Деградация модели (Model degradation)

> Синонимы: Model performance deterioration, AI aging, Ухудшение характеристик модели, Старение искусственного интеллекта

При реализации большинства (по некоторым оценкам — более 90%) аналитических проектов, использующих модели машинного обучения (ML-модели), возникает ситуация, когда после их ввода в промышленную эксплуатацию предсказательная эффективность модели начинает постепенно, а в некоторых случаях и быстро, ухудшаться. Это приводит к снижению качества управленческих решений, принимаемых на основе ее результатов, и соответствующим потерям бизнеса. Данное явление в машинном обучении и бизнес-аналитике получило название **деградация моделей**.

В широком смысле, под деградацией ML-модели понимается любое ухудшение качества ее работы после передачи в промышленную эксплуатацию, относительно качества, наблюдаемого по результатам обучения на обучающих данных.

Причин, по которым качество работы ML-моделей в промышленной среде оказывается ниже ожидаемого по результатам обучения, может быть множество. В качестве основных, как правило, выделяют:

- отсутствие репрезентативности обучающей выборки;
- переобучение;
- утечка данных;
- дрейф данных.

Первые два фактора не связаны с бизнес-средой, в которой развертывается модель, и могут быть учтены и скомпенсированы на этапе ее обучения. Утечка данных обычно приводит к тому, что модель сразу начинает работать не так, как ожидалось. А вот дрейф данных, который в большинстве случаев представляет собой длительный, постепенный процесс, вызывающий ухудшение работы модели, как правило и упоминается в качестве основной причины ее деградации.

Однако, как показали исследования, дрейф данных не всегда приводит к деградации модели, и наоборот, деградация модели может происходить в отсутствии значимого дрейфа данных. Таким образом, деградация моделей представляет собой сложное и еще плохо изученное явление.

Признаками деградации модели могут быть:

- рост ошибки модели или частоты неправильных классификаций с течением времени;
- высокая вариативность ошибки, когда модель работает то хорошо, то плохо;
- явная неадекватность предсказаний модели текущему состоянию бизнес-процессов.

Деградация модели может носить как временной, так и пространственный характер. В первом случае качество работы модели снижается по мере увеличения времени, прошедшего с момента ее ввода в промышленную эксплуатацию. Во втором — по мере перемещения места эксплуатации модели в другую локацию, где в бизнес-среде действуют зависимости и закономерности, отличные от тех, на которых обучалась модель.

Среди способов предотвращения пространственной деградации можно выделить **федеративное обучение**, когда модель обучается на данных, собранных во всех локациях, где потенциально будет использоваться модель. Для борьбы с временной деградацией может использоваться **трансферное обучение** — дообучение модели на реальных данных после ее разворачивания в промышленной среде, т.е. адаптация к реальным условиям, ее синхронизация с бизнес-средой.

Наиболее общий подход к борьбе с деградацией ML-модели включает следующие мероприятия:

1. Включить в конвейер анализа данных, в котором работает модель, механизмы сигнализации о снижении качества модели ниже критического уровня.
2. Разработать и реализовать средства автоматического повторного обучения.
3. Обеспечить постоянный доступ к данным, которые объективно описывают бизнес-процессы, анализируемые моделью (такие данные часто называют **ground truth** — основная, наземная истина).

При мониторинге деградации модели и повторном обучении часто используют различные способы взвешивания наблюдений. При этом величина веса обратно пропорциональна его возрасту. Поэтому «старые» наблюдения при повторном обучении модели будут обеспечивать меньший вклад в ее настройку, чем более новые и актуальные.

Очевидно, что использование этих мер приводит к увеличению затрат на создание, развертывание и эксплуатацию аналитических проектов. Однако, учитывая, что деградация модели может привести к ее полной непригодности и значительным потерям в бизнесе, такие затраты, как правило, оправдывают себя.

Подробнее ознакомиться с такими явлениями в машинном обучении как утечка и дрейф данных можно в статьях «Утечка данных в машинном обучении» и «Дрейф данных».