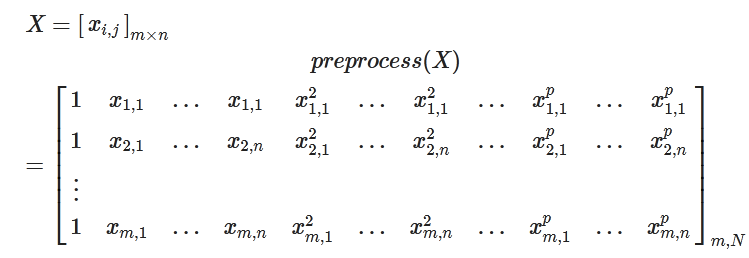
**Лабораторна робота 2 - Лінійна та поліноміальна регресія.**

**Завдання:** одне з безлічі завдань, яким займається сучасна фізика - це пошук матеріалу для виготовлення надпровідника, що працює за кімнатної температури. Крім теоретичних методів є і підхід з боку статистики, який передбачає аналіз бази даних матеріалів для знаходження залежності критичної температури від інших фізичних характеристик. Саме цим Ви і займетеся.

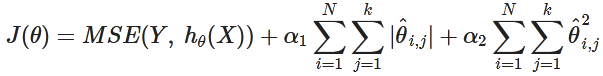
У файлі data.csv міститься весь датасет. Разом маємо 21 тисячу рядків і 169 колонок, з яких перші 167 - ознаки, колонка critical\_temp містить величину, яку треба передбачити. Колонка material - містить хімічну формулу матеріалу, її можна відкинути.

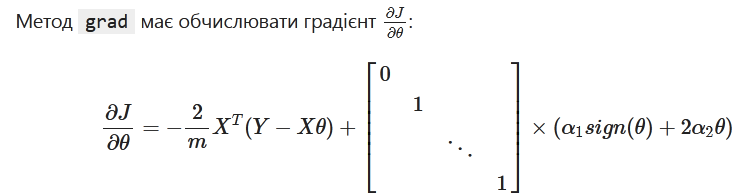
Реалізуйте методи з позначкою #TODO класу PolynomialRegression:

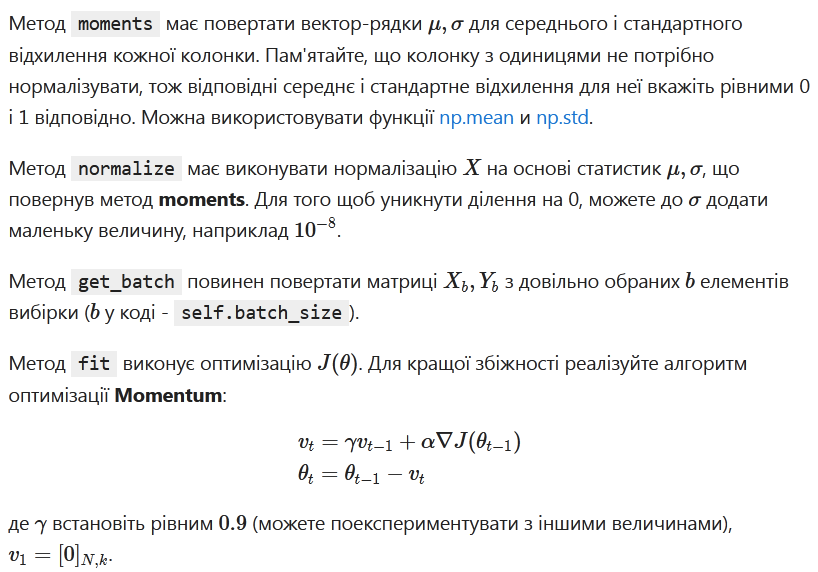
Метод preprocess повинен виконувати таке перетворення:



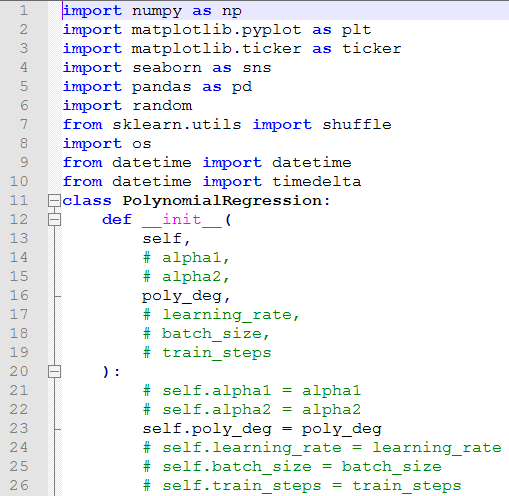
де p - ступінь полінома (self.poly\_deg у коді). Таким чином, preprocess додає поліноміальні ознаки до X. Метод J має обчислювати оціночну функцію регресії:

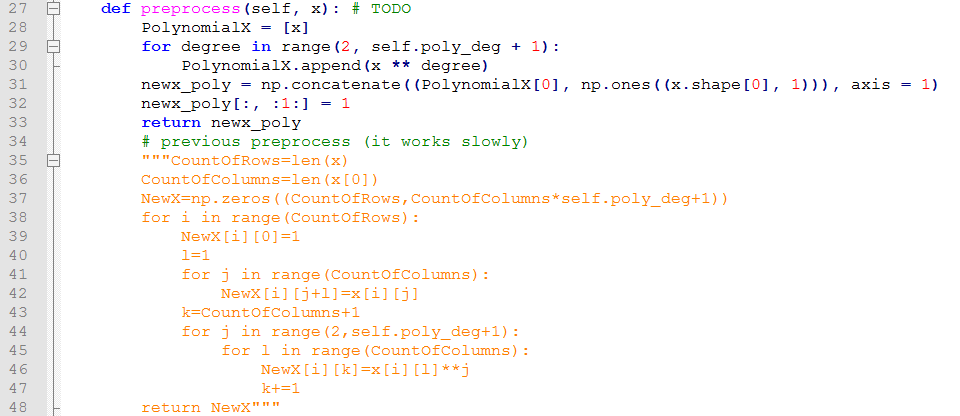


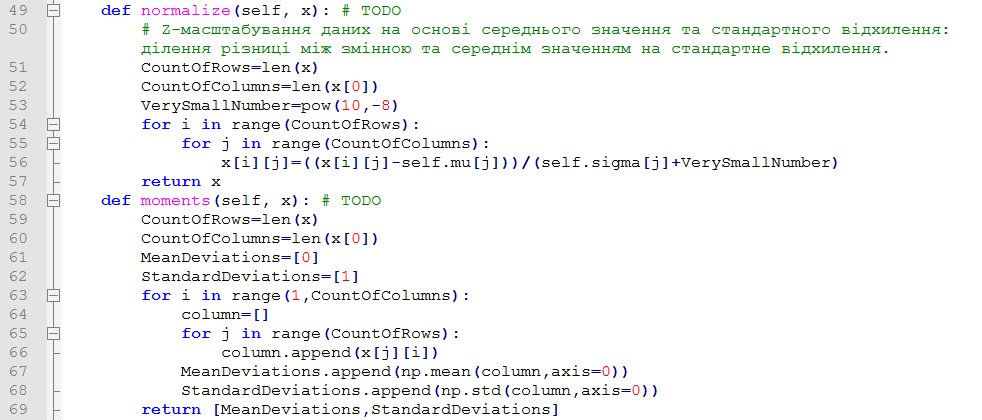


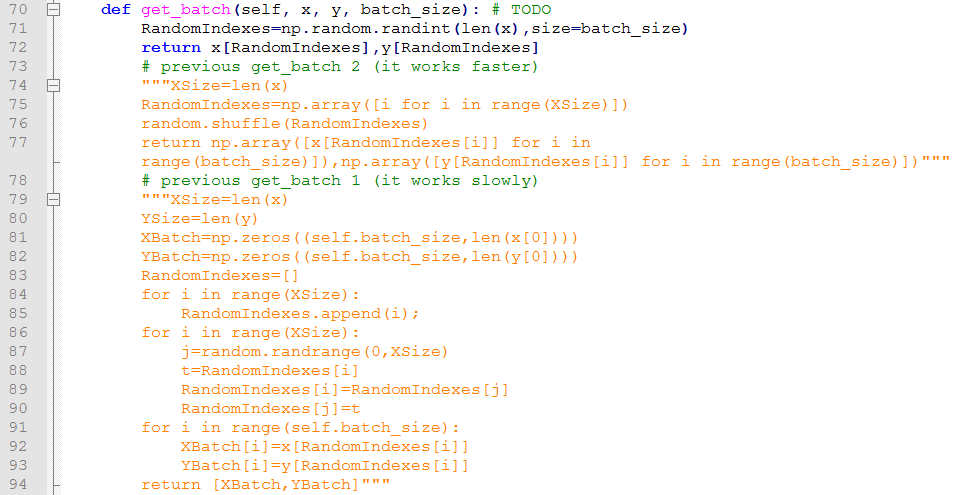


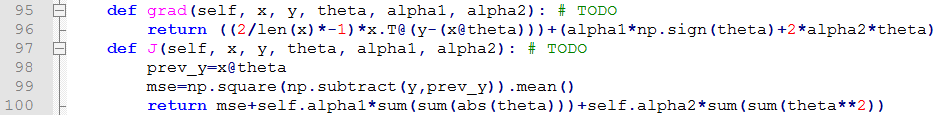
**Код класу PolynomialRegression:**

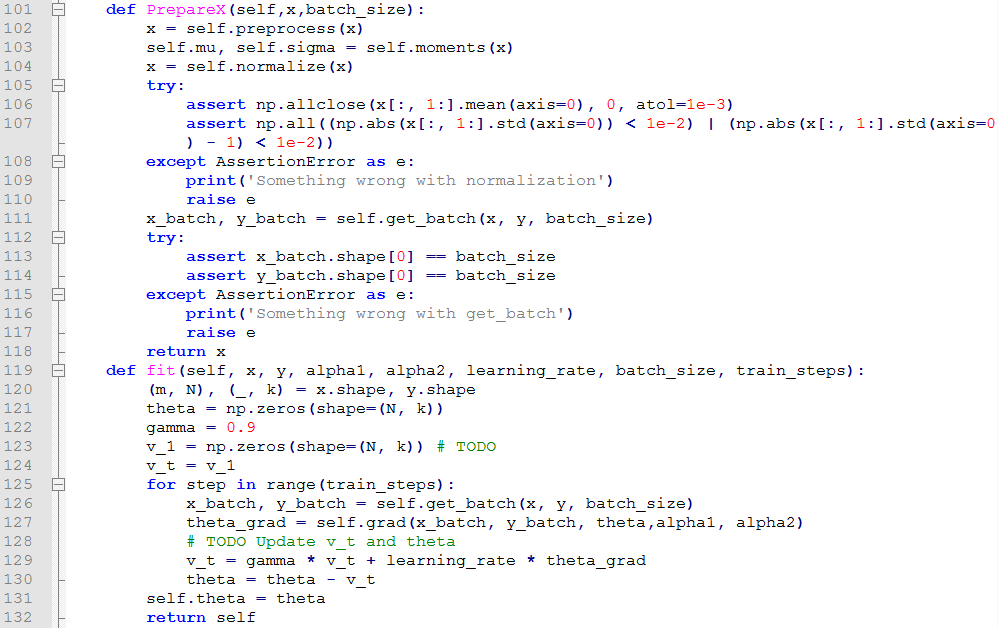


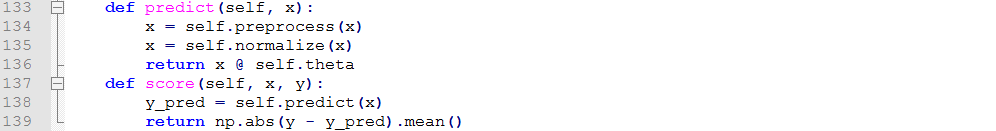












**Опис проведених досліджень**

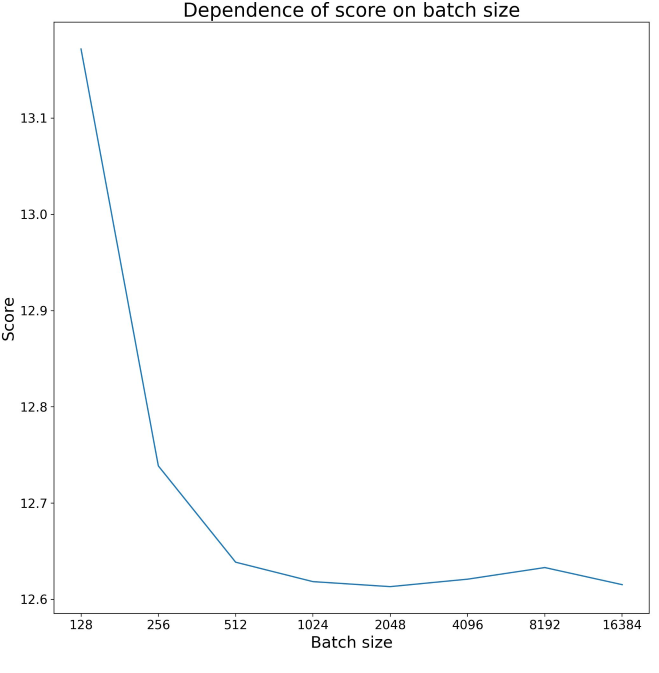
Запустивши код із стандартними параметрами ми отримали наступний вивід:



У рамках даної лабораторної роботи ми також виконали численні експерименти з різними значеннями таких основних параметрів моделі як ступінь поліноміальності (poly\_deg), коефіцієнти alpha1 та alpha2, коефіцієнт швидкості навчання (learning rate), розмір частини вибірки (batch\_size), кількість кроків навчання (train\_steps). Вони підбирались так, щоб побачити межу між недостатнім навчанням та перенавчанням.

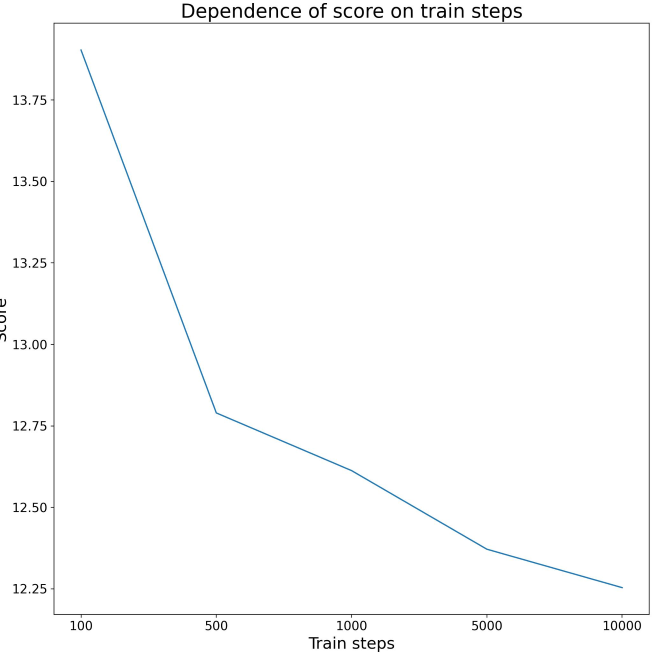
Ми створили декілька версій методів preprocess та get\_batch, задіявши під час реальних випробувань їхні найбільш швидкодіючі версії. Також слід відмітити зміну конструктора та виокремлення деяких підготовчих інструкцій у процедуру PrepareX з метою оптимізації часових витрат на попередню підготовку даних, яка виконується лише один раз перед певним набором експериментів. Окрім тестів з різними параметрами програма також містить код для визначення часових витрат, фіксації результатів у вигляді графіків, діаграм різних типів та текстового csv-файлу, який при необхідності може бути проаналізований табличним процесором.

Тестування створеного рішення проводилось на ПК з процесором  
Intel Core i5-6600 без використання окремого GPU та хмарних сервісів. Тому деякі набори експериментів виконувались у послідовності за спаданням обчислювальної складності, використовуючи в кожній наступній серії експериментів оптимальні параметри з попередніх. Спочатку ми проаналізували залежність score від різних розмірів частини вибірки (batch\_size):

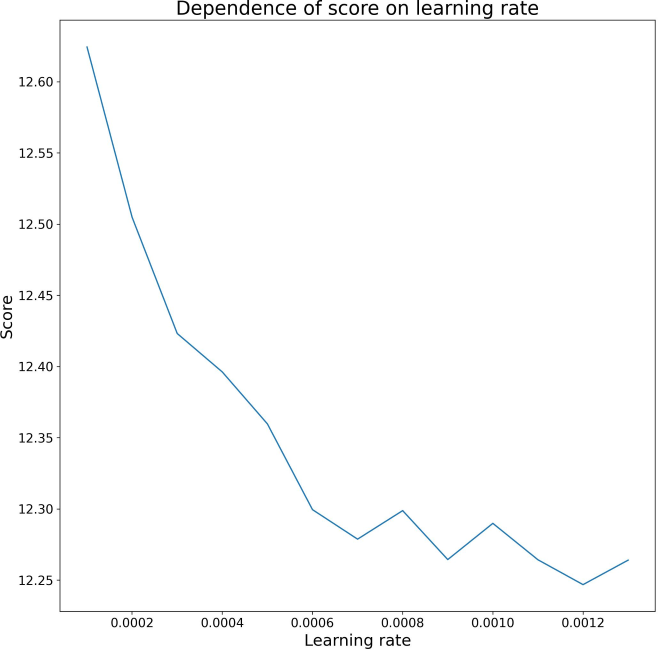


Можна зробити висновок, що найкращим значенням є 2048, проте при подальших експериментах, можливо, знайдеться більш кращий розмір.

Після цього, вже з використанням кращого batch\_size, був проаналізований параметр train\_steps (кількість кроків навчання). Звісно, чим він більший, тим точнішою є модель, проте визначення можливої межі початку перенавчання потребує значних обчислювальних потужностей або великих часових витрат:



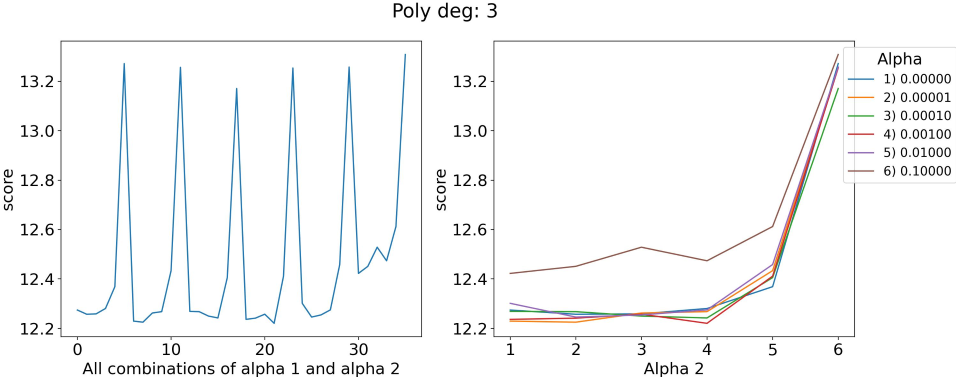
Після цього були досліджені різні значення learning\_rate (коефіцієнт швидкості навчання):

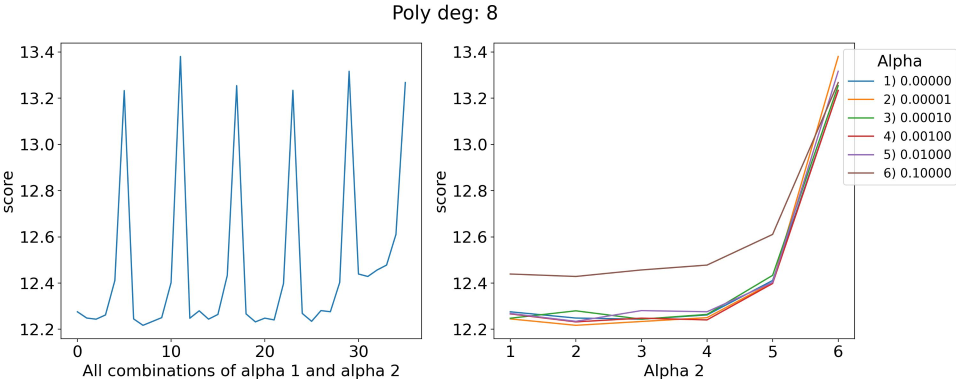


Найкращим значенням із досліджених виявилось 0.0012. Можна стверджувати, що його збільшення призводить до покращення моделі, проте визначення межі, де починається стрімке перенавчання, потребує більшої кількості додаткових експериментів.

