A Paradigm for Assessing the Scope and Performance of Predictive Analytics

Jeffrey T. Prince*

February 2018

PRELIMINARY AND INCOMPLETE

PLEASE DO NOT CITE

Abstract

In this paper, I outline possibilities and limitations for the scope and performance of predictive analytics. I do this by first bifurcating predictive analytics into two categories, passive and active. I contrast this categorization with current alternatives and highlight its relative merits in terms of clarity in boundaries, as well as appropriate methods for different types of prediction. I then describe the range of suitable applications, as well as the possibilities and limitations with regard to prediction accuracy, for each type of prediction. I conclude with a discussion of some primary concerns when prediction limitations are not heeded.

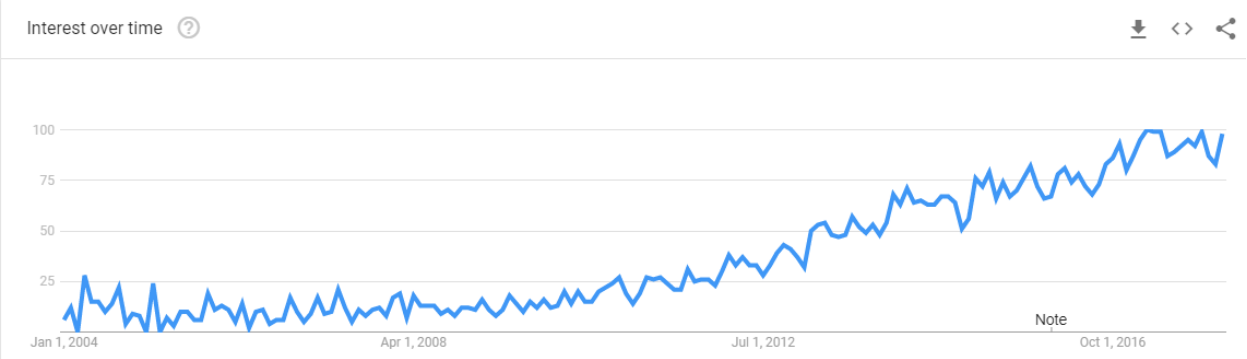
* Indiana University, Department of Business Economics and Public Policy, Kelley School of Business (jeffprin@indiana.edu). I thank Scott Wallsten and the Technology Policy Institute for their support. I am responsible for all errors.

1. Introduction

Predictive analytics is a growing field, with widening influence. As evidence, one need only observe the trend in interest over time, as measured by Google Trends (Figure 1).

Figure 1

Trend in Interest for Predictive Analytics from 2004 to 2017



With such a growth pattern, it is natural to ask where the field is headed, or how far it can go. Of course, there are many ways to interpret such questions; here, we focus on assessment in terms of scope and performance. However, to properly make such assessments, we first must have a clear understanding of, and agreement on, what prediction is.

Many existing characterizations and classification schemes concerning prediction place a narrow definition on prediction, which can lead to notable challenges. For example, suppose we asked how a firm's sales would change next month with a ten percent increase in its price. Answering such a question clearly requires a prediction – we must predict what will happen to sales with such a price change in a future period. However, this type of prediction is not treated as such by some classifications. This can lead to confusion, and worse, misapplication of predictive models.

In this paper, we present a paradigm for classifying prediction, where we bifurcate it in a simple, clearly defined, way. The bifurcation we propose is active and passive prediction, where roughly speaking the key difference is whether the prediction involves intervention. We then detail what types of applications are suitable for each type of prediction (scope) and how we measure the accuracy of each (performance). With this backdrop, we explore possibilities and limitations for the scope and performance of each type of prediction. We conclude with a discussion of consequences when predictive models are mis-applied and/or misunderstood.

2. Prediction

2.1. Prediction Defined

We begin our prediction section by simply defining what prediction is. As we are interested in the use of data and analytical methods toward making prediction, the definition we provide concerns predictive analytics (we leave for elsewhere other types of prediction, such as supernatural prediction). We define predictive analytics as any use of data analysis designed to form predictions about future, or unknown, events or outcomes. The typical understanding of predictive analytics defined this way is that there is some event or outcome (generically the dependent variable) whose realization we don't know but would like to predict using available data (generically predictor variables). In the context of machine learning (and hence within the general umbrella of artificial intelligence), this concept of prediction aligns with what is known as supervised learning, where the analyst picks the event or outcome to be predicted (in his/her role as supervisor), and the machine learns to predict it.

