Промежуточная аттестация 3

Формулировка задания

1. Выбрать данные и выполнить загрузку данных;
2. Провести подготовку данных для алгоритмов машинного обучения;
3. Провести машинное обучение минимум 5 алгоритмами и оценить по критериям качества. Гиперпараметры алгоритмов настроить и подобрать под данные;
4. Выбрать один из инструментов AutoML и применить к данным. Сравнить работу AutoML с результатом машинного обучения без AutoML;

Описание плана работы

Задача 1 Загрузка данных

Для проведения исследования требуется выбрать табличные данные, которые не содержат персональных характеристик реальных людей.

Рекомендуется выбирать данные из наборов, доступных на Kaggle или в других источниках данных.

Использование выбранного набора данных необходимо согласовать с вашим наставником.

Задача 2 Подготовка данных к проведению машинного обучения

Подготовить данные для работы с алгоритмами машинного обучения (предварительная обработка и исследовательский анализ данных):

- Посмотреть основные статистики по столбцам и сделать вывод по данным;
- Обработать пропущенные значения;
- Обработать выбросы в данных;
- Обработать категориальные значения;
- Если признаков очень много, воспользуетесь методами отбора признаков;
- Если признаков всё ещё много, воспользуйтесь методами понижения размерности (например, PCA).
- Выбрать алгоритмы для задачи 3. Если алгоритмы требуют нормализации данных, то провести нормализацию.

Результат подготовки записать в файл .csv

Задача 3. Построение моделей машинного обучения

Изучить данные и выбрать целевой столбец. Определить какая будет решаться задача - классификация или регрессия.

Построить минимум 5 моделей машинного обучения с использованием следующих алгоритмов (использовать разные алгоритмы). Для каждой модели сначала подобрать гиперпараметры, затем построить саму модель.

Для решения задачи регрессии:

- Линейная регрессия
- Полиномиальная регрессия
- Регрессия с регуляризацией (Ridge, LASSO, ElasticNet)
- Регрессия knn
- SVR
- Регрессия дерева решений
- Random forest regressor
- Регрессор Градиентного бустинга (можно любой фреймворк, либо sklearn, xgboost, lightgbm, catboost)

Для решения задачи классификации:

- Логистическая регрессия
- Метод knn
- SVC
- Классификатор дерева решений
- Random forest классификатор
- Классификатор Градиентного бустинга (можно любой фреймворк, либо sklearn, xgboost, lightgbm, catboost)

Результат оценить критериями качества для классификации или для регрессии. Сравнить результаты по моделям (построить таблицы сравнения и графики).

Лучшие модели выгрузить и сохранить с помощью библиотеки joblib или аналогов.

Результат анализа записать в файл .csv

Задача 4. Применение инструментов AutoML

Выбрать один из инструментов AutoML и применить к данным.

Список возможных AutoML для выбора:

- H2O AutoML
- AutoSklearn
- Pycaret
- flaml AutoML
- LightAutoML
- FEDOT

- AutoGluon
- LAMA

Выполнить подготовку и настройку инструмента для табличных данных:

- Подготовить данные
- Настроить AutoML. Выбрать встроенные модели для машинного обучения. Определить параметры для автоматической работы моделей
- Оценить проведенное исследования критериями качества
- Построить визуализацию по проведенному исследованию
- Выгрузить лучшие модели машинного обучения

Сравнить работу AutoML с результатом машинного обучения без AutoML.

Входные данные:

- Табличные данные, выбранные по согласованию с наставником.

Выходные данные:

- Отчет со снимками экрана хода работы
- Код на python выполнения работы
- Очищенные и подготовленные данные в формате .csv
- Выгруженные лучшие модели машинного обучения вручную и с помощью AutoML

Перечень необходимых инструментов

- Python
- venv
- h2o
- autosklearn
- pycaret
- lightautoml
- scikit-learn
- xgboost
- catboost
- lightgbm
- pandas
- matplotlib
- seaborn
- plotly
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданиям и решение задач. Решения должны быть представлены в формате .ipynb или .py. Выгруженные промежуточные результаты добавить в архив

Шкала оценивания

- **1.0 – отлично**

Решены все задачи. Учтены требования, код структурирован. Продемонстрирован навык работы с подготовкой данных для машинного обучения. Продемонстрирован навык работы с настройкой гиперпараметров алгоритмов машинного обучения. Выбрана лучшая модель. Результаты проходят по критериям качества.

- **0.7–0.9 – хорошо**

Решены 1-3 задачи на высоком уровне. 4 задача выполнена минимально. Код соответствует большинству требований, но может содержать небольшие ошибки или недочеты. Есть небольшие замечания по настройке гиперпараметров алгоритмов и по выводам. Есть к чему стремиться в изучении темы.

- **0.5–0.6 – удовлетворительно**

Решена только 4 задача с множеством замечаний. Код соответствует большинству требований, но может содержать значительные ошибки или недочеты. Требуется дополнительное изучение тем модуля и исправление программы.

- **Менее 0.5 – задание не выполнено**

Задание выполнено на очень низком уровне или решено 0 задач. Требуется дополнительное изучение темы и решение заданий для практического изучения темы.