УДК 004.94

Д. В. Чорнобривець,
К. І. Сергійчук, Т. А. Ліхоузова

МОДЕЛІ ДЛЯ АНАЛІЗУ УСПІШНОСТІ СТАРТАПІВ ТА ПРОГНОЗУВАННЯ ЇХ ВИЖИВАННЯ НА РИНКУ

Анотація: Робота присвячена аналізу та прогнозуванню успішності стартапів. Метою є виявлення факторів та залежностей, що впливають на можливості фінансування старапу, а також розробка моделей для прогнозування його успіху і майбутніх тенденцій. У роботі проведено ґрунтовний аналіз історичних даних стартапів з різних країн, їх показники фінансування та пройдені віхи розвитку. Дослідження базувалось на даних за більш ніж 100 років. Очікується, що результати дослідження дозволять отримати глибше розуміння майбутніх перспектив стартапів.

Ключові слова: інтелектуальний аналіз даних, модель прогнозування, k-nearest neighbors, decision tree, gradient boosting, random forest, метрики точності.

Вступ

В сучасному світі стартапи відіграють ключову роль у стимулюванні інновацій та економічного зростання. Вони пропонують нові продукти, послуги та бізнес-моделі, які можуть змінювати галузі та створювати нові ринки. Однак, незважаючи на потенціал стартапів, їх виживання на ринку залишається однією з головних проблем. Лише невеликий відсоток стартапів досягає довготривалого успіху, тоді як більшість з них стикаються з труднощами і закриваються протягом перших кількох років діяльності.

Успішність стартапу залежить від багатьох факторів [1], включаючи ринкові умови, інноваційність продукту, якість команди, доступ до фінансування, стратегії маркетингу та багатьох інших. Розуміння та аналіз цих факторів може допомогти підприємцям краще підготуватися до викликів та збільшити шанси на успіх.

Матеріали та методи

Одним із ключових факторів, що визначають успіх стартапів, є інноваційність продукту, яка характеризується рівнем новизни та корисності продукту або послуги, що пропонується стартапом. Ринкові умови також мають значний вплив, включаючи попит на продукт, конкурентне середовище та макроекономічні показники. Якість команди, яка включає досвід, компетентність та мотивацію засновників і співробітників, відіграє важливу роль у досягненні успіху. Доступ до фінансових ресурсів, таких як венчурний капітал, гранти та інші джерела фінансування, є ще одним важливим фактором. В рамках даної роботи досліджено вплив зазначених

факторів на успішність стартапів та запропоновано моделі для прогнозування їхнього виживання на ринку. Для цього проаналізовано дані про різні стартапи, їхні характеристики та результати діяльності.

Для поставленого завдання були обрані дані [2], що включають в себе інформацію про стартапи, інвестиції, венчурні угоди, організації та людей. Даний набір даних складається з 100000+ записів про різноманітні дати/статистики, що містять багато аспектів, пов’язаних зі світом стартапів з 1901 до 2014 року.

Набори даних представлені у вигляді декількох CSV файлів (таблиця 1).

Таблиця 1

Назва	Опис
acquisitions.csv	інформація про куплені стартапи
degrees.csv	освіта людей, залучених у світ стартапів
funding_rounds.csv	раунди фінансування стартапів
funds.csv	фонди венчурного капіталу, які здійснюють інвестиції
investments.csv	інвестиції, зроблені венчурними капіталістами
ipos.csv	первинні публічні пропозиції
milestones.csv	важливі події в екосистемі стартапів
objects.csv	основний файл, що містить базову інформацію
offices.csv	інформація про офіси стартап-компаній
people.csv	інформація про людей у світі стартапів
relationships.csv	дані про відносини, які пов’язують компанії з окремими особами та їхніми посадами

Попередня обробка даних

Для роботи з даними на мові Python буде використано бібліотеку pandas [3,4]. Після завантаження кожної таблиці вибирається лише підмножина змінних, які будуть корисні в аналізі. Це робиться для спрощення даних та зменшення їх обсягу. Деякі стовпці даних конвертуються в інші типи для кращої обробки. Наступний крок – побудова моделі даних для об’єднання всіх таблиць. Після цього дані перевіряються на наявність дублікатів, дублікати видаляються. На цьому етап попередньої обробки даних можна вважати завершеним.

Первинний аналіз вхідних даних

Залежна змінна Status є категоріальною, складається з 4 рівнів, які вказують статус кожного стартапу. Ці рівні:

CLOSED: не вдалося запустити стартап (failed startup) – 2584;

ACQUIRED: придбаний стартап (acquired startup) – 9393;

IPO : стартап у списку (listed startup) – 1134;

OPERATING: стартап успішно запущений – 183387.

Слід відмітити, що кількість зареєстрованих і придбаних стартапів дуже мала, в той час як більшість з них працюють.

