ДИПЛОМНА РОБОТА

“ПЕРЕДБАЧЕННЯ РЕАКЦІЇ КЛІЄНТІВ БАНКУ НА НОВУ МАРКЕТИНГОВУ ПРОПОЗИЦІЮ ЗА ДОПОМОГОЮ АМАЗОНІВСЬКОГО СЕРВІСУ МАШИННОГО НАВЧАННЯ”

Виконала: Мулява Ольга Ярославівна, ННК “ІПСА”, гр. КА-31

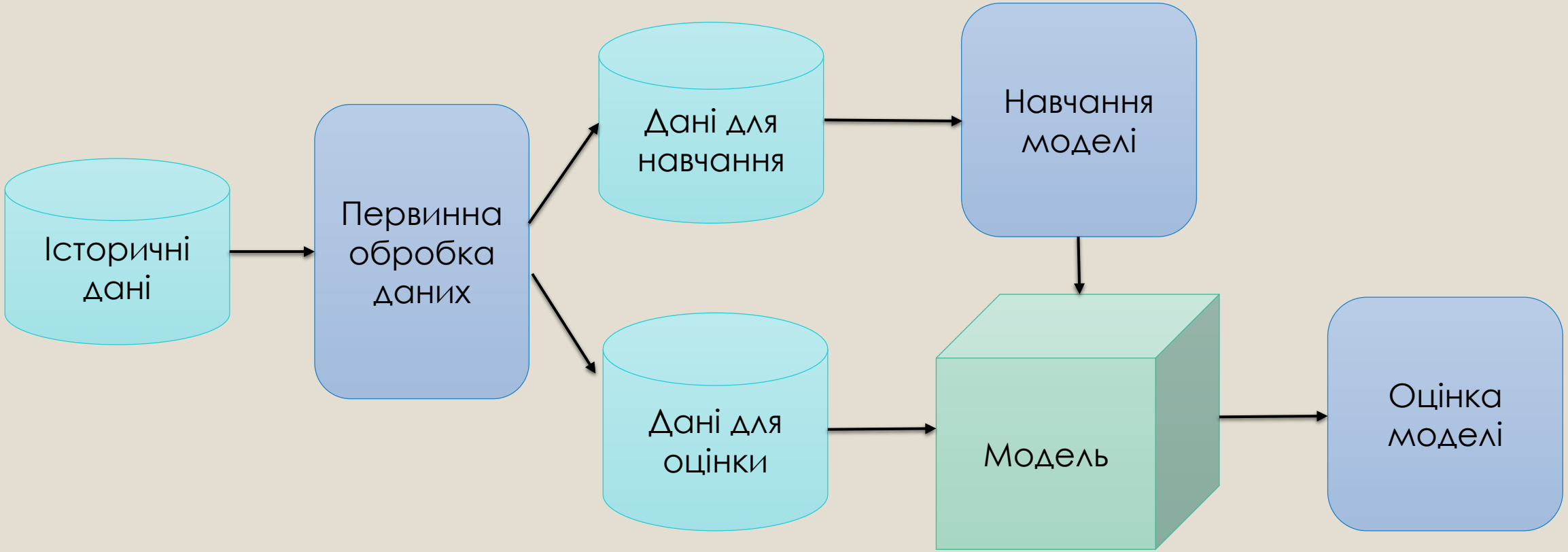
Науковий керівник: Кухарєв С. О.

Актуальність

Машинне навчання у хмарі

- МЕТА ДОСЛІДЖЕННЯ є дослідження і реалізація існуючого методу побудови математичної моделі за допомогою методів машинного навчання у хмарі, а також розробка програмного продукту, який дає змогу передбачити реакцію клієнтів банку на нову маркетингову пропозицію.
- ОБ'ЄКТ ДОСЛІДЖЕННЯ є історичні дані щодо характеристик клієнтів банку, які після машинного навчання дадуть змогу робити прогнози для всіх клієнтів.
- ПРЕДМЕТ ДОСЛІДЖЕННЯ є математична модель логістичної регресії, скоринг та інші математичні методи, застосовані для реалізації практичної проблеми.

