Using AI Deep Learning to Leverage Big Data for Investment Attraction

GAZELLE.AI

I. ABSTRACT

The explosion of AI in consumer products and in our everyday lives testifies to how useful this technology is and how widespread its applications can be. We have seen AI applications everywhere from home assistants to facial recognition to software designed to develop self-guiding cars and combat drones. The benefits of AI for socioeconomic analysis are that it is a more versatile tool, compared to statistics, that can handle large amounts of imperfect data. AI can also help tackle large-scale difficult to understand problems that may emerge from the interactions of individual agents. Further, when coupled with big data, researchers can analyze more recent and substantial amounts of a greater volume of detailed information compared to static, old, and aggregated government data. The benefit of this is that we can obtain more insights into things like spending patterns and consumer and business behavior, and one can track activities spatially or for individual agents; this can be very helpful for planning. The effects of big data on economic development are being felt on multiple levels but nowhere so acutely as in targeting strategy and lead generation.

In the Economic Development world, the Gazelle.ai investment attraction platform is unique in that it accesses big-data including industry, spatial, and firm-level characteristics going back more than 10 years. This data is combined with AI techniques like dynamic deep learning, neuroplasticity, dynamic input attention, concurrent AIs, and multi-GPU processing. The result is a platform that can quickly and comprehensively investigate the direct and interactive factors associated with firm actions, including expansions, in ways uniquely different from traditional statistical methods. Machine learning allows for continuous improvement in the structure of the algorithms via experimentation and model comparisons, and our team's continuous API-based data acquisition and curation process provides an expanding and a more accurate dataset for the AI to work with. Over time, this allows for increasingly sophisticated algorithms and more precise predictions. In practical terms for Economic Developers, that means fewer attempts being made to find the fast-growing, likely expanding organizations referred to as Gazelle companies. This whitepaper provides a broad description of the evolution of AI, some of the specifics of our approach, and our predictions for the coming year.

II. LITERATURE REVIEW

The concept of AI first achieved recognition in the period between 1950 and 1970. During this era, AI achieved prominence in psychology through the development of algorithms that mimic the human thought process. The term mimic has had multiple interpretations in this context including one that focused on internal validity or making algorithms representative of the actual human thought process, to the extent that this can actually be known. This was referred to as information processing psychology. Another school of thought, which influenced the approach employed on Gazelle.ai, focused on external validity. External validity focuses on being able to reproduce the decisional outcomes of humans. This field became known as cognitive science.

Within the broad AI field at this time, a difference of opinion arose about the degree to which algorithms should be representative of the human cognitive process. This difference is in large part responsible for determining the adjectives used to describe the two approaches - artificial neural networks (NN) versus machine learning (ML).

Some psychologists believed algorithms should be representative of the human cognitive process. These were mostly called the neural networks. Economists have a similar belief regarding their models, namely that models should be restricted to representations of existing economic theory or should only test static/predefined relationships. This contrasts with psychologists (and some economists) who favored unrestricted models, mostly called machine learning algorithms. In these models you can toss in everything including the ‘kitchen sink’ and rather than cooking the books by initial assumptions, you let the data tell you about the relationships.

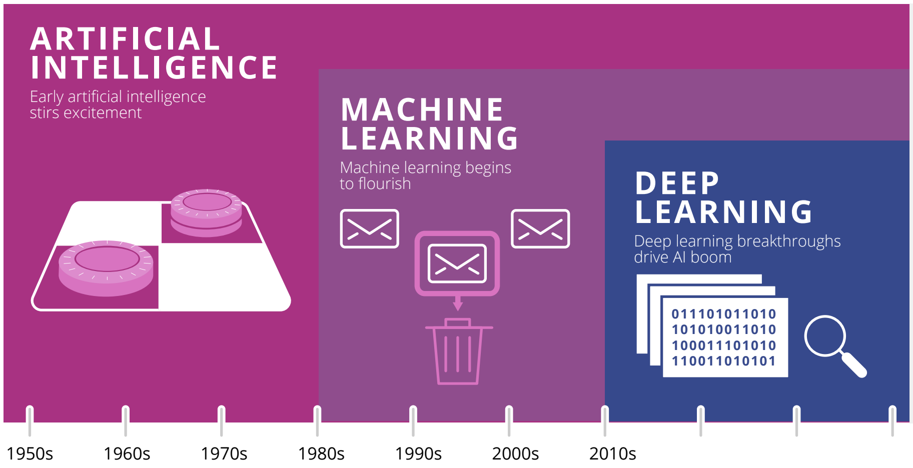
During and after the 1970s, three significant extensions occurred:

1. Improvements to basic principles (i.e., algorithms’ mathematical structures, the elements included, and assumptions required).
2. Building AI algorithms that allow corresponding psychological theories of decision making to be built.
3. The development and application of AI algorithms without being required to be representative of biological systems. This is where the Gazelle.ai approach resides (i.e., in the interdisciplinary field of cognitive science).

During the 80s and 90s, we saw a renaissance of AI applications that were less concerned with the theoretical and academic debates and battles that marked their predecessors. One of the many uses of these algorithms was in the laboratory investigating human and animal decision-making. There was also a renaissance in applications for the private sector, industry, and public-sector institutions. For example, AI was used to predict variables such as firm or farm output, land use patterns, international trade patterns, and asset prices. Some other applications included applications to economics, finance, and business (HaiHan2009), as well as private sector forecasting of GDP growth, currency in circulation, electricity demand, construction demand, and exchange rates (Fernandez-Rodriguez, Gonzalez-Martel and Sosvilla-Rivero, 2000; Redenes and White, 1998). It was also used in Government sector forecasting focused on various macroeconomic indicators and was used by the Czech National Bank (Marek Hlavacek, Michael Konak and Josef Cada, 2005), Bank of Canada (Greg Tkacz and Sarah Hu, 1999), and Bank of Jamaica (Serju, 2002). Other finance sector forecasts considered stock returns and stock selection, bond ratings, credit assignments, and property evaluation. (see Chen, Racin and Swanson, 2001; Swanson and White, 1997; Stock, and Watson, 1998)

The current era, loosely described as 2012 and beyond (due to hardware innovations in 2012), is characterized by applications with ‘Deep Learning’ or multiple layers of dynamically determined ‘hidden nodes’. Hardware improvements such as better GPUs and TPUs allowed theories developed in the 1950s and 1960s to finally find real-world applications. The availability and exploitation of big data also allowed large-scale training and massive accuracy improvements.

Today, many consider AI to be the general term that encompasses all aspects of artificial decision making. Michael Copeland, a well-known technology writer, describes this field as a series of concentric boxes, with smaller boxes being a subset of the larger ones.



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence — first machine learning, then deep learning, a subset of machine learning — have created ever larger disruptions.

Source:
M. Copeland 7/29/16

The largest box represents the broad field of AI encompassing early and recent developments from 1956 (the date of a seminal conference and meeting of minds) to today. The second box is machine learning, or ML as we'll refer to it from here on in, aka statistical AI or depending upon your branch of the AI religion, artificial neural networks. These were applications of AI theories to cases as earlier described. For the most part, deep learning was not or only marginally applied in this work (due to software and hardware limitations and associated costs). Lastly, the third box is deep learning, which has only recently been open to widespread application.

III. WHAT DISTINGUISHES GAZELLE.AI'S USE OF AI

Given this background, an important question to ask is: what distinguishes Gazelle.ai's use of AI? Aside from being the only AI application specifically designed for Economic Developers, there are 5 specific areas where Gazelle.ai sets itself apart.

1. Over multiple years and working with many experts, we have developed a unique custom database that integrates and validates multiple public and private data sources. This, coupled with extensive interpolation/ extrapolation of suppressed and missing data, allowed us to develop one of, if not the most, comprehensive US and international databases describing firms of interest to economic development agencies and business attraction and retention officers and institutions. As of July 2018, we have records for more than 6 million firms, and the list is continually growing and updating.

2. Gazelle.ai runs on multiple custom-designed applications of our machine learning algorithms, based on the work of peer-reviewed, published academics, and field practitioners with over 20 years of experience. First and foremost, we apply these algorithms to the creation of our Gazelle Score (GScore™). We also utilize them for interpolation of missing/ suppressed data, and for extrapolation or forecasting of future industry, regional (County and State), and firm-level data. Furthermore, as anyone attempting to apply these methods to flawed or incomplete data knows, these approaches cannot be robustly accurate using plug-and-play, prepackaged software. With this in mind, we developed custom ML algorithms.

We only use algorithms that are human subject or macro-scale data validated. We apply both shallow and deep learning and both static and dynamic models for deep learning.

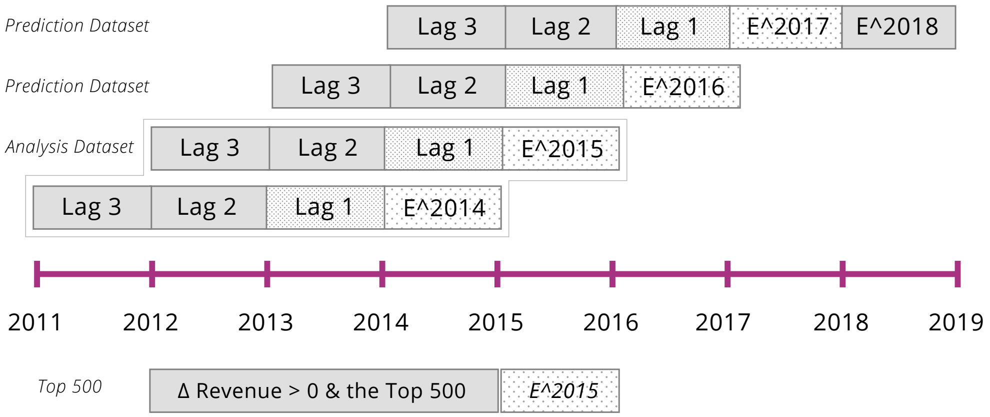
3. Although it may seem less exciting, it is essential to have a solid baseline to compare the AI algorithms' predictions to. We always apply standard/traditional statistical methods to get a baseline for predictions. The predictions of the ML algorithms are then compared to the statistical approaches to guard against, or at least be aware of, unanticipated variance.

4. Gazelle.ai runs on state-of-the-art, in-house hardware and software. This allows us to harness vast pools of big data and to produce predictions and update predictions at scale and in a timely fashion.

5. We take validation very seriously. All our training and predictions are out-of-sample and include extensive model comparisons. This allows us to improve performance, nail down the most relevant architecture for our algorithms, and update the AI architecture as real-world conditions change. As an added bonus, our parent company ROI Research on Investment employs a large team of Economic Development Associates who continually collect survey and other information directly from the firms and their websites giving us access to firm-level data on a larger scale. In the cases where neither of these approaches works, our ML missing data estimation algorithms fill in the remaining holes. This makes our dataset among the most complete and accurate for economic development, and it keeps improving. It also means our algorithms can be better trained than any single public, private, or data source.

IV. DATA

A graphical representation can help clarify how we use our data, particularly how we partition it into three separate samples (Analysis, Validation, and Prediction datasets). The image illustrates how our predictions compare to others (e.g. Top Lists). Although the exact variables we use and their sources are proprietary, the method we use to create these datasets and how they are used can be informative. Below is a representation of how the data was partitioned for the 2017 Gazelle.ai launch.



