Міністерство освіти і науки
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра системного аналізу

«Прогнозування спортивних івентів за допомогою дискретних ланцюгів Маркова та логістичної регресії»

Текстова частина до курсової роботи
за спеціальністю «Системний аналіз» 6.050103

Керівник курсової роботи:
доцент,
Чорней Р. К.
_____(підпис)
“___” _____ 2020 р.
Виконав студент:
Коваленко Р. В.
“___” _____ 2020 р.

Київ 2020

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри системного аналізу,

проф., д.ф.-м.н.

_____ Р. К. Чорней

“____”_____201_ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студенту Коваленку Руслану _____

3 _____ курсу _____ факультету інформатики

ТЕМА: Прогнозування спортивних івентів за допомогою дискретних ланцюгів Маркова та логістичної регресії

Вихідні дані:

- Теоретичні частина
- Практична частина

Зміст теоретичної частини до курсової роботи:

Індивідуальне завдання

Анотація

Вступ

1. Теоретичні основи

2. Методологія. Підходи в розробці моделі

Висновок

Список використаних джерел

Дата видачі “___ ” _____ 201_ р.

Керівник _____ Завдання отримано _____

Календарний план виконання курсової роботи

Тема: Прогнозування спортивних івентів за допомогою дискретних ланцюгів Маркова та логістичної регресії

№ п/п	Назва етапу курсового проекту (роботи)	Термін виконання етапу	Приміт ка
1.	Отримання завдання на курсову роботу	1 листопада 2019 р.	
2.	Огляд літератури за темою роботи	5 квітня 2020 р.	
3.	Огляд теоретичних відомостей	10 квітня 2020 р.	
4.	Аналіз схожих моделей прогнозування	15 квітня 2020 р.	
5.	Опис теоретичних фреймворків	20 квітня 2020 р.	
6.	Пошук та впорядкування даних	22 квітня 2020 р.	
7.	Виконання практичної частини	25 квітня 2020 р.	
8.	Опис моделі по результатах практичної частини	30 квітня 2020 р.	
9.	Оцінка результатів.	3 травня 2020 р.	
10.	Оформлення результатів та слайдів	10 травня 2020 р.	
11.	Захист курсової роботи	25 травня 2020 р.	

Студент Коваленко Руслан _____

Керівник Чорней Р. К. _____

“_____” _____ _____ р.

Зміст

Анотація

Коваленко Р. В. Прогнозування спортивних івентів за допомогою дискретних ланцюгів Маркова та логістичної регресії.

У курсовій роботі було розглянуто підхід, за допомогою якого була побудована модель, здатна прогнозувати результат матчів Української Прем'єр Ліги з футболу, будувати список команд, впорядкований за ймовірнісним показником успішності в будь-який момент сезону.

В першому розділі було розглянуто теоретичні фреймворки, за допомогою яких будувалася та тренувалася модель. Система спроектована на дискретні ланцюги Маркова. Було описано один із видів регресії - логістичний.

Другий розділ містить закріплення практичної частини: інформацію про підхід побудови моделі на базі ланцюгів Маркова; способи застосування логістичної регресії для класифікації матчів та прогнозування транзитивної ймовірності станів ланцюгів. Впорядкування команд турніру за стаціонарною ймовірністю.

У висновку ми підсумовуємо результати, яких вдалося досягти під час виконання практичної частини.

Ключові слова: ланцюги Маркова, дискретний, стаціонарний розподіл, стаціонарна ймовірність, властивість Маркова, транзитивна діаграма, транзитивна матриця, регресія, логістична регресія, сигмоїдна функція, python, sklearn.

Вступ

Актуальність теми. Проекція моделі прогнозування футбольних матчів на дискретні ланцюги Маркова дозволяє користуватися однією із найважливіших характеристик даного підходу - властивістю Маркова, а саме фактом, що стохастичний процес є безпам’ятним. Оскільки релевантна інформація знаходиться в поточному стані, і процес не враховує "історичних стрибків" між станами, це дає нам можливість створити умови для нашої системи близькі до реальності, де результати матчів не залежать від результатів минулих матчів. Прогнозування ймовірності на базі відомих методів логістичної регресії є значною перевагою, оскільки інструменти дозволяють встановити фактори, які напряму впливають на результат, оцінити похибку прогнозів. На базі цього побудувати вектор стаціонарного розподілу ймовірностей, який дозволяє впорядкувати команди турніру за стаціонарними ймовірностями, елементами знайденого вектора.

Метою курсової роботи є структуризація та проекція підходів дослідження спортивних івентів на футбольний турнір. Розробка системи, яка здатна впорядкувати команди турніру за стаціонарною ймовірністю, беручи за основу результати навчання регресії для передбачення ймовірностей переходу для того, щоб заповнити транзитивну матрицю станів.

Для досягнення поставленої мети визначені такі завдання:

- пошук датасету, який містить статичну інформацію про результати матчів будь-якого футбольного турніру. Впорядкування даних;
- обрахунок параметрів: запас перемоги на домашньому полі, ймовірність перемоги на домашньому полі;
- виведення ймовірності нейтрального поля;

- перетворення даних для навчання логістичної регресії, для передбачення ймовірності перемоги на домашньому полі;
- побудова транзитивної матриці. Обрахування стаціонарної ймовірності.

Структура курсової роботи. Робота складається із анотації, вступу, двух розділів, висновку та списку використаних джерел (включає в себе 13 найменувань).

Розділ 1. Теоретичні відомості

1.1 Марківська теорія

1.1.1 Ознайомлення із фундаментальною теорією

Фундаментальною для цієї роботи є математична модель марківських процесів, а конкретніше ланцюги Маркова. Основними поняттями марківської теорії є *стани* системи та *переходи* з одного стану в інший. За замовчуванням, система знаходиться в певному стані, якщо її можна описати деяким набором змінних, які задають цей стан. Також система виконує перехід із одного стану в інший, якщо змінні, що описували її поточний стан змінюються на значення, що описують відповідний інший стан. Ланцюги Маркова - це марківський процес з дискретним часом та дискретним простором станів. Тобто ланцюги Маркова - це дискретна послідовність станів, де кожний стан належить дискретному простору станів. Формально ланцюги Маркова можна визначити таким чином:

$$\{ X_0, X_1 \ldots \} = X,$$

де X_t - це стан, який розглядаємо в момент часу t[1].

