Московский авиационный институт
(национальный исследовательский университет)

Отчет по летней практике

Студент: Королёв Е.В.
Руководитель: Игнатов А.Н.
Дата: 03.07.2020

Москва, 2020

Постановка задачи:

Периодически на железнодорожной дороге происходят сходы грузовых поездов. Сходы могут происходить по различным причинам. Одна из этих причин - излом боковой рамы. Нужно спрогнозировать, сколько сойдет вагонов с рельсов в зависимости от того или иного набора значений факторов с использованием линейной регрессии и метода наименьших квадратов. Оценить качество полученной модели с помощью коэффициента детерминации (R^2).

Решение:

Для начала импортируем данные:

```
data = pandas.read_excel('Практика.xlsx')
```

Для удобства обозначим названия столбцов переменными:

```
x1 = 'Количество вагонов'
x2 = 'Максимальное число вагонов в сходе'
x3 = 'Общее количество вагонов'
x4 = 'Количество сшедших вагонов (вместе с локомотивом)'
x5 = 'Скорость'
x6 = 'Вес'
x7 = 'Загрузка'
x8 = 'Наличие стрелочного перевода (съезда) в месте схода'
x9 = 'Кривизна'
x10 = 'Профиль пути(в единицах: спуск с отрицательным знаком, подъем с
положительным)'
x11 = 'Режим движения ( 1 - тяга, 2 - выбег, 3 - торможение)'
```

Введем список таргета и список параметров:

```
parameters = [x7, x9, x10]
answers = [x4]
```

Составим выборку из параметров и таргета и удалим из них те данные, где есть пропуски:

```
new_data = data[parameters + answers].dropna()  # убираем пропуски из выборки
```

Для краткости обозначим:

```
X = new_data[parameters]
Y = new_data[answers]
```

Создание и обучение модели:

```
model = linear_model.LinearRegression()
model.fit(X, Y)
```

Прогнозирование:

```
Y_pred = model.predict(X)
```

Вывод результатов:

```
print('MSE: ', mean_squared_error(Y, Y_pred))  # среднеквадратическая ошибка
print('R^2: ', r2_score(Y, Y_pred))  #доля данных, которую модель смогла объяснить
print("Coefficients: ", model.coef_[0])  # вывод коэффициентов
```

```
print("Intercept: ", model.intercept_)  # свободный коэффициент
```

MSE (Mean Square Error) - среднеквадратичная ошибка.

Построим различные модели с целью максимизации R^2.

Модель 1.

Факторы:

загрузка

кривизна

Вывод программы:

```
MSE:  40.22623295487267
R^2:  0.030632855765773903
Coefficients:  [   3.40423946 -619.0604631 ]
Intercept:  [1.5617265]
```

Модель 2.

Факторы:

кривизна

загрузка

Профиль пути (в единицах: спуск с отрицательным знаком, подъем с положительным)

Вывод программы:

```
MSE:  44.300075090514746
R^2:  0.03430483829688824
Coefficients:  [   3.61358847 -389.97255961  134.68818591]
Intercept:  [1.47548921]
```

Точность прогноза уменьшились, охват данных незначительно вырос.

Модель 3.

Факторы:

x1, x2, x3, x5, x6, x7, x8, x9, x10, x11, x12

Вывод программы:

```
MSE:  2.0352707733679414
R^2:  0.27499643695051523
Coefficients:  [ 2.08353825e-02 -8.81525521e-03  3.28281381e-03  9.50371799e-03
 -1.31939983e-03  4.46692946e+00 -7.51841156e+02  7.96499857e+00
  3.53734054e-01  4.71325758e+00]
Intercept:  [2.77602281]
```

Среднеквадратическая ошибка уменьшилась почти в 22 раза; R^2 вырос более чем в 8 раз, однако в выборке осталось только 21 строка.

Модель 4.

Факторы:

В качестве параметров возьмем параметры предыдущей модели и добавим к ним собственный - $\text{загрузка} * \text{скорость}^2$. Обозначим новый параметр как x13.

Вывод программы:

```
MSE:  1.9770755513918483
```

```
R^2:  0.29572672200015737
Coefficients:  [ 7.78783568e-02 -1.13174454e-02 -1.18711498e-01  1.07490113e-01
 -2.24180604e-05  1.93413837e+00 -8.09513922e+02  3.28155250e+01
  3.83538710e-01  8.48862304e+00 -1.06438584e-03]
Intercept:  [0.16410376]
```

MSE уменьшилось, R^2 увеличилось.

Модель 5.

Во второй модели точность прогноза снизилась из-за фактора “Профиль пути” попробуем убрать этот фактор из выборки в модели 4.

Вывод программы:

```
MSE:  1.7338856232580164
R^2:  0.2769237006030726
Coefficients:  [ 6.78776740e-02 -1.45902581e-02 -6.02187508e-02  3.15554557e-02
 -9.00869724e-04  4.29208732e+00 -7.31327840e+02  1.75064800e-01
  5.18398764e+00 -3.29963343e-04]
Intercept:  [2.11163987]
```

Как видим, точность модели возросла, однако охват данных уменьшился.

Модель 6.

В качестве Факторов возьмем все Факторы из модели 4 и добавим что-то вроде импульса: x14 = ‘Режим движения’ * ‘Вес’ * ‘Скорость’.

Вывод программы:

```
MSE:  1.400418243112245
R^2:  0.5011434206684167
Coefficients:  [-8.15310931e-01 -3.41863674e-02  6.70999772e-01  2.23206961e-01
  2.43804404e-03 -2.93598216e+00 -1.97157929e+03  2.95296995e+01
  4.22583847e+00  3.87848913e+01 -1.00077915e-03 -1.54710292e-05]
Intercept:  [-8.69962787]
```

Значительно вырос охват данных, также уменьшилась среднеквадратическая ошибка.

Модель 7.

Добавим фактор : x15 = ‘Количество вагонов’ * ‘Кривизна’ * ‘Скорость’.

Удалим часть факторов из выборки чтобы минимизировать вероятность переобучения.

$X = \{x5,\ x6,\ x7,\ x9,\ x10,\ x11,\ x13,\ x14,\ x15\}$

Вывод программы:

```
MSE:  1.7718830506540209
R^2:  0.34284184174977295
Coefficients:  [ 7.18455581e-02  6.56605936e-01 -3.70436582e+02  3.89360961e+01
  3.35959736e+00  3.04393783e-04 -1.04672724e-05  7.56689276e-02]
Intercept:  [-4.15138444]
```

Выводы:

Так как целевая переменная имеет дискретную природу, то применение метода наименьших квадратов в сочетании с линейной регрессией ожидаемо приводит к результатам с не самой высокой точностью. В дальнейшем данный недостаток будет

устранен путем использования более сложных регрессионных моделей, например геометрической регрессией.