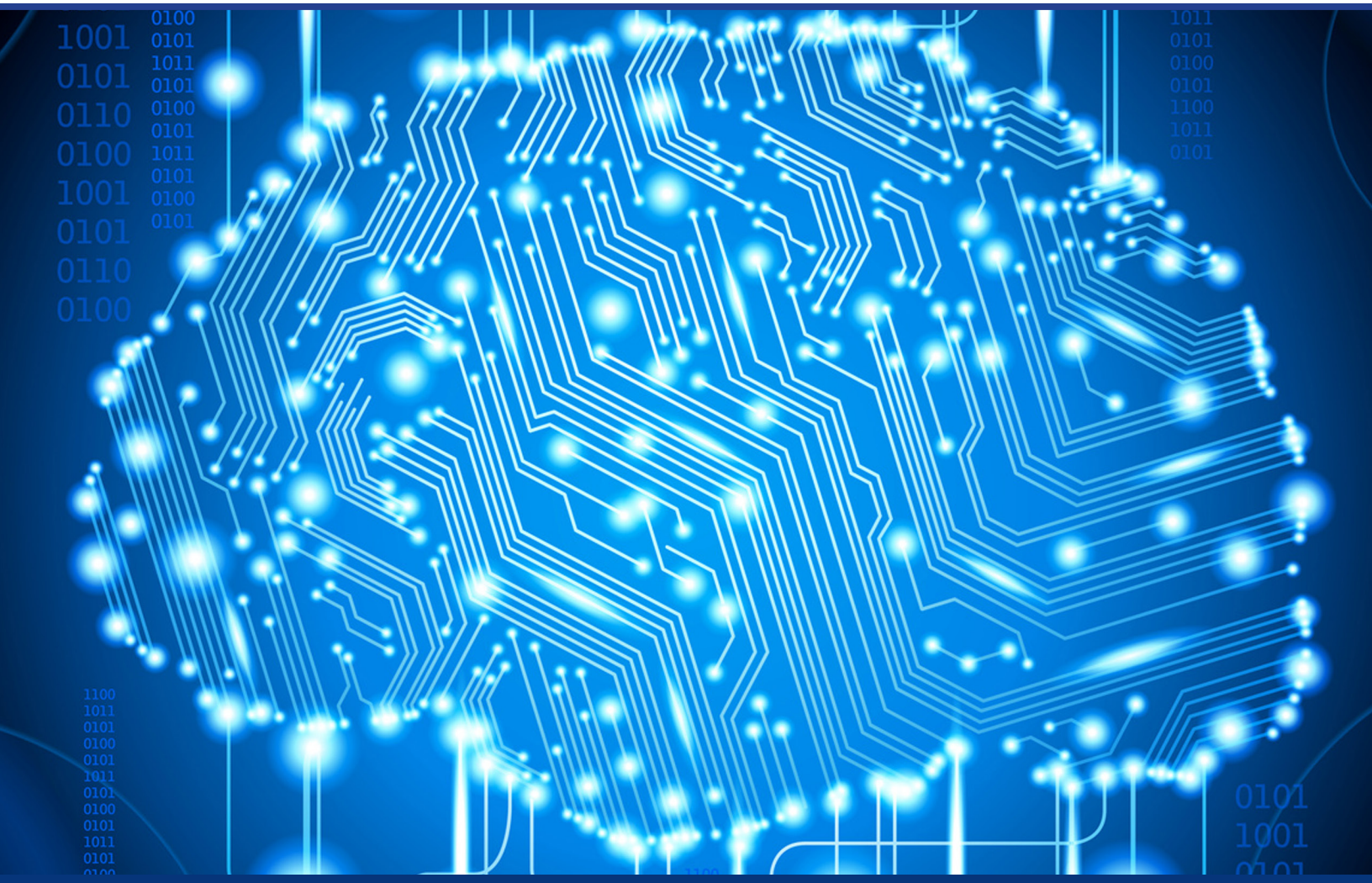
M SQUARED CONSULTING
A SolomonEdwards Company



A Simple Framework for Building Predictive Models

A Little Data Science Business Guide

Authored by:
Hal Kalechofsky, Ph.D.

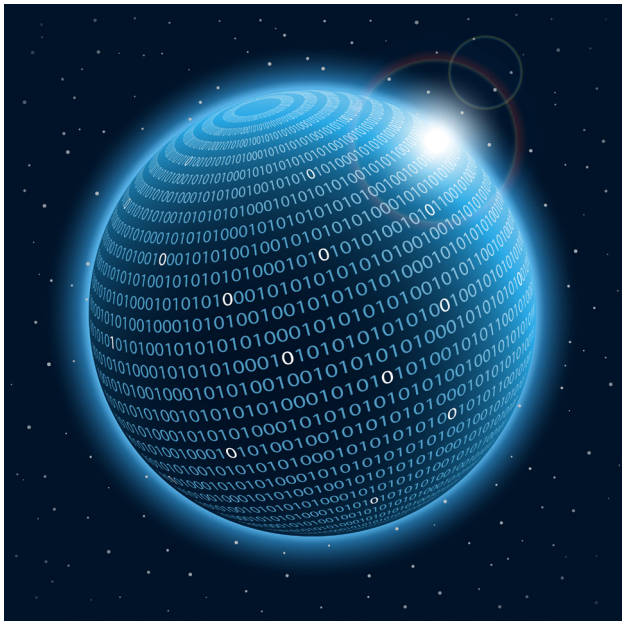
September 2016

Table of Contents

1. Introduction

"Statistics are like a bikini: What they reveal is interesting, but what they conceal is vital."

Most humans seem to want to predict the future. It is a natural human desire. To know in advance about the weather, the stock market, the next card coming up in blackjack – what a power to have over the natural world!



After decades of research and development, computer science and information technologies are now reaching a point where predictive models are an important, if not indispensable, part of the business ecosystem. Many technologies, including fast computing power, inexpensive storage, cloud computing, voice recognition, mobile devices, artificial intelligence (cognitive computing), and advanced application software are combining to make this possible. The world is generating vast amounts of data. To cite just one example, the amount of data that will be generated by the Internet of Things (IoT) in the years to come will dwarf what we today call "Big Data." Recent trends are allowing predictive analytics and models to be democratized and spread to smaller organizations and individual users, avoiding the need for large software budgets or armies of data scientists to create and analyze the insights generated.

Business leaders in all industries will want to not only be aware of the data science forces shaping our future economy, but also to be well versed in how best to use and to make the most of these coming opportunities.

As human beings, unlike most other animals, we have the power to shape our environment. One main way of doing this is to assess situations based on data and evidence and then plan and strategize for the future. One might go so far as to say this is an evolutionary necessity for survival, or at least for improved chances for survival. It is interesting how our tools (of which information technology is one powerful example) reflect what we need (or think we need) to survive throughout history. Just think back through the wheel, fire, steel, weapons, travel, medicine and information technology.

Viewed in this way, predictive computing models are an extension of our natural survival instinct. In this data-intensive world, predictive models are more important than ever in order to make sense out of what is around us and to estimate, assess, or plan for what might happen in the future.

1.1 Working Definition of Predictive Models

"It's hard to make predictions, especially about the future."

A short definition of a predictive model is:

- Using data to make decisions.

A longer definition might be:

- Using data to make decisions and to take actions using models that are empirically derived and statistically valid.

Predictive modelling is a commonly used statistical technique to predict future behavior. Predictive modelling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

Predictive models are created whenever data is used to train a predictive modelling technique. To put it another way: data + predictive modelling technique = predictive model.

A predictive model is the result of combining data and mathematics, where learning can be translated into the creation of a mapping function between a set of input data fields and a response or target variable.

1.2 Problem Statement

"I know that half my marketing spend works ... I just don't know which half."

Business metrics do a great job at summarizing the past. However, if you want to predict how customers will respond in the future, there is one place to turn — predictive analytics.

There are many business reasons why it is important to make use of predictive models. Any time you would like to know something about the future, it is useful to have a predictive model. Although an educated guess is fine, imagine being right more than half the time! Imagine being right most of the time or all of the time.

Here are just a few reasons to consider using predictive analytics and models in your business:

- Future Potential Revenue or Cost Drivers
  - How can I know my product is right for this market?
  - What is the best way to continuously lower my cost of goods sold?
- Risk Management
  - How can I identify fraud or increase trust in financial transactions?
  - Which business scenario is the likeliest to win out?
  - How can I build and maintain advantage over my competitors?
- Operations
  - How can I save money, resources and time by anticipating when and how physical assets are likely to break down, and preventing it?
- Customer Relationships
  - What will my customers want in the future?
  - How can I solve customer problems before they happen?

It is sometimes said that decisions made on wrong data are better than decisions made on no data at all. This is because even if the data are wrong, at least one can use human skepticism and questioning techniques to learn from it. Without data, this is not possible.

Business leaders often use intuition, a "gut feeling" or revenue streams to forecast future market conditions. Sometimes they are right, sometimes not. This could be called an example of an emotional or gut-level predictive model. Given all the data we have gathered in our high-tech world, it can be useful to supplement this instinct with predictable information so that one can evaluate the decisions to make.

1.3 Goal of this Publication

The goal of this paper is to present a simple framework for developing predictive or statistical models for modern business purposes.

Modern means the first few decades of the 21st century, with all of our high-tech measurements and computational apparatus brought to bear. What we were able to do with predictive modelling a generation or more ago and what we can do at the present time may share mathematical techniques, yet are different in their historical anchor points.

This is not intended to be an academic or university research type of paper; this paper is presenting a framework, not an exposition of deep data science techniques. In the modern world, we have educated many competent analysts and data scientists and have built many software systems to apply the necessary

predictive modelling techniques, so it is not needed to cover that here.

This paper is meant to be a primer, not a detailed essay. It does not instruct people on how to build models, but covers the steps involved and the practical issues to consider. This paper is designed to be relatively short and to deliver information efficiently in a short amount of the reader's time.

2. Audience

This paper is written for a variety of audiences. The reader should be familiar with modern computer systems and have some level of curiosity. Many business people these days are well versed in technology and technical reasoning. Trained statisticians and data scientists who are practitioners of predictive models may know the theory and the mathematics, but may get something out of the business discussions.

