MANAGEMENT SCIENCE Vol. 36, No. 8, August 1990 Printed in U.S.A. DATABASE MODELS AND MANAGERIAL INTUITION: 50% MODEL + 50% MANAGER\*

ROBERT C. BLATTBERG AND STEPHEN J. HOCH University of Chicago, Graduate School of Business, 11 F1E . 58th Street, Chicago, Illinois 60637 We focus on ways of combining simple database models with managerial intuition. We present a model and method for isolating managerial intuition. For five different business forecasting situations, our results indicate that a combination of model and manager always outperforms either of these decision inputs in isolation, an average R2 increase of 0.09 (16%) above the best single decision input in cross-validated model analyses. We assess the validity of an equal weighting heuristic, 50% model + 50% manager, and then discuss why our results might differ from previous research on expert judgment. (FORECASTING; DECISION MAKING; EXPERTISE; DECISION SUPPORT SYSTEMS)

Introduction Everybody's got so much information all day long that they lose their common sense. (Ger-trude Stein) A very important problem in sales forecasting is combining the wisdom of experienced businessmen with statistical analysis. (Lorie 1957)

With the availability of more and better sources of data, decision makers must begin to systematically incorporate such information into the decision process. At present, however, we know much more about building and processing databases than about how experts might use such data to improve decision quality. We show how firms can take advantage of the data explosion by combining simple database models with managerial intuition, viewing the two decision inputs in combination rather than in competition. We identify differences between models and intuition, and develop a way to isolate and evaluate intuition. We analyze several on-line business forecasting situations and show that decision quality could have been improved dramatically by relying on both models and intuition. We find that a 50% Model + 50% Manager heuristic improves forecast quality. The essential message is that both statistical and human inputs should guide final decisions, at least given the current state of model building and the difficulty in quantifyinig expert intuition. The paper is applicable to a large constituency in the business community. Forecasters can use "econometric" models effectively only if they have a built-in adjustment mech-anism to capture the changing environment. We argue that managerial forecasts can fill this role and improve predictive accuracy. Thus we will not simply conclude that com-bination rules are best. This is of course the overwhelming conclusion reached in the forecasting literature (Clemen 1989), though most of the empirical work in that literature has focused on a model-model (Granger and Ramanathan 1984) or expert-expert (Ashton and Ashton 1985) combinations. Because the model-expert case is arguably the most common, it deserves special attention. In previous judgment research, experts often have provided little predictive power beyond that contained in models (Camerer 1981), so it was not obvious a priori that model-expert combinations would actually work. Our psy-chological view highlights how the dynamic interplay between model and expert improves \* Accepted by Robert L. Winkler, former Departmental Editor; received September 8, 1988. This paper has been with the authors 4 months for 2 revisions. 887 0025- 1909/90/3608/0887$01.25 Copyright C) 1990, The Institute of Management Sciences This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

888 ROBERT C. BLATTBERG AND STEPHEN J. HOCH forecast accuracy. Models and experts have shared but also unique forecasting aptitudes; models are too consistent and experts are too flexible. By integrating these decision inputs, we can exploit strengths and compensate for weaknesses. 2. A Comparison of Model Predictions and Expert Judgment How do forecasts based on statistical models and expert judgment differ? Both methods operate on target information, but while models do so in a consistent, mechanical fashion, experts rely on intuition. Dictionaries define intuition as the act or process of coming to direct knowledge or certainty without reasoning or inferring, a keen and quick insight. Opinions about intuition have varied greatly over time, glorified as the only certain form of knowledge (the view of Descartes and Spinoza) and castigated for its unreliability (Noddings and Shore 1984). But despite disagreement over its value, there is consensus that purely intuitive judgments represent global decisions about which the decision maker is usually not able to offer a clear and complete justification. The expert has difficulty articulating the bases for intuition beyond a reliance on "gut-feel." The illusiveness of intuition is problematic for managers who have to rationalize their decisions, but (valid) "tacit" knowledge may result from the automatization of decision processes. Evidence on the validity of intuition is equivocal. We briefly review the literatures on (a) comparisons of experts and novices and (b) comparisons of experts to simple quan-titative models in order to highlight the relative strengths and weaknesses of models and expert judgment. Experts versus Novices versus Simple Models Research in cognitive science (Larkin et al. 1980; Lesgold et al. 1988) suggests that experts have highly organized, domain-specific knowledge that allows them to encode complex information; this knowledge results in faster and more accurate performance (Chi, Glaser, and Rees 1981). The judgment and decision making literature presents a less flattering picture of the expert. Experts often have performed no better than novices (Einhorn 1974; Goldberg 1959; Hoch 1988). These results are puzzling since, by defi-nition, experts should be more skillful. Research suggests that experts are better at knowing what questions to ask (diagnosis) than at predicting the future. Research comparing experts to simple actuarial models is much clearer. In studies of medical diagnosis, psychological assessment, financial forecasting, student admissions, and security analysis, actuarial models always produce more accurate forecasts than the experts (Dawes, Faust, and Meehl 1989; Sawyer 1966). The most common paradigm has been to present multiattribute profiles of a target stimulus to the expert and then ask for relevant judgments about some criterion. The models typically have been built using multiple regression, regressing the criterion onto the various predictor variables that make up the multiattribute description. In an early influential study, Meehl (1959) had 29 clinical psychologists rate the mental health of 861 patients described by profiles consisting of 11 scores from a commonly used psychological test (MMPI). The actuarial model (r = 0.46) fit better than the average (r = 0.28) and the best clinician (r = 0.39). Another body of research, known as judgment bootstrapping, has demonstrated an even more interesting finding. A model of the expert's judgment policy is constructed regressing the expert's forecasts onto the multiattribute profiles. Multiple studies (Camerer 1981) have shown that the predictions from the model of the expert (fitted values from the regression) are more highly correlated with the criterion than are the original judgments on which the expert's model is based. Bootstrapping works when the residuals from the model of the expert consist mainly of random variance in judgment; use of the fitted values from the expert's model eliminates a source of variance not predictive of outcomes. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 889 Bootstrapping systematizes judgment, but also discards any intuition the expert may have that is not compatible with the model. The robustness of the bootstrapping result has led some to conclude that model consistency more than compensates for any valid intuition experts might have not captured by the simple model (Dawes et al. 1989). Some valid aspects of judgment, however, are not easily captured by a linear model. Meehl (1954) coined the term "broken leg" cue to describe highly diagnostic cues that occur so infrequently that they are difficult to incorporate into a statistical model. He gave the example of a sociologist predicting whether a professor would go to the movies on a particular night. A sociometric model could be built containing a variety of relevant variables (tenure, children). If, however, the sociologist found out that the professor had just broken his leg, then he could confidently ignore the model and still make an accurate forecast based on a single fact. Models have difficulty incorporating broken leg cues. There may not be enough historical evidence showing the effect of these rare cues; more-over, the modeler could exhaust the available degrees of freedom by attempting to in-corporate all the potential candidates, a problem not faced by experts. Johnson (1988) found that without broken leg cues experts and novices performed at comparable levels and were outperformed by both an actuarial and bootstrapping model. With access to broken leg cues, however, experts improved dramatically, outperforming the bootstrapping model (which ignores broken leg cues) though not the actuarial model (which also ignores broken leg cues). Combining Models and Experts Over the last 20 years an extensive literature on combining forecasts has accumulated (Clemen 1989). In searching for the single "best" combining method (Granger and Ramanathan 1984), the lasting conclusion is that almost any combination of forecasts proves more accurate than the single inputs. Most research has focused on combinations of multiple models or multiple experts, but not model and expert (cf. Conroy and Harris 1987; Lawrence, Edmundson, and O'Connor 1986; Pankoff and Roberts 1967). Models from different sources (e.g., competing econometric forecasting firms) often will have different predictor variables and be calibrated on databases that are only partially redun-dant (so sample size effectively increases). In the face of random error, averaging these different model specifications can improve accuracy. Combining models can also work because of difficulties in consolidating all exogenous variables into a single comprehensive model (Bunn 1988). If such a model is estimated on an unstable covariance matrix, then a combination of submodels which ignore intercorrelations and treat forecast errors as independent can do better. Combinations of experts, on the other hand, increase accuracy because the inconsistencies of one judge tend to cancel out the inconsistencies of another (Hogarth 1978). Studies (Goldberg 1970) have found that an average of experts' forecasts performs comparable to bootstrapping models; both methods eliminate random variance. But what about model-expert combinations? One might argue that combining experts with models is simply another form of combining forecasts. However, in light of the bootstrapping literature, it is not clear that experts would add anything to the model forecasts, since their residual expertise (above that captured by a model) typically has been very small. Alternatively, experts and models have quite different strengths and weaknesses, which might be beneficial for combining purposes; these respective strengths and weaknesses are summarized below. Where Experts Are Weak and Models Are Strong: > Experts display decision biases of perception and evaluation. Models are unbiased. > Experts suffer from overconfidence and are influenced by organizational politics. Models take base-rates into account and are immune to social pressures for consensus. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

890 ROBERT C. BLATTBERG AND STEPHEN J. HOCH > Experts get tired, bored, and emotional. Models do not. > Experts do not consistently integrate evidence. Models optimally weight the evidence. Where Models Are Weak and Experts Are Strong: > Models know only what the expert has told the model-builder. Experts know what questions to ask and can identify new variables. Experts diagnose and predict, models only predict. > Experts are proficient at attribute valuation, providing subjective evaluations of variables that are difficult to measure objectively (Einhorn 1974). > Models are consistent, but as a consequence are also rigid. Experts are inconsistent but are flexible in adapting to changing conditions. > Experts have highly organized, domain-specific knowledge. They may be able to recognize and then interpret abnormal cases containing "broken leg" cues, cues that are very diagnostic but so rare that they are difficult to anticipate and therefore include in a model. Experts and models are both substitutes and complements; substitutes because both take into account much of the same decision relevant information; complements because where one decision input is weak the other is stronger and vice versa. In the next section we offer a method for taking advantage of the consistency offered by a model and the intuition that only can come from an expert. 