Advancing the Data Product

By Michael Lieberman, January 19, 2002

Beginning 2022 the marketing research industry will expand into the data strategy business. Companies store fields of data within their silos—large retailers like Kroger's and Walmart are already selling access to their data. They would like to refine that data into either an actionable predictive model or a commercial product. I believe they will turn to analytics translators, quantitative professionals now working in marketing research, to accomplish this.

The Data Company Conundrum

Savant Datex is a mock data company. On their website they advertise the company "provides advanced predictive business intelligence using sophisticated analytical software to examine our dataset to provide answers to executives for "forward-looking" companies such as retailers, brands, marketers and financial service companies."

In reality, though, Savant Datex sells access to their data and monthly consumer sentiment summarizations. They would like to introduce a more lucrative income stream with the recourses they have at hand. They want a competitive advantage in today's emerging data-driven milieu. In other words, Savant Datex wants to build a predictive model based on their proprietary database and hawk it on the data market as a subscription service.

There are many companies like Savant Datex. A glance at Amazon Web Services (AWS) shows a large number of them in many and varied industries. The most common of are consumer sentiment, brand behavior, and financial modeling. They report with headlines such as 'Nearly 180 Million Shop Over Thanksgiving Holiday Weekend', 'Internet Enables Care Providers to Gain Rapid Access to Information That Can Aid in the Diagnosis Of Health Conditions', or 'Global Brands and Influencers Emphasize Importance of Personalization for the Future of Fitness'.

BEHIND THE DATA

A closer look at the services provides shows that the vast majority of these offer little more than simple predictive percentages. The Savant Datex Consumer Confidence Survey reflects prevailing business conditions and likely developments for the months ahead. For example, sales of motor vehicles according to US government outputs are stable, so we expect these sales to be similar this year. The consumer credit confidence index rises, so we expect a rise in consumer spending within, say, retail sales.

It works well enough, until there is a profound disruption such as a worldwide pandemic.

This is commonly known as 'Nassim Taleb's Turkey Problem', when a turkey is asked to determine the population of turkeys in December based on the growth of the turkey population from January one year before. One problem—Thanksgiving. The turkey being surveyed is not aware that he, and most of his fellow turkeys, face a bleak future given the Thanksgiving holiday.

The Corona virus pandemic has presented the marketing research industry and data companies, with a real-world Taleb turkey problem. Data companies who tout their predictive power did not take into account the global disruption of the pandemic. Thus, their data from pre-March 2020 has lost its predictive power to explain consumer outputs during and after the pandemic. However, the construction of a more predictive analytics model can incorporate and digest the pandemic disruption. This mathematical model is a 'data product'.

In this article we are going to navigate how a data company can take their inventory, collected survey data, shopping history, or large consumer sentiment studies, and create a subscription service much like the one Savant Datex seeks. This is especially relevant as data companies seek to re-engage their vast amounts of historical data that has now become stale. We will begin with the stated goals, and walk through a flow chart of the process. After a case study we will list examples of other common predictive data models.

BEHIND THE DATA

The Savant Datex Structure

Savant Datex collects monthly consumer surveys in such areas as consumer sentiment, brand preference, abd financial behavior. They collect data using any number of collection methodologies, from traditional surveys, accumulated sales performance, and brand propensity measurement. The larger data companies such as Kroger's, Amazon, or Wal-Mart sell their customer databases.

Clients use this data to predict consumer behavior using attributes such as consumer emotion, happiness, impulsivity, and mood so you can create a strategy that accurately reflects demand and drives more revenue. Many measure financial product behavior.

Summarized in Table 1 are the various sections of the Savant Datex Survey of the US Consumer.

Table 1

Savant Datex Survey Components	Consumer Product 90-Day Intentions
What State Is Your Primary Residence	Retail and Food Services Sales
Head of Household	Health and Personal Care Stores
Consumer Spending: Consumer Goods	Clothing Purchases
Consumer Spending: Durable Goods	Furniture Stores
Consumer Spending: Discretionary	Restaurants and Other Eating Places
Consumer Spending: Staples	Sporting Goods
Economic Optimism	Toys
Consumer Mood - Housing	Electronics
Consumer Mood - Credit	Groceries
Consumer Mood - Economic Security	Home Improvement and Hardware Supplies
Consumer Mood - Political	Internet Shopping
Key Demographics	Travel

In addition, Savant Datex produces quarterly consumer indexes. These are summarized in Table 2.

Table 2

BEHIND THE DATA

Savant Datex Proprietary Indexes
Consumer Sentiment Current Index
Consumer Sentiment Future Index
Housing Index Baseline
Automobile Index Baseline
Economic Optimism Index
Credit Card Index
Political Optimism Index
Job Satisfaction/Security Index

Determining the Dependent Variable

The most challenging part of an effective predictive model is to determine the 'dependent variable'. In other words, the key indicator to be measured—the 'forecast US output' of the commodity in real US dollars.

The business world cannot survive without them. The most prominent example is price (fluctuating prices were absent in planned economies and the key reason that all planned economies have failed). For example, the price of beef is the dependent variable when a farmer determines how many heads of cattle to slaughter. The price of electronics determines the demand for devices and the prediction of consumers' response to price changes. The adage that 'everything has its price' is, in fact, true.

It is the US sales or commodity output that Savant Datex seeks to predict for its data product.

Savant Datex Case Project

Each month the US Census Service publishes the *Advance Monthly Sales for Retail and Food Services*. These are estimates of U.S. retail and food services sales. The www.census.gov website describes this as, "The advance estimates are based on a subsample of the Census Bureau's full retail and food services sample. A stratified random sampling method is used to select approximately 5,500 retail and food services firms whose sales are then weighted and

benchmarked to represent the complete universe of over three million retail and food services firms.".

Here is a key component to the Savant Datex strategy. It publishes its monthly prediction two months *before* the government's Advanced Monthly Sales Estimates. The proposed Savant Datex Prediction Model is the definitive data product. We will explain.

Figure 1 illustrates the process.

Figure 1
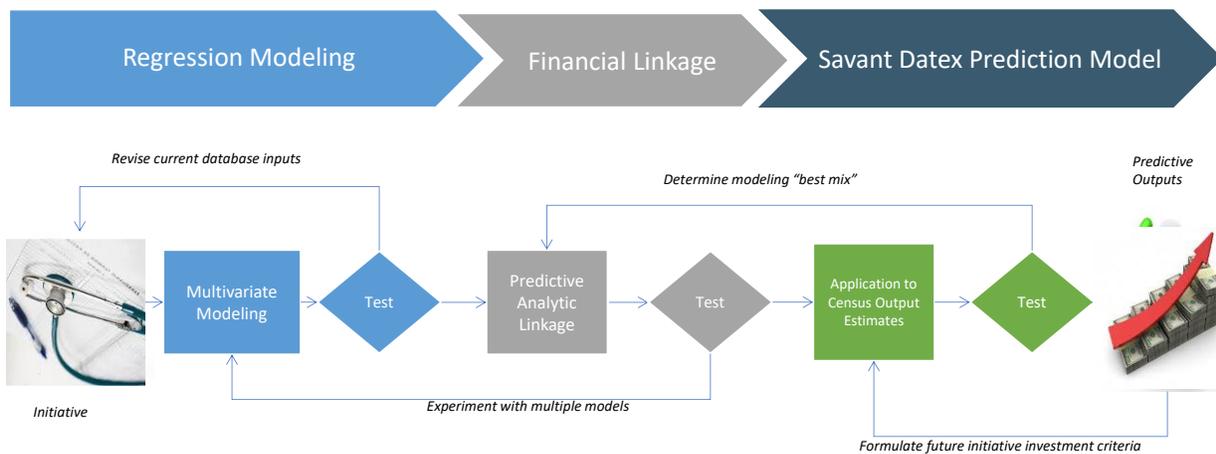
Savant Datex will produce several Prediction Models for subscription. Among them:
- Retail and Food Services
- Clothing and Clothing Access
- Department stores
- Retail sales and food services excl motor vehicle and parts

Figure 2 show a comparison of Savant Datex Predictive Model with US Census output for Clothing and Clothing Access.

BEHIND THE DATA

Figure 2

Savant Datex Predictive Model - Clothing and Clothing Access
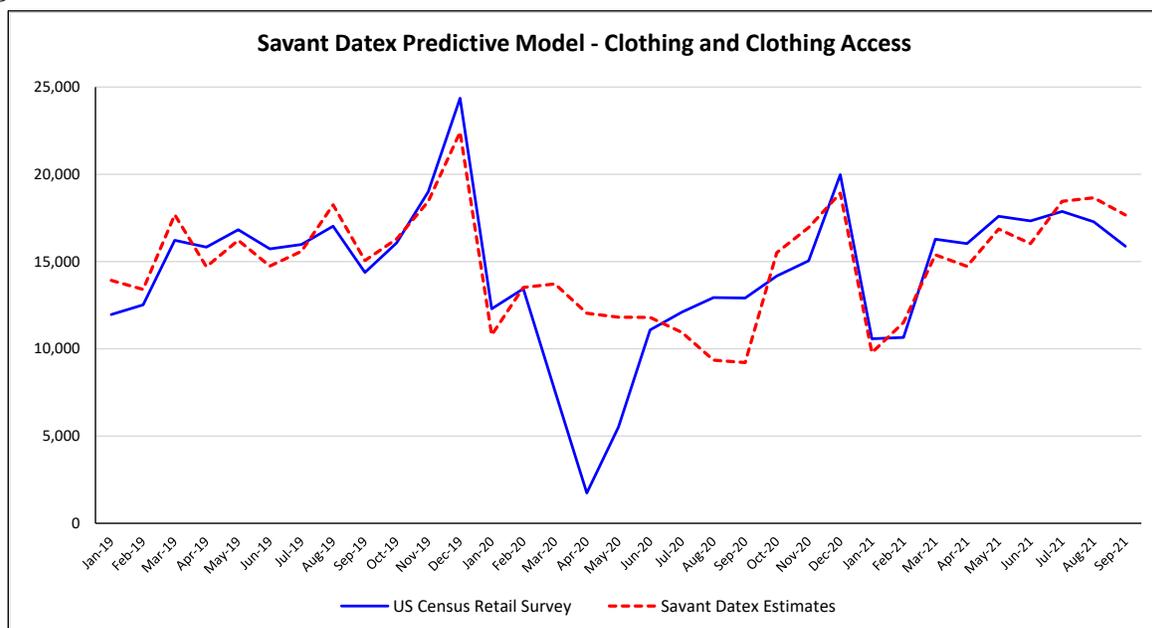
As we can see, the Savant Datex model is fairly predictive (correlation .80 pre-Covid) until the disruption. As expected, the predictions go haywire due to the pandemic. As time passes and more government data becomes available (this estimate's dependent variable), the model claws its way back. It will become more predictive as time continues.

There are over 9000 hedge funds in the world managing trillions in assets. In order to improve their performance, hedge fund managers began to seek ways to seek new sources of untapped data to improve investment decisions. These are the intended target of this particular Savant Datex Predictive Model.

Advantages of the Predictive Model Over Trend Summarizations

While trend summarizations are enormously important, they leave much information on the table. Yes, it is possible to use correlation between various indexes. However, correlation is not proof of causation.

The Savant Datex Prediction Model uses predictive analytic regression modeling, and thus establishes drivers of the dependent variable. For example, we can determine for clothing sales whether sales of, say, groceries or sporting goods drive clothing sales consumption. We are able to append external data—like PRIZM (Potential Rating Index for Zip Markets) category data—to help explain what other factors are driving clothes consumption. Figure 3 illustrates descriptive depth that would be tremendously useful for potential clients, the hedge funds.

Figure 3
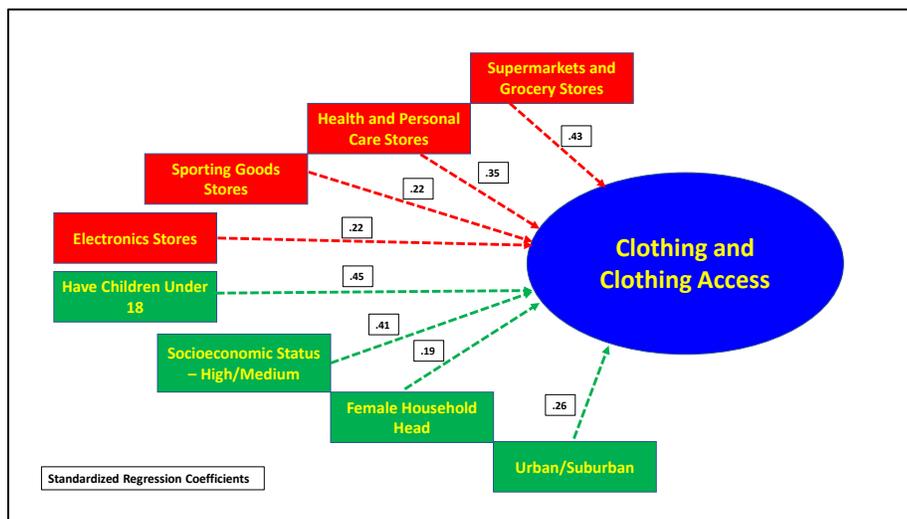
By examining the chart we can see that there exists a direct and strong link between clothing consumption and supermarket purchases, socioeconomic status, and the presence of children under 18. The predictive association with clothing sales is less potent for sporting goods. Whether the household is lead by a man or woman is the least predictive variable in our example.

BEHIND THE DATA

Potential Predictive Models Using This Methodology

We have illustrated one example of the power of a predictive model. This methodology is useful for a range of predictive subscriptions. Among them:

- Assigning a 'racial or gender bias' score to adverting campaigns or ads
- Predicting a 'cost per click' per banner ad based on 'likes'
- Estimating customer churn
- Calculating customer lifetime value
- Scoring brand purchase propensity
- Approximating employee absenteeism and theft

Conclusion

Companies store fields of data within their silos—large data retailers like IBM® Cloud Solutions and the Schwarz Group (Germany) are already selling access to their databases. They would like to refine that data into either an actionable predictive model or a commercial product. I believe they will turn to analytics translators, quantitative professionals now working in marketing research, to accomplish this.

Editor's Note: Michael Lieberman is founder and president of Multivariate Solutions, a New York based research consulting firm. He can be reached at 646-257-3794 or at michael@mvsolution.com.