Домашнее задание
по теме «RF Regressor, Feature engineering»

Формулировка задания

Разобрать на практике работу с отбором признаков с помощью дополнительных библиотек (tsfresh, featuretools или feature-engine) и алгоритма Случайного леса.

Проведите генерацию и отбор признаков. Проведите регрессию методом случайного леса на наборе данных без генерации признаков и с генерацией признаков. Качество оценить минимум по 3 критериям качества для регрессии: MAE, MSE, RMSE, MAPE, RMSLE, R^2 и др. Сравнить результат и сделать вывод по важности сгенерированных признаков.

Для генерации признаков использовать библиотеки tsfresh, featuretools или feature-engine. Для регрессии случайного леса и оценки качества использовать библиотеку scikit-learn.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

1) Загрузите данные как в задаче по теме «Алгоритм Random Forest» из дополнительных материалов или по ссылке:
https://www.kaggle.com/competitions/playground-series-s4e12

Провести исследование на части данных. Отобрать 5000 - 10000 строк. Использовать случайное сэмплирование или другие методы.

2) EDA (Exploratory Data Analysis) использовать из задания «Алгоритм Random Forest». Учесть, что данных стало меньше. Нормализовывать данные не нужно.

4) Обучить модель регрессии RandomForestRegressor.

5) Оценить модель по критериям качества. Сделать выводы по критериям.

6) Применить к набору данных одну из библиотек генерации признаков.

7) Изучить сгенерированные столбцы. Исключить столбцы со значениями NaN и большим числом нулевых значений.
8) Разделить сгенерированные и отфильтрованные данные на тестовую и тренировочную выборки. Тестовая выборка - 20%;

9) Обучить регрессор Случайного леса на тренировочной выборке с отфильтрованными данными.

10) Оценить классификатор по критериям качества на тестовой выборке. Сделать выводы по критериям.

11) Отобразить список важности признаков из модели в виде таблицы и в виде графика. Сделать вывод.

Перечень необходимых инструментов

- Python
- scikit-learn
- tsfresh
- featuretools
- feature-engine
- pandas
- venv
- 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 заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено