Predictive vs. 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

Abstract

Explanatory models test hypotheses that specify how and why certain empirical phenomena occur. Predictive models are aimed at predicting the future or new observations with high accuracy. A literature review of MISQ and ISR shows that predictive modeling is scarce in mainstream IS research. We also find that although in some cases the stated goal is predictive, the modeling is explanatory. We therefore describe the main differences between predictive and explanatory modeling, focusing on practical issues that confront an empirical researcher in the data modeling process.

"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)

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!. However, despite prediction being a core scientific activity, we show in this paper that an explanatory focus

dominates mainstream empirical IS research and that 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, there has been very little in the way of predictive modeling in mainstream IS journals, as we will show later. By predictive models we mean models that, instead of explaining existing phenomena, are aimed at predicting the future or new observations with high accuracy. Examples are predicting the price of ongoing eBay auctions (Wang et al. 2007) or predicting future box-office sales based on online movie ratings (Dellarocas et al. 2006). The key message of this paper is twofold: 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, yet this distinction is often ignored in empirical IS research.

This conflation of explanatory and predictive modeling has its 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 (e.g. the Delphi method, Dalkey & Helmer (1963)). 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).

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 could not or did not deal with. In this new era, where data are plentiful but theories are scarce, 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)).

The second reason is that predictive modeling provides a way out of the rigor-relevance conundrum. For practitioners, 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 action. Even in the presence of a properly specified explanatory model, high predictive power is not guaranteed, as the magnitude of the causal effect might not be sufficient for obtaining practically-useful levels of predictive accuracy. Thus, predictive modeling can serve as a statistically rigorous "reality check" to test the relevance of theories and the strength of explanatory causal models.

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 model. Articles that 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: first, the model is being assessed with a specific measure of predictive

accuracy (e.g., RMSE, MAPE and other measures computed from a holdout set) instead of p-values or an explained variance measure such as R^2 and second, the use of cross-validation or a holdout sample to test the predictive accuracy of the model (note that these are two necessary, but not sufficient criteria, section 3 goes into more detail about the various properties of a predictive 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 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 are classified as being both explanatory and predictive, whereas according to our criteria almost all are in fact purely explanatory.

Our major finding was that although predictive claims for models are frequently made (showing the increasing concern and awareness to the need for predictive models) they 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 (see Table 1 for the overall search results and see Table 2 in the appendix for a collection of examples and quotes from the literature where explanatory modeling is used when one of the stated goals is predictive).

---------- Table 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 situations, the method would have fit the

goal. Yet when the stated goal is predictive, the method employed should be predictive instead of explanatory. Keil et al. (2000) provide a good illustration: 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" (p.653), which nicely illustrates the match between the goal of the model and the statistical method. Keil et al. (2000) 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." (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 factors that predict escalation are different from the theoretical factors that predict non-escalation, our explanation of why a project did escalate will require a different theory compared to 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 the value that predictive modeling can have for theory-building.

The distinction between explanatory models and predictive models is not trivial: 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. Hitchcock & Sober, 2004). In other words, an explanatory model might 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 the modeling process that corresponds to the task identified. It appears from the literature review that the distinction is under-appreciated, which leads to ambiguity in matching methods to goal. We thus 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 1 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 over-fitting.

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 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 are described next.

- *Theoretical Metrics vs. Empirical Performance:* Most textbooks describe R^2 as a measure of predictive power of a regression model. Zheng & Agresti (2000) describe three types of "measures of predictive power": those based on residuals, on a variation function, and likelihood-based measures. Common to all such metrics is that they are computed from the data to which the model was fitted. Although theoretically they might be indicators of predictive power, in practice they are over-optimistic: "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, common in marketing, is predicting the top tier of a population in terms of a measurement of interest. An example is identifying the 10% of customers with the highest chance of responding to a direct mailing. 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. Note that 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). A good model in this context is one that minimizes costs, but it need not coincide with the model with highest predictive accuracy. In some cases, and especially when a decision theoretic approach is taken, costs are integrated into explanatory

models. 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. 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). Because explanatory models are primarily concerned with model misspecification, the process of model selection 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).

A related factor is the use of automated model selection methods (e.g., stepwise) and the extra level of bias that it introduces into performance metrics. Picard & Cook (1984) study 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". They describe simulations by Berk (1978a) that show a 20% bias in estimating σ^2 in regression models chosen via stepwise algorithms, as well as biases in measures such as R^2, Cp, and PRESS.

Finally, the treatment of multicollinearity is different: Whereas in explanatory models the inflated standard errors hinder the possibility of testing hypotheses regarding model parameters, for predictive purposes "multicollinearity is not quite as damning" (Vaughan & Berry 2005).

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, whether the underlying relationship between the output and set of inputs becomes clearer through the model is not detrimental. For this reason 'black-box' models such as neural networks, or even simple algorithms such as Naive Bayes or k-nearest neighbors tend to be very useful in predictive modeling, but practically absent from explanatory modeling.

- *Data size and signal-to-noise ratio:* The choice of analysis method depends on considerations of data size, data structure, and the signal-to-noise ratio in the data. Generally, data-driven methods require much larger datasets than model-driven methods, because they learn "everything" from the data. Perlich et al. (2003) compare a logistic regression model to a classification tree and find that "logistic regression is better for smaller training sets and tree induction for larger data sets." They also find that a logistic regression model outperforms a classification tree when the signal is weak, but under-performs in the presence of a strong signal. In explanatory models, a sufficient amount of observations is required to achieve statistical power. But the opposite occurs with very large datasets which yield statistical significance that is too high to be practical. Standard performance metrics based on

p-values are then of no use. Examining effect sizes is useful, but there is no general guideline to what constitutes a sound model and when over-fitting is taking place. It is precisely here where predictive power of explanatory models can assist in avoiding over-fitting.

- *Global vs. local structure:* Model-driven methods (e.g., linear regression) define a global model over the entire range of the data. To capture local behavior with such methods, the local area and the exact type of relationship must be specified (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.
- *Model Complexity:* Due to the bias-variance tradeoff of the model error, many machine-learning methods have superior predictive accuracy because they introduce some bias, but reduce the variance. Resampling and ensemble methods are examples.

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. In contrast, no such requirement is necessary in explanatory modeling, and many explanatory models include ex-post input variables. 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 it. We can therefore use proxies and even intervening variables for prediction modeling.

- *Fixed/random effects:* Treating factors as fixed effects in predictive modeling is only feasible if the observations to be predicted fall within the fixed levels.

3.6. Data preprocessing

An initial data preprocessing step involves 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 or Missing-Completely-At-Random, whereas in predictive modeling this distinction is not important (Sarle 1998), but rather whether the missingness depends on Y (Ding & Simonoff 2006). Finally, Sarle (1998) compares a set of imputation methods and shows that those most useful for explanatory modeling are either inappropriate or not useful for predictive modeling.
- *Data Partitioning:* A popular solution for avoiding over-optimistic predictive accuracy is to evaluate performance not on the data used to build the model but on a holdout sample which the model "did not see". Picard & Cook (1984} note that "The concept of splitting the data into two parts… is by no means new. Historical background is provided by Stone (1974, 1978) and Geisser (1975), who also present their own viewpoints on assessment of predictive ability." In today's data environment, large datasets are the rule rather than the exception and therefore data partitioning is common practice in data mining. With a large dataset the

reduction in sample size for the training set will not be substantial. The practice of data partitioning is especially important in predictive modeling. For explanatory models it can be used for robustness checking and more so to strengthen model validity by showing its predictive power. 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. Finally, alternatives to data partitioning for small datasets are cross-validation or resampling methods (e.g. bootstrap). Thus predictive modeling is not limited to large datasets.

3.7.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. These models can be used in both explanatory and predictive modeling. 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

Our literature survey indicates the dominance of explanatory modeling in the two top IS journals, with hardly any predictive models being published (with the exception of data mining applications that appear in specialized journals as opposed to general IS journals), despite the value that predictive models have for theory-building and practical relevance. This imbalance might be indicative of an early stage of the field, where researchers are simply trying to make sense out of the new online environment using 'offline' theories (although similar imbalances also exist in established fields such as economics and psychology). It might also be indicative of a research community where performance is not measured by profits. When it comes to industry, there is often more predictive work performed in an attempt to create predictive applications that yield profit. Although explanatory modeling might sound more academic, there is an important place for predictive modeling (in the marketing literature, for instance, predictive modeling is much more prevalent). The gap between industry and academia need not be as large as it is when it comes to 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 reality check on explanatory models. However, the benefits are not just on the relevance side, but also on the rigor side: predictive models can also highlight in a methodologically rigorous fashion new phenomena that can serve as a trigger for further theorizing. As the Keil et al. (2000) example 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, although the balance between the two needs to be redressed. Moving towards predictive modeling is another step in the direction of empirically rigorous and relevant research, because 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

Ariely D, 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

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

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

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.

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.

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

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

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

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

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

Figure 1: Main steps in the data modeling process

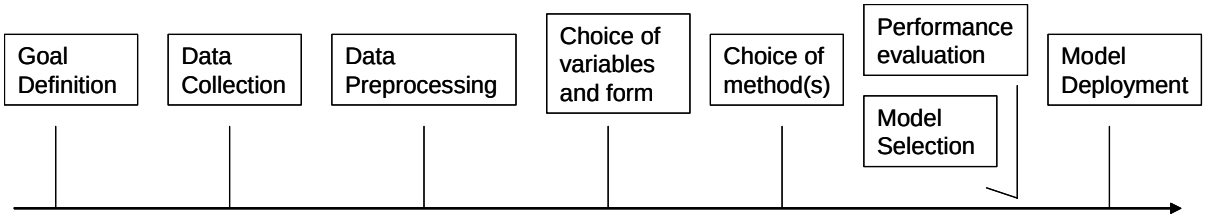

Table 1: Summary statistics of literature review

	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
Choice of independent variables (X)	Based on theory/hypotheses; causal relationship assumed (with Y)	based on association; availability at time of prediction
Data preprocessing	Visualization, summaries, outlier detection, imputation	Same (but impute differently)
Data partitioning (training/holdout)	Not typical, except for robustness testing	Always required
Data size considerations	Number of variables, model complexity; too much data problematic	Number of variables, model complexity, data partitioning; never too much data
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 inflated standard errors	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
Model use (practice)	Determine important factors	Score new data