A. Analysis dataset

To simplify, imagine that we only have data describing actual expansions or non-expansions for 2014 and 2015. In reality, this data spans 2011 all the way to 2017. This means that we can use previous years' information to predict events for these two years and train our algorithms. **Lag 1** for a 2015 event means we have predictor variable information for 2014. Similarly, **Lag 1** information for a 2014 event means we have predictor information for 2013. **Lag 2** information is 2 years back from the event data, and **Lag 6** would be 6 years back. Our ML algorithms are trained with a sample of 35K firm events for these event years and their lagged information by estimating appropriate learning and neuroplasticity rates that maximize the fit for the event years (training using these event years is based upon an out of sample structure). After an analysis run is complete, a set of ML parameters and association weights, and other program control parameters, are output.

B. Validation dataset

The second step in the process involves taking those parameters and weights for use in predicting expansions in 2016 (this is data not used in the analysis). This 2016 data is also for firms where we are aware of the outcome. By applying the analysis dataset parameters to the validation dataset lagged predictor data, we can recover the predicted propensity of 'expansion or no expansion' for the 2016 firms. These predictions are then compared to 2016 reality, and a validity measure is generated. In fact, this analysis and validation occurs "on the fly" in the AI program where hundreds of thousands of runs using different lags, mathematical structure of the algorithms, and endless combinations of hidden nodes (deep learning) are tested and validated or discarded.

C. Prediction dataset

Once the analysis goodness of fits and validation accuracies reaches desired levels, and the results are robust to perturbations in parameters, predictions can be made. This is accomplished by assuming the event date is the current year (2017 in the above case), lagged predictor information can be collected. This information is then combined with association weights to generate propensity of expansion in 2018 for each firm. Using calibrated program control parameters, these propensities can then be converted into GScores™.

D. How Gazelle.ai compares to top lists

The approach discussed above can also help investigate the hypothesis that revenue growth predicts expansions, which is the basis of many featured lists, such as the Top 500. The approach can also determine how such predictions compare to the process used by Gazelle.ai. In academic literature, there is no clear conclusion about the validity of this relationship, despite the widespread use of this idea by practitioners.

Assuming that such an approach is valid, typically a firm would find itself on such a list if it self-reported that revenue growth was positive or positive beyond some threshold in the previous years. How the data are used and related to the creation of a top list for 2015 is represented in the lower panel of the figure above. Intuitively, presence on a top list in 2015 requires 2012, 2013, and 2014 to have positive revenue growth. In most cases, there is not a statistical analysis that leads to a goodness indicator or threshold for a firm being on a list, it is rather an accounting relationship (i.e., if something surpasses an arbitrary threshold, then you are on the list). Further, many of these lists involve self-reporting by the firm, who report in good years, but not necessarily in bad years. Statistically speaking this leads to several biases such as the “lemons problem / asymmetric information” bias and the “stated preference” strategic bias. Though it may be good marketing, using Top Lists for targeting is more accounting than analysis, and as such is fundamentally different than our approach.

V. HOW DOES THE GAZELLE.AI PROCESS WORK?

Gazelle.ai uses a multiple method comparative analysis among various types of concurrent AIs, and validation via traditional statistics. It is worth noting that although we validate against traditional statistics, the predictions on our platform are entirely based on the AI approach.

For the statistics method, the goal is to determine how a broad set of firm, industry, and geographical predictor data X are associated with firm expansions E=1. Challenges for this method include violations of statistical assumptions such as: missing / suppressed data, different variable distributional properties, co-linearity, endogeneity, outliers, large numbers of predictors, and variable data scales to name a few. To resolve/minimize these issues we utilize a number of steps which have identical or analogous applications for the AI approach.

A. Traditional statistics method

The first step toward resolving these issues is data normalization. Since the data are in multiple scales ranging from decimals (for percentages) to billions of dollars for revenue, we need to rescale all data into a common range that preserves the sign of the predictor variable. This facilitates the identification of relatively more important drivers, avoids attending to large values only, and allows estimated parameters on predictors to have intuitive signs. While a variety of methods are available, we apply the approach used in Kelley Van Rensburg, Jeserich (2016); Van Rensburg, Kelley, Jeserich (2015); and Jeserich, Kelley, Toft, Cole (2012). In these works, both z-scores and custom designed median scores are used to normalize all variables in terms of their standard deviations or median average deviations. Intuitively, a value of 2 for a predictor indicates the value lies 2 standard deviations (or MADs) above the mean (median) for that predictor.

A major limitation of traditional statistics is that one may not use every variable of interest as this reduces statistical power to such an extent that results are not believable. To address this, we compress the hundreds of variables and lags into a few dozen composite variables called components (or factors). This can be accomplished using factor analysis or the more straightword principal components analysis. In an earlier cited work, the results needed to be conveyed to European Commission policy makers, landowners, and farmers and other stakeholders. To maximize the ability of our clients to understand the mathematics underlying this approach we relied on the older and simpler, but easier to explain principal components analysis. This approach essentially looks at all the variables and tries to extract out the unique variation associated with each variable and discards the common sources of variation (also eliminating co-linearity and to some extent endogeneity). Variables may also be grouped or clustered into components by similarity if needed (i.e., rotated). By using this method, we can maintain statistical power while also retaining as much of the unique variation in the variables as is possible.

Due to the fact that our dependent variable is dichotomous, (i.e., binary: the firm expanded or not), we need to use a regression approach designed for binary variables. Again, a number of approaches are available: Logit and Probit would be the most common. For this application, we employ a Probit approach using the previously identified principal components. The final steps involve isolating driver importance and making statistics-based GScore™ predictions. Due to the simple linear relationship among expansions and predictor variables posited when using principal components analysis and Probit analysis, one can recover individual variable importances. With the weights/factor loadings from the principal components analysis and parameters from the Probit analysis, one can simply feed the predictor variable values for new firms (without known expansions) into the linear representation to derive a predicted probability of expansion. After performing this algebraic operation, one obtains distributions of GScores™ for various industrial groupings such as manufacturing, professional services, technical, and education, etc. GScores™ are then assigned by parsing the distribution space based on calibrated program control parameters. These parameters indicate how far above or below the mean a firm must lie to be classified as a call or no call (due to low predicted expansion probability), and for GScores™ of 1 to 5, based on how far up or down in the tails of the distribution a particular firm lies. This last step is common to both statistical and AI approaches, and limitations of this approach are well known and described in the cited works.

B. Machine Learning AI method

The way the Gazelle.ai machine learning approach works is another point of interest. We use an analysis dataset to obtain predictor variable parameters association weights and learning rates that we use to make propensity of expansion predictions. These predictions can be transformed into GScores™. This approach uses similar transformed/normalized predictor data, although we have found that a custom designed and calibrated hyper-tan function works best for this method. The algorithms are used to train predictor variable parameters as input values change. Each predictor variable (e.g., venture capital) may have a unique set of association, learning, and neuroplasticity rates, or may have common learning and neuroplasticity rates but still unique association weights. Association weights tell us the nature of the relationship among a variable and firm expansions. Learning rates dictate how the association weights change given new information about firm characteristics and expansion outcomes across time. Also, neuroplasticity governs the memory horizon for the algorithm or the extent to which past information is discounted or over-weighted. We harness both cross-sectional and time-series variations in firms' predictor variable values for training due to the vastly larger number of firms relative to the time series information.

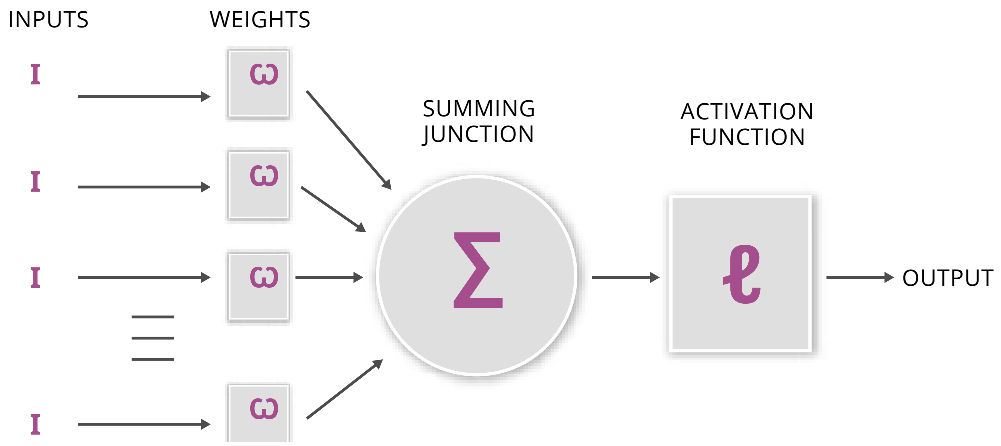
The three main steps required for the simplest instance of this method, called a 1-layer feed forward NN, are:

1. initializing association weights in various ways
2. estimating/training association, learning and neuroplasticity rates
3. linking predictor value changes and expansion outcomes by minimizing the deviation among the analysis dataset predicted and observed expansions.

The fitting of the models is performed with an out-of-sample prediction requirement. In other words, while fitting the model, only predictor information from before the expansion date (e.g., 2008-2015 variables to predict 2016 expansions) is used. The next step entails conducting a second 'out-of-sample' validation test on a validation dataset (a dataset scrolled one year forward of the analysis dataset, e.g., using 2009-2016 information to predict 2017) and using weights obtained from the analysis dataset. This validation process involves multiple model comparison and generalization tests using comparisons to statistical methods, variations in the mathematical structure of the algorithms, and variations in various other program control parameters (i.e., number of lags included, industrial sector subgroupings, how far above the mean before designated an expanding firm, etc.). This process continues until accuracy reaches a predetermined target threshold. Last but not least, we apply the analysis dataset prediction variable parameters to the prediction dataset, or the entire 6 million (and growing) firm database.

A graphical representation can often be helpful for understanding the structure of an ML algorithm. A typical visual representation authors have used over the years assumes our simplest ML algorithm may be broken down into three parts: input connections, summing and activation functions, and output connections.

ARTIFICIAL NEURON MODEL



Starting from the left, the input (I) are the various normalized predictor variables or principal components representing characteristics of a firm. These are interacted (with arbitrary mathematical formulae) with the machine learning association weights (ω). If deep learning were represented in this figure, the various I's would be multiplied (or otherwise interacted together using some function) in the space between the I's and ω's. Next, the individual variables and association weight pairs are multiplied and summed together with another arbitrarily structured summation function represented by sigma (Σ). This summing function output is then transformed into an output activation (i.e., propensity to expand) which represents the prediction. This output can then be compared to what actually happened since we are applying this in the analysis dataset, and we can establish accuracy as well as a teaching signal (i.e., how good the prediction was). The teaching signal and how it propagates back to the weights updating process and hidden node determination are not represented above. However, the back-propagation feedback takes this teaching signal and compares it to the level or change in the prediction variables and adjusts the association weights and selected hidden nodes based on the learning and neuroplasticity rates. This process is repeated until learning and neuroplasticity rates converge to maximize accuracy. Then all these parameters are output for use in the validation dataset and prediction dataset where the output activation function combines the most recent lagged predictor values to obtain expansion predictions.

VI. SOME GAZELLE.AI RESULTS FOR 2017

So what did the platform tell us about 2017 gazelle companies from the perspective of 2016 and 2017 firm information? Firms in industries with the highest GScores™ are the ones predicted to have a high propensity to expand, but not necessarily to your area.

