КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

імені ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики

Звіт

з дисципліни «Аналіз даних»

про виконання завдання

«Побудова математичної моделі методами аналізу даних»

Виконав:

студент групи К-28

Миронов Данiiл Олександрович

Київ-2022

**Зміст**

[Опис вхідної інформації 3](#_Toc103857352)

[Постановка задачі 5](#_Toc103857353)

[Передмова до аналізу 6](#_Toc103857354)

[Клас апроксимуючих параметричних функцій 7](#_Toc103857355)

[Дослідження наявності множинного зв’язку 7](#_Toc103857356)

[Аналіз моделі 7](#_Toc103857357)

[Перебудова моделі 8](#_Toc103857358)

[Структура моделі 9](#_Toc103857359)

[Оцінка якості отриманої математичної моделі 10](#_Toc103857360)

[Виявлені недоліки та шляхи покращення моделі 13](#_Toc103857361)

[Підсумок по лабораторній роботі 14](#_Toc103857362)

[Список використаних джерел 15](#_Toc103857363)

Опис вхідної інформації

**Основнi данi**

Для аналiзу вибрано набiр даних «TikTok Trending Tracks».

Посилання на джерело даних: <https://www.kaggle.com/yamqwe/tiktok-trending-tracks>

Набір даних містить популярні треки, представлені в сервiсi TikTok. Цей набір даних містить близько 7000 треків разом із даними про них, оцінкою характеристик, таких як гучнiсть, енергiчнiсть, iншi данi та технічнe інформацію.

Набiр був створений командою Team Dan. Посилання на автора набору: <https://github.com/romeoben/DSC7-Sprint2-TeamDan>

**Опис змiнних**

У наборі даних присутні 23 змінні:

|  |  |
| --- | --- |
| **Назва змінної** | **Опис** |
| track\_id | ID треку в системі TikTok |
| track\_name | Назва треку |
| artist\_id | ID виконавця в системі TikTok |
| artist\_name | Ім’я виконавця |
| album\_id | ID альбому в системі TikTok |
| duration | Тривалість треку в мілісекундах |
| release\_date | Дата виходу треку |
| popularity | Популярність треку – ціле число на шкалі від 0 до 100 включно |
| danceability | Танцювальність треку – оцінка від 0 до 1 |
| energy | Енергічність треку – оцінка від 0 до 1 |
| key | Ключ виконання треку |
| loudness | Гучність треку |
| mode | Режим треку |
| speechiness | Мовленнєвість треку – оцінка від 0 до 1 |
| acousticness | Акустичність треку – оцінка від 0 до 1 |
| instrumentalness | Інструментальність треку – оцінка від 0 до 1 |
| liveness | Живучість треку – оцінка від 0 до 1 |
| valence | Музична позитивність, яку передає трек – оцінка від 0 до 1 |
| tempo | Темп треку – удари в хвилину |
| playlist\_id | ID плейлисту в системі TikTok |
| playlist\_name | Назва плейлисту |
| duration\_mins | Тривалість треку в хвилинах |
| genre | Жанр треку |

Постановка задачі

Потрібно провести побудову математичної моделі методами аналізу даних для обраних скалярних змінних.

Залежна змінна – **energy**.

Незалежні змінні: **popularity**, **loudness**, **tempo**, **popularity**, **danceability**, **acousticness**, **speechiness**, **valence**, **duration**.

Для цього, максимально використовуючи результати отримані у попередніх лабораторних роботах:

● визначитися з класом апроксимуючих параметричних функцій для правої частини моделі,

● обчислити точечні та множинні оцінки невідомих параметрів моделі та їх характеристики, які вважаєте потрібними,

● уточнити структуру Вашої математичної моделі, з обгрунтуванням,

● з'ясувати якість отриманої математичної моделі, чисельний та графічний супровід рекомендується,

● по отриманій математичній моделі сформулювати:

висновки, вказати на виявлені її недоліки та шляхи її покращення.

Передмова до аналізу

Цей звіт, виконувані файли з розширенням .R, результати виконання та інші лабораторні роботи з поточним набором даних доступні за посиланням: <https://github.com/leirimnad/data-analysis-labs>

Перед виконанням коду з наступних розділів, матимемо на увазі, що перед кожним з фрагментів виконується наступний код:

options(scipen=5)

songs <- read.csv(file = 'tiktok.csv')

Фрагмент коду options(scipen=5) використовується для того, щоб числові дані виводилися не в експоненціальному записі.

