Применение ансамблевых методов для прогнозирования трудозатрат выполнения проектов при проектировании объектов обустройства месторождений

Т. М. Мурзагалеев

Общество с ограниченной ответственностью «РН-Проектирование Добыча»;

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

tmurzagaleev@yandex.ru

Н. А. Жукова

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

nazhukova@mail.ru

Аннотация. **Прогнозирование трудозатрат и продолжительности выполнения проекта имеет решающее значение и является основой для определения стоимости и прогнозирования сроков сдачи проекта. Успешное внедрение методов машинного обучения в управление проектами для прогнозирования трудозатрат проекта может обеспечить преимущество и существенно улучшить данную область. Целью исследования является применение ансамблевых методов для прогнозирования трудозатрат при выполнении проектов для объектов обустройства месторождений на основании исторических данных, в частности, проектов автомобильных дорог 3 и 4 категории, инженерной подготовки площадок на нефтегазовых месторождениях и линий электропередач и волоконно-оптической системы. В качестве исходных данных были использованы исторические данные проектов обустройства нефтяных и газовых месторождений и трудозатраты специалистов, участвующих в разработке проектов, собранные за последние 7 лет.**

Ключевые слова**:** ***прогнозирование трудозатрат выполнения проекта; ансамблевые методы; машинное обучение; регрессионная модель; управление проектами; объекты обустройства месторождения***

I. Актуальность

Управление проектами является одной из ключевых практик, которая помогает организациям обеспечить эффективное управление реализуемыми ими проектами, при решении которой должно учитываться множество внешних и внутренних факторов. Управление проектами включает в себя много аспектов работы с проектами, среди которых оценка трудозатрат и длительность проекта выделяются как одни из наиболее значимых [1].

В проектировании объектов обустройства месторождений, как, например, и в ИТ-сфере, наиболее значимым ресурсом является человеческий, в связи с этим оценка трудозатрат является наиболее значимой составляющей проекта. В связи с этим точное прогнозирование трудозатрат влияет на планирование, составлении смет, распределение ресурсов, обеспечение качества и производительности, удовлетворение потребностей заказчиком, обмен информацией, совместное принятие решений и общее управление проектом [2]. При этом с ростом масштабов и количества проектов растет объем информации, что усложняет задачу управления проектами. В то же время оценка трудозатрат и сроков проекта при проектировании, которое является одним из этапов строительного проекта, представляет собой задачу, которая часто осложняется недостатком данных, неопределенностью и потенциальными рисками. Тем не менее, оценка трудозатрат и сроков остается одним из важнейших элементов обеспечения успеха строительного проекта, а качественное планирование позволяет компаниям развиваться значительно быстрее [3, 4].

II. Постановка проблемы

Каждый проект в начале своего жизненного цикла требует определения затрат и временных рамок для определения бизнес-обоснования и получения одобрения от заказчика. Для этой цели наиболее широко используются несколько традиционных подходов к оценке трудозатрат, основанные на экспертных знаниях и аналогах. Тем не менее, эти методы обычно подвержены ошибкам, так как экспертные знания базируются на субъективной оценке руководителей проектов, основанной на опыте, а поиск близкого аналога или не представляется возможным, или является достаточно сложной задачей, учитывая количество влияющих на проект факторов [5, 6].

Поэтому, чтобы устранить упомянутые недостатки, в течение последних двух десятилетий проводились обширные исследования методов интеллектуального анализа данных для оценки трудозатрат проекта с помощью современных прогнозирующих алгоритмов машинного обучения. Благодаря автоматизированному процессу прогнозирования, основанному на исторической информации, они, как правило, снижают предубеждения людей и психологическое влияние со стороны [6]. Среди методов машинного обучения ансамблевые методы часто проявляются как более эффективные методы и обеспечивают более точные результаты по сравнению с существующими индивидуальными моделями машинного обучения или моделями глубокого обучения [7, 8].

Целью настоящего исследования является применение ансамблевых методов для прогнозирования трудозатрат при выполнении проектов для объектов обустройства месторождений на основании исторических данных, в частности, проектов автомобильных дорог 3 и 4 категории, инженерной подготовки площадок на нефтегазовых месторождениях и линий электропередач (ВЛ), волоконно-оптической системы (ВОЛС) и кабельных эстакад (КЭ). В качестве исходных данных были использованы исторические данные проектов обустройства нефтяных и газовых месторождений и трудозатраты специалистов, участвующих в разработке проектов, собранные за последние 7 лет.

