UNIVERSITY OF MARYLAND 18 56

ROBERT H. SMITH
SCHOOL OF BUSINESS

Leaders for the Digital Economy

Working Paper No. RHS-06-058

March 24, 2008

Contrasting Predictive and Explanatory Modeling in IS Research

Galit Shmueli
Smith School of Business, University of Maryland, College Park MD, USA

Otto Koppius
Rotterdam School of Management, Erasmus University, The Netherlands,

CONTRASTING PREDICTIVE AND EXPLANATORY MODELING IN IS RESEARCH

Galit Shmueli, Smith School of Business, University of Maryland, College Park MD, USA,
gshmueli@rhsmith.umd.edu
Otto Koppius, Rotterdam School of Management, Erasmus University, The Netherlands,
okoppius@rsm.nl

March 24, 2008

Abstract

Explanatory models test hypotheses that specify how and why certain empirical phenomena occur. Predictive models are aimed at predicting future observations with high accuracy. The distinction between explanatory and predictive models is key, as both types of models play a different yet essential role in advancing scientific research. A literature review of MISQ and ISR shows that predictive goals are scarce but growing in mainstream IS research. However, in most cases where the stated goal is predictive, explanatory modeling is employed. This practice is questionable, since often the best explanatory model is different from the best predictive model. In light of this distinction and current practice in IS, we highlight the main differences between predictive and explanatory modeling, focusing on practical issues that confront an empirical researcher in the data analysis process.

CONTRASTING PREDICTIVE AND EXPLANATORY MODELING IN IS RESEARCH

1. Introduction

Empirical research in information systems (IS), and in particular electronic commerce, has been constantly growing in the last years as researchers take advantage of large, high-quality, and publicly available datasets from websites such as Amazon, eBay, and Yahoo! (Bapna et al. 2006). However, despite prediction being a core scientific activity (Kaplan 1964; Dubin 1969), we show in this paper that an explanatory focus dominates mainstream empirical IS research and that even when the goal is predictive, it is often accompanied by explanatory methods instead of more appropriate predictive methods. By explanatory we mean that the purpose of the analysis is to test hypotheses that specify how and why certain empirical phenomena occur (Gregor 2006). Examples of explanatory goals that have been pursued in the IS literature are finding determinants of auction prices (Ariely & Simonson 2003); explaining the diffusion and non-diffusion of e-commerce among SMEs (Grandon & Pearson 2004); explaining attitudes towards online security and privacy (Malhotra et al. 2004); understanding the antecedents and consequences of online trust (Gefen et al. 2003) and explaining the impact of overlapping auctions (Jank & Shmueli 2007). In contrast to the proliferation of explanatory models, we will show later that there has been very little in the way of predictive modeling in mainstream IS journals. By predictive models we mean models that, instead of explaining existing phenomena, are aimed at predicting new/future observations with high accuracy. Examples are predicting the price of ongoing eBay auctions (Wang et al. 2008) or predicting future box-office sales based on online movie ratings (Dellarocas et al. 2006). In line with

Dubin (1969), we argue that predictive and explanatory models 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" (Robert Dubin, "Theory Building", 1969: p9)*

In statistical terms, the fundamental difference between explanatory and predictive models is the metric being optimized: whereas explanatory models seek to minimize bias in order to obtain the most accurate representation of the underlying theory, predictive models seek to minimize the sum of bias and variance, which can sometimes imply sacrificing theoretical accuracy in return for improved empirical precision (Hastie et al., pp. 197-200).

The contribution of this paper is threefold: first, the value of predictive modeling for both rigorous theory-building as well as for achieving practical relevance is underappreciated and second, although explanatory and predictive goals of theories are by no means mutually exclusive, they do require quite different data-analytic modeling strategies, and third, this distinction is often ignored in empirical IS research.

Although our main argument is true for any scientific discipline, there are two reasons why predictive goals (and models) are especially important in IS. First is its value for theory-building in fast-changing environments, such as the online environment that poses many challenges for the economic, psychological, and other theoretical models traditionally employed in IS. One example is auctions, where classical auction theory has only found limited applicability in the move from offline to online auctions, with online auctions raising many new theoretically and practically relevant questions that classical auction theory did not deal with (Seidmann et al. 2003; Bajari & Hortacsu 2004; Bapna et al. 2008). In this new era, where data are plentiful but theories are scarce, the data driven nature of predictive modeling can play a major role in theory-building. Predictive modeling can show new patterns and behaviors and help uncover potential new causal mechanisms, which in turn could lead to new theories being developed, provided the model is interpretable (cf. the discussion between Gurbaxani & Mendelson (1990, 1994) and Collopy et al. (1994)).

