

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ/SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

DOI: https://doi.org/10.60797/IRJ.2026.168.1 EDN: CQYGGC

АНАЛИЗ ВЛИЯНИЯ КОНТЕКСТА ПРОЕКТА НА МОДЕЛЬ ПРОГНОЗИРОВАНИЯ ТРУДОЗАТРАТ

Научная статья

Цыварев И.В.[1, *], Филиппов Ф.В.[2]

[1, 2] Санкт-Петербургский государственный университет телекоммуникаций имени профессора М.А. Бонч-Бруевича, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (cyvarev.ilya156[at]gmail.com)

Предложена: 14.03.2026; Принята: 04.06.2026; Опубликована: 17.07.2026

Аннотация

Целью исследования является анализ переносимости регрессионных моделей, обученных на данных, полученных из многокомпонентных систем. В качестве исследуемого параметра состояния системы выступают трудозатраты на выполнение элементов (задач) в подсистемах (проектах разработки программного обеспечения). Методология эксперимента базируется на открытом наборе данных TAWOS (более 450 тыс. наблюдений из 39 подсистем) и включает сравнение двух регрессионных моделей — градиентного бустинга XGBoost и нейронной сети прямого распространения — как с использованием PCA-эмбеддингов текстовых описаний задач, так и без них. Оценка эффективности проводится в двух сценариях, моделирующих различную степень репрезентативности обучающей выборки относительно генеральной совокупности: обучение на смешанных данных всех подсистем; обучение и тестирование в рамках одной изолированной подсистемы. Статистическая неоднородность данных проверяется с помощью теста Левина и однофакторного дисперсионного анализа (ANOVA).

Ключевые результаты подтвердили статистическую неоднородность данных: для всех числовых признаков выявлены значимые различия как дисперсий, так и средних между проектами ($p < 0,001$). Модель, обученная на смешанных данных, показала низкое качество прогноза (MAE = 3,37, Accuracy = 28,67% для лучшей конфигурации XGBoost без эмбеддингов). При обучении и тестировании в рамках изолированных подсистем метрики существенно возрастают, однако демонстрируют высокую вариативность (MAE от 0,53 до 5,91; Accuracy от 14,29% до 54,76%), что отражает «статистическую индивидуальность» каждого проекта. Сравнение моделей выявило преимущество XGBoost на большинстве проектов при отсутствии универсального лидера; добавление текстовых эмбеддингов не дало однозначного улучшения.

Ключевые слова: гетерогенность данных, инвариантность модели, многокомпонентные системы, обработка информации, доменная адаптация, регрессионный анализ, нейронные сети.

ANALYSIS OF THE INFLUENCE OF THE PROJECT CONTEXT ON THE LABOUR COST FORECASTING MODEL

Research article

Tsyvarev I.V.[1, *], Filippov F.V.[2]

[1, 2] Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russian Federation

* Corresponding author (cyvarev.ilya156[at]gmail.com)

Suggested: 14.03.2026; Accepted: 04.06.2026; Published: 17.07.2026

Abstract

The aim of the study is to analyse the transferability of regression models trained on data obtained from multi-component systems. The studied parameter of the system's state is the labour costs associated with completing elements (tasks) within subsystems (software development projects). The experimental methodology is based on the open-source TAWOS dataset (over 450,000 observations from 39 subsystems) and involves a comparison of two regression models — the gradient boosting method XGBoost and a feedforward neural network — both with and without the use of PCA embeddings of textual task descriptions. Performance evaluation is carried out in two scenarios simulating different degrees of representativeness of the training sample relative to the population: training on mixed data from all subsystems; training and testing within a single isolated subsystem. Statistical heterogeneity of the data is tested using the Levene's test and one-way analysis of variance (ANOVA).

The key results confirmed the statistical heterogeneity of the data: for all numerical features, significant differences were identified in both variances and means across projects ($p < 0.001$). The model trained on mixed data demonstrated poor prediction quality (MAE = 3.37, Accuracy = 28.67% for the best XGBoost configuration without embeddings). When trained and tested within isolated subsystems, the metrics improve significantly but exhibit high variability (MAE ranging from 0.53 to 5.91; Accuracy from 14.29% to 54.76%), reflecting the 'statistical individuality' of each project. A comparison of models showed the advantage of XGBoost on most projects, with no universal leader; the addition of text embeddings did not yield a definite improvement.

Keywords: data heterogeneity, model invariance, multi-component systems, information processing, domain adaptation, regression analysis, neural networks.

Введение

Процесс разработки программного обеспечения может быть рассмотрен как сложная организационно-техническая система, состоящая из множества взаимодействующих подсистем (проектов, доменов, команд). Каждая такая подсистема генерирует потоки данных, характеризующих её состояние и динамику, включая метрики выполнения задач (трудозатраты), которые являются ключевым параметром для анализа эффективности системы в целом.

При построении математических моделей для обработки информации и прогнозирования состояния подобных систем широко используется подход, основанный на объединении данных от различных подсистем в единый обучающий набор данных. Такой подход подразумевает статистическую гомогенность (однородность) этих данных и предполагает, что модель способна выделить инвариантные закономерности, общие для всего набора. Однако на практике каждая подсистема обладает уникальными характеристиками: спецификой решаемых задач, распределением ролей, внутренними процессами и метриками учета.

Целью настоящего исследования является проверка гипотезы о гомогенности данных в многокомпонентной системе разработки ПО и оценка степени влияния доменного контекста (принадлежности к конкретной подсистеме) на метрики и инвариантность регрессионной модели машинного обучения. В работе решается задача количественной оценки вариативности характеристик модели при смене домена, что имеет фундаментальное значение для развития методов обработки информации в условиях гетерогенных источников данных. Полученные результаты направлены на обоснование необходимости предварительного анализа структуры данных и применения методов доменной адаптации как обязательного этапа системного анализа.

Обзор предметной области

Задача прогнозирования трудозатрат при разработке программного обеспечения остаётся одной из наиболее актуальных в области программной инженерии. По данным масштабных исследований, охватывающих тысячи крупных ИТ-проектов, значительная их доля завершается с существенным превышением бюджета и сроков. Согласно анализу статистики управления проектами за 2023 год, около 47% Agile-проектов сталкиваются с задержками, перерасходом средств или снижением удовлетворённости заказчика [11]. Это свидетельствует о сохраняющейся практической значимости разработки надёжных методов оценки трудозатрат.

В современных Agile-проектах оценка трудозатрат традиционно выражается в единицах Story Points, отражающих относительную сложность задачи. Обзорные работы последних лет фиксируют устойчивый переход от экспертных методов (Planning Poker, оценка по аналогии) к автоматизированным подходам на основе машинного обучения [12], [13]. Uc-Cetina (2023) в обзоре современных методов машинного обучения для оценки трудозатрат подчёркивает, что накопление корпоративных данных о проектах открывает возможности для построения персонализированных прогностических моделей, однако межпроектная обобщаемость таких моделей остаётся ограниченной [12].

Среди ансамблевых методов, применяемых для оценки Story Points, особое место занимают алгоритмы градиентного бустинга. Широкомасштабное сравнительное исследование Yalciner et al. (2024), опубликованное в журнале Applied Sciences, продемонстрировало, что гибридная модель на основе SBERT и алгоритмов градиентного бустинга достигает MAE = 2,15 и MdAE = 1,85, превосходя ряд существующих методов [11]. Систематический литературный обзор Cabral et al. (2023), опубликованный в Journal of Systems and Software, зафиксировал, что ансамблевые методы стабильно превосходят одиночные алгоритмы на задачах оценки трудозатрат [13]. Применение текстовых эмбеддингов для обогащения признакового пространства исследовалось в работе Molla (2024): авторы установили, что предобученные представления слов (FastText, SBERT) повышают точность прогноза на одних наборах данных, однако на других наборах заметного улучшения не наблюдается [14].

Проблема гетерогенности данных и доменной адаптации в контексте программной инженерии рассматривается в ряде актуальных публикаций. Singhal (2023) в обзоре «Domain Adaptation: Challenges, Methods, Datasets and Applications», опубликованном в IEEE Access, систематизируют подходы к сокращению доменного сдвига, включая методы согласования признакового пространства, перевзвешивания экземпляров и многозадачного обучения [15]. Авторы указывают, что игнорирование статистической неоднородности при объединении данных из разных источников закономерно ведёт к снижению обобщающей способности модели. Параллельно исследования по прогнозированию Story Points с использованием гетерогенных графовых нейронных сетей (HeteroSP, 2022) продемонстрировали ценность учёта структурных взаимосвязей между задачами проекта, однако также зафиксировали существенную зависимость качества прогноза от доменного контекста [16].