Алгоритм реалізації машинного навчання



Вхідні дані, їх первинна обробка

- Age - вік
- Job - професія
- Marital – сімейний стан
- Education - освіта
- Default – наявність кредиту в дефолті
- Housing – наявність кредиту на житло
- Loan - наявність особистого кредиту
- Y – бінарне {0,1} чи згодився клієнт на депозит

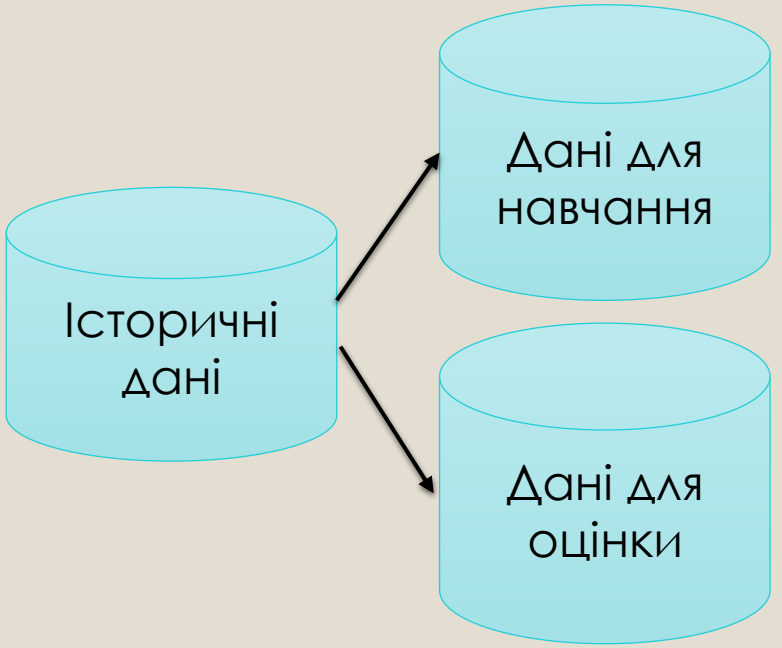
41 189 клієнтів

Первинна обробка даних:

- Відсутні значення
- Невірні значення
- Кореляція вхідних змінних з вихідною змінною

Поділ вхідних даних

Три варіанти поділу



- **Попередній розділ даних**

Користувач розділює дані на дві групи: для навчання (історичні), для оцінки

- **Послідовний поділ даних**

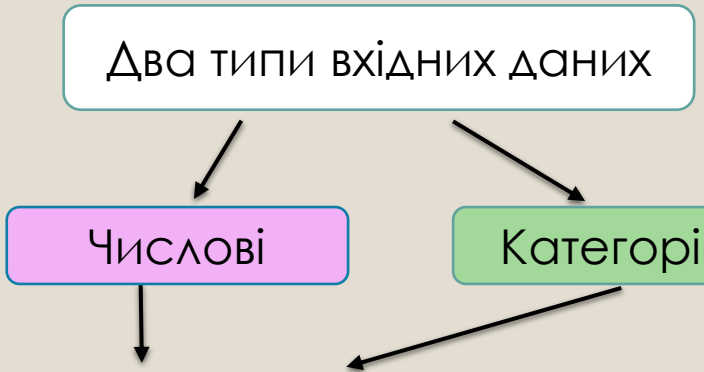
Перші 70 % даних - для навчання,

а решта 30 % даних для оцінки

- **Випадковий поділ даних** (рандомний)

Коли не потрібно, щоб зберігся порядок вхідних даних.

