Распределение баллов: 35 баллов

Название модуля: Соревнования на Kaggle 1. Регрессия

# Регистрация на Kaggle

РЕГИСТРАЦИЯ НА KAGGLE

**Для участия в соревнованиях вам необходимо зарегистрироваться на сайте:** [**https://www.kaggle.com**](https://www.kaggle.com/)**.**

**Важно, чтобы по имени пользователя вас можно было однозначно определить. Используйте при регистрации своё имя и фамилию.**

**Если у вас возникнут проблемы с верификацией телефона на Kaggle, попробуйте вводить телефон в формате 79xxxxxxxxx (даже несмотря на то, что код страны уже введён).**

**Если данный способ не поможет — напишите в тех. поддержку:** [**https://www.kaggle.com/contact#/account/activate/phone**](https://www.kaggle.com/contact#/account/activate/phone)**. Опыт показывает, что после 1–2 обращений в тех. поддержку аккаунт верифицируется.**

**Удачи в соревнованиях!**

# Соревнование № 1. Регрессия

СОРЕВНОВАНИЕ № 1

**В этом соревновании вам предстоит показать навыки обучения моделей машинного обучения для регрессии. Для успешного прохождения соревнования вам также придется генерировать признаки для модели.**

**Для участия в соревновании вам необходимо:**

* **Пройти по ссылке на [соревнование](https://www.kaggle.com/t/46230b18e5774be0ab2af7f7fd16f16a).**
* **Ознакомиться с описанием соревнования (нажать «Участвовать в соревновании»).**
* **Ознакомиться с содержанием** [**блокнота**](https://drive.google.com/file/d/1sSZ9miNy2Ds1xfOBtK2vzhPBYSu_4hv-/view?usp=sharing) **для знакомства с [Kaggle API, Join, Pipeline для предварительной обработки](https://colab.research.google.com/drive/1NrjLYISj81ZSvXQnNibpsRsS3YaI7I3n?usp=sharing). Рекомендуется хотя бы один раз полностью пройтись по блокноту и загрузить базовое решение на платформу Kaggle.**
* **Обучить модель машинного обучения:**
  + **Загрузить данные соревнования (через Kaggle API, как показано в блокноте примере или «скачав» их из Kaggle с дальнейшей загрузкой в Colab Research, что не рекомендуется).**
  + **Выполнить предварительный анализ данных (напомним, что подобное вы выполняли в рамках [первого практического задания в 1 семестре](https://apps.skillfactory.ru/learning/course/course-v1:Skillfactory+URFUML2023+SEP2023/block-v1:Skillfactory+URFUML2023+SEP2023+type@sequential+block@4d952a3ad913418eb1e71373aa4c54a9/block-v1:Skillfactory+URFUML2023+SEP2023+type@vertical+block@de98e0c578044022badc0747bd026709)) .**

**Ожидается, что вы не сразу начнёте крутить fit-predict , а выполните предварительный анализ. Работа включает в себя визуализацию данных и их предварительный статистический анализ (в том числе агрегацию для проверки начальных гипотез), расчёт корреляции признаков с целевой переменной (т.к. сравнение мультиколлинеарности важно для задачи регрессии), а также предварительную обработку данных.. Для визуализации данных не забудьте присоединить (Join) целевую переменную к столбцу с признаками, чтобы у вас была возможность увидеть потенциальную со-зависимость целевой переменной и признаков.**

**Также для большой таблички (studentVle) желательно извлечь признаки: например, сделать сводную таблицу с использованием методов groupby и agg.**

* **Обучить модель на ваш выбор:**
  + **Можно начать с простой модели линейной регрессии с регуляризацией, как мы делали в примере, и попытаться оптимизировать её гиперпараметры, как показано в [6 модуле 1 семестра](https://drive.google.com/file/d/1EOlrpXy3o9Znl1E5zGw43XtNwZY7cxGI/view?usp=sharing).**
  + **Можно воспользоваться метрическими моделями, продвинутыми линейными моделями, деревьями решений, случайным лесом или бустингами. Все эти модели рассматриваются во 2 семестре.**
  + **Можно использовать модели, не рассматриваемые в рамках данного курса.**
* **Получить предварительные оценки модели с использованием кросс-валидации и/или отложенной выборки. Метрика на платформе mse (среднеквадратичная ошибка).**

**ВАЖНО! На платформе Kaggle стоит ограничение на 5 загрузок в день. Поэтому вам следует не “спамить” загрузками в надежде подобрать предсказания вручную. Сначала получаете предварительные оценки, которые показывают что новая модель предсказывает лучше предыдущей, и только потом отправляете данные на платформу.**

* **Когда будете уверены в своей модели, подготовьте итоговый файл для загрузки, применив предсказание обученной модели к тестовым данным..**
* **Загрузить файл (через Kaggle API или напрямую на страницу соревнования).**
* **Если метрика на платформе недостаточно хороша, то попытайтесь добавить новые признаки, поменять модель, оптимизировать гиперпараметры (предпочтительно именно в таком порядке) и повторить предварительную оценку и загрузку..**
* **Загрузить ссылку (предварительно убедившись в открытом доступе) или ipynb-файл своего решения на платформу Skillfactory. В блокнот необходимо добавить скриншот метрики, из которого понятно, что он принадлежит вам (ваш ник на платформе), а ваша метрика действительно ваша (рекомендуется добавить обозначение, показывающее, что вы из Skillfactory).**

**Оцениваться будет как итоговая метрика, оформление блокнота и реализация базовых вещей (визуализация данных, кросс-валидация, анализ признаков, комментарии), так и оригинальность: например, дизайн графиков, random\_state, порядок метрик при оценке валидации.**

**Блокнот необходимо прилежно оформить:**

* **в ответе должна присутствовать структура (заголовки разного уровня);**
* **ожидается видеть не только код, но и комментарии к нему;**
* **не должно быть запущенных ячеек с ошибками;**
* **ожидается, что в блокноте есть финальное решение;**
* **рисунков адекватное количество и они полезные (т.е. вы не просто построили множество графиков для формального выполнения критериев, а проанализировали их и оставили только наиболее информативные);**
* **рисунки должны быть наглядно и понятно оформлены — немаленького размера, с названием и легендой при необходимости..**

**Критерий 1. Метрика на Kaggle MSE (Mean Square Error)**

|  |  |  |
| --- | --- | --- |
| **Вариант 1** | **MSE<90** | **20 баллов** |
| **Вариант 2** | **90<MSE<100** | **18 баллов** |
| **Вариант 3** | **100<MSE<150** | **15 баллов** |
| **Вариант 4** | **150<MSE<200** | **10 баллов** |
| **Вариант 5** | **200<MSE<250** | **7 баллов** |
| **Вариант 6** | **250<MSE<300** | **6 баллов** |
| **Вариант 7** | **300<MSE<400** | **5 баллов** |
| **Вариант 8** | **400<MSE<500** | **4 балла** |
| **Вариант 9** | **500<MSE<600** | **3 балла** |
| **Вариант 10** | **600<MSE<700** | **2 балла** |
| **Вариант 11** | **700<MSE<800** | **1 балл** |
| **Вариант 12** | **800<MSE** | **0 баллов** |

**Критерий 2. Наполнение блокнота**

|  |  |  |
| --- | --- | --- |
| **Вариант 1** | **В дополнение к визуальному анализу и подбору гиперпараметров модели добавлен анализ модели и её интерпретация. Сделаны выводы о том, какие гиперпараметры оказались наиболее значимыми для модели.** | **10 баллов** |
| **Вариант 2** | **Дополнительно к визуальному анализу добавлен подбор гиперпараметров модели.** | **7 баллов** |
| **Вариант 3** | **Есть небольшое количество графиков, добавлены комментарии к ним.** | **4 балла** |
| **Вариант 4** | **Запущен лишь базовый пример с небольшими изменениями.**  **Или в блокноте слишком много графиков добавленных просто так, без анализа.** | **1 балл** |

**Критерий 3. Общее оформление**

|  |  |  |
| --- | --- | --- |
| **Вариант 1** | **Ячейки выполнены без ошибок.**  **Почти ко всем ячейкам есть оригинальные комментарии, описывающие действия.** | **5 баллов** |
| **Вариант 2** | **Ячейки выполнены без ошибок. Но как правило, это просто код без комментариев.**  **Авторство сомнительно.** | **3 балла** |
| **Вариант 3** | **Блокнот есть, но это просто набор невыполненных ячеек.**  **Авторство сомнительно.** | **1 балл** |

# Оценивание

Данное задание оценивается в рамках промежуточной аттестации.

**Максимальное количество баллов — 35.**

**Домашнее задание проверит ментор и поставит оценку. Если возникнут сложности, обращайтесь в канал модуля. Желаем успехов! ;)**