However, it would be wrong to see predictive modeling as merely a sophisticated approach to exploratory data analysis during the theory-building phase of research, as predictive modeling also has an important role to play in confirmatory data analysis during the theory-testing phase of research. Even in the presence of a properly specified explanatory model, high predictive power is not guaranteed, as the precision or magnitude of the causal effect might not be sufficient for obtaining levels of predictive accuracy that are practically meaningful. This is also important for practitioners, for whom accurately predicting future behavior of customers or competitors is more important than merely explaining past behavior without any reference to future behavior, since it is anticipated future behavior that guides managerial action. This leads to our second argument favoring the use of predictive models specifically for IS research: in a discipline plagued by debates regarding rigor and relevance, predictive modeling can serve as a statistically rigorous "reality check" to test the relevance of theories and the strength of explanatory causal models, thus providing a potential way out of the rigor-relevance conundrum.

In sum, predictive models have an important role to play in both novel as well as established theoretical environments. The question now becomes: is the value of predictive modeling, as well as the distinction from explanatory modeling, recognized in IS research?

2. Predictive Modeling in the IS literature

To investigate the extent to which predictive modeling is embedded within the current IS research, we conducted a literature search. Using EBSCO's Business Source Premier, we searched all fulltext articles in MIS Quarterly and Information Systems Research between 1990-2006 for one of the search terms "predictive OR predicting OR forecast*". Initial pre-testing of the search string revealed that expanding the search to use additional terms such as 'predict', 'prediction' or 'predictor' yielded many more hits, but none of the additional hits were relevant for our purposes, and the ones that were had already been uncovered by the more restrictive search terms. Every article was then checked to see whether or not the article had an explicit predictive goal and/or predictive claims were made about the statistical model. Articles that only used predictive language in a more generic sense (e.g. 'based on theory ABC, we are predicting that X will be associated with Y') or articles that were qualitative or purely theoretical were excluded. All remaining articles were subsequently checked for two criteria that are distinguishing features of all predictive models:

1. Is the model correctly being assessed with a specific measure of predictive accuracy (e.g., RMSE, MAPE and other measures computed from a holdout set)? Or is predictive accuracy incorrectly inferred from p-values or an explained variance measure such as R^2 ?
2. Is cross-validation or a holdout sample being used to test the predictive accuracy of the model?

Note that these are two necessary, but not sufficient criteria for a predictive model. In section 3 we go into more detail about the various properties of a predictive model that contrast with an explanatory model. If the article made predictive claims, yet did not satisfy both of these criteria, the article was classified as incorrectly predictive, i.e. explanatory. This is also where we depart from the otherwise very useful distinctions made by Gregor (2006): Whereas Gregor used only the stated goal of the article for purposes of classification, we use the additional criterion of the type of modeling actually employed. For this reason the majority of articles surveyed in Gregor were classified by her as being both explanatory and predictive, whereas according to our criteria almost all of these were purely explanatory.

The major findings from this literature study can be seen in Figure 1 and are twofold: first, although predictive claims for models are few, they are increasingly made, showing the growing awareness of the need for predictive models, yet, second, these predictive claims are often not accompanied by appropriate predictive modeling techniques. When examining each of the individual articles in detail, we found that almost 90% of articles that claimed predictive properties for their models, arrived at these claims by building and assessing their model using techniques appropriate for explanatory modeling instead of those appropriate for predictive models, with no discernible increase in correct application over time. Table 1 shows the counts of the overall search results and Table 2 in the appendix lists several illustrative examples and quotes from articles where explanatory modeling is used when one of the stated goals is predictive.

-------------- Table 1 around here --------------

-------------- Figure 1 around here --------------

We would like to emphasize that for many of these papers, had the goal been aimed solely at explaining without any claim to predict future observations, the method would have fit the goal and all would be fine. Yet when the stated goal is predictive, the method employed should be predictive instead of explanatory. Keil et al. (2000) provide a good illustration of this fit between goal and method: after validating an explanatory logistic regression model to test several factors that explain why some projects escalate and others do not, they go on to say: "To assess the predictive validity of each model, we examined its classification performance on both the estimation sample and a separate holdout sample" (Keil et al. (2000), p.653), which nicely illustrates the match between the goal of the model and the statistical method. They conclude from their predictive model "In summary, constructs derived from approach avoidance theory and agency theory perform well in classifying both escalated and non-escalated projects. On the other hand, constructs derived from self-justification theory and prospect theory perform well in classifying escalated projects, but do not perform well in their classification of non-escalated projects." (Keil et al. (2000), p.653). While the authors used this conclusion in the "Implications for Practice" section, we would argue that it also has important *theoretical* implications. Since the theoretical factors that predict escalation are different from the theoretical factors that predict non-escalation, explaining why a project *did* escalate will require a different theory compared to a theory explaining a project that *did not* escalate. Such a theoretical nuance was not easily available from the explanatory metrics derived from the logistic regression model, which illustrates our earlier point regarding the value that predictive modeling can have for theory-building.