 ISSN 1560-8956

Категорії стартапів. Category_code — це категоріальна змінна, що складається з 42 невпорядкованих змінних, яка вказує на сегмент ринку, на якому працює компанія.

```
# Display the result
print(status_settore)
```

	category_code	n_acquired	n_closed	n_ipo	n_ipo_acquired	\
35	semiconductor	154	34	53	0	
3	biotech	455	117	247	0	
14	hardware	200	75	90	0	
4	cleantech	115	74	56	0	
19	manufacturing	30	5	13	0	
34	security	127	17	21	0	
21	messaging	24	22	5	0	
11	finance	40	18	22	0	
24	nanotech	1	1	1	0	
2	automotive	5	4	4	0	
16	hospitality	13	7	10	0	
25	network_hosting	201	66	30	0	
9	enterprise	409	68	54	0	
20	medical	16	26	13	0	
31	public_relations	215	65	30	0	
26	news	27	10	7	0	
32	real_estate	8	3	4	0	
22	mobile	416	172	49	0	
15	health	27	8	11	0	
37	software	1466	314	114	0	
28	other	313	103	73	0	
10	fashion	11	11	3	0	
5	consulting	136	31	21	0	
12	games_video	342	201	31	0	
...						
29	0.0	1.000000				
13	0.0	1.000000				
38	0.0	0.979259				
6	0.0	0.975089				

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Рисунок 1. Розподіл стартапів по сегментам ринку, з метриками рівнів

Розподіл частоти стартапів для різних секторів показано нижче, де ми можемо помітити, що найбільш популярними є сектори, що стосуються технологій і цифрових технологій (рис.2).

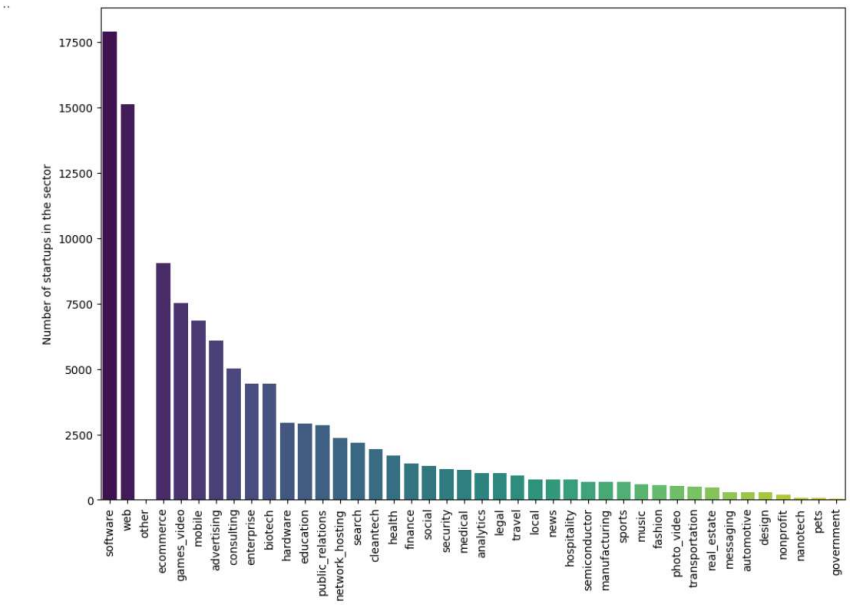

Рисунок 2. Графік розподілу частоти стартапів для різних секторів

Нижче наведено графіки (рис.3), що аналізують різноманітні аспекти діяльності стартапів за допомогою відсоткових показників, які відображаються для кожної категорії стартапів окремо.

Аналіз цих графіків дозволяє зробити висновки про те, які категорії стартапів є найбільш успішними в різних аспектах їхньої діяльності на ринку. Наприклад, високий відсоток стартапів, що вийшли на біржу, може свідчити про сильний інтерес ринку до цих секторів або їхню успішність у залученні інвестицій. З іншого боку, високий відсоток закритих стартапів може вказувати на високий рівень конкуренції або складність підтримки стабільної діяльності в цих секторах.

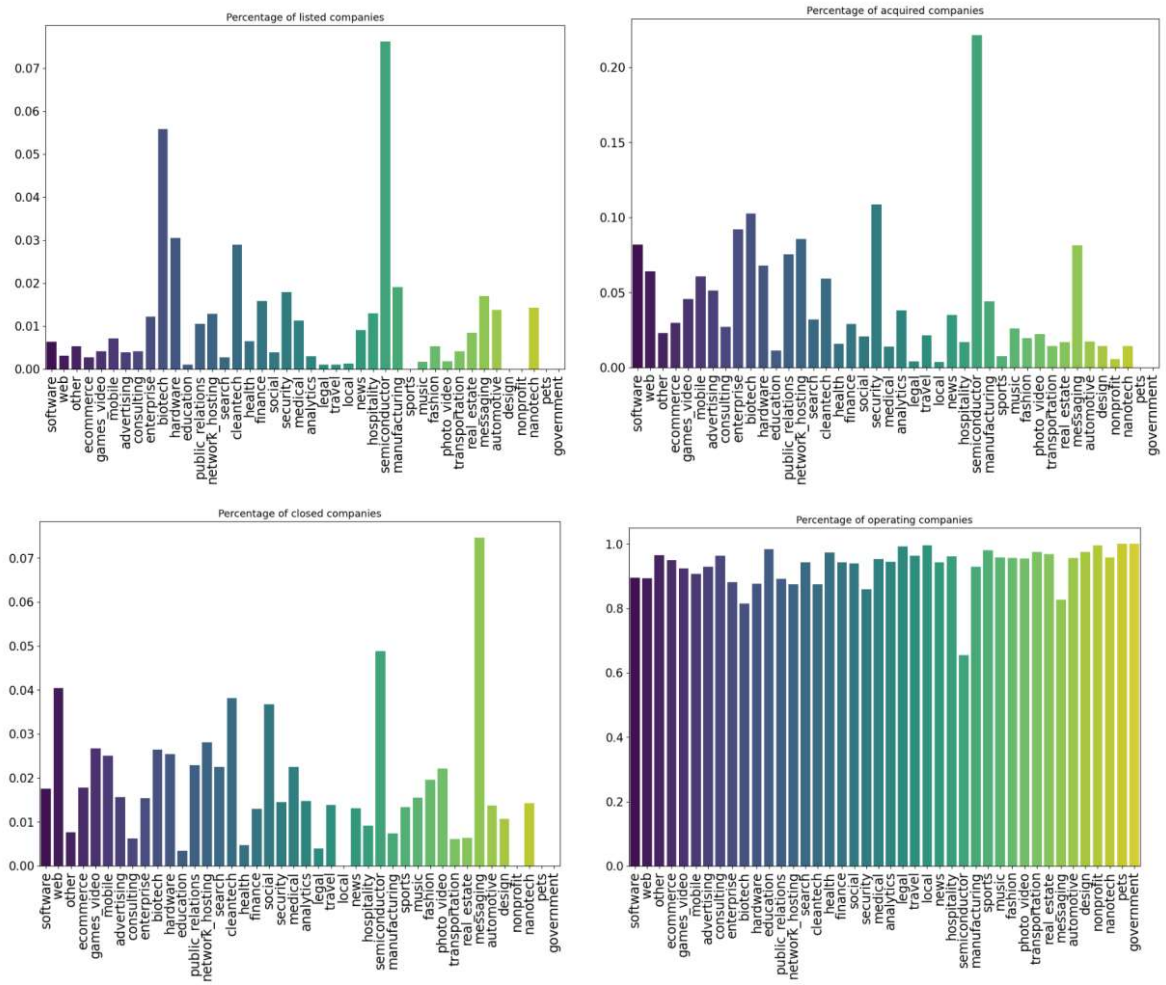

Рисунок 3. Аналіз впливу секторів ринку

Можна помітити, що більшість компаній працюють; справді, майже в усіх секторах коефіцієнт діючих компаній/загальна кількість компаній становить приблизно 1.

Сектор напівпровідників має найвищий рівень IPO та рівень придбань, а також найнижчий рівень операцій. У сфері біотехнологій багато стартапів вийшли на IPO та/або були придбані.

 ISSN 1560-8956

Різні значення цих відсотків у різних секторах є чітким показником того, що сегмент ринку, до якого належить стартап, може вплинути на його провал/успіх.

Географічне положення. На наступній карті показано розташування штаб-квартир приблизно 80 000 стартапів (в інших відсутні значення широти та/або довготи). Кожне коло, розміщене на карті, представляє окремий стартап із визначеним статусом (IPO, Closed, Operating, Acquired). Колір кола відповідає конкретному статусу стартапу, що дозволяє візуально порівняти розподіл стартапів за їхнім географічним розташуванням і статусом. Такий підхід дозволяє швидко оцінити, де зосереджені стартапи з різними видами діяльності та статусами.

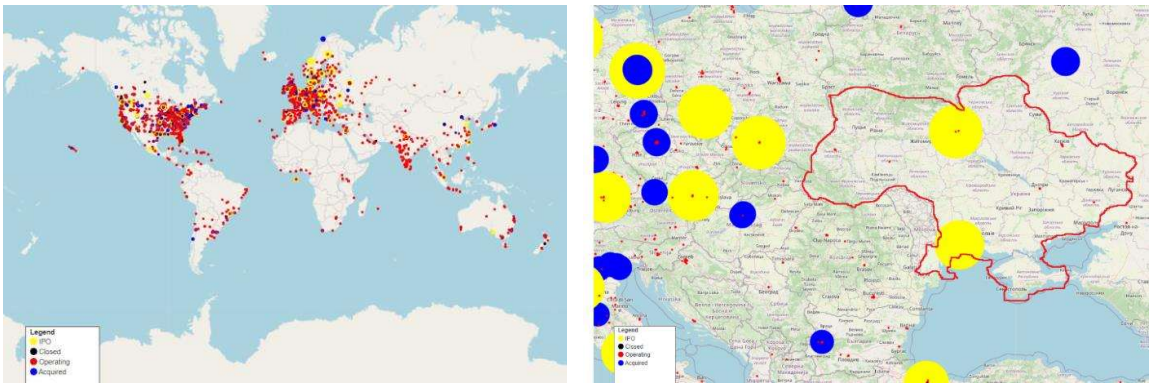

Рисунок 4. Географічне розташування стартапів, з їхнім статусом

Більшість із них у Європі та США. «Найгустішими» районами є Нью-Йорк, Бостон, Вашингтон (східне узбережжя), Кремнієва долина та Лос-Анджелес (західне узбережжя). Далі йдуть такі міста, як Денвер, Атланта, Даллас. Також показано в Україні. (рис. 4)

Той факт, що в країні було засновано багато стартапів, але жоден із них не котирується на фондовій біржі, може означати, що відповідна країна є важкою територією для розвитку такого типу бізнесу, або що чинне законодавство та бюрократія змушують процес розміщення на фондовій біржі складний і повільний, але також те, що стартапи в цих країнах були нещодавно засновані та ще не встигли зріти, особливо для країн з високим рівнем компаній, які все ще працюють.

Серед країн із понад 125 стартапами, жоден з яких не є IPO, ми аналізуємо частку діючих, закритих і придбаних компаній від загальної кількості компаній у кожній країні (рис.5). Можемо помітити, що більшість стартапів все ще працюють у своїх країнах, тоді як невелика частина з них закрита або придбана. У Данії та Бельгії високий відсоток придбаних компаній (7,12% і %), тоді як у Румунії дуже високий відсоток закритих компаній (0). З цього аналізу можна сказати, що, ймовірно, географічний регіон заснування стартапу впливає на його успіх/невдачу.

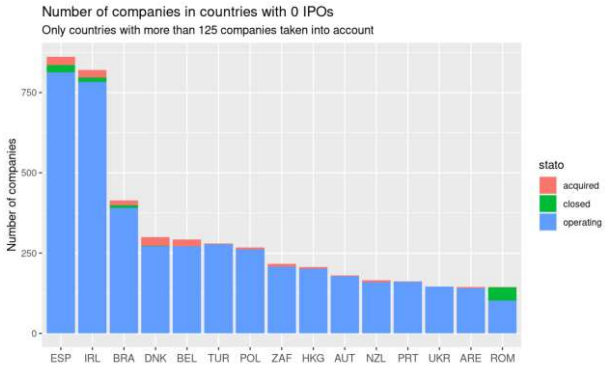

Рисунок 5. Частка діючих, закритих і придбаних компаній

Дата заснування стартапу також може вплинути на успіх/провал компанії. Наступні графіки показують розподіл цієї функції залежно від статусу(рис.6). Діаграми показують, що компанії, зареєстровані на біржі і придбані, в середньому старші за закриті або діючі.

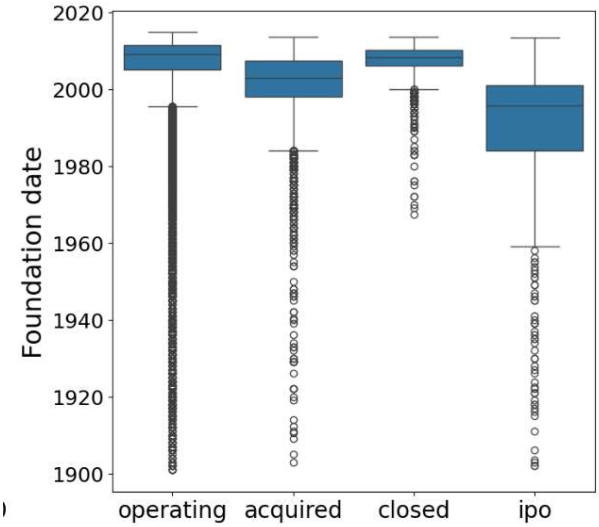

Рисунок 6. Вплив Foundation date

Ці міркування змушують нас думати, що дата заснування впливає на ймовірність успіху компанії (IPO або придбання), або, скоріше, враховуючи, що старі стартапи, як правило, котируються на фондовій біржі або придбані, і, отже, чим довше минув час з моменту заснування , вищі шанси на успіх.

Кількість інвестиційних раундів. Компанії, зареєстровані на біржі, в середньому беруть участь у більшій кількості інвестиційних раундів, тоді як збанкрутілі компанії беруть участь у кількох раундах (максимум 3). Діючі стартапи в середньому беруть участь у меншій кількості раундів, ніж у IPO, але діапазон набагато більший, ніж у останнього, із спостереженнями, що досягають 478 інвестиційних

 ISSN 1560-8956

раундів (проти максимум 37 компаній у IPO). У будь-якому випадку, основна відмінність розподілів полягає в хвостах. Насправді для всіх значень статусу 75% спостережень дорівнюють 0, але хвіст операцій набагато довший і важчий за інші.

Раунди фінансування. Більшість компаній (83,6%) не проходили навіть 1 інвестиційний раунд. 10,3% з них склали 1, з яких 61,2% не склали. Цікаво відзначити, що після 2-х раундів фінансування відсоток компаній на IPO завжди перевищує відсоток компаній-банкротів. Крім того, компанії, зареєстровані на біржі, також мають найвищий відсоток у значеннях funding_rounds.

Загальна зібрана сума. Listed та придбані стартапи отримують у середньому більше інвестицій, ніж невдалі та діючі. 32,7% невдалих стартапів не отримали фінансування, значення, яке несподівано зростає до 55,5% для компаній, що пройшли IPO, і 75,2% для придбаних. Не отримали фінансування 87,25% діючих компаній.

Досягнення компаній. Змінна має 8 рівнів, кожен з яких представляє кількість віх. Більшість стартапів (53,18%) не досягають жодної віхи, тоді як лише 6,9456% досягають більше ніж 1 віхи.

Відносини — це змінна, яка вимірює кількість відносин, які стартап має із зовнішніми суб’єктами, якими можуть бути інші компанії, інвестиційні фонди, фонди венчурного капіталу, фізичні особи тощо. Зведена таблиця показує медіану, середнє значення та діапазон (максимум і мінімум) зв’язків, обумовлених рівнями статусу. Зрозуміло, що стартапи в IPO мають набагато вищі значення, ніж інші дистрибутиви, і в ньому навіть немає відсутніх значень. Придбані компанії мають у середньому більше відносин, ніж ті, що закриті та діючі, навіть якщо в середньому закриті компанії мають на одну співпрацю більше. Різниця полягає в тому, що правий хвіст набагато важчий і довший, ніж закритий, який має найменший діапазон з 4 категорій. (рис.7)

	mean	min	25%	50%	75%	max
ipo	23.254	0.0	3.0	6.0	18.0	1189.0
acquired	3.422	0.0	0.0	1.0	4.0	363.0
closed	2.311	0.0	0.0	2.0	3.0	49.0
operating	1.664	0.0	0.0	1.0	2.0	666.0

Рисунок 7. Аналіз зв’язків

Тип фінансування раундів – це категоріальна змінна, що має 9 невпорядкованих рівнів, що вказують на тип фінансування, отриманого компанією.

Як і очікувалося, більшість фінансування є венчурним/ангельським, оскільки це перші кроки, які проходить стартап, і більшій кількості компаній вдається дожити до цих кроків. Наступні раунди: серія a, серія b і серія c. У цьому випадку ми бачимо,

що кількість компаній поступово скорочується, враховуючи, що це кроки, які повинні бути досягнуті з часом і зрілістю. Прямий капітал і краудфандинг мають низьку частоту не тому, що це ще більш просунуті кроки, а тому, що вони є «підтипами» фінансування, які можуть мати місце на інших етапах. З іншого боку, інвестиції після IPO дуже нечисленні, оскільки вони припускають, що компанія досягла IPO.

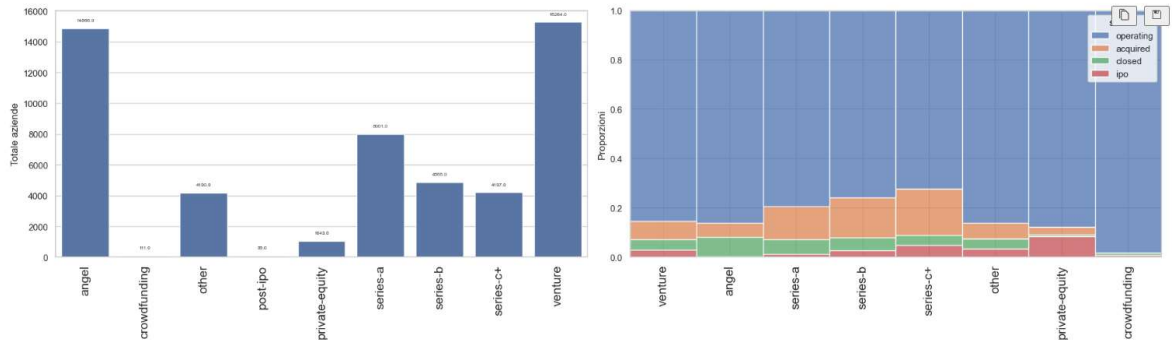

Рисунок 8. Аналіз впливу фінансування

Другий графік показує пропорції статусу для кожного виду фінансування. Тенденційно кількість придбаних компаній зростає з кожним кроком (від підприємства до серії c), тоді як кількість IPO збільшується від серії a до серії c. Значний відсоток стартапів, які отримують фінансування прямим капіталом, досягає IPO, тоді як практично всі ті, що фінансуються за допомогою краудфандингу, працюють (дуже мала частина закрита або знаходиться на IPO).

Стартапи, які закриваються, практично не відносяться до тих, що фінансуються за рахунок прямих інвестицій, у той час як вони зберігають однаковий відсоток на всіх етапах, від angel до серії c.

Результати та обговорення

Для аналізу впливу різних факторів на успішність стартапів та прогнозування їхнього виживання на ринку обрано чотири моделі: Random Forest, k-Nearest Neighbors (kNN), Decision Tree та Gradient Boosting [5-9]. Оскільки різні класи цільової змінної мають суттєво різну кількість екземплярів, це може вплинути на точність моделей класифікації. Для певних завдань знадобитися використання методів балансування класів для покращення прогнозування для менш представлених класів.

Вхідні дані містять широкий спектр характеристик стартапів, які потенційно можуть впливати на їхній статус. Для ефективного аналізу необхідно визначити найбільш релевантні ознаки, які будуть використані в моделі. Обчислена кореляція між кожною характеристикою і статусом стартапу дозволяє визначити, наскільки сильно кожна характеристика пов'язана зі статусом (рис.9).

 ISSN 1560-8956

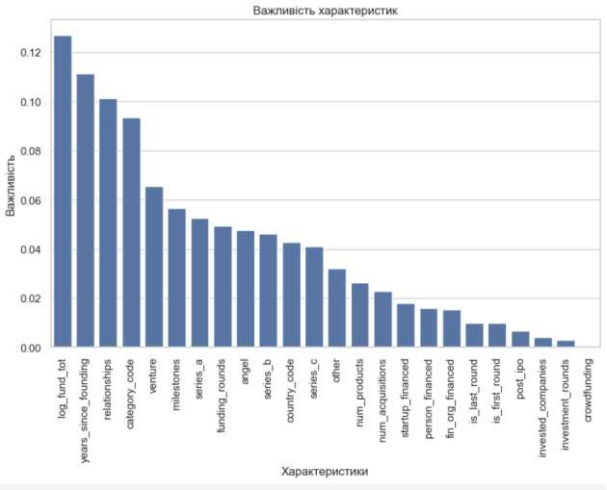

Рисунок 9. Оцінка важливості характеристик

Random Forest Classifier. За допомогою методу Grid Search з крос-валідацією були вибрані гіперпараметри для моделі Random Forest, а саме: n_estimators (кількість дерев) та max_depth (максимальна глибина дерева). Більша кількість дерев покращує продуктивність моделі, але після певного моменту додаткові дерева можуть призвести до незначного збільшення продуктивності за високу обчислювальну вартість. Параметр max_depth визначає максимальну глибину кожного дерева. Якщо цей параметр встановлено як None, дерево може рости до повної чистоти, що може призвести до перенавчання моделі. При збільшенні кількості дерев (n_estimators) F1 score зростає, але при певному значенні стабілізується. Також максимальна глибина дерева (max_depth) значно впливає на F1 score, причому None (без обмеження) дає кращі результати, ніж max_depth = 10. Отже, в результаті аналізу найкращими параметрами виявилися max_depth = None, n_estimators = 70. Найкраща модель далі була протестована на тестових даних. Отримані значення можна побачити на рис.10.

```
Best F1 Score: 0.9177412085029675
Accuracy: 0.9338253084926204
Precision: 0.9305898113935481
Recall: 0.9338253084926204
F1 Score: 0.9287059924696727
              precision    recall  f1-score   support

           0       0.92      0.70      0.80       818
           1       0.75      0.47      0.58       353
           2       0.99      0.83      0.90       173
           3       0.94      0.99      0.96      6922

    accuracy                           0.93      8266
   macro avg       0.90      0.75      0.81      8266
weighted avg       0.93      0.93      0.93      8266
```

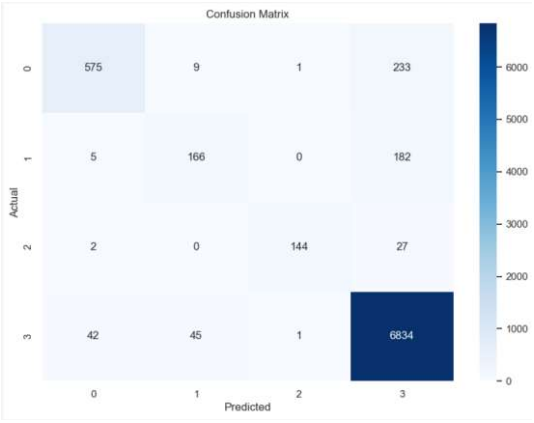

Рисунок 10. Результати для методу RandomForestClassifier

k-Nearest Neighbors. Використовуючи метод Grid Search з крос-валідацією, були відібрані найкращі параметри для моделі k-Nearest Neighbors (KNN), включаючи n_neighbors, який визначає кількість найближчих сусідів для класифікації. Після вибору оптимальних гіперпараметрів модель KNN була навчена на навчальних даних та протестована на тестових. Отримані значення можна побачити на рис.11.

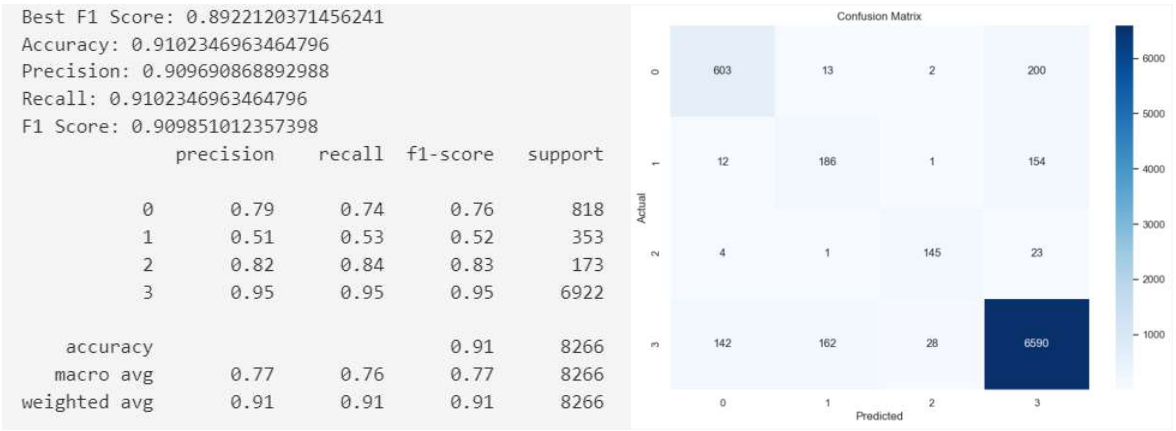

Рисунок 11. Результати для методу k-Nearest Neighbors

Decision Tree Classifier. За допомогою методу Grid Search з крос-валідацією були вибрані оптимальні гіперпараметри для моделі Decision Tree, а саме: max_depth (максимальна глибина дерева), min_samples_split (мінімальна кількість вибірок, необхідна для розгалуження вузла) та min_samples_leaf (мінімальна кількість вибірок, необхідна для листового вузла).

Зауважимо, що збільшення максимальної глибини дерева може спричинити перенавчання моделі, тоді як зменшення може призвести до недонавчання. Мінімальна кількість вибірок для розгалуження вузла (min_samples_split) і для листового вузла (min_samples_leaf) впливає на складність моделі та допомагає уникнути перенавчання. Отже, в результаті аналізу найкращими параметрами виявилися max_depth = 20, min_samples_split = 2, min_samples_leaf =1.

Після вибору оптимальних гіперпараметрів модель Decision Tree була натренована на навчальних даних, а потім протестована на тестових даних. Результати можна побачити на рис.12.

Gradient Boosting Classifier. Результати найкращої моделі GradientBoosting Classifier після використання методу Grid Search з крос-валідацією показали, що оптимальні гіперпараметри для моделі - max_depth=10, n_estimators=110. Параметр n_estimators визначає кількість базових моделей, що використовуються в градієнтному підсиленні. Зазвичай більша кількість базових моделей дозволяє підвищити ефективність моделі, але при цьому може збільшуватися час навчання та ризик перенавчання. Параметр max_depth визначає максимальну глибину кожного дерева

рішень у композиції. Встановлення цього параметра допомагає управляти рівнем складності моделі та ризиком перенавчання.

```
Best F1 Score: 0.8823689547412787
Accuracy: 0.8988628115170578
Precision: 0.8971083095679437
Recall: 0.8988628115170578
F1 Score: 0.8978647392122913
              precision    recall  f1-score   support

           0       0.73      0.68      0.70       818
           1       0.50      0.49      0.49       353
           2       0.82      0.85      0.83       173
           3       0.94      0.95      0.94      6922

    accuracy                           0.90      8266
   macro avg       0.75      0.74      0.74      8266
weighted avg       0.90      0.90      0.90      8266
```

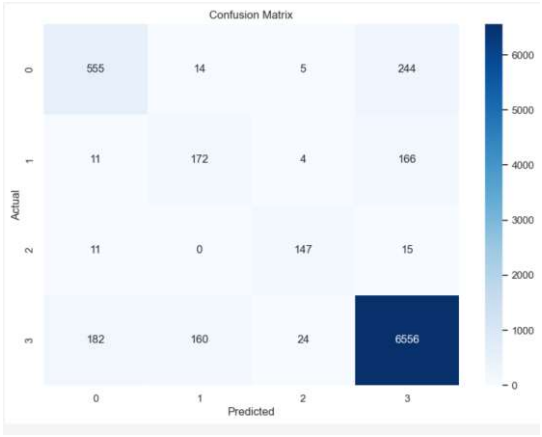

Рисунок 12. Результати для методу DecisionTreeClassifier

Найкраща модель була протестована на тестових даних для оцінки її продуктивності та точності прогнозування, результати представлені на рис.13.

```
Best F1 Score: 0.9162191127804282
Accuracy: 0.9310428260343576
Precision: 0.9280878930108792
Recall: 0.9310428260343576
F1 Score: 0.9246951691044125
              precision    recall  f1-score   support

           0       0.93      0.68      0.78       818
           1       0.79      0.43      0.56       353
           2       0.96      0.86      0.91       173
           3       0.93      0.99      0.96      6922

    accuracy                           0.93      8266
   macro avg       0.90      0.74      0.80      8266
weighted avg       0.93      0.93      0.92      8266
```

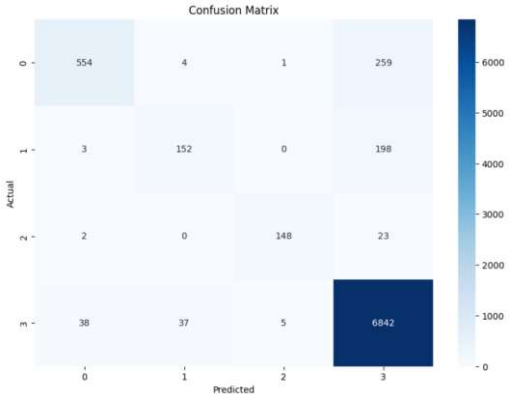

Рисунок 13. Результати для методу GradientBoostingClassifier

Порівняння ефективності методів. Основною метрикою для визначення точності моделі у цьому дослідженні є зважена F-міра. F-міра представляє собою гармонійне середнє точності і повноти, що дозволяє отримати одну числову оцінку, яка враховує як правильність передбачень, так і здатність моделі виявляти всі позитивні приклади. Зважена F-міра враховує підтримку кожного класу, що робить її більш корисною для незбалансованих наборів даних, у порівнянні зі звичайною точністю. Використання звичайної точності в таких ситуаціях може бути недостатнім, оскільки модель може показувати високу точність, передбачаючи лише найбільш чисельний клас, ігноруючи менш представлені класи.

На основі отриманих результатів f1-score можна зробити висновок, що найоптимальнішим методом для прогнозування успішності стартапу є модель Random

Forest з результатом f1-score = 0.93. Проте загалом кожна модель показала досить гарні результати.

Висновки

Сьогодні стартапи відіграють важливу роль у світовій економіці, виступаючи як каталізатори інновацій та економічного зростання. Стартапи представляють собою нові підприємства, які прагнуть створювати інноваційні продукти, послуги чи бізнес-моделі, часто з метою заповнення ринкових ніш або кардинальної зміни існуючих ринкових умов. Вони мають потенціал не лише для значних фінансових прибутків, але й для формування нових галузей та зміни економічного ландшафту.

Проте, високий рівень невизначеності та ризиків, які супроводжують стартапи, призводять до того, що більшість з них зазнають невдач у перші кілька років свого існування. Згідно з дослідженнями [2], лише близько 10% стартапів виживають і досягають довгострокового успіху. Тому аналіз факторів, що впливають на успішність стартапів, є критично важливим для розробки стратегій їх підтримки та розвитку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Startup Investor Academy. [Електронний ресурс] – URL: https://www.startupinvestoracademy.com/

2. Kaggle Learn. [Електронний ресурс] – URL: https://www.kaggle.com/datasets/justinas/startup-investments

3. Seaborn: statistical data visualization. [Електронний ресурс] – URL: https://seaborn.pydata.org/

4. Pandas documentation – pandas 2.2.2 documentation. pandas - Python Data Analysis Library. [Електронний ресурс] – URL: https://pandas.pydata.org/docs/

5. Т. А. Ліхоузова, Теорія імовірностей та математична статистика – 2018, КПІ ім. Ігоря Сікорського. [Електронний ресурс] – URL: https://ela.kpi.ua/handle/123456789/23168

6. Rajender Kumar, Mastering Data Analysis with Python. – Jamba Academy, 2023.

7. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning – Springer, 2017.

8. Steven S. Skiena, The Data Science Design Manual – Springer, 2017.

9. Tomas Hrycej, Bernhard Bermeitinger, Matthias Cetto, Siegfried Handschuh, Mathematical Foundations of Data Science – Springer, 2017.