In 2017, the manufacturing sector high-growth areas included oil and natural gas extraction, heavy and civil engineering and construction, specialty trade contractors, computer and electrical component manufacturing, and electrical appliance and component manufacturing. If these are the main targets, an EDO should strive to reach out to the fewest number of companies within these industries to organize face-to-face meetings and attract them to their region.

In the services sector, growth areas included: wholesale; information processing, publishing, data processing and hosting, finance and insurance; management of companies (including holding companies); professional, scientific, and technical devices; administrative support services; and waste management and remediation.

Gazelle.ai also had a lot to say about the expansion drivers for 2017. Variables common to both manufacturing and services included: the amount of awarded expansion venture capital, liabilities and debt relative to firm value, and inventory turnover. Drivers particular to manufacturing included: patent activity, magnitude of particular forms of export activity, the degree of concentration of a particular industry, return on equity if public, sales relative to assets, assets relative to liabilities and inventories, availability of credit/lending, and interest expenses. Finally, drivers particular to services included: employment levels, level of benefits, magnitude of export activity of an alternative type, number of establishments in an industry, royalties received; characteristics of accounts receivables, and magnitude of executive compensation.

Signs of the effects of these variables often change across lags. For example, recent past employment or employment growth might have a positive effect, but more distant past employment could have a negative effect. Further, there may be quadratic or cubic effects, as well as important interaction terms (i.e., hidden nodes). The relative magnitudes of the variables change from year to year, indicating that relying too heavily on past results may not necessarily be to a development agency's best interests (i.e., you need to stay up-to-date if you want to keep efficiently organizing meetings and attracting firms).

Gazelle companies may have little interest in expanding to your area depending upon your region's main assets. They may already have an existing branch, your strengths may not suit their needs, your distance from their primary markets may not be optimal, etc. This means that targeting firms with GScores™ ranging from 2 to 5, and not just the 5s, is a smart approach (GScore™ rating is on a scale from 1-5, with 5 being the strongest rating). Keep in mind companies with 2 or 3 scores could still be in expansion mode, and may be more willing to consider a variety of areas.

VII. CONCLUSION

The future of Gazelle.ai as the tool of choice for Economic Developers and anyone who needs to identify fast-growing companies looks bright. We are continuously improving our algorithms - training and retraining alternative structures to find the most relevant approaches for each year. The Gazelle.ai team is also constantly working to expand our data sources and further filter and clean existing data. As more data is acquired and cleaned, we can investigate deeper learning to identify the important interactions among variables (and local regions characteristics) and improve accuracy. This also involves considering more existing learning functions, as well as developing more of our own custom structures. Our clients and internal researchers have integrated and embedded Gazelle.ai into their research processes. The ongoing feedback we get from them is also helping us continually improve the platform.

Stay tuned for exciting new developments on our new site at www. gazelle.ai.

About the author

DR. HUGH KELLEY. PH.D.

Hugh received his Ph.D. in International Economics from the University of California. Hugh has been a professor in the department of Accounting, Finance, and Economics at the Oxford Brookes University, U.K. Hugh's research has been highly inter-disciplinary, spanning from applied and theoretical contexts, and reflects an integrated economic, psychological, and natural sciences methodology. The economic elements of his work involve multiple forms of modeling including CGE and statistical modeling, microsimulation, and frontier agent-based methods. His forecasting experience includes traditional univariate time series and multivariate panel approaches, and behavioral approaches. Hugh has an extensive history as a pioneer in applying AI to various fields within economics.

As Gazelle.ai Chief Economist, Dr. Kelley is responsible for the Data Creation, Forecasting, and Analysis of the platform. He manages our team of economists and data scientists who are responsible for developing and training our AI algorithms.

www.gazelle.ai
Tel:1-866-625-5136
Address: 416 de Maisonneuve Blvd. W. Suite 1000,
Montreal, Quebec, H3A 1L2 Canada.