Таким образом, анализ современной литературы свидетельствует о следующем. Во-первых, задача автоматической оценки трудозатрат в Agile-проектах остаётся актуальной и широко исследуемой, при этом методы машинного обучения последовательно вытесняют экспертные подходы. Во-вторых, проблема межпроектной переносимости моделей и статистической гетерогенности данных признаётся исследовательским сообществом существенной, однако комплексный количественный анализ её проявлений на крупных многопроектных наборах данных (таких как TAWOS) представлен в литературе недостаточно полно. Настоящая работа заполняет этот пробел, предоставляя систематические эмпирические свидетельства статистической индивидуальности проектов и количественно оценивая снижение качества прогнозирования при игнорировании доменного контекста.

Методы исследования

В качестве объекта исследования выбрана многокомпонентная система, представленная открытым набором данных TAWOS, который содержит сведения о более чем 450 тысячах задач из 39 различных проектов по разработке программного обеспечения [1], [2], [3]. Каждое наблюдение (задача) описывается набором признаков, отражающих её атрибуты и активность в рамках подсистемы.

В таблице 1 представлены данные о признаках задач, использованных для обучения регрессионной модели.

Таблица 1 - Признаки набора данных

DOI: https://doi.org/10.60797/IRJ.2026.168.1.1

Колонка	Тип	Описание
Project_ID	Целое число	Идентификатор проекта
Creation_Date	Дата	Дата создания задачи
Description	Текст	Текстовое описание задачи
Type	Целое число	Тип (Задача, ошибка, история, анализ, тестирование)
Priority	Целое число	Приоритет (Самый низкий, низкий, средний, высокий, критический)
Reporter_ID	Целое число	Идентификатор создателя задачи
Assignee_ID	Целое число	Идентификатор исполнителя задачи
Sprint_ID	Целое число	Идентификатор спринта
Story_Point	Целое число	Целевая переменная — условная единица измерения времени выполнения задачи

Для обучения моделей были взяты только те данные, у которых заполнено значение Story_Point (всего 59906 записей). Кроме того, из выборки удален проект с идентификатором 32, поскольку лишь 6 из 232 имеют значение, отличное от 1.

Методология эксперимента построена как исследование сходимости регрессионной модели в условиях различной степени репрезентативности обучающей выборки. Алгоритм обработки информации и построения модели представлен на рисунке 1.

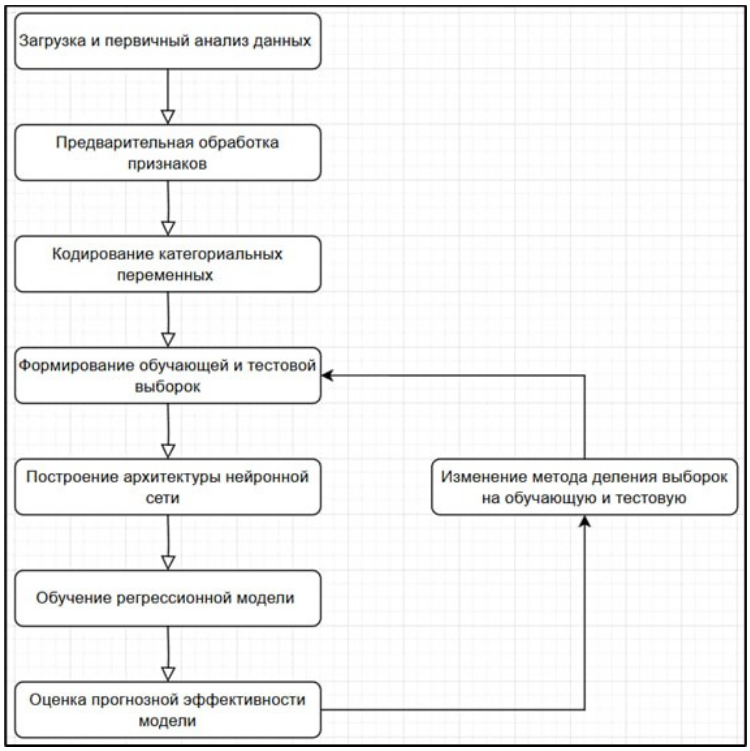

Рисунок 1 - Алгоритм экспериментального исследования влияния гетерогенности данных на метрики регрессионной модели

DOI: https://doi.org/10.60797/IRJ.2026.168.1.2

На этапе предварительной обработки выполняется унификация входного пространства признаков. Категориальные переменные с умеренной кардинальностью преобразуются методом LabelEncoder [4], [5], [6], [7]. После обработки данных с использованием LabelEncoder в наборе данных осталось 9 признаков.