Although seldom emphasized or clarified, the definition of predictive analytics implies prediction need not always concern the future. Rather, prediction can also extend into the present and past, where we use data and analysis to make a best guess as to the realization of an unknown outcome (e.g., "Is that a person's face in a given picture?", or "Was there fraudulent record keeping last month?", respectively). We can call this subset of prediction, diagnostics. As we will discuss, the methods used to make predictions in the form of diagnostics heavily overlap with those used for many predictions about the future; however, making this distinction can be useful when ascertaining the limits and possibilities of prediction in general.

2.2. Prediction Bifurcated

There are many ways to categorize prediction, e.g., across different types of methods, different types of outcomes, etc. However, there is one simple categorization scheme – a bifurcation – that helps to draw a clear line between different types of prediction approaches and applications. Such a clear line can be particularly helpful given the rampant examples of predictions that either actually (through their execution) or via their public interpretation (e.g., media summaries) inappropriately cross the line.

Before describing our bifurcation for prediction, it is useful to review the definition of a data-generating process. We define the data-generating process as the underlying mechanism that produces the pieces of information contained in a dataset. As will soon be clear, the data-generating process plays a pivotal role in the prediction bifurcation we describe. As a simple example, suppose we wish to predict sales of the iPhone X. We may (roughly) characterize the data-generating process as Sales = f(Price, Features, Comp Prices, Comp Features, Income) + ε.

Here, we've specified that iPhone X sales depend on its price and features, the prices and features of competing smartphones, income (perhaps deflated measures of income at each decile), and "other factors" (ε). Specified this way, we are saying that, given own price and features, competitors' prices and features, and relevant income information, we can feed these into our function f(.), and the output of this function added to the conglomeration of other factors (represented by ε) generates the sales of the iPhone X.

The bifurcation for prediction that we consider is active vs. passive prediction. Passive prediction is the use of predictive analytics to make predictions based on actual and/or hypothetical data for which no variables are exogenously altered. Here, we define a variable to be exogenously altered if it changes due to factors outside the data-generating process that are independent of all other variables within the data-generating process. In more practical terms, passive prediction concerns predictions where there is no intervention in the data-generating process; hence, the name passive – we passively observe and predict.

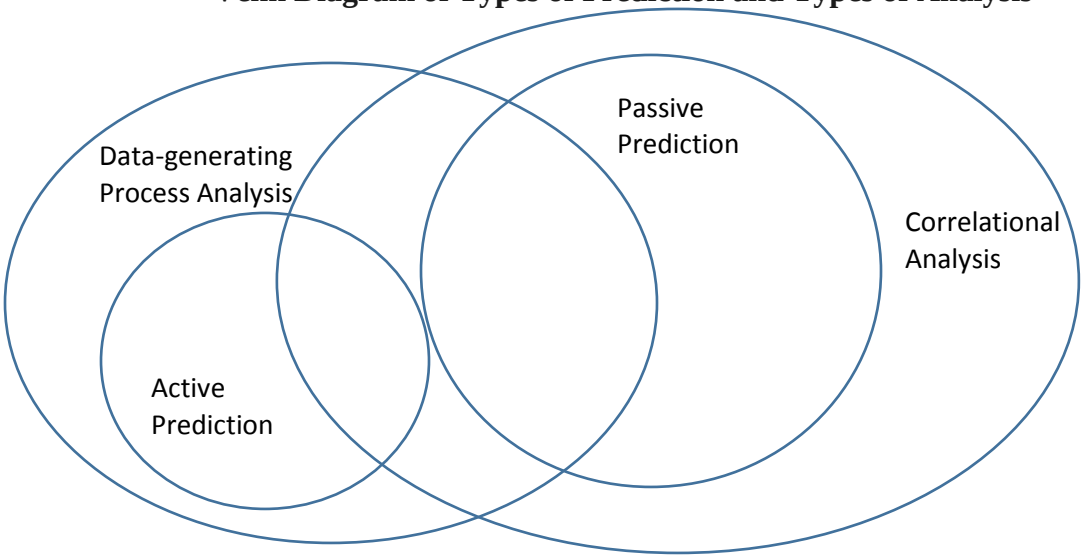
Active prediction is the use of predictive analytics to make predictions based on actual and/or hypothetical data for which one or more variables experience an exogenous alteration. It is the complement to passive prediction, as now it involves predictions where there is intervention in the data-generating process. Hence, the name active – we act (on the data-generating process) and predict.

