

# lab2

## Условие

*Данная лабораторная работа будет представлять собой соревнование на Kaggle.*

***Минздрав предупреждает: курение вредит вашему здоровью!***

В рамках этого соревнования вам предстоит обучить модель классификации на основе датасета из «train.csv». Предсказываемой переменной является «*smoking*». Часть выборки догенерирована нейронной сетью на основе существующего датасета, для улучшения качества модели вы также можете дополнить выборку изначальными данными.

После обучения вы сможете проверить эффективность вашего алгоритма на тестовом наборе данных из «test.csv», а затем отправить в тестирующую систему. В этом соревновании после его завершения все решения будут протестированы на приватной тестовой выборке, после чего лучшие решения будут награждены дополнительными баллами в рейтинг. Более подробную информацию по данным и посылкам решений можно найти на странице соревнования.

В данной работе разрешено использование любых моделей *классического* машинного обучения.

## Оценивание

### Разбалловка

Работа оценивается в 20 баллов и складывается из двух компонент:

**15р.** *Обязательные задания по коду.*

**5р.** *Защита лабораторной работы.*

### Задания по коду

Список обязательных заданий по коду (15 баллов) и их оценивание в баллах:

- 1р.** *Провести разведочный анализ данных (EDA): построить графики зависимости некоторых признаков друг от друга, график целевой переменной и матрицу корреляций, сделать выводы. Подготовить на основе выводов данные для обучения.*
- 3р.** *Реализовать класс бэггинга, который для обучения методом *fit* обучает *n* базовых моделей *b*. Сравнить полученные после обучения результаты с реализацией из *sklearn*.*
- 5р.** *Реализовать класс градиентного бустинга, который для обучения методом *fit* обучает *n* базовых моделей *b*. Сравнить полученные после обучения результаты с реализацией из *sklearn* и других библиотек.*

- 1р. Сравнить результаты работы алгоритмов градиентного бустинга: реализация из *sklearn*, *LightGBM*, *XGBoost*, *CatBoost*. Выбрать алгоритм с лучшей метрикой на данных.
- 2р. Подобрать оптимальные гиперпараметры для лучшей модели с помощью *Optuna*.
- 0.5р. Реализовать метрику *Accuracy*, протестировать и сравнить полученный результат с *Accuracy* из *sklearn*.
- 0.5р. Реализовать *Precision*, протестировать и сравнить с метрикой из *sklearn*.
- 0.5р. Реализовать *Recall*, протестировать и сравнить с метрикой из *sklearn*.
- 0.5р. Реализовать *F1-score*, протестировать и сравнить с метрикой из *sklearn*.

Дополнительные 2 балла можно получить за реализацию метрик AUC-ROC и AUC-PR и последующее сравнение с готовыми реализациями.

## Основная метрика

Для получения баллов за код и последующего допуска к защите необходимо получить приемлемое значение основной метрики задачи.

В качестве основной метрики используется ROC-AUC (Area Under the Receiver Operating Characteristic Curve). Метрика вычисляет площадь под кривой ROC, которая отражает зависимость между True Positive Rate (TPR) и False Positive Rate (FPR) при различных порогах классификации.

**Минимальное пороговое значение ROC-AUC, которое можно получить при предсказаниях на публичной тестовой выборке и при котором можно допустить к защите работы составляет 0.855.**

Важно отметить, что хотя ROC-AUC является основной метрикой для оценки качества модели и окончательный скор будет вестись только по ROC-AUC, вы должны использовать и другие метрики для самопроверки, такие как Precision, Recall, F1-score, PR-AUC.

## Загрузка ноутбука

После получения приемлемой метрики ( $\text{ROC-AUC} > 0.855$ ) после отправки на Kaggle, необходимо загрузить ваш ноутбук («.ipynb»-файл) на свой GitHub для оценивания заданий по коду, а также для проверки на оригинальность. Затем нужно оставить комментарий в Google-таблице с оценками в столбце «lab2» в строке со своей фамилией о том, что вы выполнили работу с указанием ника на Kaggle.

## Защита работы

После оценки заданий по коду, вы должны защитить лабораторную работу. За защиту можно получить до 5 баллов. На самой защите необходимо кратко рассказать про своё решение, а также ответить на вопросы по теории, связанные с логистической регрессией, метриками классификации, решающих деревьев, ансамблями (бэггинг и бустинг) и подбором гиперпараметров.

Оценку за задания по коду можно повышать до дедлайна, оценку за защиту поднять нельзя.