УДК 303.732.4
DOI: 10.20535/SRIT.2308-8893.2019.3.02

ДОСЛІДЖЕННЯ І ПРОГНОЗУВАННЯ УСПІШНОСТІ СТАРТАПІВ ПЛАТФОРМИ KICKSTARTER

Н.В. КУЗНЄЦОВА, Я.В. ГРУШКО

Анотація. Основна мета дослідження, проведеного у роботі, — виявлення та прогнозування успішності нових проектів-стартапів. Розв'язано задачу прогнозування факту успішності того чи іншого стартапу, застосовано різні методи інтелектуального аналізу даних, такі як методи екстремального градієнтного бустингу та k -найближчих сусідів, що дало змогу з високою точністю передбачити успішність проекту, а найефективнішим виявився метод екстремального градієнтного бустингу. Використання моделей виживання дозволило оцінити середній час роботи над успішним стартапом і визначити ключові галузі, для яких стартапи стають ефективними; спрогнозовано для кожного з них необхідний час роботи для втілення прогресивної ідеї в успішний бізнес. Виявлено найбільш успішні категорії проектів-стартапів та спрогнозовано час, необхідний як у цілому для досягнення успішності (виживання) проектів, так і для окремих категорій проектів. Для цього побудовано моделі виживання на основі пропорційних ризиків Кокса та моделі Каплан–Майєра.

Ключові слова: прогнозування, метод ектремального градієнтного бустингу, метод *k*-найближчих сусідів, моделі виживання, стартапи, успішність проектів, платформа kickstarter.

ВСТУП

Сучасний світ є настільки відкритим і динамічним, що будь-яка ідея, відкриття, технологія можуть бути реалізовані та впроваджені дуже швидко. Це зумовлює не лише швидкий розвиток і передавання нових тенденцій між різними галузями, людьми, народами, континентами, а і зміну навколишнього буття, світогляду та уподобань людей, зацікавленість їх у нових продуктах та послугах, появу нових потреб та розвиток нових звичок і захоплення новими трендами, напрямами, які до цього навіть не мали жодної перспективи для розвитку. Саме відкритість світу і можливість використання інтернет-технологій спричиняють такий розвиток та появу платформ і майданчиків для презентації та обміну ідеями, спеціальні фонди, підтримання і залучення всесвітньо відомих компаній і корпорацій, інвесторів, краудфандингу тощо. Виникнення і поширення однієї технології може не лише досягти великого успіху за короткий час, але і «згаснути», тобто втратити ключові позиції на ринку або основних прихильників.

Найпоширеніший і найбільш відомий термін для визначення прогресивних ідей та думок наразі є стартап. Стартапами часто називають невеликі бізнес-компанії, ядром яких є якась така «особлива» ідея, що гіпотетично зможе перетворити бізнес у великий мільярдний (у доларах США) бізнес. Реалізація стартапів стала дуже поширеним рухом; його розвивають студенти, молодь, прогресивні люди, захоплені і натхненні певною чудовою ідеєю,

які не мають достатнього фінансування для негайного впровадження її у широке виробництво, але зацікавлені її реалізувати як власну, а не під брендом певної відомої компанії.

Компанії, основним зерном яких була «особлива» ідея, створювалися і існували понад 100 років. Є навіть статистика, виконана аналітиками з оцінювання і прогнозування стартапів, яка вказує, що насправді ймовірність поразки стартапу дуже велика. У 2017 р. американський діловий журнал Fortune [1] оцінив такі проекти стартапів і навів вражаючі факти: 90% стартапів в кінцевому підсумку зазнають невдачі. Тобто основна частина таких «ідейних» бізнесів не реалізується, або не знаходить достатньої кількості споживачів, підтримання, фінансування. Це вже не перші такі дослідження ділового журналу. Ця тема є актуальною і досі, а у своєму дослідженні Fortune відзначає [2], що основними причинами невдачі стартапів у 2014 р. були: відсутність потреби продукту у користувачів; нестача коштів у засновників; недостатня згуртованість команди розробників та генераторів ідей; порушення певних законів, наприклад, порушення прав конфіденційності користувачів тощо.

ПОСТАНОВКА ЗАДАЧІ

Головною ідеєю роботи стали дослідження та прогнозування популярних і найбільш «успішних» проектів-ідей, реалізованих у вигляді стартапів та зареєстрованих на платформі kickstarter. Мета дослідження — виявлення основних тенденцій, притаманних успішним стартап-проектам, оцінювання та прогнозування періоду їх успішності та необхідні витрати на їх розвиток; на підставі виконаного аналізу успішних проектів визначення ключових галузей та напрям ідей, які можуть бути перспективними для розвитку і впровадження у великий бізнес. Для такого аналізу обрано реальний набір даних і визначено методи інтелектуального аналізу даних (ІАД), такі як методи класифікації, градієнтного бустингу із застосуванням дерев та методи аналізу виживання для прогнозування часу успішності стартапів. Спрогнозувати, чи буде проект успішним, і якщо так, то який період часу потрібен для фінансування і в якій галузі.

МЕТОДИ ТА ІНСТРУМЕНТИ ІАД, ВИКОРИСТАНІ У РОБОТІ

Особливості поставленого завдання зумовлюють необхідність вирішення одразу декількох завдань різними методами ІАД, тому важливим є обрання коректного інструментарію і математичного апарату. Для визначення успішності проекту для обраного набору даних має бути розв’язана задача класифікації, тобто віднесення проекту до успішних або провальних. Для задачі класифікації пропонується використати найбільш відомий і достатньо ефективний метод k-найближчих сусідів та метод екстремального градієнтного бустингу. Для визначення оптимального часу, необхідного для реалізації стартапу, тобто для орієнтування розробників на необхідний час роботи і інвестування проекту, щоб не зупинитися «за крок до реалізації мрії», вирішується задача прогнозування часу з використанням моделей виживання.

Метод *k*-найближчих сусідів (*k*-NN)

У розпізнаванні образів метод *k*-найближчих сусідів (*k*-nearest neighbor method) є непараметричним методом, що використовується для класифікації та регресії [3]. В обох випадках вхід складається з k найбільш близьких прикладів навчання у просторі ознак. Вихід залежить від того, чи використовується *k*-NN для класифікації або регресії:

- У класифікації *k*-NN вихід є класом, до якого належить вхідний об’єкт. При цьому об'єкт призначається класу, найбільш поширеному серед його найближчих сусідів (k — ціле додатне число, зазвичай невелике). Якщо $k=1$, то об'єкт просто присвоюється класу того самого найближчого сусіда.
- У *k*-NN регресії вихід є значенням властивості для об'єкта. Це значення є середнім значенням для k найближчих сусідів.

Тобто в основу методу покладено використання відстані (зазвичай евклідової метрики) між вхідним об’єктом та вже промаркованими (навченими) сусідами (об’єктами).

Приклад класифікації методом *k*-NN зображено на рис. 1. Вхідний об’єкт (точку) слід класифікувати або до першого класу квадратів, або до другого класу трикутників. Якщо $k=3$ (суцільне коло лінії), то він присвоюється другому класу, оскільки у внутрішньому колі є 2 трикутники і лише 1 квадрат. Якщо $k=5$ (коло пунктирної лінії), то він призначається першому класу (3 квадрати проти 2 трикутників всередині зовнішнього кола).

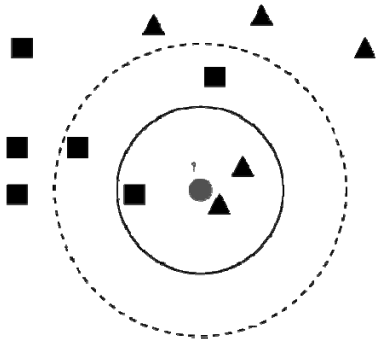

Рис. 1. Приклад класифікації *k*-NN

Цікаво, що алгоритм *k*-NN є одним з найпростіших алгоритмів машинного навчання і при цьому він є одним з найефективніших (відповідне порівняння, наведено в документації [4] популярної бібліотеки *sklearn* для мови програмування *python*).

Метод екстремального градієнтного бустингу (XGBoost)

Метод екстремального бустингу XGBoost (eXtreme Gradient Boosting) [5] реалізований у вигляді бібліотеки з відкритим вихідним кодом, доступний на найбільш поширених і широко вживаних мовах програмування, таких як C ++, Java, Python [6], R і Julia.

Особливістю методу є те, що він забезпечує паралельний бустинг дерев (також відомий як GBDT, GBM) і дозволяє розпаралелювати, а тим самим пришвидшувати розв’язання задачі порівняно з відомим методом градієнт-

ного бустингу, а сам метод ще називають градієнтним бустингом із застосуванням дерев.

Градієнтний бустинг використовують у регресійних і класифікаційних задачах як техніку, суть якої полягає в тому, щоб збудувати з ансамблю слабких моделей прогнозування (зазвичай моделей дерев рішень) одну, але точну та ефективну модель.

Алгоритм реалізації методу градієнтного бустингу можна подати у такому вигляді [7, 8].

Задано: навчальну вибірку $\{(x_i, y_i)\}_{i=1}^n$, функцію витрат $L(y, F(x))$, кількість ітерацій (кількість слабких моделей) — M.

1. Ініціалізуємо модель константою ($F_0(x)$):

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) .$$

2. Від $m = 1$ до M :

- обчислюємо псевдозалишки (pseudo-residuals):

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)};$$

- навчаємо чергову слабку модель (наприклад, дерево) $h_t(x)$ отриманими псевдозалишками, тобто навчаємо такою вибіркою: $\{(x_i, r_{im})\}_{i=1}^n$;

- обчислюємо множник γ_m, розв'язуючи однорозмірну оптимізаційну задачу:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)) ;$$

- оновлюємо модель:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) .$$

3. Виводимо $F_m(x)$.

Бібліотекою XGBoost почала користуватися велика кількість розробників, що сприяло популярності методу серед спільноти Kaggle (платформа, де змагаються вчені з науки про дані). Метод дозволяє отримати одні з найкращих результатів прогнозів, а тому цей метод обрано в цій роботі для прогнозування успішності проектів стартапів [10].

Для аналізування та прогнозування необхідного часу роботи для успішного запуску стартапу використаємо різні моделі з теорії виживання.

