Predictive modelling

Predictive modelling uses statistics to predict outcomes.[1] Most often the event one wants to predict is in the future, but predictive modelling can be applied to any type of unknown event, regardless of when it occurred. For example, predictive models are often used to detect crimes and identify suspects, after the crime has taken place.[2]

In many cases, the model is chosen on the basis of detection theory to try to guess the probability of an outcome given a set amount of input data, for example given an email determining how likely that it is spam.

Models can use one or more classifiers in trying to determine the probability of a set of data belonging to another set. For example, a model might be used to determine whether an email is spam or "ham" (non-spam).

Depending on definitional boundaries, predictive modelling is synonymous with, or largely overlapping with, the field of machine learning, as it is more commonly referred to in academic or research and development contexts. When deployed commercially, predictive modelling is often referred to as predictive analytics.

Predictive modelling is often contrasted with causal modelling/analysis. In the former, one may be entirely satisfied to make use of indicators of, or proxies for, the outcome of interest. In the latter, one seeks to determine true cause-and-effect relationships. This distinction has given rise to a burgeoning literature in the fields of research methods and statistics and to the common statement that "correlation does not imply causation".

Models

Nearly any statistical model can be used for prediction purposes. Broadly speaking, there are two classes of predictive models: parametric and non-parametric. A third class, semi-parametric models, includes features of both. Parametric models make "specific assumptions with regard to one or more of the population parameters that characterize the underlying distribution(s)".[3] Non-parametric models "typically involve fewer assumptions of structure and distributional form [than parametric models] but usually contain strong assumptions about independencies".[4]

Applications

Uplift modelling

Uplift modelling is a technique for modelling the *change in probability* caused by an action. Typically this is a marketing action such as an offer to buy a product, to use a product more or to re-sign a contract. For example, in a retention campaign you wish to predict the change in probability that a customer will

remain a customer if they are contacted. A model of the change in probability allows the retention campaign to be targeted at those customers on whom the change in probability will be beneficial. This allows the retention programme to avoid triggering unnecessary churn or customer attrition without wasting money contacting people who would act anyway.

Archaeology

Predictive modelling in archaeology gets its foundations from Gordon Willey's mid-fifties work in the Virú Valley of Peru.[5] Complete, intensive surveys were performed then covariability between cultural remains and natural features such as slope and vegetation were determined. Development of quantitative methods and a greater availability of applicable data led to growth of the discipline in the 1960s and by the late 1980s, substantial progress had been made by major land managers worldwide.

Generally, predictive modelling in archaeology is establishing statistically valid causal or covariable relationships between natural proxies such as soil types, elevation, slope, vegetation, proximity to water, geology, geomorphology, etc., and the presence of archaeological features. Through analysis of these quantifiable attributes from land that has undergone archaeological survey, sometimes the "archaeological sensitivity" of unsurveyed areas can be anticipated based on the natural proxies in those areas. Large land managers in the United States, such as the Bureau of Land Management (BLM), the Department of Defense (DOD),[6][7] and numerous highway and parks agencies, have successfully employed this strategy. By using predictive modelling in their cultural resource management plans, they are capable of making more informed decisions when planning for activities that have the potential to require ground disturbance and subsequently affect archaeological sites.

Customer relationship management

Predictive modelling is used extensively in analytical customer relationship management and data mining to produce customer-level models that describe the likelihood that a customer will take a particular action. The actions are usually sales, marketing and customer retention related.

For example, a large consumer organization such as a mobile telecommunications operator will have a set of predictive models for product cross-sell, product deep-sell (or upselling) and churn. It is also now more common for such an organization to have a model of savability using an uplift model. This predicts the likelihood that a customer can be saved at the end of a contract period (the change in churn probability) as opposed to the standard churn prediction model.

Auto insurance

Predictive modelling is utilised in vehicle insurance to assign risk of incidents to policy holders from information obtained from policy holders. This is extensively employed in usage-based insurance solutions where predictive models utilise telemetry-based data to build a model of predictive risk for claim likelihood. Black-box auto insurance predictive models utilise GPS or accelerometer sensor input only. Some models include a wide range of predictive input beyond basic telemetry including advanced driving behaviour, independent crash records, road history, and user profiles to provide improved risk models.

