**Техническое задание на подготовку соревнования по анализу данных и машинному обучению (в частности решению контейнерной и табличной задач) Data Fusion Contest 2025**

[Data Fusion Contest 2025](https://data-fusion.ru/contest.html) – это соревнование по анализу данных и машинному обучению (в частности решению контейнерной и табличной задач).

Заказчик – ООО «Т1», Исполнитель – ODS (ООО «Соревнования анализа данных»).

**Сроки проведения соревнования для участников:** февраль - апрель 2025

**Термины и определения:**

**Платформа соревнования** – специализированное программное обеспечение, функционал которого предназначен для проведения онлайн-соревнований по тематике машинного обучения, анализа данных, data science.

**Датасет** – массив обезличенных синтетических данных, предоставленный Заказчиком, необходимый для решения задач соревнования.

**Задача соревнования** – описание конкретной проблемной ситуации, для решения которой необходимо использовать инструменты машинного обучения и представленный Заказчиком датасет.

**Метрика** – мера оценки качества решения исходя из эталонной разметки с монотонной зависимостью значения метрики от уровня качества решения (модели)

**Уязвимости задач –** недостатки в задачах, используя которые можно намеренно нарушить целостность и вызвать неправильную работу. Виды потенциальных уязвимостей перечислены в пункте 8 Технического задания.

**Пайплайн решения задачи –** последовательность действий участника, который приводит к результату выполнения задачи.

**Лидерборд -** инструмент, с помощью которого визуализированы персонализированные соревновательные таблицы с результатами лучших участников и местом участников в рейтинге.

**Чат с участниками –** публичный канал для коммуникации участников с организаторами по техническим и организационным вопросам соревнования.

**Лендинг** - веб-страница для вовлечения и информирования пользователей о соревновании Data Fusion Contest 2025.
**Онлайн-митап** – онлайн-формат встречи участников и организаторов соревнования для обсуждения методов и подходов к решению задач от участников и организаторов соревнования.

1. **Общие требования к Исполнителю по подготовке соревнования «под ключ»:**

1.1. Подготовка данных (данные предоставляются Заказчиком), определение всех связанных с данными ограничений, проведение необходимой для соревнования обработки/чистки данных. Обогащение данных из внешних источников при необходимости.

1.2. Исследование аналогичных данных, постановок и соревнований на российских и международных платформах. Подготовка формулировки заданий для соревнования на основе исследования. Необходимо изучить аналогичные постановки задач, в случае если ранее уже проводились соревнования по выбранной теме. Также необходимо изучить релевантные научные статьи по теме задачи, чтобы корректно предложить и сформулировать метрики соревнования.

1.3. Определение основных и специальных номинаций, их количества, критериев, целесообразность и распределение вознаграждения между победителями и финалистами.

1.4. Проработка возможных уязвимостей задач со стороны данных, метрик и формулировок, а также злоупотреблений со стороны участников и их предотвращение.

1.5. Передача Заказчику полного набора протоколов, закрывающей документации и полного набора данных участников соревнования, описание решений победителей, исходный текст и воспроизводимый код решения победителей.

1. **Требования к экспертной команде, разрабатывающей соревнование на стороне Исполнителя**

2.1. Требуется наличие релевантного опыта у членов команды для решения самой задачи соревнования.

* 1. Индустриальный опыт. У команды должен быть не только релевантный опыт в решении задач машинного обучения и анализа данных, но и опыт решения либо схожих индустриальных задач, либо аналогичных исследовательских задач.
	2. Опыт в подготовке подобных соревнований по анализу данных и машинному обучению.

2.2. Наличие у команды итеративного процесса работы в рамках подготовки задачи, включая оценку и коммуникацию планируемых работ внутри итераций. Продуктом работы команды является подготовленное и проработанное соревнование со всеми его артефактами. У соревнований имеется универсальный тест в виде проверки обработки базовых решений на развернутом на платформе соревновании.

2.3. Наличие у команды опыта оформления соревнований и/или подготовки интерактивных образовательных материалов по практике задач DS/ML:

1. В рамках документации, необходимо оформить страницы всех ключевых составляющих соревнования.
2. Оформление как самих базовых решений, так и при надобности, раздела с метриками, следует реализовывать с уклоном в образовательные материалы. В частности, абсолютно все необходимые для полного пайплайна решения задачи элементы должны быть сопровождены достаточным количеством комментариев.
3. Необходимо заранее проработать серию подсказок и рекомендаций для участников о том, какими открытыми инструментами и техниками анализа можно было бы воспользоваться, чтобы улучшить свои результаты.

