Предсказательная аналитика (Predictive analytics)

> Синонимы: Предиктивная аналитика, Прогнозная аналитика
> Разделы: Бизнес-задачи

Предсказательная аналитика — направление в интеллектуальном анализе данных, использующее методы статистики и машинного обучения с целью предсказания значений неизвестных, но представляющих интерес значений признаков, описывающих объекты и процессы, на основе известных.

Типичными примерами задач предсказательной аналитики могут быть:

- **кредитный скоринг** — на основе известных значений таких признаков, как доход, возраст, образование и др. требуется предсказать уровень кредитоспособности заемщика;
- **определение уровня лояльности клиентов** — используя признаки, описывающие поведение клиента (число обращений и интервал между ними, сумма, потраченная на покупки, реакция на маркетинговые акции) требуется предсказать вероятность его оттока;
- **предсказание сбоев на производстве** — на основе данных о работе оборудования предсказывается вероятность его выхода из строя.

Общая постановка задачи предсказательной аналитики может быть описана следующим образом. Пусть некоторый объект или процесс, закономерности поведения которого требуется определить, характеризуется набором признаков с известными значениями $X = (x_1, x_2, \ldots, x_n)$. Кроме этого, имеется набор признаков, значения которых неизвестны, но представляют интерес для аналитика $Y = (y_1, y_2, \ldots, y_m)$.

При этом предполагается, что Y зависит от X и если построить модель, которая сможет обнаружить и аппроксимировать эту зависимость, то ее можно будет использовать для предсказания Y по X. Такие модели называются **предсказательными**, а технология их построения известна как **предсказательное моделирование**.

К числу наиболее известных задач предсказательного моделирования относятся:

- **численное предсказание** — имеет место, когда X и Y представляют собой непрерывные (вещественные) данные. Для его реализации используются такие виды предсказательных моделей как нейронные сети и линейная регрессия;
- **классификация** — в этом случае Y представляют собой метку класса. Реализуется с помощью нейронных сетей, деревьев решений, дискриминантного анализа, SVM, логистической регрессии, алгоритм k-ближайших соседей;

- **кластеризация** — Y принимает значения номеров кластеров. Решается с использованием сетей Кохонена, алгоритмов k-средних, CLOPE, DBSCAN, EM, иерархических и др.
- **ассоциация** — использует поиск ассоциативных правил и последовательных шаблонов.

Методы и модели, используемые в предсказательной аналитике, существенно зависят от характера анализируемых данных. Например, если данные исторические (т.е. временные ряды), то на их основе могут предсказываться будущие события (т.е. зависимости, обнаруженные в ретроспективных данных экстраполируются на будущее). В этом случае имеет место задача прогнозирования.

Следует отметить, что некоторые авторы рассматривают предсказательную аналитику именно как прогнозную, т.е. использующую исторические данные для предсказания поведения объектов и процессов в будущем, т.е. ответа на вопрос "Что, скорее всего, произойдет в будущем?". Однако предсказательная аналитика может работать с информацией, не связанной со шкалой времени и не образующей временных рядов.

Например, пространственные данные, в которых наблюдения представляют собой географически распределенные объекты (скажем, торговые точки), значения признаков которых регистрировались в фиксированный момент времени. Что дает возможность отвечать на вопрос "Что происходит сейчас?". Проблемами предсказания будущих событий в анализе данных занимается другое направление, известное как прогнозирование (forecasting)

Типичный проект в области предсказательной аналитики содержит следующие этапы:

1. **Сбор данных**. Он может производится как из источников внутри компании (учетных систем, CRM, ERP, отчетов, шаблонов поведения клиентов при посещения сайта), так и из внешних (соцсетей, публикаций органов статистики, открытых баз данных и т.д.).
2. **Предобработка данных**. В свою очередь содержит трансформацию и очистку данных.
3. **Предсказательное моделирование**. Модели обучаются и тестируются, а затем применяются к собранным и подготовленным данным.
4. **Интерпретация результатов**. Результаты обычно выражаются в вероятностях или сценариях поведения (например, «вероятность ухода клиента — 72%»). Аналитика должна быть понятна бизнесу, иначе прогнозы теряют свою ценность.
5. **Внедрение и автоматизация**. Предсказательные модели интегрируются в бизнес-процессы: CRM, ERP, маркетинговые платформы, системы управления рисками и т.д. Автоматизация позволяет использовать предсказательные модели в реальном времени, например для онлайн-оценки кредитоспособности заемщика.
6. **Непрерывное обновление моделей**. Данные и скрытые в них закономерности со временем меняются, что приводит к деградации моделей, поэтому они должны регулярно тестироваться и при необходимости дообучаться.

Таким образом, предсказательную аналитику следует рассматривать не как разовый проект, а как рабочий процесс с достаточно протяженным жизненным циклом.