In the next two sections, we go into significantly greater detail on both types of prediction. However, we conclude this subsection by placing them in the context of the closely related general concept of correlation vs. causality. First, note that active predictions require an estimation of a data-generating process, which models causal impacts of predictors on a dependent variable. Hence, active prediction is within the subset of applications to which causal

analysis can be applied. Second, correlation can be sufficient for making passive predictions. That is, we can make satisfactory passive predictions using only variable co-movement, without specifying or estimating a data-generating process. Lastly, passive predictions can also be made using an estimated data-generating process. As we discuss further in Section 3, it is sometimes the case that using a model designed for causality and so appropriate for measuring effects of actions, is also the optimal choice for making predictions when action/intervention is not anticipated. Figure 2 graphically illustrates these distinctions.

Figure 2

Venn Diagram of Types of Prediction and Types of Analysis[1]



[1] Figure 2 suggests seven categories of analysis. We list each here (with an example in parentheses): 1. Data-generating process analysis only (theoretical model, without accompanying data analysis, that explains, e.g., cross-sectional differences); 2. Data-generating process analysis & Active prediction only (theoretical model, without accompanying data analysis, that makes predictions with intervention); 3. Data-generating process & Correlational analysis only (theoretical model estimated using data and correlational model (e.g., regression) that explains cross-sectional differences); 4. Data-generating process & Active prediction & Correlational analysis (theoretical model estimated using data and correlational model that makes predictions with intervention); 5. Correlational analysis only (correlational model estimated using data but without a theoretical model, and used to explain cross-sectional differences); 6. Correlational analysis & Passive prediction only (correlational model estimated using data but without a theoretical model, and used to make predictions without intervention); 7. Correlational analysis & Passive prediction & Data-generating process analysis (theoretical model estimated using data and a correlational model, used to make predictions without intervention).

As Figure 2 illustrates, active/passive prediction and correlation/causality do not fully coincide. As we highlight in the next subsection, while they are similar concepts, the varying notions of prediction in the current lexicon alone warrant added clarity. Additional clarity is also important due to the rampant mis-application of predictive analytics, as we detail below.

2.3. Alternative Characterizations of Prediction

Before taking a closer look at passive and active prediction, it is useful to contrast our categorization with some relevant alternatives.

A popular characterization of business analytics builds three levels: descriptive analytics, predictive analytics, and prescriptive analytics (HaloBI). Within this paradigm and put succinctly: descriptive analytics is typically backward looking, telling us about what happened; predictive analytics assesses what is likely to happen; prescriptive analytics delves into why a prediction will occur and prescribes decision options that can best take advantage of opportunities and/or mitigate risks. Defined this way, predictive analytics only contains passive prediction, while active prediction would be a part of prescriptive analytics.

This tri-level breakdown of business analytics can be useful, but it invites at least two points of confusion. First, it creates confusion regarding the concept of prediction. A natural question in business is something like: What will happen to sales if I raise my price 10% next month? Answering this question requires us to make a prediction (an active one), but within the tri-level paradigm, answers to such questions appear relegated to a subset of prescriptive analytics. Active prediction is prediction, and so separating it from the category of predictive

analytics can generate confusion, per se. Further, if we classify passive prediction as the only type of prediction, it can suggest passive prediction encompasses all prediction, inviting its use for questions that are active in nature, as with the above question. Examples abound of such misuse of passive prediction, as we highlight below. Second, it suggests that establishment of "why" variables move the way they do lives outside prediction (and instead in prescription). However, much of prediction involves measuring and utilizing a data-generating process, which often is constructed based on why variables co-move. This isn't just true for active prediction; it can hold for passive prediction as well. For example, weather prediction at least partially relies on physics equations describing how and why weather variables move together.

Consider now a well-known characterization of statistical learning, which concerns estimation of a function that relates input variables to an output variable. The standard characterization is:

$$Y = f(X) + \varepsilon$$

Here, Y is the output variable, X represents a set of input variables, and ε is an "error term." In trying to estimate f, as noted in James et al. (2015), we may be interested in "prediction" or "inference." Similar to the tri-level characterization of business analytics, prediction in this context only includes passive prediction, and again leaves out the possibility of passive prediction that utilizes a data-generating process. It instead treats f(.) in the context of prediction as constituting a "black box." In contrast, inference concerns the way the output is affected as the inputs change. Described this way, inference clearly requires us to measure the data-generating process. Further, by doing so, we can address a range of questions, including includes questions surrounding active prediction. This flavor of classification is not unique to James et al. (2015); Schmueli (2010) has a very similar classification, using the terms predictive

modeling and explanatory modeling roughly analogously to prediction and inference, respectively.

