Predictive Analytics in Information Systems Research

Galit Shmueli
University of Maryland - Department of Decision, Operations & Information Technologies

O. Koppius
Rotterdam School of Management, Erasmus University; Erasmus Research Institute of Management (ERIM) - Joint Research Institute of Rotterdam School of Management (RSM) and Erasmus School of Economics (ESE), EUR

Predictive Analytics in Information Systems Research

July 31, 2010

This research essay highlights the need to integrate predictive analytics into information systems (IS) research, and shows several concrete ways in which this can be accomplished. Predictive analytics include empirical methods (statistical and other) that generate data predictions as well as methods for assessing predictive power. Predictive analytics not only assist in creating practically useful models, they also play an important role alongside explanatory modeling in theory building and theory testing. We describe six roles for predictive analytics: new theory generation, measurement development, comparison of competing theories, improvement of existing models, relevance assessment, and assessment of the predictability of empirical phenomena. Despite the importance of predictive analytics, we find that they are rare in the empirical IS literature. The latter relies nearly exclusively on explanatory statistical modeling, where statistical inference is used to test and evaluate the explanatory power of underlying causal models. However, explanatory power does not imply predictive power and thus predictive analytics are necessary for assessing predictive power and for building empirical models that predict well. To show the distinction between predictive analytics and explanatory statistical modeling, we present differences that arise in the modeling process of each type. These differences translate into different final models, so that a pure explanatory statistical model is best tuned for testing causal hypotheses and a pure empirical predictive model is best in terms of predictive power. We "convert" a well-known explanatory paper on TAM to a predictive context to illustrate these differences and show how predictive analytics can add theoretical and practical value to IS research.

Predictive Analytics in Information Systems Research

INTRODUCTION

In the last decade, the IS field has made great strides in employing more advanced statistical modeling techniques to support empirical research. It is now common to see IS researchers use structural equation modeling (Marcoulides and Saunders 2006) and increased attention is being paid to issues such as formative constructs (Petter et al. 2008) and selection bias (Li and Hitt 2008). At the same time, many opportunities for further improvement remain. In this research essay we address a particularly large gap, namely, the near-absence of predictive analytics in mainstream empirical IS literature. This gap presents an important opportunity, because predictive analytics are useful for generating new theory, developing new measures, comparing competing theories, improving existing theories, assessing the relevance of theories, and assessing the predictability of empirical phenomena.

Predictive analytics include statistical models and other empirical methods that are aimed at creating empirical predictions, as well as methods for assessing the quality of those predictions in practice, i.e., predictive power. Aside from their practical usefulness, predictive analytics play an important role in theory building, theory testing, and relevance assessment. Hence, they are a necessary component in scientific research (Kaplan 1964; Dubin 1969).

We show that despite prediction being a core scientific activity, empirical modeling in IS has been dominated by *causal-explanatory statistical modeling*, where statistical inference is used to test causal hypotheses and to evaluate the explanatory power of underlying causal models. Yet, contrary to common belief, explanatory power does not imply predictive power (Dawes 1979; Forster and Sober 1994). In addition, when statistical explanatory models are built for the purpose of testing

2

hypotheses rather than for generating accurate empirical predictions, they are less useful when the main goal is high predictive power.

The dominance of causal-explanatory statistical modeling and rarity of predictive analytics for theory building and testing exists not only in IS but in the social sciences in general, as well as in other disciplines such as economics and finance. In contrast, in fields such as computational linguistics and bioinformatics, predictive analytics are commonly used and have lead to theoretical advances. In computational linguistics: *"the mathematical and computational work has given us deep insights into the working of language… [and] will contribute to psycholinguistic research which studies the human processing of language"* (Joshi, 1991). In bioinformatics: *"A predictive model represents the gold standard in understanding a biological system and will permit us to investigate the underlying cause of diseases and help us to develop therapeutics"* (Gifford, 2001).

We begin by defining the terms 'explanatory statistical model' and 'predictive analytics' and then describe sources of differences between them. Next, the role of predictive analytics in scientific research is discussed, followed by the results of an IS literature search indicating the rarity of predictive analytics. The last part of the paper presents methods for assessing predictive power and for building predictive models. The methods are illustrated by 'converting' a well-known explanatory study of TAM into a predictive context. We conclude with a discussion.

DEFINITIONS: "EXPLANATORY STATISTICAL MODELS" AND "PREDICTIVE ANALYTICS"

We define 'explanation' and 'prediction' in the context of empirical modeling and the related terms *explanatory statistical model*, *explanatory power*, *predictive analytics*, and *predictive power*.

3

Empirical Models for Explanation

In the context of empirical modeling, we use the term *explanatory statistical model* to describe a statistical model that is built for the purpose of testing causal hypotheses that specify how and why certain empirical phenomena occur (Gregor 2006). Causal theoretical models being at the core, a set of causal hypotheses are then derived and tested using statistical models and statistical inference.

Explanatory statistical modeling includes two components:

(1) *Explanatory statistical models*, which include any type of statistical model used for testing causal hypotheses. In IS, as in the social sciences in general, it is common to assume causality at the theoretical level and then test causal hypotheses using association-type statistical models[1] such as regression models and structural equation models that rely on observational data.

(2) Methods for *evaluating the explanatory power* of a model (e.g., statistical tests or measures such as R^2), which indicates the strength of the relationship.

Examples of explanatory-oriented research in the IS literature, studied via explanatory statistical modeling, include finding determinants of auction prices (Ariely and Simonson 2003); explaining the diffusion and non-diffusion of e-commerce among SMEs (Grandon and Pearson 2004); explaining attitudes towards online security and privacy (Malhotra et al. 2004); and understanding the antecedents and consequences of online trust (Gefen et al. 2003).

---

[1] The use of association-type models for causal inference is common in the social sciences, although it is frowned upon by many statisticians under the common saying "association does not imply causation". The justification for using such models for causal inference is that given a significant association that is *consistent* with the theoretical argument, causality is inherited directly from the theoretical model.

[2] We use the terms *predictive power* and *predictive accuracy* interchangeably, as is common in predictive analytics.

[3] Predictive models rely on association rather than causation, and assume that the prediction context is probabilistically identical to the context under which the model was built. Hence, if an important causal factor is omitted, which causes the prediction context to change (termed "population drift" by Hand, 2006),

4

Empirical Models for Prediction

Predictive analytics include two components:

(1) Empirical predictive models (statistical models and other methods such as data mining algorithms) designed for predicting new/future observations or scenarios.

(2) Methods for evaluating the *predictive power* of a model.

Predictive power (or predictive accuracy[2]) refers to a model's ability to generate accurate predictions of new observations, where 'new' can be interpreted temporally, i.e., observations in a future time period, or cross-sectionally, i.e., observations that were not included in the original sample used to build the model. Examples of predictive-oriented research using predictive analytics in the context of IS include predicting the price of ongoing eBay auctions (Wang et al. 2008a); predicting future box-office sales based on online movie ratings (Dellarocas et al. 2006); and predicting repeat visits and the likelihood of purchase of online customers (Padmanabhan et al. 2006).

Note that the above definition of prediction refers to *empirical prediction* rather than *theoretical prediction*, where the latter describes an assertion that arises from a causal theory (e.g., "based on theory ABC, we predict that X will be associated with Y" or "hypothesis H1 predicts that…"). In the remainder of the paper, we use the terms "models", "modeling", and "prediction" in the sense of empirical models, empirical modeling, and empirical prediction.

Empirical Models for Explanation and Prediction

Another theory type by Gregor (2006) is "Explanation and Prediction". Both of these goals are desirable, and many empirical models indeed aim to achieve both. However, explanation and prediction are perhaps best thought of as two separate modeling goals not entirely mutually exclusive, but with a tension between them. Since the best explanatory statistical model will almost

---

[2] We use the terms *predictive power* and *predictive accuracy* interchangeably, as is common in predictive analytics.

always greatly differ from the best predictive model (Forster and Sober 1994, Konishi and Kitagawa 2007, Shmueli 2010), any model that tries to achieve both goals will have to somewhat compromise. Such comprises are common and can take several forms. For instance, when the main purpose is causal explanation but a certain level of predictive power is desired, one can build an explanatory statistical model and then, in a second stage, assess its predictive power using predictive analytics, perhaps modifying the model if it does not achieve the minimum desired level of predictive power. Or, when the main purpose is prediction but a certain level of interpretability is required (e.g., because the logic underlying the model needs to be explained to stakeholders), then predictive analytics can focus on predictors and methods that produce a relatively transparent model, while perhaps sacrificing some predictive power. Hence, designing a model for both causal explanation and empirical prediction requires understanding the tensions between the two goals and the difference between explanatory and predictive power.

In the remainder of the paper we focus on the distinction between explanatory statistical modeling and predictive analytics. While we recognize the existence of modeling for a dual goal as described above, the exposition is eased if we present both types in their respective canonical forms to more clearly dissipate the current ambiguity between them. This approach also helps highlight the roles that predictive analytics play in scientific research, roles that differ yet complement those of explanatory statistical modeling.

