A13

Vai al contenuto multimediale

Marco Aleandri

Data Science and Machine Learning in Insurance

A gentle introduction for actuaries

Preface by
Fabio Grasso, Susanna Levantesi

Afterword by
Giampaolo Crenca

Aracne editrice

www.aracneeditrice.it
info@aracneeditrice.it

www.gioacchinoonoratieditore.it
info@gioacchinoonoratieditore.it

via Vittorio Veneto, 20
00020 Canterano (RM)
(06) 45551463

ISBN 978-88-255-2865-7

I^st^ edition: December 2019

Ai miei genitori

Contents

Preface

Fabio Grasso[*], Susanna Levantesi[**]

With the explosion of big data, data science, analytic solutions and machine learning have become a very hot topic in areas such as banking, finance and insurance as well as in the scientific community, where it is highly debated. Indeed, those topics have already proven their potential in a wide range of business fields such as marketing and operational research. Among financial institutions, the perception that data science could create value by increasing efficiency and profitability is rising day after day. This is leading to a growing demand for data–related skills in the job market, requiring a multi–disciplinary approach across quantitative methods and business knowledge.

Many universities have included data science and analytics in master's degree courses and second level master's programs, or created dedicated undergraduate and postgraduate degree courses to teach how to collect, manage and analyze big data. In Italy, for example, Sapienza – Università di Roma has just introduced the MSc in Data Science, the first MSc program in data science established in the country, besides some advanced programs in "Deep Learning with Python for Big Data Analysis" and "Big data: Statistical Methods for the Knowledge Society". Not surprisingly, data science and machine learning are topics covered by the Ph.D program in Statistics as well, namely "School of Statistical Sciences", where the author developed his studies from 2015 to 2019.

An increasing number of students and researchers are putting considerable effort toward enhancements of machine learning techniques and relevant business applications of data science. This book, arising

[*] Prof. Fabio Grasso, president of the Master's degree in Actuarial and financial sciences at the Dipartimento di Scienze statistiche of Sapienza – Università di Roma.

[**] Prof. Susanna Levantesi, coordinator of the curriculum in Actuarial Sciences at the Scuola di Dottorato in Scienze statistiche of the Dipartimento di Scienze statistiche at the Sapienza – Università di Roma.

from the doctoral thesis of the author, represents an outcome of this academic effort. It contributes to develop data–driven decision making in insurance as well as lay the foundation of data science in actuarial education. Beyond the traditional techniques of actuarial practice, these pages aim to familiarize actuaries with data science, its terminology and underlying concepts. A first step toward a new way of tackling typical actuarial questions.

Premise

The main goal of this book is to represent a systematic introduction to data science for actuaries, leveraging some of those topics that could benefit from it. The work is structured as a data science handbook, but the applications are purely actuarial. Obviously, this manual does not account for each and every aspect of data science or actuarial science, but it may be easily enhanced to include new algorithms and examples in the future.

Similar works have been published in the last decade, trying to bridge between daily actuarial practice and more advanced techniques. One of the very first examples is represented by Parodi (2009), where a large number of risk evaluation techniques are introduced, and applied to actuarial topics. To some extent, its structure is similar to that of this book. On the one hand, it is broader and goes beyond data science itself, introducing a large range of computational methods. On the other hand, however, it is limited to general insurance, which lends itself to advanced statistics in a more natural way.

After the aforementioned work, an even higher level of detail is reached in Frees et al. (2014) and Frees et al. (2016), respectively dealing with relevant quantitative techniques in the actuarial field and a variety of applications from several papers. That is the result of a huge effort from many researchers, providing a comprehensive view of quantitative actuarial science in theory and practice. Frees et al. (2014) introduces regression models and other parametric methods, which are then used in Frees et al. (2016). Even if it is not strictly about data science and machine learning algorithms, it has surely inspired this book.

Wüthrich et al. (2018) represents a more recent attempt to connect data science and actuarial practice. It introduces all the main machine learning techniques with a remarkable level of detail, focusing on a variety of motor insurance pricing applications. Once again, the presentation of the methods is extremely comprehensive, but the applications are still limited to a very specific field, that is, pricing in motor insurance. Of course, that represents a typical actuarial topic where analytics can be successfully applied in its several forms, but it is not the only one.

In the last years, researchers were not the only ones to produce these type of works. Indeed, actuarial associations began to promote data science topics in their working groups, reporting on the opportunities offered by machine learning and big data to the actuarial world.