1.1.2 Memoryless або Марківська властивість

Процес, який ми вище описали, має одну дуже характерну властивість:

X_{t+1} залежать тільки від X_t, і не залежить від $X_0, X_1 \ldots X_{t-1}$.
Тобто, майбутнє залежить лише від теперішнього положення, і не залежить від минулого. Дана характеристика називається memoryless або

Марківською властивістю. Така властивість формально може бути записаною наступним чином:

$$P(\ X_{t+1} = \ s \mid \ X_t \ = s_t, X_{t-1} = s_{t-1}, \ldots, \ X_0 = s_0) \ = \ P(\ X_{t+1} = \ s \mid \ X_t \ = s_t),$$

виконується в будь-який момент часу і для будь-якого стану. Таким чином, з останнього рівняння: розподіл ймовірностей для наступного стану залежить від поточного стану, але не від минулих станів. Це означає, що релевантна інформація знаходиться в поточному стані, і процес не враховує "історичних стрибків" між станами. Щойно запропоноване визначення ілюструє основний елемент того, що становить ланцюги Маркова, а саме властивість Маркова, яка означає, що стохастичний процес є безпам’ятним [1].

1.1.3 Транзитивна ймовірність. Матриця

Одним із найзручніших способів задання ланцюгів Маркова є транзитивні діаграми. Транзитивна діаграма або транзитивна діаграма станів - це орієнтований граф, який будується за певними правилами:

- кожна нода відповідає стану;
- ноди p та q з’єднуються направленим ребром a, якому відповідає значення транзитивної функції $\delta(p, a) \ = p$.

Таким чином, діграма називається транзитивною, оскільки вона відображає переходи між станами [2].

Отже, повернемося до ланцюгів Маркова. На транзитивній діаграмі, X_t відповідає ноді, в якій ми знаходимося в момент часу t.

Хоча існує зовсім інший підхід: можна узагальнити ймовірності із транзитивної діаграми в матрицю. Це представлення називається транзитивною матрицею. Елементи якої мають вигляд:

$$p_{ij} = P(\ X_{t+1} = \ i \mid \ X_t \ = j) \qquad i, j \in X$$

Це головнний інструмент, який використовується для аналізу Марківських ланцюгів [1]. Матриця має вигляд:

$$\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.7 & 0.2 & 0.1 \\ 0.8 & 0.1 & 0.1 \end{bmatrix}$$

1.1.4 Властивості транзитивної матриці

В транзитивній матриці:

- рядки відповідають поточному стану, із якого відбудеться наступний крок;
- стовпчик відповідає ймовірному стану, тобто X_{t+1}.
- на перетині i-го та j-го елементів знаходиться значення умовної ймовірності, що наступним станом буде j, враховуючи, що поточний стан i.

Деякі зауваження, про які варто пам’ятати при побудові транзитивної матриці:

- матриця має включати всі можливі стани із простору станів X ;
- матриця є квадратною, оскільки X_{t+1} та X_t мають однаковий простір станів, до який система може перейти;
- сума рядка матриці P має дорівнювати 1 за означенням ймовірності:

$$\sum_{j}^{N} p_{ij} = \sum_{j}^{N} P(\ X_{t+1} = \ j \mid \ X_t \ = i) = \sum_{j}^{N} P_{\{X_t = i\}}(\ X_{t+1} = j) = 1$$

- сума стовпчиків необов’язково дорівнює 1 [1,3].

1.1.5 Стаціонарний розподіл ланцюга Маркова

Як частина поняття ланцюгів Маркова існує деякий розподіл ймовірностей на станах в момент часу 0. На кожному кроці ланцюга відбувається еволюція розподілу: деякі стани можуть ставати більш імовірними, а інші - менш ймовірними, на це впливає матриця транзитивності.

Нехай маємо ланцюг Маркова із простором станів X, матриця транзитивності P. Тоді π - стаціонарний розподіл. Якщо $\pi = (\pi_i, i \in X)$ - це розподіл на просторі станів X(де π - це вектор-рядок з $|X|$ компонент, таких, що $\sum_i^N \quad \pi_i = 1, \pi_i \geq 0$). Тоді, припускаючи що початковим розподілом X_0 рівний π, ми зробимо наш ланцюг Маркова стаціонарним, якщо

$$\pi = \pi P \tag{1}$$

[4,5]

1.2 Регресійний аналіз

Регресійний аналіз - це надійний метод виявлення змінних, які впливають на предметну область, що ми досліджуємо. Процес виконання регресії дозволяє точно визначити: які фактори мають найбільше значення при навчанні, які з них можна ігнорувати та як ці фактори впливають один на одного. Тобто, задачею цього статистичного підходу є оцінювання взаємозв'язків між залежними змінними, вони також називаються output. Існує велика кількість ситуацій, в яких IO (input-output) "відносини" є важливими, наприклад, коли output є дискретним, а не безперервним[6,7].

1.2.1 Логістична регресія

Цікавим є випадок, коли вихід є бінарним, наприклад, чи перемогла команда на домашньому полі чи ні. В нашому дослідженні нас цікавить окремий вид регресії - логістичний, оскільки нам потрібно, щоб модель в деякий момент часу із мінімальною похибкою вгадувала бінарний вихід. Цей процес називається класифікацією.