Были реализованы и сравнены две регрессионные модели для оценки Story Point’ов по задачам: градиентный бустинг (XGBoost и CatBoost) и нейронная сеть (MLP). Для XGBoost и CatBoost использовались фиксированные гиперпараметры: 300 деревьев решений, максимальная глубина 9, скорость обучения 0.1, функция потерь — среднеквадратичная ошибка. Для нейронной сети была использована следующая архитектура: входной слой (размерность признакового пространства), два скрытых слоя на 128 и 64 нейрона соответственно, выходной линейный слой [8], [9]. Выбранная архитектура с двумя скрытыми слоями является компромиссом для табличных данных средней размерности согласно практическим рекомендациям, для таких задач достаточно 2–3 слоя для улавливания нелинейностей без риска переобучения на малых проектах; предварительный подбор гиперпараметров (число слоёв, количество нейронов) не показал значимого выигрыша от усложнения сети.

Обе модели обучались на двух наборах признаков — базовом (без эмбеддингов) и расширенном, включающем эмбеддинги текстовых описаний задач, полученные с помощью языковой модели bge-small-en-v1.5, и сжатые с помощью метода PCA [10].

Перед обучением для каждого проекта независимо удалялись выбросы целевой переменной: отсекались значения ниже 5-го и выше 95-го процентилей. Обучение выполнялось отдельно по каждому проекту, с разбиением выборки на обучающую (80%) и тестовую (20%). Качество оценивалось по метрикам MAE и accuracy (доле предсказаний, попадающих в интервал ±25% от истинного значения). Сравнение двух моделей на проектах с разным объёмом задач позволяет оценить вклад текстовых эмбеддингов и устойчивость каждого алгоритма к шуму и малым выборкам.

Эксперимент включает два сценария:

1. Оценка на смешанной выборке: стандартное разделение всех данных в пропорции 80/20 для обучения и тестирования. Этот сценарий моделирует ситуацию, когда данные всех подсистем рассматриваются в единой совокупности.

2. Оценка в рамках изолированной подсистемы: для каждого проекта в наборе данных модель обучается и тестируется исключительно на её внутренних данных (разбиение 80/20).

Основные результаты

Результаты первого сценария представлены в таблице 2.

Таблица 2 - Результаты моделей, обученных по сценарию 1

DOI: https://doi.org/10.60797/IRJ.2026.168.1.3

Модель	MAE	Accuracy
XGBoost (без эмбеддингов)	3,37	28,67
XGBoost (с эмбеддингами)	3,92	27,61
CatBoost (без эмбеддингов)	3,52	17,40
CatBoost (с эмбеддингами)	4,01	16,12
Нейронная сеть (без эмбеддингов)	4,67	15,52
Нейронная сеть (с эмбеддингами)	4,46	18,67

Как можно увидеть, модели, построенные в рамках первого сценария, показали низкие значения метрик, что свидетельствует об их неприменимости для прогнозирования трудозатрат по задачам.

Однако результаты второго сценария (табл. 3) выявили принципиально иную картину. При обучении и тестировании в рамках изолированных подсистем качество предсказаний ожидаемо возрастает, но главное — обнаруживается высокая вариативность метрик между проектами. Значения MAE варьируются от 0,53 (проект 31) до 5,91 (проект 34) для градиентного бустинга без эмбеддингов, а accuracy — от 14,29% (проект 17) до 54,76% (проект 31).

Таблица 3 - Результаты моделей, обученных по сценарию 2

DOI: https://doi.org/10.60797/IRJ.2026.168.1.4

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
1	3427	1,55	29,74	1,61	30,90	1,58	28,28	1,52	32,22	1,72	25,66	1,67	29,74
2	205	1,74	21,95	1,58	26,83	1,81	21,95	1,69	19,51	1,76	26,83	1,44	21,95

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
3	1763	0,85	32,01	0,79	33,71	0,74	31,73	0,73	32,29	0,75	31,73	0,74	32,01
4	3206	1,36	34,74	1,34	31,78	1,29	35,98	1,29	34,42	1,36	35,20	1,31	34,74
5	470	0,92	51,06	0,79	54,26	0,79	55,32	0,76	59,57	0,92	50,00	0,89	51,06

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
6	197	1,98	42,50	2,08	35,00	1,90	37,50	1,83	45,00	1,77	42,50	2,07	42,50
7	292	1,54	40,68	1,64	30,51	1,51	30,51	1,47	32,20	1,57	33,90	1,41	40,68
8	819	3,32	31,10	3,21	41,46	3,03	40,85	3,28	42,07	3,77	35,37	3,98	31,10

