Predictive analytics in the era of big data

openaccessgovernment.org/article/predictive-analytics-in-the-era-of-big-data/185046

13 November 2024

Jacob Zahavi, from The Coller School of Management at Tel Aviv University, discusses the components of predictive analytics (PA) and the increasing complexity of PA problems in the world of Big Data

This is the era of Big Data. Big Data are datasets that are too large and complex to handle using traditional data-processing tools and software. Data nowadays comes in huge quantities (volume), in different forms (velocity), and in a multitude of formats (variety). Some also add ‘value’ to this list. While raw data doesn’t have a value of its own, it carries a lot of hidden information essential for decision-making. For example, in marketing, data reflects the characteristics of potential buyers; in healthcare, which symptoms affect certain diseases; in insurance, which customers are more prone to risk; in banking, which customers are more likely to return their loans; and more.

However, this information doesn’t surface on its own, and one needs to dig out this information from the Big Data using Data Mining and Machine Learning technologies. In this article, we focus on Predictive Analytics (PA), one of the hottest technologies of data mining, which is concerned with predicting future events, usually responses, from historical data (e.g., in marketing, whether a customer will respond to a solicitation to buy a product or a service). The variables used to predict the response are predictors (also referred to as features, explanatory variables, or attributes). The large number of observations, which may run in the millions, and the large number of attributes, which may run in the thousands, make PA problems excessively complex.

Predictive analytics and data preparation

Indeed, PA is a process that comprises three major components: data preparation, modeling, and implementation. Data preparation is often the most time-consuming of the PA process. The starting point is extracting the relevant data for the application involved (the ‘universe’) from the organizational databases, for which the response values are known, e.g., in marketing the audience of a previous campaign for the same or “similar” offer, (the characteristics of which usually depend on the application domain), and then randomly selecting a sample from this audience to create the testing dataset. Data preparation often requires preprocessing the universe to standardize attributes, handle missing values and outliers, enrich the universe from outside sources (e.g., demographics), and so on. The final objective is to create a flat file that includes the response variable and all the relevant predictors.

Predictive analytics and data modeling

The following step is modeling. A multitude of models, the leading models of which are regression-driven, have been developed to predict response. By and large, prediction problems are divided into two main categories – classification of an observation to one of several discrete predefined classes (e.g., buyer and non-buyer) and estimation – predicting the value of a continuous response (e.g., donation amounts in a charity campaign). Particular attention should be given to time-dependent events in which observations are censored, which should be analyzed by means of models from the realm of survival analysis. It is beyond this article to discuss the modeling issues, but they are widely discussed in the literature. [1]

Building a stable and generalized model

PA is a proactive process with the goal of building models that are generalized enough to allow for the prediction of responses for unseen data and new observations. The major problem afflicting PA in Big Data is the over- fitting phenomenon, namely a model that yields very accurate predictions when applied to the testing dataset but wrong predictions when applied to new observations. The standard method to detect over-fitting is to split the testing dataset into two exhaustive and mutually exclusive subsets – a training dataset for building the model and a validation dataset for evaluating the model's accuracy. Since the validation dataset also contains the actual response values, the idea is to apply the modeling results in the form of an equation, tree, pattern, formula, or other (which depends on the particular model involved) to predict response on the validation dataset and then compare the predicted and the actual response. If the results are pretty close, the model is regarded as stable enough to apply to predict responses for new observations. The clue to building a stable and generalized model is feature selection, namely selecting the most influential attributes affecting response from the larger set of potential predictors. The feature selection problem is definitely the most challenging issue of the PA process, and a variety of methods have been devised to address this problem. [2][3] Various metrics have been developed to assess the prediction accuracy of the model.

Predicting response for unseen data

The final step of the PA process is predicting response for unseen data (scoring), by applying the modeling results to the new observations – often the occurrence probabilities for classification problems and the expected response values for estimation problems. Note that for scoring, the new observations must have the same set of features as in the testing dataset. The prediction model needs to be calibrated over time to cope with the dynamics of real-world problems, ensure it is still generalized enough, and meet the accuracy standards.

PA has diverse applications in various fields, including marketing, finance and risk, telecommunication, healthcare and medical, social media, supply chain management, and more.

Future trends

PA is, indeed, a very complex process primarily because of the excessive number and the variety of attributes involved. As data is becoming more complex and data analysis tools more advanced, modern PA problems are becoming more and more challenging, coping with the complexity of the prediction issues and the diverse data sources. More sophisticated technologies that are capable of dealing with messy and unstructured data are therefore required to 'remove the chaff from the hey' and render accurate and stable prediction models. A recent example is the highly spoken about ChatGPT, which is basically a prediction problem of the next word to insert in a sentence or a paragraph. ChatGPT was trained using deep learning technology based on the entire English language vocabulary (about 55 thousand words). As PA problems are getting more challenging and diverse, advanced AI and Gen AI tools are likely to take an increasing role in addressing the larger complexity of PA problems in the world of Big Data.

References

1. Hastie T, et al., The Elements of Statistical Learning, Springer, 2009
2. Miller A. J., Subset Selection in Regression, Chapman Hall, 2002
3. Zahavi J, Predictive Analytics for Targeting Decisions, Machine Learning for Data Science Handbook, Rokach L, et al., editors, 2023, p. 751-777.

Contributor Details

- Article Categories
- Technology

- Article Tags
- Computing
- Technological Innovations

- Publication Tags
- OAG 045 - January 2025

- Stakeholder Tags
- SH - Coller School of Management