Нехай один із класів називається "1", а інший "0". Часто перший означає позитивну відповідь, а другий - негативну. Нехай, p - це пропорція відповідей із результатом 1, таким чином, $1 - p$ буде ймовірністю негативної відповіді, а саме відповіді 0. Відношення ймовірностей $p/(1 - p)$ називається *odds*, а процес логарифмування попереднього значення має назву *logit***.** Математично логіт перетворення можна записати таким чином:

$$l = logit(p) = ln(\frac{p}{1-p}) \tag{2}$$

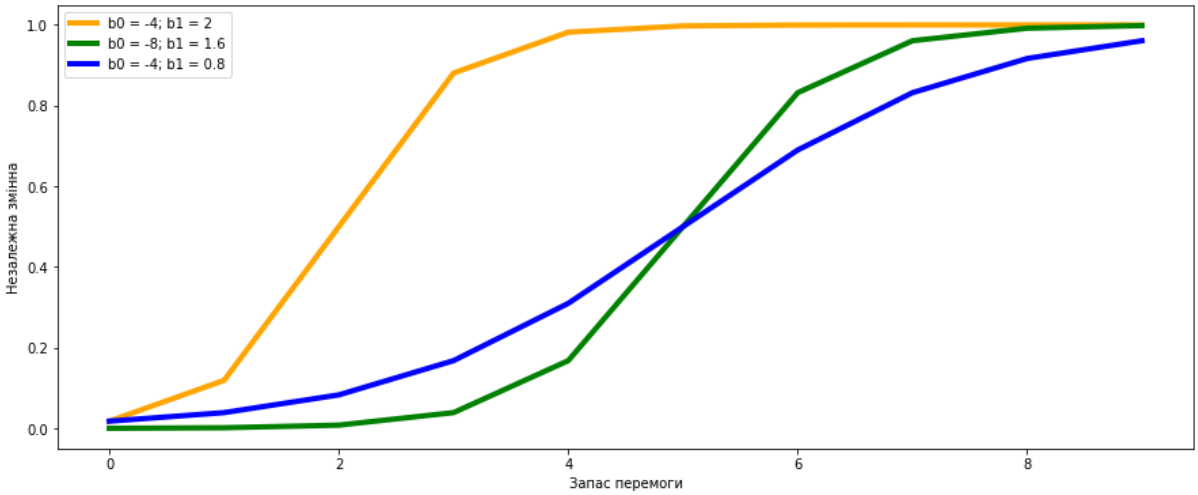
Варто зауважити, що в той час, коли правила ймовірності обмежуть значення p від 0 до 1 включно, тоді як логіт може набувати значення на проміжку $(-\infty; +\infty)$. Цікавим є момент, коли p дорівнює 0.5, в такому випадку logit буде рівний 0. Існує також зворотній процес logit, який називається логістичною трансформацією або logistic. Для математичного перетворення будемо користуватися наступним рівнянням:

$$p = logistic(l) = \frac{e^l}{1 - e^l} \tag{3}$$

Як було сказано вище, ми розглядаємо той випадок логістичної регресії, коли вихід $Y = \{ 0,1 \}$ - бінарний. На даному етапі базових понять достатньо, щоб побудувати модель логістичної регресії. Ми моделюємо *log odds*, як лінійну функцію:

$$ln(\frac{p}{1-p}) = b0 + xb1 \tag{4}$$

Дана функція демонструє лінійну залежність між p та x, на Мал. 1.2.1_1 видно, як змінюється функція залежно від параметрів b0 та b1.



Мал. 1.2.1_1. Як похибка під час визначення b0 та b1 вплине на прогнозування ймовірності

Для логістичної регресії ми використовуємо натуральний логарифм. Процес пошуку коефіцієнтів називається *fitting* моделі. Бібліотека sklearn на мові програмування Python, яка була використана під час навчання нашої моделі, дозволяє отримати коефіцієнти лінійного рівняння. Отже, маючи коефіцієнти, за допомогою простих математичних перетворень:

$$\frac{p}{1-p} = e^{b0+xb1} \tag{5}$$

Виразимо змінну p:

$$p = \frac{e^{b0+xb1}}{1-e^{b0+xb1}} = \frac{1}{1+e^{-(b0+xb1)}} \tag{6}$$

Помітно, останній вираз набуває вигляду сигмоїда (7). Саме функція, яку ми будемо використовувати в подальшому аналізі турніру. Отже, на даному етапі ми маємо повний набір інструментів, який дозволить користуватися логістичною регресією [7].

1.2.1.1 Сигмоїдна функція

Окремим етапом логістичної регресії є перетворення прогнозованого значення в ймовірність. В такому випадку ми використовуємо функцію сигмоїд. Дана процедура дозволить нам перетворити будь-яке значення в проміжок від 0 до 1.

Функція має вигляд:

$$f(x) = \frac{1}{1+e^{-(x)}} \tag{7}$$

[7,8]

Розділ 2. Методологія. Підходи в розробці моделі

На даному етапі ми розглянемо основне питання: як точно оцінити команди, використовуючи лише основні вхідні дані. Задачу, яку ми перед собою ставимо, це побудова моделі, яка буде в змозі оцінити ймовірнісні характеристики команд будь-якого спортивного турніру. Наша модель використовуватиме ланцюг Маркова та логістичну регресію, яку ми описали в розділах 1.1 Марківська теорія та 1.2 Регресійний аналіз відповідно.

2.1 Збір даних

Інформація, котра потрібна для нашого дослідження, була запозичена із сервісу thesportsdb.com. Впорядкована за сезонними турнірами.

	A	B	C	D	E	F
1	Date	Round	Home_Team	Score_1	Visit_Team	Score_2
2	16.07.2017	Round 1	Illichivets	0	Veres Rivne	0
3	16.07.2017	Round 1	Karpaty	1	Zirka	1
4	16.07.2017	Round 1	Zorya	0	Stal Kamianske	1
5	16.07.2017	Round 1	FC Olexandria	0	Olimpik Donetsk	2
6	18.07.2017	Round 1	Dynamo Kiev	2	Ch. Odessa	1
7	18.07.2017	Round 1	Vorskla	0	Shakhtar Donetsk	3

Мал. 2.1_1. Приклад датасету

Для обрахунку транзитивної ймовірності марківських ланцюгів використовується логістична регресія, при цьому, будуючи модель за підходом П. Квама та Д. С. Сокола в роботі “Логістична регресія / Модель ланцюга Маркова для баскетболу NCAA” [9], нам потрібна незначна частина інформації: команда, яка грає на домашньому полі, команда-суперник та рахунок, а також динамічна інформація, яка полягає в місці в

турнірній таблиці, а саме кількість очок після зіграного матчу, що обраховується по мірі проходження по датасету.