The Keil et al. (2000) example illustrates that the distinction between explanatory models and predictive models is not trivial, and indeed this point holds more generally: although a good explanatory model will often exhibit some predictive power as well, the large 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 dataset (e.g. Stone 1974, Copas 1983; Hastie et al. 2001). In other words, an explanatory model may have poor predictive power, while a predictive model that might be based on the same original data would have higher predictive power. Most importantly, however, is that the modeling requirements can differ according to the task at hand. We therefore emphasize the importance of correctly specifying the modeling task and following the modeling process that corresponds to the task identified. It appears from the literature review that the distinction is under-appreciated, which not only leads to ambiguity in matching methods to goal, but at worst may result in incorrect conclusions for both theory and practice (e.g., Dawes 1979). Thus, we now turn to a more detailed look at the process of developing a predictive model vs. that of an explanatory model, highlighting the differences between the two.

3. Modeling Process

Determining the goal of the study upfront as either explanatory or predictive is essential to conducting adequate data analysis. Differences arise at each of the modeling steps, from the early stage of data collection and processing, through the choice of analysis methods, model selection, and final model usage (see Figure 1). In the following we describe differences at each step. We discuss the steps from last to first because differences at later steps motivate and affect issues at earlier stages.

-------------- Figure 2 around here --------------

3.1. Model Deployment

An explanatory model is used to support or refute an existing theory. The main concern is model misspecification and type I and II errors (Bayesians would consider a profit function related to such risks). In contrast, a predictive model is deployed by predicting new observations. The risk is a function of prediction inaccuracies and thus the main concern is of overfitting to the current dataset without taking into account the uncertainty associated with future data (Stone 1974, Copas 1983; Hastie et al. 2001). Picard and Cook (1984) refer to this as the "optimism principle", where "a model chosen via some selection process provides a much more optimistic explanation of the data used in its derivation than it does of other data that will arise in a similar fashion".

3.2. Model evaluation

A good predictive model is one that accurately predicts new data. A good explanatory model is one where the hypothesized model approximates the existing data well. These warrant different performance metrics. In explanatory modeling we use "goodness of fit" measures that measure closeness of the data to a pre-specified model. In contrast, predictive models are evaluated by their ability to predict new observations accurately. Three particular issues with evaluating predictive models, each closely tied to the specific purpose for which the model is deployed, are described next.

- *Theoretical Metrics vs. Empirical Performance:* Most textbooks describe R^2 as a measure of explanatory and predictive power of a regression model. However, since it is measured from the data that were used to fit the regression model, R^2 is really an explanatory metric and is "over-optimistic" in measuring predictive accuracy. The same is true for other metrics computed from the data to which the model was fit. Zheng & Agresti (2000) describe three types of "measures of predictive power": those based on residuals, on a variation function, and likelihood-based measures. When such metrics are computed from the data to which the model was fitted (as is typically done), they provide measures of goodness of fit. Although theoretically they might be indicators of predictive power, in practice they are over-optimistic due to shrinkage: "Testing the procedure on the data that gave it birth is almost certain to overestimate performance" (Mosteller & Tukey 1977).
- *Predicting the Top Tier:* A special type of predictive goal, particularly common in marketing and personnel psychology, is predicting the top tier of a population in terms of a measurement of interest. Examples would be identifying the 10% of customers with the highest chance of responding to a direct mailing, selecting the top 5 students for admission to graduate school or selecting the best applicant for a job. IS examples are identifying customers most likely to switch purchase channel or users most likely to benefit from adopting a new technology. A good model

here is one that correctly scores the top tier, while the remaining predictions do not matter. Performance is therefore measured directly with respect to this top tier, with the most popular tool being the lift chart (for details on constructing a lift chart and illustrations see Shmueli et al. 2007). Note that due to its focus on a particular segment of the data, a model with good lift need not necessarily exhibit high overall predictive accuracy.

- *Costs:* Costs play a major role mainly in predictive tasks. Often there are costs associated with predictive inaccuracy, which tend to be asymmetric (e.g., they are heftier for some types of errors than for others). For example, the costs of erroneously selecting a low-quality applicant for a job are likely to be much higher than the (opportunity) costs of failing to select the single best applicant. A good model in this context is one that minimizes costs, which need not coincide with the model with highest predictive accuracy per se. In some cases, and especially when a decision theoretic approach is taken, costs can be integrated into explanatory models as well. In such cases, the performance metric to consider is a cost function rather than ordinary goodness-of-fit.

3.3. Model Selection

The different goals of explanatory and predictive models affects the function to optimize: In explanatory modeling the focus is on minimizing the bias (i.e., the specification error), whereas in predictive modeling we minimize the combined bias and variance. Large variance is associated with low predictive accuracy (Hastie et al. 2001), and therefore a key approach for improving predictive accuracy is to tolerate some bias if the gain in variance reduction is large. In Appendix A we show an example where an under-specified regression model achieves higher predictive accuracy than a correctly specified model. This bias-variance balance means that predictive models tend to be simpler and smaller ("Typically the more complex we make the model, the lower the bias but the higher the variance", Hastie et al. 2001), although in some cases predictive models are more complicated in order to capture small nuances that improve predictive accuracy (Breiman, 2001). Because explanatory models are primarily concerned with model misspecification, the process of model selection for explanatory purposes is aimed at reducing bias by removing input variables with statistically insignificant coefficients. In contrast, model selection in predictive modeling might lead to shrinking or setting some coefficients to zero, thereby removing inputs with small coefficients, even if they are statistically significant (Wu et al. 2007).

