DataScience®
foundation
MAKE BETTER DECISIONS

Propensity Modelling for Business

Author, Tim Royston-Webb

A Data Science Foundation White Paper

April 2018

----

www.datascience.foundation

PivotalIQ
sourcing intelligence redefined

Propensity Modelling for Business

Propensity modelling is a statistical approach and a set of techniques which attempts to estimate the likelihood of subjects performing certain types of behaviour (e.g. the purchase of a product) by accounting for independent variables (covariates) and confounding variables that affect such behaviour. Normally, this likelihood is estimated as a probability known in propensity models as propensity score. By assigning different probabilities to different subjects based on shared features and covariates, a propensity model allows for the creation of reasonably accurate predictions of future behaviour. This functionality makes propensity modelling a popular technique in various fields including economics, business, education, healthcare, marketing, and more.

Why Propensity Modelling?

Why should we use propensity modelling to infer the causes and implications of behaviours if we have other alternatives such as conventional randomized trials or A/B testing? The thing is that we can't always rely on these statistical methods in the real world. There might be several scenarios where real experiments are not possible:

- sometimes management may be unwilling to risk short-term revenue losses by assigning sales to random customers.
- a sales team earning commission-based bonuses may rebel against the randomization of leads.
- real-world experiments may be impractical and costly in certain cases when the same data or participants can be modelled through quasi-experimental procedures or when historical data is enough to produce actionable insights.
- real-world experiments may involve ethical or health issues, for example, studying the effect of certain chemicals.

Propensity modelling as proposed by Paul Rosenbaum and Donald Rubin in 1983 is an effective solution in the above-mentioned scenarios. It represents a subset of *quasi-experimental research techniques* which are similar to experimental research except that subjects are not randomly assigned to participate in it. They may instead self-select themselves into the experiment or be assigned non-randomly by the administrator.

Propensity modelling includes several approaches and techniques among which **Propensity Score Matching (PSM)**, **Propensity Score Stratification (PSS) and Propensity Score Weighting (PSW)** are the most prominent. In this paper, we will focus on PSM the main technique is used in propensity modelling for business.

In a nutshell, **PSM** attempts to reduce the bias created by confounding variables, that is, factors which affect both independent and dependent variables and cause spurious associations. For example, confounding variables may appear if we compare features and outcomes among customers who used a product and those who did not, without accounting for essential differences between them. And the fact that the apparent difference in outcome between these two groups may depend on the characteristics of units rather than the action itself.

PSM tries to avoid the bias of quasi-experimental research by creating a balance between these two groups; those who used a product and those who did not. The balance is created by basing comparisons on shared covariates; age, gender or product use. In this way, the model attempts to mimic an experimental design *after the fact*. The computed variable - a *propensity score* – captures how the differences in the mentioned variables contribute to a statistical probability of applying to one group or another. The propensity score thus refers to a composite variable that summarizes important group differences. Subjects with a similar propensity score resemble each other. In a business context, propensity scoring helps to identify potential customers. If a person has a propensity score similar to a person who is already a customer, then there is a high probability that they too will become a customer. In this way, we can identify subjects in non-customer populations who are likely to make a purchase given the correct stimulus.

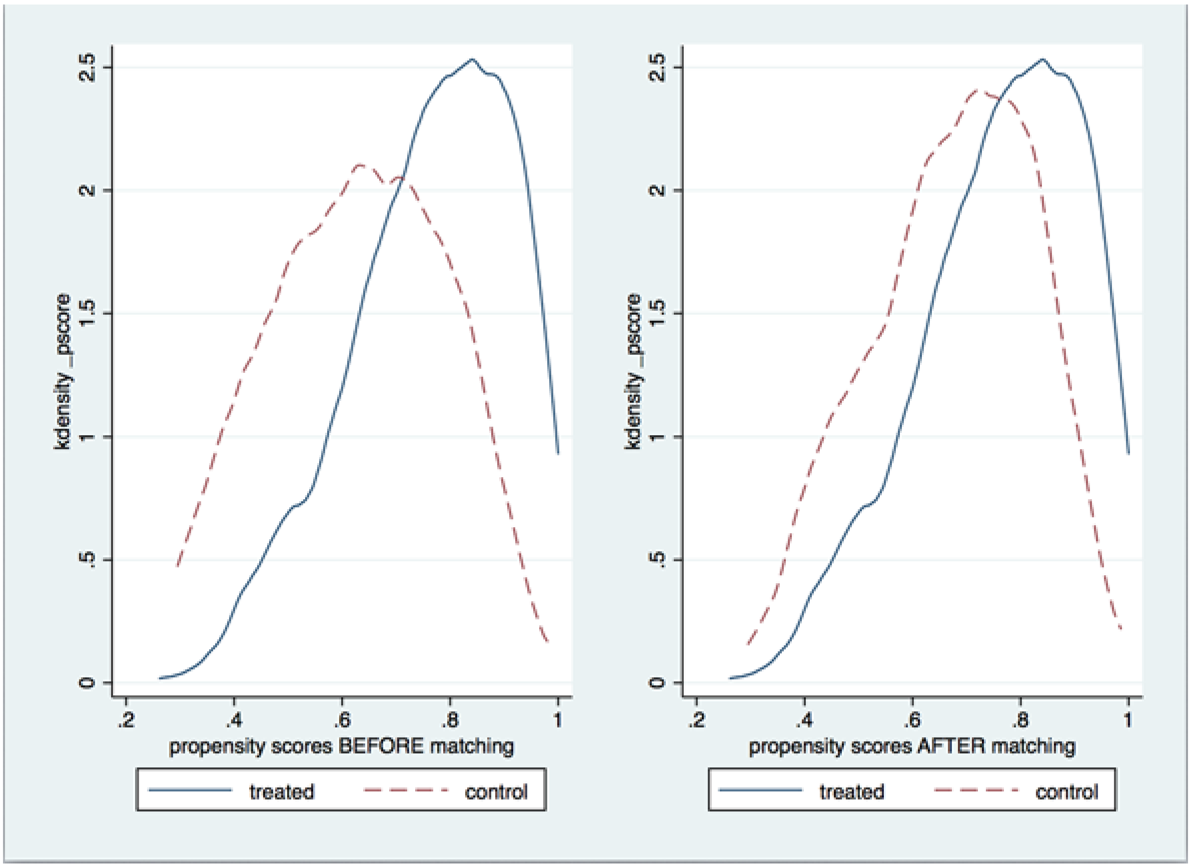

Figure # 1 Propensity Scores Before and After the Matching

The Motivation for Propensity Scoring for Casual Inference: Business Case

By matching observations from each group based on the propensity score, we can achieve a better confounding variable balance between users and non-users of your product. In this case, propensity will be based on shared covariates rather than simple outcomes which are not shareable in both groups.

Formal Definition of the Propensity Score

The propensity score may be defined as a probability that a subject performs a certain action. The action could for example be to make a purchase.

$$Si = P(A=1|Xi)$$

Where A=1 is the probability that a person ends up in a group of buyers given a set of covariates (independent variables) X.

Suppose that age is the only X variable and that older people are more likely to buy your product. Then, the *propensity score* is larger for older people. If a person has a PS of .30, this means that she has 30% chance of being in a group of buyers given a set of covariates.

It turns out that propensity can be the same among people in buyer and non-buyer groups. This understanding can help us identify people who are likely to buy a product if they have the same propensity score as those who have already purchased. To expand on this, think about two subjects who have the same propensity score value, but have different covariates values of X. Despite different covariates, they are both likely to *buy a product*. This means that both subjects X are as likely to be found in the buyer group as in the non-buyer group. If we then restrict the analysis to a subpopulation who have the same propensity score value, there should be a balance between the two groups. Thus, the propensity score is the *balancing score*.

