



Дипломна робота

на тему:

«Прогноз відтоку клієнтів підприємства за допомогою алгоритмів
машинного навчання»

Виконав:

Студент групи КА-43

Смирнов Олександр Олександрович

Керівник:

Недашківська Надія Іванівна

Машинне навчання:

Навчання без учителя

Мета: опис множини об'єктів, виявлення внутрішніх взаємозв'язків, залежностей, закономірностей, що існують між об'єктами.

Задачі, що вирішуються:

- Кластеризація
- Зменшення розмірності даних

Приклади алгоритмів:

- K-means
- PCA

Навчання з учителем

Мета: підбір оптимальних параметрів функції прогнозування для передбачення значенні цільової функції, базуючись на вхідних даних.

Задачі, що вирішуються:

- Задачі регресії
- Класифікація об'єктів
- Розпізнавання

Приклади алгоритмів:

- Лінійна регресія
- Random forest
- SVM

Навчання з підкріпленням

Мета: побудова алгоритму в умовах «агент-середа», за допомогою введення функції корисності і моделювання взаємодії агента з середовищем.

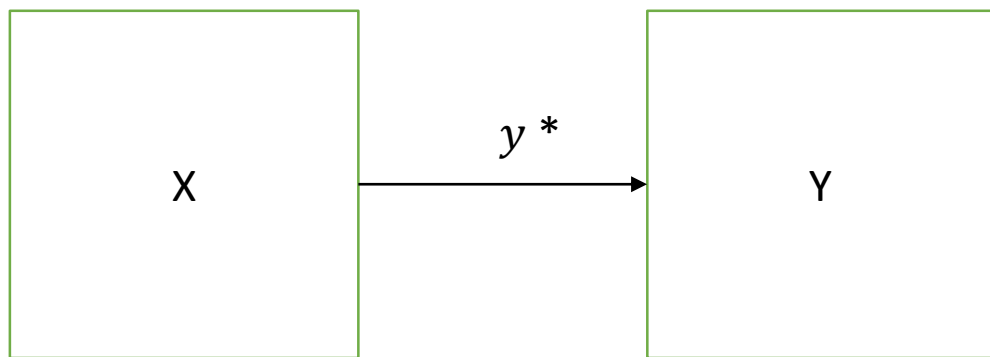
Задачі, що вирішуються:

- Автопілот машини
- Alpha GO
- Покер

Постановка задачі: навчання з учителем

Вхідні дані:

Задається множина об'єктів X та множина допустимих відповідей Y



y^* - цільова функція
Значення: $y_i = y^*(x_i)$
 $\{x_1, \dots, x_\ell\} \subset X$
 $\{y_1, \dots, y_k\} \subset Y$

Необхідно:

За вибіркою відновити залежність y^* , тобто побудувати вирішальну функцію

$$a: X \rightarrow Y$$

Вирішальна функція повинна наближати y^* не лише на об'єктах навчальної вибірки, а і на всій множині X

Тобто потрібно мінімізувати, де L – функція втрат

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^l L(a, x_i)$$

Постановка задачі: прогнозування відтоку

Задача прогнозування відтоку клієнтів:

Мета:

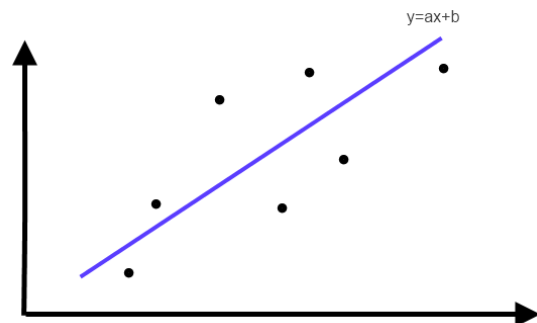
- Зменшення витрат на залучення нових клієнтів
- Зменшення затрат на використання людського ресурсу (працівники, що виявляють нелояльних клієнтів)
- Збільшення клієнтської бази

Конкретизація:

