уявленнями.

Приклад відображення пояснення прогнозу у вебзастосунку наведено на рисунку 3.6.



Рисунок 3.6 – Пояснення прогнозу ринкової вартості за допомогою SHAP

Окрім спрощеного пояснення, вебдодаток надає окремий блок детального аналізу, призначений для більш досвідчених користувачів. Він надає повний графік впливу всіх ознак на прогноз у вигляді SHAP-діаграми. Такий двопідхідний метод представлення дозволяє одночасно зробити результат зрозумілим для широкого кола користувачів та зберегти можливість поглибленого аналізу для спортивного аналітика.

На відміну від спрощеного пояснення, детальний аналіз відображає внесок кожної ознаки окремо, включаючи й ті, що мають незначний вплив на кінцеву оцінку та не потрапляють до основного пояснення. Це дозволяє досвідченому користувачеві простежити повну логіку формування прогнозу, побачити відносну вагу всіх характеристик і, за необхідності,

перевірити обґрунтованість впливу окремих ознак на підсумковий результат.

Приклад детального аналізу наведено на рисунку 3.7.



Рисунок 3.7 – Детальний аналіз впливу ознак (SHAP) для аналітиків

Структура вебзастосунку відповідає загальній архітектурі інтелектуальної системи. Спочатку користувач вводить характеристики футболіста, після чого система виконує підготовку вхідних даних, формує прогноз за допомогою навченої моделі, визначає можливий діапазон оцінки та пояснює результат. Всі ці етапи приховані від користувача за графічним інтерфейсом, що робить систему зручнішою у використанні.

Реалізований вебзастосунок має демонстраційний характер, але відображає основні функціональні можливості інтелектуальної системи.

Його можна використовувати для тестування роботи моделі на окремих прикладах, аналізу впливу характеристик футболіста на прогноз та демонстрації підходу до оцінки ринкової вартості за допомогою машинного навчання.

Отже, в рамках роботи було реалізовано вебзастосунок, який забезпечує взаємодію користувача з навченою моделлю машинного навчання. Інтерфейс дозволяє вводити характеристики футболіста, отримувати прогноз приблизної ринкової вартості, переглядати можливий діапазон оцінки та аналізувати пояснення результату. Завдяки цьому розроблена система набуває практичної форми та може бути використана як допоміжний інструмент для спортивної аналітики.

4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Умови проведення експериментального дослідження

Було проведено експериментальне дослідження з метою оцінки якості розроблених моделей машинного навчання та визначення моделі, яка найбільше підходить для прогнозування приблизної ринкової вартості футболістів. Для цього було використано підготовлений набір даних рівня «гравець-сезон», сформований на попередніх етапах роботи. Після очищення та зіставлення таблиць набір містив 52 254 записи про 15 965 футболістів за сезони 2012–2025 років.

Вхідними даними для моделей були особисті, позиційні, статистичні та контекстуальні характеристики футболістів. До них належать вік, зріст, національність, основна позиція, детальна позиція, ліга, сезон, ігровий час, кількість голів, результативних передач, жовтих та червоних карток, а також кількість турнірів. Цільовою змінною є ринкова вартість футболіста в євро.

Перед навчанням моделей цільову змінну було трансформовано за допомогою $\log_{1p}$ -перетворення. Це дозволило зменшити вплив дуже великих ринкових значень та зробити процес навчання стабільнішим. Після генерації прогнозів було виконано зворотне перетворення для представлення результату у вихідному масштабі, тобто в євро.

Для оцінки якості моделей набір даних було розділено на навчальну та тестову вибірки у співвідношенні 80% до 20%. З метою зменшення ризику витоку даних було застосовано групування за ідентифікатором гравця. Такий підхід запобігає ситуації, коли записи одного й того ж гравця за різні сезони одночасно потрапляють до навчальної та тестової вибірок, що дозволяє більш реалістично оцінити здатність моделі працювати з новими гравцями.

В рамках експерименту було навчено та порівняно три регресійні моделі: Ridge Regression, Random Forest та XGBoost. Всі моделі були навчені та протестовані за однакових умов – на однакових навчальних та тестових вибірках та з однаковим набором ознак, що забезпечує коректність їх порівняння. Для оцінки якості прогнозування використовувалися метрики MAE, RMSE та R^2 . MAE показує середню абсолютну похибку, RMSE сильніше враховує великі відхилення, а R^2 дозволяє оцінити, яка частина мінливості цільової змінної пояснюється моделлю.