Health care

In 2009 Parkland Health & Hospital System began analyzing electronic medical records in order to use predictive modeling to help identify patients at high risk of readmission. Initially, the hospital focused on patients with congestive heart failure, but the program has expanded to include patients with diabetes, acute myocardial infarction, and pneumonia.[8]

In 2018, Banerjee et al.[9] proposed a deep learning model for estimating short-term life expectancy (>3 months) of the patients by analyzing free-text clinical notes in the electronic medical record, while maintaining the temporal visit sequence. The model was trained on a large dataset (10,293 patients) and validated on a separated dataset (1818 patients). It achieved an area under the ROC (Receiver Operating Characteristic) curve of 0.89. To provide explain-ability, they developed an interactive graphical tool that may improve physician understanding of the basis for the model's predictions. The high accuracy and explain-ability of the PPES-Met model may enable the model to be used as a decision support tool to personalize metastatic cancer treatment and provide valuable assistance to physicians.

The first clinical prediction model reporting guidelines were published in 2015 (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)), and have since been updated.[10]

Predictive modelling has been used to estimate surgery duration.

Algorithmic trading

Predictive modeling in trading is a modeling process wherein the probability of an outcome is predicted using a set of predictor variables. Predictive models can be built for different assets like stocks, futures, currencies, commodities etc. Predictive modeling is still extensively used by trading firms to devise strategies and trade. It utilizes mathematically advanced software to evaluate indicators on price, volume, open interest and other historical data, to discover repeatable patterns.[11]

Lead tracking systems

Predictive modelling gives lead generators a head start by forecasting data-driven outcomes for each potential campaign. This method saves time and exposes potential blind spots to help client make smarter decisions.[12]

Notable failures of predictive modeling

Although not widely discussed by the mainstream predictive modeling community, predictive modeling is a methodology that has been widely used in the financial industry in the past and some of the major failures contributed to the 2008 financial crisis. These failures exemplify the danger of relying exclusively on models that are essentially backward looking in nature. The following examples are by no mean a complete list:

1. Bond rating. S&P, Moody's and Fitch quantify the probability of default of bonds with discrete variables called rating. The rating can take on discrete values from AAA down to D. The rating is a predictor of the risk of default based on a variety of variables associated with the borrower and historical macroeconomic data. The rating agencies failed with their ratings on the US$600 billion mortgage backed Collateralized Debt Obligation (CDO) market. Almost

the entire AAA sector (and the super-AAA sector, a new rating the rating agencies provided to represent super safe investment) of the CDO market defaulted or severely downgraded during 2008, many of which obtained their ratings less than just a year previously.
2. So far, no statistical models that attempt to predict equity market prices based on historical data are considered to consistently make correct predictions over the long term. One particularly memorable failure is that of Long Term Capital Management, a fund that hired highly qualified analysts, including a Nobel Memorial Prize in Economic Sciences winner, to develop a sophisticated statistical model that predicted the price spreads between different securities. The models produced impressive profits until a major debacle that caused the then Federal Reserve chairman Alan Greenspan to step in to broker a rescue plan by the Wall Street broker dealers in order to prevent a meltdown of the bond market.

Possible fundamental limitations of predictive models based on data fitting

History cannot always accurately predict the future. Using relations derived from historical data to predict the future implicitly assumes there are certain lasting conditions or constants in a complex system. This almost always leads to some imprecision when the system involves people.

Unknown unknowns are an issue. In all data collection, the collector first defines the set of variables for which data is collected. However, no matter how extensive the collector considers his/her selection of the variables, there is always the possibility of new variables that have not been considered or even defined, yet are critical to the outcome.

Algorithms can be defeated adversarially. After an algorithm becomes an accepted standard of measurement, it can be taken advantage of by people who understand the algorithm and have the incentive to fool or manipulate the outcome. This is what happened to the CDO rating described above. The CDO dealers actively fulfilled the rating agencies' input to reach an AAA or super-AAA on the CDO they were issuing, by cleverly manipulating variables that were "unknown" to the rating agencies' "sophisticated" models.

