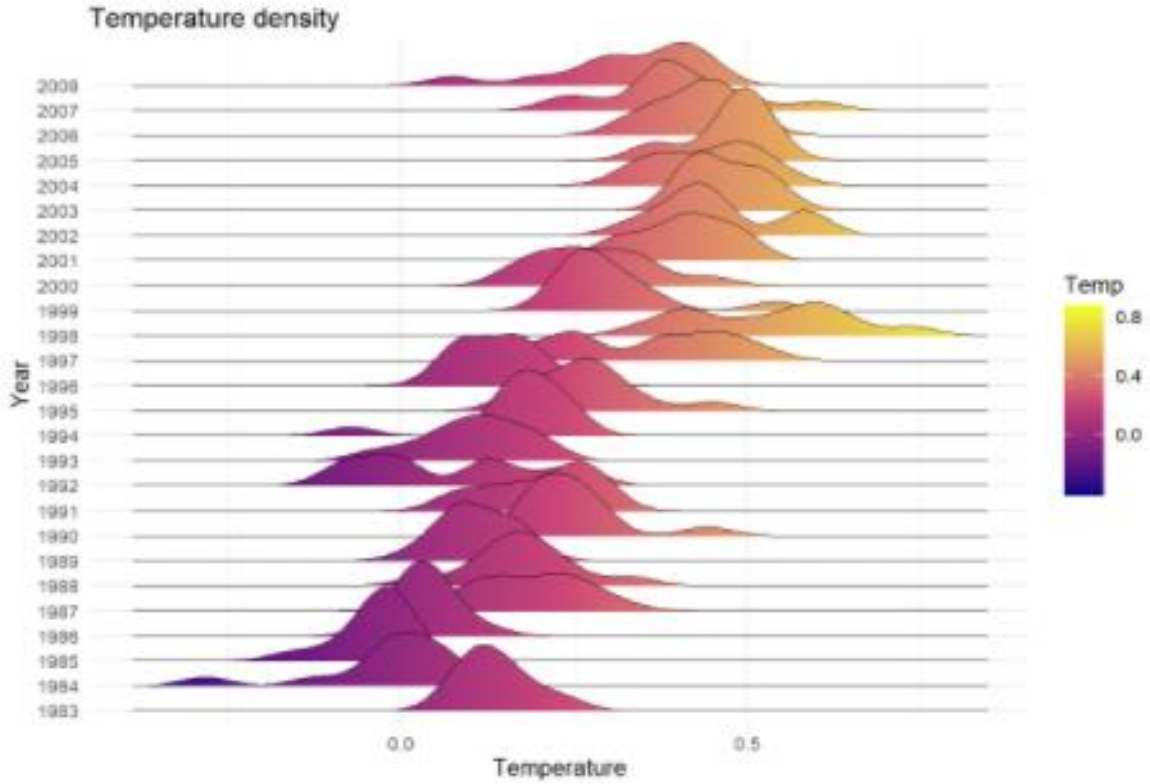

Рисунок 4 – Розподіл температурної щільності

Висновок: Отже використання аналізу даних дає можливість візуалізувати дані та створити діаграми для більш наглядного зберігання даних. Даний дата-сет дає нам можливість створення статистики і розуміння того як температурний індекс та інші речовини в атмосфері планети чинять зміни в кліматі нашої планети.

Список використаної літератури

1. «Climate» статистика – [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vageeshabudanur/riseintemp-dataset

2. Gareth J., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: With Applications in R. - Edition: 1st ed. 2013 - 426 p.

УДК 004.82:004:85

Войтко Б. С., *студент 1 курсу спеціальності 122 «Комп'ютерні науки», СО Магістр*
Нескородєва Т. В., *канд. техн. наук, доцент, завідувач кафедри комп'ютерних наук та інформаційних технологій*

ПОБУДОВА МОДЕЛЕЙ ПЕРЕДБАЧЕННЯ ЗРОСТАННЯ ЦІНИ КОМПЮТЕРІВ НА ОСНОВІ ДАТАСЕТУ BASIC COMPUTER DATA

Донецький національний університет імені Василя Стуса, м. Вінниця, Україна

Аналіз даних про ринок персональних комп'ютерів, дозволяє ознайомитись з тим, які пропозиції на ньому переважають, та зробити висновок які фактори впливають на формування ціни та оцінити точність передбачення їх можливого зростання.