Моделі виживання

У загальному вигляді функція виживання подається у вигляді [11]

$$S(t) = P(T > t) , \text{ де } 0 \le S(t) \le 1 .$$

Предметом дослідження є визначення ймовірності виживання та провалу (або смерті) проекту, що відбувається в момент часу t з урахуванням того, що подія смерті не відбувалася до часу t. Математично це подано таким чином [11, 12]:

$$\lim_{\Delta \to 0} P(t \le T \le t + \Delta \mid T > t).$$

Ця подія відбувається у безмежно малий проміжок часу (від t до $t + \Delta t$), тож отримуємо формулу, що є функцією ризику:

$$h(t) = \lim_{\Delta \to 0} \frac{P(t \le T \le t + \Delta \mid T > t)}{\Delta}.$$

Це ідентично такому виразу:

$$h(t) = -\frac{\frac{dS(t)}{dt}}{S(t)}.$$

У наведених позначеннях та припущеннях можна виконати формалізацію реальної практичної задачі оцінювання ризиків.

Відома модель Д. Кокса [13], запропонована в 1972 р., заснована на припущенні, що функцію ризику можна факторизувати, тобто подати у вигляді добутку двох функцій:

$$h_i(t) = h_0(t) \cdot \psi(x_{i1},...,x_{ik}),$$

де $h_0(t)$ — базова функція інтенсивності, що включає фактор часу, але не включає коваріанти, а $\psi(x_{i1},...x_{ik})$ — лінійна функція досліджуваних ознак, яка не включає фактор часу.

Досить часто модель записують у такому вигляді [12, 13]:

$$h_i(t) = h_0(t) e^{\{\beta_1 x_{i1} + ... + \beta_k x_{ik}\}};$$

$$\ln h_i(t) = \ln h_o(t) + \beta_1 x_{i1} + ... + \beta_k x_{ik},$$

де $\beta_1,...,\beta_k$ — невідомі параметри, а x_i є вхідними змінними (тобто стовпчиками) у вибірці.

Модель пропорційних ризиків Кокса у вигляді функції умовного виживання $S(t \mid x)$ передбачає оцінку сукупної умовної функції ризику $L(t \mid x)$ з використанням максимальної правдоподібності.

ПІДГОТОВКА ТА ОБРОБЛЕННЯ ВХІДНИХ ДАНИХ ДО МОДЕЛЮВАННЯ

Вхідними даними для моделювання обрано набір даних [10] за стартапами, зареєстрованих на платформі кікстартер [14]. Набір містить 13 змінних (стовпчиків):

ID: ID кожного клієнта;

name: назва стартапу;

category: детальний опис категорії, у якій функціонує стартап (наприклад, їжа, документальна література);

main_category: загальна категорія діяльності стартапу, ширша ніж просто категорія (наприклад, східна кухня, література);

currency: валюта проекту;

launched: дата та час початку роботи над стартапом;

deadline: останній термін (дата та час) закінчення роботи над стартапом.

Оскільки час початку роботи над проектом (launched) у всіх стартапів різний (у когось 2014 р., у когось 2017 р.), то задля кращої репрезентатив-

ності даних будемо створювати агрегований показник, тобто сформуємо нову змінну «термін роботи» як різницю між кінцем і початком роботи над проектом: **time_spent = deadline – launched ,** а саме:

goal: сума грошей, яку має на меті зібрати стартап;

pledged: сума грошей, яку вже вкладено в стартап;

state: результат стартапу (0 — провалений, 1 — успішний);

backers: кількість людей в команді;

country: країна, у якій зареєстровано стартап;

usd_pledged: сума доларів США, яку вже вкладено в стартап.

Вхідними даними обрано для початку 3743 записів із загальних 378600, які містять 9 стовпчиків. Усі дані переведено в числа для розв'язання задачі класифікації (табл. 1).

Таблиця 1. Перші рядки з набору даних

Category	Main_ category	Currency	Goal	Pledged	State	Backers	Country	Usd_ pledged
108	12	5	7	0	0	0	9	0
93	6	13	3908	24883	0	670	22	2015
93	6	13	5325	22364	0	2014	22	39948

Проаналізуємо вхідну вибірку з 3743 унікальних стартапів:

- середнє значення суми грошей, що вже вкладено в стартап (usd_pledged), — 43911 дол. Стандартне відхилення велике (34961 дол.), максимальне значення 109021 дол.;
- середня кількість людей у команді стартапу: 1585;
- середній час, витрачений на роботу команди над стартапом: 38 днів; максимальний час: 1284 дні; мінімальний — 2 год (так, є і такі проекти!);
- оскільки значення 0 для state означає «стартап провалився», а значення 1 означає «стартап успішний», середнє значення 0,375 означає, що лише 37,5% стартапів з усієї вибірки були успішними;
- пропущених даних у вибірці немає.

Матрицю кореляції між вхідними змінними зображено на рис. 2.

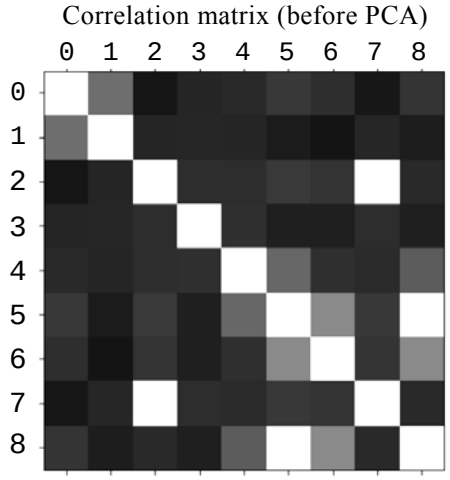

Рис. 2. Матриця кореляції між вхідними змінними

Оскільки існує сильна кореляція (чим світліші квадратики, тим більша кореляція) між деякими вхідними значеннями, тому застосуємо метод головних компонент (Principle Component Analysis (PCA)) для зменшення кореляції та кількості вхідних змінних. Матрицю кореляції вхідних змінних після застосування PCA зображено на рис. 3.

Після застосування методу суттєві сплески кореляції між вхідними значеннями згладились, а кількість змінних зменшилась із 10 до 4, але при цьому, як було перевірено далі, точність значно знизилась (показник становив ROC_AUC=0,85, а став 0,66), але швидкість оброблення збільшилась. Тож було прийнято рішення не використовувати PCA у подальшій роботі.

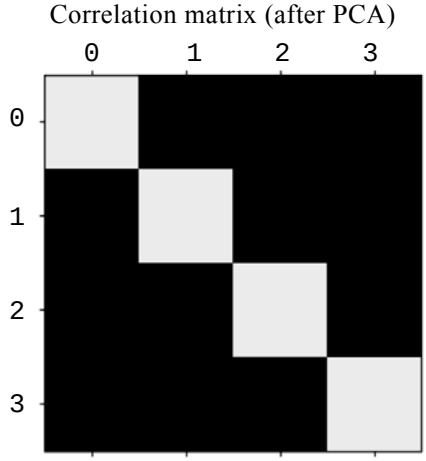

Рис. 3. Матриця кореляції між вхідними змінними після застосування PCA

Вхідну вибірку розділено у співвідношенні 70% для навчальної та 30% для тестової. Для перевірки наближеності тестової вибірки до прогнозованої використовується індекс Жаккара, що є стандартним критерієм у мові програмування python.

Коефіцієнт подібності Жаккара обчислюється за формулою

$$K_j = \frac{c}{a+b-c},$$

де a — кількість видів на першому пробному майданчику; b — кількість видів на другому пробному майданчику; c — кількість видів, спільних для першого та другого майданчиків.

РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ

Для прогнозування факту успішності стартапу застосовано метод KNeigboursClassifier (*k*-NN), результати роботи якого подано у вигляді матриці помилок (Confusion Matrix) на рис. 4.

Для оцінювання точності прогнозування використовувався показник площі під кривою ROC, який для даного методу визначився на рівні 0,74. Поріг, обраний для оцінювання і побудови матриці помилок, становив 0,6, тобто якщо ймовірність успіху більша за 0,6, то стартап вважається успішним.

Можна бачити, що в методі *k*-NN відношення помилок неправильної класифікації позитивних випадків до неправильно виявлених негативних становить FP(False Positive)/FN(False Negative) $=1,3$. Це означає, що класифікатор прогнозуватиме помилково успіх частіше, ніж помилково провал стартапу, а тому така модель є некоректною для розгляданої задачі.

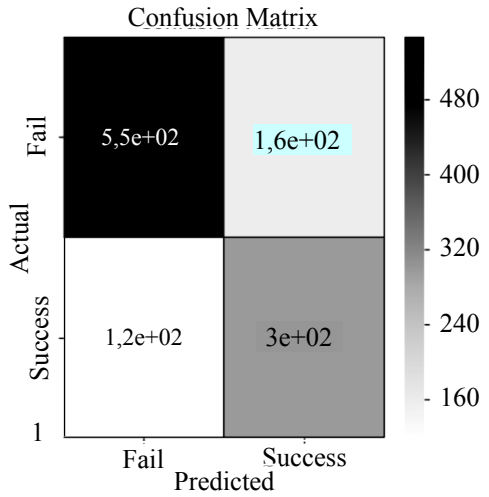

Рис. 4. Confusion Matrix для методу *k*-NN з порогом 0,6

Виконано також велику кількість експериментів з різними порогами і емпірично виявлено, що оптимальний поріг для прогнозування ймовірності успіху стартапу становить 0,7. Матрицю помилок для *k*-NN і порога 0,7 зображено на рис. 5.

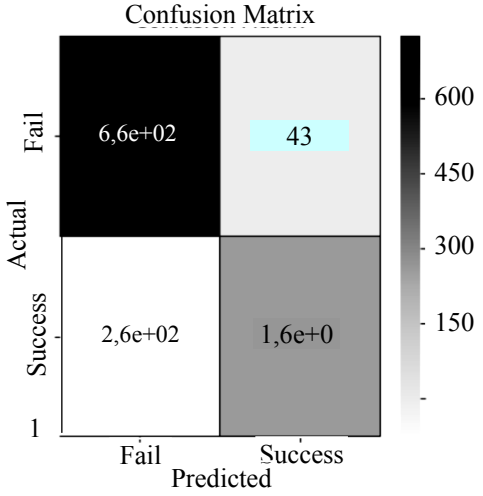

Рис. 5. Confusion Matrix для методу *k*-NN з порогом 0,7

Площа під кривою ROC зменшилась до 0,65, але це вимушений крок, оскільки довелося накласти додаткові жорсткіші вимоги, щоб наблизити модель до реальних умов.

За методом екстремального градієнтного бустингу XGBoost визначено модель, для якої індекс Жаккара виявився вищим, ніж у попереднього кла-

сифікатора, і дорівнює 0,85, що свідчить про більшу коректність моделі XGBoost.

Показник ROC_AUC на тестовій вибірці становив 0,92, що вищий**,** ніж у класифікатора *k*-NN, а сам індекс свідчить про високу предикативну здатність такої моделі.

Для моделі XGBoost установлено кількість ітерацій на рівні 70, а кількість епох, за якими неможливе покращення, EARLY_STOP $= 70$, при цьому ROC_AUC на тестовій вибірці становив 0,85. Емпірично підібрано кращі параметри: кількість ітерацій збільшено до 1000, а EARLY_STOP зменшено до 50. За таких параметрів на тестовій вибірці ROC_AUC $= 0{,}92$.

Для методу екстремального градієнтного бустингу також експериментально підібрано поріг для покращення репрезентативності моделі прогнозу успіху стартапів і оптимальне значення порога виявилось 0,6, проте точність моделі за індексом ROC_AUC дещо знизилась до 0,83, але є вищою порівняно з попереднім методом. Відношення FP(False Positive)/FN(False Negative) для XGBoost становить 0,6, що є більшим порівняно з методом *k*-NN зі значенням 0,16 (рис. 6). Порівняння обох методів наведено у табл. 2.

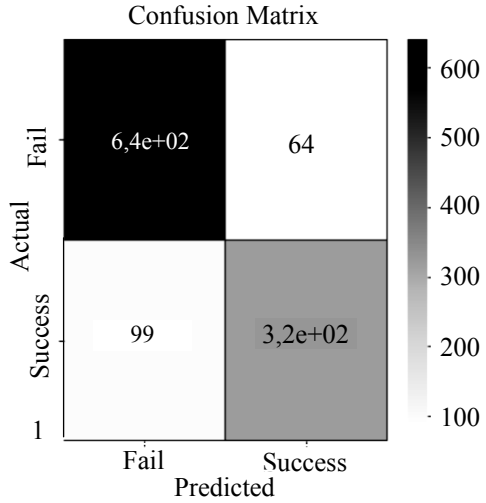

Рис. 6. Матриця помилок для методу екстремального градієнтного бустингу з порогом 0,6

Таблиця 2. Порівняльна таблиця результатів методів за різними критеріями ефективності та порогом

Method	**Index**				
	Jaccard Index	**ROC_AUC**	**ROC_AUC_ optimum**	**FP/FN**	**Поріг**
KNeigbours	0,75	0,74	0,65	0,16	0,7
XGBoost	0,85	0,92	0,83	0,6	0,6

Отже, класифікатор XGBoost виявився кращим за основними критеріями і його доцільно застосовувати для прогнозування факту успішності проекту.

ПРОГНОЗУВАННЯ ЧАСУ УСПІШНОСТІ СТАРТАПУ МОДЕЛЯМИ ВИЖИВАННЯ

Для вхідного набору даних необхідно спрогнозувати час успішності стартапу, для цього ключовими змінними будуть час time_spent (який відповідає часу в моделі Кокса) та state (має значення {0;1}, 1 — якщо відбулась подія успішного стартапу і 0 — якщо провал).

Для побудови моделі Кокса використано python-бібліотеку lifelines. Результати оцінювання значень вхідних змінних та коефіцієнтів для моделі виживання Кокса наведено у табл. 3.

Таблиця 3. Значення вхідних змінних для моделі Кокса

Variable	Coefficient							
	coef ex	**p(coef)**	**se(coef)**	***z***	***P***	**-log2(p)**	**lower 0.95**	**upper 0.95**
category	0	1	0	0,72	0,47	1,09	0	0
main_category	-0,02	0,98	0,01	-3,36	<0,005	10,3	-0,04	-0,01
currency	0,1	1,1	0,05	1,94	0,05	4,26	0	0,2
goal	0	1	0	-3,06	<0,005	8,82	0	0
pledged	0	1	0	2,17	0,03	5,06	0	0
backers	0	1	0	11,61	<0,005	101,04	0	0
country	-0,05	0,96	0,03	-1,51	0,13	2,92	-0,11	0,01
usd_pledged	0	1	0	3,5	<0,005	11,09	0	0

Для перевірки предикативної здатності та валідації моделі виживання використовується показник Concordance Index, а прийнятними вважаються моделі, що мають значення індексу від 0,55 до 0,75. Для моделі Concordance Index = 0,63 відношення правдоподібності — 306,07, а — $\log2(p) = 201{,}56$, тому модель Кокса може використовуватись для подальшого аналізу.

Значення коефіцієнтів у побудованій моделі Кокса подано у вигляді рис. 7.

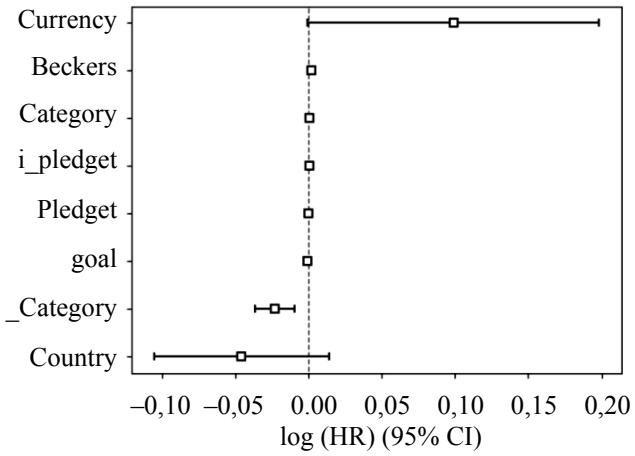

Рис. 7. Значення коефіцієнтів у побудованій моделі Кокса

Далі побудовано модель Каплана–Майєра, графік функції виживання якої зображено на рис. 8.

Порівняння функцій виживання моделей Каплан-Майєра та базового рівня виживання Кокса проілюстровано на рис. 9.

Бачимо, що криві збігаються, хоча з часом дещо розходяться, що є коректним, оскільки чим більше базова лінія в моделі Кокса відрізняється від моделі Каплан–Майєра, тим краще, адже тим більший корисний внесок роблять вхідні змінні, що містяться під експонентою у відповідній формулі моделі Кокса.

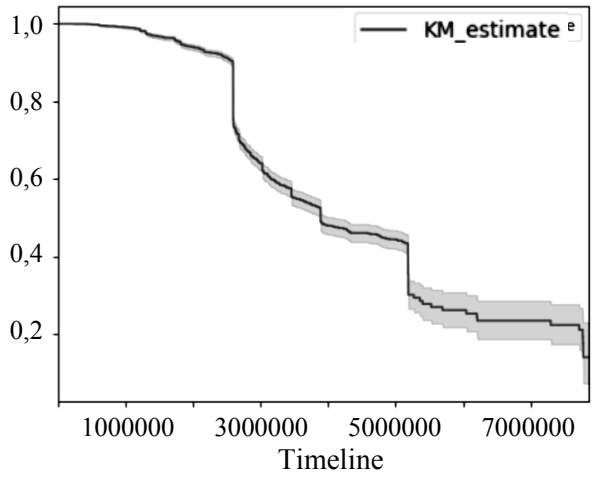

Рис. 8. Графік виживання моделі Каплана–Майєра в часі (у секундах)

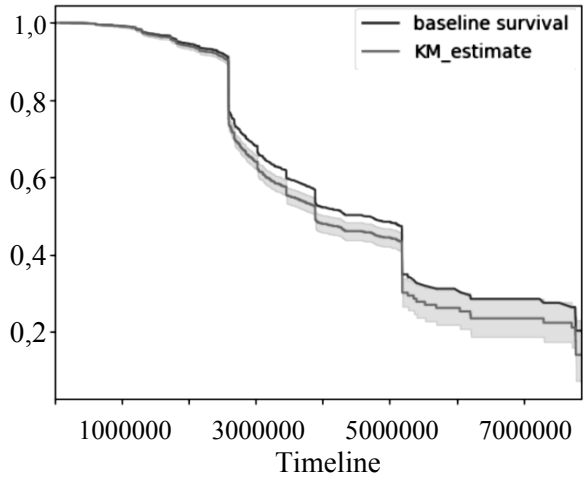

Рис. 9. Порівняння моделей виживання

Оскільки необхідно було виявити стартапи, які найімовірніше стають успішними, то виконувалася стратифікація моделі Кокса за основними категоріями (рис. 10).

Отже, найбільш вдалими є стартапи, реалізовані у таких категоріях, як Crafts, Journalism, Dance, Fashion, Comics, а ймовірності виживання для них становлять відповідно від 0,65 до 0,5. Слід відзначити, що після 60 днів стартапи із цих категорій перестають «вмирати», тобто можемо зробити висновок, що час успішності для стартапів категорії становить понад 60 днів.

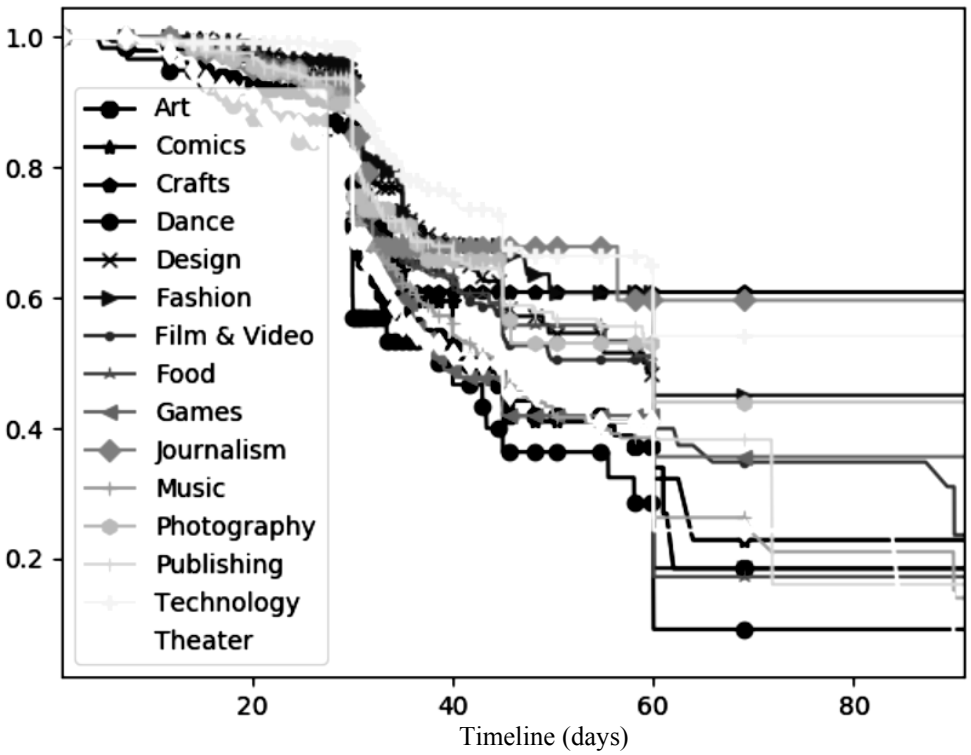

Рис. 10. Стратифікована за стовпчиками *main_category* модель Кокса

МОДЕЛЮВАННЯ ЧАСУ УСПІШНОСТІ СТАРТАПІВ ЗА КАТЕГОРІЯМИ

Для визначення часу, необхідного для розроблення і запуску стартапу, який досягне успіху, було вирішено виконати подальше моделювання. Оскільки вхідна вибірка на платформі велика — 378 000 різних стартапів, то було обрано перші 7486 унікальних стартапів, які розбито в рівних частинах по 3743 — на навчальну та тестову вибірки.

Для аналізу і побудови моделі Кокса використано навчальну вибірку 3743 стартапи і спрогнозовано час життя стартапів за когортами (категоріями/ напрямами стартапів) для другого набору для перевірної вибірки.

Для кожної категорії стартапів спрогнозовано час життя для тестового набору (рис. 11); реальні значення функції виживання для тестового набору подано у вигляді рис. 12.

Отже, результати прогнозування часу для різних категорій стартапів за допомогою функцій виживання показали, що найтриваліший час життя мають такі категорії стартапів з тестового набору, як технології, фільми і відео, фотографія, мода та публікації. Імовірність виживання для цих категорій є не нижчою, ніж 0,4 протягом усього часу спостереження. Крім цього, після 60 днів стартапи таких категорій перестають «вмирати», тобто починають окупатися і переходити в успішний бізнес, що дає дохід розробникам. Аналогічні результати спостерігались і на навчальному наборі, проте для дещо інших категорій (технології, фільми та відео, фотографії, мода та публікації). Моделлю Кокса спрогнозовано, що всі категорії стартапів проходять відмітку 40 днів, найуспішнішими стартапами, що проходять відмітку 60 днів, є такі категорії: Journalism, Technology, Crafts, Fashion, Photography.

Порівнявши з реальними даними, маємо, що прогноз більш-менш точний, оскільки 3 з 5 збігаються з реальними значеннями (Technology, Fashion, Photography), а ще 2 з 5 збігаються з попереднім набором даних (Journalism, Fashion). Отже, одними з найуспішніших категорій для стартапів за проведеним дослідженням можна вважати Fashion, Technology, Journalism, Photography. Цікавим є те, що виділився серед них Fashion, потрапивши до всіх вибірок успішних категорій (і до навчальної, і до тестової, і до реальної).

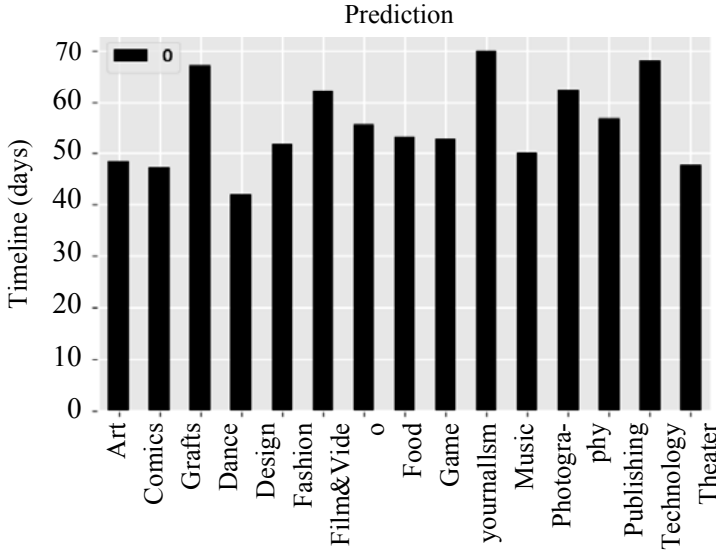

Рис. 11. Прогноз часу життя за категоріями для тестового набору

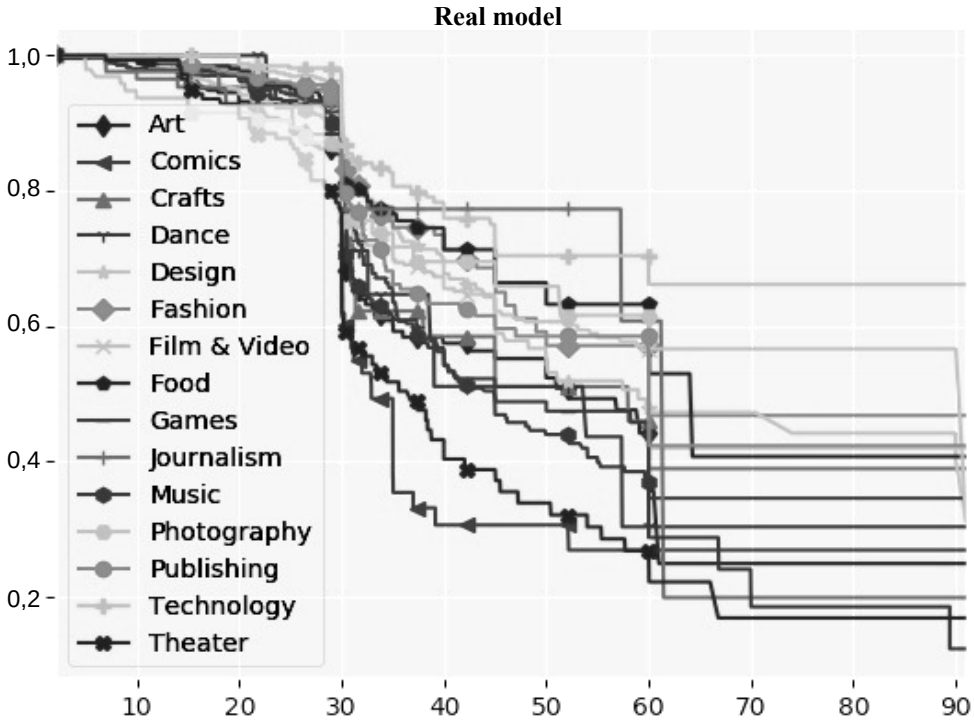

Рис. 12. Реальні значення функції виживання для тестового набору даних за категоріями

ВИСНОВКИ

Динаміка зміни технологій і уподобань клієнтів викликає появу великої кількості нових стартапів. Успіх тієї чи іншої бізнес-ідеї істотно залежатиме від удало підібраного моменту, відсутності подібних конкурентних пропозицій, часу її реалізації у вигляді стартапу та навіть галузі, у якій її буде запропоновано. Проведене у роботі дослідження сучасних успішних та провальних проектів дає змогу зорієнтуватись, для якої галузі натепер є актуальними нові стартапи. За допомогою сучасних методів інтелектуального аналізу даних визначено кращі методи для оцінювання і прогнозування успішності стартапів. Застосовані методи екстремального бустингу та *k*-найближчих сусідів дозволили з високою точністю передбачити успішність проекту, а моделі виживання — оцінити середній час роботи над успішним стартапом і визначити саме ті ключові галузі, для яких стартапи стають ефективними, спрогнозувавши для кожного з них необхідний час роботи для втілення прогресивної ідеї в успішний бізнес. Проведене дослідження буде корисним для молодих винахідників, які хочуть втілити власну ідею у життя і планують започаткувати власний бізнес, а також для інвесторів, які прагнуть підтримувати нові проекти і хотіли б швидкого повернення та примноження вкладених коштів, а отже, зацікавлені у пошуку потенційних проектів, які не матимуть конкурентів.

У КПІ імені Ігоря Сікорського також наявна відома широкому загалу платформа для розвитку стартапів і потенційних інвесторів [15], яка є сходинкою для залучення початкового фінансування на розроблення власних ідей. Для виявлення галузей, що потребують змін та втілення нових ідей, напрацювання рекомендацій новачками, які тільки хочуть організовувати власний бізнес для реалізації своєї прогресивної ідеї, також може бути корисним таке дослідження вже реалізованих стартапів і аналізу їх доцільності й ефективності за різними напрямами та сферами.

ЛІТЕРАТУРА

1. *Conventional* Wisdom Says 90% of Startups Fail. Data Says Otherwise // Fortune. — Updated June 2017. — Available at: http://fortune.com/2017/06/27/startup-advice-data-failure/
2. *Why startups* fail, according to their founders // Fortune. — Updated September 2014. — Available at: http://fortune.com/2014/09/25/why-startups-fail-according-to-their-founders/
3. *Altman N.S.* An introduction to kernel and nearest-neighbor nonparametric regression / N.S. Altman // The American Statistician. — 1992. — P. 175–185.
4. *Classifier* comparison // Scikit-learn. — Updated 2018. — Available at: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
5. *XGBoost* (eXtreme Gradient Boosting) // Distributed (Deep) Machine Learning Community. — Updated 2016. — Available at: https://github.com/dmlc/xgboost.
6. *Xgboost* 0.82 // Python Package Index (PyPI). — Updated 2019. — Available at: https://pypi.org/project/xgboost/.

7. *Friedman J.H.* Greedy Function Approximation: A Gradient Boosting Machine / J.H. Friedman // Reitz Lecture. — 1999.
8. *Hastie T.* 10. Boosting and Additive Trees / T. Hastie, R. Tibshirani, J.H. Friedman // The Elements of Statistical Learning. — 2009. — N 2. — P. 337–384.
9. *XGBoost* (eXtreme Gradient Boosting) // Distributed (Deep) Machine Learning Community. — Updated 2016. — Available at: https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions.
10. *Kickstarter* projects // Kaggle. — Updated 2018. — Available at: https://www. kaggle.com/kemical/kickstarter-projects/version/3#ks-projects-201801.csv
11. *Kuznietsova N.V.* Information Technologies for Clients' Database Analysis and Behaviour Forecasting / N.V. Kuznietsova // Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017). — 2017. — P. 56–62. — Available at: http://ceur-ws.org/Vol-2067.
12. *Allison P.D.* Survival Analysis Using SAS / P.D. Allison // Cary. — 2010. — 324 p.
13. *Cox D.R.* Regression Models and Life-Tables / D.R. Cox // Journal of the Royal Statistical Society, Series B. — 1972. — Vol. 34, N 2. — P. 187–220.
14. *Kickstarter* // PBC. — Updated 2019. — Available at: https://www.kickstarter.com/.
15. *Sikorsky* Challenge. — Updated 2019. — Available at: https://www. sikorskychallenge.com/.

Надійшла 08.07.2019