WHY EMPIRICAL EXPLANATION AND EMPIRICAL PREDICTION DIFFER

In the philosophy of science literature, there has been much debate over the difference between "explaining" and "predicting" (e.g., Forster and Sober 1994; Forster 2002; Sober 2002; Hitchcock

and Sober 2004; Dowe et al. 2007). Dubin (1969) argued that predictive and explanatory goals are distinct, yet both are essential to scientific research:

> *Theories of social and human behavior address themselves to two distinct goals of science: (1) prediction and (2) understanding. It will be argued that these are separate goals […] I will not, however, conclude that they are either inconsistent or incompatible. (Dubin, "Theory Building", 1969, p.9)*

In the context of IS research, Gregor (2006) proposed a taxonomy of five theory types, among them "explanation", "prediction", and "explanation and prediction".

We do not intend to contribute to the discussion at the philosophical level. Instead, we consider the difference between explaining and predicting in the context of *empirical modeling*. Within this realm, we emphasize two differences: (1) between explanatory and predictive modeling, and (2) between explanatory power and predictive accuracy.

Statisticians recognize that statistical models aimed at explanation are different from those aimed at prediction, and that explanatory power and predictive accuracy are two distinct dimensions of empirical models. For example, Konishi and Kitagawa (2007, p.2) note,

> *There may be no significant difference between the point of view of inferring the true structure and that of making a prediction if an infinitely large quantity of data is available [and] if the data are noiseless. However, in modeling based on a finite quantity of real data, there is a significant gap between these two points of view, because an optimal model for prediction purposes may be different from one obtained by estimating the 'true model'.*

In other words, the goal of finding a predictively accurate model differs from the goal of finding the true model (see also Sober (2006, p.537)). Why does the goal of analysis lead to such differences at the empirical level? The reason is the different level on which the two types of empirical models operate, as well as the notion of causality. Whereas explanatory statistical models are based on

7

underlying *causal relationships between theoretical constructs*, predictive models rely on *associations between measurable variables*. The operationalization of theoretical models and constructs into empirical models and measurable data creates a disparity between the ability to explain phenomena at the conceptual level and to generate accurate predictions at the observed level.

A related fundamental difference between explanatory and predictive empirical modeling is the metric optimized: whereas explanatory modeling seeks to minimize *model bias* (i.e., specification error) to obtain the most accurate representation of the underlying theoretical model, predictive modeling seeks to minimize the *combination of model bias and sampling variance*. However, there exists a tradeoff between model bias and sampling variance (Geman et al. 1992; Friedman 1997), which implies that improving predictive power sometimes requires sacrificing theoretical accuracy (higher bias) for improved empirical precision (lower variance) (Hastie et al. 2008, p.57). Although a properly specified explanatory statistical model will often exhibit some level of predictive power, the large statistical literature on cross-validation, shrinkage, and over-fitting shows that the best-fitting model for a single dataset is very likely to be a worse fit for future or other data (e.g., Stone 1974; Copas 1983; Hastie et al. 2008). In other words, an explanatory model may have poor predictive power, while a predictive model based on the same data may well possess high predictive power[3]. Finally, the prospective nature of predictive modeling, where a model is built for predicting new observations, is different from explanatory empirical modeling, where a model is built to retrospectively test a set of existing hypotheses. One implication, for example, is that in a predictive model all predictor variables must be available at the time of prediction, while in explanatory modeling there is no such constraint. Consider the example of a linear regression model: although it

---

[3] Predictive models rely on association rather than causation, and assume that the prediction context is probabilistically identical to the context under which the model was built. Hence, if an important causal factor is omitted, which causes the prediction context to change (termed "population drift" by Hand, 2006), predictive power might drop drastically. See also the discussion in footnote 9.

8

can be used for building an explanatory statistical model as well as a predictive model, the two resulting models will differ in many ways. The differences are not only in the statistical criteria used to assess the model, but are prevalent throughout the process of modeling: from the data used to estimate the model (e.g., variables included and excluded, form of the variables, treatment of missing data), to how performance is assessed (model validation and evaluation), and how results are used to support research. We discuss and illustrate these and other issues in later sections.

Shmueli (2010) summarizes the aforementioned sources of differences between empirical explanatory modeling and predictive analytics into four dimensions: causation-association, theory-data, retrospective-prospective, and bias-variance. The theory-data dimension means that predictive modeling relies more heavily on data whereas explanatory modeling relies more heavily on theory. However, in the context of scientific research, the data-driven nature of predictive analytics is integrated with theoretical knowledge throughout the entire model building and evaluation process, albeit in a less formal way than in explanatory statistical modeling (see the Discussion for further details and examples).

In summary, the different functions of empirical explanatory modeling and predictive analytics, and the different contexts in which they are built and later operate (testing causal-theoretical hypotheses versus generating data predictions), lead to many differences in the model building process, which translate into different final models. The final models will differ in terms of explanatory power as well as predictive power. Table 1 summarizes key differences that arise in explanatory and predictive empirical modeling. A more detailed discussion of the differences that arise in the model building process is presented in the section on Building Predictive Models.

Table 1: Differences between empirical explanatory modeling and predictive analytics

Step	Description
Analysis goal	Explanatory statistical models are used for testing causal hypotheses. Predictive models are used for predicting new observations and assessing predictability level.
Variables of interest	Explanatory: operationalized variables are only used as instruments to study the underlying conceptual constructs and the relation between them. Predictive: The observed, measurable variables are the focus.
Model building: Optimized Function	In explanatory modeling the focus is on minimizing model bias. Main risks are type I and II errors. In predictive modeling the focus is on minimizing the combined bias and variance. The main risk is over-fitting.
Model building: Constraints	Explanatory: empirical model must be interpretable, must support statistical testing of the hypotheses of interest, must adhere to theoretical model (e.g., in terms of form, variables, specification). Predictive: must use variables that are available at time of prediction.
Model evaluation	Explanatory power is measured by strength-of-fit measures and tests (e.g., R^2 and statistical significance of coefficients). Predictive power is measured by accuracy of out-of-sample predictions.

THE ROLES OF PREDICTIVE ANALYTICS IN SCIENTIFIC RESEARCH

We now focus on the value of predictive analytics to theory building, theory testing, and relevance assessment. We show that predictive analytics help develop and examine theoretical models through a different lens than explanatory statistical models, and are therefore necessary in addition to explanatory statistical models in scientific research. In particular, we describe six roles of predictive analytics in research.

Role 1: Generating New Theory

The important role of predictive models in theory building is closely related to Glaser and Strauss's argument, in the context of grounded theory, that both quantitative and qualitative data can be used for theory building. The authors stress the importance of using quantitative data for generating new theory: "quantitative data are often used not for rigorous demonstration of theory but as another way to discover more theory" (Glaser and Strauss 1980, p. 235).

10

Predictive analytics are valuable for theory building especially in fast-changing environments, such as the online environment, which poses many challenges for economic, psychological, and other theoretical models traditionally employed in IS. An example is auctions, where classical auction theory has only found limited applicability in the move from offline to online auctions, and where empirical research of online auctions has raised new theoretical and practical questions that classical auction theory does not address (Bajari and Hortacsu 2004; Bapna et al. 2008; Pinker et al. 2003).

The new types of datasets available today are rich in detail: they include and combine information of multiple types (e.g., temporal, cross-sectional, geographical, and textual), on a large number of observations, and with high level of granularity (e.g., clicks or bids at the seconds level). Such data often contain complex relationships and patterns that are hard to hypothesize, especially given theories that exclude many of newly-measurable concepts. Predictive analytics, which are designed to operate in such environments, can detect new patterns and behaviors and help uncover potential new causal mechanisms, in turn leading to the development of new theoretical models. One example is the use of predictive analytics for forecasting prices of ongoing online auctions. The predictive approach by Jank and Shmueli (2010, Chapter 4) relies on quantifying "price dynamics", such as price velocity and price acceleration patterns, from the auction start until the time of prediction, and integrating these dynamics into a predictive model alongside other common predictors (e.g., item characteristics and auction properties). While the concept of "price dynamics" is nonexistent in classic auction theory, including such empirical measures in predictive models has been shown to produce significantly more accurate price predictions across a range of items, auctions formats, and marketplaces than models excluding such information. The predictive approach thus discovered the new concept of "price dynamics" and its role in online auctions.

A second example is the study by Stern et al. (2004), in which predictive analytics were used to detect factors affecting broadband adoption by Australian households, resulting in the discovery of a

11

new construct called "technophilia". A third example is the work by Wang et al. (2008b), who studied the relationship between how firms disclose security risk factors in a certain period and their subsequent breach announcements. Using predictive analytics with textual data, the textual content of security risk factors was found to be is a good predictor of future breaches, shedding light on a relatively unexplored research area.

Role 2: Developing Measures

A second aspect of how predictive analytics support theory building is in terms of construct operationalization. This aspect is a more specific instance of new theory generation, since the development of new theory often goes hand in hand with the development of new measures (Van Maanen et al. 2007; Compeau et al. 2007).

Predictive analytics can be used to compare different operationalizations of constructs, such as 'user competence' (e.g., Marcolin et al. 2000) or different measurement instruments. Szajna (1994) notes, in the context of technology assessment instruments, that predictive validity provides a form of construct validation. The study by Padmanabhan et al. (2006) used predictive analytics to show the advantage of multi-source (user-centric) measures of user behavior over single-source (site-centric) measures for capturing customer loyalty.

Role 3: Comparing Competing Theories

Given competing theoretical models, explanatory statistical models can be used as a means of comparison. However, unless the theoretical models can be formulated in terms of nested statistical models (i.e., one model contains another as a special case), it is difficult to compare them statistically. Predictive analytics offer a straightforward way to compare models (whether explanatory or predictive), by examining their predictive accuracy. The study on project escalation by Keil et al. (2000) provides a good illustration of this aspect. They compared four explanatory (logistic

regression) models for testing the factors affecting project escalation, each model using constructs from one of four theories (self-justification theory, prospect theory, agency theory, and approach avoidance theory). All models exhibited similar explanatory power. The authors then proceeded to test the predictive accuracy of the models using predictive analytics. They discovered that the models based on approach avoidance and agency theories performed well in classifying both escalated and non-escalated projects, while models based on self-justification and prospect theories performed well only in classifying escalated projects, but did not perform well in their classification of non-escalated projects. They further examined the different factors through the predictive lens and discovered that the completion effect construct, derived from approach avoidance theory, had high discriminatory power. Another example is the aforementioned study by Padmanabhan et al. (2006), who used predictive analytics to identify factors impacting the gains from user-centric data. A third example is the study by Collopy et al. (1994), who compared diffusion models with simpler linear models for forecasting IS spending, and showed the higher predictive power of linear models. Finally, Sethi and King (1999) used predictive analytics to compare linear and nonlinear judgment models for obtaining user information satisfaction (UIS) measures.

Role 4: Improving Existing Models

Predictive analytics can capture complex underlying patterns and relationships, and thereby improve existing explanatory statistical models. One example is Ko and Osei-Bryson's study (2008) examining the impact of investments in IT on hospital productivity. The authors chose predictive analytics to resolve the mixed conclusions of previous explanatory models, and found that the impact of IT investment was not uniform and that the rate of IT impact was contingent on the amounts invested in the IT Stock, non-IT Labor, non-IT Capital, and possibly time. Their predictive approach enabled them to capture the more complex non-linear nature of the relationship, which in turn can be used to improve existing theoretical models (e.g., by including

13

moderated relationships). Another example is Keil et al.'s (2000) study on determining the factors that explain why some projects escalate and others do not. The authors, using predictive analytics to test an explanatory model, discovered that using factors from self-justification and prospect theories accurately predicted escalation, but poorly predicted non-escalation. Their finding indicates that separate theoretical models are needed for escalation phenomena and non-escalation phenomena. Such a theoretical nuance was not easily available from the explanatory metrics derived from the explanatory statistical model, i.e. the statistical significance of the model and the coefficients for the variables representing the different constructs.

Role 5: Assessing Relevance

Scientific development requires empirically rigorous and relevant research. In the words of Kaplan (1964, p.350): "It remains true that if we can predict successfully on the basis of a certain explanation, we have good reason and perhaps the best sort of reason, for accepting the explanation". Predictive analytics are useful tools for assessing the distance between theory and practice. Although explanatory power measures can tell us about the strength of a relationship, they do not quantify the empirical model's accuracy level in predicting new data. In contrast, assessing predictive power can shed light on the actual performance of an empirical model.

The Keil et al. (2000) study described above also illustrates how predictive analytics can be used to assess practical relevance. The authors found that the best model correctly classified 77% of the escalated projects and 71% of the non-escalated projects. These values are practically meaningful, as they give an idea of the impact of applying the theory in practice: how often will a project manager be able to 'see escalation coming' when using this model? When costs estimates of escalation and non-escalation are available, practical relevance can be further quantified in monetary terms, which could be used to determine the financial feasibility of preventive or corrective actions.

14

The study by Padmanabhan et al. (2006) also evaluates relevance: In addition to showing the practical usefulness of multi-source data, the authors quantified the magnitude of the gains that can be achieved by using user-centric data. In addition, they identified measures of user loyalty and browsing/buying intensity that accurately predict online purchase behavior, illustrating the practical use of a theory (related to measurement development). Another example is the study by Wu et al. (2005), who developed an explanatory model for studying the effect of advertising and website characteristics on sales. The authors used predictive assessment to validate their model and to assess its practical relevance for managerial consideration.

Besides assessing the relevance of the model as a whole, predictive analytics can also be used for assessing the practical relevance of individual predictors. For example, Collopy et al. (1994) showed that adding a price-adjustment predictor to models for IS spending greatly improves predictive power (i.e., reduces prediction error on out-of-sample data). It is worth re-emphasizing that this predictive assessment is fundamentally different from assessing statistical significance. In some cases, including statistically significant predictors can decrease predictive accuracy in at least two ways. First, additional predictors increase the variance, which may outweigh the predictive gain from their inclusion. Second, large sample sizes might inflate statistical significance of effects in an explanatory model, even if their addition to a predictive model worsens predictive accuracy due to increased variance from measurement error or over-fitting (Lin et al. 2008). Data collection costs also play a role here: a variable with a statistically significant but small standardized beta in a regression model, may suggest only a marginal increase in predictive accuracy, not worth the cost and effort of collecting that predictor in the future.

15

Role 6: Assessing Predictability

Predictive models play an important role in quantifying the level of predictability of measurable phenomena (Ehrenberg & Bound, 1993), by creating benchmarks of predictive accuracy. Knowledge of predictability (or un-predictability) is a fundamental component of scientific knowledge (see, e.g., Taleb 2007; Makridakis and Taleb 2009; Makridakis et al. 2009). A very low level of predictability can spur the development of new measures, collection of data, and new empirical approaches. Predictive models can also set benchmarks of potential levels of predictability of a phenomenon: if newer models with more sophisticated data and/or analysis methods result in only small improvements in predictive power, that level represents the meanwhile predictability of the phenomenon (at least until proven otherwise).

A predictive accuracy benchmark is also useful for evaluating the difference in predictive power of existing explanatory models. On the one hand, an explanatory model that is close to the predictive benchmark may suggest that our theoretical understanding of that phenomenon can only be increased marginally[4]. On the other hand, an explanatory model that is very far from the predictive benchmark would imply that there are substantial practical and theoretical gains to be obtained from further research. For example, Collopy et al. (1994) compared the predictive power of explanatory diffusion models for IS spending with that of predictive models, showing the superiority of the latter. While Gurbaxani and Mendelson (1994) criticized the predictive models as being "atheoretical blackbox" methods, Collopy et al's work nevertheless provided a predictability benchmark for IS spending behavior, which lead Gurbaxani and Mendelson to further develop improved explanatory empirical models for IS spending (thereby also supporting Role 4).

---

[4] For instance, Venkatesh et al. (2003, p.471) claim "Given that UTAUT explains as much as 70 percent of the variance in intention, it is possible that we may be approaching the practical limits of our ability to explain individual acceptance and usage decisions in organizations." While we do not necessarily disagree with their conclusion, ideally such statements would be couched in terms of predictive accuracy instead of explained variance.

PREDICTIVE ANALYTICS IN THE INFORMATION SYSTEMS LITERATURE

To investigate the extent to which predictive analytics are integrated into mainstream empirical IS research, a search of the literature was conducted. Using EBSCO's Business Source Premier, we searched all fulltext articles in MIS Quarterly (MISQ) and Information Systems Research (ISR) between 1990 - 2006[5] for one of the search terms "predictive OR predicting OR forecasting". Initial pre-testing of the search string revealed that although expanding the search to use additional terms such as 'predict', 'prediction' or 'predictor' yielded many more hits, none of the additional hits were relevant for our purposes; All relevant items had already been captured by the more restrictive search terms. The search returned a total of over 250 papers. Every article was then manually examined for an explicit predictive goal, or for predictive claims made based on the empirical model. We excluded articles that used predictive language in a generic sense (e.g., 'based on theory ABC, we predict that X will be associated with Y' or 'hypothesis H1 predicts that…') as well as articles that were qualitative or purely theoretical. We also excluded articles that, although explanatory in nature, used the term "predictors" to denote covariates. These last comprised a majority of the papers found. The total number of relevant predictive articles after the above filtering produced 52 articles (18 in ISR and 34 in MISQ).

We subsequently investigated whether empirical papers with predictive claims evaluated predictive power properly. The 52 articles were therefore checked for two distinguishing criteria of predictive testing:

1.   Was predictive accuracy based on out-of-sample assessment? (e.g., cross-validation or a holdout sample). This criterion is well-established in predictive testing (see, e.g., Mosteller and Tukey 1977, and Collopy et al. 1994)

---

[5] During this period there were a total of 692 articles published in MISQ and 380 in ISR.

2. Was predictive accuracy assessed with adequate predictive measures (e.g., RMSE, MAPE, PRESS[6], overall accuracy or other measures computed from a holdout set), or was it incorrectly inferred from explanatory power measures? (e.g., p-values or R^2)

It should be noted that both criteria are necessary for testing the predictive performance of any empirical model, as they test predictive performance regardless of whether the goal is explanatory and/or predictive (see also next section on assessing predictive power).

Based on these criteria, each of the 52 articles was classified as one of four types (see Table 2):

- *Predictive Goal - Adequate:* predictive goal stated; adequate predictive analytics used

- *Predictive Goal - Inadequate :* predictive goal stated; inadequate predictive analytics used

- *Predictive Assessment - Adequate :* explanatory goal stated; predictive power properly assessed

- *Predictive Assessment - Inadequate :* explanatory goal stated; predictive power incorrectly inferred from explanatory power

Table 2: Summary of Literature Search. Breakdown of "predictive" articles in ISR and MISQ (1990-2006) according to predictive goal/claims and use of predictive analytics

	ISR	MISQ	Total
Initial hits (predictive OR predicting OR forecasting, 1990-2006)	95	164	259
Relevant papers (empirical with predictive goal or claims of predictive power)	18	34	52
Predictive Goal - Adequate	4	1	5
Predictive Goal - Inadequate	8	10	18
Predictive Assessment – Adequate	1	1	2
Predictive Assessment – Inadequate	5	22	27

---

[6] *RMSE* = Root mean squared error. *MAPE* = Mean absolute percentage error. *PRESS* = predicted residual sum of squares

Two major findings emerge from this literature study:

1.   Empirical predictive goals and claims are rare: From over 1,000 published articles, only 23 of the empirical articles stated one or more goals of analysis as predictive, and only 29 made predictive claims regarding their explanatory model.

2.   Predictive analytics are rare: Only 7 papers (out of the 52) employed predictive analytics in one form or the other. The remaining 45 papers, although stating a predictive goal or making predictive claims, did not employ predictive analytics and instead inferred predictive power from explanatory power. Table 4 in the appendix lists several illustrative quotes from articles where measures of explanatory power are used for supporting predictive claims.

In summary, it can be seen from the literature search that predictive analytics are rare in mainstream IS literature, and even when predictive goals or statements about predictive power are made, they incorrectly use explanatory models and metrics. This ambiguity between explanatory and predictive empirical modeling and testing leads not only to ambiguity in matching methods to goal, but at worst may result in incorrect conclusions for both theory and practice (e.g., Dawes 1979). Hence, we next describe how predictive power should be evaluated and then describe the main steps and considerations in building predictive models.

ASSESSING PREDICTIVE POWER (OF ANY EMPIRICAL MODEL)

Predictive power refers to an empirical model's ability to predict new observations accurately. In contrast, explanatory power refers to the strength of association indicated by a statistical model. A statistically-significant effect or relationship does not guarantee high predictive power, because the precision or magnitude of the causal effect might not be sufficient for obtaining levels of predictive accuracy that are practically meaningful. To illustrate a practical IS setting where this phenomenon might occur, consider a TAM-based study on the acceptance of a radically new information system.

In such a setting, potential users have great uncertainty evaluating the usefulness of the system (Hoeffler 2002), resulting in a much larger variance for the Perceived Usefulness (PU) construct. While PU may still be statistically significant as in almost all TAM studies, its larger variance will substantially reduce the gains in predictive accuracy from including it in the model, perhaps even to the point of reducing predictive accuracy.

Most importantly, since the same data were used to fit the model and to estimate explanatory power, performance on new data will almost certainly be weaker (Mosteller and Tukey 1977, p.37).

The first key difference between evaluating explanatory versus predictive power lies in the data used for the assessment. While explanatory power is evaluated using in-sample strength-of-fit measures, predictive power is evaluated using out-of-sample prediction accuracy measures. A popular method to obtain out-of-sample data is to initially partition the data randomly, using one part (the 'training set') to fit the empirical model, and the other (the 'holdout set') to assess the model's predictive accuracy (Hastie et al. 2008, p.222; Berk 2008, p.31). In time series, the holdout set is chosen to be the last periods of the series (see Collopy et al. 1994). With smaller datasets, where partitioning the data can significantly deteriorate the fitted model (in terms of bias), methods such as cross-validation are used. In cross-validation the model is fitted to the large majority of the data, and tested on a small number of left-out observations. The procedure is then repeated multiple times, each time leaving out a different set of observations, and finally the results from all repetitions are aggregated to produce a measure of predictive accuracy (see Chapter 7.10 in Hastie et al. 2008 for further details on cross-validation).

Low predictive power can result from "over-fitting", where an empirical model fits the training data so well that it underperforms in predicting new data (see Breiman 2001a, p.204). Hence, besides avoiding fitting the training data too closely (Friedman, 2006), it is also important to compare the

20

model's performance on the training and holdout sets; a large discrepancy is indicative of over-fitting, which will lead to low predictive accuracy on new data.

The second difference between explanatory and predictive power assessment is in the metrics used. In contrast to explanatory power, statistical significance plays a minor role or no role at all in assessing predictive performance. In fact, it is sometimes the case that removing predictors with small coefficients, even if they are statistically significant (and theoretically justified), results in improved prediction accuracy (see Wu et al. 2007; for a simple example see Appendix A in Shmueli 2010).

Similarly, R^2 is an explanatory strength-of-fit measure, but does not indicate predictive accuracy (see, e.g., Copas 1983, p.237; Berk 2008, p. 29). We especially note the widespread misconception of R^2 as a predictive measure, as seen in our literature survey results (see Table 4) and even in textbooks (e.g., Mendenhall and Sinich 1989, p.158). A model with a very high R^2 indicates a strong relationship within the data used to build that model, but the same model might have very low predictive accuracy in practice (Barrett 1974).

In terms of predictive performance measures, popular metrics are out-of-sample error rate and statistics such as PRESS, RMSE and MAPE or cross-validation summaries. A popular metric for variable selection is the Akaike Information Criterion (AIC)[7]. Akaike derived the AIC from a predictive viewpoint, where the model is not intended to accurately infer the "true distribution", but rather to predict future data as accurately as possible (see, e.g., Konishi and Kitagawa 2007 and Berk 2008). AIC is useful when maximum likelihood estimation is used, but otherwise too complicated to compute.

---

[7] Although an in-sample metric, AIC is based on estimating the discrepancy between the in-sample and out-of-sample error rate, and adding this discrepancy to the in-sample error (Hastie et al., 2001, p. 203).

Specialized predictive measures: When asymmetric costs are associated with prediction errors (i.e., costs are heftier for some types of errors than for others), a popular measure is the "average cost per predicted observation". When the goal is to accurately predict the "top tier" of a population rather than the entire population (a goal particularly common in marketing and personnel psychology but often of interest in IS, e.g., predicting the most likely adopters of a new technology or predicting the biggest barriers to successful IS implementation), then model building relies on all observations, but predictive accuracy focuses on "top tier" observations, which will lead to a different final model. Lift charts are commonly used in this context (see, e.g., Padmanabhan et al. 2006 and Shmueli et al. 2010). Note that due to its focus on a particular segment of the population, a model with good lift need not necessarily exhibit a low overall error rate.

In short, since metrics for assessing predictive power are only based on the observed values and the predicted values from the model, they can be evaluated for *any empirical model* that can generate predictions. In contrast, since explanatory power assessment relies on statistical estimation and statistical inference, assessing explanatory power is straightforward only with statistical models, but not with every empirical model.

BUILDING PREDICTIVE MODELS

In this section we present a brief overview of steps and considerations in the process of building a predictive model, which differ from explanatory statistical model building. We illustrate these in the next section, by "converting" a known TAM explanatory study to a predictive context. For a detailed exposition of the differences between predictive and explanatory model building from a statistical methodological perspective see Shmueli (2010).

A schematic of the model building steps in explanatory and predictive modeling is shown in Figure 1. Although the main steps are the same, within each step a predictive model dictates different operations and criteria. The steps will now be described in more detail.

Figure 1: Schematic of the steps in building an empirical model (predictive or explanatory)

Define Goal	Design Study & Collect Data	Prepare Data	EDA	Choose Variables & Form	Choose Potential Method(s)	Evaluate, Validate, & Model Selection	Use Model & Report

Goal Definition

One common goal in predictive modeling is to accurately predict an outcome value for a new set of observations. This goal is known in predictive analytics as *prediction* (for a numerical outcome) or *classification* (for a categorical outcome). A different goal, when the outcome is categorical (e.g., adopter/non-adopter), is to rank a new set of observations according to their probability of belonging to a certain class, for the purpose of detecting the "top tier" (as mentioned earlier). This goal is known in predictive analytics as *ranking*.

Data Collection and Study Design

*Experimental versus observational setting*s: Observational data can be preferable to "overly clean" experimental data if they better represent the realistic context of prediction in terms of the uncontrolled factors, the noise, the measured response and other factors. This is unlike explanatory studies, where experiments are preferable for establishing causality (e.g., Rosenbaum 2002, p.11).

Data collection instrument: Focus is on measurement quality and relation to data at time of prediction. In predictive analytics, closeness of the collected data (used for modeling) to the prediction context is a main consideration. Ideally, the data used for modeling and for prediction consist of the same variables and are drawn in a similar fashion from the same population. This consideration often

23

overrides explanatory considerations. For instance, whereas obtrusive collection methods are disadvantageous in explanatory modeling due to the bias that they introduce, in predictive analytics obtrusiveness is not necessarily problematic if the same instrument is employed at the time of prediction. Similarly, secondary data (or even primary data) can be disadvantageous in predictive analytics if they are too different from the measurements available at the time of prediction, even if they represent the same underlying construct.