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
9	350	1,63	30,00	1,72	24,29	1,47	41,43	1,45	41,43	1,57	31,43	1,58	30,00
10	229	4,47	21,74	3,98	23,91	3,95	32,61	3,75	30,43	3,96	26,09	4,48	21,74
11	1219	1,92	26,23	1,87	33,20	1,89	28,69	1,81	30,74	2,12	25,41	2,07	26,23

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
12	4191	2,22	32,90	2,22	35,88	2,17	35,40	2,20	35,76	2,39	30,99	2,36	32,90
13	3257	1,83	44,17	1,74	48,31	1,72	46,78	1,74	46,93	1,85	45,40	1,90	44,17
14	499	1,04	35,00	1,02	30,00	0,92	40,00	0,91	38,00	0,96	36,00	1,00	35,00

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
15	359	2,34	22,22	2,16	25,00	2,09	15,28	1,94	25,00	2,56	8,33	2,05	22,22
16	209	1,01	23,81	0,84	21,43	0,83	28,57	0,82	26,19	0,83	30,95	0,88	23,81
17	209	1,25	7,14	1,18	16,67	1,49	16,67	1,24	11,90	1,33	11,90	1,45	7,14

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
18	361	1,78	32,88	1,60	36,99	1,31	54,79	1,30	54,79	1,75	43,84	1,69	32,88
19	281	1,90	28,07	1,71	31,58	1,56	31,58	1,71	31,58	1,55	24,56	1,67	28,07
20	309	1,67	37,10	1,85	27,42	1,66	30,65	1,72	30,65	1,64	37,10	1,68	37,10

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
21	333	1,26	37,31	1,05	50,75	1,09	46,27	1,00	52,24	1,18	44,78	1,25	37,31
22	639	1,25	46,09	1,21	46,88	1,16	42,97	1,16	42,97	1,44	42,19	1,28	46,09
23	346	0,90	48,57	0,87	44,29	0,89	40,00	0,82	52,86	0,90	47,14	0,90	48,57

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
24	613	1,17	30,08	1,11	48,78	1,04	42,28	1,00	45,53	1,38	20,33	1,17	30,08
25	672	1,66	29,63	1,56	34,81	1,47	35,56	1,51	35,56	1,56	34,07	1,64	29,63
26	695	1,57	33,09	1,52	38,13	1,58	32,37	1,50	33,09	1,66	31,65	1,61	33,09

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
27	946	0,91	43,16	0,91	40,00	0,83	44,21	0,84	40,53	0,87	43,16	0,91	40,00
28	18412	2,74	24,52	2,69	25,20	2,81	21,94	2,74	23,76	3,08	19,90	3,02	21,97
29	635	0,86	40,16	0,79	33,86	0,84	31,50	0,77	35,43	0,82	28,35	0,74	31,50

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
30	499	1,45	36,00	1,42	38,00	1,33	39,00	1,21	45,00	1,34	39,00	1,28	34,00
31	208	0,53	54,76	0,55	47,62	0,56	42,86	0,50	47,62	0,67	28,57	0,64	40,48
33	725	1,27	31,72	1,23	29,66	1,16	32,41	1,14	35,86	1,23	31,03	1,22	33,10

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
34	1451	5,91	26,46	5,64	25,77	5,33	24,40	5,37	23,37	6,24	19,59	6,15	19,59
35	451	1,76	26,37	1,51	32,97	1,58	31,87	1,48	26,37	1,70	25,27	1,67	28,57
36	3556	2,19	28,23	2,16	29,21	2,14	31,18	2,09	29,49	2,23	28,79	2,19	28,09

Project_ID	Кол-во задач	XGBoost				CatBoost				Нейронная сеть			
		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами		Без эмбеддингов		С эмбеддингами	
		MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy	MAE	Accuracy
42	2504	0,65	47,11	0,64	45,91	0,61	43,11	0,63	43,51	0,68	42,32	0,69	41,12
43	5024	0,71	38,01	0,70	38,61	0,68	38,21	0,68	36,92	0,73	36,62	0,73	36,32
44	345	1,52	47,83	1,67	39,13	1,30	57,97	1,30	57,97	1,44	47,83	1,48	49,28

Наиболее показательным является сравнение результатов смешанного и изолированного подходов. В первом сценарии лучшая модель (XGBoost без эмбеддингов) демонстрирует MAE = 3,37 и accuracy = 28,67%. Во втором сценарии для многих проектов достигается существенно более высокая точность. Например, проект 5: MAE = 0,92 и accuracy = 54,26%, проект 31: MAE = 0,53, accuracy = 54,76%. проект 42: MAE = 0,65, accuracy = 47,11%. Эти результаты демонстрируют, что модель способна достаточно точно прогнозировать трудозатраты внутри отдельной подсистемы, но переносимые на всю совокупность закономерности отсутствуют.

Объём обучающей выборки не является единственным фактором успеха. Так, проект 28 содержит 18412 задач — самый большой массив данных, однако его метрики (MAE = 2,74, accuracy = 24,52%) заметно хуже, чем у проектов с гораздо меньшим количеством наблюдений. Это указывает на внутреннюю неоднородность процессов даже внутри крупных проектов или на высокий уровень шума в данных.

Сравнение двух типов моделей показывает, что градиентный бустинг XGBoost в целом превосходит нейронную сеть как по MAE, так и по accuracy на большинстве проектов. Например, для проекта 24 XGBoost без эмбеддингов даёт accuracy = 30,08%, тогда как нейронная сеть без эмбеддингов — лишь 20,33%. Однако есть исключения: проект 14 (нейросеть без эмбеддингов: MAE = 0,96 против 1,04 у бустинга). Это говорит об отсутствии универсального лидера — выбор оптимальной модели также зависит от доменного контекста.

Сравнение CatBoost с XGBoost показывает схожую картину: градиентный бустинг (обе реализации) стабильно превосходит нейронную сеть на большинстве проектов, однако универсального лидера между XGBoost и CatBoost не наблюдается. Например, на проекте 18 CatBoost без эмбеддингов достигает accuracy = 54,79% против 32,88% у XGBoost, тогда как на проекте 31 XGBoost показывает лучший результат (54,76% против 42,86% у CatBoost). Таким образом, выбор оптимального алгоритма также зависит от доменного контекста, что усиливает основной вывод о статистической индивидуальности проектов.

Добавление PCA-эмбеддингов текстовых описаний не даёт однозначного улучшения. В ряде проектов (5, 21) эмбеддинги повышают точность, в других (1, 31, 43) — снижают, а в некоторых (27, 42) — не меняют метрики. Например, для проекта 5 accuracy бустинга возрастает с 51,06% до 54,26%, а для проекта 1 — падает с 35,71% до 30,90%. Следовательно, информативность текстовых описаний существенно различается между подсистемами: в одних проектах текст задачи содержит сильный сигнал о трудозатратах, в других — преимущественно шум или дублирует уже имеющиеся категориальные признаки.

Отдельного внимания заслуживают проекты с высокими значениями метрик при малой выборке, как например проект 31: при всего 208 задачах XGBoost без эмбеддингов демонстрирует MAE = 0,53 и accuracy = 54,76% — аномально высокий результат для малой выборки. Предварительный анализ показывает, что в этом проекте значения Story Point’ов сосредоточены в узком диапазоне (в основном 2) и сильно коррелируют с типом задачи (Type), что создаёт зависимость, легко улавливаемую бустингом.

Проведенный тест Левена показал выраженную неоднородность дисперсий между 38 проектами. Для всех проанализированных числовых признаков, включая тип задачи (Type), приоритет (Priority), оценку Story Point и дату создания (Creation_Date), значение p-value оказалось практически нулевым ($p < 0.001$). Доля признаков с неоднородными дисперсиями составила 100%, что убедительно свидетельствует: разброс значений внутри проектов статистически значимо различается. Это означает, что вариативность трудозатрат и сопутствующих атрибутов задач существенно зависит от контекста конкретного проекта.

Результаты однофакторного дисперсионного анализа (ANOVA) полностью подтверждают выводы теста Левина, дополняя их картиной различий в центральных тенденциях. Для каждого из признаков нулевая гипотеза о равенстве средних между проектами была отвергнута при $p < 0.001$, а доля признаков с различными средними также составила 100%. Особенно высокие значения F-статистики (например, 5275 для Priority) указывают на то, что межгрупповая вариация многократно превышает внутригрупповую. Следовательно, проекты смещены относительно друг друга не только по разбросу, но и по типичным значениям ключевых характеристик.