One of the first contributions was provided by the Belgian actuarial association publishing IABE Information Paper (2015). Although it is quite focused on the business perspective of big data rather than the quantitative aspects of data science, it brings together a lot of interesting ideas on the future of the actuarial profession in the era of data explosion. It covers all the sectors involving actuarial activities, from life to non–life, from pricing to reserving. Emphasizing business–related problems such as customer management, claim reserving and policyholder behaviour, it has somehow suggested the three applications we will present. At the same time, in spite of its lack of quantitative analysis, it has inspired the conclusion to this book in the last chapter.

Another, more recent contribution is represented by IFoA (2018) published by the Modelling, Analytics and Insights in Data working party of the Institute and Faculty of Actuaries. Even if it is much briefer than this work, it shares some of its goals, which include, among others, introducing data science to actuaries and applying machine learning to actuarial case studies. While the former is reached by handling most of the concepts handled in this work as well, the latter is reached through different applications. In particular, IFoA (2018) uses supervised learning for interest rate prediction, marine hull pricing, catastrophe exposure management, and suicide rate estimations.

In IFoA (2018), the business perspective is still present and sometimes predominant, but the discussion encompasses many fundamental data science ideas. Part of the concepts expressed in this book are introduced there as well, although from a high–level perspective and without the necessary details to understand what is really behind each algorithm. Even if the applications do not represent the heart of that work, they touch various actuarial sectors just like we will do here, aiming to demonstrate the potential of alternative techniques against traditional methods. Those aspects make IFoA (2018) different to all the aforementioned works, raising interest among actuaries from any background and sector.

In this work, we aim to pick the best features from the cited works, in order to build a comprehensive, detailed and actuary–oriented introduction to data science. It is essentially structured to answer the following questions:

— *what?* That is: data, dataset partitions, performance measures, software, supervised learning, unsupervised learning, algorithms, etc. Some of those topics will be first presented in Chapter 1, especially concepts representing the foundations of data science, which will be then useful throughout the book. By contrast, the peculiarities of algorithms will be covered in detail as soon as we will need them for the actuarial applications of the central chapters, in order to better connect machine learning technicalities and specific case studies. Considerable effort will be devoted to maintain the typical structure of data science manuals in terms of sequence of the topics;

— *how?* That is: actuarial applications for different sectors of the actuarial practice. More specifically, we will tackle three case studies: customer management in Chapter 2, individual reserving in Chapter 3 and policyholder behaviour in Chapter 4. They will be respectively handled by using unsupervised learning, supervised learning and ensembles (i.e., combinations of different methods). Data availability is not the only reason why we choose those topics. First, we want to tackle things that are potentially part of the daily actuarial practice such as renewal rates, claim reserves or lapse probabilities. Second, we want to tackle things involving different actuarial areas such as underwriting, non–life business and life business. These two targets aim to raise the interest of actuaries regardless of their specific background and role in the industry;

— *why?* That is: accuracy or inaccuracy, stability or instability, interpretability or black–box effect, etc. Highlighting relevant reasons to prefer data science over traditional approaches is crucial. As a first step, Chapter 3 will demonstrate the importance of unsupervised learning in data manipulation as well as its potential in detecting clusters and improving accuracy. However, this will come at the cost of increase in model complexity. Instead, Chapter 2 will show that more flexible machine learning techniques such as decision trees may outperform regression models. Even if this is not a statement that holds in general, it will provide actuaries with suitable alternative methods to boost model performance. Actually, Chapter 4 will start illustrating that those alternatives may miserably fail because of instability due to data flaws, algorithmic issues or other problems. That will justify the usage of ensembles: in particular, bagging trees will imply more stability and outperformance over logistic regression.

This is just a first step to make data science methods relevant to the whole actuarial world, in both academia and industry.

Introduction

In the light of the increasing importance of data in insurance and, as a consequence, in the actuarial practice, this dissertation has a two–fold objective. First of all, it should provide actuaries with an introduction to data science, including basic concepts, terminology, data mining issues, performance measures, machine learning techniques and software. That is necessary as those topics are definitely outside the typical toolkit of actuaries. Secondly, it should suggest significant applications of data science to actuarial problems (e.g., pricing, reserving, ratemaking, etc.). We will look at a number of case studies, investigating the extent to which machine learning can enhance or outperform more traditional approaches.

The dissertation is indeed structured to reach these two goals. While Chapter 1 consists of an introduction to the main concepts of data science, Chapters 2, 3 and 4 describe three different applications in actuarial practice tackled with machine learning techniques. The details on those techniques are outlined just before the applications themselves, in order to keep data science and actuarial practice parallel all the way through the dissertation.