2.2 Проекція моделі на Марківські ланцюги

В цьому підрозділі розглядаються Марківські ланцюги, а саме підхід, який дозволяє створити систему, головною задачею якої є симулювання поведінки спортивних турнірів, а також передбачати результати матчів по окремості. Тобто система, яка використовує великий обсяг даних для навчання самої себе, також оперує своїми процесами подібними на Марківські ланцюги.

Як було зазначено вище, основою нашої моделі є Марківські ланцюги: кожній команді відповідає один стан. Інтуїтивно, переходи між станами керуються деяким агентом в одному із напрямків. Поточний стан, в якому знаходиться агент, відповідає команді, на думку агента яка є кращою. В кожний момент часу агент робить рішення, в якому напрямку йому рухатися далі наступним чином: враховуючи, що він вважає поточну команду i найкращою, він випадковим чином обирає гру, де команда і зустрічається з певним суперником $\boldsymbol{j}$. З ймовірністю p наш агент переходить в стан, який відповідає переможцю даної гри; з ймовірністю $1 - p$ агент рухається до стану команди, яка програла.

2.2.1 Транзитивна ймовірність

Нехай, команда i зіграла N_i ігор, де k-ті ($k \leq N$) матчі були здійснені проти опонента θ_k. Визначимо індикаторну функцію (8).

$$I_{ij} = \begin{cases} 1, & \text{i won j} \\ 0, & \text{i lose match with j} \end{cases} \tag{8}$$

Таким чином, транзитивна ймовірність t_{ij} із стану i за визначенням ланцюгів Маркова, дорівнює:

$$\begin{aligned} t_{ij} &= \frac{1}{N_i}\sum_{k=1}^{N_i}\left[I_{ik}(1-p)+(1-I_{ik})p)\right] \qquad i \neq j \\ t_{ij} &= \frac{1}{N_i}\sum_{k=1}^{N_i}\left[I_{ik}p+(1-I_{ik})(1-p)\right] \qquad i = j \end{aligned} \tag{9}$$

Якщо ми визначимо W_i та L_i, як кількість виграних та програних ігор командою i відповідно, а значення w_{ij} та l_{ij} як кількості перемог і поразок, які команда i отримала проти суперника j. Таким чином, формула транзитивної ймовірності може бути перевизначена в більш інтуїтивній формі:

$$\begin{aligned} t_{ij} &= \frac{1}{N_i}\left[w_{ij}(1-p)+l_{ij}p)\right] \qquad i \neq j \\ t_{ij} &= \frac{1}{N_i}\left[W_{ij}p+L_{ij}(1-p)\right] \qquad i = j \end{aligned} \tag{10}$$

Згідно (9) та (10) рівняння, перехід може бути відтворений підкиданням N-стороннього кубика для вибору гри і підкиданням монети для того, щоб визначити чи буде відповідати наступний стан переможцю матчу з ймовірністю p чи команді, яка програла - $1-p$[9].

2.2.2 Альтернативна транзитивна ймовірінсть

Транзитивний параметр p інтерпретується наступним чином: значення p - це відповідь моделі на питання: “Враховуючи, що команда А перемогла команду В, яка ймовірність, що команда А краща за команду В?”. Оперуючи інформацією із детального опису кожного матчу, можливо отримати відповідь на попереднє питання. Відомий факт, що в подібних до до футболу видах спорту, таких як баскетбол, сокер, хокей чи волейбол,

домашня гра є перевагою для господарів. Саме цим фактором ми скористаємося для побудови моделі. Введемо поняття “запас перемоги”, визначимо його як різницю між рахунком команди-переможця та команди, яка програла. Сенс цього терміну полягає в наступному: результат матчу із ширшим запасом перемоги є ймовірно позитивним для команди із більшою кількістю очок в турнірній таблиці.

В контексті цих понять, ми можемо знайти ймовірність, яка буде давати відповідь на питання: “Враховуючи, що команда А перемогла команду В на домашньому полі, яка ймовірність, що команда А краща за команду В?”.

Нехай, функція $g(x)$ дорівнює різниці кількості очок в турнірній таблиці команди, яка грає на домашньому полі і команди, що грає на чужому полі після матчу x. Визначимо r_g^H як ймовірність, що команда, яка переважає свого опонента на $g(x)$ одиниць, краща за свого опонента. Відповідно r_g^H= 1 - r_g^R - ймовірність (road), яка програє своєму опоненту за одиницями в турнірній таблиці, є кращою за опонента.

Важливо розуміти, що величина $g(x)$ може набувати від’ємних значень у випадку непервної кількості програшів на домашньому полі.

Таким чином, якщо ми визначимо кожний матч як впорядковану пару (i, j) команд, тоді рівняння транзитивної ймовірності для кожної команди i можна записати наступним чином:

$$\begin{aligned} t_{ij} &= \frac{1}{N_i}\left[\sum_{g(i,j)}(1-r^R_{g(x)}) + \sum_{g(j,i)}(1-r^H_{g(x)})\right] && i \neq j \\ t_{ij} &= \frac{1}{N_i}\left[\sum_{j}\sum_{g(i,j)} r^R_{g(x)} + \sum_{j}\sum_{g(j,i)} r^H_{g(x)}\right] && i = j \end{aligned} \tag{11}$$

Поразки, перемоги та запаси перемог, значення яких легко групуються та аналізуються. Значення ймовірностей r_g^H та r_g^R складно оцінити і, водночас, майже неможливо порахувати точно. Саме тому, в наступному підрозділі розглянемо підхід, за допомогою якого ми можемо оцінити значення ймовірностей для кожного g [9].

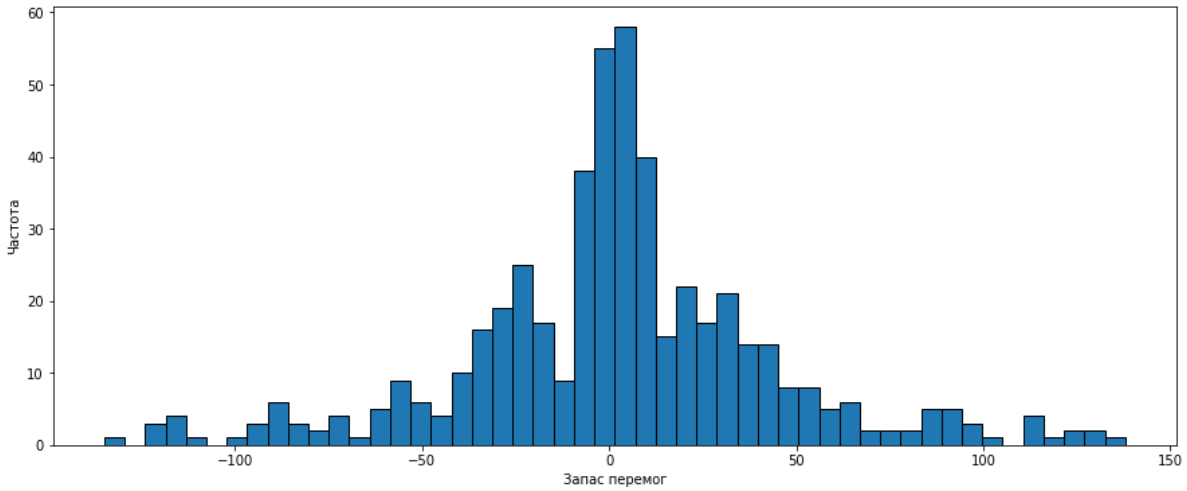
2.3 Логістична регресія для оцінки транзитивної ймовірності

В цьому розділі ми розглянемо логістичну регресію, як інструмент, за допомогою якого ми оцінюємо значення ймовірності r_g^H, а саме ймовірність того, що команда із запасом перемоги g на своєму полі є кращою за свого опонента. Очевидно, що оцінка переваги однієї команди над іншою є складною для точного підрахунку.

Отже, приймаємо за умову, що будь-які дві команди i та j, зіграють лише два рази за сезон турніру, один раз на території i-ої команди, інший - на j-ої. Всі матчі, які відбулися на нейтральному полі не будуть враховані при навчанні нашої моделі, оскільки це суперечить концепції моделі про запас перемог на домашньому полі. Таким чином, питання яке ми піднімаємо: “Враховуючи, що команда А має запас перемог g на домашньому полі проти команди В, яка ймовірність (s_g^H) того, що А переможе на полі противника?”. На наступному етапі, на базі попередньої ймовірності, ми обрахуємо r_g^H. Таким чином, відповімо на питання: “Враховуючи, що команда А має запас перемог в g одиниць над командою В, яка ймовірність, що команда А є кращою за команду В?” [9].

Як і було сказано вище, головною умовою участі матчу в навчанні нашої моделі є факт наявності двох “протилежних” ігор, а саме однієї гри на домашньому полі першої команди і гри на рідному полі іншої команди і

тільки. Для кожної такого матчу ми зберігаємо інформацію про господаря поєдинку, гостя, а також різницю в очках турнірної таблиці після того, як матч здійснився. Нагадаємо, що матчі, проведені на нейтральному полі, не беруться до уваги. Відсутність таких матчів на результати не вплине, а наявність - суперечить концепції.

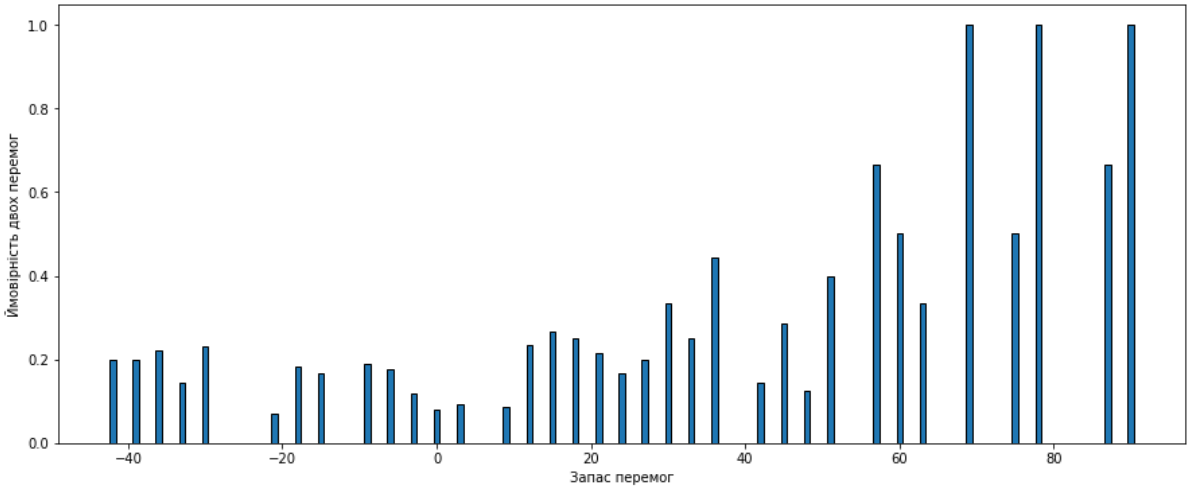


Мал. 2.3_1. Гістограма, яка демонструє частоту запасів перемог

На Мал. 2.3_1 можна спостерігати залежність: при збільшенні різниці між очками команд в турнірній таблиці, зменшується частота. Наступним кроком, знайдемо відповідь на питання: “Враховуючи, що команда A із запасом перемог g перемогла команду B на домашньому полі, яка ймовірність, що команда A переможе другий матч на полі суперника B?”. Для цього обрахуємо частку команд, які отримали перемогу на домашньому та полі суперника від загальної кількості команд, які перемогли на домашньому полі із запасом перемог g.

$$P_g = \frac{K_g}{N_g} \tag{12}$$

На Мал. 2.3_2 відобразимо залежність частоти P_g від запасу перемог g. Наприклад, приблизно 50% команд із запасом перемог 60 одиниць, отримали обидві перемоги: на домашньому полі та полі суперника.



