Predictive Modelling

Document 218081

Turning Big Data into Big Opportunities

Canadian Institute of Actuaries

Institut canadien des actuaires

Essays on Predictive Modelling: Turning Big Data into Big Opportunities

In recent years, data has become a key driver of economic growth and the foundation on which industries are being built. For many, the question of what to do with this complex raw material has become a key organizational challenge. That's where actuaries come in. Using predictive modelling, they can help turn your big data into big opportunities.

Predictive modelling is the analysis of sets of data to identify meaningful relationships, and the use of these relationships to better predict outcomes and make better, faster, actionable decisions. It uses historical information to describe past relationships, from which to draw insights about the future. These insights can apply to several aspects of a business, such as consumer, provider, and distributor behaviour.

Predictive modelling draws on many disciplines, including statistics, modelling, optimization, clustering, market research, and computer programming. Its application generally relies on substantial computer power and overlaps with fields such as machine learning and artificial intelligence.

Why Actuaries?

Predictive modelling shares similarities with actuarial science. Actuaries examine the relationships within large data sets and relate them to real-world business problems, traditionally in the context of an insurance company, a pension plan, or a risk management function. The models they develop are implemented in several key business functions that have an impact on the bottom line.

Actuaries are trained in mathematics with a focus on building models and solving complex problems with financial consequences. With the right analytical skills, the rigour of a professional, an aptitude for computer science, and an ability to look beyond the complex math, actuaries excel at developing business solutions in an uncertain and changing environment.

Beyond the traditional settings they are familiar with, actuaries are sought by organizations in banking, investment management, e-commerce, weather risk management, transportation, energy, and social programs.

Learn more about how predictive modelling, and actuaries, can help your organization.

The Canadian Institute of Actuaries' Predictive Modelling Committee would like to recognize the contributions of our committed volunteers, the many valued authors, and the staff at the CIA Head Office to this project.

Why now?

Predictive modelling is not new. In fact, business has used core techniques such as logistic regression for decades. There are many reasons why predictive modelling has gained momentum now:

- Companies face immense competitive pressure to differentiate and provide better customer experience and to streamline processes.
- Big data revolution. We now create 2.5 quintillion bytes of data each day[1], suitable for analytics.
- Availability of new data types such as social media data, web data, sensor data, audio, and images.
- Significant improvements in computing power with high-speed and distributed parallel processing at low cost.
- Decrease in data storage cost.
- Highly scalable new technology to store and manage structured and unstructured data, e.g., Hadoop.

- Cloud storage and cloud computing.
- New innovations in machine learning, deep learning, and artificial intelligence.
- Availability of open-source data sets.
- Open source, free software (R and Python).

What problems can analytics solve?

The insights gleaned from predictive models can apply to several aspects of a business, including consumer, provider, and distributor behaviour. Predictive modelling can enhance many business processes including the following:

- **Sales and marketing:** identify target sales groups, identify individual characteristics correlated with purchase decision, understand purchase behaviours and recommend the right product, match prospective clients with the most appropriate sales agent;
- **Customer experience:** provide tailored services and relevant information to customers;
- **Current business management:** identify and retain clients, offer additional products to current customers, profile customers;
- **Pricing:** improve pricing accuracy, project impact of deviations from pricing parameters;
- **Risk management:** determine range of outcomes of key performance metrics, capital/equity modelling;
- **Fraud detection:** identify likely fraudulent activities, respond quickly to fraud suspicion, find fraud patterns; and
- **HR analytics:** direct employee to best functions, improve employee retention, assess impact of human resources policies on performance.

In insurance companies, in addition to the above applications, predictive modelling can enhance the following:

- **Underwriting:** identify best risks and prioritize acceptance efforts, identify applicants for whom additional underwriting is needed, support simplified underwriting;
- **Claims:** predict claim frequency and severity, claims triage, prioritize claims management resources;

- **Reserving:** reserve more accurately; and
- **Experience analysis:** identify experience drivers, improve mortality/lapse assumption modelling.