Актуальність полягає в тому, щоб на основі багатьох факторів виявити закономірності та розробити моделі для оцінки точності передбачення ціни комп'ютерів.

Розглянемо приклад такого аналізу на основі даних датасету BCD. Ми використовували набір даних BCD (Basic Computer Data)[1], що містить основні комп'ютерні дані. У датасеті наявно 6259 записів та 10 показників. Для побудови моделей використовувалися наступні показники:

- price – вартість комп'ютера (представлена у доларах США $).
- speed – тактова частота в МГц .
- hd – розмір жорсткого диска в МБ.
- ram – розмір оперативної пам'яті в МБ.
- screen – розмір екрану в дюймах.
- cd – чи містить cd привід чи ні.
- premium – чи входить даний комп'ютер у преміум сегмент.
- multi – чи є процесор багатоядерним.
- ads – чи наявна реклама.
- trend - якому трендові відповідає.

Для побудови передбачення було використано 5 моделей, а саме: Модель найменших квадратів [2]; Гребнева модель; Лассо-модель; PCR модель; PLS модель. Під час дослідження ми дійшли до висновку, що найменша помилка перехресної перевірки спостерігається при використанні лише 6 компонентів. Дослідження відбувалися у середовищі обчислень R.

Для кращого розуміння приводиться приклад результатів функції lm.fit, що використовується при знаходженні оцінки метода найменших квадратів.

```
> summary(lm.fit)

Call:
lm(formula = price ~ ., data = PC[train, ])

Residuals:
    Min      1Q  Median      3Q     Max
-920.12 -188.37  -25.96  131.40 1944.60

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -195.64590   93.77845  -2.086    0.037 *
speed          9.09022    0.29157  31.176   <2e-16 ***
hd             0.74547    0.04418  16.874   <2e-16 ***
ram           46.62060    1.65530  28.164   <2e-16 ***
screen       122.37943    6.39212  19.145   <2e-16 ***
ads            0.95064    0.07930  11.988   <2e-16 ***
trend        -47.61666    0.95263 -49.984   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 304.6 on 3122 degrees of freedom
Multiple R-squared:  0.7238,    Adjusted R-squared:  0.7233
F-statistic:  1364 on 6 and 3122 DF,  p-value: < 2.2e-16
```

Рисунок 1 – Результат використання функції lm.fit

Середньоквадратична помилка побудованих моделей наведена на графіку та нижче:

Модель найменших квадратів - 101709.2

Гребнева модель - 4873275.1

Лассо-модель - 4873275.1

PCR модель - 4873721.3

PLS модель – 4873721.3

Рисунок 2 - Середньоквадратична помилка побудованих моделей

З отриманих результатів робимо висновок, що найкращою моделлю для даного набору даних є модель найменших квадратів. Найгіршими є PCR та PLS моделі.

Для знаходження точності передбачення зростання ціни комп'ютерів, використовуємо модель найменших квадратів. Для знаходження оцінки точності використовуємо наступну формулу: $1 - \frac{|MeanError|}{\bar{y}_{test}}$. Дана формула у середовищі R із потрібними змінними та результатом показана на рисунку 3. Отже точність передбачення ціни дорівнює 70%.

```
>  avg_price=mean(PC[,"price"])
>  1 - mean((PC[test, "price"] - lm.pred)^2) /mean((PC[test, "price"] - avg_price)^2)
[1] 0.7004061
```

Рисунок 3 – Знаходження оцінки точності передбачення ціни

Висновок. Отже, завдяки використання Методу найменших квадратів, Гребневої моделі, Лассо, PCR та PLS моделей ми можемо визначити фактори, які впливають на ціну при її формуванні та знайти оцінку точності передбачення ціни комп'ютерів.

Список використаної літератури

1. «Basic Computer Data» статистика – [Електронний ресурс]. Режим доступу: https://www.kaggle.com/kingburrito666/basic-computer-data-set

2. Джеймс Г., Уиттон Д., Хасти Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С. Э. Мастицкого - М.: ДМК Пресс, 2017. - 456 с.: ил.