3. Quantifying Managerial Intuition From the perspective of engineering higher quality decisions, it is not necessary to completely understand intuition in order to use it to good advantage. In fact, if experts could tell us the information on which they base their intuition, we would include it in our models if quantifiable. In situations where experts do have valid intuition, all that is necessary is that we can isolate that part of subjective judgment that contains the intuition. Our approach shares similarities with the Lens Model analysis developed by Brunswik ( 1952) used in judgment bootstrapping (Goldberg 1970). We start with a target event (Y) that is a probabilistic function of multiple correlated cues or environmental infor-mation (Xi's). The Xi's are observable predictor variables available to both decision maker and modeler. The expert also makes a prediction (P) of Y, the target event. We begin by building the best fitting model of Y given the Xi cues using multiple regression, Y= X+E. (1) Let M = X,8 be the vector of model predictions. We consider only linear models, though our approach could be adapted to any estimable nonlinear form. On the estimation data, the decision maker cannot integrate (linearly) the same information in a more efficient manner; the model predictions (M) represent an upper bound on the linear information extractable from these environmental cues. The question is whether the expert adds any predictive power above and beyond that of the model. Operationally, we define intuition as the residual portion of an expert's prediction controlling for the predictor variables (the Xi's). Intuition can be isolated by regressing the expert's predictions (P) onto the model's predictions (M), P= yM+ U, (2) with all variables in standardized form (including the residuals), equation (2) can be rewritten as p =aihm+ 1l-ao2u, (3) This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 891 where a equals the correlation between p and m,, and u represents the standardized residuals (Hoch 1987). The residuals, u, contain the "unique" part of the expert's forecast composed of both valid intuition and random error. The valid intuition could result from the expert's ability to pick up omitted variables or nonlinearities and interactions not in the model. Because u and mn are orthogonal, the expert's predictive accuracy, the correlation between y and p [ r( y, p)], can be expressed as, r(y, p) = ar(y, m) + 1-aa2r(y, u), (4) where the correlation r( y, mt) represents the accuracy of the model and the correlation r(y, u) represents the validity of intuition (Hoch 1987).1 With a means of isolating intuition, the question becomes whether we can put u to better use than the experts have in their raw predictions; in other words, can we combine the model (mt) and intuition (u) together in a better manner? On the estimation data, the answer is obviously yes. An optimal combination of model and intuition can be no less accurate than the best of the inputs in isolation. Adding any new variable would increase the overall fit, though this may not hold true when moving on to out-of-sample forecasting. Consider first the optimal combination (y3) of model predictions with expert predictions that is obtained from regressing y onto mh and p, y=b1rm+b2P- (5) The overall fit of the optimal combination of model and expert can be written as R2(y, 9) = b1r(y, mh) + b2r(y, p), (6) where bI and b2 are the relative weights for model and expert. The trade-off between model and intuition is revealed by substituting equation (4) into equation (6), R2(y, 9) = (b, + b2a)r(y, mi) + b2l1 - a2r(y, u). (7) Because the orthogonality of mn and u allows for a unique variance partitioning between model and intuition, equation (7) further simplifies to R2(y, 9) = R2(y, mh) + R2(y, u). (8) Whenever experts have valid intuition [ r( y, u) # 0], model-expert combinations will be more accurate than either of the single inputs. Two useful statistics fall out of this analysis. The first is r( y, u), the validity of expert intuition, which is equivalent to the semipartial correlation between y and p after partialling the model ( m) out of p (Cohen and Cohen 1975). Regular partial correlations, where the model is controlled for in both y and p, can also be calculated [equivalent to the correlation between the residuals in equations ( 1) and (3), r( E, u)]. When squared, the regular partial correlation represents the percent of outcome variance unexplained by the model that can be explained by expert intuition. 4. Database Models and Managerial Intuition Forecasting Situations Two different forecasting situations were studied: (a) buyers' predictions of catalog sales of fashion merchandise, and (b) brand managers' predictions of coupon redemption rates. Statistical models were built and then contrasted against managerial forecasts and ' The residuals here are estimated by controlling for the actuarial model (Ye in Lens Model terms). The Lens Model residuals, z,, are estimated by controlling for the model of the judge (j)). We use this formulation because in combining models and intuition, we wish to use the best available model and the actuarial model always will be at least as good as the model of the judge. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