Для достижения поставленной цели требовалось выполнить следующие задачи:

- на 3 наборах данных провести исследования с применением базовых методов машинного обучения, используемых другими авторами исследований для прогнозирования трудозатрат выполнения проектов;
- применить ансамблевые методы voting ensemble и stacking ensemble, сравнить полученные результаты с результатами базовых методов машинного обучения.

III. Материалы и методы

В качестве исходных данных использовались данные по проектам автомобильных дорог, инженерной подготовки площадок на нефтегазовых месторождениях, ВЛ, ВОЛС и КЭ обустройства месторождений северных регионов Российской Федерации различных недропользователей за период с 2018 по 2025 год. Были выделены 3 набора данных, из которых первый набор данных – это данные по проектам автомобильных дорог и трудозатратам отдела автомобильных дорог, второй набор данных содержит данные по проектам инженерной подготовки площадок на нефтегазовых месторождениях, в которые входят данные по проектам инженерной подготовки площадок кустов скважин и вспомогательных (инфраструктурных) площадок и данные о трудозатратах отдела генеральных планов, и третий набор данных состоит из общих данных по проектам автомобильных дорог, инженерной подготовки площадок на нефтегазовых месторождениях, ВЛ, ВОЛС и КЭ, и данные о трудозатратах отдела землеустроительных работ (далее набор данных по общим проектам). Полный перечень и описание признаков указанных наборов данных приведены в таблицах 1-3, целевыми переменными наборов данных являются трудозатраты специалистов отделов по разработке проектов. Первый набор данных состоит из 1417 векторов данных, второй и третий наборы данных состоят из 1455 векторов данных с различными характеристиками из ранее выполненных проектов. Данные были получены из 109, 97 и 97 ранее выполненных проектов, соответственно, к исходным признакам был применен метод масштабирующего коэффициента для улучшения обучения моделей на больших данных, а также для расширения диапазона данных. Все признаки имеют числовые значения за исключением признаков «Наличие основных проектных решений» и «Опыт исполнителей», которые являются категориальными и которые необходимо было перевести в числовые, для возможности создания корректной прогнозирующей модели. Поэтому признак «Наличие основных проектных решений» имел значение «0» при значении «нет», и «1» – при значении «да». Признак «Опыт исполнителей» рассчитывался из опыта двух основных исполнителей проекта, количество трудозатрат которых вносило наибольший вклад в общие трудозатраты проекта, поэтому учитывая различные уровни специалистов от «инженера без категории» или «техника» до «ведущего инженера» или «главного специалиста» получилась сетка из 16 уровней опыта.

ТАБЛИЦА I. Признаки набора данных по проектам автомобильных дорог

№№	Наименование признака	Единицы измерения	Тип переменной
X1	Протяженность автодороги	км	Независимая
X2	Количество автодорог	шт.	Независимая
X3	Количество изменений технического задания на инженерные изыскания	шт.	Независимая
X4	Количество изменений задания на проектирование	шт.	Независимая
X5	Наличие основных проектных решений	да/нет	Независимая
X6	Количество внутри площадочных дорог	шт.	Независимая
X7	Количество водопропускных труб	шт.	Независимая
X8	Количество изгибов автодорог	шт.	Независимая
X9	Количество примыканий дорог	шт.	Независимая
X10	Опыт исполнителей	уровень	Независимая
X11	Количество специалистов	шт.	Независимая
y	Трудозатраты по проекту	человек/день	Целевая

ТАБЛИЦА II. Признаки набора данных по проектам инженерной подготовки нефтегазовых месторождений

№№	Наименование признака	Единицы измерения	Тип переменной
X1	Количество изменений технического задания на инженерные изыскания	шт.	Независимая
X2	Количество изменений задания на проектирование	шт.	Независимая
X3	Наличие основных проектных решений	да/нет	Независимая
X4	Количество водонагнетательных, водоутилизирующих, вододобывающих скважин	шт.	Независимая
X5	Общая площадь площадок кустов скважин	га	Независимая
X6	Количество сооружений	шт.	Независимая
X7	Общая площадь вспомогательных площадок	га	Независимая
X8	Количество этапов	шт.	Независимая
X9	Опыт исполнителей	уровень	Независимая
X10	Количество специалистов	шт.	Независимая
y	Трудозатраты по проекту	человек/день	Целевая