More formally

$$P (X=x)| S(x) = p, A=1) = P (X=x)| S(x) = p, A=0)$$

This means that we can have the same type of Xs in both buyer and non-buyer groups. The implication is that if we match the propensity score, we can achieve a balance and predict propensity to buy without making an offer or contacting potential customers.

Building Propensity Scores in Business

Propensity scores are typically computed using *logistic regression*, with group status regressed on observed baseline characteristics / features such as age, gender and other parameters specified by a study's hypothesis. Logistic regression is not the only approach available with probit analysis, discriminant analysis, tree-based methods, and machine learning (ML) models being other alternatives. ML models for propensity scoring gathered momentum recently with the growth of computing power and advances in the Artificial Intelligence. Propensity scores can be calculated using ML methods such as neural networks and Support Vector Machines (SVM).

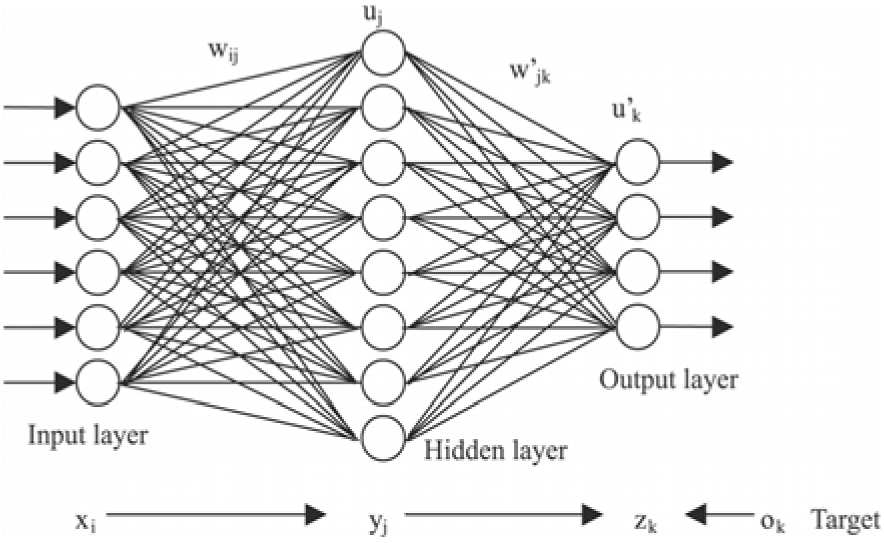

Figure #2 Neural Network Architecture

A neural network is a ML architecture that models the functioning of the human brain. It contains input and output layers and a number of 'hidden' layers, with neurons and connections that process and pass information and configurable parameter weights between each other and gradually adjust those weights and parameters until they match the training data.

It's has been demonstrated that by using neural networks it is possible to develop complex models and abstractions that capture hidden patterns in complex and unstructured data such as images and speech.

Neural networks can be used as classifiers for propensity scoring and have a number of advantages when compared to logistic regression:

- neural networks are better than logistic regression in working with high-dimensional data (data with a lot of covariates and features).
- neural networks with multiple 'hidden' neurons and layers can be extremely fast.
- a neural network of sufficient complexity (i.e., enough internal nodes) can approximate any smooth polynomial function, notwithstanding the order of the polynomials and the number of interaction terms. Such capacity frees the investigator from a priori determining which interactions and functional forms are likely to exist, as they would with logistic regression.

These advantages, along with the fact that stable neural network implementations are available in most data science and ML libraries, make them an important component of propensity modelling.

Support Vector Machines (SVMs)

Linear classifiers such as SVMs make classification decisions based on linear combinations of features of the data points (i.e. covariates). For instance, in the two-class case, the decision rule learned by the classifier may be considered to be a piding hyperplane in the feature space separating those data points into two classes. The main category of ML linear classifiers is represented by SVMs. The major difference between SVMs and logistic regression is that while the latter attempts to explicitly model the probability, SVMs try to directly find the best piding hyperplane (or hyperplanes) regardless of the actual probability of class membership. Thus, an SVM could be used to directly construct propensity categories.