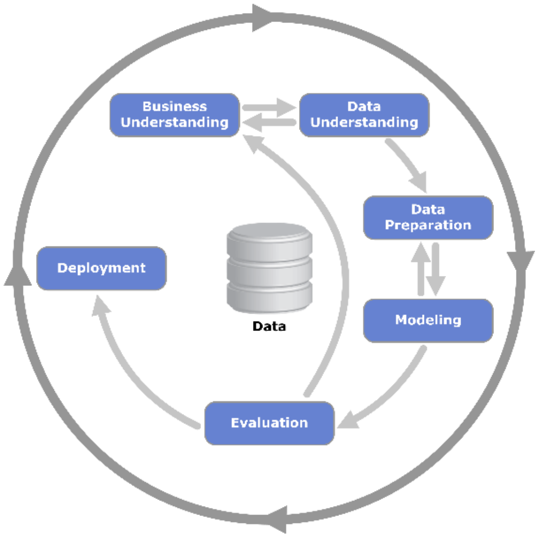
What does a typical predictive modelling process involve?

Developing the right solutions to business problems using predictive modelling requires close collaboration with business subject matter experts from start to finish. The following are central to the process:

1. Identify a problem where predictions of future outcomes or behaviour can enhance the accuracy or efficiency of business decision-making.

2. Understand the business: know the products, the needs of the stakeholders, the resources and data available, identify assumptions and constraints, and the means of implementing a predictive modelling solution.

3. Clearly define the outcome to be predicted (the response variable) by the predictive model.



Understanding Data

The first step in data understanding is to identify the data necessary to develop the predictive model. This involves understanding the data sources and data flow so that the data can be collected. Once the data is gathered, initial data exploration needs to be performed to get familiar with the data, discover insights, and to identify and rectify data quality issues. Data is often not perfect, and it is important that the limitations and solutions to data issues are well understood and documented.

Data collection and preparation are crucial steps. Data must be explored, cleansed, combined, transformed, and formatted. The resulting data set is then split into two or more partitions, so that part of the data can “train” (calibrate) the model, and the remainder of the data can evaluate the model’s performance. A data governance structure helps ease the process and avoids wasting time at further iterations. This usually involves automated data collection processes from internal and external sources, documented data treatment and transformation, data quality threshold rules, and data warehousing.

The Art of Science

Modelling is where art meets science. The art comes from the a priori intuition and assumptions on the relationship between each predictor variable—i.e., the variables used to predict the outcome—and the response variable. The science comes from running quantitative analysis on the data and using a large collection of models. Model selection involves limiting the number of parameters and ultimately selecting a model that is understandable and actionable by stakeholders.

The model should be evaluated on data that the model has never seen before. A predictive model must be generalizable to unseen data, not just describe the data used to calibrate the model. The model may be inapplicable, incorrect, unstable, or used inappropriately. Model evaluation conducted regularly helps keep trust in the model and improves on it by doing new iterations of the process.

Deployment

If the model performs well and business stakeholders accept it, the next step is model deployment. Usually, the IT department interacts with the business people to integrate the predictive model in the organization's processes. The deployment usually includes a testing period in a non-production environment before moving on to full implementation. This testing period allows modellers to identify and correct discrepancies.

Production models should be reviewed on a regular basis, comparing actual data to model predictions to ensure that performance has not substantially deteriorated. The model may need to be refreshed or recreated if there is evidence of deterioration.

What are the techniques?

Predictive modelling techniques can be classified in a number of ways. The summary below, while not a comprehensive list of all techniques available, provides a broad categorization of those used.

I. Supervised learning model category

In a supervised learning problem, the modeller uses a data set where the values of the response variable are known. For example, to predict whether a policyholder will lapse, a modeller may use a set of historical policy data in which each customer's decision to lapse is known and already coded in the data. The majority of the current applications in predictive modelling for insurance are based in supervised learning techniques.

