Mathematics in Modern Technical University
ISSN 2664-4258

Mathematics in Modern Technical University, 2025(2), 65-82

10.20535/mmtu-2025.2-065

Аналіз результатів опитування з використанням stepwise method в моделі логістичної регресії на платформі MS Excel (опитування проводилось серед українських студентів у лютому 2024 р.)

О.О. Врубльовська [1], О.В. Мулик[2]

[1,2]Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, [2]ORCID ID: https://orcid.org/0000-0003-0710-9693
([1,2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" , Kyiv, Ukraine)

Анотація

Дана стаття має на меті продемонструвати дослідження на платформі MS Excel результатів опитування методом покрокового виключення змінних в моделі логістичної регресії для мінімізації кількості факторних ознак, пов'язаних з вихідною змінною. Наведено базові формули логістичної регресії та як саме вони реалізуються на платформі MS Excel (Real Statistics, Logistic and Probit Regression). Розглянуто широко вживані статистичні міри продуктивності тестів бінарної класифікації: чутливість (Sensitivity), специфічність (Specificity), прогностичність позитивного результату (PPV) моделі, прогностичність негативного результату (NPV) моделі, поріг поширеності (Prevalence) та інші. Подано детальний аналіз кривої операційних характеристик (Receiver Operating Characteristic – ROC curve analysis) та застосування J-статистики Юдена.

Ключові слова: логістична регресія, багатовимірна модель, покроковий метод, MS Excel, ROC-крива, AIC, BIC, AUC, p-значення.

УДК 004.85

mulyk.olena@gmail.com

Cite as: O.O. Vrublovska & O.V.Mulyk (2025). Analysis of survey results using the stepwise method in the model of logistic regression on the platform of MS Excel (the survey of Ukrainian students in February 2024). *Mathematics in Modern Technical University*, 2025(2), 65-82. 10.20535/mmtu-2025.2-065

1 Вступ

В умовах глобальних змін, зокрема воєнних дій та економічної нестабільності, студенти стикаються з численними стресовими факторами, які можуть негативно впливати на їхні академічні досягнення та психологічний стан. Особливо це стосується українських студентів, які в період 2022-2023 років зазнали впливу негативних подій, що серйозно змінили їхнє звичне життя. Психологічна тривожність, соціальна ізоляція, фінансові труднощі та загальна нестабільність створюють додатковий тиск, який ускладнює процес навчання та досягнення успіху.

Проблема полягає в тому, що до сьогодні бракує чіткого розуміння того, як саме різні фактори впливають на особистісні почуття студентів в умовах широкомасштабної агресії. Логістична регресія дозволяє вивчити зв'язки між стресовими факторами, академічною успішністю та визначити рекомендації, що саме може стати вирішальними факторами для підтримки студентів у їх рішенні залишатись навчатись і працювати в Україні.

2 Методи дослідження

Для вирішення поставленої проблеми було обрано кількісне дослідження з використанням логістичної регресії. Основна мета полягала в аналізі впливу поточного військового стану в Україні на можливість прийняття рішення студентами II-III курсів залишитись в Україні для подальшого навчання та роботи за фахом і знаходження ключових факторів, що можуть визначати цей результат.

Дослідження базувалося на результатах анонімного опитування, яке було проведене серед студентів 2-3 курсів у лютому 2024 року. Опитування включало 14 питань, серед яких було відібрано 11 ключових для подальшого аналізу. Кожне питання оцінювалося за п'ятибальною шкалою (від 1 до 5), що дозволило дослідити рівень тривожності, фізичного, морального стану, мотивації студентів до навчання та працевлаштування в Україні.

3 Основні етапи методології

3.1 Збір даних.

Студентам було запропоновано пройти анкетування, що включало питання про їхні відчуття та академічні результати в умовах кризових подій. Загалом було зібрано 40 відповідей, які були розподілені за кількома ключовими ознаками.

3.2 Підготовка даних

Після збору даних вони були очищені від неповних відповідей. Було виконано стандартизацію змінних для зменшення їхнього впливу на модель. Була підготовлена бінарна змінна $\boldsymbol{Y}$ для визначення прийняття рішення: $\boldsymbol{y_i = 1}$ якщо студент відповідав на запитання на 4 чи 5 з 5 балів; $\boldsymbol{y_i = 0}$, якщо 3, і менше.

3.3Опис бінарних змінних та їх використання в моделі

Бінарні змінні широко використовуються в статистиці для моделювання ймовірності певного стану чи події ($P(E)$), наприклад ймовірності перемоги/поразки, здоров'я/хвороби, тощо. Тобто, якщо бінарна змінна $\boldsymbol{Y}$ є результатом певної події E: $\boldsymbol{y_i = 1}$ подія відбулась, а $\boldsymbol{y_i = 0}$, метою стає передбачення ймовірності настання події E: $P(Y_i = 1) = p_i$ чи не настання події E: $P(y_i = 0) = 1 - p_i$ на основі результатів попередніх відомостей. Тоді логістична модель (логіт-модель) є найбільш придатною і будується мультиплікативна модель залежності шансів результуючої бінарної змінної $\boldsymbol{Y}$ від факторних змінних $X_1, X_2, \dots X_n$ у вигляді лінійної комбінації однієї чи кількох незалежних змінних:

$$\ln\big(Odds(E)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon,$$

де $\beta_i, i = \overline{1,n}$ коефіцієнти моделі, функція $\ln(Odds(E)) = \ln\left(\frac{P(E)}{1-P(E)}\right)$ – є ймовірністю того, що подія E відбудеться і, також, вводиться функція logit:

$$Logit(P) = \ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \sum_{k=1}^{n} \beta_k x_k.,$$

для позначення логарифму шансів. До того ж бінарні змінні можна узагальнити до категоріальних змінних, якщо існує більше двох можливих значень, тобто бінарну логістичну регресію узагальнити до мультиноміальної логістичної регресії.