Sample size: In predictive analytics, required sample sizes are often larger than in explanatory modeling for several reasons. First, predicting individual observations has higher uncertainty than estimating population-level parameters (for instance, a confidence interval for the mean is narrower than a prediction interval for a new observation). Second, the structure of the empirical model is often learned directly from the data using data-driven algorithms rather than being constructed directly by theory. Third, predictive analytics are often used to capture complex relationships. Hence, increasing sample size can reduce both model bias and sampling variance. Finally, more data are needed for creating holdout datasets to evaluate predictive power. Guidelines for minimum sample size needed in predictive analytics are difficult to specify, as the required sample size depends on the nature of the data, the properties of the final model, and the potential predictive power, all of which are typically unknown at the start of the modeling process. Moreover, setting the sample size apriori would limit the researcher's ability to use the wide range of available predictive tools or to combine the results of multiple models, as is commonly done in predictive analytics.

Data dimension: The initial number of variables is usually large, in an effort to capture new sources of information and new relationships. Justification for each variable is based on combining theory, domain knowledge, and exploratory analysis. Large secondary datasets are often used in predictive analytics due to their breadth.

Hierarchical designs: In hierarchical designs (e.g., a sample of students from multiple schools), sample allocation for predictive purposes calls for increasing group size at the expense of the number of groups (e.g., sample heavily in a small number of schools). This strategy is the opposite when the goal is explanatory (Afshartous and de Leeuw 2005).

Data Preparation

Missing values: Determining how to treat missing values depends on (1) whether the missingness is informative of the response (Ding and Simonoff 2010) and (2) whether the missing values are in the training set or in the to-be-predicted observations (Saar-Tsechansky and Provost 2007). Missingness can be a blessing in a predictive context, if it is sufficiently informative of the response. For instance, missing data for perceived usefulness in a TAM survey might be caused by a basic unfamiliarity with the technology under investigation, which in turn increases the likelihood of non-adoption. Methods for handling missing values include removing observations, removing variables, using proxy variables, creating dummy variables that indicate missingness, and using algorithms such as classification and regression trees for imputation. Note that this treatment of missing values in a prediction context is different from that in the explanatory case, which is guided by other principles (see Little and Rubin (2002)).

Data partitioning: The dataset is randomly partitioned into two parts. The 'training set' is used to fit models. A holdout set is used to evaluate predictive performance of the final chosen model. A third dataset ('validation set') might be used for model tuning and model selection (Hastie et al 2008, p. 222). If the dataset is too small for partitioning, cross-validation techniques can be used.

Exploratory Data Analysis (EDA)

EDA consists of summarizing data numerically and graphically, reducing their dimension, and handling outliers.

25

Visualization: In predictive modeling, EDA is used in a free-form fashion to support capturing relationships that are perhaps unknown or at least less formally formulated. This type of exploration is called "exploratory visualization", as opposed to the more restricted and theory-driven "confirmatory visualization" (Fayyad et al. 2002). Interactive visualization supports exploration across a wide and sometimes unknown terrain, and is therefore useful for learning about measurement quality and associations that are at the core of predictive modeling.

Dimension reduction: Due to the often large number of predictors, reducing the dimension can help reduce sampling variance (even at the cost of increasing bias), and in turn increase predictive accuracy. Hence, methods such as principal components analysis (PCA) or other data compression methods are often carried out initially. The compressed variables can then be used as predictors.

Choice of Variables

Predictive models are based on association rather than causation between the predictors and the response. Hence variables (predictors and response) are chosen based on their observable qualities. The response variable and its scale are chosen according to the predictive goal, data availability, and measurement precision. Two constraints in choosing predictors are their availability at the time of prediction (ex-ante availability[8]), and their measurement quality. The choice of potential predictors is often wider than in an explanatory model, due to the objective of discovering new relationships. Predictors are chosen based on a combination of theory, domain knowledge, and empirical evidence of association with the response. Although in practical prediction the relation between the predictors and underlying constructs is irrelevant, construct consideration can be relevant in some theoretical development research (see Discussion section). Note that although improving construct validity reduces model bias, it does not address measurement precision which affects sampling variance; and

---

[8] For instance, including the number of bidders in an online auction as a covariate is useful for *explaining* the final price, but cannot be used for *predicting* the price of an ongoing auction (because it is unknown until the auction closes).

prediction accuracy is determined by both model bias and sampling variance. For this reason, when proxy variables or even confounding variables can be measured more precisely and are more strongly correlated with the measured output than "proper" causal variables, those can be better choices for a predictive model than the "theoretically correct" predictors. For the same reason, in predictive models there is typically no distinction between predictors in terms of their causal priority as in mediation analysis, and considerations of endogeneity and model identifiability are irrelevant. In fact, under-specified models can produce better predictions (Wu et al. 2007). For instance, Montgomery et al. (2005) showed that it is often beneficial to exclude the main effects in a model even if the interaction term between them is present.

Choice of Potential Methods

Data-driven algorithms: Predictive models often rely on non-parametric data mining algorithms (e.g., classification trees, neural networks and k-nearest-neighbors) and non-parametric smoothing methods (e.g., moving average forecasters, wavelets). The flexibility of such methods enables them to capture complex relationships in the data without making restricting statistical assumptions. The price of this flexibility is lower transparency (*"Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict."* Breiman 2001a, p.206). However, correct specification and model transparency are of lesser importance in predictive analytics than in explanatory modeling.

Shrinkage methods: Methods such as ridge regression and principal components regression (Hastie et al. 2008, Chapter 3) sacrifice bias for a reduction in sampling variance, resulting in improved prediction accuracy (see, e.g., Friedman and Montgomery 1985). Such methods "shrink" predictor coefficients or even set them to zero, thereby effectively removing the predictors altogether.

Ensembles: A popular method for improving prediction accuracy is using ensembles, i.e., averaging across multiple models that rely on different data or reweighted data, and/or employ different

models or methods. Similar to financial asset portfolios ('ensembles'), where a reduction of portfolio risk can be achieved through diversification, the underlying idea of ensembles is that combining models reduces the sampling variance of the final model, which results in better predictions. Widely-used ensemble methods include bagging (Breiman, 1996), random forests (Breiman, 2001b), boosting (Shapire, 1999), and variations of these methods.

Evaluation, Validation and Model Selection

Model Evaluation: To evaluate the predictive performance of a model, predictive accuracy is measured by applying the method to a holdout set and generating predictions.

Model Validation: Over-fitting is the major focus in predictive analytics (Stone 1974, Copas 1983, Hastie et al. 2008). Assessing over-fitting is achieved by comparing the performance on the training and holdout sets, as described earlier.

Model Selection: One way to reduce sampling variance is to reduce the data dimension (number of predictors). Model selection is aimed at finding the right level of model complexity that balances bias and variance, in order to achieve high predictive accuracy. This consideration is different from explanatory considerations such as model specification. For instance, for purposes of prediction "*multicollinearity is not so damning*" (Vaughan and Berry 2005, online). Variable selection and stepwise-type algorithms are useful, as long as the selection criteria are based on predictive power (i.e., using predictive metrics as described in the section on Assessing Predictive Power).

Model Use and Reporting

Studies that rely on predictive analytics focus on predictive accuracy and its meaning. Performance measures (e.g., error rates and classification matrices) and plots (e.g., ROC curves and lift charts) are geared towards conveying predictive accuracy and if applicable, related costs. Predictive power is compared against naive and alternative predictive models (e.g., Armstrong 2001). In addition, the

28

treatment of over-fitting is often discussed. An example of a predictive study report in the IS literature is Padmanabhan et al. (2006). Note the overall structure: the placement of the section on 'Rational behind Variable Construction' in the Appendix; the lack of causal statements or hypotheses; the reported measures and plots; the emphasis on predictive assessment; reporting model evaluation in practically relevant terms; and the translation of results into new knowledge.

EXAMPLE: PREDICTIVE MODEL FOR TAM

To illustrate how the considerations mentioned above affect the process of building a predictive model, and to contrast that with the explanatory process, we will 'convert' a known IS explanatory study into a predictive one in the context of the TAM model (Davis 1989). In particular, we chose the study on "Trust and TAM in Online Shopping: An Integrated Model" by Gefen et al. (MISQ, 2003) – further denoted GKS. In brief, the study examines the role of trust and IT assessment (perceived usefulness and ease of use) in online consumers' purchase intentions (denoted as behavioral intention, or BI). The authors collected data via a questionnaire, filled by a sample of 400 students considered to be "experienced online shoppers". Responders were asked about their last online purchase of a CD or book. The final relevant dataset consisted of 213 observations and was used to test a set of causal hypotheses regarding the effect of trust and IT assessment on purchase intentions. The goal of the GKS study was explanatory, and the statistical modeling was correspondingly explanatory.

We now approach the same topic from a predictive perspective.

Goal Definition

Possible research goals include benchmarking the predictive power of an existing explanatory TAM model, evaluating the survey questions' ability to predict intention, revealing more complicated relationships between the inputs and BI, and validating the predictive validity of constructs.

In terms of empirical goal, consider the goal of predicting BI for shoppers that were not part of the original sample. The original GKS data can be used as the training set to build (and evaluate) a model that predicts BI. This model can then be deployed in a situation where a similar questionnaire is administered to potential shoppers from the same population, but with the BI questions excluded (whether to shorten questionnaire length, to avoid social desirability issues in answers, or for another reason). According to his/her responses, the shopper's BI is predicted (and, for instance, an immediate customization of the online store takes place).

The overall net benefit of the predictive model would be a function of the prediction accuracy and, possibly, of costs associated with prediction error. For example, we may consider asymmetric costs, such that erroneously predicting low BI (while in reality a customer has high BI) is more costly than erroneously predicting high BI. The reason for such a cost structure could be the amount of effort that an e-vendor invests in high-BI customers. Or, the opposite cost structure could be assumed, if an e-vendor is focused on retention.

An alternative predictive goal could be to rank a new set of customers from most likely to least likely to express high BI, for the purpose of identifying, say, the top or bottom 10% of customers.

For the sake of simplicity, we continue with the first goal described above, without considering costs. Table 3 summarizes the main points and compares the explanatory and predictive modeling processes that are described next.

Data Collection and Study Design

*Experimental versus observational setting*s: Due to the predictive context, the GKS observational survey is likely preferable to an experiment, because the "dirtier" observational context is similar to the predictive context in the field than a "clean" lab setting would be.

Instrument: In choosing a data collection instrument, attention is first given to its relation to the prediction context. For instance, using a survey to build and evaluate the model is most appropriate if a survey will be used at the time of prediction. The questions and measurement scales should be sufficiently similar to those used at the time of prediction. Moreover, the data to be predicted should be from the same population as the training and evaluation data and should have similar sample properties[9], so that the training, evaluation and prediction contexts are as similar as possible. Note that bias created by the obtrusive nature of the survey or by self-selection is irrelevant, because the same mechanism would be used at the time of prediction. The suitability of a retrospective questionnaire would also be evaluated in the prediction context, e.g., whether a retrospective recount of a purchase experience is predictive of future BI. In designing the instrument, the correlation with BI would also be taken into account (ideally through the use of pre-testing). For instance, the seven-point Likert scale might be replaced by a different scale (finer or coarser) according to the required level of prediction accuracy.

Sample size: The final usable sample of 213 observations is considered small in predictive analytics, requiring the use of cross-validation in place of a holdout set and being limited to model-based

---

[9] The definition of 'same population' is to some extent at the researcher's discretion, e.g., is the population here 'US college students', 'college students', 'experienced online shoppers', 'online shoppers' and so on. The population to which the predictive model is deployed should be similar to the one used for building and evaluating the predictive model, otherwise predictive power is not guaranteed. In terms of sampling, if the same biases (e.g., self-selection) are expected in the first and second datasets, then the predictive model can be expected to perform properly. Finally, predictive assessment can help test the generalizability of the model to other populations by evaluating predictive power on samples from such populations where the BI questions are included, thereby serving as holdout samples.

methods. Depending on the signal strength and the data properties, a larger sample that would allow for use of data-driven algorithms might improve predictive power.

Data dimension: Using domain knowledge and examining correlations, any additional information beyond the survey answers that might be associated with BI would be considered, even if not dictated by TAM theory (e.g., the website of most recent purchase, or # of previous purchases).

Data Preparation

Missing data: GKS report that the final dataset contained missing values. For prediction, one would check whether the missingness is informative of BI, e.g., if it reflects less trusting behavior. If so, including dummy variables that indicate the missingness might improve prediction accuracy.

Data partitioning: Due to the small dataset, the data would not be partitioned. Instead, cross-validation methods would be used. When and if another sample is obtained (perhaps as more data are gathered at the time of prediction), then the model could be applied to the new sample, which would be considered a holdout set.

Exploratory Data Analysis (EDA)

Data visualization and summaries: Each question, rather than each construct, would be treated as an individual predictor. In addition to exploring each variable, examining the correlation table between BI and all the predictors would help identify strong predictor candidates and information overlap between predictors (candidates for dimension reduction).

Dimension reduction: PCA or a different compression method would be applied to the predictors in the *complete* training set, with predictors including individual questions and any other measured variables such as demographics (this procedure differs from the explanatory procedure, where PCAs were run separately for each construct.) The resulting compressed predictors would then be used in the predictive model, with less or no emphasis on their interpretability or relation to constructs.

32

Choice of Variables

Ex-ante availability: To predict BI, predictors must be available at the time of prediction. The survey asks respondents retrospectively about their perceived usefulness and ease of use as well as BI. Given the predictive scenario, the model can be used for assessing the predictability of BI using retrospective information, for comparing theories, or even for practical use. In either case, the BI question(s) in the original study should be placed last in the questionnaire, to avoid affecting earlier answers (a clickstream-based measure of BI, e.g. Hauser et al. (2009), would be another way of dealing with this issue), and to obtain results that are similar to the prediction context. In addition, each of the other collected variables should be assessed as to its availability at the time of prediction.

Measurement quality: The quality and precision of predictor measurements are of key importance in a predictive version of GKS, but with a slight nuance: while a unidimensional operationalization of constructs such as trust, PU and PEOU is desirable, it should not come at the expense of measurement precision and hence increased variance. Unobtrusive measures such as clickstream data or purchase history (if available) would be particularly valued here. Even though they might be conceptually more difficult to interpret in terms of the underlying explanation, their measurement precision can boost predictive accuracy.

Choice of Potential Methods

Data-driven algorithms would be evaluated (although the small dataset would limit the choices). Shrinkage methods could be applied to the raw question-level data, before data reduction. If we are predicting BI for people who have survey-answering profiles that are different from those in the training data (i.e., extrapolation), then shrinkage methods should be used. The issue of extrapolation is also relevant to the issue in GKS of generalizing their theory to other types of users.

33

Ensembles would be considered. In particular, the authors mention the two competing models of TAM and TBP which can be averaged to produce an improved predictive model. Similarly, if clickstream data were available, one could average the results from a survey-based BI-model and a clickstream-based BI-model to produce improved predictions. If real-time prediction is expected then computational considerations will affect the choice of methods.

Evaluation, Validation and Model Selection

Predictive variable selection algorithms (e.g., stepwise-type algorithms) could be used to reduce the number of survey questions, using criteria such as AIC or out-of-sample predictive accuracy. Predictive accuracy would be evaluated using cross-validation (due to small sample size), and compared to competing models and the naïve prediction "*predict each BI by the overall average BI*".

Model Use and Reporting

The results of the predictive analytics can be used here for one, or more of several research goals:

(1) Benchmarking the predictive power of existing explanatory TAM models: The paper would present the predictive accuracy of different TAM models and discuss practical differences. Also, an indication of overall predictability could be obtained.

(2) Evaluating the actual precision of the survey questions with respect to predicting BI: A comparison of the predictive accuracy of different models which rely on different questions.

(3) Revealing more complicated relationships between the inputs and BI, such as moderating effects: Comparing the predictive power of the original and more complex model and showing how the added complexity provides a useful contribution.

(4) Validating assertions about the predictive validity of concepts: GKS (p.73) remark: "the TAM construct PU remains an important predictor of intended use, as in many past studies". Such an

assertion in terms of actual prediction would be based on the predictive accuracy associated with

PU (e.g., by comparing the best model that excludes PU to the model with PU).

These are a few examples of how the predictive analytics complement explanatory TAM research.

Table 3: Building explanatory versus predictive models: Summary of the Gefen et al. (2003) example

Modeling Step	Explanatory Task	Predictive Task
Goal Definition	Understand the role of trust and IT assessment (perceived usefulness and ease of use) in online consumers' purchase intensions	Predict the intention of use (BI) of new B2C website customers, or, predict 10% of those most likely to express high BI. (Might include asymmetric costs).
Study Design and Data Collection	Observational data Survey (obtrusive) **Sample size:** 400 students (213 usable observations) **Variables:** operationalization of PU and PEOU, demographics **Instrument:** Questionnaire; seven-point Likert scale **Pre-testing:** for validating questionnaire	Observational data – similar to prediction context; Variables must be available at prediction time Survey (obtrusive) – with identical questions and scales as at prediction time **Sample Size:** Larger sample preferable **Variables:** Predictors that strongly *correlate* with BI (questions, demographics, other information) **Instrument:** Questionnaire; BI questions last; non-retrospective would be better; scale for questions - according to required prediction scale, and correlations with BI **Pre-testing:** for trouble-shooting questionnaire
Data Preparation	**Missing Values:** some missing values reported, action not reported **Data Partitioning:** none	**Missing Values:** Is missingness informative of BI? If so, add relevant dummy variables; is missingness in training data or to-be-predicted data? **Data Partitioning:** sample size too small (213); cross-validation used
Continued on Next page		