A list of role/resource types that may benefit from this paper are listed below:

- Project Leaders: who desire to have a more detailed understanding of predictive modelling methods and techniques to better manage and interact with their practitioners
- Business Analysts: who must develop and interpret the models, communicate the results and make actionable recommendations
- Big Data Analysts: who are under increasing pressure to transform their deluge of data from a liability to an asset
- Functional Analytic Users: Customer Relationship Managers, Risk Analysts, Business Forecasters, Statistical Analysts, Social Media and Web Data Analysts, Fraud Detection Analysts, Audit Selection Managers, Direct Marketing Analysts, Medical Diagnostic Analysts, Market Timers
- Data Scientists: who desire to extend the scientist aspect of the role with formal process and hands-on methodological practice
- IT Professionals: who wish to gain a better understanding of the data preparation, analytics and analytic sandbox development requirements to more fully support the growing demand for analytic IT support
- Anyone overwhelmed with data and starved for actionable insights

3. Business Usage of Predictive Models

"If you predict it, you own it."

This section describes some of the business usages of predictive analytics and predictive models.

Predictive analytics encompasses a variety of statistical techniques from predictive modelling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events. When most lay people discuss predictive analytics, they are usually discussing it in terms of predictive modelling. Indeed, predictive modelling is at the heart of predictive analytics.

Predictive analytics is used in marketing, ecommerce, financial services, actuarial science and insurance, telecommunications, internet security, retail, travel, medical and healthcare, child protection, pharmaceuticals, capacity planning, supply chain, and other fields.

3.1. Use Cases

The table below lists predictive analytics business applications. The first column names the type of business benefit, and the second column identifies the type of customer prediction required – that is, which behavior or action must be predicted to undertake each business application. As there are many such applications, this list includes only the most pervasive in commercial deployment to date.

Business Application	What is Predicted	Reference or Company Case Study
Customer retention	Customer defection/churn/ attrition	Reed Elsevier, and Telenor, Published article
eCommerce	How to sell successfully online	Amazon.com
Direct marketing	Customer response	Charles Schwab Published article
Product recommendations	Customer wants/likes	Netflix Prize leader, HSBC & Amazon.com
Behavior-based advertising	Which ad customer will click	Google, Yahoo! and Click Forensics, "$1 million" case study
Email targeting	Message that will generate customer response	World Wildlife Fund
Credit scoring	Debtor risk or fraud	Wells Fargo
Fundraising for nonprofits	Donation amount	NRA
Insurance pricing and selection	Applicant response, insured risk	Insurance Journal, Pinnacol Assurance

There are many more applications of predictive analytics, including collections, supply chain optimization, human resource decision support for recruitment and human capital retention, and market research survey analysis. The way predictive scores help your business depends on the customer behavior predicted – just aim predictive analytics towards the right customer prediction goal and fire away. This is why it is sometimes said, "If you predict it, you own it!"

4. Modelling Framework

"The best way to predict the future is to create it."

4.1 Recommended Predictive Modelling Business Requirements

While they may share many characteristics and techniques, building statistical or predictive models for a scientific or university research context can be different than a similar exercise in a business context. For business-oriented projects, here are some basic core modelling requirements and considerations.

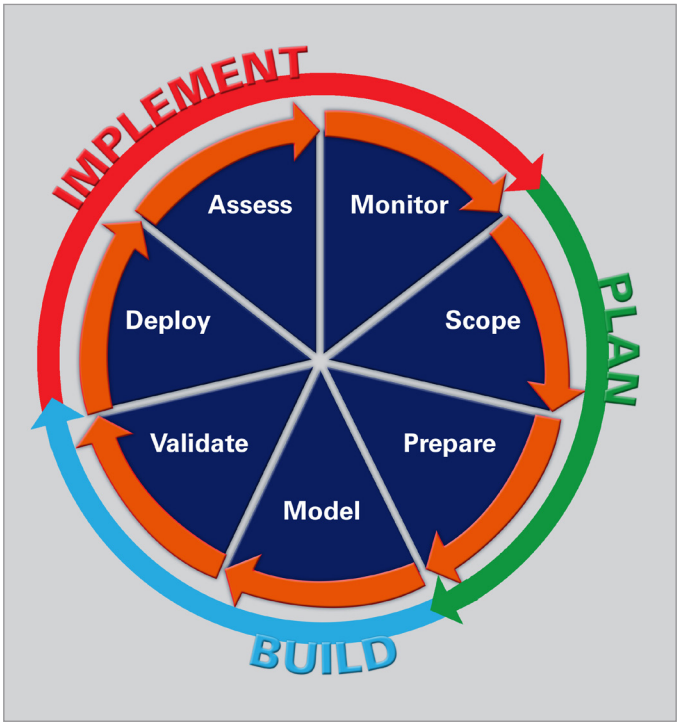
- Formulate a clear and transparent framework:
  - What are the objectives? What are we modelling for?
  - Predictive models alone do not create business value, but rather need to be effectively deployed either into a decision-making or business process.
- Make room for using business knowledge in the process:
  - Involve business users in the modelling process to ensure clear requirements, communication and purpose verification.
  - Business experience and heuristics may contain valuable nuggets for how to use the data.
  - Do solutions pass the "Does it make business sense" test?
- Make sure that data are appropriate for modelling:
  - Quality, scope and quantity suitable to meet objectives
  - Structured in line with underlying processes being modeled to maximize the signal
- Don't just make best statistical fits to historical data:
  - Test predictive power and consistency of patterns over time and across validation samples
- Think about managing the model:
  - How to integrate the model into existing business systems
  - After implementing, how to track and manage the interactions, and how to maintain, tweak, or update the model going forward in an efficient manner

4.2 The Modelling Process

"The devil is in the details."

Let's take a look at the process of predictive modelling. There is a wealth of information on the Web about this. However, to understand the strategic areas, it is useful to break down the process of prediction into just a few essential components.

The below diagram is a good depiction of the predictive modelling process across a wide range of businesses.



Broadly speaking, predictive analysis and modelling can be divided into three parts: Plan, Build, and Implement.

4.2.1 Plan the Model

The next step includes scoping and planning. In the Planning section, the sub-phase activities here are called Scope and Prepare. This part of the exercise might take around 40% of the total time.

To build a predictive model, you first need to assemble the datasets that will be used for training. You must formulate clear objectives, cleanse and organize the data, perform data treatment including missing values and outlier fixing, make a descriptive analysis of the data with statistical distributions, and create data sets used for the model-building.

Scope

A clear specific objective for the model is required. Each model is developed for its own specific purpose and cannot be used effectively in another situation. For example a model that predicts ecommerce online customer churn cannot be used to predict credit card churn. An example of a clearly defined model objective contains the event or action that the model is to predict and the period when it is likely to happen. For example, the objective could be to predict customers that are likely to miss a credit card payment within the next month.

Prepare the Dataset

The dataset created to use for the model might broadly cover diverse information, such as product, case, behavioral, demographic, geographic, competitor details or weather. The variables that are not included in the dataset will not form part of the prediction. Variables cover both static fields such as income and triggers such as change in spend. Both technical and business people need to be involved in the decisions relating to the contents of the dataset. Focus needs to be on the behavioral information as this is more powerful for predictions than demographic data. It is useful to give consideration to which customers to exclude from the model build process.

Customers need to be excluded if they are going to impair the performance of the model. Potential exclusions include things like bad debt, staff and new customers as they have insufficient history. The actual exclusions applied relates to the specific purpose of the model. For example, if the model is to predict customers that are likely to carry bad debts, bad debt customers would be included in the model. In practice, exclusions are applied as filters when creating the dataset and should be noted down.

4.2.2 Build the Model

Here you will write model code, build the model, calculate scores, and validate the data. In the Build section, the Sub-phases are known as Model and Validate. This might take around 20% of the overall time. This is the part that can be left with the data scientists or technical analysts. The technical aspects of building models are not covered in this paper.

The model will be built using a sample from the data set created. The resulting model will contain a subset of the original list of variables considered for the model. This is acceptable and happens because some of the variables considered for the model will be correlated with each other, for example, a product type and product family, or the floor number and height of building. Other variables will have been discarded as they add little or nothing to the model's predictive power.

Calculate a Score

The model developed will be an equation that, when applied to the customer base, will allocate a score to each customer. The score represents a customer's likelihood 'to do' whatever the model is predicting, such as predict a customer's expected order value or likelihood to churn within a month. If you are building a model for product failures, you may have a score that calculates the likelihood of a product to fail in a given time frame.

Validate the Model

Typically models are validated against a hold out group. This group contains customers that have not been included in the development of the model. As such, they represent a group of previously unseen

customers that are representative of the customer base. To achieve an accurate prediction of lift, the hold out group must not be made up of the customers that have been excluded from the model development process. The same idea applies if you are building a product failure score model.

4.2.3 Implement the Model

Here you will deploy and apply the model, rank customers or products, use the model outcomes for some business purpose, estimate model performance, assess and monitor the model, and drive initiatives based on the model. The Sub-phases here are known as Deploy, Assess, and Monitor.

You need to think about accessing, storing, and using the data. This might take around 40% of the overall time, require IT department work, and can be an ongoing operational exercise when the model is used continuously for the business.

Deploy the Model

It's time to apply the model. The model will be built on a subset of data. Once the model is complete and has been validated, it will be run over the customer, product, or case base.

Assess

It is typical to generate rankings from your model – depending what you are modelling – customers or products. You will want to understand the performance of your model. The ranking scores allow a customer or product base to be ranked in order of the predicted score, such as from most likely to least likely to churn or from most popular product to least. In reality, some customers will churn and some customers will not. Therefore, in absolute terms, these predictions will not be accurate. The ranked list provides a superb base upon which to vary the treatment, and therefore the level of service or marketing spend, to groups of customers.

Monitor

You want to consider how often to run the model across the base and generate scores. You will want to monitor the performance of the model on an ongoing basis, and perhaps take advantage of new data or techniques as they become available.

4.3 Modelling Longevity and Considerations

"The simplest solution is the best."

It is useful to think of the deployment of predictive models in a continuous improvement scenario. You don't just build the model and leave it alone. The model may take nourishment, watering, monitoring, and care and can improve over time as conditions change and new data come in.

A model does not have to be extremely complicated in order to make good predictions. Like in coding, simplicity is a good guideline. It is useful to keep Occam's Razor in mind when considering how to build a model: Some people say Occam's Razor is "the simplest solution is the best." However, if you look at the original text, it says "Do not multiply entities beyond necessity."

Generally, the analyst may make assumptions when building a model. However, nature does not assume anything before forcing an event to occur. The fewer assumptions there are in a predictive model, the greater

will be the predictive power. Clearly, attributes that are not in the model will have no effect on the model's predictions.

The speed of changing business conditions is a factor in how well the model will perform over time. For example, information technology is a faster-moving industry than insurance. Once a model is in use and driving actions, there is a requirement for tracking and managing the interactions. It is a good idea to refresh model scores over time, sometimes frequently depending on the industry.

Control files will be required to test initiatives and to test the model performance. The results from this will help determine how frequently the model needs to be refreshed. As a rule of thumb, a model needs to be reviewed and possibly rebuilt annually.

4.3.1 Types of Predictive Models

"There are three types of lies -- lies, damn lies, and statistics."

Many types of predictive models have been developed over the years that are useful for different classes of problems. Applying these techniques is the domain of statisticians and data scientists, and is intentionally not covered in this short paper.

1. Business Rules

A business rule is a rule that defines or constrains some aspect of business and always resolves to either true or false. Business rules are intended to assert business structure or to control or influence the behavior of the business.

2. Classification and Decision Trees

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.

3. Naive Bayes

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The technique constructs classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

4. Linear Regression

In statistics, regression analysis is a statistical process for estimating the relationships among variables. Linear regression is an approach for modelling the relationship between a scalar dependent variable Y and one or more explanatory variables (or independent variables), denoted X. The case of one explanatory variable is called simple linear regression. More than one variable is called multivariate.

5. Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical or binary.

6. Neural Networks (NNs)

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information.

7. Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

8. Support Vector Machines (SVMs)

In Machine Learning, an SVM is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. An SVM maps input data vectors into a higher dimensional space, where an "optimal hyperplane" that separates the data is constructed. An SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

9. Natural Language Processing (NLP)

Since NLP (content analytics) feeds predictive and prescriptive analytics, it is included here. NLP fact

extraction can be used with descriptive statistical formulae to make use of the wealth of unstructured data.

4.4 Supervised and Unsupervised Learning

"Statistics can be made to prove anything - even the truth."

SVMs, decision trees, NNs and regression models use supervised learning to create the mapping function between a set of input data fields and a target variable. The known outcome is then used as a teacher who supervises the learning of her pupil. Whenever the pupil makes a mistake, the teacher provides her with the right answer in the hopes that the pupil will eventually get it right. For instance, when presented with a specific set of inputs, her output will match the target.

Unsupervised learning requires no teacher or target. Clustering techniques fall into this category. Data points are simply grouped together based on their similarity. In an ecommerce website analysis, online shoppers might be grouped into window shoppers or power buyers. In case of customer churn, a clustering technique could potentially assign different clusters to churners and non-churners even though the outcome is not available during model training.

Black-box is a term used to identify certain predictive modelling techniques that are not capable of explaining their reasoning because you can't really know what is happening in the Black-box. Although extremely powerful, techniques such as NNs and SVMs fall into this category. Very often, an explanation or reason for the model decision is required; for example, when a risk score is used to decline a loan application or a credit card transaction.

Whenever explaining is a must, you need to consider using a predictive modelling technique that clearly describes the reasons for its decisions. Scorecards fit such criteria very well. Based on regression models, scorecards are a popular technique used by financial institutions to assess risk. With scorecards, all data fields in an input record are associated with specific reason codes. Thus, it is possible to explain to the customer why a given decision was rendered.

4.5 Practical Considerations When Working with the Data

"Correlation does not imply causality."

1) Asking the Right Questions

What are the business objectives of the project? Does it make business sense?

Awash in reams of data, it is critical that companies ask the right questions. Being specific is important. Predictive analytics is most efficient when used to determine the answer to a narrow inquiry, such as the likelihood of customer A to buy product X at time Y for price Z, rather than the likelihood of customers buying product X.

2) Identify Available Data

It is one thing to have someone say, "Build me a predictive model to mint money." Is that realistic? If the goal looks realistic, one must find out if there are even data available to do something reasonable. Sometimes the data will not be accessible, or it will not be ordered or structured well enough.

3) Data Accuracy and Cleanup

Data scientists must be aware that not all data is accurate, arrive at an estimate of bad data, and correct for it in their studies. Data can be bad for any number of reasons, including self-reporting errors, corrupted files, poorly phrased questions, incomplete data aggregation, and poor standardization methods.

It is critical that data scientists quickly recognize and filter bad data from their data sets. They must also make sure they do not create bad data themselves, i.e., through an imperfectly calculated transformation function. Since bad data can play through the model-building process, you have to have a feel for what will cause problems and what you can accept. In general, it is useful to have procedures in place to both filter out bad data and to prohibit data pollution in the future.

4) What Assumptions are Being Made?

Generally speaking, the more assumptions one makes, the less predictive the model may be. For example, if you assume that every customer is from an urban area,

what will that tell you about people who live in the countryside?

Big Data is messy, consisting of everything from social media mentions to traffic camera images to website logs. Predictive analytics, being a set of statistical techniques, requires all data to be standardized and quantified. Quantifying non-numeric data has its own risks and creates uncertainty.

Further, data is unpredictable, especially dynamic data. A model that accurately forecasts future events could be thrown into disarray by a sudden unanticipated cascade of events, which were not initially estimated. In 2007, the failure the majority of financial services firms to incorporate the possibility of sudden credit defaults triggered a series of other events that prior to 2007 would have been improbable.

5) Privacy and Security

It is always crucial to respect customer data privacy and data security. Predictive models bear the added weight of collecting and using information about individuals in order to predict future behavior. Some privacy advocates find such data usage invasive and alarming. Many people feel there is intrusive about firms collecting information about individuals in order to predict their behavior. Advocacy efforts include lobbying for limitations to data collection types, amounts and methods in nations across the globe. Executives and data managers must be aware of the ever-changing Big Data regulatory landscape.

