**Индивидуальный проект**

Предлагается провести полный анализ данных и построение прогнозной модели для какого-то из наборов данных репозитория UCI, Kaggle или любого другого ([Kaggle datasets](https://www.kaggle.com/datasets) – прекрасный источник всякого рода данных). Можно приносить и свои данные. Баллы можно зачесть либо за этот проект либо за "Элис". Оцениваться проекты будут по схеме Peer-review, критерии оценивания перечислены ниже. По каждому пункту надо будет поставить +, +/-, -/+ или -. Максимум за проект – 40 баллов. Проверка 3 проектов своих товарищей – необходимое условие получения баллов. Времени до 22 апреля, а до 25 апреля надо проверить 3 чужих проекта. Контактное лицо – Илья Барышников a.k.a. @lucidyan.

Опубликовать проект следующим образом:

1. Заранее написать свою предполагаемую тему в [гугл-доке](https://goo.gl/ZzNuUN)
2. Сделать pull request с проектом в виде Jupyter-тетрадки [сюда](https://github.com/Yorko/mlcourse_open/tree/master/jupyter_russian/projects_individual). При пул реквесте соблюдайте структуру репозитория (данные – в data, картинки – в img), большие файлы (скажем, больше 5 Mb) не надо комитить – лучше ссылку на данные дать в тетрадке. И естественно, перед этим обновите свою версию репозитория
3. Дождаться от @yorko nbviewer-ссылку, кинуть в канал #mlcourse\_open ссылку, а также краткое описание, чему посвящен проект. Поставить тег *#proj\_indiv* (обязательно), запинить (т.е. отправить в pinned items)
4. Оцениваться проекты, в отличие от тьюториалов, будут не по плюсам, а в процессе взаимной проверки (peer-review), про которую сказано чуть выше.

**План и критерии оценки**

**1. Описание набора данных и признаков (2 балла)**

* (+) Описан процесс сбора данных (если применимо), есть подробное описание решаемой задачи, в чем ее ценность, дано описание целевого и прочих признаков;
* (+/-) Сказано, какая задача решается, откуда данные, что есть целевой признак. Даны названия признаков;
* (-/+) Сказано, какая задача решается, откуда данные и что есть целевой признак;
* (-) Описание отсутствует и дано только название датасета или решаемой задачи, скажем, "прогноз оттока".

**2. Первичный анализ данных (4 балла)**

* (+) Исследованы признаки, их взаимодействия, влияние на целевой признак. Исследовано распределение целевого признака (в случае задачи регрессии проведены стат-тесты на нормальность и скошенность (skewness) распределения). Если необходимо, объясняется, почему и как можно преобразовать целевой признак. Изучены выбросы и пропуски в данных;
* (+/-) Присутствует часть пунктов полного исследования. Даны пояснения;
* (-/+) Нет большей части пунктов полного исследования и пояснений;
* (-) Отсутствует.

**3. Первичный визуальный анализ данных (4 балла)**

* (+) Построены визуализации (распределения признаков, матрица корреляций и т.д.), описана связь с анализом данным (п. 2). Присутствуют выводы;
* (+/-) Построены визуализации (распределения признаков, матрица корреляций и т.д.). Присутствуют выводы с небольшими ошибками;
* (-/+) Недостает важных визуализаций и/или присутствует много ошибок в выводах;
* (-) Отсутствует.

**4. Инсайты, найденные зависимости (4 балла)**

* (+) Найдены и выдвинуты предположения о природе различных корреляций/пропусков/закономерностей и выбросов, найденных в предыдущих пунктах. Есть пояснение, почему они важны для решаемой задачи;
* (+/-) Найдены закономерности, есть пояснения;
* (-/+) Найдены закономерностей, но значимость инсайтов не показана или слабо выражена;
* (-) Отсутствуют.

**5. Выбор метрики (3 балла)**

* (+) Есть разумное обоснование выбора метрики качества модели. Описаны моменты, влияющие на выбор метрики качества (решаемая задача, цель решения, количество классов, дисбаланс классов, прочее);
* (+/-) Присутствует описание, но оно неполное или нет связи (слабая связь) между обоснованием и принимаемым решением; (-/+) Обоснование выбора метрики неверное или имеет серьезные ошибки;
* (-) Нет обоснования выбора метрики.

**6. Выбор модели (3 балла)**

* (+) Произведен выбор модели. Описан процесс выбора и связь с решаемой задачей;
* (+/-) В процессе выбора допущены неточности, присутствует объяснение;
* (-/+) В процессе выбора допущены серьезные ошибки и/или отсутствует объяснение;
* (-) Выбор модели не обоснован.

**7. Предобработка данных (4 балла)**

* (+) Проведена предобработка данных для конкретной модели. При необходимости есть и описано масштабирование признаков, заполнение пропусков, замены строк на числа, OheHotEncoding, обработка выбросов, отбор признаков с описанием используемых для этого методов. Корректно сделано разбиение данных на обучающую и отложенную части;
* (+/-) Описание каких-то шагов пропущено, хотя в коде есть. Например, есть выкидывание выбросов, но оно не прокомментировано. Либо есть мелкие недочеты;
* (-/+) Каких-то шагов не хватает, например, масштабирования признаков в случае линейных или метрических моделей. Или выделения отложенной выборки. Либо есть ошибки типа числового кодирования (LabelEncoder) категориальных признаков там, где это неуместно. Либо даже подобные действия могут быть оправданы, но не дается никаких комментариев;
* (-) Отсутствует.

**8. Кросс-валидация и настройка гиперпараметров модели (4 балла)**

* (+) Кросс-валидация выполнена технически верно, нет утечек данных. Разумно выбрано количество фолдов и разбиение (Random/Stratified или иное), зафиксирован seed. Присутствует объяснение. Объяснены гиперпараметры модели и способ их выбора. Выбор основан на некотором исследовании гипрепараметров модели для данной задачи;
* (+/-) Присутствуют незначительные ошибки (например, не зафиксирован seed) или отсутствует объяснение кросс-валидации. Но гиперпараметры модели и способ их выбора должны быть объяснены;
* (-/+) Кросс-валидация выполнена со значительными ошибками (например, преобразования данных проводится на всей выборке, таким образом возникает утечка данных из тестовой части выборки и, соответственно, результат кросс-валидации может иметь слишком оптимистичное значение). Гиперпараметры модели и способ их выбора не объяснены;
* (-) Отсутствуют.

**9. Создание новых признаков и описание этого процесса (4 балла)**

* (+) Созданы новые признаки. Дано обоснование: логическое (например, у птиц температура тела на несколько градусов выше человеческой, значит вирус ХХХ не выживет в такой среде), физическое (например, радуга означает, что источник света расположен сзади; расчет величины по физическому закону с использованием данных признаков) или другое (скажем, признак построен после визуализации данных). Обоснование разумно описано. Полезность новых признаков подтверждена статистически или с помощью соответствующей модели;
* (+/-) Созданные признаки малозначимы для используемой модели и/или есть ошибки при формировании признаков. Присутствует обоснование выбора признаков;
* (-/+) Созданные признаки ухудшают качество и/или отсутствует обоснование;
* (-) Новые признаки не созданы.

**10. Построение кривых валидации и обучения (4 балла)**

* (+) Построены кривые валидации и обучения. Дана верная интерпретация;
* (+/-) Заметны ошибки при построении кривых или кривые не информативны;
* (-/+) Дана неверная интерпретация или заметны другие серьезные ошибки;
* (-) Отсутствуют кривые обучения и валидации.

**11. Прогноз для тестовой или отложенной выборке (2 балла)**

* (+) Указаны результаты на тестовой выборке или LB score. Результаты на тестовой выборке сравнимы с результатами на кросс-валидации. Если тестовая выборка создавалась автором проекта, то механизм создания должен быть непредвзят и объяснен (применен разумный механизм выборки, в простейшем случае – рандомизация);
* (+/-) Значения метрик на тестовой выборке не сильно отличаются от значений метрик на кросс-валидации и/или тестовая выборка создана предвзято, но есть разумное обоснование этому (пример: заказчик взял тестовую выборку из другого распределения);
* (-/+) Значения метрик на тестовой выборке сильно отличаются от значений метрик на кросс-валидации и/или тестовая выборка создана предвзято;
* (-) Прогноз для тестовой или отложенной выборки отсутствует.

**12. Выводы (2 балла)**

* (+) Описана ценность решения, возможности применения, дальнейшие пути развития и улучшения решения;
* (+/-) Выводы похожи на формальное описание выполненного решения и есть описание ценности решения, дальнейших путей развития и возможностей применения;
* (-/+) Выводы похожи на формальное описание выполненного решения;
* (-) Выводы отсутствуют.