Within supervised learning, there are two primary subsets of model:

i. Classification models

In a classification problem, the objective is to predict a categorical outcome. This could include a binary response (i.e., 1 if policy lapsed; 0 if policy did not lapse), or could include a broad categorization/multiple states (i.e., 1 if healthy, 2 if disabled, 3 if retired, 4 if deceased).

ii. **Regression models**

In a regression problem, the objective is to predict a continuous outcome; for example, the severity of an auto insurance collision claim.

There are modelling techniques that can handle both classification and regression problems; however, some models are better suited for one or the other. For example, one can use a linear regression model effectively to predict a continuous variable (a regression problem), but it is not as effective when predicting a binary response variable (a classification problem).

II. Unsupervised learning model category

In an unsupervised learning problem, the modeller does not attempt to predict a certain outcome, but rather seeks to uncover latent structure or attributes within the data. For example, a modeller may analyze a company's customer base to detect its major customer segments.

Within unsupervised learning there are two primary subsets of model:

i. **Clustering models**

In a clustering problem, the objective is to group the data into similar categories or clusters. Since this is unsupervised, clustering algorithms will attempt to find patterns in the underlying data that provide more information for the modeller. Examples include k-means clustering and density-based spatial clustering of applications with noise (DBSCAN).

ii. **Dimensionality reduction models**

In a dimensionality reduction problem, the objective is to condense the number of variables that are being considered. Again, since this is unsupervised, these algorithms attempt to find the lowest number of variables that provide the highest amount of information. Examples include principal component analysis (PCA) and linear discriminant analysis (LDA).

III. Semi-supervised learning

In a semi-supervised learning problem, the modeller is likely faced with a data set where only a partial amount of the response variable is known. One option in this scenario is to use the unsupervised portion of the data to enhance a supervised model. For example, this might be a relevant technique if you are attempting to predict the effectiveness of a continuing sales strategy that has been in place for five years but only started gathering data on the sales results beginning last year.

What are the privacy considerations?[2]

Some applications of predictive models require data sets with granular information about individuals. Examples of personal data collected can include personally identifiable information (PII), social media data, browsing data, consumer purchasing habits, and location tracking.

One should assess the privacy considerations for each situation in the context of the objectives of developing the predictive model and how they are used. Using genetic or personal medical information in a de-identified data set for research in developing socially valuable decisions for the public is very different from prediction of customer pregnancy for advertising.

In any application of a predictive model which uses potentially sensitive information, one should give careful thought to many aspects of privacy including the knowledge and consent of the consumer, transparency in the use of data, ethical use of analytics, and accountability.

In Canada, private sector companies developing predictive models using personal information must ensure their practices comply with the principles contained in the Personal Information Protection and Electronic Documents Act (PIPEDA).

What about disruption?

Predictive modelling is at the heart of major business and technological disruptions.

Disruptive businesses or ideas seek to exchange data without human intervention to feed a predictive model that will trigger a business decision.

Telematics integrates collection, transmission, and storage of data of a remote object, such as a vehicle. Auto insurance uses predictive models to relate driving habits to claim risks. The shipping business uses them to manage labour and fuel costs of a fleet.

Wearables are smart electronic devices that can be worn. The insurance industry is looking into this technology to relate life habits and behaviours with claim information. You may not realize it, but your smart phone provides similar information to your service provider and the companies whose applications you use.

Stock markets and trading are also impacted by predictive modelling, via algorithmic trading. For example, a model will “read” tweets to find information about a stock and trade accordingly in a matter of seconds.

Predictive models can use data gathered from power plant turbines to anticipate when they may require repairs. This allows for potentially less-costly preventive maintenance rather than the costs associated with emergency repairs. Models are also used in other areas of the energy sector to determine wind turbine output. Perhaps you receive a monthly update on your electricity usage: that information, and suggestions for improving your energy consumption, come from predictive modelling.