ТАБЛИЦА III. ПРИЗНАКИ НАБОРА ДАННЫХ ПО ОБЩИМ ПРОЕКТАМ

№№	Наименование признака	Единицы измерения	Тип переменной
X1	Протяженность автодорог	км	Независимая
X2	Количество сооружений	шт.	Независимая
X3	Наличие основных проектных решений	да/нет	Независимая
X4	Общая площадь кустовых площадок	га	Независимая
X5	Общая площадь вспомогательных площадок	га	Независимая
X6	Количество этапов инженерной подготовки	га	Независимая
X7	Длина ВЛ, ВОЛС и КЭ	км	Независимая
X8	Количество изменений технического задания на инженерные изыскания по автодорогам	шт.	Независимая
X9	Количество изменений технического задания на инженерные изыскания по инженерной подготовке	шт.	Независимая
X10	Количество изменений технического задания на инженерные изыскания по ВЛ, ВОЛС и КЭ	шт.	Независимая
X11	Опыт исполнителей	уровень	Независимая
X12	Количество специалистов	шт.	Независимая
y	Трудозатраты по проекту	человек/день	Целевая

Для преобразования признаков к единому масштабу без искажения распределения была выполнена нормализация наборов данных с преобразованием значений признаков с использованием скалярного параметра MinMax в масштаб, который находится от 0 до 1. Нормализация также улучшает скорость и устойчивость сходимости градиентных методов оптимизации, стабильность и обобщающую способность регуляризованных моделей и качество некоторых алгоритмов (например, методы, основанные на расстояниях, и использующие скалярное произведение) и интерпретируемость коэффициентов. Затем предварительно обработанные наборы данных были разделены на обучающие (тренировочные) и тестовые данные в соотношении 90:10.

В исследовании были использованы ансамблевые методы stacking ensemble и voting ensemble, включающих различное количество базовых моделей, а также одиночные методы машинного обучения, которые ранее были применены у других авторов исследований прогнозирования трудозатрат в проектах строительной отрасли или разработки программного обеспечения: полиномиальная регрессия, методы ridgeCV, support vector regression (SVR), random forest (RF), k-nearest neighboring (KNN), gradient boosting trees (GBT), eXtreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), categorical boosting (CatBoost), decision trees (DT), multilayer perceptron (MLP), extra trees regressor (ETR), обобщённая регрессионная нейронная сеть (GRNN) [4, 5, 9, 10]. Voting ensemble представляет из себя ансамблевую модель, которая агрегирует прогнозные результаты базовых моделей, подаваемых на вход ансамблевой модели, и усредняет результаты каждой базовой модели. В исследовании для voting ensemble были использованы модели в различном количестве от трех до семи, которые индивидуально показали лучшие метрики, полученных в результате экспериментов, из всех моделей, примененных к набору данных. Stacking ensemble представляет из себя метод, который создаёт новые признаки из предсказаний базовых моделей и обучает мета-модель на этих предсказаниях, таким образом, мета-модель учится комбинировать сильные стороны базовых моделей, повышая качество прогнозных результатов. В качестве регрессионной мета-модели в ансамблевом методе stacking ensemble использована модель по умолчанию RidgeCV. В исследовании для stacking ensemble были использованы модели в различном количестве от трех до восьми, которые также индивидуально показали лучшие метрики из всех моделей, примененных к набору данных.

Для оценки эффективности и сравнения методов между собой были использованы следующие оценочные метрики для возможности идентификации их эффективности и сравнения между собой: средняя абсолютная ошибка (MAE), коэффициент детерминации (R2), средняя величина относительной ошибки (MMRE), медианная величина относительной погрешности (MdMRE), точность прогнозирования (Pred(25)), сумма остатков квадратов (SSR). Указанный набор метрик образует сбалансированную систему для оценки регрессионных моделей, покрывая различные аспекты ошибок и качества модели. Также, были рассчитаны метрики с помощью кросс-валидации (с 5 K-Fold), такие как, стандартное отклонение модели (standard deviation (StdDev)) и коэффициент вариации (coefficient of variation (CV)), которые дают информацию об устойчивости и надежности прогнозов модели.

