Москва, 2025

Искусственный интеллект и машинное обучение в аудите национальных проектов

Аннотация

В статье представлено описание отдельных примеров использования алгоритмов машинного обучения и искусственного интеллекта при проведении Счетной палатой Российской Федерации аудита национальных проектов, ключевые выводы и направления дальнейшего развития данных подходов.

Колеров Сергей Борисович

к.э.н., главный инспектор,
Счетная Палата Российской Федерации

Шишлин Андрей Владимирович

главный инспектор,
Счетная палата Российской Федерации

Ключевые слова

DFOG (Duplication, Fragmentation, Overlap, and Gap), BERT (Bidirectional Encoder Representations from Transformers), Machine Learning, Random Forest, K-nearest neighbors, KNN, Probabilistic Latent Semantic Analysis, PLSA, ROC-анализ.

Следуя международным стандартам (ISSAI 130; ISSAI 300) Счетная палата Российской Федерации в своей аудиторской деятельности, наряду с базисными принципами критического подхода, объективности и профессионального скептицизма, также открыта новому и стремится внедрять инновации. Одной из ключевых инноваций в работе Счетной палаты является применение алгоритмов, задействующих искусственный интеллект (AI) и в частности, машинное обучение (ML).

В последнее время получило развитие направление внешнего государственного аудита национальных проектов, однако считалось, что в нем превалируют нетиповые задачи, в которых алгоритмизация не дает необходимого по масштабу эффекта или несет неприемлемые последствия ошибочных решений. Вместе с тем, учитывая тот факт, что информация, подвергаемая анализу в рамках аудита национальных проектов, уже достигла масштабов «big data», а также с учетом нарастающего дефицита ресурсов для ее обработки, использование AI-алгоритмов становится практически неизбежным.

Среди конкретных примеров применения AI и ML в последний год в Счетной палате в рамках аудита национальных проектов выделим следующие.

Анализ портфеля на предмет поиска дублирующих сущностей

При работе с достаточно крупным портфелем проектов, не имеющим стандартизированной структуры декомпозиции работ (справочника мероприятий/результатов/целей), вероятно возникновение дублирования данных категорий между собой в рамках одного проекта («вертикальное» дублирование), а также между проектами («горизонтальное» дублирование).

Поиску и анализу дублирований посвящен большой пласт аудиторской работы, в частности в рамках концепции DFOG (Duplication, Fragmentation, Overlap, and Gap)[1].

В отдельных случаях дублирование не является негативной практикой, поскольку может отражать, к примеру, выполнение типовых проектных активностей применительно к различным объектам (например, работа «проведение инженерных изысканий» в независящих друг от друга строительных проектах), либо декомпозицию агрегируемого показателя (например, «прирост выручки»).

В общем случае дублирование должно быть поводом для пристального внимания, поскольку является нарушением принципа MECE[2] и может приводить к таким отклонениям, как параллельное финансирование одной работы, появление активностей-«клонов», искажение отчетности (наполнение ее избыточной информацией, появление сущностей, не соответствующих иерархическому уровню и т.п.). Все перечисленное несет угрозу необоснованных трат, а также принятия ошибочных управленческих решений.

Проблема поиска дублей может быть решена различными прикладными методами. Пожалуй, самый простой в реализации подход связан с токенизацией (разбивкой на составляющие элементы) наименований работ/мероприятий/результатов/целей по отдельным символам алфавита с последующим сравнением по принципу «всё со всем». В этом случае метрикой схожести выступает разность между количеством конкретных символов в сравниваемых наименованиях. Преимуществами данного алгоритма является его понятность и интерпретируемость, а также то, что он без каких-либо дополнительных издержек может быть реализован в Excel. Среди недостатков можно отметить то, что данный алгоритм чувствителен к использованию слов-синонимов, сокращений, размыванию сути словами, не несущими смысловой нагрузки, а также к амфиболии (к двусмысленности – например, «Казнить нельзя помиловать»).

Более сложные и более совершенные алгоритмы поиска дублирований опираются на другие метрики, в том числе известное в теории информации расстояние Левенштейна. В качестве практической реализации продвинутых алгоритмов приведем пример модуля «Fuzzy lookup» для Excel, который основан на использовании

1 https://asosai.org/asosai/upload/file/202308/62322.pdf

2 «Mutually Exclusive, Collectively Exhaustive» - принцип управления, предполагающий, что совокупность элементов должна быть исчерпывающей и взаимно непересекающейся.

нечеткой логики (Fuzzy Logic). Несколько более удобным образом может быть организован поиск дублей с помощью инструмента слияния запросов с нечетким сопоставлением в подсистеме обработки данных Excel - Power Query. Еще более мощный инструмент для анализа дублирования – использование языковых моделей на основе ИИ, например, BERT (Bidirectional Encoder Representations from Transformers), которые устраняют многие недостатки более простых алгоритмов, в частности касательно поиска слов-синонимов.

В случае, если предполагается наличие множественного дублирования (в частности, одновременно «горизонтального» и «вертикального» дублирования), для выявления «уникальных» сущностей целесообразно проводить кластеризацию. Если при этом опираться только на одну метрику (например, расстояние Левенштейна), кластеризация может быть проведена путем вычленения связных графов. Например, если обнаружено дублирование между мероприятиями «А» и «Б», а также (независимо) между мероприятиями «Б» и «В», то, представляя буквы вершинами графа, можно говорить о связности графа А-Б-В. Таким образом осуществляется переход от вычисления попарного дублирования к выявлению групп (кластеров), в которых наблюдается определенная схожесть между всеми входящими в них сущностями. Далее ставится управленческий вопрос о причинах появления дублей и, возможно, об их исключении. Если же речь заходит о кластеризации с использованием нескольких признаков (например, помимо расстояния Левенштейна – учет соответствия плановых или базовых значений показателей), могут применяться специализированные алгоритмы кластеризации, в том числе использующие ML – K-means, вероятностный латентный семантический анализ (англ. Probabilistic Latent Semantic Analysis, PLSA) и другие.

Стоит признать, что все перечисленные алгоритмы не устраняют полностью необходимости дальнейшего «ручного» анализа, однако, по опыту Счетной палаты, кратно снижают его трудоемкость в силу сокращения объема анализируемой информации.

Система ключевых индикаторов рисков реализации национальных проектов

Важным этапом оперативного анализа национальных проектов является оценка вероятности достижения в срок входящих в них мероприятий, что в конечном итоге предопределяет успешность реализации проекта в целом. Для решения данной задачи нами был разработан комплект индикаторов, рассчитываемых на основе достоверной информации из периодических отчетов о ходе реализации проектов – например, доля выполненных контрольных точек федерального и регионального уровня, тенденция их достижения, наличие проблем в реализации мероприятий в прошлые годы и т.д. В этой связи, во-первых, возникает задача оценки важности (веса, значимости) каждого из индикаторов, а во-вторых задача прогнозирования вероятности выполнения мероприятия (или, наоборот, риска его невыполнения).

Для решения первой задачи возможно исследование ретроспективной информации об успешности реализации мероприятий в увязке с соответствующими индикаторами на основе различных подходов, таких как построение регрессий. Вторая задача может быть решена с помощью ML алгоритма случайного леса (Random Forest - RF)[3]. Суть метода состоит формировании ансамбля решающих деревьев и вычислении по ним ряда метрик, в том числе – MDI[4] (mean decrease impurity), которую можно трактовать как улучшение качества модели при добавлении в решающие деревья какого-либо фактора.

После определения перечня риск-индикаторов и получения по ним соответствующей информации в рамках периодического мониторинга возникает задача прогнозирования риска невыполнения мероприятий. Для качественной оценки может применяться скорингованя модель, суть которой состоит в подсчете количества «сработавших» индикаторов. Для количественной оценки вероятности выполнения мероприятия необходимо использовать более продвинутые алгоритмы, например ML-алгоритм K ближайших соседей (K-nearest neighbors, KNN), который используется преимущественно для задач классификации и состоит в следующем. Если в пространстве факторов, среди K ближайших соседей неклассифицированной точки больше точек одного класса, чем другого (например, «красных» больше чем «лиловых»), значит и неклассифицированная точка скорее всего «красная».

С помощью изложенных алгоритмов и в рамках принятого Счетной палатой риск-ориентированного подхода создана система ключевых индикаторов рисков, позволяющая заблаговременно выявлять проблемные мероприятия национальных проектов и оценивать соответствующие меры реагирования. Визуализация информации об индикаторах рисков в разрезе проектов и их мероприятий осуществлена благодаря дешборду.

Первые результаты анализа системы ключевых индикаторов рисков, проведенного на основе методологи ROC-анализа[5] по итогам 2024 года, указывают на хорошие прогностические возможности модели. Вместе с тем по ряду параметров модель требует доработки, а именно – повышения избирательности, что может быть достигнуто уточнением перечня используемых индикаторов рисков.

Развитие системы ключевых индикаторов рисков направлено на расширение спектра ее применения, в частности в составе критериев аудиторской существенности, и включает в себя задачи повышения точности и своевременности определения мероприятий проектов под риском неисполнения - разработку новых, научно обоснованных индикаторов (в том числе учитывающих специфику отдельных типов мероприятий – строительства, закупок и т.д.), увеличение частоты оценок и операционализацию их пост-контроля.

---

3 Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

4 https://scikit-learn.org/1.5/auto_examples/inspection/plot_permutation_importance.html

5 См., например, Davis, Jesse & Goadrich, Mark. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, ACM. 06. 10.1145/1143844.1143874.