Після цього експериментальним шляхом було встановлено, що коефіцієнт поліноміальності (poly\_deg) більший ніж 6 призводить до перенавчання, тобто до збільшення score. Ми намагались підібрати його у різних комбінаціях разом із параметрами alpha1 та alpha2, отримавши схожі один на одного графіки та діаграми:

|  |  |
| --- | --- |
|  |  |





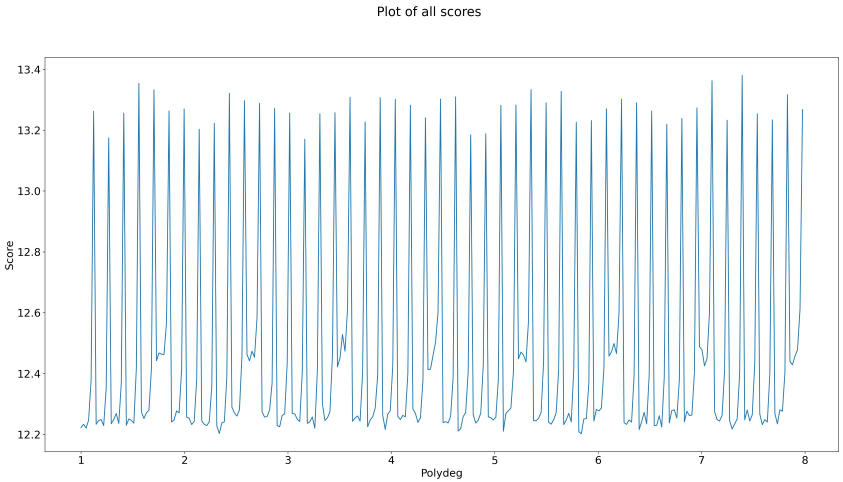
|  |  |
| --- | --- |
|  |  |

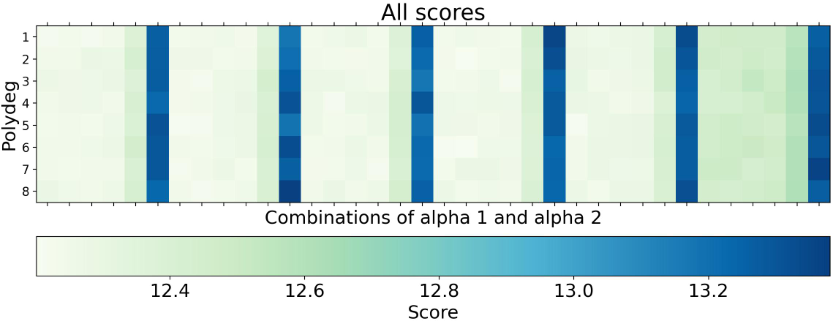
Можна зробити висновок, що оптимальні значення alpha1 та alpha1 знаходяться в межах до 0.01, вище – стрімке зменшення якості моделі. Найкращим alpha1 виявилось 0.001, alpha2 – 0.00001, poly\_deg – 6.

Найгіршим значенням score є 13.90356, отримане при таких параметрах:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| poly\_deg | alpha1 | alpha2 | learning\_rate | batch\_size | train\_steps |
| 1 | 0.00000 | 0.00000 | 0.001 | 128 | 100 |

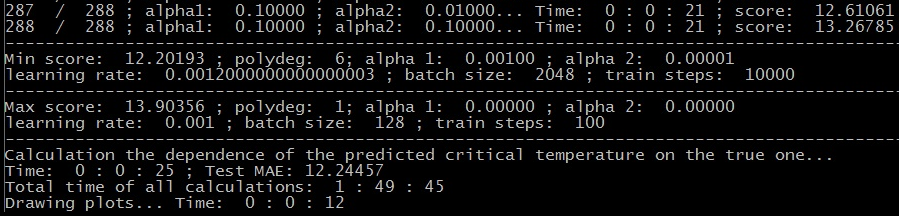
Наведемо глобальний графік та діаграму типу matshow для всіх комбінацій коефіцієнтів (результат пошуку параметрів alpha1 та alpha2 разом з poly\_deg):





Можна зробити висновок, що коефіцієнти alpha1 та alpha2, які використовуються при обчисленні градієнту, є досить впливовими параметрами.

Після завершення роботи програми ми отримуємо наступний вивід на екран:



Візуалізація залежності передбаченої критичної температури від істинної (обчислення виконані з параметрами кращої моделі):