The consequent points of confusion for this second characterization are similar to the first. Prediction only includes passive prediction, again inviting its use for questions that are active in nature. Prediction again also does not account for the possibility that using the data-generating process can be helpful even for these types of predictions. Further, the use of inference explicitly allows for the possibility of making predictions for the output, making it particularly confusing that such predictions are in a category other than the one titled "prediction."

This division of statistical learning, like the aforementioned characterization of business analytics, can be useful toward our general understanding of analytics on the whole. However, it is in their treatment and clarity regarding prediction that our active/passive classification stands to provide useful distinctions.

3. Passive Prediction

3.1. Overview

As noted above, passive prediction concerns predictions where there is no intervention in the data-generating process, i.e., we passively observe and predict. Importantly, passive prediction does not require the estimation of a data-generating process, but does not preclude it. For passive prediction, we are essentially leveraging our best understanding of how variables move together, and this understanding may or may not involve knowledge of *why* such co-movement occurs.

For the remainder of this section, we discuss applications and accuracy for passive prediction. For the former, we give current examples of where it is applied and discuss possibilities and limitations for additional applications. For the latter, we discuss ways the accuracy of passive prediction is measured and with this context, follow with possibilities and limitations in terms of how accurate passive predictions can become.

3.2. Applications

3.2.1. Examples

Passive prediction has a wide range of applications. Rather than attempt to provide an exhaustive, detailed list, we highlight a couple examples and describe them at a very high level. We subsequently emphasize their key components, and what makes them passive in nature.

The first application we highlight is weather prediction. In general, weather forecasts use current weather conditions along with mathematical models to predict the weather for several days into the future. The models divide the relevant area into a 3-D grid, using as inputs information on measurable weather features such as winds and relative humidity. Using these inputs, they solve systems of differential equations that yield weather predictions.

Weather prediction has a couple key features. First, we don't intervene in the weather, generally speaking. Rather, we passively observe current weather conditions and use our models to predict what they will be in the future. Consequently, the predictions we see in our seven-day forecasts are passive predictions. Climate is a different story, as we note in the next subsection. Second, weather prediction at least partially incorporates a data-generating process. The differential equations being solved as part of the process incorporate laws of physics (e.g., the ideal gas law), which describe data-generating processes. Hence, weather forecasting is a case

where we are making passive predictions using models that are at least partially built using our understanding of the data-generating process.

The second application we highlight is prediction of customer churn. A major concern for many, particularly subscription-based, firms is customer non-renewal. The first step toward stemming these customer losses is trying to predict which customers are most likely to leave. To make such predictions, we can employ a range of models (regression, logistic, random forests) to tell us, based on, e.g., customer demographics and past behavior, which customers are of greatest concern.

Predicting customer churn also has a couple key features. First, as described, we again are not intervening in the customers' decision processes. Rather, we observe who they are and what they've been doing, and passively predict what they are going to do next. Consequently, such churn predictions again fall squarely into passive prediction. Second, if these predictions are our end goal, there is no need per se to try to model a data-generating process. Unlike weather prediction, the underlying mechanisms driving customer churn are not as well established. Therefore, it is not clear that attempts to model the data-generating process will be helpful toward making these predictions. Instead, we can forego such attempts and seek models that yield the best predictions, regardless of whether they accurately reflect the underlying data-generating process.

3.2.2. Possibilities and Limitations

In terms of applications, the possibilities for passive prediction are essentially endless. Any variable whose future value is uncertain, or current value is unknown, and where we have

collectible information that may help resolve some of that uncertainty, is a potential candidate for passive prediction. The data collection and models do not have to tease out the role of intervention, as all variables are being passively observed. Of course, how good these predictions can become is another matter (discussed in the next subsection).

