



ОТКРЫТАЯ НАУКА  
издательство

Международный журнал информационных технологий и  
энергоэффективности

Сайт журнала: <http://www.openaccessscience.ru/index.php/ijcse/>



УДК 519.23:004.94:339.13

## СИСТЕМНЫЙ ПОДХОД В ПРОГНОЗИРОВАНИИ РЕЗУЛЬТАТОВ СПЛИТ-ТЕСТИРОВАНИЯ

**Жданов Я.Д.**

*ФГБОУ ВО МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ТЕХНОЛОГИЙ И УПРАВЛЕНИЯ ИМЕНИ К.Г. РАЗУМОВСКОГО ", Москва, Россия, (109004, город Москва, ул. Земляной Вал, д.73 ), e-mail: yaroslavd.zhdanov@gmail.com*

В статье рассматриваются основные инструменты системного анализа, позволяющие наиболее эффективно выполнить проект по разработке алгоритма машинного обучения для прогнозирования результатов сплит-тестирования. Среди таких инструментов применяются следующие: диграф связи проблем, X матрица Хосин Канри и древовидная диаграмма для ключевой проблемы, выявленной благодаря диграфу связи. Результаты анализа позволяют выстроить последовательный список связанных шагов, позволяющих как можно быстро и эффективно выполнить основные задачи поставленной цели.

Ключевые слова: Системный анализ, сплит-тестирование, машинное обучение, декомпозиция задач, анализ ключевых проблем.

## A SYSTEMATIC APPROACH TO FORECASTING SPLIT TESTING RESULTS

**Zhdanov Ya.D.**

*MOSCOW STATE UNIVERSITY OF TECHNOLOGIES AND MANAGEMENT NAMED AFTER K.G. RAZUMOVSKY ", Moscow, Russia, (109004, Moscow, Zemlyanoy Val str., 73 ), e-mail: yaroslavd.zhdanov@gmail.com*

The article discusses the main tools of system analysis that make it possible to most effectively carry out a project to develop a machine learning algorithm for predicting split-test results. Among such tools, the following are used: the problem communication digraph, the Hoshin Kanri X matrix, and the tree diagram for the key problem identified by the communication digraph. The analysis results allow you to build a consistent list of related steps that allow you to quickly and efficiently accomplish the main objectives of the goal.

Keywords: System analysis, split testing, machine learning, task decomposition, key problem analysis.

### Введение

Внедрение алгоритмов прогнозирования в процесс сплит-тестирования является одним из ключевых инструментов цифровой оптимизации. Алгоритм позволит с высокой точностью и достоверностью прогнозировать результаты будущих тестов, что позволит снизить издержки и ускорить процесс принятия решений. Однако результат разработки подобного инструмента сильно зависит от большого количество факторов, среди которых качество сырых данных, правильно подобранный критерий оценки качества модели, интерпретация результатов.

В данной работе описываются и применяются некоторые инструменты системного анализа, которые позволили определить ключевые проблемы проекта, понять стратегические приоритеты и декомпонировать основную проблему.

Основной проблемой проекта является низкое качество и нерепрезентативность данных. Такое вывод можно сделать благодаря построению диграфа связи проблем, где были выявлены все основные проблемы. Приоритетность выполнения задач и корреляции со стратегией и тактикой помогла определить Х матрица Хосин Канри. Древовидная диаграмма в свою очередь наглядно показала проблемы, из которых состоит основная сложность проекта.

### **1. Ключевые проблемы проекта**

Диграф связи проблем – это инструмент визуализации, который отображает причинно-следственные связи между различными проблемами в системе. Данный инструмент позволяет выявить взаимозависимости проблем и определить среди них ключевую, устранение которой воздействует на полное или частичное решение связанных сложностей. Это преимущества позволяют с помощью диграфа связи проблем строить наиболее эффективную стратегию по выполнению проектов.[1]

Среди основных проблем проекта «Прогнозирование результат сплит-тестирования с использованием алгоритмов машинного обучения» можно выделить:

- Низкое качество и нерепрезентативность данных

Основная проблема в разработке алгоритма – это качество данных и низкая репрезентативность данных.