Заключение

В ходе исследования выполнен анализ влияния гетерогенности данных, генерируемых различными подсистемами сложной организационно-технической системы, на инвариантность регрессионных моделей машинного обучения. На примере открытого набора данных TAWOS экспериментально подтверждена статистическая неоднородность данных как по дисперсиям, так и по средним значениям ключевых признаков. Гипотеза о возможности построения единой модели, способной прогнозировать трудозатраты без учёта доменного контекста, была опровергнута: модель, обученная на агрегированной выборке, демонстрирует низкое качество (MAE = 3,37, Accuracy ±25% = 28,67 % для лучшей конфигурации).

Проведённый эксперимент показал, что приемлемая точность прогноза достигается исключительно при обучении и тестировании в пределах одной подсистемы, однако метрики крайне вариативны (MAE от 0,53 до 5,91; Accuracy ±25% от 14,29% до 54,76%), что отражает «статистическую индивидуальность» каждого проекта. Тесты Левена и однофакторный дисперсионный анализ подтвердили значимую неоднородность дисперсий и средних ($p < 0.001$) для всех рассмотренных числовых признаков. Сравнение градиентного бустинга и нейронной сети выявило преимущество XGBoost на большинстве проектов, однако универсально лучший алгоритм отсутствует — выбор оптимальной модели также зависит от доменного контекста. Добавление текстовых эмбеддингов не привело к однозначному улучшению, что свидетельствует о различной информативности описаний задач в разных подсистемах.

Полученные результаты имеют следующее значение для развития методов системного анализа и обработки информации в многокомпонентных средах:

1. Показана необходимость обязательной предварительной проверки статистической однородности выборок, объединяемых из разнородных источников. Игнорирование гетерогенности ведёт к построению моделей, не отражающих истинные закономерности отдельных подсистем, и делает их неприменимыми для практических задач поддержки принятия решений.

2. Результаты обосновывают необходимость внедрения методов доменной адаптации или многоуровневого моделирования (с общими и доменно-специфичными компонентами) как обязательного этапа предобработки данных при смене объекта исследования.

Таким образом, исследование демонстрирует, что проблема переносимости моделей в многокомпонентных системах является не техническим ограничением, а фундаментальным свойством, обусловленным гетерогенностью порождаемых данных. Учёт этого свойства, включая предварительный анализ гомогенности выборки и адаптацию к условиям конкретной подсистемы, представляет собой необходимое условие корректной обработки информации и принятия обоснованных управленческих решений в сложных организационно-технических системах.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. The TAWOS Dataset: Tasks, Worklogs and Sprints / SOLAR Research Group. — URL: https://github.com/SOLAR-group/TAWOS (accessed: 01.10.2025).

2. The TAWOS Dataset // University College London (UCL) Research Data Repository. — URL: https://rdr.ucl.ac.uk/articles/dataset/The_TAWOS_dataset/21308124 (accessed: 01.10.2025).

3. Tawosi A. A Versatile Dataset of Agile Open-Source Software Projects / A. Tawosi, A. Al-Subaihin, M. Moussa [et al.] // Proceedings of the 19th International Conference on Mining Software Repositories (MSR 2022). — 2022. — P. 1–5.

4. Pandas 2.3.3 Documentation. — URL: https://pandas.pydata.org/docs/ (accessed: 13.11.2025).

5. scikit-learn 1.7.2 User Guide. — URL: https://scikit-learn.org/stable/user_guide.html (accessed: 13.11.2025).

6. Синьков Д.В. Кодирование категориальных данных для использования в машинном обучении / Д.В. Синьков, А.Д. Ваничкин // Молодой учёный. — 2020. — № 21 (311). — С. 70–72.

7. TensorFlow v2.16.1 API Documentation. — URL: https://www.tensorflow.org/api_docs (accessed: 13.11.2025).

8. Hyndman R.J. Another look at measures of forecast accuracy / R.J. Hyndman, A.B. Koehler // International Journal of Forecasting. — 2006. — № 22 (4). — P. 679–688.

9. Willmott C.J. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance / C.J. Willmott, K. Matsuura // Climate Research. — 2005. — № 30 (1). — P. 79–82

10. BAAI/bge-small-en-v1.5. — URL: https://huggingface.co/BAAI/bge-small-en-v1.5 (accessed: 13.11.2025).

11. Yalçıner B. Enhancing Agile Story Point Estimation: Integrating Deep Learning, Machine Learning, and Natural Language Processing with SBERT and Gradient Boosted Trees / B. Yalçıner [et al.] // Applied Sciences. — 2024. — Vol. 14. — №. 16. — P. 7305.