To pass on the message that data science can be relevant to any actuary regardless of his/her specific field, the applications involve very different topics. Chapter 2 focuses on marketing and customer behaviour in motor insurance to highlight the importance of data preparation and unsupervised learning. Chapter 3 describes the most common supervised learning techniques as an alternative to regression models in a traditional non–life topic like claim reserving. Finally, Chapter 4 illustrates an example of data science application in life practice, that is, predicting lapse rates to improve asset–liability–management models.

To conclude, Chapter 5 provides a brief overview about the next steps in insurance business, actuarial education, actuarial profession and, more generally, the role of actuaries in the future.

Chapter I

Data Science Basics

As suggested by the title, this first chapter will represent a brief introduction to the main concepts used in data science. They encompass general and soft notions (e.g., classification and prediction, supervised learning and unsupervised learning, etc.) as well as more precise and technical definitions (e.g., bias and variance, performance measures, etc.).

The first part will provide the reader with an overview of the current role of data science in business, describe the data mining process through its main standards, and highlight the importance of data availability. Subsequently, the second part will introduce some typical performance evaluation tools and other criteria for model selection, define generalized linear models that are relevant to actuarial practice, and list the most common machine learning techniques.

Most of those concepts will be used throughout the book on several occasions, although they are not entirely part of the traditional actuarial background. Therefore, the next pages aim to build the necessary foundation for the following chapters.

1.1. Some terminology

When it comes with statistical methods, there are quite a lot of terms that are spreading among actuaries nowadays. Nonetheless, actuaries are not really statistical experts since their education is focused on statistics to the extent they are effectively applicable to insurance. This first section aims to introduce some basic concepts that will turn out to be useful throughout the book. However, it is not meant to be fully comprehensive, and it can easily get out of date as new concepts come out. For our aims, however, it should be enough.

The very first concept to introduce is that related to *data science*. It denotes the scientific field about the entire range of systems, processes and methods used for *data mining*, that is,

> the process of exploration and analysis, by automatic or semi–automatic means, of large quantities of data in order to discover meaningful patterns and rules

as defined in Berry et al. (1997). Therefore, data science is much more than statistics. Among others, it covers data integration, data architecture, data visualization, data engineering, data–driven business analysis and of course the whole range of tools to mine data.

Some of these tools come from classical statistics, for instance, sampling methods, confidence intervals, hypothesis tests and regression, just to mention the major ones. To some extent, all of them are based on analytical assumptions and mathematical formalization. It guarantees a strong, theoretical foundation to these tools, so that they may be used in any relevant application.

Other tools lack such a theoretical strength, but gain much more flexibility to recognize pattern in data. More specifically, they are structured to automatically adapt their input parameters in order to catch more and more information from a given dataset. This is the reason why it is often said that such tools "learn" from data. The statistical field that encompasses all of them is called *machine learning*, the object of this work.

In machine learning, two types of "learning" are usually mentioned:

— *supervised learning*, that is, learning about the relations between a range of *predictors* (so–called *explanatory variables* in regression) and a determined *target* (so–called *response variable* in regression);

— *unsupervised learning*, that is, learning about the relations between a range of variables in order to group "similar" records.

In both the cases, the algorithm catches information, and uses it to interpret new data. In supervised learning, it will use new data predictors to predict the related target variable. In unsupervised learning, it will use new data variables to assign new records in previously identified clusters. Given a dataset, one may use supervised tools or unsupervised tools depending on the goal — prediction for the former, segmentation for the latter — but there is no difference in the data. Just remember that, in supervised learning, we distinguish variables between a range of predictors and one single target, while there is not such a distinction in unsupervised learning.

Nonetheless, a further distinction is relevant in supervised learning, depending on the nature of the target variable. A specific tool is used for:

— *classification* if the target variable is categorical;
— *prediction* if the target variable is numerical;
— *forecasting* if the target variable is a time series.

The great majority of machine learning tools can be adapted for both classification and prediction, while the tools for forecasting time series are usually considered in isolation. In our analyses, we will only consider tools for classification and prediction such as decision trees and neural networks. The map in Figure 1.1 shows the most common ones (sometimes, there is a lack of uniformity in naming different tools among researchers, so one could find various names for essentially the same tool). However, on a daily basis, brand new algorithms or improvements to old tools are created by data scientists to tackle specific problems, especially in the last years. Rather than explaining all of them, we are going to focus on the most widely used.