Експерименти проводилися на Python з використанням бібліотек pandas, scikit-learn та XGBoost у середовищі Jupyter Notebook. Цей інструментарій забезпечив відтворюваність експерименту: підготовлений набір даних, навчені моделі та проміжні результати зберігаються на диску, що дозволяє отримати ті самі результати знову без перенавчання.

Тому експериментальне дослідження мало на меті об'єктивне порівняння моделей за однакових умов. Це дозволяє визначити, яка з моделей забезпечує найкращу якість прогнозування та може бути використана як основа розробленої інтелектуальної системи.

4.2 Порівняння якості моделей машинного навчання

Після навчання моделей їх порівнювали на тестовій вибірці. Для оцінки якості прогнозу використовувалися метрики MAE, RMSE та R^2 , які дозволяють оцінити середній рівень похибки, наявність великих відхилень та здатність моделі пояснювати мінливість цільової змінної. Оскільки модель навчалася на логарифмічно перетвореній цільовій змінній, метрики розраховувалися за двома шкалами: логарифмічною шкалою, в якій проводилося навчання, та вихідною шкалою (мільйони євро), що зручно для практичної інтерпретації.

Результати порівняння моделей машинного навчання наведено в таблиці 4.1.

Таблиця 4.1 – Порівняння якості моделей машинного навчання

Модель	MAE (log)	RMSE (log)	R ² (log)	MAE, млн €	RMSE, млн €	R ² (€)
Ridge Regression	0,6690	0,8549	0,6387	3,40	8,61	0,4234
Random Forest	0,6288	0,8124	0,6737	3,25	8,21	0,4766
XGBoost	0,5905	0,7666	0,7095	3,06	7,55	0,5567

Рідж-регресія використовується як базова лінійна модель для порівняння. Вона дозволяє оцінити, наскільки добре можна вирішити задачу за допомогою лінійного зв'язку між ознаками та цільовою змінною. Випадковий ліс відображає можливості ансамблевого підходу, заснованого на деревах рішень, а XGBoost дозволяє враховувати складніші нелінійні зв'язки між характеристиками футболістів.

Як видно з таблиці 4.1, якість прогнозування прогресує від базової лінійної моделі до ансамблевих методів. Рідж-регресія показала найслабші результати ($R^2 = 0,6387$ у логарифмічній шкалі), що не змогло подолати обмеження чисто лінійного підходу для цієї задачі. Випадковий ліс підвищив якість ($R^2 = 0,6737$), а найкращу продуктивність продемонструвала модель XGBoost: найнижчі значення помилок (MAE = 0,5905 та RMSE = 0,7666 у логарифмічній шкалі) та найвищі значення коефіцієнта детермінації ($R^2 = 0,7095$). Таку ж перевагу XGBoost підтверджують і метрики у вихідній шкалі: середня абсолютна помилка становить близько €3,06 млн, а коефіцієнт детермінації – 0,5567.

Постійне вдосконалення від Ridge до XGBoost підтверджує доцільність використання градієнтного бустингу для вирішення цієї проблеми. Це пояснюється тим, що ринкова вартість футболіста залежить від складної взаємодії ознак (включаючи вік, позицію, лігу та

результативність), яку лінійна модель не може відтворити, а градієнтний бустинг враховує це ефективніше, ніж ансамбль незалежних дерев.

Для наочного подання результатів порівняння побудовано графік значень метрик якості, наведений на рисунку 4.1.