IV. ЭКСПЕРИМЕНТЫ И ОЦЕНКА РЕЗУЛЬТАТОВ

В табл. 4–6 представлены метрики, полученные в результате применения прогнозирующих моделей к тестовому набору данных по проектам автомобильных дорог, инженерной подготовки, а также общего проекта, с трудозатратами отдела автомобильных дорог, отдела генеральных планов и отдела землеустройства, соответственно. Ключевыми метриками для оценки моделей являются коэффициент детерминации (R2) и сумма остатков квадратов (SSR), поэтому ранжирование результатов методов машинного обучения основывалась на них, остальные метрики являются дополнительными для оценки методов, но также дают важную качественную информацию.

Из таблиц с результатами на тестовых данных видно, что из одиночных моделей по большинству метрик наилучшими оказались модели XGBoost, CatBoost и Extra Trees Regressor, обобщённая регрессионная нейронная сеть и Random forest, поэтому данные одиночные модели в большинстве случаев были использованы в качестве базовых моделей для ансамблевых моделей voting ensemble и stacking ensemble. Gradient boosting trees, k-nearest neighboring, полиномиальная регрессия, lightGBM и decision trees показали средние результаты, худшими моделями со всеми наборами данных оказались метод опорных векторов, ridgeCV и многослойный персептрон, поэтому

они не использовались в качестве базовых моделей для ансамблевых методов.

Также из таблиц с результатами можно увидеть, что ансамблевый метод stacking ensemble с тремя базовыми моделями показывает одни из лучших метрик MAE, R^2 и SSR из всех представленных моделей, причем при увеличении количества базовых моделей наблюдалось улучшение метрик ансамблевого метода и только с восемью базовыми моделями метрики перестали заметно улучшаться, за исключением набора данных с общим проектом, в котором stacking ensemble с восемью базовыми моделями показал лучшие результаты из всех представленных методов.

Ансамблевый метод voting ensemble на наборе данных по проектам автомобильных дорог и наборе данных по общим проектам дает усредненные результаты входящих в него методов, при этом результаты метода снижаются с увеличением количества входящих в него базовых методов и лучшие результаты показывает voting ensemble с тремя базовыми методами. Voting ensemble с набором данных по проектам инженерной подготовки площадок показал метрики лучшие, чем входящие в него базовые модели, также данный метод по метрикам превзошел метод stacking ensemble, что является более редким случаем. Такие результаты voting ensemble с набором данных по проектам инженерной подготовки площадок могут объясняться тем, что произошло уменьшение дисперсии прогнозных данных (сглаживание ошибок, а также тем, что разные модели имеют разные смещения и усреднение их скомпенсировало, т.е. voting ensemble улучшил обобщающую способность за счёт компенсации переобучения входящих в нее моделей и уменьшения разброса предсказаний).