Quite interesting is the relationship between machine learning and *artificial intelligence*. The two concepts share something similar and tend to be confused with each other, but are quite different in reality. Artificial intelligence was born in the 1960s as a subfield of computer science with the main goal of programming computers to perform human tasks (e.g., speaking, listening, writing, translating and many more). Actually, some of them are so complex that artificial intelligence needs specific machine learning tools. However, the same tools could just be as successful in any other field, say, predicting stock prices.

At the same time, one might use non–learning tools in artificial intelligence, if they are sufficient to replicate some human behaviours (for example, imagine a very complex algorithm with some fixed pa-

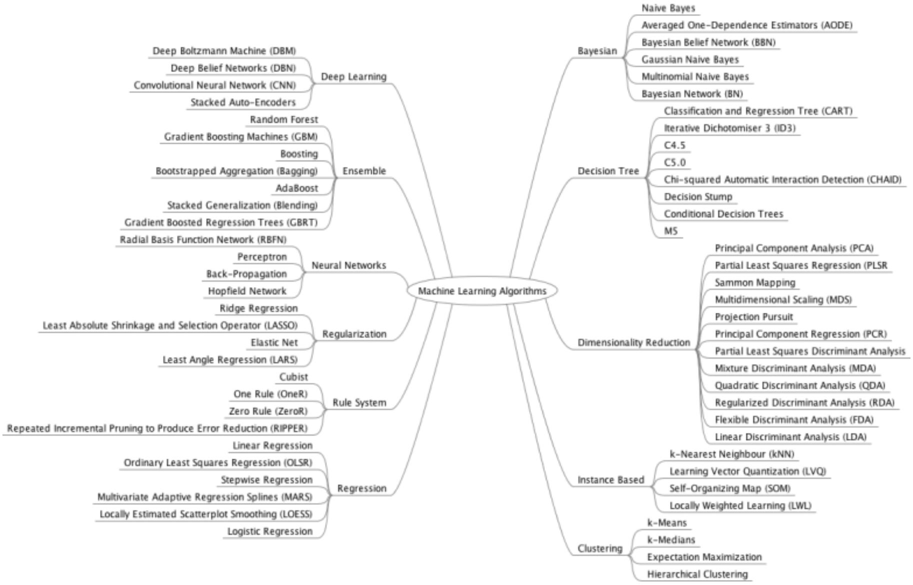
Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Figure 1.1. A map of machine learning techniques (see Brownlee (2013))

rameters: it will run always the same way, without learning anything from new data). In reality, human beings are so complex that machine learning tools are considered a "must" in artificial intelligence. Contemporarily, artificial intelligence is seen as a main goal for machine learning. This is the reason why they get often confused.

Nonetheless, artificial intelligence is not the only field where machine learning techniques have been proven to be useful. A less "noble", but more practical field is *business intelligence*. Generally, the concept of business intelligence comprises all the data processes, data technologies and data insights aimed to the improvement of specific business activities and performances. Forrester.com reports the following definition:

> Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision–making.

If we compare this definition to that of data mining, it will be clear that business intelligence is just data mining from a business perspective, in an industry environment. As such, it also focuses on the practical aspect of data mining, for instance, data visualization (e.g., reporting tools and executive dashboards), decision–making process and so on. While data mining and machine learning are rather standardized concepts, business intelligence features are quite customizable by industry. For example, a specific industry professionals usually prefer certain machine learning tools or software, because they have been already proven to be effective in that field. Here is a list of actual, specific and successful examples of business intelligence:

- e–commerce companies need machine learning to draw data insights about a number of daily issues, for instance, which products would be preferred by which customers, which customers would go for which promotions, which products would be purchased together or in a short time–frame and so on (this is an example of *market basket analysis*, introduced later in the book);
- social network developers use machine learning to analyse users emotions after status updates (this is an example of *sentiment analysis*);
- search engines such as Google and Yahoo cluster "similar" web pages by using unsupervised learning techniques;
- e–mail spam filtering always rely on some simple machine learning tools to detect spam;
- financial institutions use machine learning to forecast the stock market, and thus direct investment decisions;
- banks classify loan and mortgage applicants by different level of riskiness using the default probability predicted by some machine learning tools;
- healthcare industry relies on machine learning in several applications, for instance, detecting patients who will likely develop a chronic disease, and predicting possible adverse drug reactions in patients;
- transport companies use machine learning to forecast customer behaviour and thus real needing in transportation in order to reduce costs;
- automobile industry companies predict the failure or breakdown of mechanical parts by using machine learning;