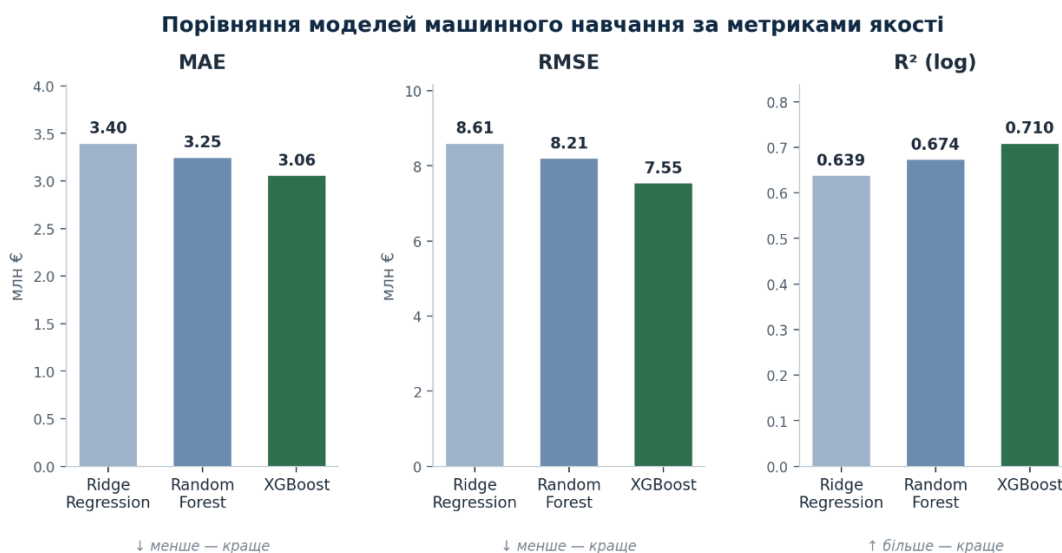


Рисунок 4.1 – Порівняння моделей машинного навчання за метриками якості

Характер похибки обраної моделі слід пояснити окремо. Середня абсолютна похибка XGBoost у вихідній шкалі (близько €3,06 млн) є середньою по всій вибірці та суттєво зміщена невеликою кількістю дуже дорогих футболістів, для яких абсолютні відхилення вимірюються десятками мільйонів євро. Для переважної більшості гравців з помірною вартістю фактична похибка моделі значно менша. Тому показник R^2 у логарифмічній шкалі (0,7095) точніше відображає реальну якість моделі, ніж абсолютна похибка в євро, яку слід інтерпретувати з урахуванням нерівномірного розподілу цільової змінної.

Таким чином, порівняння моделей дозволило визначити найбільш підходящий алгоритм для прогнозування приблизної ринкової вартості футболістів. Найкращі результати показала модель XGBoost, яка була

обрана за основу інтелектуальної системи. Надалі ця модель використовується для генерації прогнозу у вебзастосунку, а також для побудови пояснення результатів за допомогою SHAP-інтерпретації.

4.3 Інтерпретація результатів прогнозування

Після вибору моделі, яка показала найкращу якість прогнозування, її результати були інтерпретовані. Для цього було використано метод SHAP, який дозволяє оцінити внесок окремих ознак у сформований прогноз ринкової вартості футболіста.

Інтерпретація результатів є важливою, оскільки користувачеві недостатньо отримати лише числову оцінку значення. Для практичного застосування системи необхідно розуміти, які характеристики футболіста найбільше вплинули на результат: вік, ігровий час, позиція, ліга, результативність чи інші характеристики.

Глобальна інтерпретація дозволяє визначити, які ознаки мають найбільший вплив на прогнози моделі в цілому, та перевірити, чи відповідає поведінка моделі логіці предметної галузі. Слід зазначити, що деякі категоріальні ознаки (такі як громадянство та ліга) мають високу кардинальність, тобто містять велику кількість можливих значень. Під час кодування такі ознаки розкладаються на багато окремих стовпців, і їхній загальний внесок згідно з SHAP накопичується. Тому для правильного аналізу доцільно розглядати глобальну важливість не на рівні окремих ознак, а на рівні їх груп.

Для побудови узагальненої оцінки всі ознаки об'єднані у чотири групи за їхньою природою: особистісні характеристики футболіста, позиційні ознаки, показники ігрової статистики та контекстуальні фактори. Для кожної групи розраховується загальний середній внесок ознак за модулем значень SHAP, що дає змогу порівняти вплив різних категорій ознак на прогноз та простежити, які аспекти найбільше визначають оцінку ринкової

вартості. Такий груповий підхід забезпечує більш стабільну та зрозумілу картину, ніж аналіз окремих ознак, оскільки виключає вплив технічних особливостей кодування категоріальних змінних. Результати такого аналізу наведено на рисунку 4.2.

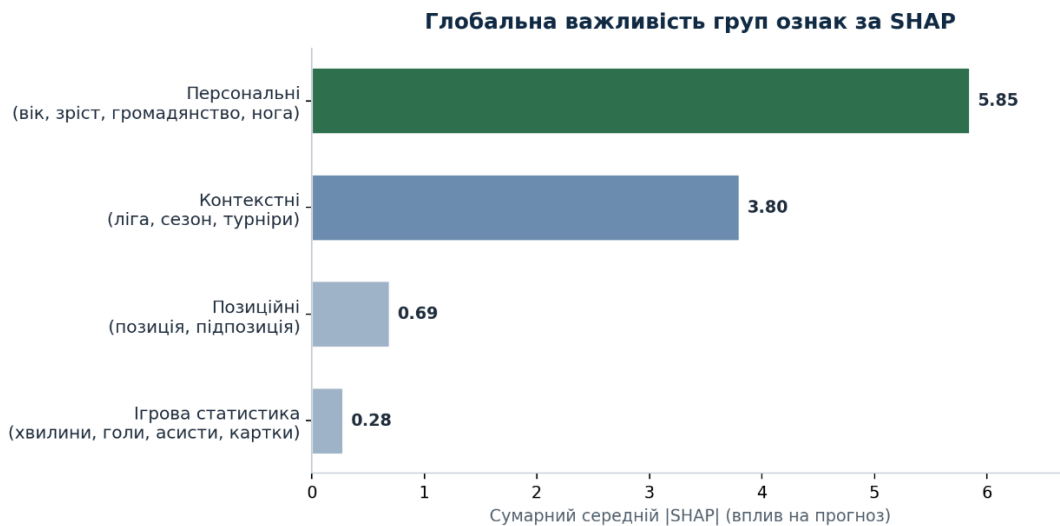


Рисунок 4.2 – Глобальна важливість груп ознак за результатами SHAP-інтерпретації

Як видно з рисунка 4.2, найбільший вплив на прогноз мають особисті характеристики футболіста (вік, зріст, національність) та контекстуальні особливості, серед яких ключову роль відіграє рівень ліги. Це узгоджується з логікою футбольного ринку, де базовий рівень цінності гравця значною мірою визначається його віком та рівнем змагань, у яких він грає. Позиційні характеристики та статистика гри мають менший загальний вплив, що частково пояснюється відсутністю поглиблених показників (наприклад, xG) у відкритих даних та впливом рівня ліги, який опосередковано відображається й на самій статистиці.

Особливий інтерес представляє аналіз впливу результативності залежно від позиції гравця. Він дозволяє перевірити припущення, висунуте в розділі 1, що одні й ті ж статистичні показники оцінюються по-різному

для гравців різних амплуа. Для цього середній внесок голів у прогноз було розраховано окремо для нападників, півзахисників та захисників. Результати наведено на рисунку 4.3.