3.4 Побудова логістичної моделі та розрахунки

При переході до аналізу логіт-моделі, заснованої на спостережуваній вибірці — значення $\boldsymbol{Y}$ замінюються їх вибірковими оцінками P, коефіцієнти β_i замінюються на вибіркові оцінки $b_i,$ а значення помилки ε відкидається. Тоді:

$$Odds(E) = \frac{P(E)}{1-P(E)} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n+\varepsilon} \Longrightarrow$$

$$\Longrightarrow P(E) = P(y=1) = \frac{1}{1+e^{-(b_0+\sum_{k=1}^{n} b_k x_k)}},$$

де $b_i, i = \overline{1,n}$ коефіцієнти моделі, які обчислюються на платформі MS Excel (Real Statistics, Logistic and Probit Regression, https://real-statistics.com/).

Важливо відмітити уведення функції «відношення шансів» $R_{x_i x_j}$ (ВШ, **odds ratio**), яка показує в якому відношенні ймовірність одного випадку буде більше/менше, ніж другого при всіх інших однакових умовах:

$$R_{x_i x_j} = \frac{Odds(x_{i1},\cdots,x_{in})}{Odds(x_{j1},\cdots,x_{jn})} = \frac{e^{\beta_0+\sum_{k=1}^{n}\beta_k x_{ik}}}{e^{\beta_0+\sum_{k=1}^{n}\beta_k x_{jk}}} = e^{\sum_{k=1}^{n}\beta_k(x_{ik}-x_{jk})}.$$

Якщо записати через функцію logit для випадку $n = 1$ при $P(y=1) = \frac{1}{1+e^{\beta_0+\beta_1 x_1}}$:

$$Logit\left(\frac{p_{x+1}}{p_x}\right) = \ln\left(\frac{p_{x+1}}{1-p_{x+1}}\Big/\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1(x+1) - \beta_0 + \beta_1 x = \beta_1,$$

то:

$$\frac{Odds(x+1)}{Odds(x)} = \frac{p_{x+1}}{1-p_{x+1}}\Big/\frac{p_x}{1-p_x} = e^{\beta_1},$$

тобто, якщо є $x = 1$ та $x = 0$, то величина e^{β_1} представляє відношення шансів між цими значеннями. Якщо маємо подію E_i, то її ймовірність буде тим вище, чим більше значення $Odds(E_i)$:

$$Odds(E_i) = \frac{event}{not\ event}.$$

Для логістичної моделі не можна використовувати метод найменших квадратів для розрахунку значень коефіцієнтів $b_i, i = \overline{1,n}$. Натомість обчислення проводиться методом максимальної правдоподібності. Модель, що використовується, заснована на біноміальному розподілі, тобто функція щільності ймовірності $f(x;\theta) = C_n^k p^k (1-p)^{n-k}$ події для вибіркових даних визначається формулою:

$$L = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \Longrightarrow LL = \ln L = \sum_{i=1}^{n}(y_i \ln p_i + (1-y_i)\ln(1-p_i))$$

де y_i — спостережувані значення, а p_i — відповідні теоретичні значення. Метою є максимізувати функцію LL, припускаючи, що $p_i = \frac{1}{1+e^{-(b_0+\sum_{k=1}^{n} b_k x_k)}}$. Це дозволить знайти значення коефіцієнтів b_i. Процес можна зробити вручну, застосовуючи Solver (Data Analysis). Також обчислюється значення LL_0, яке вважається початковим для L, оскільки включає лише значення коефіцієнта перетину (intercept b_0) і обчислюється за формулою:

$$LL_0 = \ln L_0 = n_0 \ln \frac{n_0}{n} + n_1 \ln \frac{n_1}{n},$$

де n_0 – кількість спостережень із значенням 0; n_1- кількість спостережень зі значенням 1 і $n = n_0 + n_1$.

Пакет ресурсів Real Statistics надає інструмент аналізу даних логістичної та пробіт регресії (Logistic and Probit Regression). Дана платформа приймає як вхідні дані діапазон, що містить вибіркові незалежні дані у вигляді стовпців $X_1, X_2, \ldots X_n$ та стовпець з бінарною результуючою змінною $\boldsymbol{Y}$ успішних і невдалих випадків (див. Рис.1 і Рис.2)

4 Використання платформи MS Excel (Real Statistics) для обчислень

Пропонується детально розглянути використання платформи MS Excel (Real Statistics) на прикладі аналізу результатів анонімного опитування у лютому 2024 року для студентів 2-3 курсів (N=40). Мета - виявлення найбільш впливових факторних ознак (з 14 поставлених запитань, для дослідження відібрані 11 запитань за шкалою від 1 до 5), що можуть стати визначальними на вибір студентів продовжувати навчання і фахову роботу в Україні. В процесі аналізу якості моделі дослідимо, які саме числові характеристики наводить платформа MS Excel і які буде доцільно розрахувати додатково для максимально інформативного представлення результатів дослідження.