Підготування даних

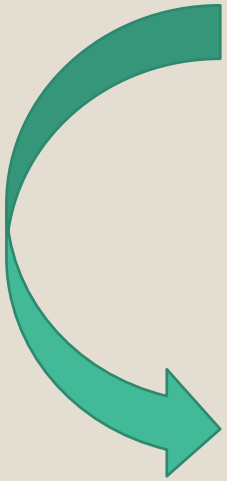


Категорізація числових даних

age	job	marital	education	default	housing	loan	y
44	blue-collar	married	basic.4y	unknown	yes	no	0
53	technician	married	unknown	no	no	unknown	0
28	management	single	university.degree	no	unknown	no	1
39	services	married	high.school	no	no	no	0
55	retired	married	basic.4y	no	yes	no	1
30	management	divorced	basic.4y	no	yes	no	0
37	blue-collar	married	basic.4y	no	yes	no	0
39	blue-collar	divorced	basic.9y	no	yes	no	0
36	admin.	married	university.degree	no	no	no	1

Категорізація числових даних

Об'єднання



Категорії	Згодились "1"	Не згодились "0"	d1	d0	WOE	
до 25	3530	3001	0,12	0,28	0,88	0,14
25 - 30	1118	1471	0,04	0,14	1,31	0,13
30 - 35	3413	1824	0,11	0,17	0,41	0,02
35 - 40	3589	1236	0,12	0,11	-0,03	0,00
40 - 45	1765	471	0,06	0,04	-0,28	0,00
45 - 50	3648	824	0,12	0,08	-0,45	0,02
50 - 55	2824	471	0,09	0,04	-0,75	0,04
55 - 60	3354	824	0,11	0,08	-0,37	0,01
60 - 65	3883	412	0,13	0,04	-1,21	0,11
від 70	3295	235	0,11	0,02	-1,60	0,14
Всього	30420	10768	*IV* - information value ->			0,59

Категорії	Згодились "1"	Не згодились "0"	d1	d0	WOE	
до 30	4648	4472	0,15	0,42	1,00	0,26
30 - 35	3413	1824	0,11	0,17	0,41	0,02
35 - 40	1471	647	0,05	0,06	0,22	0,00
40 - 60	13710	3177	0,45	0,30	-0,42	0,07
від 60	7178	647	0,24	0,06	-1,37	0,24
Всього	30420	10768	*IV* - information value ->			0,62

Оцінка результату категорізації

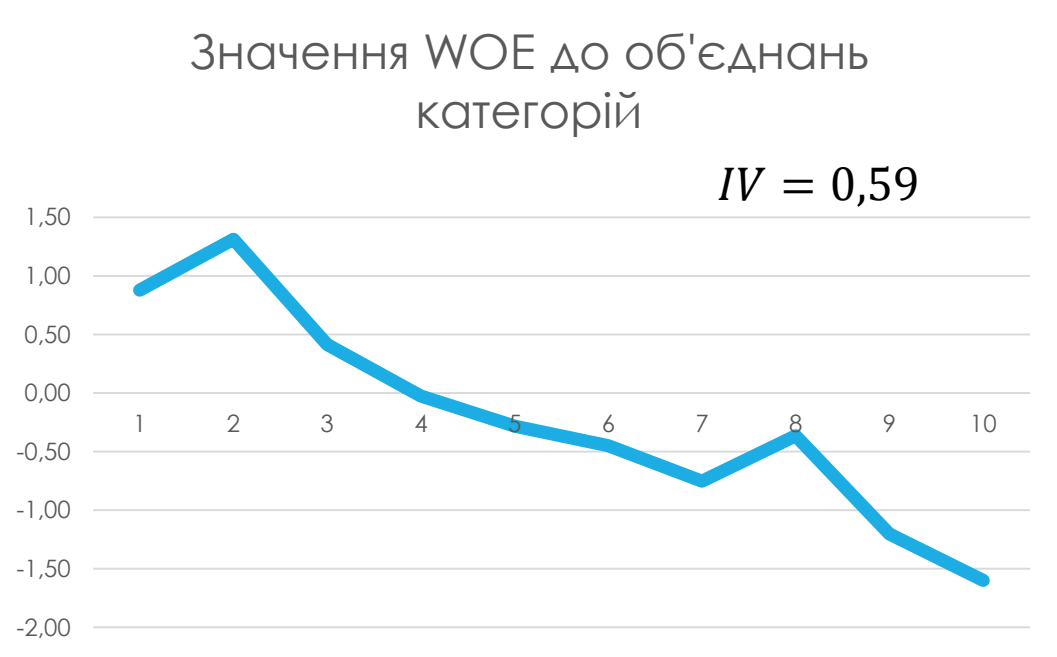
$$WOE_i = ln\frac{d_i^{(0)}}{d_i^{(1)}}$$