See also

- Calibration (statistics)
- Prediction interval
- Predictive analytics
- Predictive inference
- Statistical learning theory
- Statistical model

References

1. Geisser, Seymour (1993). *Predictive Inference: An Introduction*. Chapman & Hall. p. . ISBN 978-0-412-03471-8.
2. Finlay, Steven (2014). *Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods* (1st ed.). Palgrave Macmillan. p. 237. ISBN 978-1-137-37927-6.

3. Sheskin, David J. (April 27, 2011). *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press. p. 109. ISBN 978-1-4398-5801-1.
4. Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press. p. 2.
5. Willey, Gordon R. (1953), "Prehistoric Settlement Patterns in the Virú Valley, Peru", Bulletin 155. Bureau of American Ethnology
6. Heidelberg, Kurt, et al. "An Evaluation of the Archaeological Sample Survey Program at the Nevada Test and Training Range", SRI Technical Report 02-16, 2002
7. Jeffrey H. Altschul, Lynne Sebastian, and Kurt Heidelberg, "Predictive Modeling in the Military: Similar Goals, Divergent Paths", Preservation Research Series 1, SRI Foundation, 2004
8. "Hospital Uses Data Analytics and Predictive Modeling To Identify and Allocate Scarce Resources to High-Risk Patients, Leading to Fewer Readmissions" (https://innovations.ahrq.gov/profiles/hospital-uses-data-analytics-and-predictive-modeling-identify-and-allocate-scarce-resources). Agency for Healthcare Research and Quality. 2014-01-29. Retrieved 2019-03-19.
9. Banerjee, Imon; et al. (2018-07-03). "Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met) Utilizing Free-Text Clinical Narratives" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030075). *Scientific Reports*. **8** (10037 (2018)) 10037. Bibcode:2018NatSR...810037B (https://ui.adsabs.harvard.edu/abs/2018NatSR...810037B). doi:10.1038/s41598-018-27946-5 (https://doi.org/10.1038%2Fs41598-018-27946-5). PMC 6030075 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030075). PMID 29968730 (https://pubmed.ncbi.nlm.nih.gov/29968730).
10. Collins, Gary; et al. (2024-04-16). "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11019967). *BMJ*. **385** e078378. doi:10.1136/bmj-2023-078378 (https://doi.org/10.1136%2Fbmj-2023-078378). PMC 11019967 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11019967). PMID 38626948 (https://pubmed.ncbi.nlm.nih.gov/38626948).
11. "Predictive-Model Based Trading Systems, Part 1 - System Trader Success" (http://systemtradersuccess.com/predictive-model-based-trading-systems-2/). *System Trader Success*. 2013-07-22. Retrieved 2016-11-25.
12. "Predictive Modeling for Call Tracking" (https://phonexa.uk/call-logic/predictive-modeling-call-logic/). *Phonexa*. 2019-08-22. Retrieved 2021-02-25.

Further reading

- Clarke, Bertrand S.; Clarke, Jennifer L. (2018), *Predictive Statistics*, Cambridge University Press
- Iglesias, Pilar; Sandoval, Mônica C.; Pereira, Carlos Alberto de Bragança (1993), "Predictive likelihood in finite populations" (https://www.researchgate.net/publication/259975170), *Brazilian Journal of Probability and Statistics*, **7** (1): 65–82, JSTOR 43600831 (https://www.jstor.org/stable/43600831)
- Kelleher, John D.; Mac Namee, Brian; D'Arcy, Aoife (2015), *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, worked Examples and Case Studies*, MIT Press
- Kuhn, Max; Johnson, Kjell (2013), *Applied Predictive Modeling*, Springer
- Shmueli, G. (2010), "To explain or to predict?", *Statistical Science*, **25** (3): 289–310, arXiv:1101.0891 (https://arxiv.org/abs/1101.0891), doi:10.1214/10-STS330 (https://doi.org/10.1214%2F10-STS330), S2CID 15900983 (https://api.semanticscholar.org/CorpusID:15900983)