

Для моделі з 3 значущими предикатами

Multiple R-squared: 0.9043, Adjusted R-squared: 0.9012

Виходячи з цих регресій, можна сказати, що вони однаково відповідають даним, при цьому лінійна регресія, яка складається тільки зі значущих предикатів трохи краще підходить для опису даних. Про це свідчить менша стандартна помилка та кількість рівнів свободи.

Висновок. Отже використання моделей даних дає можливість виявити саме ті фактори, які найбільшим чином впливають на рейтинг автомобіля, а також на ціну. Саме тому, використання аналізу даних може допомогти підібрати саме той автомобіль, який займає лідируючі місця в рейтингах, виходячи тільки з його ключових характеристик.

Список літературних джерел

1. Ярош О.Л., Нескородева Т.В. Аналіз даних про ринок житла в Україні засобами мови R Матеріали всеукраїнської науково-практичної конференції для студентів, аспірантів та молодих вчених "Прикладні інформаційні технології" (29 квітня 2020 року) - Вінниця: ДонНУ імені Василя Стуса. С.74-76.
2. Новицький М.О., Нескородева Т. В. Аналіз даних про рівень щастя населення в країнах світу. Матеріали всеукраїнської науково-практичної конференції для студентів, аспірантів та молодих вчених "Прикладні інформаційні технології" (29 квітня 2020 року) - Вінниця: ДонНУ імені Василя Стуса. С.43-45.
3. «Cars» статистика – [Електронний ресурс]. Режим доступу: <http://kaggle/data/cars.csv>
4. Джеймс Г., Уиттон Д., Хасті Т., Тибширани Р. Введение в статистическое обучение с примерами на языке R. Пер. с англ. С. Э. Мاستицкого - М.: ДМК Пресс, 2017. - 456 с.: ил.

УДК 004.82:004:85

*Данілевич Б. С., студент 1 курсу спеціальності 122 «Комп'ютерні науки»
Федоров Є.Є., докт. техн. наук, доцент
кафедри комп'ютерних наук та
інформаційних технологій*

ІНДЕКСАЦІЯ І ТИПИ ІНДЕКСІВ В БІБЛІОТЕЦІ PANDAS

Донецький національний університет імені Василя Стуса, м. Вінниця

Індекс – це інструмент, призначений для оптимізації пошуку значень в об’єкті Series або в об’єкті DataFrame. Він дуже схожий на ключ реляційної бази даних, але дає більше можливостей. Індекс дозволяє виконати вирівнювання для декількох наборів даних, а також зрозуміти, як виконати різні операції з даними.

Важливість застосування індексів

Індекси бібліотеки pandas дозволяють ефективно знаходити значення. Якщо індекс відсутній, потрібно виконати лінійний пошук по всіх наших даних. Індекси створюють оптимізаційні ярлики для конкретних значень, використовуючи малий пошук.

Типи індексів бібліотеки Pandas

Бібліотека pandas пропонує масу встроєних індексів. Кожен з цих типів індекса призначений для оптимізаційного пошуку на основі певного типу даних чи структури даних.

Основний тип Index

Цей тип індекса є найбільш загальним і представляє з себе впорядкований набір значень, на основі якого можливо створити зрізи. Значення, які він приймає, повинні бути хешуєними пайтоновими об’єктами. Це обумовлено тим, що індекс буде використовувати хеш для ефективного пошуку елементів, зв’язаних з значеннями цього об’єкту. Хоча пошук хешу переважніше, ім лінійний пошук, існують і інші типи індексів, які можна додатково оптимізувати.

Зазвичай таким типом являється індекс стовпців. Наступний програмний код демонструє використання цього типу індексу в якості стовпців датафрейму.

```
#покажем, що стовпці фактично є індексом
temps = pd.DataFrame( {"City": ["Missoula", "Philadelphia"],
                        "Temperature": [70, 80] })
```

```
temps
```

```
Out
```

	City	Temperature
0	Missoula	70
1	Philadelphia	80

```
#ми бачим, що стовпці – це індекс
```

```
temps.columns
```

```
Out:
```

```
Index(["City", "Temperature"], dtype= "object")
```

Хоча цей тип індекса зазвичай добре працює для буквенно-цифрових імен стовпців, при бажанні в якості індекса стовпців можна використовувати індекси других типів.

Індекси Int64Index і RangeIndex використовують цілі числа

Int64Index представляє з себе незмінний масив, що складається з 64-бітових цілих чисел, які зіставляються значенням. До недавнього часу цей тип індекса використовувався в бібліотеці pandas за замовчуванням в тих випадках,

коли індекс не був вказаний чи були використані цілі числа, як показано в наступному фрагменті кода:

```
#створюємо Int64Index
df_i64 = dp.DataFrame(np.arange(10, 20), index=np.arange(0, 10))
df_i64[:5]
Out:
0
      0      10
      1      11
      2      12
      3      13
      4      14
```

При використанні цього типу індекса пошук строк в об'єкті DataFrame стає дуже ефективним, оскільки він виконується за допомогою неперервного масиву, оброблюваного в оперативній пам'яті.

IntervalIndex

Окремі інтервали міток можна представити за допомогою IntervalIndex. Інтервал закривається або зліва, або справа, тобто відповідний кінець включається в даний інтервал. Згідно прийнятої в нотації інтервалів, кругла скобка означає, що відповідний кінець не включається (відкритий), а квадратна – що включається (закритий).

Індексація по даті і часу за допомогою DatetimeIndex

DatetimeIndex використовується для представлення набору дати і часу. Він взагалі використовується в даних, представлених в виді часових рядів, коли виміри для нагляду беруться через певні проміжки часу.

IntervalIndex

Окремі інтервали міток можна представити за допомогою IntervalIndex. Інтервал закривається або зліва, або справа, тобто відповідний кінець включається в даний інтервал. Згідно прийнятої в нотації інтервалів, кругла скобка означає, що відповідний кінець не включається (відкритий), а квадратна – що включається (закритий).

Індексація по даті і часу за допомогою DatetimeIndex

DatetimeIndex використовується для представлення набору дати і часу. Він взагалі використовується в даних, представлених в виді часових рядів, коли виміри для нагляду беруться через певні проміжки часу.

Індексація періодів часу за допомогою PeriodIndex

Крім того, іноді потрібно представити в виді міток індекса періоди часу типу дня, місяця чи року. Це дуже схоже на роботу з інтервалами, але тепер мова йде про період часу. Данні сценарії можна змодельовати, використовуючи PeriodIndex і задавши конкретну частоту періодів в індексі.

Підводячи підсумок можна сказати, що більша частина процесу підготовки даних, виконується за допомогою бібліотеки Pandas. Правильно створений

індекс покращить продуктивність і стане важливим інструментом для виконання аналізу.

Список літературних джерел

1. Хейдт М. Изучаем Pandas/ Пер. с англ. А.В. Груздева. –М.: ДМК Пресс, 2018. – 438 с.
2. Індексція Пандас [Електронний ресурс] – Режим доступу до ресурсу:https://pandas-docs.github.io/pandas-docs-travis/user_guide/indexing.html.