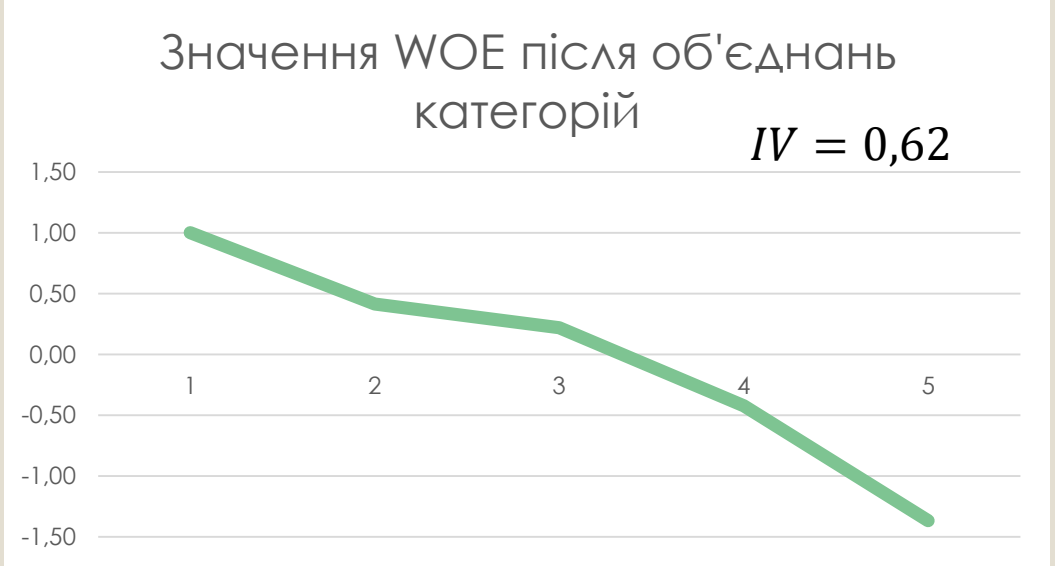
$d_i^{(0)}$ - відносна частота “0” в і-тій категорії

$d_i^{(1)}$ - відносна частота “1” в і-тій категорії

$$IV = \sum_{i=1}^{k}(d_i^{(1)} - d_i^{(2)})\, WOE_i$$

Information Value – значення інформації





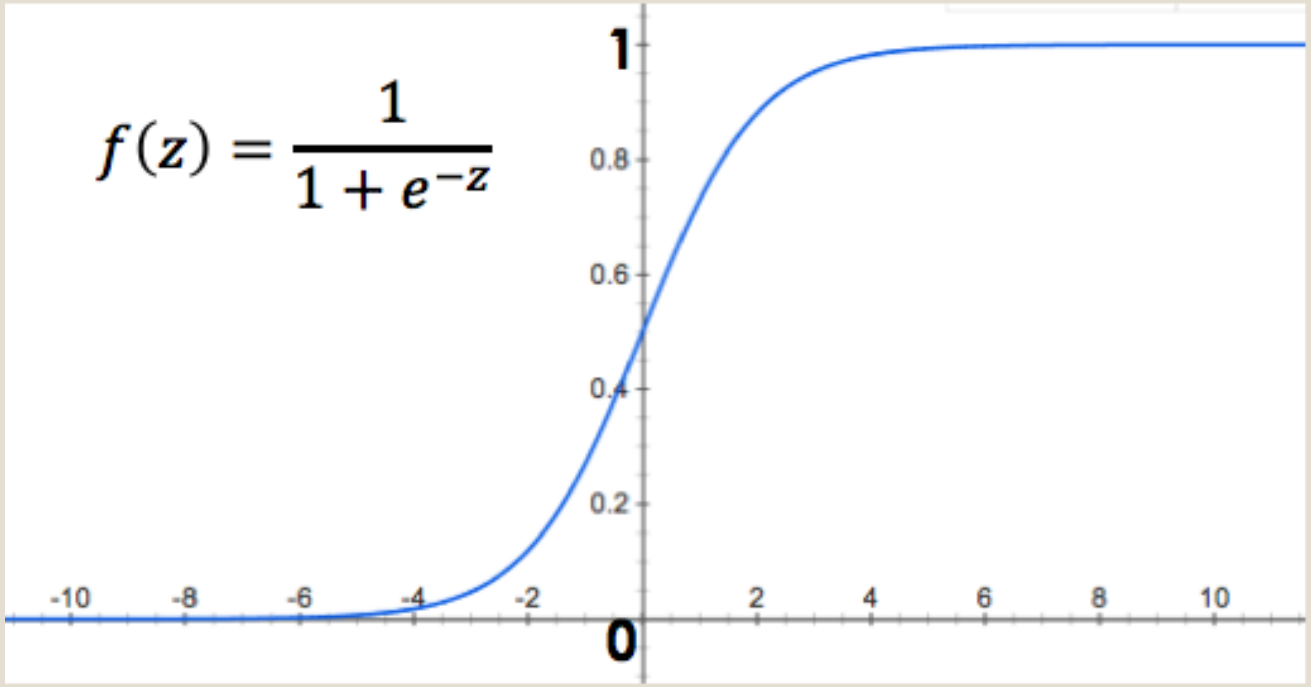
Постановка задачі логістичної регресії

Логістична регресія – окремий випадок лінійної регресії

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$z = ln\frac{f(z)}{1 - f(z)} = ln\frac{p}{1 - p} = logit(p)$$



Знаходження параметрів регресії θ_1

Метод максимальної правдоподібності

- Нехай X — випадкова змінна з дискретним розподілом імовірності p, який залежить від параметра θ.
- Функція правдоподібності

$$\mathcal{L}(\theta, x) = P(X = x; \theta) \rightarrow max, \text{при якому } \theta^*$$

- Максимізуємо не саму функцію $\mathcal{L}(\theta, x)$, а її натуральний логарифм

$$ln\ \mathcal{L}(\theta, x) \rightarrow max, \text{при якому } \theta^*$$

Результати і оцінка моделі

Фактичний клас	Передбачений клас	
	True(1)	False (0)
True (1)	TP – істинно позитивний 581	FN – хибно негативний 782
False (0)	FP – хибно позитивний 316	TN - істинно негативний 10 676