12. Uc-Cetina V. Recent Advances in Software Effort Estimation using Machine Learning / V. Uc-Cetina. — 2023. — URL: https://arxiv.org/abs/2303.03482 (accessed: 28.05.2026).

13. Cabral J.T.H.A. Ensemble effort estimation: An updated and extended systematic literature review / J.T.H.A. Cabral, A.L.I. Oliveira, F.Q. da Silva // Journal of Systems and Software. — 2023. — Vol. 195. — P. 111542.

14. Atoum I. Enhancing software effort estimation with pre-trained word embeddings: A small-dataset solution for accurate story point prediction / I. Atoum, A.A. Otoom // Electronics. — 2024. — Vol. 13. — № 23. — P. 4843.

15. Singhal P. Domain adaptation: challenges, methods, datasets, and applications / P. Singhal [et al.] // IEEE access. — 2023. — Vol. 11. — P. 6973–7020.

16. Phan H. Heterogeneous graph neural networks for software effort estimation / H. Phan, A. Jannesari // Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. — 2022. — P. 103–113.

Список литературы на английском языке / References in English

1. The TAWOS Dataset: Tasks, Worklogs and Sprints / SOLAR Research Group. — URL: https://github.com/SOLAR-group/TAWOS (accessed: 01.10.2025).

2. The TAWOS Dataset // University College London (UCL) Research Data Repository. — URL: https://rdr.ucl.ac.uk/articles/dataset/The_TAWOS_dataset/21308124 (accessed: 01.10.2025).

3. Tawosi A. A Versatile Dataset of Agile Open-Source Software Projects / A. Tawosi, A. Al-Subaihin, M. Moussa [et al.] // Proceedings of the 19th International Conference on Mining Software Repositories (MSR 2022). — 2022. — P. 1–5.

4. Pandas 2.3.3 Documentation. — URL: https://pandas.pydata.org/docs/ (accessed: 13.11.2025).

5. scikit-learn 1.7.2 User Guide. — URL: https://scikit-learn.org/stable/user_guide.html (accessed: 13.11.2025).

6. Sinkov D.V. Kodirovanie kategorialnikh dannikh dlya ispolzovaniya v mashinnom obuchenii [Encoding Categorical Data for Use in Machine Learning] / D.V. Sinkov, A.D. Vanichkin // Molodoi uchyonii [Young Scientist]. — 2020. — № 21 (311). — P. 70–72. [in Russian]

7. TensorFlow v2.16.1 API Documentation. — URL: https://www.tensorflow.org/api_docs (accessed: 13.11.2025).

8. Hyndman R.J. Another look at measures of forecast accuracy / R.J. Hyndman, A.B. Koehler // International Journal of Forecasting. — 2006. — № 22 (4). — P. 679–688.

9. Willmott C.J. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance / C.J. Willmott, K. Matsuura // Climate Research. — 2005. — № 30 (1). — P. 79–82

10. BAAI/bge-small-en-v1.5. — URL: https://huggingface.co/BAAI/bge-small-en-v1.5 (accessed: 13.11.2025).

11. Yalçıner B. Enhancing Agile Story Point Estimation: Integrating Deep Learning, Machine Learning, and Natural Language Processing with SBERT and Gradient Boosted Trees / B. Yalçıner [et al.] // Applied Sciences. — 2024. — Vol. 14. — №. 16. — P. 7305.

12. Uc-Cetina V. Recent Advances in Software Effort Estimation using Machine Learning / V. Uc-Cetina. — 2023. — URL: https://arxiv.org/abs/2303.03482 (accessed: 28.05.2026).

13. Cabral J.T.H.A. Ensemble effort estimation: An updated and extended systematic literature review / J.T.H.A. Cabral, A.L.I. Oliveira, F.Q. da Silva // Journal of Systems and Software. — 2023. — Vol. 195. — P. 111542.

14. Atoum I. Enhancing software effort estimation with pre-trained word embeddings: A small-dataset solution for accurate story point prediction / I. Atoum, A.A. Otoom // Electronics. — 2024. — Vol. 13. — № 23. — P. 4843.

15. Singhal P. Domain adaptation: challenges, methods, datasets, and applications / P. Singhal [et al.] // IEEE access. — 2023. — Vol. 11. — P. 6973–7020.

16. Phan H. Heterogeneous graph neural networks for software effort estimation / H. Phan, A. Jannesari // Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. — 2022. — P. 103–113.