ФГАОУ ВО «СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт космических и информационных технологий
Кафедра «Информатика»

Методы анализа данных

Практическая работа № 4

Регрессионный анализ

Красноярск, 2024

Цель: знакомство с теоретическими основами регрессионного анализа, формирование навыков применения регрессионного анализа для решения задачи восстановления функциональных зависимостей с помощью языка программирования Python.

Задачи:

Выполнение практической работы предполагает решение следующий задач:

1. Предварительная обработку исходных данных
2. Обучение базовых регрессионных моделей
3. Подбор оптимальных параметров регрессионных моделей
4. Оценка качества построенных моделей на валидационной/тестовой выборке

Ссылка на соревнование

https://www.kaggle.com/c/krasnoyarsk-flat-price-prediction

Используемые регрессионные модели:

1. Линейная регрессия (МНК)
2. Лассо регрессия
3. Гребневая регрессия
4. Elastic-Net
5. Метод наименьших углов (Least-angle regression)
6. Байесовская регрессия
7. Обобщенная линейная регрессия (обобщенный МНК)
8. Взвешенный МНК
9. Полиномиальная регрессия
10. Сплайны
11. Непараметрическая регрессия

Реализация данных регрессионных моделей представлена в двух следующих библиотеках:

1. `statsmodels`: https://www.statsmodels.org/stable/api.html
2. `scikit-learn`: https://scikit-learn.org/stable/modules/linear_model.html

Общая последовательность действий

1. Ознакомиться с описанием соревнования.
2. Загрузить данные для обучения и для теста.
3. Выполнить предварительную обработку исходных данных (в случае необходимости)
4. Построить регрессионные модели с параметрами, подобранными на перекрестной проверке (cross validation).

5. Спрогнозировать значение выходной переменной для тестовой выборки.
6. Набрать необходимый score.

Используемая метрика для оценки качества – RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Предварительно исследовать данные на коллинеарность. Построить и визуализировать матрицу корреляции (в виде тепловой карты). Удалить признаки, 1) которые слабо коррелируют с зависимой переменной, 2) сильно коррелируют друг с другом.

При кодировании категориальных переменных избегать «ловушки фиктивной переменной».

Для моделей без регуляризации использовать пошаговые алгоритмы отбора признаков.

Для построенных моделей выводить сводную информацию о значимости регрессионной модели и её коэффициентах.

Для модели рассчитать значения следующих критериев: коэффициент детерминации R^2, скорректированный коэффициент детерминации $Adj.R^2$, информативный критерий Акаике AIC, Байесовский информативный критерий BIC.

Для коэффициентов модели рассчитать значение статистики Стьюдента, проверить выполняется ли гипотеза относительно незначимости коэффициентов ($p_{value} > |t|$) и найти 95%-доверительный интервал.

Для получения данной сводной информации о значимости регрессионной модели и её коэффициентах модели можно использовать библиотеку `statsmodels`.

```
import statsmodels.formula.api as smf

results = smf.ols('Y ~ X', data=dataframe).fit()
print(results.summary())
```

Требования к выполнению практической работы:

1. Написание программного кода и формирование результатов согласно заданию и установленному варианту.

2. Составление отчета, содержащего описание решаемых задач методов решения и полученных результатов.

3. Воспроизводимость полученного результата.

Программный код и отчет должны быть выполнены в среде Jupyter notebook.