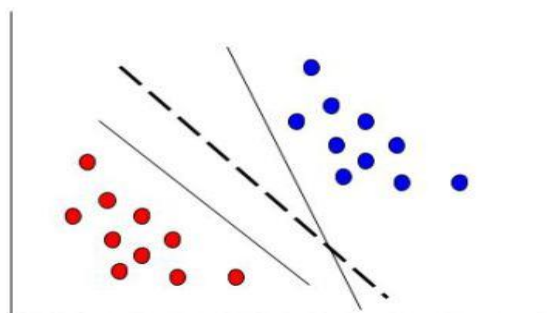
- У дипломній роботі розглянуто український маркетплейс цільовою аудиторією якого є B2B клієнти.
- За бізнес процесом компанії, вважається, що клієнт відтік, якщо не було зроблено повторної оплати протягом 1 календарного місяця
- В середньому час для утримання клієнта – 2 місяці, тому було обрано саме такий горизонт прогнозування

Існуючі методи рішення задачі навчання з учителем

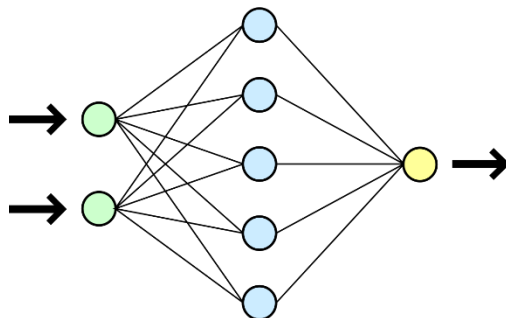
- Регресійні: метод відновлення залежності між двома змінними



- SVM: побудова роздільної гіперплощини



- Нейронні мережі



- Ансамблі:
 - Беггінг (Bootstrap aggregating)
 - Бустінг (Boosting)

Беггінг – random forest – голосування дерев рішень

Бустінг – лінійна комбінація дерев рішень, кожне з яких компенсує помилки композиції передуючих

Методи оцінки якості

У задачах машинного навчання для оцінки якості моделей і порівняння різних алгоритмів використовуються метрики. Їх вибір і аналіз залежить від типу та мети задачі.

Приклади:

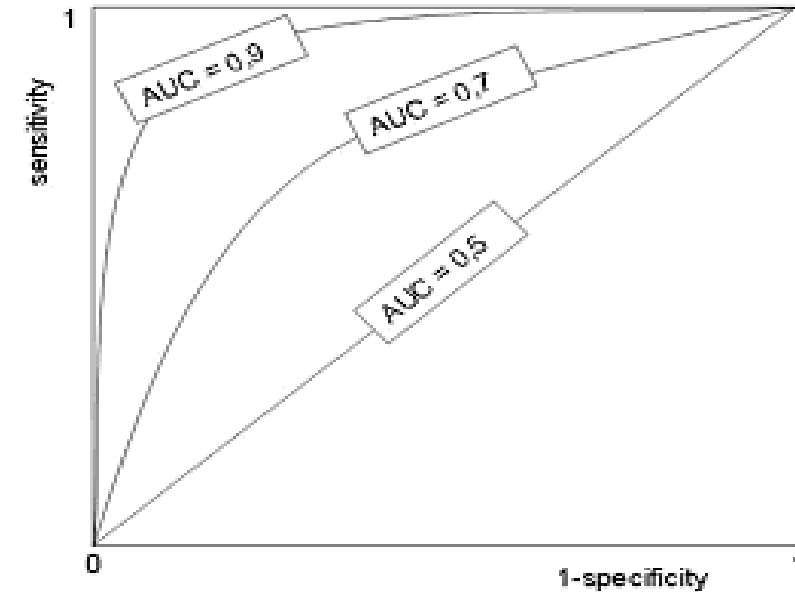
Для задач регресії:

- $MSE = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2$
- $MAE = \frac{1}{l} \sum_{i=1}^l |a(x_i) - y_i|$