Modeling Step	Explanatory Task	Predictive Task
Exploratory Data Analysis	**Summaries:** Numerical summaries for constructs; Pairwise correlations between questions; Univariate summaries by gender, age and other variables. **Plots:** None **Data Reduction:** PCA applied separately to each construct for purpose of construct validation (during pre-testing)	**Summaries:** Examine numerical summaries of all questions and additional collected variables (such as gender, age), correlation table with BI. **Plots:** Interactive visualization **Data Reduction:** PCA or other data reduction method applied to complete set of questions and other variables; applied to entire data (not just pre-test)
Choice of Variables	Guided by theoretical considerations	Predictors chosen based on their association with BI; BI chosen according to prediction goal
Choice of Methods	Structural equations model (after applying confirmatory factor analysis to validate the constructs)	Try an array of methods: **Model-driven and data-driven** methods, ideally on a larger collected sample: machine-learning algorithms, parametric and non-parametric statistical models **Shrinkage methods** - for reducing dimension (instead of PCA); for robust extrapolation (if deemed necessary); for variable selection **Ensemble methods** – combine several models to improve accuracy, (e.g., TAM and TBP)
Model Evaluation, Validation and Selection	Questions removed from constructs based on residual variance backed by theoretical considerations; constructs included based on theoretical considerations **Explanatory power** based on theoretical coherence, strength-of-fit statistics, residual analysis, estimated coefficients, statistical significance	Variable selection algorithms applied to original questions **Predictive power** Predictive accuracy assessed on holdout set (use cross-validation if small sample); evaluate over-fitting (compare performance on training and holdout data)
Model Use and Reporting	**Use:** Test causal hypotheses about how trust and TAM affect BI **Statistical Reporting:** explanatory power metrics (e.g., path coefficients), plot of estimated path model	**Use:** Discover new relationships (e.g., moderating effect; unexpected questions or features that predict BI), evaluate magnitude of trust and TAM effects in practice, assess predictability of BI **Statistical Reporting:** predictive accuracy, final predictors, method used, over-fitting analysis

36

DISCUSSION

In this essay we discussed the role of predictive analytics in scientific research; how they differ from explanatory statistical modeling; and their current under-representation in mainstream IS literature. We also described how to assess the predictive power of any empirical model and how to build a predictive model. Predictive models can lead to the discovery of new constructs, new relationships, nuances to existing models, and unknown patterns. Predictive assessment provides a straightforward way to assess the practical relevance of theories, to compare competing theories, to compare different construct operationalizations, and to assess the predictability of measurable phenomena.

Predictive analytics support the extraction of information from large datasets and from a variety of data structures. Although they are more data-driven than explanatory statistical models, in the sense that predictive models integrate knowledge from existing theoretical models in a less formal way than explanatory statistical models, they can be useful for theory development provided that a careful linkage to theory guides both variable and model selection. It is the responsibility of the researcher to carefully ground the analytics in existing theory. The few IS papers that use predictive analytics demonstrate the various aspects of linking and integrating the predictive analytics into theory. One such link is in the literature review step, discussing existing theories and models and how the predictive study fits in. Examples are the study by Ko and Osei-Bryson (2008) that relies on production theory and considers existing IT productivity studies and models; The predictive work by Wang et al. (2008b), which was linked to the body of literature in the two areas of management and economics of information security and disclosures in accounting. And finally, the study by Stern et al. (2004), which examined existing theoretical models and previous studies of broadband adoption and used them as a basis for their variable choice. This study also directly specified the potential theoretical contribution: "*Findings that emerge from the data can be compared with prior theory and any unusual findings can suggest opportunities for theory extension or modification*" (p.453). A second link to

37

theory is at the construct operationalization stage. In studies that are aimed at generating new theory, the choice of variables should of course be motivated by and related to previous studies and existing models. However, if the goal is to assess the predictability of a phenomenon or to establish a benchmark of potential predictive accuracy, then construct considerations are negligible. Finally, research conclusions should specifically show how the empirical results contribute to the theoretical body of knowledge. As mentioned earlier, the contribution can be in terms of one or more of the six roles: discovering new relationships potentially leading to new theory, contributing to measure development, improving existing theoretical models, comparing existing theories, establishing the relevance of existing models, and assessing predictability of empirical phenomena.

In light of our IS literature survey, a question that arises is whether the under-representation of predictive analytics in mainstream IS literature indicates that such research is not being conducted within the field of IS, or that such research exists but does not get published in these top two journals. A related question is why most published explanatory statistical models lack predictive testing. We do not aim to answer these questions, although we suspect that the situation is partly due to the traditional conflation of explanatory power with predictive accuracy. Classic statistical education and textbooks focus on explanatory statistical modeling and statistical inference, and very rarely discuss prediction other than in the context of prediction intervals for linear regression. Predictive analytics are taught in machine learning, data mining and related fields. Thus, the unfamiliarity of most IS researchers with predictive analytics may be another reason why we see little of it so far in the IS field. We hope that this research essay convinces IS researchers to employ more predictive analytics, but not only when the main goal is predictive. Even when the main goal of the modeling is explanatory, augmenting the modeling with predictive power evaluation is easily done and can add substantial insight. We therefore strongly advocate adopting and reporting predictive

power as accepted practice in empirical IS literature. We predict that increased application of predictive analytics in the IS field holds great theoretical and practical value.

Acknowledgements

REFERENCES

Afshartous D and de Leeuw J. 2005. Prediction in Multilevel Models. *Journal of Educational and Behavioral Statistics*. **30**(2): 109-139.

Ariely D and Simonson I. 2003. Buying, bidding, playing, or competing? Value assessment and decision dynamics in online auctions. *Journal of Consumer Psychology* **13**(1-2): 113-123.

Armstrong JS. 2001. *Principles of Forecasting – A Handbook for Researchers and Practitioners*, Springer.

Bapna R, Jank W and Shmueli G. 2008. Price Formation and its Dynamics in Online Auctions. *Decision Support Systems*, **44:** 641-656.

Bajari P and Hortacsu A. 2004. Economic Insights from Internet Auctions, *Journal of Economic Literature,* **42**(2): 457-486.

Barrett JP. 1974. The Coefficient of Determination - Some Limitations. *The American Statistician,* **28**(1):19-20.

Berk RA. 2008. *Statistical Learning from a Regression Perspective.* Springer.

Breiman L. 1996. Bagging Predictors. *Machine Learning*, **24**:123-140.

Breiman L. 2001a. Statistical Modeling: The Two Cultures. *Statistical Science* **16**:199-215

Breiman L. 2001b. Random Forests. *Machine Learning Journal.* **45:**5-32.

Compeau DR, Meister DB and Higgins CA. 2007. From Prediction to Explanation: Reconceptualizing and Extending the Perceived Characteristics of Innovation, *Journal of the Association of Information Systems,* **8(**8):409-439.

Copas JB. 1983. Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society B.* **45:**311-354.

Collopy F, Adya M, and Armstrong JS. 1994. Principles for Examining Predictive-Validity - The Case of Information-Systems Spending Forecasts. *IS Research* **5**(2):170-179.

Davis FD. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly.* **13**(3):319-340.

Dawes RM. 1979. The robust beauty of improper linear models in decision making. *American Psychologist*, **34**(7): 571-582.

Dellarocas C, Awad NF, and Zhang X. 2007. Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures. *Journal of Interactive Marketing,* **21**(4):23-45.

Ding Y and Simonoff JS. 2010. An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data. *Journal of Machine Learning Research* **11**:131-170.

Dowe DL, Gardner S and Oppy GR. 2007. Bayes not Bust! Why Simplicity is No Problem for Bayesians. *British Journal for the Philosophy of Science*, 58(4):709-754.

Dubin R. 1969. *Theory building.* New York: The Free Press.

Ehrenberg ASC and Bound JA. 1993. Predictability and Prediction. *Journal of the Royal Statistical Society Series A*, **156**(2): 167-206

Fayyad UM, Grinstein GG, and Wierse A. 2002. *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann.

Forster MR and Sober E. 1994. How To Tell When Simpler, More Unified, or Less Ad-Hoc Theories Will Provide More Accurate Predictions. *British Journal for the Philosophy of Science* **45**(1):1-35

Forster MR. 2002. Predictive Accuracy as an Achievable Goal of Science. *Philosophy of Science* **69**(3): S124-S134

41

Friedman DJ and Montgomery DC. 1985. Evaluation of the Predictive Performance of Biased Regression Estimators. *Journal of Forecasting*, **4**:153-163.

Friedman JH. 1997. On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, **1**: 55-77.

Friedman JH. 2006. Comment: Classifier Technology and the Illusion of Progress, *Statistical Science.* **21**(1): 15-18

Gefen D, Karahanna E, and Straub DW. 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly* **27**(1):51-90

Geman S, Bienenstock E, and Doursat R. 1992. Neural Networks and The Bias/Variance Dilemma. *Neural Computation* **4**:1-58.

Gifford DK. 2001. Blazing Pathways Through Genetic Mountains. *Science.* **293**(5537):2049-2051.

Glaser BG and Strauss AL. 1980. *The Discovery of Grounded Theory: Strategies for Qualitative Research.* 11th edition. New York: Aldine Publishing Company.

Grandon EE and Pearson JM. 2004. Electronic Commerce Adoption: An Empirical Study of Small and Medium US Businesses. *Information & Management.* **42**(1):197-216.

Gregor S. 2006. The Nature of Theory in IS. *MIS Quarterly.* **30**(3):611-642.

Gurbaxani V and Mendelson H. 1994. Modeling vs. Forecasting: The Case of Information Systems Spending. *Information Systems Research.* **5**(2):180-190.

Hand DJ. 2006. Classifier Technology and the Illusion of Progress. *Statistical Science.* **21**(1):1-14.

Hastie T, Tibshirani R, and Friedman JH. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition.

Hauser JR, Urban GL, Liberali G and Braun M. 2009. Website Morphing. *Marketing Science.* **28**(2):202-223.