Privacy is a huge concern for another reason – security. Hackers may target companies for financial gain. If hackers can glean customer or product information, this can lead to data compliance issues, customer dis-satisfaction, or unfavorable press.

Knowing your company's data privacy and data security policies in advance is a big help when doing predictive analysis of any type.

6) Pre-model Exploration – Statistical Distributions

This typically involves data mining: univariate, bivariate, sets, graphs, and so forth. There are generally four basic questions that can help in the characterization of the data and constructing data distributions:

- Are the data discrete values or whether the data is continuous?
- Are the data symmetric in any way?
- Are there upper or lower limits on the data?
- What is the likelihood of observing extreme values in the distribution?

7) Determine Predictors

It is common sense that one should identify as many of the main predictive variables as possible. It is important to distinguish dependent and independent variables, as well as correlations between variables that could skew model results. Additionally, it is good to find if there is an identifiable signal to noise ratio in the data. For example, if you are trying to predict incomes in a demographic population, geolocation and job type may be important predictors.

8) What Model to Select and Why

There are many statistical, data mining, and machine-learning algorithms available for use in your predictive analysis model, often available in enterprise analytics or open-source software. Once you have defined the objectives of your model and selected the data you'll work on, you should be in a good position to choose which algorithms might apply best.

Here are a few general rules of thumb to decide which algorithms can address various business concerns.

- For customer segmentation and/or community detection in the social sphere, you will need clustering algorithms.
- For customer retention or to develop a recommender system, you'd use classification algorithms.
- You can use decision trees when you have a linear decision boundary, for example, classifying people on the basis of their IQ.
- For credit scoring or predicting the next outcome of time-driven events, you'd use a regression algorithm. Use regression when you want to predict continuous values, instead of classifying. Regression can be used for traffic prediction, for instance.
- You can use the Naive Bayes classifier when the features are conditionally independent. For example, it has been used for really simplistic object recognition in RGB where the three channels were assumed to be uncorrelated.

- Machine Learning (ML) is not a solution for every type of problem. There are certain cases where robust solutions can be developed without using ML techniques. For example, you don't need ML if you can determine a target value by using simple rules, computations, or predetermined steps that can be programmed without needing any data-driven learning.
    - Machine learning can be used for the following situations:
        - It is hard to code the rules or there are many rule factors: Many human tasks (such as recognizing whether an email is spam or not spam) cannot be adequately solved using a simple (deterministic), rule-based solution.
        - The solution cannot scale. You might be able to manually recognize a few hundred emails and decide whether they are spam or not. However, this task becomes tedious for millions of emails. ML solutions are effective at handling large-scale problems.
        - Machine learning is a great approach for many text classification problems, like the email spam case above. However, with other problems, such as classifying a job title into a rank, there may be an approach difficulty in that, unless the training set is very large and sufficiently diverse, a machinelearning solution can significantly overfit it. The term "overfit" means "the learned model does not work adequately well on titles not seen during training". In other words, it is crucial to measure how the model performs on cases not seen during the training phase.

As time and resources permit, it is okay to run several algorithms of the appropriate type. Comparing different runs of different algorithms can bring surprising findings about the data or the business intelligence embedded in the data. Doing so gives you more detailed insight into the business problem, and helps you identify which variables within your data have predictive power.

Some predictive analytics projects succeed best by building an ensemble model, a group of models that operate on the same data. An ensemble model uses a predefined mechanism to gather outcomes from all its component models and provide a final outcome for the user.

9) Deployment Path – Integrated Architecture – Analytic Infrastructure

Consider the outputs of your model and how this will be incorporated into the enterprise or business infrastructure. It is very advisable to be familiar with the data systems architecture and how the model will be managed over time.

10) Performance - Test , Train, Validate

Validating the predictive model you have built is critical. You can generate scores on a test sample and generalize to a whole similar population, bearing in mind that there will be statistical confidence factors to consider.

Broadly speaking, one can talk about several types of validation:

- Apparent: performance on sample used to develop model
- Internal: performance on population underlying the sample
- External: performance on related but slightly different population

Good modelers and data scientists are able to explain in detail how the model was validated and can show the statistical performance of the model (Regression coefficients, chi squares, significance of functions). Analysis of variance (ANOVA) is often used for this purpose. ANOVA is a collection of statistical models used to analyze the differences among group means and their associated procedures, such as variation among and between groups. Below is a quick overview of the main regression calculations for performance purposes.

R Square:

R Square describes the goodness of the fit in terms of the independent variables. For example, 96% of the variation in Quantity Sold is explained by the independent variables Price and Advertising. The closer to 1, the better the regression line fits the data.

Significance F:

To check if your results are reliable (statistically significant), look at Significance F (0.001). If this value is less than 0.05, you're okay. If Significance F is greater than 0.05, it's probably better to stop using this set of independent variables. Most or all P values should be below 0.05. Delete a variable with a high P-value (greater than 0.05) and rerun the regression until Significance F drops below 0.05.