Для задач класифікації:

- $accuracy = \frac{\sum_{i=1}^l [a(x_i) = y_i]}{l}$
- $precision = \frac{TP}{TP + FP}$
- $recall = \frac{TP}{TP + FN}$
- $F = \frac{2 * precision * recall}{precision + recall}$

- ROC-крива - графік, що дозволяє оцінити якість бінарної класифікації



Опис алгоритму розв'язку задачі

Алгоритм класифікації у вигляді композиції дерев рішень

$$F_M(x) = \sum_{m=1}^M b_m h(x; a_m), b_m \in R, a_m \in A$$

Задача вирішується жадібним пошуком

$$F_K(x) = F_{K-1}(x) + b_k h(x; a_k), b_k \in R, a_k \in A$$

Вводиться функція втрат

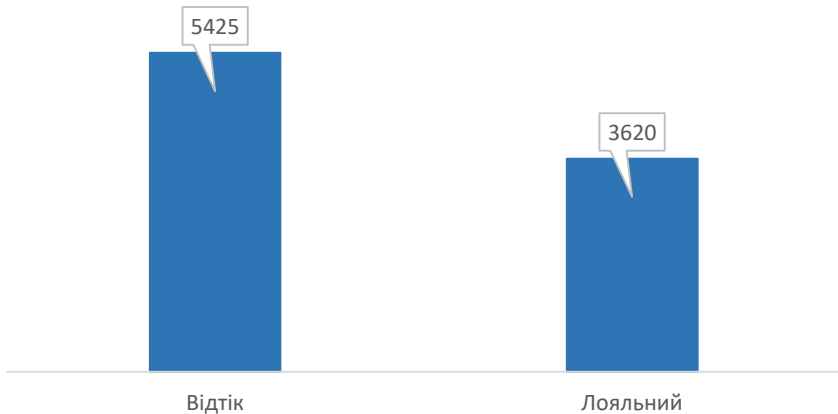
$$L(y_i, F_m(x_i)), i = 1..N$$

Мінімізується функціонал помилки

$$Q = \sum_{i=1}^N L(y_i, F_m(x_i)) \rightarrow \min$$

Попередня обробка даних

- Кількість відвідувань особового кабінету на ресурсі
- Частоту переглядів товарів клієнта на сайті
- Кількість заявок на перегляд контактів клієнта
- Історію клієнта по покупкам послуг
- Кількість зв'язків клієнта з менеджером
- Вид господарського товариства компанії
- Кількість працівників
- Посаду контакту, з якою ведуться переговори



			next_account_start	next_account_end	payed_cost	clicks	users	cnt_msg	cnt_phones
tld	account_start	account_end							
		2015-07-27			1500.00	622.400000	351.600000	1.800000	95.000000
					2750.00	449.600000	277.000000	1.800000	27.200000
				2017-07-31	2750.00	509.200000	305.800000	3.200000	24.000000
					3850.00	284.727273	158.272727	2.272727	26.090909
1	2015-03-02				4919.11	341.000000	191.636364	2.909091	61.818182

Побудови моделі

Результати застосування відомих алгоритмів класифікації

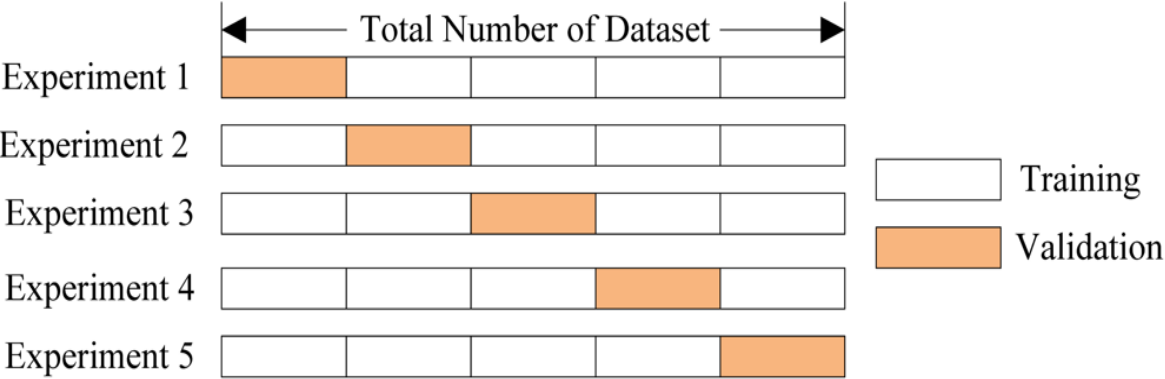
	Accuracy	Precision	Recall
Logistic regression	67%	68%	83%
Random forest	64%	71%	67%
Gradient boosting	71%	72%	82%

Параметри які піддавалися навчанню:

1. Кількість дерев
2. Глибина дерева
3. Мінімальна кількість елементів для розбиття
4. Мінімальна кількість елементів у листі

Перевірка стійкості моделі кросс-валідацією

Кросс-валідація

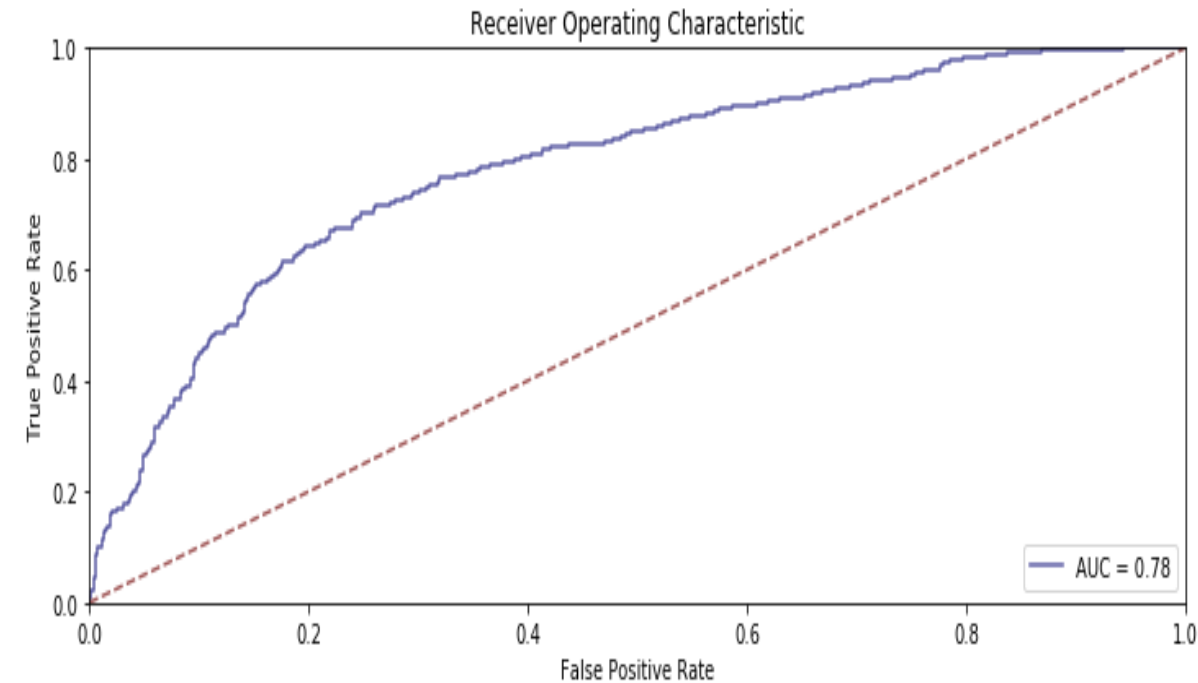


	Доля правильних відповідей	Точність	Повнота
1	0.7	0.73	0.82
2	0.71	0.72	0.81
3	0.71	0.73	0.83
4	0.71	0.71	0.84
5	0.7	0.73	0.83

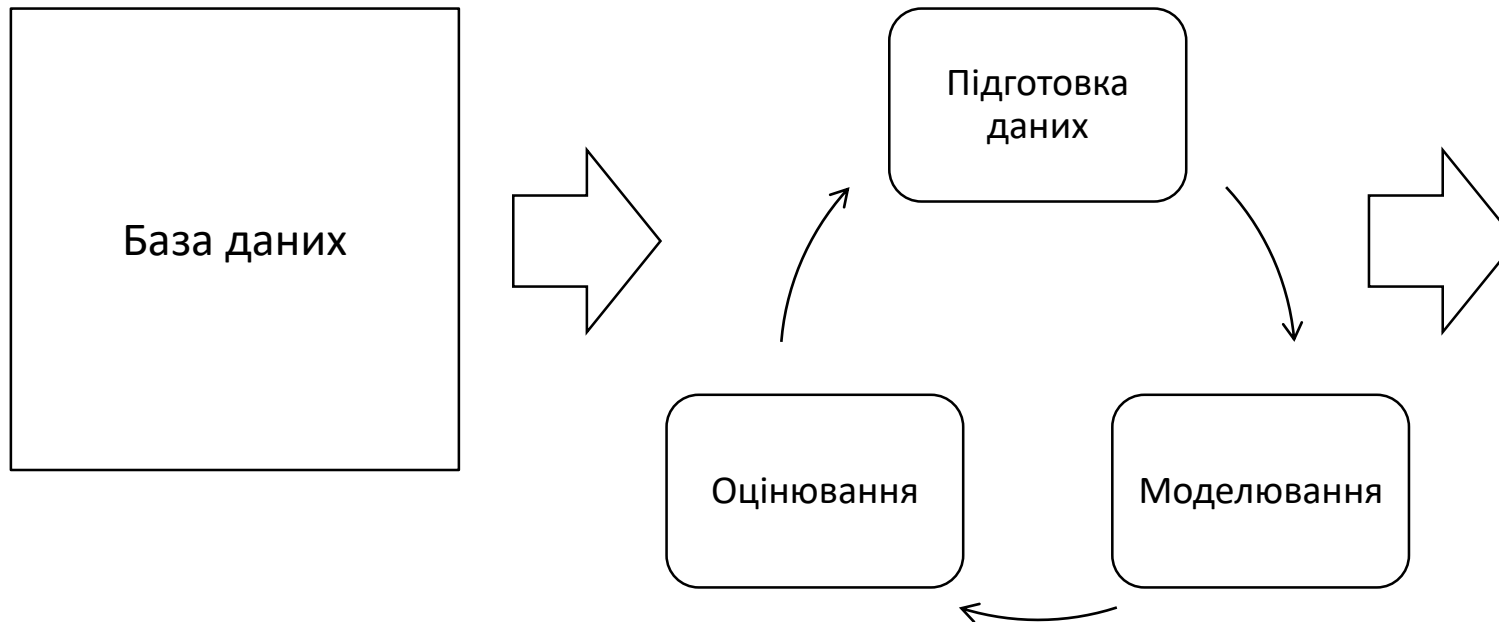
Аналіз отриманих результатів

	$y = 1$	$y = 0$
$a(x) = 1$	667	230
$a(x) = 0$	146	314

Метрика	Значення
Доля правильних відповідей (accuracy)	0.73
Точність (precision)	0.74
Повнота (recall)	0.82
ROC-AUC	0.78