ТАБЛИЦА IV. ПОКАЗАТЕЛИ РАБОТЫ МОДЕЛЕЙ С ДАННЫМИ ПО ПРОЕКТАМ АВТОМОБИЛЬНЫХ ДОРОГ НА ТЕСТОВЫХ ДАННЫХ

Модель	**MAE**	**R^2**	**MMRE**	**MdMRE**	**Pred(25), %**	**StdDev**	**CV, %**	**SSR**
Полиномиальная регрессия (PR)	8,55	0,9420	0,3502	0,1940	58,45	0,83	8,91	21137,4
Multilayer perceptron (MLP)	12,46	0,8529	0,4366	0,2423	50,70	0,73	5,91	53592,5
Support vector regression (SVR)	16,47	0,8613	1,1534	0,3865	37,32	1,54	8,82	50540,0
Random forest (RF)	6,32	0,9630	0,2145	0,1192	72,54	0,92	10,75	13481,5
K-Nearest Neighboring (KNN)	7,79	0,9384	0,2485	0,1462	71,13	0,82	8,51	22437,9
eXtreme Gradient Boosting (XGBoost)	5,53	0,9745	0,1811	0,1306	78,17	0,50	6,26	9275,7
Light Gradient Boosting Machine (LightGBM)	6,96	0,9504	0,2208	0,1474	69,72	0,55	6,05	18076,7
Categorical Boosting (CatBoost)	4,58	0,9789	0,1744	0,0858	78,87	0,58	9,56	7671,0
RidgeCV	12,10	0,8504	0,4224	0,2347	52,11	0,74	5,61	54499,8
Gradient boosting trees (GBT)	8,39	0,9439	0,3170	0,1894	60,56	0,89	8,42	20439,0
Decision Trees (DT)	9,60	0,7979	0,2875	0,1211	64,79	0,83	7,52	73627,2
Extra Trees Regressor (ETR)	4,87	0,9786	0,2016	0,0989	72,54	0,8	11,64	7802,6
Обобщённая регрессионная нейронная сеть (GRNN)	4,45	0,9743	0,1883	0,0775	82,39	0,43	8,39	9373,0
Voting ensemble (CatBoost+ETR+XGBoost)	4,62	0,9807	0,1767	0,0904	74,65	0,59	8,92	7041,8
Voting ensemble (CatBoost+ETR+XGBoost+GRNN+RF)	4,73	0,9792	0,1843	0,0952	73,94	0,58	8,79	7584,9
Voting ensemble (CatBoost+ETR+XGBoost+GRNN+RF+LightGBM+GBT)	5,41	0,9738	0,2018	0,1106	73,94	0,59	8,13	9546,5
Stacking ensemble (CatBoost+ETR+XGBoost)	4,91	0,9794	0,1730	0,0971	75,35	0,54	8,65	7488,0
Stacking ensemble (CatBoost+ETR+XGBoost+GRNN+RF)	4,67	0,9808	0,1661	0,0982	77,46	0,48	8,64	6991,1
Stacking ensemble (CatBoost+ETR+XGBoost+GRNN+RF+LightGBM+GBT)	4,27	0,9829	0,1535	0,0862	78,87	0,51	9,67	6229,1
Stacking ensemble (CatBoost+ETR+XGBoost+GRNN+RF+LightGBM+GBT+PR)	4,64	0,9823	0,1682	0,0985	74,65	0,53	9,85	6430,0

ТАБЛИЦА V. ПОКАЗАТЕЛИ РАБОТЫ МОДЕЛЕЙ С ДАННЫМИ ПО ПРОЕКТАМ ИНЖЕНЕРНОЙ ПОДГОТОВКИ НА ТЕСТОВЫХ ДАННЫХ

Модель	**MAE**	**R^2**	**MMRE**	**MdMRE**	**Pred(25), %**	**StdDev**	**CV, %**	**SSR**
Полиномиальная регрессия (PR)	11,32	0,9153	0,3271	0,2304	50,68	0,49	4,34	68396,3
Multilayer perceptron (MLP)	16,17	0,9001	0,6053	0,4321	36,30	0,82	5,48	80709,5
Support vector regression (SVR)	47,64	0,5238	3,4694	1,7907	13,01	4,42	22,30	384756,9
Random forest (RF)	6,87	0,9803	0,2202	0,1535	68,49	0,99	11,56	15889,9
K-Nearest Neighboring (KNN)	6,30	0,9757	0,1821	0,1206	74,66	1,08	13,76	19607,6
eXtreme Gradient Boosting (XGBoost)	6,35	0,9797	0,1825	0,1424	73,97	1,18	15,29	16415,4
Light Gradient Boosting Machine (LightGBM)	8,13	0,9499	0,2206	0,1666	70,55	2,10	18,59	40441,6
Categorical Boosting (CatBoost)	6,03	0,9825	0,1879	0,1101	76,03	0,89	12,92	14133,4
RidgeCV	14,81	0,8918	0,4345	0,3036	43,15	0,79	4,92	87387,3
Gradient boosting trees (GBT)	9,45	0,9602	0,2748	0,1735	61,64	0,78	7,80	32182,7
Decision Trees (DT)	6,98	0,9735	0,2545	0,1041	68,49	0,99	9,87	21432,0
Extra Trees Regressor (ETR)	6,39	0,9799	0,2058	0,1407	67,12	0,95	12,27	16199,6
Обобщённая регрессионная нейронная сеть (GRNN)	5,35	0,9776	0,2094	0,1104	76,03	1,11	18,58	18081,8
Voting ensemble (CatBoost+ETR+RF)	6,13	0,9823	0,1908	0,1161	73,29	0,92	12,39	14282,4
Voting ensemble (CatBoost+ETR+RF+XGBoost+GRNN)	5,28	0,9872	0,1749	0,0944	76,03	0,86	12,69	10314,3
Voting ensemble (CatBoost+ETR+RF+XGBoost+GRNN+KNN+DT)	5,10	0,9878	0,1705	0,0923	76,03	0,80	11,82	9831,3
Stacking ensemble (CatBoost+ETR+RF)	5,99	0,9779	0,1999	0,1183	67,12	1,26	17,33	17849,8
Stacking ensemble (CatBoost+ETR+RF+XGBoost+GRNN)	5,17	0,9867	0,1708	0,0999	75,34	0,97	14,18	10747
Stacking ensemble (CatBoost+ETR+RF+XGBoost+GRNN+KNN+DT)	5,02	0,9868	0,1699	0,1034	71,92	1,03	15,2	10650,7
Stacking ensemble (CatBoost+ETR+RF+XGBoost+GRNN+KNN+DT+GBT)	5,33	0,9856	0,1771	0,1151	71,23	1,09	16,13	11670,8