Фрагмент songs <- read.csv(file = 'tiktok.csv') обробляє .csv-файл з даними та зберігає їх у змінну songs.

Клас апроксимуючих параметричних функцій

Для побудови мат. моделі зв’язків між виключно кількісними змінними використаємо регресійний аналіз. В якості класу апроксимуючих функцій візьмемо клас функцій лінійних по вектору невідомих параметрів.

Дослідження наявності множинного зв’язку

Побудуємо лінійну модель за допомогою функції lm(…) та віднайдемо рівень значущості.

|  |
| --- |
| model <- lm(energy ~  popularity + loudness + tempo + popularity + danceability  + acousticness + speechiness + valence + duration,  data = songs)  cor.test(model$model$`energy`, model$fitted.values) |
|  |

За результатами, бачимо, що p-value малий, отже існує вагома залежність змінної energy від незалежних змінних.

Аналіз моделі

За допомогою функції summary(…) виведемо основні підсумки по побудованій моделі:

|  |
| --- |
| summary(model) |
|  |

У колонці Pr(>|t|) відображено рівень значущості при перевірці гіпотези .

У випадку зі змінною **duration** від досягає значення 0.55, тому приймається гіпотеза , і можемо відкинути змінну duration з моделі.

Перебудова моделі

Відкинемо змінну **duration**, перебудуємо модель та виведемо підсумки по ній:

|  |
| --- |
| model <- lm(energy ~  popularity + loudness + tempo + popularity + danceability  + acousticness + speechiness + valence,  data = songs)  summary(model) |
|  |

Структура моделі

В підсумках моделі в колонці «Estimate» отримали оцінку вектора невідомих параметрів. Отже, отриману модель можна записати наступним чином:

Оцінка якості отриманої математичної моделі

* По-перше, для кожної з незалежних змінних гіпотезу відкидаємо через отримані значення рівня значущості (Pr(>|t|) < 0.05). Всі незалежні змінні, використані в моделі, мають значний вплив на energy.
* Для вектора невідомих коєфіцієнтів можемо отримати довірчий інтервал (95%):

|  |
| --- |
| confint(model) |
|  |

* Середня варіація значень від лінії регресія наступна:  
  Residual standard error: 0.1083 on 6738 degrees of freedom

Поділимо це значення на вибіркове середнє значення energy і отримаємо відсоток помилки прогнозу:

|  |
| --- |
| sigma(model)\*100/mean(songs$energy) |
|  |

Відсоток помилки: 17.3456%

* Значення Adjusted R-squared (коригований коефіцієнт детермінації) дорівнює 0.6144, є достатньо великим, і показує, що велика частина результату пояснюється регресійною моделлю.
* F-статистика є великою, а p-value дуже малим, що показує, що модель є значущою.
* Перевіримо залишки на нормальний розподіл за допомогою побудови нормального Q-Q графіку:

|  |
| --- |
| qqnorm(model$residuals)  qqline(model$residuals, col="red") |
|  |

Дійсно, залишки мають нормальний розподіл.

* Побудуємо графік, що покаже незалежність між залишками та передбаченими регресією значеннями, а також чисельно підрахуємо, чи є між ними залежність:

|  |
| --- |
| plot(model$fitted.values, model$residuals, xlab="predicted", ylab="residuals") |
|  |
| summary(lm(model$fitted.values ~ model$residuals)) |
|  |

За графіком та значенням p-value можемо прийти до висновку, що залежність відсутня.

Виявлені недоліки та шляхи покращення моделі

Хоча у моделі і досить високий коефіцієнт детермінації, однак, можливо, він міг би бути вищим. З іншої сторони, у моделі були використані майже всі доступні змінні датасету, що ставить під сумнів можливість покращення моделі.

Підсумок по лабораторній роботі

Побудовано математичну модель для наступних змінних:

Залежна змінна – energy. Незалежні змінні: popularity, loudness, tempo, popularity, danceability, acousticness, speechiness, valence, duration.

Засобами мови R отримано структуру моделі, основні її характеристики. Було покращено модель шляхом виключення зміннох без значного впливу за залежну змінну. З’ясовано якість математичної моделі з числовим та графічним супроводом.

Список використаних джерел

1. <https://en.wikipedia.org/wiki/Regression_analysis>
2. <http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/>
3. <http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r>
4. <https://uk.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BD%D0%B0%D0%B9%D0%BC%D0%B5%D0%BD%D1%88%D0%B8%D1%85_%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D1%96%D0%B2>
5. <https://www.cyclismo.org/tutorial/R/linearLeastSquares.html>
6. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm>