Лабораторная работа №2

Вы используете исходные данные из лабораторной работы №1. Вариант также за вами сохраняется.

Вы решаете задачу бинарной классификации на таргет Ybin – бинарная переменная, где 1 обозначает регион с выраженностью демографических трендов.

Требования по отправке и оформлению лабораторной работы аналогичны требованию лабораторной работы №1.

Основные этапы выполнения работы:

1. Разделение данных на обучающую (85%) и тестовую часть (15%) случайным образом (можно сделать более корректным методом – разделить в такой пропорции с сохранением распределения таргета в каждой подвыборке – **желательно, но не обязательно**).
2. Нормирование (масштабирование) исходных данных. Обратите внимание, что данные для нормализации (масштабирования) рассчитываются только на основе обучающей выборки.
3. С помощью библиотеки sklearn сделать fit-predict модели kNN. Перебрать по сетке параметр числа соседей с целью определения наилучшего на тестовой выборке.
4. С помощью библиотеки sklearn сделать fit-predict модели логистической регрессии. Перебрать по сетке параметр регуляризации с целью определения наилучшего на тестовой выборке.
5. С помощью библиотеки sklearn сделать fit-predict модели дерева решений. Перебрать по сетке параметр глубины дерева с целью определения наилучшего на тестовой выборке.
   Дополнительно (желательно, но не обязательно): с помощью библиотеки sklearn сделать fit-predict модели случайного леса. Перебрать по сетке параметр глубины дерева с целью определения наилучшего на тестовой выборке.
6. Сравнить качество всех моделей на обучающей и тестовой выборке отдельно по метрикам Accuracy, ROC-AUC, Precision, Recall, F1-мера. Обратите внимание, что 4 из 5 метрик требуют определения порога отсечения по вероятности. В качестве эвристики предлагается взять его как среднее значение полученных вероятностей (**желательно, но не обязательно:** подобрать по сетке такой порог, при котором precision и recall примерно уравниваются).
7. Проанализировать различие в качестве между моделями. Определить на основе метрик модели, в которых сильно выражено переобучение.
8. Сравнить полученную важность признаков в модели логистической регрессии, в модели деревьев решений и в случайном лесе (для древесных моделей это можно сделать с помощью ключа feature_importances у обученной модели). Проинтерпретировать полученную важность признаков.