Оцінка моделі

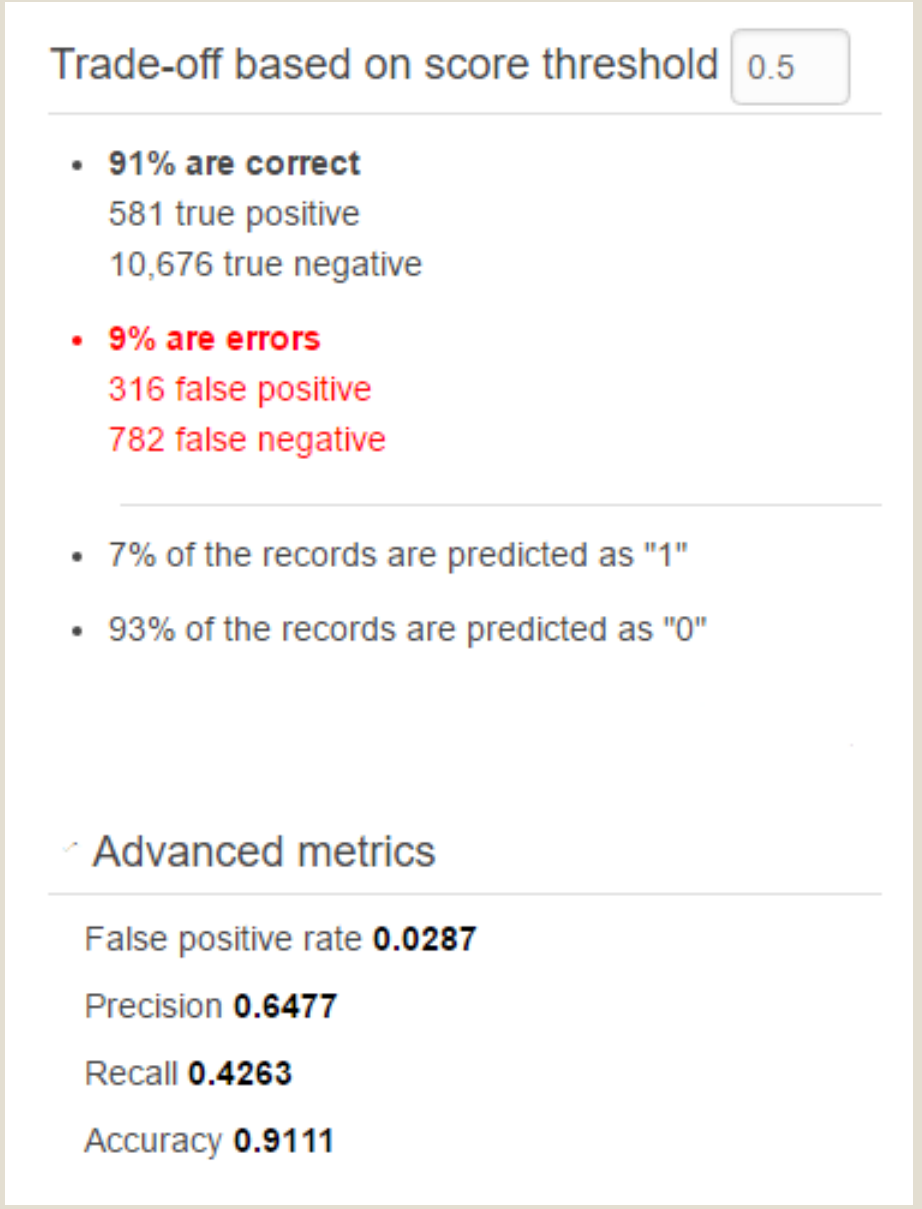
False positive rate – хибно позитивний показник

$$FPR = \frac{FP}{FP + TN}$$

Акуратність - $$Precision = \frac{TP}{TP + FP}$$

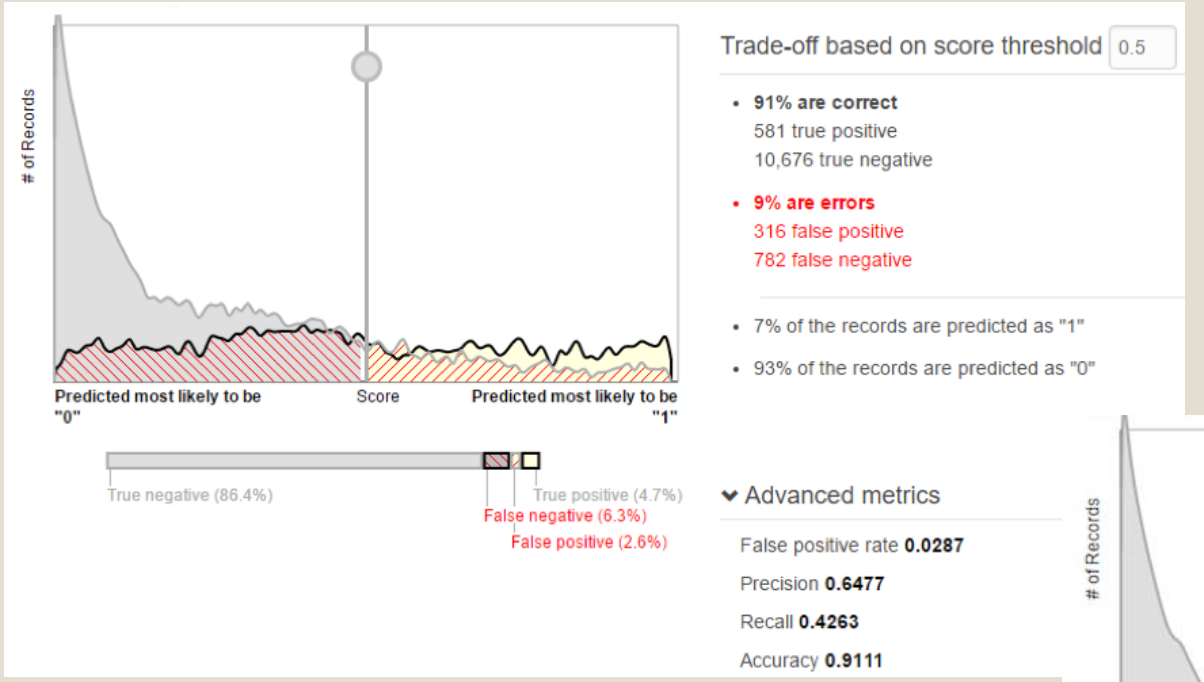
Відгук - $$Recall = \frac{TP}{TP + FN}$$

