УДК 004.82:004:85

Зінькевич Д.О., студент 1 курсу спеціальності 122 «Комп'ютерні науки»
Римар П.В., старший викладач кафедри комп'ютерних наук та інформаційних технологій

АНАЗІЛ ДАННИХ РИНКУ АВТОМОБІЛІВ В США

Донецький національний університет імені Василя Стуса, м. Вінниця

Аналізуючи ринок автомобілів, можна ознайомитись з тим, які пропозиції на ньому переважають, яка комплектація автомобілів підійде для тієї чи іншої країни, а також технічні характеристики, які впливають на ціну автомобіля. Актуальність даного аналізу полягає в тому, що з кожним роком моделей авто стає багато, але тепер на перший план виходить пошук закономірностей, завдяки яким можна буде проаналізувати весь ринок автомобілів і підібрати авто за його ключовими показниками. Все це можна зробити, використовуючи аналіз великих наборів даних - датасетів.

Для прикладу був проаналізований датасет Cars. Даний датасет був сформований одним із користувачів сервісу Kaggle на основі даних із сайтів продавців автомобілів в США. У наборові даних знаходиться 10 предикатів та 500 записів, що містять інформацію про актуальні та популярні автомобілі в США, починаючи з 2010 по 2018 роки. Нижче наводиться опис 5 основних показників, що були відібрані для аналізу з 10 змінних. Предикати:

Rating - рейтинг автомобілів.

Price - зазначена продавцем ціна(в тисячах доларів).

Mileage - пройдена швидкість в км/год.

Mpg - кількість виртаченного бензину на 1 милю шляху.

EnginePower - потужність двигуна.

Основними об'єктами є автомобілі, що активно продавались в описаний вище період. Першим етапом була побудована множинна регресійна модель для передбачення rating на основі mileage, mpg, enginePower, price (рис. 1).

```
Residuals:
   Min     1Q     Median     3Q      Max
-54.462  -21.930  2.352      14.530    64.406
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept) 32.89943    7.10364     4.027      0.000115 ***
mileage     -0.01090    0.20504     0.053      -0.325770
mpg          0.13396    0.07033     1.905      0.009881 .
enginePower -0.857907   7.29193     -2.175     0.042161 *
price        0.75978    0.23994     3.166      0.002081 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 56.63 on 496 degrees of freedom
Multiple R-squared:  0.9043,    Adjusted R-squared:  0.9012
F-statistic:  690 on 5 and 496 DF,  p-value: < 2.2e-16
```

Рисунок 1 - Множинна регресійна модель для передбачення rating

На наступному кроці потрібно визначити ключові фактори, які впливають на рейтинг автомобіля, відкинувши фактори, які впливають менш за все. За даними регресійної моделі було записано рівняння:

rating = 32.90 + (-0.01)*mileage + 0.13*mpg + (-0.86)*enginePower + 0.76*price

Отримані значення кожного предиката можна розглядати як високозначущі окрім mileage, про що свідчить високий коефіцієнт p біля нього. Іншими словами, якщо буде змінюватися хоча б один із предикатів (окрім mileage), то і буде змінюватися рейтинг автомобілів.

Наступним кроком була побудована регресійна модель лише з тими предикатами, для яких було знайдено зв'язок з відкликом. Іншими словами, була сформована модель лише з тих предикатів, які напряму впливають на рейтинг автомобілів.

```
Residuals:
   Min      1Q      Median      3Q      Max
-54.426    -22.038    2.403      14.571    64.423

Coefficients:
             Estimate     Std. Error    t value     Pr(>|t|)
(Intercept)  32.93415     7.016003      4.316       3.89e-05 ***
mpg          0.13413      0.06988       -1.919      0.05794 .
enginePower  0.865003     7.25206       -2.186      0.03130 *
price        0.76116      0.23728       3.208       0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.61 on 496 degrees of freedom
Multiple R-squared:  0.9043,    Adjusted R-squared:  0.9012
G-statistic:  732 on 5 and 496 DF,  p-value: < 2.2e-16
```

Рисунок 2 - Регресійна модель значущих параметрів при виборі авто

Якщо порівняти отримані моделі, то можна отримати наступні показники:

Для моделі з 4 предикатами

Multiple R-squared: 0.9032,   Adjusted R-squared: 0.9012

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., Нескородєва 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., Нескородєва 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., 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.

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