Вихідна статистика для першого кроку має вигляд:

egression

X2	X4	X6	X7	X8	X9	X10	X11	X12	Success	Failure	Total	p-Obs	p-Pred	Suc-Pred	Fail-Pred	LL	% Correct
2	2	4	5	5	4	3	3	4	1	0	1	1	0,32491	0,32491	0,67509	-1,1242	0
3	2	5	5	5	5	4	5	3	1	0	1	1	0,86858	0,86858	0,13142	-0,14089	100
1	1	2	5	4	3	4	3	2	1	0	1	1	0,83819	0,83819	0,16181	-0,17651	100
5	1	5	5	5	5	5	1	2	0	1	1	0	0,00779	0,00779	0,99221	-0,00782	100
3	5	5	5	3	1	4	3	4	0	1	1	0	0,01513	0,01513	0,98487	-0,01525	100
1	2	3	4	4	3	2	4	4	1	0	1	1	0,8445	0,8445	0,1555	-0,16901	100
2	1	4	5	4	4	5	5	4	1	0	1	1	0,97761	0,97761	0,02239	-0,02264	100
2	4	3	4	2	2	3	3	4	1	0	1	1	0,06984	0,06984	0,93016	-2,66151	0
2	4	4	5	2	4	1	2	3	0	1	1	0	0,26232	0,26232	0,73768	-0,30425	100
4	4	2	5	2	1	1	3	3	1	0	1	1	0,58531	0,58531	0,41469	-0,53561	100
4	3	4	4	2	2	3	4	3	0	1	1	0	0,14185	0,14185	0,85815	-0,15298	100
1	1	5	5	5	5	5	1	2	0	1	1	0	0,13234	0,13234	0,86766	-0,14196	100
2	3	5	5	5	2	4	3	3	0	1	1	0	0,06877	0,06877	0,93123	-0,07125	100
3	3	4	5	5	5	5	1	2	0	1	1	0	0,07965	0,07965	0,92035	-0,083	100
4	3	4	4	4	4	3	3	2	0	1	1	0	0,27719	0,27719	0,72281	-0,32461	100
3	1	4	5	4	2	1	2	4	0	1	1	0	0,00374	0,00374	0,99626	-0,00374	100
2	1	4	5	3	4	5	3	3	0	1	1	0	0,72153	0,72153	0,27847	-1,27845	0

Coeff
-0,92712
-1,65638
-0,74176
0,20768
-0,50152
0,43848
0,16013
0,92791
0,27973
1,32999
0,00454

LL statistics	
LL0	-25,8979
LL1	-16,8962
Chi-Sq	18,0032
df	10
p-value	0,05491
alpha	0,05
sig	no
R-Sq (L)	0,34758
R-Sq (CS	0,36242
R-Sq (N)	0,49916
AIC	55,7925
BIC	74,3702

Рис.1 Частина вихідної статистики: вибіркові незалежні дані у вигляді стовпців $X_1, X_2, \ldots X_{12}$; бінарна результуюча змінна $\boldsymbol{Y}$; передбачена ймовірність p_i; «LL statistics»; коефіцієнти моделі та характеристики даної моделі.

В результаті проведених розрахунків отримуємо точкову оцінку β_i-коефіцієнтів моделі, в таблиці **Logistic Regression** стовпець під назвою **coeff** b_i, і стандартну похибку (Se) оцінки - стовпець **s.e.** Наступний стовпець **Wald** - для всіх коефіцієнтів логістичної регресії b_i, із стандартною помилкою, обчисленою в стовпці **s.e.**, статистика Вальда визначається формулою:

$$Wald = \left(\frac{b}{Se}\right)^2,$$

Загалом - статистика Вальда є приблизно нормально розподіленою, використовується для перевірки відмінності від 0 для кожного i-го коефіцієнту моделі і відповідає на питання, чи впливає дана i-та факторна ознака (з урахуванням стандартизації за всіма іншими факторами) на вихідну змінну. Чим більше величина $\boldsymbol{Wald}$ тим значніше впливає ця ознака на результат, причому статистика Вальда приблизно повторює розподіл $\chi^2(df) \sim (Wald)^2$ і значення наступного стовпця **p-value** розраховується як: $= CHIDIST(Wald; df)$, де $df = 1$ (CHISQ.DIST.RT).

Саме на стовпець **p-value** потрібно орієнтуватись для визначення мінімального набору факторних ознак, пов'язаних з вихідною змінною використовується метод покрокового відкидання/включення змінних (Stepwise), тобто виключаємо факторну ознаку з найвищим значенням p-value.

Останні три стовці під заголовком **exp(b)**, **lower**, **upper** є показником відношенням шансів (ВШ, **odds ratio**). Перший стовпець і є $= EXP(b)$, останні два - представляють довірчий інтервал $= EXP\left(b \pm Se * NORMSINV\left(1 - \frac{\alpha}{2}\right)\right.$ для $EXP(b)$. Значення ВШ більш точно оцінюють ступень впливу кожної факторної змінної в логістичній моделі регресії (з урахуванням стандартизації за всіма іншими факторами) на вихідну змінну. Якщо значення показника i-ої факторної ознаки ВШ>1 при $b_i > 0, (p - value < 0{,}05)$ можна сподіватись на зростання шансів випадку при зростанні цієї ознаки; якщо ВШ<1 при $b_i < 0, (p - value < 0{,}05)$, це свідчить про зниження шансів випадку при зростанні цієї ознаки. Якщо ж значення показника ВШ i-ої факторної ознаки статистично значимо не відрізняється від 1, то зміна цієї ознаки не пов'язана зі зміною ризику випадку.

Повернемось до Рис.1 і розглянемо стовпець з назвою «LL statistics». З точки зору удосконалення перевірки значущості, потрібно відмітити, що стандартна помилка та пов'язана з нею статистика Вальда є завищеними і це збільшує ймовірність того, що коефіцієнти b_i буде вважатися таким, що не робить істотного внеску в модель, навіть якщо цей внесок існує (тобто помилка II роду). Щоб подолати

цю проблему, для оцінки якості побудованої моделі може бути розраховано значення критерію χ^2- квадрат:

$$2(LL_1 - LL_0) \sim \chi^2(df) \sim -2\ln\frac{L_1}{L_0},$$

де L_0 – функція правдоподібності для прогнозування вихідної змінної $\boldsymbol{Y}$ без урахування факторних ознак (тільки з коефіцієнтом b_0 і без інших коефіцієнтів), L_1 – максимальне значення функції правдоподібності для змінної $\boldsymbol{Y}$ з урахуванням факторних ознак (повної моделі); $df = k - k_0$ – кількість факторів у повній моделі мінус виключені фактори.

Якщо включення факторних ознак не змінює значення функції правдоподібності, тоді χ^2-квадрат статистично значимо не відрізняється від 0, p>0.05. Чим більше відмінність L_0 та L (тим більше значення χ^2-квадрат моделі), тим більший вплив факторних ознак на зміну результуючої ознаки. На даному прикладі маємо:

$$LL_0 = \ln L_0 = 26\ln\frac{26}{40} + 14\ln\frac{14}{40} = -25{,}8979.$$

В таблиці значення $LL_0 = -25{,}89, LL = -16{,}89$, значення LL обчислюється Logistic and Probit Regression на платформі MS Excel. Далі для обчислення χ^2 використовуються формули:

$\chi^2(10) = 2(LL - LL_0) = 18{,}003$ і p-value= $CHIDIST(18{,}003; 10) = 0{,}0549$ (CHISQ.DIST.RT). Але даний результат в нашому випадку не є значущим, оскільки p>0.05, і потрібно використати метод покрокового включення/виключення факторних змінних (stepwise method) для зменшення їх кількості, що, можливо, покращить модель.

Далі в стовпці «LL statistics» обчислені три коефіцієнта детермінації $R - Sq(L)$ $R - Sq(CS)$, $R - Sq(N)$. Нагадаємо, що для лінійної регресії коефіцієнт детермінації R^2:

$$R^2 = 1 - \frac{RSS}{TSS},$$

де RSS - сума квадратів залишків регресі; TSS - загальна сума квадратів, тобто, R^2 має зміст, який відсоток дисперсії пояснено регресійною моделлю чи наскільки отримані спостереження підтверджують модель. Для логістичної регресії подібна статистика визначена у вигляді наступних трьох псевдо- R^2 статистик:

1) Коефіцієнта детермінації **McFadden's** R^2:

$$R_L^2 = 1 - \frac{LL_1}{LL_0} = 0{,}348,$$

де LL_1 відноситься до повної логарифмічної правдоподібної моделі, а LL_0 відноситься до моделі з меншою кількістю коефіцієнтів (особливо моделі з лише з коефіцієнтом b_0 і без інших коефіцієнтів).

2) Коефіцієнта детермінації **Cox and Snell's** R^2:

$$R_{CS}^2 = 1 - e^{-\frac{2}{n}(LL_1 - LL_0)} = 0{,}362,$$

де n – розмір вибірки. Подібно до коефіцієнту **McFadden's**, **Cox and Snell's** R^2 використовує ймовірність вибраної моделі та підгонку моделі в тих самих значеннях логарифму правдоподібності.

3) Коефіцієнта детермінації **Nagelkerke's** R^2:

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{2LL_0/n}} = 0{,}499,$$

який був розроблений з властивостями, аналогічними до статистики R^2, яка використовується у звичайній лінійній регресії.

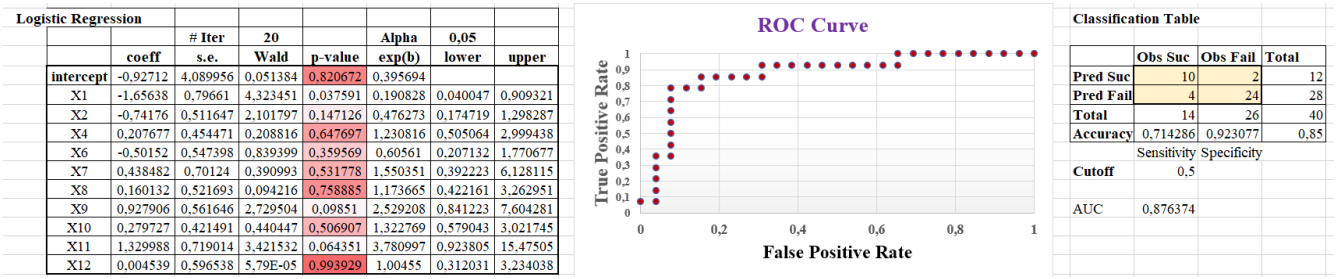
Загальне емпіричне правило для інтерпретації коефіцієнтів детермінації полягає в тому, що значення, менше, ніж 0,2 вказує на слабкий зв'язок між предикторами та результатом. Значення від 0,2 до 0,4 вказує на помірний зв'язок. Якщо значення стає більше, ніж 0,4, то можна говорити, що зв'язок є сильним. Хоча деякі дослідники вважають, що значення коефіцієнтів детермінації більше, ніж 0,16 вже є достатньо значимим – наразі все залежить від величини вибірки.

Також, важливим є граничне значення **cutoff,** що дорівнює 0,5 на Рис.2 і яке можна змінювати, як буде показано далі. На Рис. 1 прогнозовані значення (у стовпці

p-Pred), що перевищують або дорівнюють значенню **cutoff=0,5**, класифікуються як позитивні (тобто прогнозовані як успішні), значення, що є менші за 0,5, класифікуються як негативні (тобто прогнозовані як невдалі). Стовпець **Suc-Pred** – в даному випадку буде повторювати стовпець **p-Pred**, тому що метою було знаходження прогнозованих ймовірностей, відповідних до значень $\boldsymbol{Y = 1}$ та $\boldsymbol{Y = 0}$.

5 Статистичні міри продуктивності моделі. Оцінка прогностичних характеристик моделі

На Рис. 2 також представлена таблиця «Classification Table» яка є іншим способом оцінки якості моделі. Інструмент аналізу даних Real Statistics Logistic Regression створює її, і вона відображається праворуч на Рис.2 і має наступний вигляд:



Logistic Regression

		# Iter	20		Alpha	0,05	
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper
intercept	-0,92712	4,089956	0,051384	0,820672	0,395694		
X1	-1,65638	0,79661	4,323451	0,037591	0,190828	0,040047	0,909321
X2	-0,74176	0,511647	2,101797	0,147126	0,476273	0,174719	1,298287
X4	0,207677	0,454471	0,208816	0,647697	1,230816	0,505064	2,999438
X6	-0,50152	0,547398	0,839399	0,359569	0,60561	0,207132	1,770677
X7	0,438482	0,70124	0,390993	0,531778	1,550351	0,392223	6,128115
X8	0,160132	0,521693	0,094216	0,758885	1,173665	0,422161	3,262951
X9	0,927906	0,561646	2,729504	0,09851	2,529208	0,841223	7,604281
X10	0,279727	0,421491	0,440447	0,506907	1,322769	0,579043	3,021745
X11	1,329988	0,719014	3,421532	0,064351	3,780997	0,923805	15,47505
X12	0,004539	0,596538	5,79E-05	0,993929	1,00455	0,312031	3,234038

Classification Table

	Obs Suc	Obs Fail	Total
Pred Suc	10	2	12
Pred Fail	4	24	28
Total	14	26	40
Accuracy	0,714286	0,923077	0,85
	Sensitivity	Specificity	
Cutoff	0,5		
AUC	0,876374		

Рис.2 Частина вихідної статистики: обчислені коефіцієнти моделі $b_0, b_1, b_2, \ldots\ b_{12}$; ROC крива (Receiver Operating Characteristic); таблиця оцінки прогностичних характеристик.

Таблиця оцінки прогностичних характеристик моделі:

Справжній стан			
	Позитивний стан	Негативний стан	
позитивний прогнозований стан	істино позитивний **TP (true positive)**	хибно позитивний **FP (false positive)** помилка I роду	**TP+FP**

негативний прогнозований стан	хибно негативний **FN (false negative)** помилка II роду	істино негативний **TN (true negative)**	**TN+ FN**
	TP+FN	**TN+FP**	**N= TP+FN+ FP+TN**

Якість біостатистичного бінарного тестування (і діагностичного, і скринінгового) характеризується певним балансом між двома важливими статистичними мірами продуктивності моделі – чутливістю та специфічністю. Розглянемо на нашому прикладі обчислення даних характеристик і які саме з них обчислює Real Statistics Logistic Regression:

Чутливість (sensitivity, *true positive rate,* recall, hit rate) показує частку істинно позитивних об'єктів ($\boldsymbol{Y = 1}$), які визначені правильно (частка об'єктів, що справді має заданий стан $\boldsymbol{Y = 1}$, і були правильно визначені тестом як такі, що мають цей стан):

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{10}{10 + 4} = 0{,}714.$$

Чутливість є мірою того, наскільки добре тест може визначати частку чи відсоток істинно позитивних серед усіх зразків. Для нашої моделі це є 71,4%.

Специфічність (*specificity, selectivity*, *true negative rate*) показує частку істинно негативних об'єктів ($\boldsymbol{Y = 0}$), частина об'єктів, які справді не має певного стану ($\boldsymbol{Y = 0}$), яку було, правильно саме, так і визначено:

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{24}{24 + 2} = 0{,}92 \Longrightarrow 92\%.$$

Представлена модель є задовільною тим більше, що точність «Accuracy» є 85%, але похибка, таки, досягає 15%. Такі характеристики дають можливість доволі добре прогнозувати реальний стан речей. Модель буде тим кращою, якщо вона точніше визначає людей, що мають певний стан (той, що досліджується), тобто кількість визначених хибнопозитивними чи хибнонегативними має бути доволі малим. Тобто - метою діагностики є з найбільшою правдоподібністю визначення людей - носіїв певного заданого стану, як насправді мають цей стан. Велика кількість

хибнопозитивних чи хибнонегативних ідентифікацій призводить не тільки до неправильного прогнозу, але в реальному житті має такі наслідки, як необхідність додаткових тестувань (підвищення ціни обстеження) та, можливо, різне психологічне травмування чи, навпаки, при хибнонегативному прогнозі – можна пропустити наявність реальної загрози, неправильний економічний прогноз, тощо.

Додатково до чутливості та специфічності, було розраховано наступні прогностичні показники:

Прогностичність позитивного результату (***PPV***, Positive Predictive Value) є доля правильних прогнозів, для яких модель прогнозує наявність випадку:

$$PPV = \frac{\text{Sensitivity} \times \text{ Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + \text{ } (1 - \text{Specificity}) \times (1 - \text{ Prevalence})} \cdot 100\%. =$$

$$PPV = \frac{0{,}714 \cdot 0{,}252}{0{,}714 \cdot 0{,}252 + (1 - 0{,}92) \cdot (1 - 0{,}252)} \cdot 100\% = 75.04\%$$

Тобто 75.04% усіх випадків, які модель визначає як позитивні, насправді є істинно позитивними.

Прогностичність негативного результату (***NPV***, Negative Predictive Value) є доля вірних прогнозів, для яких модель прогнозувала відсутність випадку:

$$NPV = \frac{\text{Specificity} \times \text{ } (1 - \text{Prevalence})}{(1 - \text{Sensitivity}) \times \text{ Prevalence} + \text{ Specificity} \times \text{ } (1 - \text{ Prevalence})} \cdot 100\%$$

$$= NPV = \frac{0{,}92 \cdot (1 - 0{,}252)}{(1 - 0{,}714) \cdot 0{,}252 + 0{,}92 \cdot (1 - 0.252)} \cdot 100\% = 90.5\%$$

Отже, 90.5% - це кількість вірних прогнозів, для яких модель прогнозувала відсутність випадку.

Поріг поширеності - Prevalence:

$$\text{Prevalence} = \frac{\sqrt{a(1-b)} + b - 1}{a + b - 1} = \frac{\sqrt{0{,}714(1 - 0{,}92)} + 0{,}92 - 1}{0{,}714 + 0{,}92 - 1} = \frac{0{,}159}{0{,}634} = 0{,}252.$$

де a – чутливість (Sensitivity), а b – специфічність (Specificity):

Ці показники демонструють, що модель є більш точною в прогнозуванні негативних результатів, ніж позитивних. Результат може бути пояснено специфікою

вибірки та сильним впливом факторів, що перешкоджають успішності студентів у складних умовах.

Коефіцієнт невлучання (miss rate) це показник, який характеризує частку істинно позитивних випадків, які не були виявлені тестом, тобто частку хибно негативних результатів серед усіх дійсно позитивних випадків:

$$\text{Коефіцієнт невлучання} = \frac{FN}{TP + FN} = \frac{4}{14} = 0.2857$$

Отже, 28.57% дійсно позитивних випадків (коли подія мала місце) були помилково визначені як негативні (подія не відбулась) в даній моделі. Тобто приблизно 29% позитивних результатів не були розпізнані моделлю.

Побічний продукт або хибнопозитивний рівень (false positive rate) — це показник, який характеризує частку хибно позитивних результатів серед усіх дійсно негативних випадків.

$$\text{Побічний продукт або хибнопозитивний рівень} = \frac{FP}{FP + TN} = \frac{2}{26} = 0.0769$$

Для нашої моделі він склав 7.69%, що свідчить про невелику кількість хибних прогнозів позитивного результату.

Рівень хибного виявлення (false discovery rate) — це показник, який показує частку хибно негативних результатів серед усіх випадків, які були визначені як позитивні.

$$\text{Рівень хибного виявлення} = \frac{FN}{FN + TP} = \frac{4}{14} = 0.2857$$

У нашому випадку рівень хибного виявлення становить 28.57%. Це означає, що приблизно 28.57% всіх випадків, які були визначені моделлю як позитивні, насправді є хибними.

Рівень хибного пропускання (false omission rate) — це показник, який характеризує частку хибно негативних результатів серед усіх випадків, які були визначені як негативні.

$$\text{Рівень хибного пропускання} = \frac{FN}{FN + TN} = \frac{4}{28} = 0.1429$$

Це означає, що приблизно 14.29% всіх випадків, які були визначені моделлю як негативні, насправді є хибними.

6 Аналіз кривої операційних характеристик (ROC curve analysis)

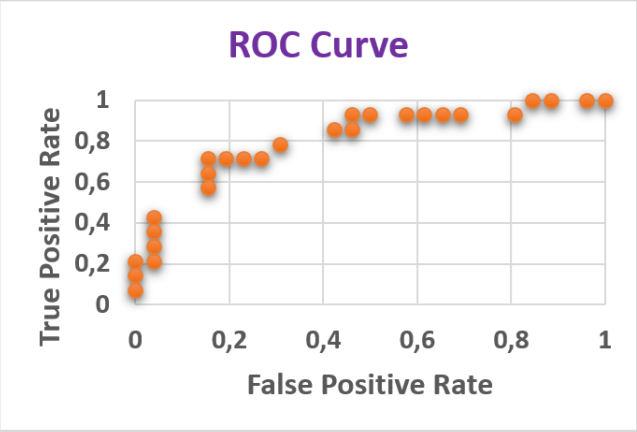
Прогностичні характеристики логістичної моделі регресії (або деякого тесту) добре описуються кривою операційних характеристик моделі (Receiver Operating Characteristic – ROC curve analysis). Інструмент аналізу даних Real Statistics Logistic Regression створює її, і вона відображається прямо на листі розрахунків у вигляді графіку залежності чутливості (**True Positive Rate**) моделі від її специфічності (**False Positive Rate**). Якість побудованої моделі може бути оцінена за площею під ROC - кривою (AUC): чим більше значення AUC, тим кращі прогностичні характеристики моделі (для ідеальної моделі – Sensitivity=100% і Specificity=100% площа під ROC - кривою, AUC=1. Модель буде адекватною експериментальним даним, якщо AUC статистично значимо ($p<0.05$) перевищує 0.5 (ROC-крива не співпадає з діагональною лінією). Якщо відрізняється від 0.5, то модель прогнозує не краще ніж при випадковому виборі прогнозу (за частотою випадків), модель не адекватна експериментальним даним.

Якість моделі умовно може бути оцінена як: відмінна (при AUC≥0.9), дуже добра (при 0.8≤AUC≤0.9), добра (при 0.7≤AUC≤0.8), задовільна (при 0.6≤AUC≤0.7), погана (при 0.5≤AUC≤0.6). Показник AUC часто використовується для порівняння якості моделей (тестів).

На Рис.3 зображено криву операційних характеристик нашої моделі з AUC=0,88, що дає можливість говорити про дуже добру її якість. Але ряд коефіцієнтів рівняння $b_i, i = \overline{1,n}$ не є значущими, тому доцільно застосувати метод вибору мінімального набору змінних, що включаються в модель (Stepwise selection based on p-value) та залишити тільки статистично значимі ($p<0.05$).

Logistic Regression

		# Iter	20		Alpha	0,05	
	coeff	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*
intercept	**-1,0213**	2,71072	0,14195	**0,70635**	0,36013		
X1	**-1,32806**	0,62269	4,54872	**0,03294**	0,26499	0,0782	0,89798
X9	**0,77311**	0,42833	3,25778	**0,07109**	2,1665	0,93575	5,016
X11	**1,1585**	0,54092	4,587	**0,03222**	3,18514	1,10331	9,19515



LL statistics

LL	**-19,2461**
LL0	**-25,8979**
Chi-sq	**13,3035**
df	**3**
p-value	**0,00402**
R-sq (L)	**0,25685**
R-sq (CS)	**0,28293**
R-sq (N)	**0,38968**
AIC	**46,4922**
BIC	**53,2478**

Рис.3 Результати розрахунків для моделі логістичної регресії опісля семи кроків метод вибору мінімального набору змінних, що включаються в модель (Stepwise selection based on p-value).

Classification Table

	Obs Suc	Obs Fail	Total
Pred Suc	**9**	**4**	**13**
Pred Fail	**5**	**22**	**27**
Total	**14**	**26**	**40**
Accuracy	**0,642857**	**0,846154**	**0,775**
Cutoff	**0,5**		
AUC	**0,832418**		

Рис.4 Результати розрахунків Таблиці оцінки прогностичних характеристик моделі опісля семи кроків метод вибору мінімального набору змінних, що включаються в модель (Stepwise selection based on p-value) Cutoff=0,5.

Якщо застосувати **J-статистику Юдена** (індекс Юдена, Youden's J statistic), яка фіксує ефективність дихотомічного діагностичного тесту:

$$\text{Youden's J index} = \text{TPR} + 1 - \text{FPR},$$

це можна зробити, поставивши клік на розрахунку ROC Table і знайти максимальне значення Youden's index, далі у стовпчику p-Pred вибрати відповідне значення ймовірності – воно буде давати найкраще значення **Cutoff**=0,462 замість стандартного 0,5. В нашій моделі отримали Youden's index =1,56, тобто чутливість

(TPR) 71,4%, специфічність (FPR) 84,6%. Можна побачити, що характеристики моделі покращились (Рис.5).