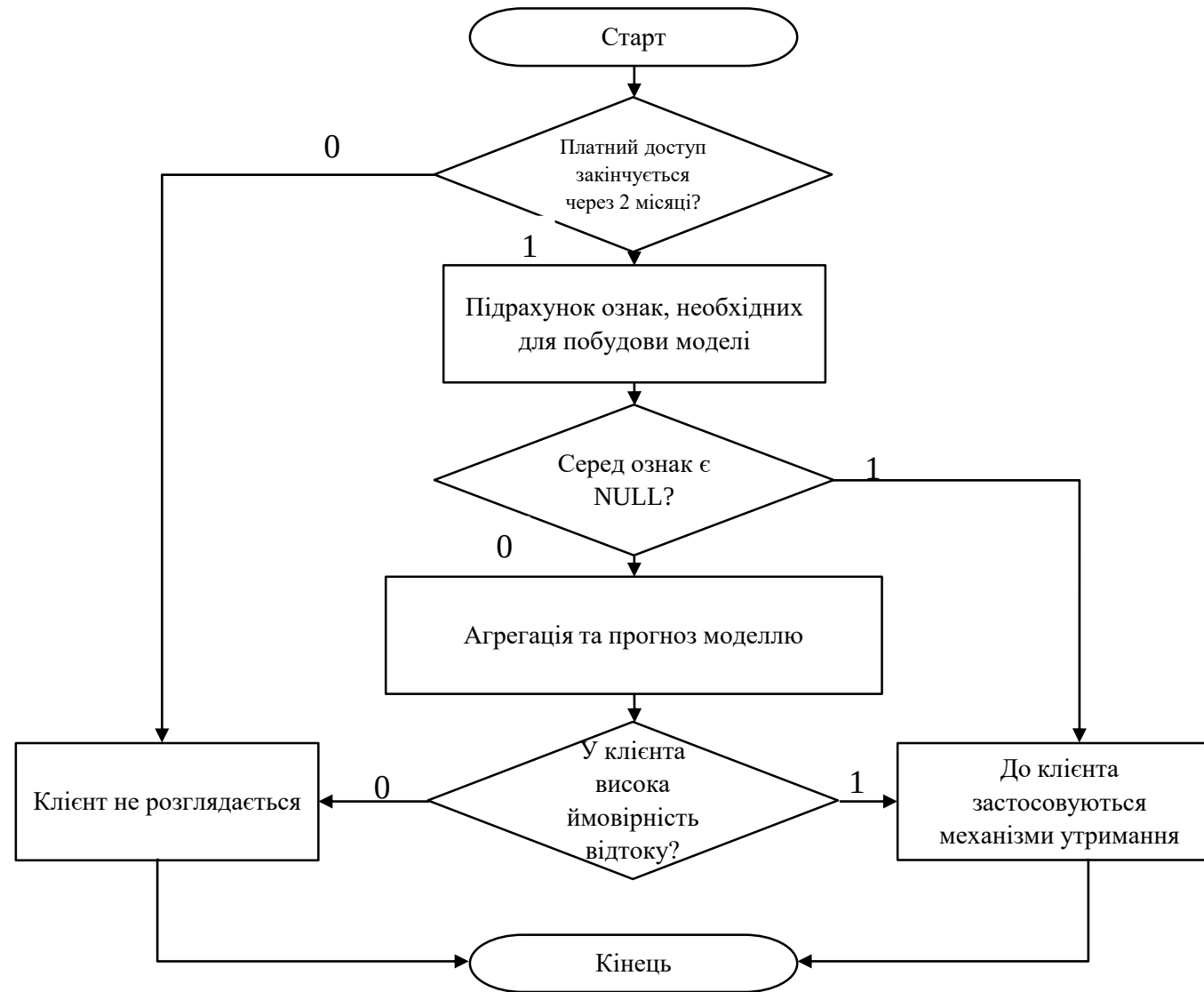


Результати роботи програми



	A	B	C
1	id	account_start	Probability
2	1	2015-09-22 00:00:00	0.610258593
3	2	2016-06-06 00:00:00	0.82536827
4	3	2015-06-11 00:00:00	0.595841314
5	4	2016-09-07 00:00:00	0.526768938
6	5	2016-05-21 00:00:00	0.265333053
7	6	2016-04-11 00:00:00	0.519719051

Схема роботи прогнозу відтоку на сервері компанії





Дякую за увагу