892 ROBERT C. BLATTBERG AND STEPHEN J. HOCH a model-expert combination. Models were developed during various consulting projects, and as such represented good faith efforts to identify the "best" model given time and money considerations. Catalog Fashion Sales. In two different firms selling apparel through direct mail catalogs, buyers were responsible for estimating demand for an item at the SKU (stock keeping unit) level. Predicting fashion is extremely difficult-fashions are constantly changing and buying decisions need to be made anywhere from 3 to 6 months in advance of the catalog drop in order to ensure adequate inventory. These catalog companies were interested in improving the ordering process. There were two types of predictor variables (Xi's): (a) characteristics of the item and the way it was merchandised in the catalog; and (b) information about each item from a consumer survey. Variables in the first category included: percent of page devoted to the item; location in the catalog; department (e.g. lingerie); price and percent markup; and other variables like the number of colors. These variables were identified both through discussions with buyers, management, and previous catalog research. Also, a sample of target consumers were shown mocked-up versions of the catalogs. After browsing through the catalog, consumers answered a series of questions about each item such as "Is the item a good value for the money?" and also could purchase any of the items at a 10% discount, allowing us to calculate sample response rates. Buyers made sales forecasts for each of the items in the context of normal decision making about item purchasing and inventory. Buyer forecasts were made after all other decision parameters (price, etc.) had been set. In each firm, multiple buyers were involved in forecasting, though for each item only one buyer made a forecast. The criterion variable (Y) was the number of orders for each item received by each firm, more appropriate than actual sales because it is unaffected by item stock-outs due to inaccurate forecasts. Models were built using OLS regression, regressing the logarithm of orders onto the predictors. Buyer forecasts were also logged. Coupon Redemption Rates. In three other firms, brand managers routinely made predictions about the redemption rates for price-off coupons on frequently purchased consumer packaged goods. Such forecasts are a common aspect of managing promotional activities; accurate forecasts are important because managers need to anticipate the prod-uct's financial liability, and for purposes of choosing among different promotional activities depending on tactical goals (e.g., inducing trial). Predictor variables for these models were identified through interviews with product management and from related published research in the area (Blattberg and Neslin 1990). Predictors included: coupon face value, percent discount, brand, duration of the offer, a product category development index, and media type (FSIs, in-packs). Managers made forecasts at the time of the coupon issue date so all decision parameters had been set. Models were built using weighted least squares regression; the observations were weighted to reflect the number of coupons dropped during the promotion. The criterion and manager forecasts were logged before estimation. Results The final fashion buying models for Company 1 (CO 1) and Company 2 (C02) each contained 11 predictors. The coupon redemption models for Companies 3-5 each con-tained over 30 predictors, most of which were dummy variables for media type and brand. To control for shrinkage, cross-validated models were tested, where the model was fit on half the data and then used to predict the remaining data. To ensure robustness of the results, 10 separate cross-validation analyses, randomly splitting the samples 10 different ways, were conducted for each data set; the reported results represent the average of the 10 analyses. The cross-validation results appear in Table 1. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 893 Model, Manager, and Model + Manager. In all cases, the statistical models fit quite well. Though some shrinkage was observed in the cross-validation analyses, the model fits were still quite good, an average R2 of over 0.55. Managers also displayed substantial expertise, predictive accuracy comparable to that of the models except for CO 1, t( 105) - 2.4, p < 0.01. Hierarchical regression analyses were conducted to assess whether the combination of the model and manager provided a significant incremental increase in accuracy over either decision input in isolation. The overall fit of "model + manager" was assessed by regressing the criterion onto the predictions of the model (M) and the expert (P). The statistical tests are F tests of the differences between the fit (R 2) of the full model (model + manager) and the two reduced models (database model or manager alone). In all cases a combination of the model and the manager led to a significant increase in predictive accuracy over the model or the manager alone, all p's < 0.0001. The fourth column labled i\ shows the increase in R2 that accompanies relying on a combination of model + manager compared to using the best single decision input. i\ averages 0.09 in the cross-validation results, increasing from 0.05 in the complete data sets. This suggests that manager forecasts assume an even more important role when model shrinkage occurs (whether due to overfitting or structural changes in the environment). Model shrinkage averaged almost 13% in the cross-validation analyses. When manager forecasts were combined with model predictions, however, shrinkage was reduced to only 5.5%, resulting in almost 60% less shrinkage in the model + manager combination forecasts. The two decision inputs are complementary. When misspecified models break down during im-plementation, expert judgment can significantly improve predictive accuracy. Several other analyses are of interest. First, there was a significant degree of overlap between manager forecasts and model predictions, average r (M, P) > 0.7 for the five firms. At the same time, however, managers displayed a significant amount of intuition about fashion buying and coupon redemptions. To examine intuition, we calculated semipartial and regular partial correlations between actual sales (Y) and manager forecasts (P) controlling for model predictions. These statistics allow us to assess the incremental contribution of the manager adjusting for the correlation between the manager and the database model. The results appear in the last two columns of Table 1. One thing is very clear: in all five data sets, managers demonstrate a substantial degree of intuition about the nonlinear aspects of their forecasting tasks. On average the validity of intuition is over 0.31 in the hold-out samples. These results differ markedly from almost all previous studies of expert judgment, where typically intuition is quite low. A reanalysis