$$Точність \quad - \quad Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$



ROC-аналіз

(receiver operating characteristic – робоча характеристика приймача)



Trade-off based on score threshold 0.5

- **91% are correct**
  581 true positive
  10,676 true negative
- **9% are errors**
  316 false positive
  782 false negative

- 7% of the records are predicted as "1"
- 93% of the records are predicted as "0"

Advanced metrics

False positive rate **0.0287**

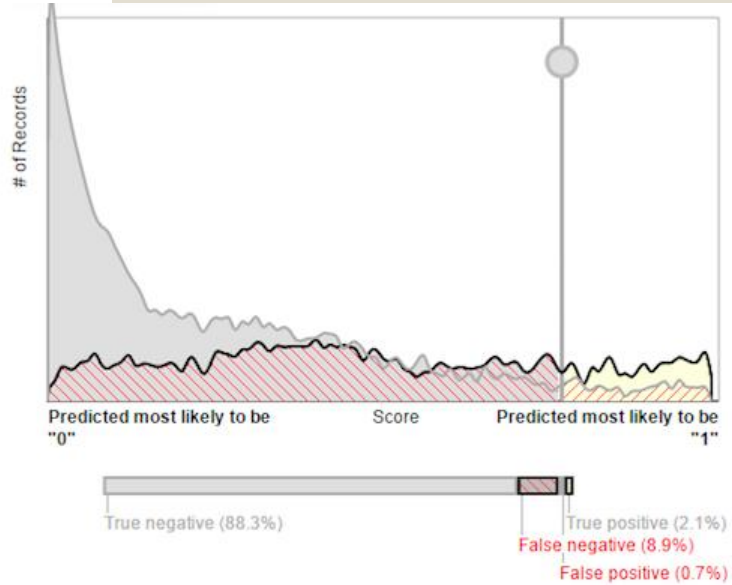
Precision **0.6477**

Recall **0.4263**

Accuracy **0.9111**

Поріг = 0,5

Поріг = 0,77



Trade-off based on score threshold 0.77

- **90% are correct**
  263 true positive
  10,906 true negative
- **10% are errors**
  86 false positive
  1,100 false negative

- 3% of the records are predicted as "1"
- 97% of the records are predicted as "0"

Advanced metrics

False positive rate **0.0078**

Precision **0.7536**

Recall **0.193**

Accuracy **0.904**

Висновки

- У даній дипломній роботі використаний підхід до методів машинного навчання, а саме за допомогою хмарних технологій.
- Була створена математична модель логістичної регресії, параметри моделі були знайдені методом максимальної правдоподібності. Результати моделі були використані для побудови скорингової карти, яка дає змогу робити прогнози для нових даних.
- Реалізовний програмний продукт за допомогою Амазонівського веб-сервісу з використанням мови програмування Python.

Дякую за увагу!