**Dataset Description**

В этом конкурсе ваша задача - разработать модель для предсказания успешности стартапа (закроется или нет)

**Данные**

Датасет состоит из двух файлов: тренировочный набор (около 53к записей) и тестовый набор (около 13к записей). Тренировочный набор содержит целевой признак status, указывающий на то, закрылся стартап или продолжает действовать. Временной период - '1970-01-01' по '2018-01-01'. Дата формирования выгрузки - '2018-01-01'

kaggle\_startups\_train\_01.csv - информация (53 000) стартапах, которые будут использоваться в качестве обучающих данных.

kaggle\_startups\_test\_01.csv - информация (13 000) стартапах, которые будут использоваться в качестве тестовых данных. Ваша задача - предсказать значение 'status' для каждого стартапа из этого датасета.

kaggle\_startups\_sample\_submit\_01.csv - файл с примером предсказаний в правильном формате.

name - идентификатор (название) стартапа в тестовом наборе.

status - целевой признак. Для каждого стартапа предскажите категориальное значение соответствующее прогнозу ['operating', 'closed'].

Описание **полей** данных

name - Название стартапа

category\_list - Список категорий, к которым относится стартап

funding\_total\_usd - Общая сумма финансирования в USD

status - Статус стартапа (закрыт или действующий)

country\_code - Код страны

state\_code - Код штата

region - Регион

city - Город

funding\_rounds - Количество раундов финансирования

founded\_at - Дата основания

first\_funding\_at - Дата первого раунда финансирования

last\_funding\_at - Дата последнего раунда финансирования

closed\_at - Дата закрытия стартапа (если применимо)

lifetime - Время существования стартапа в днях

**Description**

**В ходе работы в пройдете все основные этапы полноценного исследования:**

* загрузка и ознакомление с данными,
* предварительная обработка,
* полноценный разведочный анализ,
* разработка новых синтетических признаков,
* проверка на мультиколлинеарность,
* отбор финального набора обучающих признаков,
* выбор и обучение моделей,
* итоговая оценка качества предсказания лучшей модели,
* анализ важности ее признаков,
* подготовка отчета по исследованию.

**Дополнительное задание**

1. Для тех, кому этого покажется мало у нас есть дополнительное задание — реализовать решение с использованием технологии pipeline (из библиотеки sklearn, imblearn).
2. Выполнить полноценный исследовательский анализ - выделить важные компоненты из данных, влияющие на работу стартапов. Проще говоря на основе данных и графиков сформулировать рекомендации позволяющие повысить шанс на успех стартапа.
3. Подготовка отчета по исследованию. Если останется время можно подготовить полноценный отчет по исследовательской работе Дата Сайнтиста. Примеры хороших отчетов можно посмотреть на сайте Stanford University - отчеты по финальным проектам направления Machine Learning - <https://cs229.stanford.edu/projects2014.html>. Из примеров понятно наполнение отчета по исследованию, а так же, что главное - понятна сама структура отчета. В реальной работе именно такой формат отчета обычно завершает исследование.

Мастерская будет проходить в виде соревнования на платформе Kaggle. Длительность — 2 недели. Для участников — это отличная возможность не только выполнить интересный проект, который можно будет добавить в свое портфолио, но и познакомиться с одной из самых популярных соревновательных Интернет-площадок, на которой можно исследовать и создавать модели, взаимодействовать с другими специалистами по данным, организовывать и участвовать в конкурсах, посвященных анализу данных и машинному обучению.

**Перечень библиотек python, рекомендуемых для проведения исследования**

Для успешного проведения исследования и построения моделей рекомендуется использовать следующие библиотеки Python:

* pandas - для работы с данными и математических операций.
* matplotlib и seaborn - для визуализации данных.
* scikit-learn - для предобработки данных, разделения на тренировочные и тестовые наборы, построения и оценки моделей машинного обучения.
* catboost или lightgbm - для работы с продвинутыми моделями градиентного бустинга.
* shap - для анализа важности признаков модели
* phik - для оценки признаков на мультиколлинеарность

Соревнование нацелено не только на закрепление существующих у студентов навыков, но и на стимулирование самостоятельного изучения новых методов и подходов в области Data Science и машинного обучения.

**Evaluation**

**Оценка качества предсказаний**

Метрикой в этом соревновании является [F1-Score](https://en.wikipedia.org/wiki/F-scoref). F1-мера вычисляется на основе точности p и полноты r. Точность - это отношение true positives (tp) ко всем predicted positives (tp + fp). Полнота это отношение true positives ко всем actual positives (tp + fn). Тогда F1 задается формулой:

Метрика F1 является гармоническим средним между Recall и Precision. Хорошая модель при обучении будет максимизировать Recall и Precision одновременно. Таким образом, умеренно хорошие показатели по обоим параметрам будут предпочтительнее, чем чрезвычайно хорошие показатели по одному и плохие по другому. Функция для вычисления метрики - <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html>

**Пример файла с предсказаниями**

|  |  |
| --- | --- |
| **name** | **status** |
| Crystalsol | operating |
| JBI Fish & Wings | closed |
| COINPLUS | closed |
| Imagine Communications | operating |
| DNA13 | operating |