Hitchcock C and Sober E. 2004. Prediction Versus Accommodation and the Risk of Overfitting. *British Journal for the Philosophy of Science.* **55**(1):1-34.

Hoeffler S. 2002. Measuring Preferences for Really New Products. *Journal of Marketing Research.* **40**(4):406-21.

Jank W and Shmueli G. 2010. *Modeling Online Auctions*, John Wiley & Sons, New Jersey.

Joshi AJ. 1991. Natural Language Processing. *Science.* **523**(5025):1242-1249.

Kaplan A. 1964 *The Conduct of Inquiry: Methodology for Behavioral Science.* Chandler Publishing, New York.

Keil M, Mann J and Rai A. 2000. Why Software Projects Escalate: An Empirical Analysis and Test of Four Theoretical Models, *MIS Quarterly.* **24**(4): 631-664.

Ko M and Osei-Bryson K-M. 2008. Reexamining the Impact of Information Technology Investment on Productivity using Regression Tree and Multivariate Adaptive Regression Splines (MARS). *Information Technology and Management.* **9**(4):285-299.

Konishi S and Kitagawa G. 2007. *Information Criteria and Statistical Modeling.* Springer.

Li X and Hitt LM. 2008. Self-Selection and Information Role of Online Product Reviews. *Information Systems Research* **19**(4): 456-474.

Lin M., Lucas HJ and Shmueli G. 2008. Is more Always Better? Larger Samples and False Discoveries. *Working paper* RHS 06-068*, Robert H Smith School of Business, University of Maryland.*

Little RJA and Rubin DB. 2002. *Statistical Analysis with Missing Data*, Wiley New York, 2nd edition.

43

Makridakis S, Hogarth RM, and Gaba A. 2009. Forecasting and Uncertainty in the Economic and Business World, *International Journal on Forecasting.* **25**(4):794-812.

Makridakis S and Taleb N. 2009. Decision Making and Planning Under Low Levels of Predictability, *International Journal on Forecasting.* **25**(4): 716-733.

Malhotra NK, Kim SS and Agarwal J. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research.* **15**(4): 336-355.

Marcolin BL, Compeau DR, Munro MC and Huff SL. 2000. Assessing User Competence: Conceptualization and Measurement. *Information Systems Research.* **11**(1):37-60.

Marcoulides GA and Saunders C. 2006. PLS: A silver bullet? *MIS Quarterly.* **30**(2): III-IV.

Mendenhall W and Sinich T. 1989. *A Second Course in Business Statistics: Regression Analysis.* Dellen Publishing Company, 3rd edition.

Montgomery DC, Myers RH, Carter WH, and Vining GG. 2005. The Hierarchy Principle in Designed Industrial Experiments. *Quality and Reliability Engineering International.* **21**:197-201.

Mosteller F and Tukey JW. 1977. *Data Analysis and Regression.* Reading, Mass.: Addison-Wesley

Padmanabhan B, Zheng Z and Kimbrough SO. 2006. An Empirical Analysis of The Value of Complete Information for eCRM Models. *MIS Quarterly.* **30**(2):247-267.

Petter S, Straub D and Rai A. 2007. Specifying Formative Constructs in Information Systems Research. *MIS Quarterly.* **31**(4):623-656.

Pinker E, Seidmann A and Vakrat Y. 2003. The Design of Online Auctions: Business Issues and Current Research. *Management Science.* **49**(11):1457-1484.

Rosenbaum PR. 2002. *Observational Studies.* Springer, 2nd edition.

Saar-Tsechansky M and Provost F. 2007. Handling Missing Features when Applying Classification Models. *Journal of Machine Learning Research*. **8**(July):1625-1657.

Sethi V and King RC. 1999. Nonlinear and Noncompensatory Models in User Information Satisfaction Measurement. *Information Systems Research*. **10**(1):87-96.

Shapire RE. 1999. A Brief Introduction to Boosting. In *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*.

Shmueli G. 2010. To Explain or To Predict? *Statistical Science*, In Press.

Shmueli G, Patel NR, and Bruce PC. 2010. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, John Wiley & Sons, 2nd edition.

Sober E. 2002. Instrumentalism, Parsimony, and the Akaike Framework. *Philosophy of Science*. **69**(3):S112-S123.

Sober E. 2006. *The Philosophy of Science: An Encyclopedia*, entry "Parsimony". Eds: Sarkar S and Pfeifer J, Routledge, pp. 531-538.

Stern SE, Gregor S, Martin MA, Goode S, and Rolfe J. 2004. A Classification Tree Analysis of Broadband Adoption in Australian Households. *Proceedings of the 6th international conference on Electronic commerce:* pp. 451-456.

Stone M. 1974. Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion). *Journal of the Royal Statistical Society B.* **36:** 111-147.

Szajna B. 1994. Software Evaluation and Choice: Predictive Validation of the Technology Acceptance Instrument. *MIS Quarterly.* **18**(3): 319-324.

Taleb N. 2007. *The Black Swan*. Penguin Books.

Van Maanen J, Sorensen JB and Mitchell TR. 2007. The Interplay Between Theory and Method. *Academy of Management Review*. **32**(4):1145-1154.

Vaughn TS and Berry KE. 2005. Using Monte Carlo Techniques to Demonstrate the Meaning and Implications of Multicollinearity. *Journal of Statistics Education*. **13**(1).

Wang S, Jank W, and Shmueli G. 2008a. Explaining and Forecasting Online Auction Prices and Their Dynamics Using Functional Data Analysis. *Journal of Business and Economic Statistics*. **26**(3):144-160.

Wang T-W, Rees J and Kannan, KN. 2008b. The Association between the Disclosure and the Realization of Information Security Risk. *Working Paper, Purdue University* (Available at SSRN: http://ssrn.com/abstract=1083992)

Wu J, Cook VJ Jr. and Strong EC. 2005. A Two-Stage Model of the Promotional Performance of Pure Online Firms. *Information Systems Research*. **16**(4):334-351.

Wu S Harris TJ and McAuley KB. 2007. The Use of Simplified and Misspecified Models: Linear Case, *Canadian Journal of Chemical Engineering*. **85**:386-398.

Appendix: Table 4: Illustrative quotes from the literature review

Article	Quote
Rai, A., Patnayakuni, R., and Seth, N. "Firm performance impacts of digitally enabled supply chain integration capabilities," *MIS Quarterly* 30(2), Jun 2006, pp 225-246.	*"One indicator of the predictive power of path models is to examine the explained variance or R^2 values" (p.235)*
Pavlou, P.A., and Fygenson, M. "Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior," *MIS Quarterly* 30(1), Mar 2006, pp 115-143.	*"To examine the predictive power of the proposed model, we compare it to four models in terms of R^2 adjusted" (p.131)*
Gattiker, T.F., and Goodhue, D.L. "What happens after ERP implementation: Understanding the impact of interdependence and differentiation on plant-level outcomes," *MIS Quarterly* 29(3), Sep 2005, pp 559-585.	*"However, coordination benefits do not predict overall ERP benefits as strongly as do task efficiency and data quality (as the standardized regression coefficients in Figure 2 indicate)" (p.579)*
Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F.D. "User acceptance of information technology: Toward a unified view," *MIS Quarterly* 27(3), Sep 2003, pp 425-478.	*"With the exception of MM and SCT, the predictive validity of the models increased after including the moderating variables. For instance, the variance explained by TAM2 increased to 53 percent." (p.445)*
Wixom, B.H., and Todd, P.A. "A theoretical integration of user satisfaction and technology acceptance," *Information Systems Research* 16(1), Mar 2005, pp 85-102.	*"Usefulness and attitude again dominate in the prediction of intention, and the remaining path coefficients are generally small (8 of 13 are below 0.1). The explanatory power for intention increases marginally from 0.59 to 0.63." (p.97)*
Jones, Q., Ravid, G., and Rafaeli, S. "Information overload and the message dynamics of online interaction spaces: A theoretical model and empirical exploration," *Information Systems Research* 15(2), Jun 2004, pp 194-210.	*"Unfortunately, while the ranking and variable matching enabled regression modeling, this approach results in a loss of variance and predictive/explanatory power." (p.203)*
Jarvenpaa, S.L., Shaw, T.R., and Staples, D.S. "Toward contextualized theories of trust: The role of trust in global virtual teams," *Information Systems Research* 15(3), Sep 2004, pp 250-267.	*"The predictive power of the model (i.e., variance explained) was quite high in Study 1" (p.262)*
Bassellier, G., Benbasat, I., and Reich, B.H. "The influence of business managers' IT competence on championing IT," *Information Systems Research* 14(4), Dec 2003, pp 317-336.	*"We can also assess the completeness of our constructs by examining their ability to predict the measured overall IT knowledge and IT experience. The second order factor IT knowledge explains 71% of the variance in the overall IT knowledge." (p.331)*
Kraut, R, Mukhopadhyay, T, Szczypula, J, Kiesler, S, and Scherlis, B. "Information and Communication: Alternative Uses of the Internet in Households.Preview", *Information Systems Research* 10(4), Dec 1999, pp 287-303.	*"One can predict a participant's current e-mail use from his or her use in the prior week much better than one can predict a participant's current Web use from his or her prior Web use ($\beta_{e\text{-}mail}$ =0.73 versus β_{web}=0.56)" (p. 296)*