of 15 bootstrapping studies by Camerer (1981) found much lower nonlinear intuition, an average validity of less than 0.06 (comparable to the fifth column in Table 1) .2 Managers TABLE 1 Regression Results Comparing the Predictive Accuracy of the Model, the M1fanager, and a Combination of the Model + the Manager Unexplained R2 of Validity of Variance Picked Up Cross-validation Hold- R2 of R2 of Model Intuition by the Manager out Sample Model Manager + Manager i\ r(y, u) [r(y, p.n,)]2 Company 1 (n= 108) 0.47 0.30 0.53 0.06 0.25 12% Company 2 (n = 100) 0.63 0.67 0.74 0.07 0.33 29% Company 3 (n = 203) 0.56 0.52 0.66 0.10 0.31 22% Company4(n= 173) 0.71 0.74 0.83 0.11 0.34 40% Company 5 (n= 1008) 0.39 0.39 0.50 0.11 0.32 17% 2 Bootstrapping residuals (z5 = y- j35)a lways are more highly correlated with the criterion than the residuals (ul = Ys -Ye) used here because rz = r(ye, z5) =r(y, u)/r(zS, u). This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

894 ROBERT C. BLATTBERG AND STEPHEN J. HOCH also were able to pick up a substantial amount of the variance not explained by the model, 24% in the hold-out samples, corroborating the significant improvements in ac-curacy due to combination of model and manager. These results provide a much more flattering picture of expert judgment than most previous studies. Stabilizing Models with Manager Forecasts. The cross-validation results for both the catalog and coupon redemption data suggest that managerial forecasts can play an im-portant role in the decision making process by providing a stabilizing influence on model forecasts which serves to reduce model shrinkage. Managers may be able to take into account structural changes in the decision environment that models cannot detect. Al-though models can be reparameterized periodically, sudden changes in the data generating process (possibly signalled by broken leg cues) will be difficult to detect without the extended history required for recalibration. To examine whether managers can indeed anticipate structural changes not detected by the statistical models, temporally-based analyses of C04 and C05 were conducted using information about coupon issue date. The data sets covered 20 months. Models were initially fit to the first 10 months of data, and then these models were used to forecast coupon redemptions in the succeeding months. The results appear in Table 2. During the "fitted" period the models fit very well; however, in the forecast period, model fits decreased significantly, C04, z = 4.04, p < 0.001, and, for C05, z = 2.32, p < 0.01. This suggests nonstationarity (or trend) in the data and that the model is unable to account for whatever structural changes occurred. In contrast, manager forecasts were robust across time periods. Managers successfully picked up environmental changes, and so in the forecast period manager's forecasts explained almost 38% of the variance in re-demptions not captured by the model [r(y, p.mn)]2 By spending 10 months collecting data (and then building a model), the two firms could have increased forecast accuracy by 12% (R2 increase of 0.06) above manager forecasts. Alternatively, by combining manager forecasts with existing database models calibrated on past data, accuracy could have been increased by 38% (R2 increase of 0.17). Discussions with managers indicated that they believed that over time there had been a general decline in coupon redemption rates, whether due to increased use as a promotional tool by competitors or changes in buying patterns. Even if redemptions were declining (something not obvious in the data), it is not clear how this fact could have been incorporated into our models. Optimal and Heuristic Weighting of Model and Manager. The optimal trade-off be-tween model and manager can be expressed in percentage terms (the percentages represent a ratio of the standardized beta weights), where the percents would reflect the relative weights applied to a linear combination of the standardized forecasts of model and man-ager. On average the optimal trade-off in the hold-out samples is 50:50, ranging from 64:36 for CO 1 to 44:56 for C02 and C04. These model/manager tradeoffs can be trans- TABLE 2 Time-Based Redemption Rate Analyses Comparing Model, Manager, and Model + Manager Unexplained R2 of Validity of Variance Picked Up R2 of R2 of Model Intuition by the Manager Time Period Model Manager + Manager i\ r(y, u) [r(y, p.h)]2 Model Estimation: Company 4 (n = 174) 0.87 0.74 0.90 0.03 0.24 28% Company 5 (n = 1048) 0.42 0.41 0.51 0.09 0.31 17% Future Forecast: Company 4 (n = 172) 0.61 0.74 0.80 0.06 0.44 49% Company 5 (n = 968) 0.36 0.45 0.52 0.07 0.41 26% This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 895 lated into unique model/ intuition tradeoffs using equation (7). Part of the valid variance in manager forecasts is already contained in model predictions. The 50:50 split between model and manager translates into a 70:30 split between model and intuition. The robustness of the 50:50 heuristic was tested through a simulation where the relative weights given to model and manager were systematically varied in increments of 10% from 100% model to 100% manager. The cross-validation results are shown for each firm in Figure 1. The simulation demonstrates that the utility of combining the two decision inputs is relatively insensitive to the exact weights applied to each. Substantial drop-offs in fit occurred only at the extremes, where most or all of the emphasis was given to one of the decision inputs. Robustness with respect to the exact weighting of model and manager is expected in this case given the relatively high correlation between the two inputs. Using a 50:50 rule for each of the five firms, in lieu of the optimal weights (indicated by the dots), resulted in only about a 1% decrease in R2 on average. The Value of Less Sophisticated Models. Our analyses show that model and manager forecasts are complementary sources of information that increase in predictive accuracy when considered in tandem. Managerial judgments are more adaptable to new circum-stances and therefore can help to stabilize the performance of models in changing decision environments. Models provide a consistent information source that compensates for the inconsistency inherent in human judgment. The final set of analyses examines potential