Домашнее задание №12 «Деревья решений»

Задание

***Цель*:** изучить применение дерева решений в рамках задачи регрессии

Описание задания:

В домашнем задании нужно решить задачу регрессии. В качестве датасета необходимо взять данные о недвижимости Калифорнии из библиотеки sklearn.datasets (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html). Целевая переменная – MedHouseVal. Прочитать информацию о признаках датасета можно, выполнив следующий код – print(fetch_california_housing().DESCR). На полученных данных построить модель регрессии и дерево решений.

Этапы работы:

1. Получите данные и загрузите их в рабочую среду. (Jupyter Notebook или другую).
2. Проведите первичный анализ.
   a.Проверьте данные на пропуски. Удалите в случае обнаружения.
   b.*Нормализуйте один из признаков.
3. Разделите выборку на обучающее и тестовое подмножества. 80% данных оставить на обучающее множество, 20% - на тестовое.
4. Обучите модель регрессии (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) на обучающем множестве.
5. Для тестового множества предскажите целевую переменную и сравните с истинным значением, посчитав точность предсказания модели. Для этого используйте встроенную функцию score.

6. Обучите дерево решений (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html) на обучающем множестве.
   a. Повторите п. 5 для полученной модели.
   b. Визуализируйте часть дерева решений. Убедитесь, что график получился читабельным. Посмотрите примеры визуализации по ссылке (https://mljar.com/blog/visualize-decision-tree/).
7. Оптимизируйте глубину дерева (max_depth). *Оптимизируйте ещё один параметр модели на выбор.
   a. Повторите п. 5 для полученной модели.
8. Сформулируйте выводы по проделанной работе.
   a. Сравните точность двух моделей.
   b. Напишите свое мнение, для каких задач предпочтительнее использовать обученные в работе модели? Какие у них есть плюсы и минусы?
   Для получения зачета по этому домашнему заданию, должно быть как минимум реализовано обучение двух моделей, выведена их точность, оптимизирован один параметр дерева решений.

9. **Результат**: получены знания по работе с деревом решений

10. Форма выполнения:

- ☐ ссылка на Jupyter Notebook, загруженный на GitHub;
- ☐ ссылка на Google Colab;
- ☐ файл с расширением .ipynb.

Инструменты:

- ☐ Jupyter Notebook/Google Colab;
- ☐ GitHub;
- ☐ библиотека sklearn.datasets;
- ☐ модель регрессии;
- ☐ дерево решений.
- ☐ Срок выполнения: дедлайн приема решений на проверку

*

Рекомендации к выполнению:

- текст оформляйте в отдельной ячейке Jupyter Notebook/Google Colab в формате markdown;
- у графиков должен быть заголовок, подписи осей, легенда (опционально). Делайте графики бОльшего размера, чем стандартный вывод, чтобы увеличить читабельность;
- убедитесь, что по ссылкам есть доступ на чтение/просмотр;
- убедитесь, что все ячейки в работе выполнены и можно увидеть их вывод без повторного запуска
- прикрепите ссылку с ноутбуком (на коллабе) в курс.