Качество данных связано с различным шумом в исходных данных, который отражается в виде дублирования некоторых строк или определённых значений, пропуски в данных, ненормализованных значений и других критериев.

Низкая репрезентативность выражается в исторических данных, которые не отражают текущие поведенческие паттерны.

Сложность заключается в том, что если «мусор на входе», то будет и «мусор на выходе».

- Смещающий отбор и несбалансированность групп

Данные для обучения могут быть собраны только среди одного типа устройств или одного региона, например. В таком случае результаты прогноза можно будет экстраполировать только на группу, которая участвовала в тестировании.

- Утечка данных и переобучение модели

В моменте валидации модели информация из тестовых данных может попасть в обучающие данные, что приведёт к переобучению модели, когда модель хорошо отрабатывает на исторических данных, но выдаёт плохие результаты для новых.

- Некорректный выбор метрик и признаков

От выбора метрики зависит ценность результатов теста, поэтому при неправильном выборе метрик есть вероятность неправильно интерпретировать результаты, которые не будут отражать конечные цели бизнеса.

- Неучёт временных эффектов и зависимости от контекста [2]

Алгоритм не будет учитывать определённую сезонность и «эффект новизны» от изменения или устаревания пользовательского поведения. Помимо этого, модель не будет адаптироваться под дополнительные факторы, которые могли повлиять на изменения поведения (например, запуск рекламной кампании).

- Низкая интерпретируемость результатов модели

В большом количестве случаев нейросети генерируют «черные ящики». Это означает, что нельзя точно определить на основании каких факторов искусственный интеллект сделал определённые выводы. Сложность заключается в том, что заказчики не доверяют полученным результатам, которые не могут понять.

- Ошибочное прогнозное решение и выбор победителя

Из-за любой ошибки в предыдущих проблемах в конечном результаты могут быть искажены и принятие решений со стороны бизнеса в данном случае может повлечь за собой определённые финансовые потери.

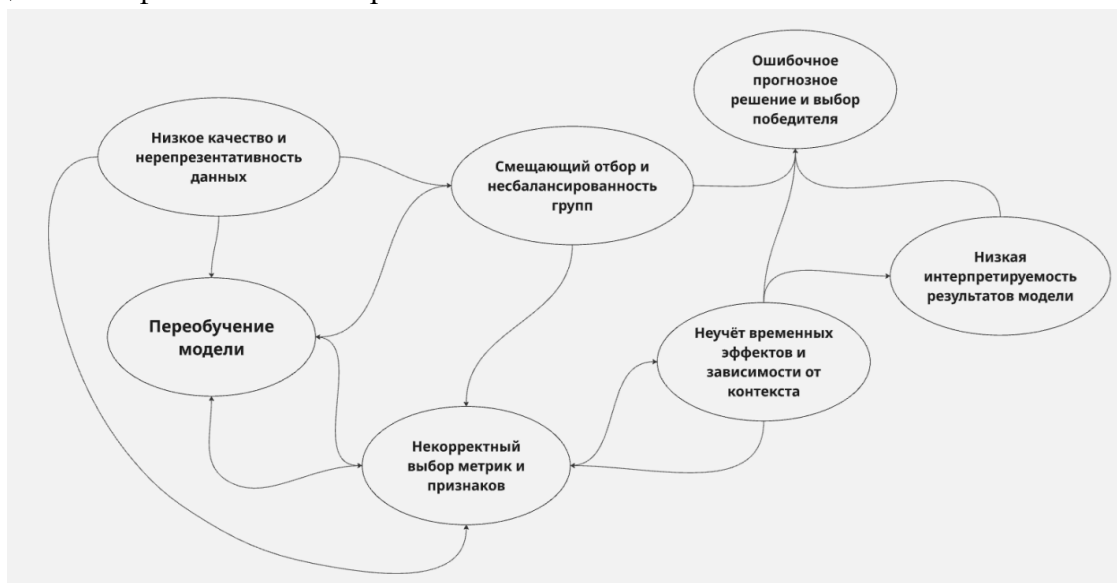


Рисунок 1 - Диграф связи проблем

## 2. Анализ стратегических приоритетов проекта

Применение X матрицы Хосин Канри позволяет установить чёткие связи и проследить за корреляцией между основными направлениями разработки проекта.

Стратегические цели проекта:

- Сокращение временных и ресурсных затрат на проведение А/В-тестов;
- Повышение точности и достоверности принимаемых решений на основе данных;
- Создание конкурентного преимущества за счет опережающего анализа гипотез.

Благодаря матрице можно сделать вывод, что основной тактической задачей является «Разработка и внедрение ML-модели для прогнозирования результатов А/В-теста», которая взаимосвязана со стратегией 1 и 2. Дополнительно необходимо учесть, что «Создание автоматизированного пайплайна для сбора и предобработки исторических данных о тестах» так же связана со стратегией 1 и это доказывает важность оптимизации процессов на этапе сбора и предобработки данных.

Взаимосвязь тактических задач и операционных процессов.

Результаты анализа полученной матрицы указали на связь между «Интеграцией системы прогнозирования в процесс принятия решений о запуске тестов» и «Постановкой задачи», что свидетельствует о необходимости учета требования интеграции уже на начальных этапах проектирования системы.[3]

Распределение ролевой ответственности и метрики успеха.

Основная доля ответственности ложится на разработчика, которому нужно выполнить технические задачи, а менеджер в свою очередь является посредником между заказчиком и остальной командой.

Система метрик успеха проекта демонстрирует сбалансированность между техническими и бизнес-показателями. Наиболее значимое влияние на ключевую метрику точности прогнозирования оказывают процессы работы с данными и моделирования, что соответствует лучшим практикам в области машинного обучения.

В итоге с помощью матрицы можно проследить взаимодействие всей системы в рамках стратегического управления, что позволяет наиболее эффективно обеспечить выполнение проекта.

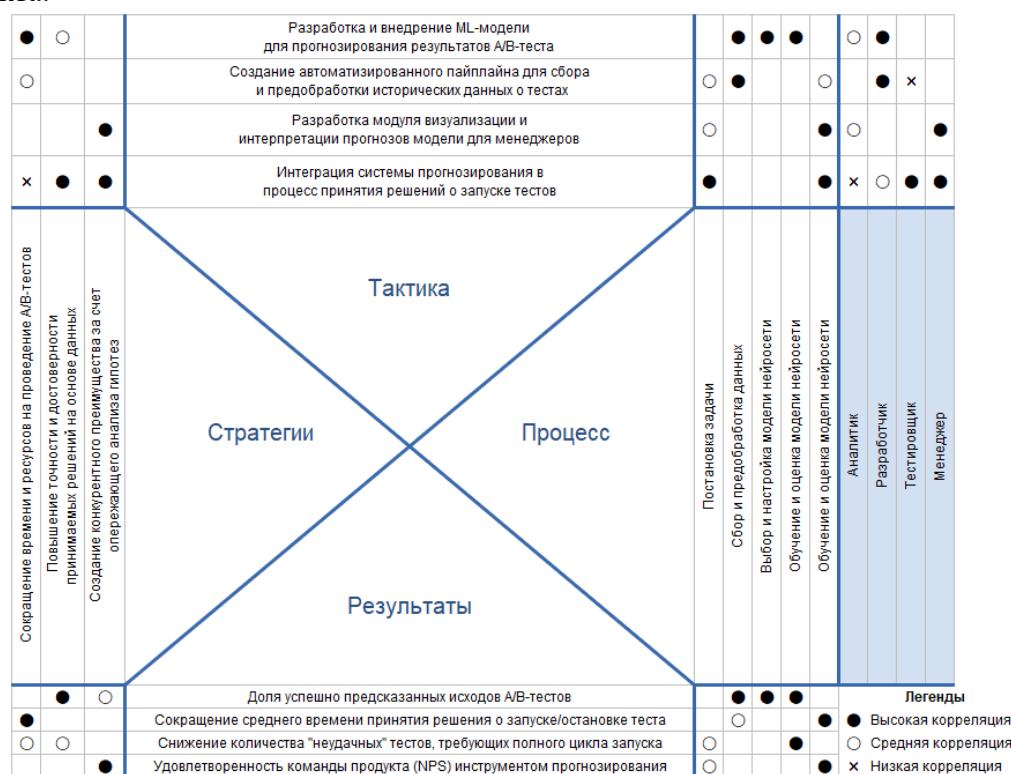


Рисунок 2 - X матрица Хосин Канри

### 3. Декомпозиция ключевой проблемы

После определения ключевой проблемы с помощью диграфа связи эту проблему необходимо решать в первую очередь, так как в большинстве случаев от её решения зависит основной результат выполняемого проекта. Для удобства решения проблемы её необходимо декомпозировать, это можно сделать с помощью инструмента «Древовидная диаграмма проблемы».

Древовидная диаграмма позволяет последовательно разбить объёмную и сложную проблему на маленькие составляющие, от общего к частному. С помощью данной диаграммы можно не только упростить решение проблемы, но и учесть все составляющие факторы, которые могут повлиять на результат.

В рамках построения древовидной диаграммы проблема «Низкое качество и нерепрезентативность данных» имеет две подпроблемы, каждая из которых имеет по три подпроблемы

1. Проблемы целостности и корректности данных отражают техническую составляющую, включающую корректность, точность анализируемых данных.

1.1. Технические сбои и ошибки регистрации

Такой сценарий может возникнуть при некорректной работе систем сбора данных, сбоев в коде сплитования трафика в рамках проводимого теста.

1.2. Аномальные данные и дублирование

Аномальные выбросы в числовых данных, дублирование строк или некоторых значений в определённых столбцах сырых данных, некорректные значения в категориальных признаках.

1.3. Низкая информационная насыщенность признаков

Отсутствие дополнительных данных, которые могут повысить потенциал прогнозирования алгоритма

2. Проблемы репрезентативности и контекста - Группа проблем из этого направления связаны с методологией сбора данных тестов

2.1. Смещение и несбалансированность выборки

Выборки в рамках распределения трафика имеют разное количество пользователей в процентном соотношении, данные собраны не со всей генеральной совокупности

2.2. Недостаточный объём данных для анализа

Низкое количество трафика, которого недостаточно для появления статистической значимости или принятия решения.

2.3. Временные искажения и эффекты взаимодействия

Сезонность теста и проведение параллельных сплит-тестов, выборки которых пересекаются с текущим тестом и тем самым искажают данные.



Рисунок 3 - Древовидная диаграмма ключевой проблемы

## Заключение

Результаты проведённого исследования указывают на то, что самым важным фактором в реализации проекта по прогнозированию результатов сплит-тестирования является решение проблемы низкого качества и нерепрезентативности данных.

Благодаря инструментам декомпозиции, структуризации и анализа приоритетов можно сделать вывод, что разработка алгоритма машинного обучения для прогноза должна быть связана с созданием последовательного плана сбора и предобработки данных.

Выстраивание подобного систематического подхода позволяет повысить успех принимаемых решений, сократить время на создание пайплайна проекта и создать конкурентное преимущество.

### **Список литературы**

1. Блауберг Игорь Викторович, Садовский Вадим Николаевич, Юдин Эрик Григорьевич. Проблемы методологии системного исследования / Мысль, 1970. – С. 205–231.
2. Перегудов Феликс Иванович, Тарасенко Феликс Петрович. Введение в системный анализ / Высшая школа, 1989. – С. 67–92.
3. Молотникова Антонина Александровна. Системный анализ. Краткий курс / Лань, 2019.

### **References**

1. Igor Viktorovich Blauberg, Vadim Nikolaevich Sadovsky, Erik Grigorievich Yudin. Problems of Systems Research Methodology / Mysl', 1970. – pp. 205–231.
  2. Felix Ivanovich Peregudov, Felix Petrovich Tarasenko. Introduction to Systems Analysis / Vysshaya shkola, 1989. – pp. 67–92.
  3. Antonina Aleksandrovna Molotnikova. Systems Analysis. A Brief Course / Lan', 2019.
-