Unfortunately, despite the many ways that passive prediction can be viably applied, it is highly common for it to be misapplied as well. A typical misapplication has the following form – notable variable co-movement is identified, and this co-movement is then used to make active predictions concerning some form of intervention (a policy change, strategy shift). Examples abound, and while it is often difficult to cite many in a business setting (businesses are not eager to share the specifics of how analytics translates into strategy shifts), it is easy to find examples from other fields (e.g., health) that are more focused on public dissemination. Below are just a few examples (listed in the form of press headlines), where the appropriateness of making active predictions based on the analysis done is suspect at best:

"Study: Marijuana Use Increases Risk of Academic Problems"

"A New Study Says Living Near a Pub Makes you Happier"

"Drinking More Coffee May Reverse Liver Damage from Booze"

"A Study Links Soda Consumption to Heart Failure"

Revisiting the applications we cited for passive prediction above, both invite active predictions in addition to the passive ones we posed (predict weather conditions over the next few days, predict who is most likely to churn). For weather, if we move to considering our climate (i.e., weather conditions over longer periods of time), we may be interested in predicting how changes in human behavior impact the climate. For churn, our passive prediction may be

just the first step toward intervention via a business strategy, and once we start considering intervention, we need to be considering active predictions.

We conclude this subsection by noting the substantial role passive prediction plays in diagnostics. For example, we may want to use patient symptoms to predict (diagnose) the presence of a disease or the status of being pregnant. Other examples include recommendation systems (e.g., diagnose tastes based on preference information) and fraud detection (diagnose a fraud has taken place based on observed distribution of financial data). These predictions concern unknown, rather than future, outcomes. However, many – including the above examples – are passive in nature. That is, when making the diagnosis, the analyst is not intervening in the data-generating process that determines the outcome. This need not always be the case, as we discuss in Section 4.

3.3. Accuracy

3.3.1. How It's Measured

For passive prediction, accuracy centers on the idea of "fit." Here, fit is typically assessed by looking at how close model predictions are to realized outcomes on "new" data. For example, we may estimate a random forest based on observed individual churn outcomes, and then use the estimated forest to make predictions for another set of customers. The rate of correct predictions is a measure of fit, which we can use to compare the suitability of competing models; the model that demonstrates the best fit (percent of correct predictions) is preferred.

For non-binary outcomes, we can use measures such as R-squared, or other statistics that are, e.g., functions of the difference between predicted and realized outcomes, as measures of fit. Again, the model that demonstrates the best fit (e.g., highest R-squared) is preferred.

3.3.2. Possibilities and Limitations

Is perfect prediction possible? Put another way, is it reasonable to aspire toward getting it right every time? As is often the case, the answer is that it depends. In some situations, perfect prediction is in fact possible; in others it isn't. We highlight some key determinants of the "ceiling" for predictive accuracy for passive prediction.

We start with an extreme example. Consider dropping a ball in a vacuum at a given distance above sea level. With little more than the aforementioned information, one could predict where that ball will be, exactly, with essentially 100% accuracy. We can do this because we essentially know the data-generating process, and it is deterministic. That is, there is essentially no randomness in the process that determines the future location of the ball.

In many other instances, e.g., movement in gas molecules or stock prices, we likely have a stochastic process. In contrast to deterministic processes, stochastic processes have some inherent uncertainty/randomness. For a process we characterize as stochastic, how good can our predictions become? To answer this, we need to assess whether the randomness in the process can be reduced via additional information or it is inherent in any representation of the data-generating process. Consider again weather prediction. Many models treat the determination of weather conditions as a stochastic process; however, one could conceive that weather conditions

are actually deterministic, and just depend on more variables than we can possibly track. Consequently, predictions, at least in theory, could become perfect for the weather (and other deterministic processes) if we had all the relevant information, including knowledge of, and ability to solve for, the full data-generating process.

Unfortunately, although theoretically possible, achieving perfect prediction for many deterministic processes isn't possible in practice, as noted by chaos theory. In short, we typically cannot perfectly measure (with infinite precision) all aspects of current weather conditions. Chaos theory tells us that such imperfect measurements – even with extremely small error – can amplify relatively quickly (the "butterfly effect"). For this reason, it is commonly believed the ceiling on accurate weather prediction is a mere fourteen days.

So how good can prediction get? If we are aspiring toward perfection, we must consider whether we can perfectly describe the data-generating process, and whether that process is deterministic or stochastic. Then, even if it's deterministic, challenges from chaos theory limit our predictions. However, a caveat concerns diagnostics; for diagnostics, we are not looking to the future, so the limitations stemming from chaos theory will not apply.

Many predictions concern human behavior, and unlike weather, we may be less certain that the underlying data-generating process is deterministic in nature. For example, is free will an irreducible error, meaning variables such as churn and stock prices are based on stochastic processes? If so, the ceiling on accuracy is lower than perfection.

Suppose instead that we forego the idea of achieving perfection, and just want to make predictions that are as accurate as possible absent the ability to fully specify the data-generating

process. Prediction for human behavior has limits not only because there is likely a stochastic underlying process but also because characterizing that process is exceedingly complex. When the data-generating process is highly complex, every model is mis-specified, so some that do not resemble at all the data-generating process may do better than others that do resemble it, as the latter may not do so sufficiently. In such cases, it is difficult to assess how good the predictions can get, but nonetheless they leave those deciding whether to act on such predictions to consider how much they are willing to bet on them despite having little understanding of the "why" behind the predictions.

A final consideration centers on predictions for the future (as opposed to diagnostics). Whatever model we are using – be it a data-generating process or "black box" – we must ask whether, and for how long, its predictive capabilities last into the future. For predictions using laws of physics, we can reasonably expect the models to work well into the future. However, other data-generating processes concerning human behavior may be less stable into the future, leaving us to ask whether there is an evolution and if so, can we model it? In contrast, it can be more difficult to assess the range, over time, of predictive capability for "black box" methods. We can continually test them on new datasets to determine if they are still producing good predictions, but given our lack of knowledge about the mechanisms behind why they are predicting well, we will be limited in our ability to assess how long into the future they will continue to predict well.

4. Active Prediction

4.1. Overview

As noted above, active prediction is the use of predictive analytics to make predictions based on actual and/or hypothetical data for which one or more variables experience an exogenous alteration. Active prediction does require us to estimate a data-generating process – or some approximation of it. The essence of intervention is to change a variable (treatment) while "holding other factors fixed." A data-generating process allows us to calculate such partial effects on the dependent variable (outcome); whereas models for fit are not suitable for partialing out the effect of one variable versus the effects of others.

As in Section 3, for the remainder of this section, we discuss applications and accuracy for active prediction. For the former, we give current examples of where it is applied and discuss possibilities and limitations for additional applications. To emphasize the key differences between active and passive prediction, we provide two new examples, but both again regarding the weather and customer churn. Regarding accuracy, we discuss ways the accuracy of active prediction is measured and follow with possibilities and limitations in terms of how accurate active predictions can become.

4.2. Applications

4.2.1. Examples

Active prediction, like passive prediction, has a wide range of applications. To help emphasize the difference between active and passive prediction, we describe applications again related to weather and customer churn, but involving active predictions. We subsequently emphasize their key components, and what makes them active in nature.

Revisiting weather, we now consider the question of climate. Of course, a major question for many decades surrounding climate concerns the impact of carbon dioxide in the atmosphere on global temperatures. Even if we take as granted that there is an impact, quantifying the impact is clearly a challenge. However, doing so is crucial toward making relevant active predictions. In particular, policies concerning carbon dioxide emissions are motivated by the idea that a reduction in carbon dioxide will have a desired impact (less warming / climate change) on the climate. Any assessment of the expected value of such policies must incorporate predictions about the change in climate with an exogenous (active) change in carbon dioxide levels.

Predicting climate response to carbon dioxide changes has a couple key features. First, we are considering interventions in the climate's data-generating process. While we could just passively observe climate, as we do with weather, policies concerning carbon dioxide emissions center on the idea of intervention. Second, predicting climate change requires us to think carefully about a data-generating process for the climate. The further the accessible data is from possessing exogenous variation in the predictor (treatment), the greater the demands on us to build a thorough representation of the data-generating process. This is exactly the challenge for climate prediction – we generally do not observe exogenous changes in carbon dioxide in the atmosphere, and so climate modelers have to have a quite thorough representation of climate processes to properly predict how exogenous changes in carbon dioxide will impact the climate. If they have a deep understanding of underlying physical processes driving the climate, this can help in constructing a data-generating process.

Revisiting customer churn, we now consider the possibility of crafting a business strategy designed to impact (reduce) customer churn. Instead of trying to predict which customers are

most likely to leave, we may want to know how an action designed to prevent churn, say, subscriber benefits that accrue over time as a customer, impacts the actual churn rate. Any assessment of the expected return on such strategies must incorporate predictions about the change in customer churn with an exogenous (active) change in subscriber benefits.

Predicting customer churn's response to a strategy shift also has a couple key features. First, as with climate, we are considering interventions in churn's data-generating process. In Section 3, we described the idea of passively observing and predicting churn, but now we are thinking about ways to actually curb it. Second, predicting churn's movement with a change in subscriber benefits requires us to think carefully about a data-generating process for churn. In particular, we have to think of ways of modeling the data-generating process such that the data we are able to acquire can adequately tell us about how exogenous variation in subscriber benefits will impact churn rates. Again, if we have a deep understanding of the underlying economic/psychological processes behind customer churn, this can aid us in constructing a data-generating process.