Below are key advantages of SVMs:

- in logistic regression, the data analyst should explicitly choose to increase the dimensionality of the feature space through the addition of polynomial terms. In SVMs, such transformations are a standard practice.
- SVMs are great in dealing with high-dimensional data.
- SVMs do not assume a parametric relationship between the model predictors and outcome. This is a good thing when the actual propensity scores themselves may be unknown and the only task is to determine group membership or its change.

Logistic regression, however, is quite efficient in dealing with the majority of business scenarios. Its main benefit is the ease of implementation and good interpretability. Therefore, logistic regression will be used in the following example.

Algorithm for Propensity Scoring

The usual algorithm for propensity score computation will include the following steps:

1. **Select variables to use as features** (e.g. gender, income, neighbourhood, age, nationality). These variables should be selected dependent on your underlying understanding of what independent factors might affect certain customer behaviour (e.g. buying vs. not buying).
2. **Build a model and prepare data**. The next step is to create a probabilistic model based on logistic regression and prepared features that will predict whether a given subject chooses a certain behaviour. The model should be trained using a dataset of people with a set of covariates and behaviours you are looking for.
3. **Calculate propensity scores for new data**. After the logistic regression step we have created an optimized and trained model, that can now be used to calculate the propensity score for new data (e.g. potential consumers).
4. **Using the model for causal inference**. The created model can then be used for causal inference. For example, to understand the differences and similarities between users and non-users of a product, we can create several buckets which cover subjects with the same propensity score. For example, we can have one bucket for subjects with 0.0-0.1 propensity score, the second bucket for users with 0.1-0.2 propensity and so on. Since propensity score is a balancing score we can then compare users and non-users in each bucket. Since both users and non-users in those buckets have the same propensity score it allows for the controlling of confounding variables and the inferring of actual causal relationships.

Propensity Scoring to Identify Customer Opportunity and Minimize Risk

Propensity scoring has a great value in identifying customer opportunities and minimizing business risk. In general, propensity models (e.g. propensity score analysis) are powerful in building three types of models:

- *Propensity to Buy* model looks at customers willing to purchase and those who need more incentive to complete the purchase.

- *Propensity to Churn* model looks for at-risk customers.
- *Propensity to Unsubscribe* model looks for the customers who have been over-saturated by marketing campaigns and are on the verge of unsubscribing from a service or platform.

In what follows, we'll closely look at how to build *Propensity to Buy and Propensity to Churn Models*.

We would like to understand which users are likely to buy a product and how they are different from those who are less willing to buy it. To develop a propensity model for this task, one has to meet several requirements.

1. **Obtain high-quality data about active and potential customers which includes features / parameters relevant for the analysis of purchasing behaviour.**

   Many companies face the problem of having bad data. Data which has a lot of missing fields, numerous duplicates, poor data entry such as typos, input errors or wrong format. Even a company has a high volume of useful data flowing into its databases, to be useful, it must be labelled, deduplicated, cleaned up, and regularized.

   Another common issue is a *lack of data*. In order to run a logistic regression that will predict consumers' propensity to buy (propensity score) and generalize a hypothesis efficiently, the study must have enough interdependent data points. How to ensure that a study has enough good consumer data that satisfies the above-mentioned criteria?

   There are several strategies to use:
   - **focus on the product**. Create a product that provides the right incentives for the users to contribute more data. By itself, good UI and UX will encourage users to fill out full profiles, enter the correct information and take surveys. This approach is very similar to the *user-in-the-loop* paradigm in which users share data merely by interacting with your platform.
   - **understand what types of data you need**. You should design your product with a good understanding of the types of data required and how this could be obtained.
   - **managing data efficiently**. To avoid data mess and the accumulation of bad data, you should deploy an efficient data warehousing strategy. It can be based on the conventional SQL solutions or advanced distributed data systems like Apache Hadoop or Apache Spark.
   - **filling the data gap using external sources**. Useful consumer data can now be obtained from commercial data providers or partner companies. If they have the data needed for training your algorithms, you should not miss out on the opportunity to use it.
2. **Select the model**. As it was noted above, businesses can select from many alternative models including both statistical and ML ones. However, logistic regression is a good choice for the vast majority of use cases and calculating the propensity to buy, in particular. *Logistic regression* is based on the sigmoid function that has an interesting property: it always maps any real number to the (0,1) interval. Which is why it can be effectively used to calculate the probability (between 0 and 1) of an observation falling within a certain category. For example, a logistic value of .6 would mean that there is a 60% chance that a person can be found in a group of buyers given a set of features and that he / she will not choose the product of your competitor/s. Since a propensity

score is an indicator of a probability of a certain behaviour, logistic regression is an excellent tool for modelling it.

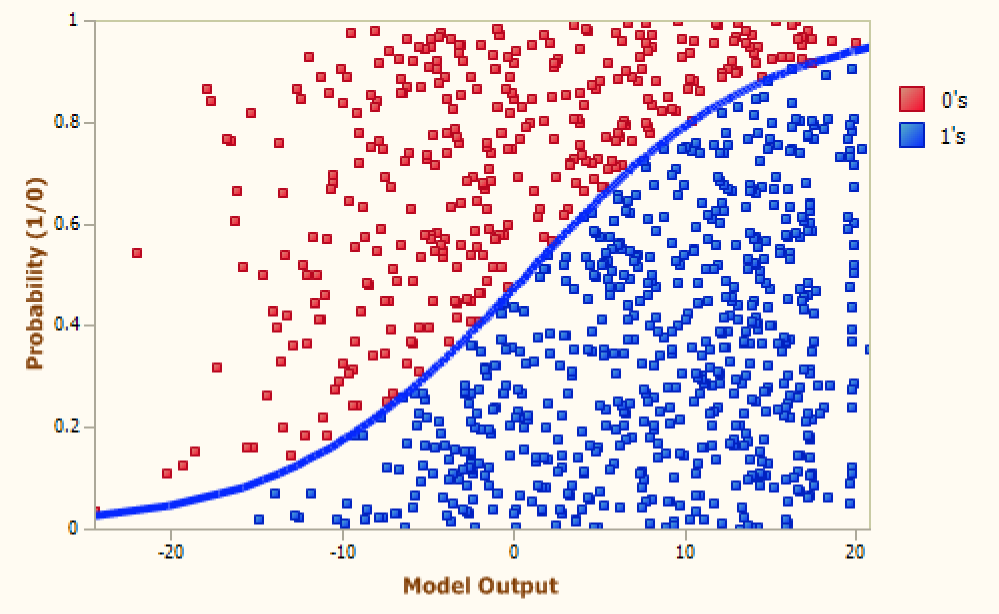

Figure #3 Logistic Regression

3. **Selecting the Customer Features**. This is perhaps the most difficult and challenging part of propensity score modelling. *Customer features* are the independent variables (covariates) we want to measure in terms of their impact on the customer's propensity to buy. Since propensity score analysis with logistic regression is generally a supervised problem, our task is to create an implicit theoretical model of consumer behaviour that will inform the selection of customer features. To define this theory, two questions should be answered. The first one is what factors affect consumer behaviour in the sector, niche or market? Who are the ideal consumers and what cultural, social, personal, psychological features they might share?