benefits from combining less sophisticated models with managerial judgment. The basic analytic strategy involved building of systematically "degraded" models. The data were reanalyzed using degraded models containing 50% of the variables contained in the full models. For each firm the "2" models represent the average fit of 10 models each constructed from random subsets of variables in the full model. The 2- models explain only 63% of the variance captured by the full models. However, a com-bination of manager + 2 -model still results in significant improvements in accuracy over the best single decision input-an average increase in R2 of 8.8% (A = 0.07). Degraded models containing less than 20% of the variables in the full model were also constructed. Because 20%-models were more likely not to contain one of the key predictors, their fit .80- . - \_ C04 .70' C0 I.. .60- 0~ ~ ~~~~~~~~~~0 >s 50- o 50:50 .30- | Co1 100% 80:20 60:40 40:60 20:80 100% Model Manager Relative Weighting of Model and Manager FIGURE 1. Predictive Accuracy for Different Weighting Combinations of Model and Manager. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

896 ROBERT C. BLATTBERG AND STEPHEN J. HOCH was poor, an average R2 of 0.18 (28% of the variance explained by the full models). These weak models still improved the accuracy of manager forecasts in isolation, increasing R2 by over 3%, and suggest that model-expert combinations may be worthwhile even in the early stages of model building. Despite an obviously naive model, the consistency inherent in such a model may improve the accuracy of forecasts based solely on intuition. 5. Discussion Our findings suggest an encouraging view of the complementarity of database models and managerial intuition. Managers displayed quite high levels of intuition; they picked up almost 25% of the variance left unexplained by our models. We still do not know where the intuition comes from, but it is clear that it would be foolish to disregard it and rely solely on a statistical model for future forecasting. Although the optimal combination of model and manager will always be more accurate on the estimation data, this need not be the case on hold-out samples or if a heuristic weighting rule is employed. The combination of model + manager increased predictive accuracy by a substantial degree (average A > 0.09 in the hold-out samples). The inclusion of the manager provided the added benefit of dramatically decreasing model shrinkage in hold-out samples. Across the five firms, the goodness of fit of the models and managers varied substantially, a range in R2 for the model of 0.24 and for managers of 0.44. But in each case the 50:50 heuristic improved predictive accuracy. Why did our experts display substantial intuition? We will discuss four possible reasons: (a) building of "naive" models; (b) use of realistic tasks where performance matters; (c) presence of certain artificial factors inflated expert performance; and (d) existence of valid intuition in business forecasting. Although our remarks are speculative, we hope that a discussion of each will put into perspective when and where one might find similar improvements in decision making quality by relying on models and intuition. Naive Models? It is possible that we built overly simplistic models that make the experts look better on a relative basis and inflate the validity of intuition. Although what constitutes a "good" model depends on the phenomenon being predicted, our models fit quite well (R2 = 0.55 on cross-validated samples), much better than most actuarial models reported in the bootstrapping research previously demonstrating the superiority of models over expert judgment. Moreover, extensive discussions with the buyers and managers did not reveal any omitted factors that could be easily incorporated into the models, only qualitative considerations ("fashion orientation"). Nonlinear transformations possibly could have improved model fits, but without prior theory, the possibilities are infinite. Better models could have been built, but given time, money, and potential payout, we feel reasonably confident that the results are not due to obviously naive modeling efforts. Realistic Task Where the Peiformance Matters? Our results were obtained in the context of everyday decision making, offering a "ve-ridical" task (Johnson 1983) where experts could tap into domain-specific knowledge. In the typical experimental study, experts are asked to make judgments about artificial stimuli, usually depicted in multiattribute rather than holistic form (cf. Phelps and Shan-teau 1978). Experts have available to them exactly the same information as does the model-they cannot take advantage of any skill they may have at identifying other in-formation not incorporated in the model. Therefore, the only way that experts can perform better than the model is if they can execute a better information integration policy. By definition, we know that experts cannot execute a better "linear" policy on the estimation sample; therefore, they must apprehend the nonlinear aspects of the given attributes. Previous research clearly has demonstrated that experts are not proficient at this. Whereas This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 897 previous studies have tended to place the burden on the expert to figure out the exact weighting of all the pre-specified cues, our study, in contrast, probably placed more of the burden on the modeler to specify all the appropriate cues available in naturally occurring decision environments. But since the burden will usually be on the modeler in most on-line forecasting situations, our results may be fairly representative. One ad-ditional point-bootstrapping of the experts did not work except for CO1; the benefits of consistency (linearization) could not compensate for the loss of a substantial amount of valid intuition. Our experts also understood the incentives for performing well, so motivation was high. In experimental situations, this may not always be the case. More importantly, fatigue, boredom, and temporary distractions were less likely to influence performance. In laboratory studies multiple judgments are required in short time spans ( 1-2 hours), whereas our experts could elect to deliberate extensively over each prediction. Also, managers in the coupon