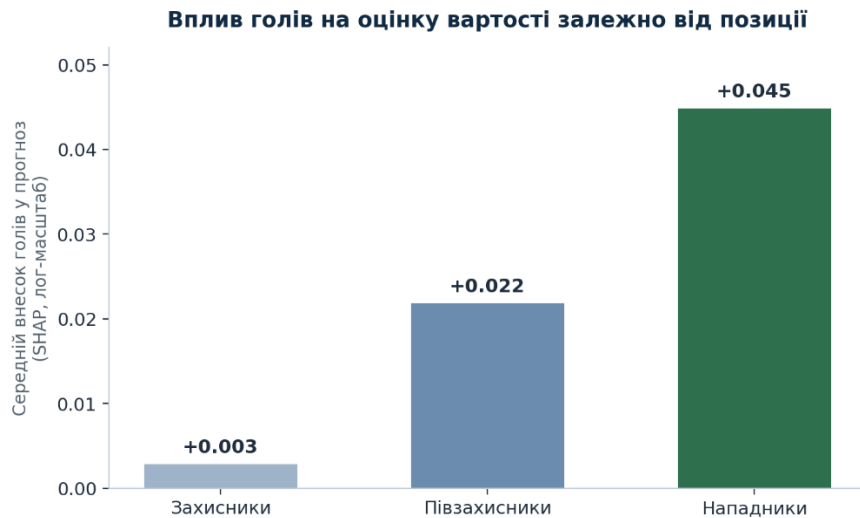


Рисунок 4.3 – Вплив голів на оцінку вартості залежно від позиції футболіста

Отримані значення підтверджують позиційно-залежний характер ціноутворення. Для нападників голи мають найбільший позитивний внесок у прогнозоване значення (в середньому +0,045 за логарифмічною шкалою), для півзахисників вплив помітно менший (+0,022), а для захисників майже нейтральний (+0,003). Це означає, що модель не сприймає голи як універсальний показник цінності, а враховує ігрову роль футболіста: результативність значно підвищує цінність нападника, але майже не впливає на оцінку захисника, для якого важливіші інші характеристики. Цей результат відповідає реальній логіці футбольного ринку та підтверджує здатність моделі вловлювати складні, контекстно-залежні закономірності.

Окрім глобального аналізу, метод SHAP також використовується для локальної інтерпретації окремого прогнозу. Він показує, які характеристики збільшили або знизили прогнозовану ринкову вартість конкретного

футболіста: позитивні значення SHAP збільшують кінцевий прогноз, а негативні – зменшують. Саме локальна інтерпретація використовується у вебдодатку для пояснення результату користувачеві (рисунки 3.6 та 3.7), що дозволяє побачити для кожного гравця список характеристик, які найбільше вплинули на оцінку.

Використання SHAP-інтерпретації робить модель більш прозорою. Користувач може не лише побачити прогнозовану ринкову вартість футболіста, але й зрозуміти, які характеристики мають найбільший вплив на результат. Це зменшує проблему «чорного ящика» та підвищує практичну корисність розробленої системи.

Отже, інтерпретація результатів прогнозування підтверджує, що система може використовуватися не лише для одержання числової оцінки, а й для аналітичного пояснення прогнозу. Особливо цінним є виявлений позиційно-залежний вплив результативності, що свідчить про здатність моделі враховувати ігрову роль футболіста. Це є важливою перевагою розробленого підходу у сфері спортивної аналітики.

4.4 Обмеження системи та напрями подальшого розвитку

Незважаючи на впровадження інтелектуальної системи оцінки ринкової вартості футболістів, її результати мають певні обмеження. Це пов'язано, перш за все, з тим, що модель оперує лише тими ознаками, які доступні у відкритому наборі даних і можуть бути представлені у структурованому вигляді. Тому прогноз слід розглядати як приблизну аналітичну оцінку, а не як точну суму майбутнього трансферу.

Одним з основних обмежень є відсутність у моделі деяких важливих ринкових та нестатистичних факторів. До них належать тривалість контракту, зарплата, історія травм, висвітлення гравця в ЗМІ, вплив агента, фінансові можливості клубів та конкретна ситуація на трансферному ринку. Ці характеристики можуть суттєво впливати на реальну вартість гравця, але

вони не завжди доступні у відкритих джерелах або їх важко кількісно оцінити. Зокрема, експериментально була перевірена можливість врахування контрактної ситуації: до моделі було додано ознаку терміну дії контракту, що залишився, але її внесок виявився незначним, а самі дані контракту були представлені лише як поточний стан без історії за сезонами, що створювало ризик часової неузгодженості. З огляду на це, ознаку було виключено з остаточної моделі.

Ще одним обмеженням є неповнота ігрової статистики. У роботі використовувалися основні статистичні показники, зокрема, ігровий час, голи, результативні передачі та дисциплінарні характеристики. Водночас, більш детальні метрики, такі як xG, xA, прогресивні передачі, успішність єдиноборств, захисні дії або показники володіння м'ячем, могли б покращити якість моделі та повніше описати внесок футболіста в гру. Саме відсутність таких метрик частково пояснює помірний загальний вплив ігрової статистики, виявлений під час інтерпретації результатів [20].

Особливу складність становить прогнозування вартості футболістів з дуже високими ринковими оцінками. Такі гравці можуть суттєво відрізнятися від більшості прикладів у наборі даних, і їхня вартість часто залежить не лише від статистики, але й від репутації, потенціалу розвитку, попиту з боку провідних клубів та впливу ЗМІ. Через це похибка прогнозування для дорогих гравців може бути більшою в абсолютному вираженні, що також відображається на загальній середній похибці моделі. Крім того, окремі ознаки виявилися слабкими предикторами: наприклад, внесок робочої ноги майже не залежить від конкретного значення, тому ця ознака не відображається у спрощеному поясненні системи.