   Once a proper understanding of customers has been gained, it is time to select their most relevant features. As an example, these could be user income, past purchases, age, gender, a mobile device with which they access your app/website etc. These metrics can provide you with a detailed background of users who buy your product. When selecting customer features, however, limit the choice to the most important ones without cluttering the feature space and try to avoid duplicate features like income / earnings, for example. Another rationale for feature limitation is that using higher order polynomials with logistic regression requires a manual setup and can lead to overfitting (your model's dependence on training data and inability to generalize to new data). Once we've defined a set of customer features, we can now place our data with these features into any appropriate format: csv, json etc. The data is now ready for model training.

4. **Running and testing the model**. We can now train our model using an efficient computational algorithm such as *gradient descent* or another similar algorithm. The aim of the training is to reach a global optimum which will make our hypothesis fit the training data. Once this is done, however, it's crucial to do back testing of the model against the *validation and test* sets to ensure that there is no overfitting of the model to the training data. In other words, we need to ensure that the model works well with the real-world data of other customers.

   That's it! Now we have a working mechanism for calculating the propensity score of any user or potential customer in our system. Each time a new user arrives, we can feed his / her data to the training algorithm to calculate his / her propensity to buy. In addition, we can make various causal inferences for customers based on our model as mentioned above.

Propensity To Churn Or Assessing Factors of Customer Engagement

The same setup as described above can be used to assess the customers' propensity to churn or factors that affect customer engagement. A high rate of churn indicates that your business is not serving its customers well. They may be willing to stop using the product because of its bad UX, poor performance, or the lack of useful features. To avoid consumer churn, we should understand the precise reasons why customers would stop using your product, software or service. As in the propensity-to-buy modelling, we need user data and a set of features for this data. This time the features could be various parameters of your web application's UI, product, platform or any other parameters that might affect the customer's feedback. For example, we could take a list of features such as installing a mobile app, signing up for a newsletter, following certain users, logging in, clicking specific pages and build a propensity model for each one. The model will rank each feature by its estimated causal impact on user engagement. Based on the propensity scores for these features, we then can determine which of them have the highest impact on user engagement and prioritize their improvement or modification. Such propensity analysis, which is, in essence, a sophisticated version of the engagement regression model, can be used as a continuous cycle for the improvement of your UX and UI which will ultimately reduce customer churn and generate better user engagement.

Other Business Advantages of Propensity Scores

As we have seen, propensity models allow managers and analysts to identify potential customer opportunities in various groups of users and to improve user engagement. Propensity scores can be also directly used to create better marketing campaigns and targeting which will ultimately increase sales.

For example, we can use results from a test mailing to score contacts in a database who have not yet received the mailing. Feeding their data into our propensity model, we can calculate the likelihood of a person responding to our marketing email and generating a lead. Marketing materials and campaigns can then be tailored to inpiduals based on their estimated propensity to purchase. This approach is much better than expensive broad-based blanket marketing or random marketing because it allows for the allocation of limited campaign resources in the most efficient way to generate more sales.

Propensity models are useful in the following situations:

Test marketing campaigns. With the propensity model, we can test marketing campaigns in the shopping locations and even cities before the fact to assess how effective the campaigns would be given a set of parameters similar to other shops or cities where our campaigns have been run already.

Measure marketing efficiency. Propensity models can analyse marketing effectiveness by channel to reveal various trade-offs between investments in different media and improve marketing ROI estimates.

Reduce costs. Propensity score can increase the ROI of your marketing campaigns if several simple rules are followed. First, contact people, for whom a campaign has a higher chance of success and, therefore, is likely to result in increased revenue from the campaign. Secondly, understanding propensity scores by customer groups you can give each customer the minimal offer needed to attract them. Such offers will be more advantageous when used with customers with a medium propensity than for those with the high propensity, as these reduced offers will lead to significant savings.

Prevent fraud. Propensity scoring can determine if a customer applying for your services is a bad debt customer or a fraudster. A propensity model embedded into an application can determine how to handle the customer – completely deny the service or offer an alternative product / package. This type of fraud prevention results in the maximization of revenue and minimization of risk.