Finally, the treatment of multicollinearity is different: Whereas in explanatory models the inflated standard errors hinder the possibility of testing hypotheses regarding model parameters, thus leading to a variety of strategies for identifying and reducing multicollinearity, for predictive purposes "multicollinearity is not quite as damning" (Vaughan & Berry 2005). As Aczel & Sounderpandian (2006) explain: "Even though individual regression parameters may be poorly estimated when collinearity exists, the combination of all regression coefficients in the regression may, in some cases, be estimated with sufficient accuracy that satisfactory predictions are possible. In such cases, however, we must be very careful to predict values of Y only within the range of the X variables…[otherwise] large error may result".

3.4. Choice of method(s)

The goal of explanatory models is to shed light on a hypothesized causal relationship between an outcome and a set of inputs. The fitted model should therefore be interpretable as well as provide insight about the importance of each of the inputs. For this reason, regression-type models are popular in explanatory modeling: they provide for each input a coefficient (with a sign and magnitude) and an associated p-value for ranking their importance. In contrast, for predictive tasks, this order of priority is reversed as the focus is first on accurately predicting new empirical observations and second on illuminating the underlying relationship between the output and set of inputs (although obviously an interpretable model will be preferred over an uninterpretable model that has the same predictive accuracy). For this reason, predictive modeling can also usefully employ quite simple models such as k-nearest neighbors or even 'black-box' models such as neural networks, even though the lack of interpretability makes them practically absent from explanatory modeling. The difference in the importance of transparency vs. predictive accuracy leads to a divergence when it comes to the choice of method: In general, data-driven algorithms are widely used and acceptable in predictive modeling, whereas they are much rarer in explanatory modeling. Data-driven methods range from simple algorithms such as k-nearest neighbors to more 'black-box' methods such as neural networks or ensembles of various methods. Some data-driven methods, which have a high level of transparency, are in fact useful for explanatory tasks. An example is classification and regression trees. However, for predictive purposes, an ensemble of such trees (called a "random forest", Breiman 2001a) can provide higher prediction accuracy, thereby sacrificing some interpretability in return for higher practical usefulness. An important aspect of data-driven methods is that they capture local behavior, i.e. behavior for a specific region of the dataset. Whereas model-driven methods require the specification of the exact type of relationship (via interaction terms, etc.), data-driven methods (and for that purpose, non-parametric methods) tend to be more flexible and can capture patterns over a wide range on the global-local spectrum. In many cases, and especially as theory becomes scarce, discovering local "pockets" of patterns can be very useful, even if not initially interpretable, as they can lead to new theories.

The usefulness of data-driven methods for predictive modeling is also related to issues of data size. Because data-driven methods learn "everything" from the data, they usually require much larger datasets than model-driven methods in order to capture relationships that can then be used for prediction. In contrast, the size of the data required for explanatory modeling is driven by considerations of statistical power. Although in theory more data would lead to better explanatory models, with very large datasets, explanatory models can backfire, as they will yield statistical significance that is driven solely by sample size and too high to be practical (Shmueli, 2008). Standard performance metrics based on p-values are of no use in that case. Examining effect sizes is useful, but there is no general guideline to what constitutes a sound model and when overfitting is taking place. It is precisely here where assessing the predictive power of explanatory models can assist in avoiding overfitting and lead to better theories.

3.5. Choice of variables

There are several aspects related to the choice of the inputs to include in the model, their role, and the form in which they are included:

- *Retrospective/prospective availability:* A fundamental requirement of a predictive model is that the input information should be available at the time of prediction, i.e. only include ex-ante variables. In contrast, no such requirement is necessary in explanatory modeling, and many explanatory models include ex-post input variables. An example would be explaining the final price in an eBay auction based on the total number of bidders. This information is only available at the end of the auction, so a predictive model of eBay auction prices cannot include the total number of bidders (Wang et al. 2008)[1]. Note that the "best" predictive model will not necessarily be the same as the "best" explanatory model without the ex-post variables.
- *Causal input variables vs. proxies:* Explanatory models tend to be based on a theoretical causal relationship, and thus the choice of inputs is driven by causal arguments. In contrast, in predictive modeling inputs are not required to be causing the output, but rather associated with

[1] When the model is to be deployed halfway through the auction, then the number of bidders up until that point can be used in building the model, since that information is available at the time the prediction is made.

it. We can therefore use proxies and even confounding variables for prediction modeling, which can be particularly beneficial if those variables can be measured with greater precision compared to the 'proper' causal variable.

3.6. Data preprocessing

Initial data preprocessing steps involve data manipulation, summarization, and visualization. We point out two manipulations that differ in explanatory vs. predictive tasks.

- *Missing Values:* Missing values require determining the extent and type of missingness, and choosing a course of action accordingly. First, in predictive tasks, if the data to be predicted have missing inputs, data imputation is a necessity, whereas in explanatory modeling often a plausible solution is to drop the missing records. Second, in explanatory modeling the type of imputation depends on whether the data are Missing-At-Random, Missing-Completely-At-Random, or Not-Missing-At-Random (Little & Rubin 2002) whereas in predictive modeling this distinction is not important (Sarle 1998), but rather what is important is whether the 'missingness' depends on the dependent variable or not (Ding & Simonoff 2006). Also, Sarle (1998) compares a set of imputation methods and shows that the methods most useful for explanatory modeling are either inappropriate or not useful for predictive modeling, again highlighting the distinction between the two modeling strategies. Finally, Saar-Tsechansky and Provost (2007) compare different approaches for dealing with missingness in data to be predicted, and show that an alternative to imputation leads to best predictions: building multiple "reduced models", each based on the subsets of the non-missing predictors for the set of observations to be predicted. This approach means that different reduced models are created for different observations. Obviously, this approach supports predictive tasks but not explanatory tasks.

- *Data Partitioning:* A popular solution for avoiding over-optimistic predictive accuracy is to evaluate performance not on the training set, i.e. data used to build the model, but rather on a holdout sample which the model "did not see". The creation of a holdout sample can be achieved in various ways, the most commonly used being a random partition of the sample into training and holdout sets. A popular alternative, especially when the data are scarce, is cross-validation (which is more computationally intensive, but avoids "bad partitions"). Another approach to partitioning was suggested by Snee (1975) who proposed an algorithm for creating partitions that are as similar as possible. With a large dataset the reduction in sample size for the training set will not be substantial and in today's data environment, with large datasets becoming the rule rather than the exception, data partitioning is now common practice in predictive modeling. Another use of data partitioning in predictive modeling (or in general when large datasets are available) is that it allows the modeler to relax assumptions about error distributions. For explanatory models data partitioning is less commonly used, but it can be used there as well for robustness checking and more so to strengthen model validity by showing its predictive power. Finally, although data partitioning is most suited to large dataset, alternatives exist specifically for small datasets, such as cross-validation or resampling methods (e.g. bootstrap). Thus predictive modeling is not limited to large datasets and can also fruitfully be used in small datasets.

3.7. *Data Collection and Study Design*

- *Fixed/random effects:* In hierarchical designs an important distinction for purposes of inference is fixed and random effects. A categorical variable is considered a fixed effect if the categories that are present in the data are the complete set of categories of interest for purpose of inference. In contrast, if the categories present in the data are just a sample of the categories to which the inference should apply, the variable is treated as a random effect. If we consider a 2-level model, such as students within schools, then at the design stage the researcher must determine the number of schools and the number of students within each school for which to collect data. Generally, for purposes of inference it is better to increase the number of groups (=schools) at the expense of the number of group size (=students). Afshartous and de Leeuw (2005) show that when the goal is predictive rather than explanatory, the exact opposite strategy leads to more accurate predictions: Increasing group size at the expense of the number of groups. They explain this effect as "the manifestation of the dangers of using a grand mean to predict at the individual level". The authors further find that the popular "prior prediction" estimation method which takes into account the hierarchical structure is inferior to a simpler approach that models each group separately (thereby ignoring the hierarchical structure) when it comes to prediction.

3.8.Example: Using Regression Models

In light of the distinctions described above, it is clear that almost every aspect of the modeling process is different depending on whether the goal is explanatory or predictive. To illustrate this in a setting that is likely to be familiar to most empirical researchers, consider the use of regression models, as these models can be used in both explanatory and predictive modeling tasks. Another commonality is that in both cases estimation is usually performed in the same way (i.e., via maximum likelihood or least squares). However, there are several important differences in the modeling process that are likely to lead to different final models. These differences affect the process from its start (data preprocessing and choice of variables), through assessing performance, model selection, and finally model choice and use. The similarities and differences between using a regression model for explanation or prediction are summarized in Table 3.

-------------- Table 3 around here --------------

4. Conclusions and Implications

Our literature survey indicates the dominance of explanatory modeling in the two top IS journals, with hardly any predictive models being published in MISQ and ISR, despite the value that predictive models have for advancing scientific theory in general and IS research in particular. Predictive models play an important role in assessing practical relevance of existing theories, and quantifying the level of predictability of phenomena (Ehrenberg & Bound, 1993). However, the benefits of predictive models are not just on the relevance side, but also on the rigor side: predictive models can -in a methodologically rigorous fashion- highlight new phenomena that can serve as a trigger for further theorizing. As the Keil et al. (2000) example earlier showed, their predictive metrics yielded more nuanced theoretical understanding than was possible solely from the explanatory metrics. Thus, the two types of models are complements rather than substitutes.

There are a few possible explanations for the current lack of predictive modeling in mainstream IS research. One reason might be that the conflation of explanatory and predictive modeling has strong roots in the early philosophy of science literature, particularly the influential hypothetico-deductive model associated with Hempel and Oppenheim (1948), who explicitly equated prediction and explanation. However, as later became clear, the type of uncertainty associated with explanation is of different nature than that associated with prediction (Helmer & Rescher 1959), which necessitated the need for developing models geared specifically towards dealing with predicting future events and trends. As statistical theory progressed, particularly in the area of model selection and the associated concept of overfitting, the distinction between the two classes of models has been further elaborated (Forster 1994, Forster 2002, Sober 2002, Hitchcock & Sober 2004) and is currently accepted in the statistics literature, although the relative merits of each class of models continues to be hotly debated (e.g. Breiman (2001) and the commentaries following it).

The lack of predictive modeling in IS might also be indicative of a research community where performance is not primarily measured by practical value. Although explanatory modeling might sound more academic and indeed some argue that "The two goals in analyzing data … I prefer to describe as "management" and "science". Management seeks profit... Science seeks truth" (Parzen, 2001), we disagree: there *is* an important place in academia for predictive modeling. Designating academia as the "explainers" and leaving the prediction to industry does not enhance the field. Growth of the research community in predictive directions will bring academic work closer to industry research, thus increasing its relevance with predictive models serving as a rigorous reality check on explanatory models, which in turn can lead to better theories. Predictive modeling can thus serve to achieve both rigor and relevance of our theories.

A final explanation may be that IS researchers have simply been unaware of the distinctions between explanatory and predictive modeling, since they have never been highlighted. With this paper, we hope to have rectified this. One implication is therefore a need to integrate predictive modeling into the IS curriculum and to highlight the differences and uses of predictive and explanatory models. Moving towards predictive modeling is another step in the direction of empirically rigorous and relevant research. As Kaplan put it: *"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" (1964, p.350).*

References

Aczel AD and Sounderpandian J. 2006. *Complete Business Statistics*, McGraw-Hill Irwin, 6th edition.

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

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

Bapna R, Goes P, Gopal R, Marsden J R. 2006. Moving from Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data. *Statistical Science* **21**(2): 116-130.

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.

Breiman L. 2001. Statistical modeling: the two cultures. Statistical Science 16: 199-215

Breiman L. 2001a. Random forests. *Machine Learning Journal.* **45:** 5-32.

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

Collopy F, Adya M, Armstrong JS. 1994. Principles for examining predictive-validity - the case of information-systems spending forecasts. *IS Research* **5**(2): 170-179

Dalkey N, Helmer O.1963. An experimental application of the Delphi method to the use of experts. *Management Science,* **9**(3): 458-467

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

Dellarocas C, Awad NF, Zhang X. 2006. Exploring the value of online product ratings in revenue forecasting: the case of motion pictures. *Working paper*, University of Maryland

Ding Y, Simonoff JS. 2006. An investigation of missing data methods for classification trees. *Working Paper SOR-2006-3, Statistics Group NYU*

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

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

Forster M, 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

Gefen D, Karahanna E, Straub DW. 2003. Trust and TAM in online shopping: An integrated model. *MIS Quarterly* **27**(1): 51-90

Grandon EE, 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, Mendelson H. 1990. An integrative model of IS spending growth. *IS Research.* **1**(1), 23-46

Gurbaxani V, Mendelson H. 1994. Modeling vs forecasting - the case of information-systems spending. *IS Research* **5**(2): 180-190

Hastie T, Tibshirani R, Friedman JH. 2001. *The elements of statistical learning: data mining, inference, and prediction*, Springer.

Helmer O, Rescher N. 1959. On the epistemology of the inexact sciences. *Management Science,* 5(June), 25-52

Hempel C, Oppenheim P. 1948. Studies in the logic of explanation, *Philosophy of Science* **15**:35-175

Hitchcock C, Sober E. 2004. Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science* **55**(1):1-34

Jank W, Shmueli G. 2007. Modeling concurrency of events in online auctions via spatio-temporal semiparametric models. *Journal of Royal Statistical Society, Series C*, **60**(1): 1-27

Kaplan A. 1964 *The conduct of inquiry: methodology for behavioral science*. Chandler Publishing, New York, NY.

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.

Kutner MH, Nachtsheim CJ and Neter J. 1994. *Applied Linear Regression Models,* McGraw-Hill, 4th ed.

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

Malhotra NK, Kim SS, and Agarwal J. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *IS Research* **15**(4): 336-355

Montgomery DC, Peck EA, Vining GG. 2001. *Introduction to Linear Regression Analysis*, 3rd Edition, Wiley.

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

Pavlou PA, Fygenson M. 2006. Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS Quarterly* **30**(1): 115-143

Parzen, E. (2001), Comment on "Statistical Modeling: The Two Cultures", *Statistical Science*, **16**(3), 224-226.

Perlich C, Provost F, Simonoff JS. 2003. Tree induction vs. logistic regression: a learning-curve analysis. *Journal of Machine Learning Research,* **4**:211-255

Picard RR, Cook RD. 1984. Cross-validation of regression models. *Journal of the American Statistical Association*, **79**(387): 575-583

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

Sarle WS. 1998. Prediction with missing inputs, in Wang PP (ed.), *JCIS 98 Proceedings,* **II** Research Triangle Park, NC, 399-402

Shmueli G, Patel NR, Bruce PC. 2007. *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, John Wiley & Sons.

Shmueli G. 2008. Statistical Inference with Large (eCommerce) Datasets. *Statistical Challenges in eCommerce Research (SCECR), NYU.*

Snee RD. 1975. Validation of Regression Models: Methods and Examples. *Technometrics* **19**(4): 415-428.

Stone, M. 1974 Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, **36:** 111-147.

Vaughn TS, 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, Shmueli G. 2007. Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics*, in press

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

Zheng B, Agresti A. 2000. Summarizing the predictive power of a generalized linear model. *Statistics in Medicine,* **19**: 1771-1781

Appendix A: The Bias-Variance Tradeoff

In the following we show the breakdown of the prediction error into three terms: bias, variance, and unexplained variability. We then show that in some cases a mis-specified model will have higher predictive accuracy than a correctly specified model.

We use the following notation:

Y	Observed output
$f(x)$	True model
$f^*(x)$	Incorrect model specification
$\hat{f}^*(x)$	Estimated model (based on incorrect model specification)

Under a quadratic loss function, the prediction error can be broken down as follows:

$$\begin{aligned} MSE = E(Y - \hat{Y}|x)^2 &= E(Y - \hat{f}^*(x))^2 \\ &= E(Y - f(x) + f(x) - f^*(x) + f^*(x) - \hat{f}^*(x))^2 \\ &= E(Y - f(x))^2 + (f^*(x) - f(x))^2 + E(\hat{f}^*(x) - f^*(x))^2 \\ &= Var(Y) + Bias^2 + Var(\hat{f}^*(x)) \end{aligned}$$

The bias is the result of misspecifying the true model; The third term is the estimation/sampling variance, which is the result of using a sample to estimate the model; The first term is the error that results even if the model is correctly specified and accurately estimated. Note that the bias and variance both decrease as the sample size increases, but not *Var(Y)*. However, model complexity leads to a tradeoff between bias and variance, with more complex models reducing bias but increasing variance (due to estimation error).

To illustrate the difference between an explanatory and predictive model, consider a true model with two predictors (and no constant) of the form

$$f(x) = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $x = [x_1, x_2]$ and *Var(ε)* = σ^2. If the estimated model is correctly specified its bias is 0 and the sampling variance is given by (Kutner et al. 2004, p. 229)

$$Var(x_1\hat{\beta}_1 + x_2\hat{\beta}_2) = \sigma^2 x'(X'X)^{-1}x$$

where X is the design matrix that includes two columns $X = [\underline{x}_1, \underline{x}_2]$. The MSE of prediction for the estimated correctly specified model is therefore given by

$$MSE_1 = E(Y - f(x))^2 = \sigma^2 + 0 + Var(x_1\hat{\beta}_1 + x_2\hat{\beta}_2) = \sigma^2\left[1 + x'(X'X)^{-1}x\right]$$

In comparison, consider the estimated under-specified form

$$\hat{f}*(x) = \hat{\gamma}_1 x_1$$

The bias of this model is given by (Montgomery et al. 2001, pp. 294-5)

$$Bias = x_1\gamma_1 - (x_1\beta_1 + x_2\beta_2) = x_1(x_1'x_1)^{-1}x_1'x_2\beta_2 - x_2\beta_2$$

And the sampling variance is given by

$$Var(X_1\hat{\gamma}_1) = \sigma^2 x_1(x_1'x_1)^{-1}x_1$$

The MSE of prediction for the estimated under-specified model is therefore given by:

$$MSE_2 = E(Y - \hat{f}*(x))^2 = \left(x_1(x_1'x_1)^{-1}x_1'x_2\beta_2 - x_2\beta_2\right)^2 + \sigma^2\left[1 + x_1(x_1'x_1)^{-1}x_1\right]$$

Although the bias of the under-specified model is larger than that of the correctly specified model, its variance can be smaller, and in some cases so small that the overall MSE will be lower for the under-specified model. Wu et al. (2007) enumerate four conditions when an underspecified linear regression model will have higher prediction accuracy than the correctly specified model:

1. When the data are very noisy (large σ);
2. When the true absolute values of the left-out parameters (in our example β_2) are small;
3. When the predictors are highly correlated; and

4. When the sample size is small or the range of left-out variables (here, X_2) is small.

The above example compares two linear regression models. However, the bias/variance tradeoff becomes even more pronounced when comparing across linear and non-linear methods and thus the study of variance reduction techniques such as ridge regression, principal components regression, and ensemble methods has been at the core of predictive modeling. The general principle is that "the more complex we make the [estimated] model, the lower the squared bias but the higher the variance" (Hastie et al., 2001, p. 197).

Figure 1: Literature search of ISR and MISQ, 1990-2006: Number of papers with stated predictive goals (top line); Number of papers with stated predictive goal and adequate predictive modeling (bottom line).

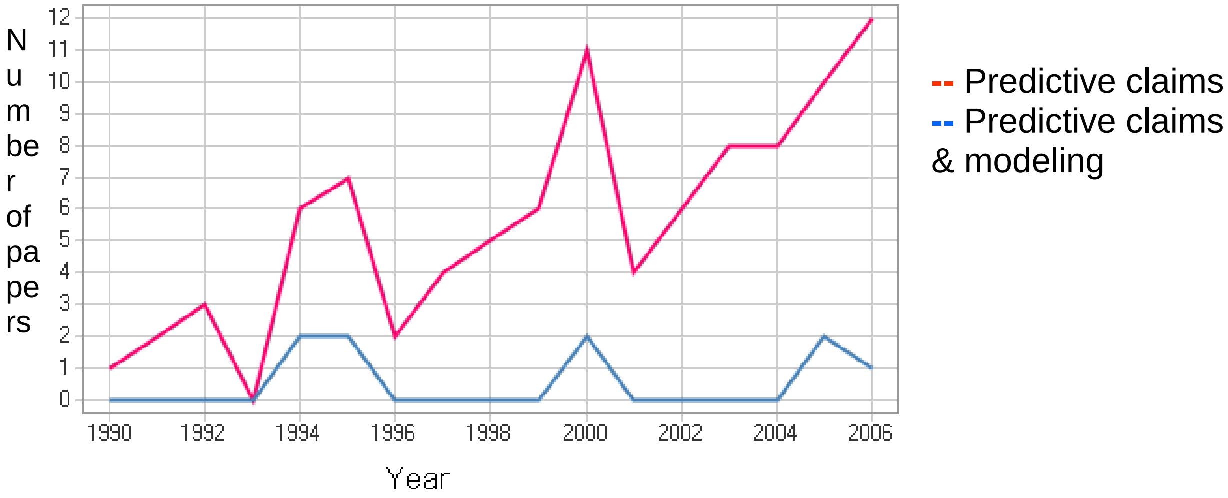

Figure 2: Main steps in the data modeling process

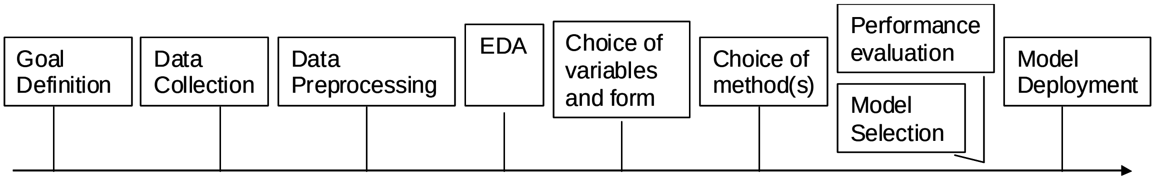

Table 1: Summary statistics of literature review

	Total	**MISQ**	**ISR**
Initial hits 1990-2006	243	144	99
Predictive aspect tangential	158	91	67
Relevant sample	85	53	32
Of which correctly predictive	9 (11%)	5 (9%)	4 (13%)
Of which incorrectly predictive (i.e. explanatory)	76 (89%)	48 (91%)	28 (87%)

Table 2: 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 R2 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)*

Table 3: Fitting a regression model: Comparing explanatory and predictive modeling

Operation	**Explanatory Task**	**Predictive Task**
Types of models	Linear, logistic, probit, etc.	Same + shrinkage models (ridge regression, PCR, ensemble methods)
Choice of independent variables (X)	Based on theory/hypotheses; causal relationship assumed (with Y)	based on association; availability at time of prediction; measurement accuracy important
Data preprocessing	Visualization, summaries, outlier detection, imputation	Same, except missing data are treated differently
Data partitioning (training/holdout)	Not typical, except for robustness testing	Always required
Software	Any statistical software (as simple as Excel)	Ordinary software requires tweaking (data partitioning, performance metrics); or data mining software (Clementine, SAS EM, XLMiner)
Estimation method	Maximum likelihood	Same
Model selection goal	Determine important factors	Dimension reduction, parsimony
Model selection methods	Stepwise, forward, etc.	Same
Multicollinearity	A serious danger, risk of incorrect inference	Not too important
Evaluation criteria	Theoretical justification, goodness of fit, statistical significance	Parsimony, predictive accuracy, costs, practical deployment
Performance metrics	R^2, MSE, residual analysis, coefficient and overall p-values	Predictive accuracy (RMSE, MAPE, lift) computed from holdout dataset
Dangers	Model misspecification, type I and II errors	Over-fitting
Model use (research)	Test hypotheses/theory	Discover new relationships, evaluate magnitude of effects, assess predictability
Model use (practice)	Determine important factors	Score new data