Якість системи також залежить від релевантності та повноти використаних даних. Оскільки ринкова інфляція враховується через сезонну функцію, а для поточного прогнозування сезон фіксується на найновішому доступному рівні, система формує оцінку відповідно до рівня цін останнього сезону навчальної вибірки. Ринок футбольних трансферів

швидко змінюється, тому модель, навчена на історичних даних, потребує періодичних оновлень. Додавання нових сезонів, оновлення ринкових оцінок та розширення набору функцій можуть покращити здатність системи працювати з поточними ринковими умовами.

Напрямки подальшого розвитку системи полягають у розширенні набору даних, додаванні більш детальних ігрових метрик, врахуванні контрактних та медіа-факторів за наявності відповідних історичних даних, а також удосконаленні інтерфейсу вебзастосунку та розширенні його функціональності, зокрема можливості порівняння гравців та виявлення недооцінених гравців. Також доцільно протестувати інші моделі машинного навчання, налаштувати гіперпараметри та порівняти результати на новіших даних.

Отже, розроблена система придатна для формування приблизної оцінки ринкової вартості футболістів та пояснення отриманого прогнозу, проте її результати слід використовувати як допоміжний аналітичний інструмент. Усвідомлення вищезазначених обмежень дозволяє правильно інтерпретувати результати, а подальше вдосконалення системи має бути спрямоване на розширення джерел даних, підвищення точності прогнозування та врахування більшої кількості факторів, що впливають на реальну вартість футболіста.

ВИСНОВКИ

У кваліфікаційній роботі було вирішено задачу розробки інтелектуальної системи оцінки приблизної ринкової вартості футболістів з використанням методів машинного навчання. Мета була досягнута, а завдання, визначені у вступі, повністю виконані.

У теоретичній частині аналізується предметна область, основні фактори формування ринкової вартості та проблеми її оцінки – багатофакторність, суб'єктивність та недостатня прозорість існуючих рішень. Огляд платформ (Transfermarkt, CIES Football Observatory, Football Benchmark) показав, що вони або спираються на краудсорсингові оцінки, або використовують закриті алгоритми, що підтвердило доцільність розробки власної системи з прозорою логікою. Розглянуто методи машинного навчання для регресійного прогнозування: моделі Ridge Regression, Random Forest та XGBoost, метрики MAE, RMSE та R^2 , квантильна регресія для оцінки невизначеності та метод SHAP для інтерпретації прогнозів.

У практичній частині було згенеровано набір даних рівня «гравець-сезон», виконано попередню обробку, логарифмічне перетворення цільової змінної та розбиття на групи за ID гравця, щоб уникнути витоку даних. Було навчено та порівняно три регресійні моделі. Найкращу якість показала модель XGBoost з $MAE \approx \text{€}3,06$ млн, $RMSE \approx \text{€}7,55$ млн та $R^2 = 0,71$ (у логарифмічному масштабі), причому якість послідовно зростала від лінійної моделі до градієнтного бустингу.

Окрім точкового прогнозу, було реалізовано оцінку діапазону значень за допомогою квантильних моделей та інтерпретацію результатів за допомогою методу SHAP. Аналіз інтерпретації підтвердив, що модель враховує ігрову роль футболіста: голи суттєво підвищують оцінку нападників, але майже не впливають на цінність захисників. Практичним результатом роботи є прототип програмного забезпечення у вигляді

вебдодатку, який поєднує прогнозування, оцінку невизначеності та пояснення результату, що відрізняє його від закритих комерційних рішень.

Розроблена система також має обмеження: вона працює лише з відкритими структурованими даними та не враховує контрактні, фінансові та медійні фактори. Подальший розвиток системи має бути спрямований на додавання більш детальних ігрових метрик (зокрема xG), врахування контрактних даних, розширення функціональності для порівняння гравців та регулярне оновлення моделі в новіших сезонах.