ТАБЛИЦА VI. ПОКАЗАТЕЛИ РАБОТЫ МОДЕЛЕЙ С ДАННЫМИ ПО ПРОЕКТАМ ОТДЕЛА ЗЕМЛЕУСТРОЙСТВА НА ТЕСТОВЫХ ДАННЫХ

Модель	MAE	R^2	MMRE	MdMRE	Pred(25), %	StdDev	CV, %	SSR
Полиномиальная регрессия (PR)	2,13	0,8959	0,3046	0,2278	54,79	0,15	6,83	1214,0
Multilayer perceptron (MLP)	2,29	0,8684	0,3466	0,2709	45,89	0,16	6,71	1535,8
Support vector regression (SVR)	3,60	0,7696	0,9291	0,3789	30,14	0,17	6,07	2687,5
Random forest (RF)	1,76	0,8986	0,2659	0,1384	69,86	0,11	6,27	1182,4
K-Nearest Neighboring (KNN)	1,62	0,9222	0,2221	0,1686	65,07	0,15	9,45	907,6
eXtreme Gradient Boosting (XGBoost)	1,49	0,9248	0,2149	0,1467	73,29	0,08	5,19	877,3
Light Gradient Boosting Machine (LightGBM)	1,73	0,8759	0,2329	0,1708	69,18	0,13	7,17	1448,1
Categorical Boosting (CatBoost)	1,18	0,9627	0,1869	0,1362	77,40	0,06	5,06	435,2
RidgeCV	2,43	0,8522	0,3484	0,2417	52,05	0,14	4,73	1724,0
Gradient boosting trees (GBT)	1,87	0,9034	0,2705	0,1747	60,27	0,10	4,87	1127,3
Decision Trees (DT)	1,81	0,8450	0,2728	0,1419	64,38	0,16	7,27	1808,3
Extra Trees Regressor (ETR)	1,57	0,9362	0,2552	0,1647	70,55	0,06	3,73	744,2
Обобщённая регрессионная нейронная сеть (GRNN)	1,09	0,9684	0,1814	0,1003	74,66	0,11	9,69	368,3
Voting ensemble (GRNN+CatBoost+ETR)	1,23	0,9623	0,2001	0,1391	74,66	0,07	5,5	439,2
Voting ensemble (GRNN+CatBoost+ETR+XGBoost+KNN)	1,24	0,9566	0,1892	0,1246	75,34	0,08	6,08	506,8
Voting ensemble (GRNN+CatBoost+ETR+XGBoost+KNN+GBT+RF)	1,38	0,9467	0,2078	0,1414	73,29	0,08	5,45	621,5
Stacking ensemble (SE) (GRNN+CatBoost+ETR)	1,07	0,9727	0,1656	0,1130	81,51	0,07	6,62	318,9
Stacking ensemble (GRNN+CatBoost+ETR+XGBoost+KNN)	1,07	0,9714	0,1604	0,1141	82,19	0,08	7,11	333,7
Stacking ensemble (SE) (GRNN+CatBoost+ETR+XGBoost+KNN+GBT+RF)	1,02	0,9740	0,1563	0,1082	84,25	0,11	10,40	303,2
Stacking ensemble (GRNN+CatBoost+ETR+XGBoost+KNN+GBT+RF+PR)	1,02	0,9743	0,1560	0,1055	83,56	0,13	11,35	300,3

V. ЗАКЛЮЧЕНИЕ

Таким образом, в данном исследовании лучшие результаты на трех наборах данных показали одиночные метод categorical boosting и обобщённая регрессионная нейронная сеть: с набором данных по проектам автомобильных дорог лучшим методом стал categorical boosting с метрикой R^2 равной 0.9789, с набором данных по проектам инженерной подготовки лучшим методом стал categorical boosting с метрикой R^2 равной 0.9825, с набором данных по общим проектам лучшей стала обобщённая регрессионная нейронная сеть с метрикой R^2 равной 0.9684.