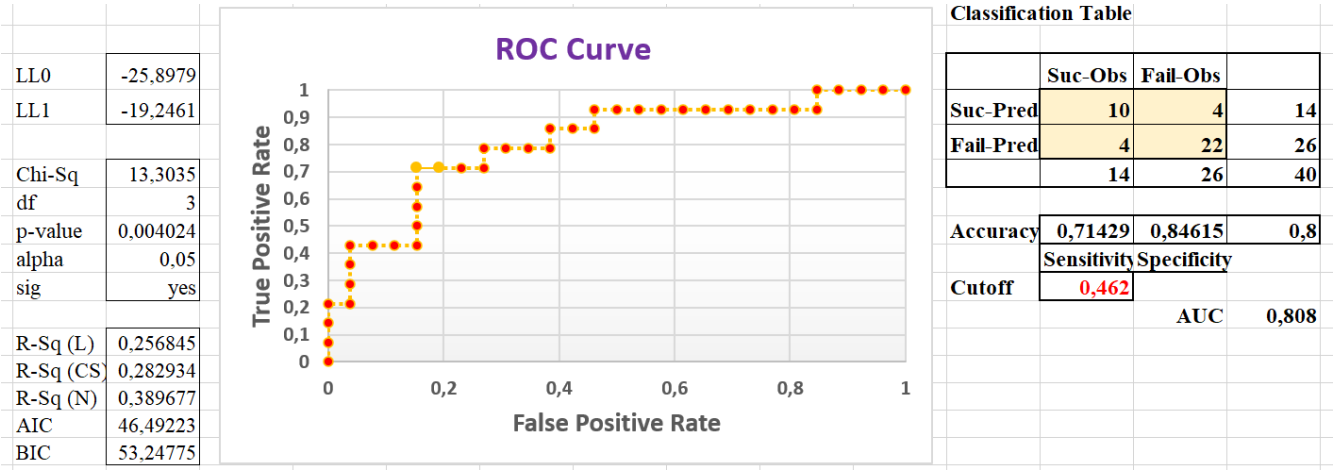


Рис.5 Результати розрахунків Таблиці оцінки прогностичних характеристик моделі опісля семи кроків метод вибору мінімального набору змінних, що включаються в модель

(Stepwise selection based on p-value) Cutoff=0,462.

Для даної моделі AIC=46,5, BIC=53,25, та площею під кривою операційних характеристик моделі (ROC-curve) AUC=0,81, це статистично значимо ($p<0.05$) перевищує 0.5, що є свідченням адекватності побудованої моделі та, за класифікацією, є дуже добрим показником. Розраховані коефіцієнти моделі для трьох факторів (Рис.4 та 5) дають наступну формулу прогнозування:

$$\ln\left(\frac{p}{1-p}\right) = -1{,}02 - 1{,}33X_1 + 0{,}77X_2 + 1{,}16X_3.$$

7 Висновки

Метою даного дослідження було знайти фактори, що можуть мати визначальний вплив на вибір студентів продовжувати навчання і фахову роботу в Україні. Такими факторами стали 1) те, що події, пов’язані із російським вторгненням в Україну, стали значним стресовим досвідом; 2) сильні фізичні реакції (серцебиття,

утруднене дихання, потіння), коли щось нагадує про події російського вторгнення (не є статистично значимим $p = 0{,}07$); 3) моральний настрій щодо навчання під час російської агресії. Аналіз результатів показує, що модель логістичної регресії здатна доволі точно (до 85%) передбачати кількість студентів, що мають зараз намір залишатись в Україні, та фактори, що можуть стати визначальними для такого вибору, хоча все може змінюватись, особливо за умов продовження агресивної війни з рф. Проте, хоча чутливість моделі можна вважати задовільною, її можна підвищити, використовуючи більшу вибірку або вдосконалюючи модель, наприклад, за допомогою значно детальнішого опитування.

Також, слід зауважити, що основна практична цінність логістичної моделі регресії полягає у визначенні факторних ознак, що пов'язані з результуючою змінною $Y_i, i = \overline{1,n}$, для оцінки сили незалежного впливу цих ознак (при урахуванні інших факторів) на результат, а не для прогнозування результату на практиці.

References:

1. McDonald, J.H. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland, 2014. 305 p.
2. The data analysis for this paper was generated using the Real Statistics Resource Pack software (Release 8.9.1). Copyright (2013 – 2023) Charles Zaiontz. www.real-statistics.com
3. Biostatistics Manual. Analysis of Medical Research Results in the EZR (R–statistics) Package / V. G. Guryanov, Yu. E. Lyakh, V. D. Pariy, O. V. Korotkyi, O. V. Chalyi, K. O. Chalyi, Ya. V. Tsekhmister: Textbook. – K.: Vistka, 2018. – 208 p. ISBN 978-617-7157-67-9
4. ROC curve analysis https://www.medcalc.org/manual/roc-curves.php

5. David C. Howell Statistical Methods for Psychology. Wadsworth, Cengage Learning, University of Vermont, 2010, 2007. 780 p. https://labs.la.utexas.edu/gilden/files/2016/05/Statistics-Text.pdf

O.O. Vrublovska, O.V. Mulyk (2025). Analysis of survey results using the stepwise method in the model of logistic regression on the platform of MS Excel (the survey of Ukrainian students in February 2024). *Mathematics in Modern Technical University*, 2025(2), 65-82.

Submitted: 2024-09-28
Accepted: 2024-11-17

Abstract. There are presented base formulas of logistic regression and as exactly they can be realized on the platform of MS Excel (Real Statistics, Logistic of and of Probit Regression). The used statistical measures of the productivity of tests of binary classification are considered widely: Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), prevalence of case (Prevalence) and other. The content of McFadden's, Cox and Snell's and Nagelkerke's coefficients of determination is considered. Also, we give the detailed analysis of curve of Receiver Operating Characteristic – ROC curve analysis, as well as application of Youden's J statistic.

The model was evaluated using performance indicators such as sensitivity (71.4%), specificity (92%), and the area under the ROC curve (AUC = 0.81), indicating high model accuracy. Additionally, the odds ratio for key factors showed that an increase in the impact of war-related stress by 1 unit decreases the probability of continuing education by 28.57%. The results obtained provide a better understanding of how external factors affect students' psychological state and academic performance, allowing for the development of recommendations to improve student support in crisis conditions.

Key words: logistic regression, multivariate model, stepwise method, MS Excel, ROC-curve, AIC, BIC, AUC, p-value.