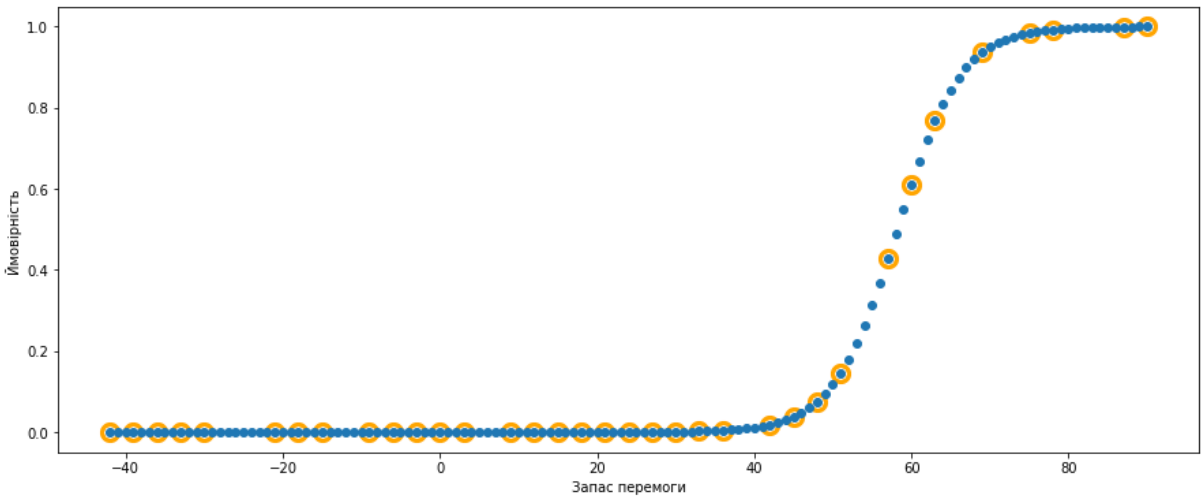
Мал. 2.3_2. Гістограма відображає частку команду, які отримали обидві перемоги для кожного запасу перемоги.

За мету собі ми ставимо навчити нашу модель прогнозувати s_g^Hймовірність. Саме для цього ми використаємо такий потужний інструмент, як логістична регресія. Саме цей підхід дозволить лінеаризувати нелінійну функцію для того, щоб оцінити параметри b0 та b1 рівняння (4) [9].

Для подальших дій із логістичною регресією будемо оперувати рівнянням (6). Як відомо, логістична регресія - це інструмент для класифікації двох чи більше типів даних. Таким чином для того, щоб підготувати дані для класифікації, потрібно визначитися із правилами, якими ми будемо керуватися при обробці. Ведемо в контекст функцію $k(s)$:

$$k(s) = \begin{cases} 1, & s \geq 0.5 \\ 0, & s < 0.5 \end{cases} \tag{13}$$

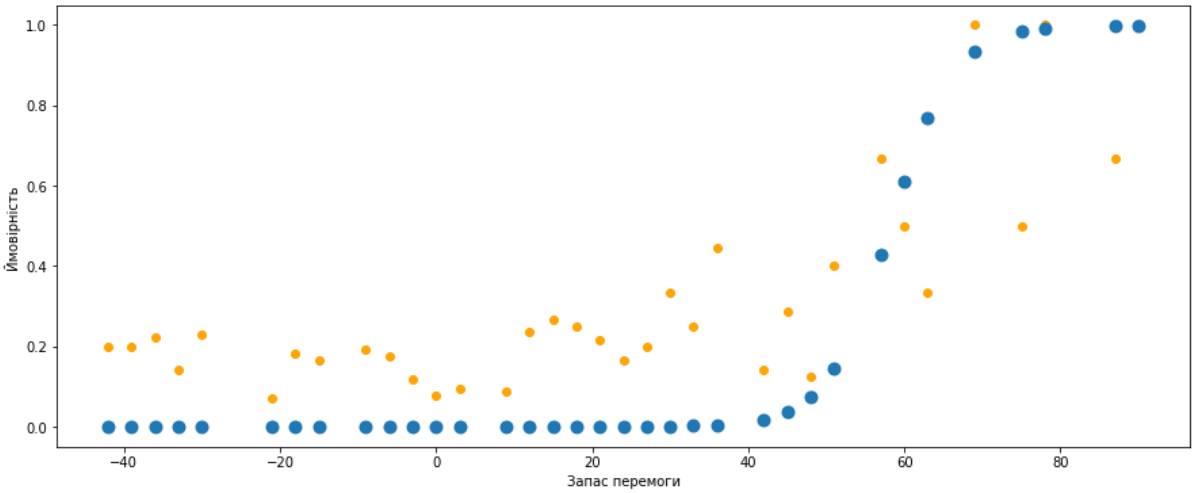
Отже, маємо запас перемоги, а також клас (0 або 1), до якого відноситься відповідна ймовірність s. Отже, результат класифікації відображений на Мал. 2.3_3. Як відомо з концепції логістичної регресії, за регулювання ймовірності відповідає логістична функція - сигмоїд. Під регулюванням мається на увазі оцінка ймовірності для кожного нового входження запасу перемоги.



Мал. 2.3_3. Результати логістичної функції

Оранжевим кольором позначені значення запасу перемог, обраховані на реальних даних. Синім кольором відмічені значення сигмоїди на тестових даних, а саме згенерована послідовність, яка обмежена знизу мінімальним значенням наших даних та зверху відповідно максимальним значенням.

Для оцінки правильності регресійної моделі зобразимо на одному графіку - Мал. 2.3_4, реальні дані, виділені оранжевим кольором, та прогнозовані - синім.



Мал. 2.3_4. Порівняння обрахованої ймовірності та прогнозованої

2.4 Виведення ймовірності нейтрального поля

До цього моменту ми розглядали ймовірність s, яка напряму залежить від місця проведення матчу. Тобто, ймовірність, що команда А переможе команду В на полі команди В, якщо в першому матчі, команда А отримала перемогу. Але це не було нашою головною метою, оскільки це не дозволяє нам заповнити транзитивну матрицю ланцюгів Маркова. Для цього нам необхідно оцінити ймовірність r, тобто ймовірність, що команда А переможе команду В на нейтральному полі, враховуючи, що команда А перемогла команду В на полі команди А. В цьому підрозділі, ми розглянемо виведення s із r.