When implementing propensity modelling in business, however, remember that a high propensity score is not the same as a high probability to buy. Instead it represents a probability that a potential customer has attributes similar to identified groups of buyers. Even though the potential customer might share the traits of historical buyers, the product and the offering must remain relevant, price competitive and at least compatible with the latest offerings of competitors. Therefore, the earnings a company generates through propensity modelling is also dependant on the quality of the offer, product, service and the competition in the market.

Summary

Propensity modelling is a powerful technique that can generate better insights from customer data, open new customer opportunities, improve marketing campaigns, minimize risk and reduce spending while ensuring better and faster decision-making. Propensity score matching can also improve the causal inference of factors that affect customer engagement, behaviour, churn, and willingness to use a service. Propensity models like propensity score matching leverage the power of statistics, computer science, and ML to connect thousands or even millions of data points to generate actionable insights. However, using propensity modelling in business is not straightforward. Its successful implementation is premised on the availability of efficient data acquisition strategies, high-quality data, and an attractive product offering. Without good data, it's hard to find associations and patterns that drive consumer behaviour and decision-making. Therefore, successful implementation of propensity modelling should be accompanied by the gradual transition of your company towards a data-driven business with operations and decision-making based on the state-of-the-art technologies of data science and machine learning.

Pivotal iQ Case Study

One 'live' example of the power of propensity modelling is given by IT intelligence provider Pivotal iQ. One of its clients, Company X, urgently needed to target customers of its rival Oracle – the multinational computer technology corporation, headquartered in California. However, Company X's visibility in the market lacked depth, meaning it was unable to get its cloud computing offering through to Oracle's incumbent customer base.

Using its data analytics tools, SpendView and InstalledView, which allow for the selection and review of many firmographic and technographic covariates, Pivotal iQ enabled Company X to build up an ideal customer profile, defining the characteristics of Oracles' customers. These characteristics representing the Ideal Customer Profile (ICP), ultimately led to the identification of the specific Oracle customers that were running legacy on premise Oracle products spending at least a projected £1m on ERP implementations over the next 12 months, the identified ICP for the use case.

Informed by the ICP and the defined competitor customer list, Company X confirmed it has "already generated many new opportunities using Pivotal iQ's IT intelligence with deal sizes ranging from £250K to £10M+" and recently closed its first deal from the campaign in less than 13 weeks open to close.

Pivotal iQ®

At Pivotal iQ we believe that 'what you see depends upon what you look for.'

Pivotal iQ's position and experience is that IT Vendors using predictive analytics to build customer propensity models are outpacing the competition and growing their market share exponentially. The intelligence sector to date has provided too narrow a view of market opportunities largely due to an inability to fully synergise critical data into actionable intelligence. Data is often of poor quality, is badly structured and not dynamic so often doesn't fit well into a company's business process further constraining what value can be derived.

Using the latest techniques, we develop large data assets on the supply and purchasing activities of the IT industry and transform them into actionable information to deliver business advantage. Our unique big data approach powered by world-class Propensity Modelling helps uncover the subtleties and previously hidden activities that point the way to real new business opportunities. Opportunities that remain hidden to competitors.

Our tools allow our clients to better understand the customer environment before they make that call by understanding what they spend, what they have purchased, what has been outsourced and their budget position. Pivotal iQ deep data sets use the latest technologies and methods that provide a deeper and more accurate view on the industries supply and procurement activities than ever before.

Pivotal iQ®

UK (Group Headquarters)

152-160 City Road, London, EC1V 2NX
Telephone: +44 (0)207 060 7080
Facsimile: +44 (0)207 060 7081

USA
42 Broadway, Suite 12, New York, NY 10004
Telephone: +1 (212) 634 4620
Facsimile: +1 (212) 634 4621

India
Pivotal Research Centre
1st Floor, Anshu Colors Building, Road Number 1, Park View Enclave, Jubilee Hills, Hyderabad, Telangana 500033

www.pivotal.iQ

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation