УДК 004.89

МЕТОДЫ ВИЗУАЛИЗАЦИИ ВАЖНОСТИ ПРИЗНАКОВ ДЛЯ ПОВЫШЕНИЯ ПРОЗРАЧНОСТИ МОДЕЛЕЙ

Воробьёва А.И., студент гр. 23.М02-мкн, II курс
Санкт-Петербургский государственный университет, г. Санкт-Петербург

Повышение прозрачности и объяснимости моделей машинного обучения, особенно в контексте их применения в чувствительных и критически значимых сферах, требует не только разработки интерпретируемых архитектур, но и создания мощных инструментов, способных визуализировать внутренние механизмы принятия решений. Одним из наиболее эффективных и востребованных направлений является визуализация важности признаков– то есть представление информации о том, какие входные переменные (факторы, признаки, токены, пиксели и др.) сыграли ключевую роль в формировании вывода модели. Такие методы становятся центральным звеном в построении доверенных систем искусственного интеллекта, позволяя делать их решения не только проверяемыми и подотчётными, но и поддающимися адаптации, диагностике и нормативной интерпретации [1].

Наиболее общая идея визуализации важности признаков заключается в оценке вклада каждого элемента входа в итоговое предсказание. Это может быть реализовано различными способами: через вычисление производных (градиентов), посредством оценки маргинального влияния при выкидывании признака, через генеративное моделирование контрпримеров или с помощью теоретико-игровых подходов. Полученная количественная информация далее представляется в виде наглядных схем: графиков, тепловых карт, линейных баров, цветных матриц, или интерактивных интерфейсов, позволяющих пользователю исследовать вклад каждого признака.

Одним из наиболее известных и широко применяемых методов визуализации важности является SHAP (Shapley Additive Explanations), основанный на концепции шеплиевских значений из кооперативной теории игр. SHAP позволяет построить аддитивное разложение предсказания модели, где каждый признак получает свой вклад, интерпретируемый как «справедливое распределение» итогового значения между всеми признаками.

Метод поддерживает как глобальную, так и локальную интерпретацию, обладает математическими гарантиями консистентности и симметрии, и визуально представляется в виде горизонтальных полос с цветовой шкалой, показывающей как направление, так и силу влияния каждого признака.

Другим популярным методом является LIME (Local Interpretable Model-agnostic Explanations), который обучает локальную аппроксимацию модели (обычно линейную) вокруг рассматриваемого примера, а затем визуализирует веса этой аппроксимации как индикаторы важности. Хотя LIME чувствителен к выбору параметров и случайности генерации окрестности, он остаётся мощным и интуитивно понятным инструментом, особенно в приложениях, где требуется объяснение «на лету» для пользовательского интерфейса.

Для моделей глубокого обучения, особенно в задачах компьютерного зрения, применяются методы визуализации внимания и градиентного анализа, такие как Grad-CAM, Guided Backpropagation, DeepLIFT и другие. Они создают тепловые карты, накладывающиеся на входное изображение и показывающие, какие участки были наиболее значимыми для вывода. Подобные визуализации применяются в медицине (например, подсветка зоны опухоли на рентгене), в промышленности (обнаружение дефектов на поверхности), в автономных системах (анализ реакции модели на дорожные знаки и объекты) и других визуальных задачах [2].

В текстовых задачах, особенно с трансформерными архитектурами, визуализация важности признаков реализуется через анализ attention-матриц: отображение того, как токены влияют друг на друга в процессе генерации ответа или классификации [3].

Методы визуализации внимания включают attention flow, attention rollout, а также техники агрегации по слоям и головам. Кроме того, используются техники визуализации вкладов слов по метрикам типа SHAP или Integrated Gradients, позволяющие подсвечивать текст в зависимости от его значимости для вывода [4].

В задачах с табличными данными используются методы пермутационного анализа важности признаков (Permutation Importance), основанные на сравнении метрик модели до и после случайной перестановки значений конкретного признака [5].

Такие методы хорошо подходят для моделей типа градиентного бустинга (XGBoost, CatBoost, LightGBM) и часто используются для построения глобальных рейтингов признаков, что особенно важно в задачах с высокой нормативной нагрузкой (финансы, страхование, кредитный скоринг), где необходимо объяснить, почему одни признаки оказывают сильное влияние, а другие – нет.

Однако визуализация важности признаков сталкивается с рядом вызовов и ограничений, особенно в контексте доверия. Во-первых, визуализированные значения могут быть нестабильными: незначительные изменения входа могут радикально менять распределение значимости. Во-вторых, существует риск когнитивной перегрузкипользователя: при большом количестве признаков, особенно в моделях с несколькими входными потоками, визуализация становится

трудноинтерпретируемой. В-третьих, существует проблема обратной интерпретации: пользователи могут приписывать визуализации смысл, которого она не несёт, особенно если не объяснено, как именно рассчитываются важности.

Для решения этих проблем предлагается контекстуализация визуализаций – то есть дополнение визуального представления пояснительным текстом, интеграция в бизнес-логику интерфейса, фильтрация и ранжирование признаков по уровню значимости, использование цветовой кодировки, адаптивных легенд и пользовательских онтологий.

Например, в медицинской системе вместо отображения «весов признаков» можно представить: «пациенту назначено обследование, поскольку в крови повышен уровень X и выявлен симптом Y, оба существенно повышают риск диагноза Z». Такой подход делает визуализацию не просто технической, а коммуникативной, то есть понятной человеку в профессиональном контексте.

Наконец, важной задачей является интеграция визуализации в процессы разработки и сопровождения моделей. Это означает автоматическое логгирование объяснений, отображение изменений важности признаков при обновлении модели, использование визуализации в CI/CD пайплайнах, а также включение в отчётность (например, Model Cards или Fairness Reports). В этом направлении развиваются фреймворки, такие как Evidently AI, Fiddler, TruEra, InterpretML, Captum, а также инструменты внутри ML-платформ (Azure ML, SageMaker Clarify, Google What-If Tool), обеспечивающие воспроизводимость, версионность и аудит визуализации.

Таким образом, методы визуализации важности признаков становятся неотъемлемым элементом построения доверенных моделей машинного обучения. Они позволяют сделать поведение модели прозрачным, поддающимся проверке, критике и адаптации, обеспечивая основу для технической, юридической и этической подотчётности.

Их развитие требует не только математической точности, но и инженерной зрелости, когнитивной адаптации и нормативной совместимости – что делает визуализацию важности не просто вспомогательной функцией, а полноценным элементом цифрового доверия.

Список литературы:

1. Simonyan, K. Deep Inside Convolutional Networks : Visualising Image Classification Models and Saliency Maps / K. Simonyan, A. Vedaldi, A. Zisserman // arXiv. – 2013. – arXiv:1312.6034.

2. Smilkov, D. SmoothGrad : Removing Noise by Adding Noise / D. Smilkov, N. Thorat, B. Kim [et al.] // arXiv. – 2017. – arXiv:1706.03825.

3. Bach, S. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation / S. Bach, A. Binder, G. Montavon [et al.] //

PLoS ONE. – 2015. – Vol. 10, № 7. – Article e0130140. – DOI 10.1371/journal.pone.0130140.

4. Zeiler, M. D. Visualizing and Understanding Convolutional Networks / M. D. Zeiler, R. Fergus // Computer Vision – ECCV 2014. – Cham : Springer, 2014. – P. 818–833. – DOI 10.1007/978-3-319-10590-1_53.

5. Fisher, A. All Models are Wrong, but Many are Useful : Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously / A. Fisher, C. Rudin, F. Dominici // Journal of Machine Learning Research. – 2019. – Vol. 20. – P. 1–81.