Припустимо факт, що гра відбувається на конкретному полі може бути оцінений числовим значенням, наприклад h. Іншими словами, за нашим припущенням, числове значення запасу перемог збільшується на h, назвемо це h- домашньою перевагою. Беручи до уваги цей факт виходить, що команди на своєму полі мають очікувану перевагу в h одиниць. Ми також припускаємо, що ймовірність перемоги команди А над командою В на полі другої команди буде дорівнювати $s = 0.5$, якщо різниця очок дорівнює 0. Якщо очікувана різниця очок після матчу на домашньому полі

команди B, в такому випадку, перевага домашнього h точно скасовує перевагу команди A над командою B. Саме тому, після матчу, очікувана перевага над командою B на домашньому полі першої команди, буде дорівнювати $2h$, оскільки команда A має властиву їй перевагу і перевагу домашнього поля.

У випадку, коли ми розглядаємо нейтральне поле, команда, котра перемагає свого суперника на g очок на домашньому полі, очікувано отримає перемогу на $g - h$ на нейтральній полі. Отже, тепер, коли ми розуміємо, що s_g^H визначало ймовірність перемоги при очікуваній різниці очок $g - 2h$, то ми можемо вивести ймовірність перемоги для очікуваної різниці очок $g - h$, таким чином:

$$r_g^H = s_{g+h}^H \tag{14}$$

[9]

2.5 Заповнення транзитивної матриці за допомогою машинного навчання

На даному етапі, ми розглянули всі деталі механізму, за допомогою, якого ми будуємо нашу модель. Підкріплюючи це все практичною частиною, маємо готовий інструмент, який допоможе нам порахувати кожний елемент транзитивної матриці, беручи до уваги, що до моменту проведення етапу навчання логістичної регресії, ми не мали змоги заповнити головну матрицю ланцюгів Маркова.

Отже, для того, щоб заповнити матрицю, скористаємося навченим агентом машинного навчання. За допомогою етапу навчання, ми встановили коефіцієнти $b0$ та $b1$. Таким чином, користуючись формулою (6), ми здатні прогнозувати ймовірність для будь-якого запасу перемог.

Якщо для навчання агента ми використовували інформацію про 5 сезонів Прем'єр Ліги України з футболу, то для побудови транзитивної матриці матиме місце лише один сезон. Зауважимо, що сезон обов'язково має містити для кожного матчу відповідний матч-відповідь. Після того, як ми узгодили всі дані, необхідні нам для побудови матриці, можемо перейти до етапу обрахування елементів. Для цього ми будемо користуватися формулами (11). Залишається одне невирішене питання: в підрозділі 2.4 Виведення ймовірності нейтрального поля, ми розібралися, яким чином, можна ігнорувати домашню перевагу h, проте не використовуємо цю додаткову інформацію. Отже, потрібно визначати очікувану величину h. Користуючись рівняння (14), ми можемо конвертувати ймовірність нейтрального поля в ймовірність перемоги на домашньому полі.

Отже, маємо транзитивну матрицю для сезону 2017- 2018 розміром *12 x 12*, оскільки маємо 12 команд.

2.6 Пошук стаціонарної ймовірності із транзитивної матриці

На наступному кроці знайдемо стаціонарні ймовірності, за допомогою яких можемо впорядкувати команди, і таким чином, утворимо рейтинг команд, де команда з найбільшою стаціонарною ймовірністю буде першою в рейтингу, і відповідно, з найменшою ймовірністю - останньою. Таке впорядкування допоможе нам передбачати результати. Очевидно, ймовірність перемоги команди буде найкращою, тобто більшою за ймовірність перемоги команди суперника.

Розглянемо підхід, який використовували Галаган Т., Портер М. і Муча П. в своєму дослідженні Random Walker Ranking for NCAA Division 1-A Football[10]. Враховуючи, що транзитивна ймовірність стану

визначена$P = [t_{ij}]$, знайдемо стаціонарну ймовірність, про яку ми говорили в підрозділі 1.1.5 Стаціонарний розподіл ланцюгів Маркова, для кожної команди за допомогою рівняння (1)[9].

Коли, ми поставимо у відповідність елементам відсортованого вектора рівняння (1) команди чемпіонату України з футболу, сезону 2017 - 2018, отримаємо прогнозований рейтинг команд:

Шахтар - Динамо - Зоря - Ворскла - ФК Маріуполь - Верес - ФК Карпати - ФК Олександрія - Олімпік - Зірка - Сталь - Чорноморець.

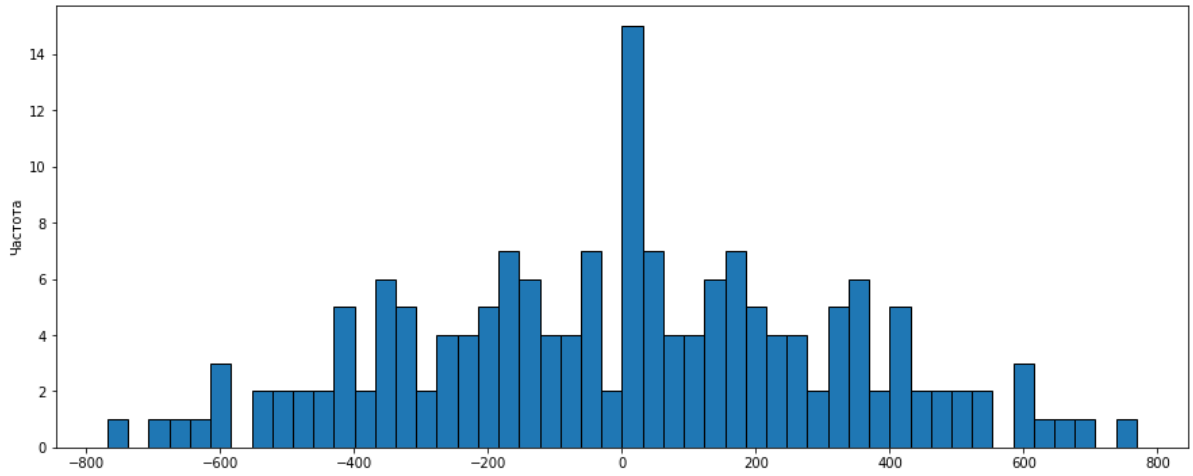
2.7 Запаси перемог на базі стаціонарної ймовірності

Дослідники Каплан та Гастка [11] або Брейтер та Карлін [12] в своїх роботах підкреслювали той факт, що вибір команди, яка оцінена за стаціонарною ймовірністю на першому місці, не обов'язково призведе до найкращої стратегії. Натомість, вони стверджували, що для того, щоб побудувати оптимальну стратегію, потрібно знайти ймовірність перемоги команди. Отже, використаємо нашу модель, а саме марківські ланцюги\ логістичну регресію, для того, щоб знайти ці ймовірності перемоги. Дослідники, про яких ми згадували вище, використали дуже простий підхід. В їхніх роботах, вчені досягли цікавого висновку, відповідно до якого, різниця в очках між двома командами, може бути легко оцінена за допомогою лінійної функції, яка базується на різниці двох стаціонарних ймовірностях:

$$x_{ij} = 9180(\pi_i - \pi_j) \tag{15}$$

Хоча, коефіцієнт помилки (9180) потребує детального аналізу. Залежить від підходу побудови моделі, а також методу, за яким обраховується запас перемог. Якщо ми обрахуємо запас перемог для

кожного матчу за допомогою лінійної функції (15) з вже відомим нам коефіцієнтом - 9180, зобразимо частоти на Мал. 2.7_1[9].



Мал. 2.7_1. Гістограма, яка демонструє частоту запасів перемог, знайдених за допомогою лінійної функції (15)

Висновок

Виконуючи дану роботу, було реалізовано повноцінну систему, яка базується на дискретних ланцюгах Маркова, транзитивна матриця станів якої була заповнена за допомогою комбінацій параметрів, що напряму впливають на матч. Логістична регресія, що здатна враховувати відносини цих факторів та прогнозувати значення ймовірності на базі проведеного заздалегідь етапу навчання. Під час обрахунку значення ймовірності були виконані певні дії щодо покращення результатів за рахунок оцінювання переваги домашнього поля. На базі фреймворку ланцюгів Маркова, обраховувався стаціонарний розподіл ймовірності, за допомогою якого будується впорядкований список команд Української Прем'єр Ліги з футболу. Таким чином, результуючий список команд можна використовувати для окремих передбачень результатів матчів так і прогнозувань результатів повноцінного турніру.

Як результат аналізу транзитивної матриці, ми отримали впорядкований вектор ймовірностей, індекси яких відповідають певним командам. Отже, ми маємо прогнозований список команд:

Шахтар - Динамо - Зоря - Ворскла - ФК Маріуполь - Верес - ФК Карпати - ФК Олександрія - Олімпік - Зірка - Сталь - Чорноморець.

Можна використати реальний рейтинг команд наприкінці сезону 2017-2018, оскільки саме для цього сезону, ми розглядаємо на етапі заповнення транзитивної матриці. Отже, дійсний рейтинг команд сезону 17-18 має вигляд:

Шахтар - Динамо - Ворскла - Зоря - ФК Маріуполь - Верес - ФК Олександрія - ФК Карпати - Олімпік - Зірка - Чорноморець - Сталь.

Як можна спостерігати, таблиці приблизно сходяться, отже результат був досягнутий. Варто зауважити, що результат логістичної регресії можливо покращити, беручи до уваги фактори, які ми не враховували, а також

збільшуючи датасет на етапі навчання - значення похибки прогнозовано має зменшуватися.

Система побудована на мові програмування Python. Були задіяні окремі пакети мови програмування: бібліотека sklearn, яка дозволяє користуватися готовим інструментом - логістичною регресією; для складних обрахунку були використані бібліотеки numpy та pandas; для зображення результатів у вигляді гістограм та графіків використовувався пакет matplotlib. Таким чином, працюючи над роботою, були вдосконалені навички роботи із вище описаними технологіями.

Список використаних джерел

1. Modelling Football as a Markov Process – Стокгольм: DEGREE PROJECT, IN APPLIED MATHEMATICS AND INDUSTRIAL ECONOMICS, 2015. – 64 с.
2. Transition Diagram [Електронний ресурс] // Javatpoint. – 2018. – Режим доступу до ресурсу: https://www.javatpoint.com/transition-diagram.
3. COURSE NOTES STATS 325 Stochastic Processes – Auckland: Department of Statistics University of Auckland, 2014. – 195 с.
4. Introduction to Stationary Distributions [Електронний ресурс] – Режим доступу до ресурсу: https://mast.queensu.ca/~stat455/lecturenotes/set3.pdf.
5. Computing Stationary Distributions of a Discrete Markov Chain [Електронний ресурс] // Stephens999. – 2016. – Режим доступу до ресурсу: https://stephens999.github.io/fiveMinuteStats/markov_chains_discrete_stationary_dist.html#rate_of_approach_to_the_stationary_distribution.
6. What is Regression Analysis and Why Should I Use It? [Електронний ресурс] // Surveygizmo. – 2018. – Режим доступу до ресурсу: https://www.surveygizmo.com/resources/blog/regression-analysis/.
7. Logistic_regression [Електронний ресурс] // Wikipedia. – 2020. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Logistic_regression.
8. Sigmoid function [Електронний ресурс] // Wikipedia. – 2020. – Режим доступу до ресурсу: https://en.wikipedia.org/wiki/Sigmoid_function.
9. Kvam P. A Logistic Regression/Markov Chain Model For NCAA Basketball / Paul Kvam. // Naval Research Logistics. – 2006. – №53. – С. 1–23.

10. Callaghan T. Random Walker Ranking for NCAA Division I-A Football [Електронний ресурс] / T. Callaghan, M. A. Porter, P. J. Mucha // American Mathematical Monthly. – 2003. – Режим доступу до ресурсу: https://arxiv.org/abs/physics/0310148.
11. Kaplan E. H. March Madness and the Office Pool / E. H. Kaplan, S. J. Garstka. // Management Science. – 2001. – №47. – С. 369–382.
12. Carlin B. Improved NCAA Basketball Tournament Modeling Via Point Spread and Team Strength Information / B. Carlin. // 1996. – 1996. – №50. – С. 39–43.