And the competition?

Marketplace competition plays a key role in encouraging organizations to innovate and find ways to continue succeeding and provide good return on investments to shareholders. Predictive modelling can help achieve this in a variety of ways across different functions.

Companies that do this well can achieve a competitive advantage, often through cost reduction or increases in sales and profitability. Organizations that employ analytics for risk selection can better understand and consequently manage risk and cost to their risk appetite. This insight also allows for improved pricing of products for consumers, matching the risk to a commensurate price. Companies with the best prices matched to the risk will eventually gain more market share and improve profitability.

Customers today demand a better consumer experience and more customized offers and engagement. Predictive modelling provides the opportunity to optimize each interaction with a consumer and predict the best individualized course of action for them.

If your competition makes better decisions as informed by predictive modelling you need to do the same to remain competitive.

Why do management and the Board care so much about analytics?

Given the cost reduction, increase in sales and profitability opportunities, and because many companies have data they can leverage and make better use of there is no question that companies need to invest time and money in analytics capabilities.

Properly applied, analytics can enhance decision-making, lower risks, and uncover insights that an organization can use to its competitive advantage. Analytics can provide the factual basis needed for better decision-making, and with time this will help businesses operate more efficiently and effectively. Analytics can minimize human error in situations where a predictive model can make a more consistent, accurate and objective decision. This can lead to cost saving or better profitability.

What do I need to use predictive modelling?

Predictive modelling is at the intersection of mathematics, statistics, and computer science. It also requires domain expertise of the business problem at hand: actuarial science, economics, engineering, medical science, and so on. Critical thinking is also necessary to exercise judgment on relationships the computer finds and to find intuition in the model.

Predictive models vary in complexity, so effective communication skills are needed for stakeholders to understand and trust the model. It is important to be able to describe in layman's terms and to be able to provide details as needed based on the demands and background of the audience. Good communication skills ensure that the model limitations are understood and that the model is not used inappropriately.

Data is the raw material of predictive models. A data governance process with a solid data infrastructure is critical to collect, store, secure, validate, access, review, and reuse data. Organizations should make efforts to store all the data they collect, even if they do not have an immediate need for it.

Predictive modelling requires specialized software, storage capacity, and computing power. Fortunately, recent advances in computer technology allow for flexibility in choosing solutions. A large organization with a big budget might decide to store all its data internally and purchase supercomputers to perform the calculations. A smaller organization might decide to use open-source programming platforms (R, Python), cloud storage, and cloud computing as a cost-effective solution with comparable results.

There are few people who have all the skills required. Consequently, an organization might hire a mix of data scientists, people with knowledge of the business, and information technology experts to implement predictive analytics solutions. Actuaries are key contributors in such a team.

Sources:

[1] https://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/

[2] https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2012/pa_201208/#toc_e5

https://en.wikipedia.org/wiki/Information_privacy_law

https://www.digitaltrends.com/social-media/why-consumers-are-increasingly-willing-to-trade-data-for-personalization/

This booklet was created by the Predictive Modelling Committee of the Canadian Institute of Actuaries.

Members:

Jean-Yves Rioux, FCIA, FSA, CERA, Chair

Jeffrey Baer, FCIA, FCAS

Patrick Duplessis, ACIA, FSA, CERA

June Quah, FCIA, FSA

Suzanna (Ying) Zhan, FCIA, FSA

The Canadian Institute of Actuaries (CIA) is the national, bilingual organization and voice of the actuarial profession in Canada. Its members are dedicated to providing actuarial services and advice of the highest quality. The Institute holds the duty of the profession to the public above the needs of the profession and its members.

Actuaries are risk management experts. They use mathematics, statistics, and probability to help ensure the financial security of Canadians. Traditional actuarial practice areas include insurance (both life and property/casualty), investments, pensions, actuarial evidence, and enterprise risk management.

Opinions expressed are those of the authors.