Coefficients:

The regression line might take the form: Y = Intercept +- A*X1 +- B*X2. This is an equation in some number of variables; Where Y is the dependent variable, and X1 and X2 are independent variables. A and B are called the coefficients. These describe the strength (increase, decrease) in the variable relationships. To use a marketing/advertising example: For each unit increase in price, Quantity Sold might decrease by with some number of units. For each unit increase in Advertising, Quantity Sold increases by some number of units. This is valuable information. You can also use these coefficients to do a forecast. For example, if you know the price and Advertising dollars, you could tell how much Quantity Sold you could achieve.

Residuals:

The residuals show how far away the actual data points are from the predicted data points (using the equation).

5. Cognitive Architecture

"I never predict anything, and I never will."

Today's enterprises have many touchpoints where predictive analytics and modelling can be brought to bear. It is increasingly important in today's data-intensive world that the business' system architecture supports predictive modelling development, processes, and outcomes. The term cognitive architecture identifies the technology infrastructure and architecture as well as executive support and proper investment that enables and supports data science practices, predictive analytics, and modelling processes.

5.1 Data Management Strategy

Building a model by itself is a good exercise. However, the real power comes from when the predictions are actually used for business decisions or customer touchpoints. As data drives everything in business today, you can't be successful without a data management strategy. There are five key areas to a successful data management strategy:

- Data quality
  - How to keep your data clean, accurate, and correct
- Data integration
  - How to integrate your data systems together for optimal usage
- Data federation (including access)
  - How to provide secure, federated access and tenancy to the data
- Data governance
  - Having understood policies and procedures for data storage and usage
- Master data management
  - How to efficiently warehouse and manage your data

In particular, when using predictive models in a corporate environment, one must consider a data architecture that allows the model to be transactionally implemented and used and tuned, configured, or updated over time. There should be simple plug-ins or API hooks that can use the predictive model output in a technology application. Enabling easy A/B testing in a live environment, for example, can be a powerful technique.

It should be said that many enterprises are not really prepared for this kind of big data architecture, even if they have modern data warehouses or business intelligence systems in place. The reason for this is that big data involves not only data volumes, but also data velocity and data variety. These are known as the three "V's" of big data. Many companies have simply not invested in an architecture that enables this modern view of data management.

5.2 Basic Model Architecture

The picture below shows a high-level data architecture that builds in predictive models and analytics as a foundational element of the business systems. There are three layers: core data, which stores all the essential data for the corporation; models/ analytics, which performs the predictive aspects; and transactions, which is everything that touches customers or end-users. Dev, QA, and Production environments are important to consider across these three basic layers.

A rough mapping back to the model process framework as described earlier is shown as well (Plan, Build, Implement). It is important to consider how to implement model inputs and outputs in a comprehensive manner, ideally without silos that would preclude certain kinds of data access or usage.It is important to see the model in this architecture as a growing element. The model may be turned on or off, upgraded, or replaced over time, and the same basic architecture can still power the business moving forward.

6. How Do You Know Your Model is Successful?

"Not everything that can be counted counts, and not everything that counts can be counted."

To an important extent, of course, the determination of what is considered a good model depends on the particular interests of the organization and is specified as the business success criterion. Generally, one sees empirical and statistical measures of success, and also more subjective measures of success.

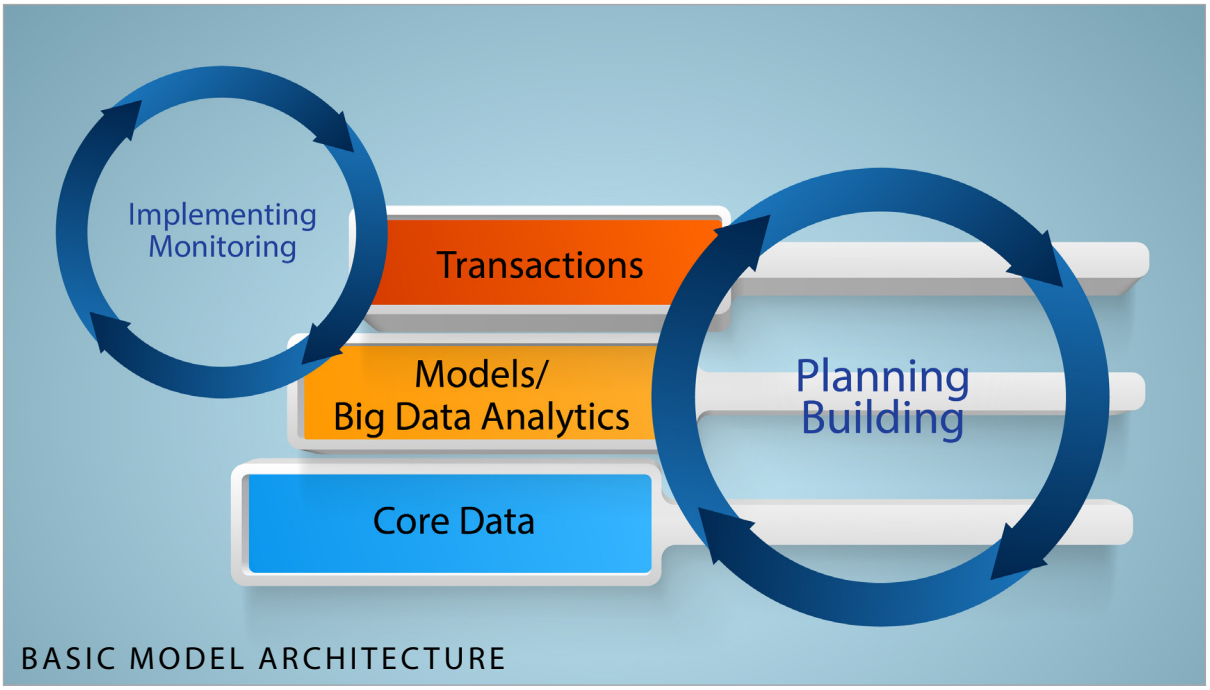
Some principal success metrics or indicators of model success are described below.

6.1 Statistical Uplift

1) Uplift from your model

- Compare predictive model performance against random results with lift charts and decile tables.
  - Using lift charts, hold out samples, and decile tables are a basic way to look at predictive model performance. Typically models are validated against a hold out group. This group contains data that have not been included in the development of the model. For example, in a target lead model, you can look at how much more successful the model is likely to be than if no predictive model was used to target leads.
- Evaluate the validity of your discovery with target shuffling.

  This method is particularly useful for identifying false positives, or when two events or variables occurring together are perceived to have a cause-and-effect relationship, as opposed to a coincidental one. This may happen because the more variables you have, the more things may seem to have a causal relationship and you may "oversearch."



- Test predictive model consistency using boot-strap sampling.

  This method tests a model's performance on certain subsets of data over and over again to provide an estimate of accuracy, not just for statistical significance, but also for operational significance.

2) If the purpose of the model is to provide highly accurate predictions or decisions to be used by the business, empirical measures of accuracy will be used, such as confidence levels or other statistical quantities.

6.2 Business Success Measures

"Success means going from failure to failure with enthusiasm."

Business success measures are important to keep in mind. Often, executives have a good feel for what is helping to drive the business. An analyst might build an elegant model that works and saves the business $80,000 a year, but that is really about the salary of the analyst, so is there really an overall benefit?

1) Is the model intended to lower costs, is this really helping to do that? If the model is supposed to enhance revenue or cross-sell, can this be checked independently? Often, other effects have to be disentangled from the model performance itself.
2) If interpretation of the business is of most interest, accuracy measures will not be used; instead, subjective measures of what provides maximum insight may be most desirable.
3) Some projects may use a combination of both empirical and subjective measures so that the most accurate model is not selected if a less accurate but more transparent model with nearly the same accuracy is available.
4) The project nature of the exercise carries its own success measures as well.

- Implementation efficiency: The ease of use, immediacy, and practicality of model implementation is important. If it is costly to deploy a smart model, that could be a downside.
- Time-to-capability (or value): The best idea in the world has no real business value unless it hits the street and people are using the results. Spending years on a huge abstract idea may be less valuable than quick low-hanging fruit.
- End-user satisfaction: If internal users, sponsors, or end customers are not happy with what they experience, this can be an uphill battle, regardless of how good the model is.

7. Appendix A: References

1. http://www.theanalysisfactor.com/7-guidelines-model-building/
2. https://en.wikipedia.org/wiki/Predictive_modelling
3. https://the-modelling-agency.com/model-development/
4. http://math.gmu.edu/~rgoldin/Articles/StatisticalModelsBookReview.pdf
5. Statistical Models: Theory and Practice (Revised Edition). By David A. Freedman. Cambridge University Press, 2009.
6. https://www3.nd.edu/~steve/Rcourse/Lecture7v1.pdf
7. http://cdn.oreillystatic.com/en/assets/1/event/85/Best%20Practices%20for%20Building%20and%20Deploying%20Predictive%20Models%20over%20Big%20Data%20Presentation.pdf
8. http://www.jerrydallal.com/lhsp/LHSP.HTM
9. http://www.gartner.com/it-glossary/predictive-modelling/
10. http://www.analyticbridge.com/profiles/blogs/the-8-worst-predictive-modelling-techniques
11. https://www.analyticsvidhya.com/blog/2015/09/perfect-build-predictive-model-10-minutes/
12. http://www.predictiveanalyticsworld.com/businessapplications.php
13. http://www.predictionimpact.com/predictive-analytics-training.html
14. http://www.datamine.com/site/datamine/files/White%20Papers/predictive_modelling_process_whitepaper.pdf
15. http://www.plottingsuccess.com/3-predictive-model-accuracy-tests-0114/
16. https://www.cleverism.com/predictive-analytics-forecast-future/
17. http://www.dummies.com/how-to/content/how-to-choose-an-algorithm-for-a-predictive-analys.html
18. http://www.ibm.com/developerworks/library/ba-predictive-analytics2/
19. http://www.kdnuggets.com/

8. About the Author

Dr. Hal Kalechofsky is passionate about using technology innovation to solve new business problems. With more than 20 years of high-tech experience, he has managed data science initiatives and has built analytical computing systems and processes at eBay, CERN and Wells Fargo Bank. As co-Founder of Appiom, Inc., Kalechofsky invented new home networking techniques for parental internet applications. Kalechofsky has led service and business innovation teams at Cisco Systems, eBay, and Coremetrics (acquired by IBM), and has consulted and designed/deployed innovative technology solutions at several Fortune 500 companies.

Dr. Hal Kalechofsky

Dr. Kalechofsky earned a Ph.D. in high energy physics from the University of Pittsburgh and a Bachelor of Arts in philosophy from Tufts University.

LinkedIn Profile:

www.linkedin.com/in/halkalechofsky

SOLOMONEDWARDS

Where strategy meets execution