We conclude this subsection by noting the connection between active prediction and judgment, as defined in Agarwal et al. (2017). They define judgment as understanding the impact different actions will have on outcomes in light of predictions. When considering different actions in light of predictions, it is important then to distinguish whether the predictions are dependent or independent of the actions. For example, consider again our customer churn example. As a manager, one may be considering actions that are comprised of business strategies, which could impact the rate of churn. Thus, judgment regarding which action to take requires consideration of active predictions about the impact of different business strategies. In contrast, as an investor, one may be considering actions that are comprised of different levels of

stock purchase for the firm in question. Here, judgment regarding which action to take requires consideration of passive predictions about the firm's churn rates into the future (the investor's stock purchase, per se, is unlikely to impact customer churn rates for the firm).

4.2.2. Possibilities and Limitations

In terms of applications, there are a great deal of important questions necessitating the use of active prediction. Questions concerning policy actions and business strategy alone are almost innumerable, and many of these require the use of active prediction. Questions about health behaviors (e.g., diet, exercise) also tend to demand active predictions.

The key limitation on active predictions is our ability to attain exogenous variation for the treatment in question. There are many ways an analyst can try to extract, and/or generate, needed exogenous variation in the treatment variable. For example, firms (particularly those in digital markets) attempt to run their own experiments, using the fundamental idea of random treatment assignment as a means to determine the effect of a strategic variable (e.g., digital ad placement) on an outcome they care about (e.g., click-through-rate). Other ways of attaining exogenous variation include: the use of natural experiments (e.g., policy shifts that transpired for reasons not related to the outcome), the use of instrumental variables (variables that move with the treatment but not the outcome), matching estimators (trying to approximate experiments), difference-in-differences, and highly detailed data-generating processes (attempting to control/account for other factors that move with the treatment). In addition, when using an observed change in a treatment over time to measure its effect, as Varian (2014) notes, we can utilize passive prediction to predict what would have happened without the change to get even

more accurate measurements of the treatment effect. However, it still is the case that we need the observed change in treatment to be exogenous to get meaningful estimates.

While there are many ways to attempt to attain exogenous variation in a treatment, it is important to recognize that such variation may not be feasible. This will typically be the case when many treatments are administered at once, and there is no realistic way of disentangling one from the other. This is the essence of an identification problem (discussed further below) – the effect of one treatment cannot be identified separately from the effect of another/others. It is situations where existing data have such identification problems that often motivate attempts at experimentation – since exogenous variation does not exist, we manufacture a situation where it does.

As with Section 3.2.2, we conclude this subsection by noting the role active prediction plays in diagnostics. When it comes to diagnosing a disease or catching fraud, the role of active intervention is not apparent – we aren't interested in actively changing, or even can't change, the symptoms or distribution of financial data, respectively. However, if the outcome is, say, perception of an image (rather than the real object the image is capturing), we can consider a role for active prediction. For example, suppose the outcome is whether an image is perceived to be a bird. We could then consider exogenous changes in features of the image (e.g., coloring, presence of triangular shapes), and try to predict how such changes affect the likelihood that the image is perceived to be that of a bird. Such an exercise is again diagnostic, as we are not predicting a future outcome, but instead predicting the unknown (a person's perception of an image).

4.3. Accuracy

4.3.1. How It's Measured

For active prediction, accuracy centers on the idea of "identification." Here, identification concerns whether or not we are able to get measures for the parameters of the data-generating process that's been specified. Note the stark difference from the goal of fit for passive prediction. We may have a great fit for the data, but if the model estimates are not telling us about a data-generating process, they are not particularly helpful toward making active predictions. In contrast, we may have what seems to be a weak fit for the data (e.g., low R-squared), but the estimates may be informing us about the data-generating process. For example, the variable whose effect we'd like to measure may do relatively little to explain the overall variation in the outcome, but if it's identified, we could make accurate predictions about how changes in that variable impact the outcome.

To illustrate this idea, let's return to our customer churn example. First, suppose we simply wanted to predict future churn, without implementing any strategy. In this case, we want a model that fits the data as closely as possible, so our predictions will be as close to what's realized as possible, as this is our primary aim. Now, suppose we want to predict how churn will change with the implementation of a strategy of added subscriber benefits. Here, we need to get a measure of impact – how the churn rate will change with a change in benefits. For this question, identification is the primary concern; even a model with relatively weak fit will be effective if it accurately tells us about this impact.

4.3.2. Possibilities and Limitations

As we've noted above, active predictions typically center on predicting how an outcome will change with a change in a treatment, typically measured as an average treatment effect. The accuracy of such a measure will depend on how close to the population our samples can get and how much exogenous variation in the treatment we can find – the more of each, the better in terms of accuracy.

However, the improved accuracy described above is for real-time predictions, i.e., what will happen with an immediate change and/or diagnostics questions. If we start asking about changes/actions that will happen in the future, we then must concern ourselves with the stability/evolution of the treatment effect we've measured; is it fixed into the future, or is it evolving? If the latter, can we predict its movement – do we need to model and estimate a data-generating process for the average treatment effect? For example, the weather won't alter its behavior based on the fact that we are forming predictions about it; however, people may. In such a case, we likely must account for how treatment effects change over time as a result of predictive activity. Further, even if we have a complete model of how treatment effects change over time, chaos theory may come back into play for predictions into the relatively near future, as we may not be able to get perfect measures of the conditions determining the treatment effect at a given point in time.

5. Discussion and Conclusions

In this paper, we provided a paradigm for predictive analytics, breaking it up into two major categories – passive and active prediction. We contrasted this paradigm with existing alternatives, and then using our paradigm, described appropriate scope and measures of

performance for each type of prediction, concluding with a discussion of possibilities and limitations for each.

Understanding predictive analytics within our paradigm can be valuable for several reasons. First, it can help avoid misapplication of predictive analysis. By casting a wider net in terms of what is a prediction while also clearly characterizing distinctions in different types of predictions, we can potentially reduce or avoid the many instances where active predictions are made using passive methods. Avoiding such misapplications can have many benefits. For example, improper analysis greatly raises the risk of multiple analyses arriving at contradictory findings. When such contradictions occur, they can erode confidence in data-driven analyses, not because they aren't useful, but because they were misapplied. Contradictory findings also invite the practice by some analytics consumers of starting with a preferred conclusion and using whatever analysis supports that conclusion as proof (while ignoring the contradicting evidence). Of course, contradictory findings are bound to occur when conducting a wide range of analyses, but let them be due to the complexity of the problem, not misapplied models.

Second, our paradigm helps clarify what makes a model "good" at the prediction it is asked to make. In particular, it makes the clear distinction between the importance of fit for passive prediction and the importance of identification for active prediction.

Next, this paradigm can help us to think about which predictions can approach perfection, and which have stronger limitations (e.g., chaos theory issues for predictions about the future). It also helps us to ask whether models that have proven effective for certain applications will continue to work for future predictions. For example, we can ask whether we are dealing with a stable data-generating process, or a black box, whose reliability over time is difficult to assess other than to keep checking. Ultimately, the hope is that this paradigm, in conjunction with

others out there, can aid in using predictive analytics to make the best, and most credible, predictions possible.

References

Agrawal, A., Gans, J., and Goldfarb, A. (2017), "What to Expect from Artificial Intelligence," *MIT Sloan Management Review*, 58, 3, pp. 23-26.

https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/

Hurst, P. (2016), "A New Study Says Living Near a Pub Makes You Happier," Vice.com, January 25.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015), An Introduction to Statistical Learning with Applications in R, Springer.

Narins, E. (2016), "Study Links Soda Consumption to Heart Failure," Redbook Magazine, January 30.

Prince, J. (2019), Predictive Analytics for Business Strategy, McGraw-Hill Education.

Reuters. (2016), "Drinking More Coffee May Reverse Liver Damage from Booze," February 18.

Richards, C. (2013), "Market Forecasting Isn't Like the Weather," *New York Times*. https://bucks.blogs.nytimes.com/2013/06/17/market-forecasting-isnt-like-the-weather/

Schick, D. (2013), "Study: Marijuana Use Increases Risk of Academic Problems," *USA Today*, June 7.

Shmueli, G. (2010), "To Explain or Predict?" *Statistical Science*, 25, 3, pp. 289-310.

Varian, H. (2014), "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 2, pp. 3-28.

Vespignani, A. (2009), "Predicting the Behavior of Techno-Social Systems," *Science*, 325, 5939, pp. 425-428. http://science.sciencemag.org/content/325/5939/425

Young, J. (2017), "A New Forecast: Human Weather," *WhereNext Magazine*.