**3. Требования к Платформе соревнования, предоставляемой Исполнителем**

3.1. Платформа соревнований должна иметь функциональные возможности по обеспечению мощностей контейнерного соревнования, а также возможностей управления ими.

3.2. Платформа должна иметь возможность использования серверных мощностей от облачных провайдеров, в том числе с использованием образов с GPU не ниже, чем NVIDIA V100 GPU 96 GB.

3.3. На платформе и/или у ее команды должна быть возможность перераспределения мощностей: наращивания числа выбранных машин, а также их сокращения, с целью перераспределения приоритета обработки очереди.

3.4. На платформе должен быть реализован функционал сбора информации о количестве решений в очереди, для связи как с организаторами и/или операторами соревнования, так и для коммуникации участникам статусов их решений.

3.5. На платформе должен быть реализован функционал по оперативной публикации лидерборда с рейтингом участников.

**4. Требования к обеспечению вычислительных мощностей**

4.1. Вычислительные мощности для не менее 1500 участников контейнерного соревнования с возможностью одновременной обработки не менее 15 решений участников в среднем не превышающее 60 минут.

4.2. Обеспечение вычислительных мощностей соревнования, обеспечивающие не менее 5000 решений по каждой задаче, в течение трёх месяцев активной стадии соревнования (февраль-апрель 2025 г.).

4.3. Расчёт числа машин необходимо делать с учётом резервирования машин для случаев скапливания решений в очереди, а также с поправкой на известные паттерны по активности отправленных решений: существенных увеличений темпов отправляемых решений в последние 7-3-1 дней, что требует последовательного наращивания числа машин для обеспечения пропускной способности проверочной системы.
4.4. Вычислительные мощности должны быть зарезервированы и обеспечены в течение соревнования исходя из согласованных расчетов вычислительных ресурсов.

4.5. До соревнования Исполнитель и Заказчик согласовывают необходимую конфигурацию машин. Все машины должны соответствовать согласованной конфигурации не ниже, чем NVIDIA V100 GPU 96 GB.

4.6. Машины должны быть доступны участникам. В случае возникновения больших очередей число машин должно под них наращиваться. Необходимо устранять очереди за менее, чем 12 часов.

**5. Требования к основной соревновательной задаче**

5.1. Подготовка основной соревновательной задачи по теме «NLP/LLM. Большие языковые модели» в контейнерном или табличном формате (финальный формат задачи согласовывается с Заказчиком).

5.2. Задача по «NLP/LLM. Большие языковые модели» формируется на предоставленном от Заказчика синтетическом датасете на обезличенных данных вида (клиент-локация) со следующими признаками: кол-во/сумма операций в локации за период; разбивка операций по временным интервалам; разбивка операций по типам продавцов; прочие статистики и агрегаты (средние, макс, мин и тд); разметка локаций по целевому действию (например, использование банкомата или домашней локации); бейзлайн модель с метрикой качества.

5.3. Сформировать постановку задачи по «NLP/LLM. Большие языковые модели» в которой участникам необходимо разметить локации по целевому действию для другого набора клиентов. Участникам должно быть доступно несколько направлений для работы (например, добавление и расчет новых фичей на имеющихся данных или включение в модель внешних открытых данных (например, open-street-maps); эксперименты с формализацией задачи для выбора архитектуры модели (может быть и простой классификатор, так и использование графов или подходов из рекомендательных систем).

5.4. Подготовка выборок для соревнования: обучающая, открытая валидационная, приватная тестовая и др. необходимые. Проверка качества полученных выборок с точки зрения воспроизводимости решений на разных её частях с сохранением качества, отсутствия утечек целевой переменной, достаточности объема данных.

5.5. Выбор оптимальных целевых метрик, обеспечивающих оптимальную чувствительность к улучшению решений участниками соревнования, разработка протокола валидации и тестирования, имплементация протокола и самих метрик под задачу.

5.6. Прорешивание задачи, поиск и устранение потенциальных уязвимостей в постановке задачи. Проработка основных путей решения задач для определения общей решаемости задач и возможностей для участников соревнования по созданию качественных решений.

5.7. Подготовка бейзлайнов и «starter kit» для участников. Подготовка/выбор «стартового» образа для запуска бейзлайна/построения начальных решений.

5.8. Подготовка описания задачи и сопроводительных материалов, и документов.

**6. Требования к дополнительной соревновательной задаче**

6.1. Подготовка дополнительной соревновательной задачи по теме «Модели оттока клиентов» в контейнерном или табличном формате (финальный формат задачи согласовывается с заказчиком).

6.2. Разработка соревновательной задачи с постановкой для участников, где необходимо разработать модель оттока из Банка клиента – физического лица на третий/шестой/девятый/двенадцатый месяц. Под оттоком понимается переход клиента из статуса «Активный» на месяц оценки в статусы, характеризующие клиента как неактивного, на третий месяц/шестой/девятый/двенадцатый, считая со следующего месяца от месяца оценки.

6.3. Задача по «Моделям оттока» формируется на предоставленном от заказчика синтетическом датасете на обезличенных данных.

6.4. Подготовка выборок для соревнования: обучающая, открытая валидационная, приватная тестовая и др. необходимые. Проверка качества полученных выборок с точки зрения воспроизводимости решений на разных её частях с сохранением качества, отсутствия утечек целевой переменной, достаточности объема данных.

6.5. Выбор оптимальных целевых метрик, обеспечивающих оптимальную чувствительность к улучшению решений участниками соревнования, разработка протокола валидации и тестирования, имплементация протокола и самих метрик под задачу.

6.6. Прорешивание задачи, поиск и устранение потенциальных уязвимостей в постановке задачи. Проработка основных путей решения задач для определения общей решимости задач и возможностей для участников соревнования по созданию качественных решений

6.7. Подготовка базовых решений и «starter kit» для участников.

6.8. Подготовка описания задачи и сопроводительных материалов, и документов.

**7. Требования к образовательной задаче**

7.1. Подготовка образовательной задачи по теме «Задача прогнозирования сбережений на сбер.продуктах» в табличном формате (финальный формат задачи согласовывается с заказчиком).

7.2. Разработка соревновательной задачи с постановкой для участников, где необходимо разработать модель оттока из Банка клиента – физического лица на третий/шестой/девятый/двенадцатый месяц. Под оттоком понимается переход клиента из статуса «Активный» на месяц оценки в статусы, характеризующие клиента как неактивного, на третий месяц/шестой/девятый/двенадцатый, считая со следующего месяца от месяца оценки.

7.3. «Задача прогнозирования сбережений на сбер.продуктах» формируется на предоставленном от заказчика синтетическом датасете на обезличенных данных.

7.4. Подготовка выборок для соревнования: обучающая, открытая валидационная, приватная тестовая и др. необходимые. Проверка качества полученных выборок с точки зрения воспроизводимости решений на разных её частях с сохранением качества, отсутствия утечек целевой переменной, достаточности объема данных.

7.5. Выбор оптимальных целевых метрик, обеспечивающих оптимальную чувствительность к улучшению решений участниками соревнования, разработка протокола валидации и тестирования, имплементация протокола и самих метрик под задачу.

7.6. Прорешивание задачи, поиск и устранение потенциальных уязвимостей в постановке задачи. Проработка основных путей решения задач для определения общей решимости задач и возможностей для участников соревнования по созданию качественных решений

7.7. Подготовка базовых решений и «starter kit» для участников.

7.8. Подготовка описания задачи и сопроводительных материалов, и документов.

**8. Требования к метрикам**

8.1. Итоговый выбор метрик должен быть обоснован и согласован с целями соревновательного проекта.

* 1. Процесс выбора метрик должен обеспечить согласованность итоговых метрик с ожиданиями по общему ходу решений участников.
	2. Ожидаемые участниками улучшения их решений по ходу соревнования должны быть согласованы с целями и концепцией всего соревновательного проекта.

8.2. Платформа соревнования должна поддерживать широкий функционал по использованию от 2 метрик для задач соревнования.

1. Должны быть возможности добавления произвольных метрик, включая составные, технические, турнирные и другие.
2. Требуется наличие возможностей по гибкому управлению и наращиванию числа метрик в рамках одной задачи.

**9. Требования к тестированию задач и проработке возможных уязвимостей**

При проработке задач Исполнителю необходимо избежать все перечисленные уязвимости:

9.1. Уязвимости, связанные с данными:

* Злоупотребление ручной разметкой тестовых данных;
* «Пробивание» лидерборда;
* Парсинг тестовых данных извне;
* Переиспользование данных с предыдущих соревнований;
* Утечки целевой переменной в данных.

9.2. Уязвимости, связанные с проверочной системой:

* Уязвимости, связанные с манипуляциями с размеченным табличным тестовым файлом;
* Уязвимости, связанные с доступом к проверочной системе;
* Уязвимости, связанные с используемыми образами.

9.3. Прочие уязвимости, которые могут возникнуть по ходу соревнований:

Поведение участников:

* Регистрация дополнительных аккаунтов для получения большего числа попыток, либо попыток пользования уязвимостями выше;
* Списывание или непубличный обмен решениями вне своих команд;
* Игнорирование правил соревнования;
* Потенциальные проблемы с используемой участниками интеллектуальной собственностью.

**10. Требования к размещению основной соревновательной задачи**

10.1. Размещение на платформе, соответствующей целям и задачам проекта основной соревновательной задачи по «NLP/LLM. Большие языковые модели». Количество участников от 1500 до 5500.

10.2. Техническая поддержка по вопросам работоспособности проверяющей системы, а также платформы с интерфейсом для пользователей. Обеспечение оперативного решения вопросов участников:

- в период с 9:00 до 22:00 в течение 1 часа;

- вопросы участников, полученные в период c 22:00 до 9:00, должны быть решены не позднее 11:00.

10.3. Размещение FAQ для участников соревнования

**11. Требования к размещению дополнительной соревновательной задачи**

11.1. Размещение на платформе, соответствующей целям и задачам проекта дополнительной соревновательной задачи по теме «Модели оттока».

11.2. Техническая поддержка по вопросам работоспособности проверяющей системы, а также платформы с интерфейсом для пользователей. Обеспечение оперативного решения вопросов участников:

- в период с 9:00 до 22:00 в течение 1 часа;

- вопросы участников, полученные в период c 22:00 до 9:00, должны быть решены не позднее 11:00.

11.3. Размещение FAQ для участников соревнования

**12. Подготовка сопроводительных документов**

12.1. Подготовка юридически корректных форм согласия на обработку и передачу персональных данных участников для организации коммуникаций Исполнителя с Участниками и последующей коммуникации Заказчика с участниками.

12.2. Организация, сбор и передача Заказчику персональных данных участников для последующей коммуникации с ними, в том числе с целью трудоустройства. Вся информация должна собираться, храниться и передаваться в соответствии законом № 152-ФЗ «О персональных данных».

12.3. Подготовка юридически и технически корректных Правил участия в соревновании.

**13. Требования к подготовке базового информационного контента**

13.1. Подготовка и согласование с Заказчиком полного комплекта текстов информационного наполнения (контента) и размещение на платформе соревнования, и странице соревнования (https://data-fusion.ru,https://data-fusion.ru/online-competition/), включая раздел FAQ

13.2. Подготовка, согласование с Заказчиком и размещение обновления информации для лендинга соревнования не позднее чем за 2 недели до старта.

13.3. Интеграция лидерборда с платформы соревнования на страницу соревнования сайта проекта Data Fusion.

13.4. Подготовка итоговых материалов:

* 1. Подготовка, согласование с Заказчиком пресс-релиза и пост-релиза о результатах соревнования для последующего размещения на ресурсах Исполнителя и Заказчика, а также внешних ресурсах.
	2. Материалы с результатами соревнования для мероприятий в рамках конференции Data Fusion – разбор решений и торжественной церемонии награждения (презентация в формате PDF).

**14. Размещение анонсов на собственных площадках**

14. Публикации анонсов и поддерживающих публикаций в сообществах, в соцсетях, в каналах Telegram, email-рассылка по базе Исполнителя.

**15. Подготовка аналитического материала**

15. Подготовка итогового материала с полным разбором соревнования для Блога Заказчика на Habr (не менее 12 000 знаков).

**16. Работа с участниками Соревнования**

Организация чата с участниками включает в себя следующие требования:

16.1. Создание и координация чата с участниками соревнования (проведение соревнования, подведения итогов).

16.2. Сопровождение чата: техническая и организационная поддержка участников 24/7.

**17. Требования к организации онлайн-митапов**

17. Организация не менее 2 (двух) онлайн-митапов (вебинаров) с разбором датасета соревнования, задач и подходов, текущих вопросов участников соревнования после первого месяца соревнования. Привлечение участников онлайн-митапа.

**18. Участие в рамках деловой программы конференции Data Fusion 2025**

Участие в рамках деловой программы конференции Data Fusion 2025 включает в себя следующие требования:

18.1. Подготовка сессии (1,5 часа) по тематике соревнования в рамках конференции Data Fusion (апрель 2025, data-fusion.ru).

18.2. Участники сессии – победители соревнования с докладами о разборе решений; а также профильные специалисты из индустрии по тематике задач («NLP/LLM. Большие языковые модели» и «Модели оттока»).