Разработаны ансамблевые методы voting ensemble и stacking ensemble с различным составом базовых моделей, которые показали лучшие метрики оценки эффективности среди всех рассмотренных методов. С набором данных по проектам автомобильных дорог лучшим ансамблевым методом стал stacking ensemble с 7 базовыми моделями с метрикой R^2 равной 0.9829, с набором данных по проектам инженерной подготовки площадок лучшим ансамблевым методом стал voting ensemble с 7 базовыми моделями с метрикой R^2 равной 0.9878, с набором данных по общим проектам лучшим ансамблевым методом стал stacking ensemble с 8 базовыми моделями с метрикой R^2 равной 0.9743.

Таким образом, метод stacking ensemble оказался наиболее стабильным по полученным метрикам и показал со всеми наборами данных лучшие результаты, уступив только voting ensemble с набором данных по проектам инженерной подготовки площадок. При этом коэффициенты вариации (CV) метода stacking ensemble во всех экспериментах были низкими и составили от 6.62 до 17.33 % со всеми наборами данных, что указывает на достаточную стабильность и надежность метода stacking ensemble.

Учитывая вышеуказанное, ансамблевые методы voting ensemble и stacking ensemble являются рекомендуемыми к использованию для прогнозирования трудозатрат по проектам автомобильных дорог, инженерной подготовки площадок и общим проектам.

СПИСОК ЛИТЕРАТУРЫ

[1] A. O. Sousa et al. Applying Machine Learning to Estimate the Effort and Duration of Individual Tasks in Software Projects. IEEE Access, vol. 11, pp. 89933-89946, 2023. https://doi.org/10.1109/ACCESS.2023.3307310.

[2] A. Jadhav, S. K. Shandilya, I. Izonin and M. Gregus. Effective Software Effort Estimation Leveraging Machine Learning for Digital Transformation. IEEE Access, vol. 11, pp. 83523-83536, 2023. https://doi.org/10.1109/ACCESS.2023.3293432.

[3] Сахнюк Т.И., Коршикова М.В., Сахнюк П.А. Российские системы управления проектами // Наука Красноярья. 2022. №4. С. 24-36. URL: https://cyberleninka.ru/article/n/rossiyskie-sistemy-upravleniya-proektami.

[4] Nassar, A.H. and Elbisy, A.M. 2024. A Machine Learning Approach to Predict Time Delays in Marine Construction Projects. Engineering, Technology & Applied Science Research. 2024. Vol.14, №5, pp. 16125–16134. https://doi.org/10.48084/etasr.8173.

[5] Shen Zhang & Xuechun Li. A comparative study of machine learning regression models for predicting construction duration. Journal of Asian Architecture and Building Engineering, 23(6), 2023. https://doi.org/10.1080/13467581.2023.2278887.

[6] Przemyslaw Pospieszny, Beata Czarnacka-Chrobot, Andrzej Kobylinski. An effective approach for software project effort and duration estimation with machine learning algorithms // Journal of Systems and Software. Volume 137, 2018, Pages 184-196, https://doi.org/10.1016/j.jss.2017.11.066.

[7] Eswara Rao, K., Pydi, B., Annan Naidu, P., Prasann, U. D., & Anjaneyulu, P. (2023). Ensemble Learning Approach for Effective Software Development Effort Estimation with Future Ranking. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 12(1), e31206. https://doi.org/10.14201/adcaij.31206.

[8] A.G. Priya Varshini, Anitha Kumari K, and Vijayakumar Varadarajan. Estimating Software Development Efforts Using a Random Forest-Based Stacked Ensemble Approach. Electronics 2021, 10, 1195. https://doi.org/10.3390/electronics10101195.

[9] Коньков В.В., Широков В.И., Жабицкий М.Г. Прогнозирование срывов сроков строительства с использованием машинного обучения на основе исторических данных о фактической продолжительности завершенных проектов // International Journal of Open Information Technologies. 2024. №8. С. 35-47

[10] Soreti M Liben, Demiss A Belachew and Walied A Elsaigh. Comparing advanced and traditional machine learning algorithms for construction duration prediction: a case study of Addis Ababa's public sector // Engineering Research Express. 2024. Volume 6. Number 4. URL: https://doi.org/10.1088/2631-8695/ad979f.