Таким чином, у межах роботи розроблено інтелектуальну систему, яка поєднує прогнозування ринкової вартості футболістів, оцінку невизначеності та інтерпретацію результатів за допомогою методів машинного навчання. Отримані результати підтверджують придатність такого підходу як допоміжного інструменту для спортивної аналітики.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Transfer reports. *Inside FIFA*. URL: <https://inside.fifa.com/transfer-system/transfer-reports> (date of access: 05.07.2026).
2. Müller O., Simons A., Weinmann M. Beyond crowd judgments: data-driven estimation of market value in association football. *European journal of operational research*. 2017. Vol. 263, no. 2. P. 611–624. URL: <https://doi.org/10.1016/j.ejor.2017.05.005> (date of access: 05.07.2026).
3. Determinants of football players' valuation: a systematic review / M. Franceschi et al. *Journal of economic surveys*. 2023. URL: <https://doi.org/10.1111/joes.12552> (date of access: 05.07.2026).
4. Majewski S. Identification of factors determining market value of the most valuable football players. *Journal of management and business administration. central europe*. 2016. Vol. 24, no. 3. P. 91–104. URL: <https://doi.org/10.7206/jmba.ce.2450-7814.177> (date of access: 05.07.2026).
5. Behravan I., Razavi S. M. A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*. 2020. URL: <https://doi.org/10.1007/s00500-020-05319-3> (date of access: 05.07.2026).
6. Football transfers, rumours, market values and news. *Football transfers, rumours, market values and news | Transfermarkt*. URL: <https://www.transfermarkt.com> (date of access: 05.07.2026).
7. Herm S., Callsen-Bracker H.-M., Kreis H. When the crowd evaluates soccer players' market values: accuracy and evaluation attributes of an online community. *Sport management review*. 2014. Vol. 17, no. 4. P. 484–492. URL: <https://doi.org/10.1016/j.smr.2013.12.006> (date of access: 05.07.2026).
8. Go to the tool - CIES football observatory. *Observatoire du football CIES*. URL: <https://football-observatory.com/Tool-Value> (date of access: 05.07.2026).

9. Monthly report 53. *Observatoire du football CIES*.
URL: <https://football-observatory.com/IMG/sites/mr/mr53/en/> (date of access: 05.07.2026).
10. Player valuation - methodology. *Football benchmark*.
URL: <https://footballbenchmark.com/player-valuation-methodology> (date of access: 05.07.2026).
11. FBref: football statistics and history. *FBref.com*. URL: <https://fbref.com> (date of access: 05.07.2026).
12. GitHub - statsbomb/open-data: free football data from StatsBomb. *GitHub*. URL: <https://github.com/statsbomb/open-data> (date of access: 05.07.2026).
13. Jiang T., Gradus J. L., Rosellini A. J. Supervised machine learning: a brief primer. *Behavior therapy*. 2020. Vol. 51, no. 5. P. 675–687.
URL: <https://doi.org/10.1016/j.beth.2020.05.002> (date of access: 05.07.2026).
14. Russell S., Norvig P. Artificial intelligence: a modern approach. 4th ed. Hoboken : Pearson, 2020. 1168 p.
15. Shwartz-Ziv R., Armon A. Tabular data: deep learning is not all you need. *Information fusion*. 2022. Vol. 81. P. 84–90.
URL: <https://doi.org/10.1016/j.inffus.2021.11.011> (date of access: 05.07.2026).
16. Scikit-learn: machine learning in Python. *Journal of machine learning research*. URL: <https://jmlr.org/papers/v12/pedregosa11a.html> (date of access: 05.07.2026).
17. Adadi A., Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*. 2018. Vol. 6. P. 52138–52160. URL: <https://doi.org/10.1109/access.2018.2870052> (date of access: 06.07.2026).
18. McHale I. G., Holmes B. Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European journal of operational research*. 2022.
URL: <https://doi.org/10.1016/j.ejor.2022.06.033> (date of access: 05.07.2026).

19. Yang Y., Koenigstorfer J., Pawlowski T. Predicting transfer fees in professional European football before and during COVID-19 using machine learning. *European sport management quarterly*. 2022. P. 1–21. URL: <https://doi.org/10.1080/16184742.2022.2153898> (date of access: 05.07.2026).
20. PlayeRank / L. Pappalardo et al. *ACM transactions on intelligent systems and technology*. 2019. Vol. 10, no. 5. P. 1–27. URL: <https://doi.org/10.1145/3343172> (date of access: 05.07.2026).