ПОСТРОЕНИЕ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ

В. Н. Ткаченко, к. т. н. П. С. Сафронов

Одесский национальный политехнический университет
Украина, г. Одесса
vitalkagrsp@gmail.com, sps@opu.ua

Определена предметная область и основные задачи, которые ставит перед собой предсказательная аналитика. Описана методология построения предсказательной модели. Выбран инструментарий для построения предсказательной модели, реализована линейная регрессионная модель, позволяющая предсказывать объем ежегодных продаж для коммерческой компании, и проведена оценка данной модели.

Ключевые слова: предсказательная аналитика, регрессионная модель, оценка модели.

Предсказательная аналитика — область статистики, которая занимается извлечением информации из данных и ее использованием для прогнозирования тенденций и моделей поведения. Ядро предсказательной аналитики полагается на захват отношений между объясняющими переменными и предсказанными переменными из прошлых случаев. Предсказательная аналитика широко применяется в таких областях и направлениях, как маркетинг, финансовые услуги, здравоохранение, страхование, розничная торговля, телекоммуникации, планирование потенциала.

В типичном сценарии присутствует измерение результата, обычно количественное (например, цена на акции) или категориальное (например, сердечный приступ / без сердечной атаки), который необходимо предсказать на основе набора признаков (например, таких как диета и клинические измерения). При этом присутствует обучающий набор данных, в котором наблюдаются результаты и характеристики объекта для набора объектов (например, людей). Используя эти данные, строится модель прогнозирования или ученика, которая позволит предсказать результат для новых, невидимых ранее, объектов [1].

Подходы и методы, используемые для проведения прогнозной аналитики, можно в целом сгруппировать в методы регрессии и методы машинного обучения.

Регрессионные модели являются основой предсказательной аналитики. Основное внимание уделяется установлению математического уравнения в качестве модели для представления взаимодействий между рассматриваемыми различными переменными. Линейные модели были в значительной степени развиты в предкомпьютерный период статистики, но даже в сегодняшнюю компьютерную эру все еще есть веские основания изучать и использовать их. Они просты и часто обеспечивают адекватное и интерпретируемое описание того, как входы влияют на выход. Для целей прогнозирования они иногда могут опережать более необычные нелинейные модели, особенно в ситуациях с небольшим количеством учебных случаев, низким отношением сигнал/шум или разреженными данными [2].

В общем процесс предсказательной аналитики состоит из таких этапов: определение проекта, сбор данных, анализ данных, статистический анализ, моделирование, развертывание модели, мониторинг модели.

Целью данной работы является построение предсказательной модели, позволяющей предсказывать объем продаж коммерческой компании на основе статистических демографических и коммерческих данных в зависимости от количества мест расположения точек продаж.

Задан входной вектор $X^T = (X_1, X_2,\ldots, X_p)$ и необходимо предсказать вещественный вывод Y. Модель линейной регрессии имеет вид

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j .$$

Линейная модель либо предполагает, что функция регрессии $E(Y \mid X)$ линейна, либо линейная модель является разумным приближением, где β_j — неизвестные параметры или коэффициенты, а переменные X_j могут исходить из разных источников:

— количественные данные;

— преобразования количественных данных, такие как логарифмирование, квадратный корень или квадрат;

— базисные разложения, такие как $X_2 = X_1^2$, $X_3 = X_1^3$, что приводит к представлению полиномов;

— числовое или «фиктивное» кодирование уровней количественных входных данных. Например, если G является входным коэффициентом пятиступенчатого фактора, то можно создать X_j, $j = 1, \ldots, 5$, так, что $X_j = I(G = j)$. Вместе эта группа X_j представляет собой элемент G с помощью набора констант, зависящих от уровня, поскольку в $\sum_{j=1}^{5} X_j \beta_j$ одно из значений X_j равно единице, а остальные равны нулю;

— взаимодействие между переменными, например $X_3 = X_2 \cdot X_1$.

Независимо от источника X_j модель линейна по параметрам.

Наиболее популярным методом оценки регрессионной модели является метод наименьших квадратов, когда выбираются коэффициенты $\beta (\beta_1, \beta_2, \ldots, \beta_p)^T$ для минимизации остаточной суммы квадратов:

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - f\left(x_i\right)\right)^2 = \sum_{i=1}^{\mathrm{N}} \left(y_i - \beta_0 - \sum_{j=1}^{p} \left(x_{ij} \cdot \beta_j\right)^2 \right).$$

Со статистической точки зрения этот критерий разумен, если учебные наблюдения (x_i, y_i) представляют собой независимые случайные розыгрыши из их популяции. Даже если x_i не были выбраны случайным образом, критерий остается в силе, если y_i являются условно независимыми, учитывая входы x_i.

В результате проведенных исследований в области предсказательной аналитики была реализована модель предсказания ежегодных продаж для коммерческой компании. С использованием статистических демографических и коммерческих данных был дан прогноз ежегодных продаж в 10000 региональных местах. Целевой или зависимой переменной в этом случае выступал годовой доход коммерческой компании. Наилучшие результаты были достигнуты при применении множественной линейной регрессии, что еще раз показало эффективность и целесообразность использования модели линейной регрессии для предсказательных задач.

Открытие новой точки продаж требует больших инвестиций времени и капитала. Но если выбрано неправильное место для ее размещения, то она он закрывается в течение 18 месяцев и приносит большие операционные убытки. Поиск математической модели для повышения эффективности инвестиций в новые точки продаж позволит компании инвестировать больше в другие важные сферы бизнеса, такие как устойчивость, инновации и обучение новых сотрудников.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — New York: Springer-Verlag New York Inc., 2009.

2. Johnson K, Kuhn M. Applied Predictive Modeling.— New York: Springer-Verlag New York Inc., 2013.

———————

V. N. Tkachenko, P. S. Safronov

Development of predictive models

The subject domain and the primary objectives of predictive analytics are determined. The methodology for development of the predictive models is described. The tools for the predictive model development are chosen and the linear regression model, which allows predicting the annual sales for a commercial company, is implemented and estimated.

Keywords: predictive analytics, regression model, model estimation.