redemption studies may have learned from experience, as they did receive some outcome feedback over time. Artificial Factors Inflate Decision Maker Accuracy? In discussing why the earnings forecasts of management are consistently superior to those of analysts and extrapolation methods (Armstrong 1983), Brown ( 1988) identifies three factors: self-selection of events to be predicted; inside information; and control over the phenomenon being forecast. Although our study required experts and models to make forecasts for all events, some self-selection may be at work-fashion items and coupon offers characterized by high uncertainty or low expectations of success could have been censored early in the process. Experts also had inside information, not the Machiavellian variety available to corporate officers, but they clearly had more information available to them than did the models. We have elected to label this inside information as intuition and see it as a valuable decision input. Finally, although expert and model forecasts were made after all decision parameters (price, media type) were established, experts probably exerted some control because expectations may have guided their setting of some of decision parameters. Expectations may produce outcomes that are at least partly self-fulfilling. These factors would be inoperative in a randomized experiment, but in most on-line forecasting situations decision makers do exert some control and as a consequence will display higher predictive accuracy. Whether one chooses to label the reasons for this improved performance as artificial or real is irrelevant to the decision to rely on a model-expert combination. We see this as another reason why model-builders cannot ignore expert forecasts. Truly Valid Intuition? If we assume that our models are adequate and that our experts had only minimal control over future outcomes, then one is left with the fact that our buyers and product managers demonstrated valid intuition. Our experts explained almost 10% incremental variance in the target events which represents almost 25% of the outcome variance not captured by the statistical models. At the same time, experts and models were highly correlated, indicating that a substantial part of judgment was linear in form and also fairly consistent. So the question becomes why do decision makers display so much valid intuition about these two forecasting situations? It could be that our experts were capable of the categorical thinking not easily simulated by a linear statistical model, possibly taking account of interactions and nonlinear predictors. They may have had access to cues that either could not be quantified or were perceived at a less conscious, pattern-recognition level. Also, experts may have been able to interpret abnormal cases ("broken leg"-like omitted variables) when encountered even though they could not anticipate them a priori. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

898 ROBERT C. BLATTBERG AND STEPHEN J. HOCH Weather forecasters also have consistently demonstrated forecasting prowess, and in a situation where the forecaster has no control over the target event. Although not fore-casting a physical system, our task had characteristics similar to those that Murphy and Brown ( 1984) believe make weather forecasting amenable to the development of expertise. The tasks were circumscribed, each target event being characterized by similar background information. Experts had substantial experience with the forecasting domain, were com-fortable with the response scales, and had access to a variety of general industry analyses and other experts' forecasts. Conclusions In two forecasting situations where managers made real-time forecasts, we found that statistical models and managerial judgment achieved about the same level of predictive accuracy. We also found that a combination of model + manager outperformed either decision input in isolation. Models and managers have complementary skills. Models combine complex data in a consistent and unbiased manner. Managers have additional insight that the model cannot incorporate, such as the state of the economy, fashion trends, idiosyncratic features of an item, and shifting coupon redemption patterns. Man-agers may pick up a "broken leg" cue, so rare, that it would never be anticipated by a model. The trick is to incorporate, model consistency and managerial insight into one forecast. Models are inflexible, making them less accurate as environments change. Man-agers, on the other hand, may tend to be too adaptive and overreact to current devel-opments. Model-manager combinations can increase adaptivity while placing a regressive, but needed, upper bound on that adaptivity. Thus, model and manager may stabilize each other. Given the well-recognized limits to human information processing capacity and the explosion of new data sources, managers need to move away from intuition as the sole basis for decision making. Intuition needs to be made "less intuitive." Until more is known about how to build better models, the 50% Model + 50% Manager decision heuristic is a nonoptimal but pragmatic solution offering three key advantages: (a) sim-plicity-managers do not need to understand or develop models, so the natural orga-nizational separation of modelers and managers can continue; (b) palatability-managers retain a considerable amount of control over the decision making process; and (c) ac-curacy-a combination of model + manager will be more accurate than the individual decision inputs. Model-manager combinations are easily incorporated into existing decision support systems. If managers (possibly working with modelers) can identify the informational basis for exceptions to the model, the model refinement process could be improved. Future research might investigate the efficacy of a procedure where, like weather forecasters who modify model predictions on the fly when not jibing with subjective assessments (Murphy and Brown 1984), managers make adjustments to model predictions rather than making independent forecasts. Whether or not an interactive procedure such as this would lead to an improvement (or decrement) over a mechanical combination rule is an open question, but the opportunity for training managers to use the system is intriguing given the success of weather forecasters.3 Order of authorship is alphabetical. The authors thank Ken Hammond, Scott Hawkins, Robin Hogarth, Josh Klayman, George Loewenstein, and Jay Russo for comments. References ARMSTRONG, J. S., "Relative Accuracy of Judgmental and Extrapolative Methods in Forecasting Annual Earn-ings," J. Forecasting, 2(1983), 437-447. ASHTON, A. H. AND R. H. ASHTON, "Aggregating Subjective Forecasts: Some Empirical Results," Management Sci., 31 (1985), 1499-1508. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 899 BLATTBERG,R. C. AND S. NESLIN, Sales Promotions, Prentice-Hall, Englewood Cliffs, NJ, 1990. BROWN, L. D., "Comparing Judgmental to ExtrapolativeF orecasts:I t's Time to Ask Why and When," Internat. J. Forecasting, 4 (1988), 171-173. BRUNSWIK, E., "The Conceptual Framework of Psychology," in International Encyclopedia of Unified Science, Vol. 1, No. 10, University of Chicago Press, Chicago, 1952. BUNN, D. W., "Combining Forecasts," European J. Oper. Res., 33 (1988), 223-229. CAMERER, C., "General Conditions for the Success of Bootstrapping Models," Organizational Behavior and Human Performance, 27 (1981), 411-422. CHI, M. T. H., R. GLASER AND E. REES, "Expertise in Problem Solving," in R. J. Sternberg (Ed.), Advances in the Psychology of Intelligence, Erlbaum, Hilldale, NJ, 1981. CLEMEN, R. T., "Combining Forecasts: A Review and Annotated Bibliography," Internat. J. Forecasting, 4 ([989), 559-584. COHEN, J. AND P. COHEN, Applied Multiple Regression/Correlation Analysis of the Behavioral Sciences, Erlbaum, Hillsdale, NJ, 1975. CONROY, R. AND R. HARRIS, "Consensus Forecasts of Corporate Earnings: Analysts' Forecasts and Time Series Methods," Management Sci., 33 (1987), 725-738. DAWES, R. M., D. FAUST AND P. E. MEEHL, "Clinical Versus Actuarial Judgment," Science, 243 (1989), 1668-1674. EINHORN, H. J., "Cue Definition and Residual Judgment," Organizational Behavior and Human Performance, 12 (1974), 30-49. GOLDBERG, L., "The Effectiveness of Clinician's Judgments: The Diagnosis of Organic Brain Damage from the Bender-Gestalt Test," J. Consulting Psychology, 23 (1959), 25-33. , "Man vs. Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences," Psychological Bulletin, 73 (1970), 422-432. GRANGER, C. W. J. AND R. RAMANATHAN, "Improved Methods of Forecasting," J. Forecasting, 3(1984), 197-204. HOCH, S. J., "Perceived Consensus and Predictive Accuracy: The Pros and Cons of Projection," J. Personality and Social Psychology, 53 (1987), 221-234. , "Who Do We Know: Predicting the Interests and Opinions of the American Consumer," J. Consumer Res., 15 (1988), 315-324. HOGARTH, R. M., "A Note on Aggregating Opinions," Organizational Behavior and Human Petformance, 21 (1978), 40-46. JOHNSON, E. J., "Expertise and Decision Under Uncertainty: Performance and Process," in M. T. H. Chi, R. Glaser, and M. J. Farr (Eds.), The Nature of Expertise, Erlbaum, Hillsdale, NJ, 1988. JOHNSON, P. E., "What Kind of Expert Should a System Be?," J. Medicine and Philosophy, 8 (1983), 77-97. LARKIN, J., J. McDERMOTT, D. P. SIMON AND H. A. SIMON, "Expert and Novice Performance in Solving Physics Problems," Science, 208 (June 1980), 1335-1342. LAWRENCE, M. J., R. H. EDMUNDSON, AND M. J. O'CONNOR, "The Accuracy of Combining Judgmental and Statistical Forecasts," Management Sci., 32 (1986), 1521-1532. LESGOLD, A., H. RUBINSON, P. FELTOVICH, R. GLASER AND D. KLOPFER, "Expertise in Complex Skills: Diagnosing X-Ray Pictures," in M. T. H. Chi, R. Glaser, and M. J. Farr (Eds.), The Nature of Expertise, Erlbaum, Hillsdale, NJ, 1988. LORIE, J. H., "Two Important Problems in Sales Forecasting," J. Business, 30 (1957), 172-179. MEEHL, P. E., Clinical Versus Statistical Prediction, University of Minnesota Press, Minneapolis, 1954. ,("A Comparison of Clinicians with Five Statistical Methods of Identifying Psychotic MMPI Profiles," J. Consulting Psychology, 6 (1959), 102-109. MURPHY, A. H. AND B. G. BROWN, "A Comparative Evaluation of Objective and Subjective Weather Forecasts in the United States," J. Forecasting, 3 (1984), 369-393. NODDINGS, N. AND P. J. SHORE, Awakeningt he InnerE ye: Intuitioni n Education, Teacher's College Press, New York, 1984. PANKOFF, L. D. AND H. V. ROBERTS, "Bayesian Synthesis of Clinical and Statistical Prediction," Psychological Bulletin, 80 (1968), 762-773. PHELPS, R. H. AND J. SHANTEAU, "Livestock Judges: How Much Information Can An Expert Use?," Orga-nizational Behavior and Human Petformance, 21 (1978), 209-219. SAWYER, J., "Measurements and Prediction, Clinical and Statistical," Psychological Bulletin, 66 (1966), 178- 200. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

Database Models and Managerial Intuition: 50${\tt\%}$ Model + 50${\tt\%}$ ManagerAuthor(s): Robert C. Blattberg and Stephen J. HochSource: Management Science, Vol. 36, No. 8 (Aug., 1990), pp. 887-899Published by: INFORMSStable URL: http://www.jstor.org/stable/2632364 .Accessed: 26/02/2014 18:04Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at .http://www.jstor.org/page/info/about/policies/terms.jsp .JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range ofcontent in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new formsof scholarship. For more information about JSTOR, please contact support@jstor.org. .*INFORMS* is collaborating with JSTOR to digitize, preserve and extend access to *Management Science.*http://www.jstor.org This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

MANAGEMENT SCIENCE Vol. 36, No. 8, August 1990 Printed in U.S.A. DATABASE MODELS AND MANAGERIAL INTUITION: 50% MODEL + 50% MANAGER\* ROBERT C. BLATTBERG AND STEPHEN J. HOCH University of Chicago, Graduate School of Business, 11 F1E . 58th Street, Chicago, Illinois 60637 We focus on ways of combining simple database models with managerial intuition. We present a model and method for isolating managerial intuition. For five different business forecasting situations, our results indicate that a combination of model and manager always outperforms either of these decision inputs in isolation, an average R2 increase of 0.09 (16%) above the best single decision input in cross-validated model analyses. We assess the validity of an equal weighting heuristic, 50% model + 50% manager, and then discuss why our results might differ from previous research on expert judgment. (FORECASTING; DECISION MAKING; EXPERTISE; DECISION SUPPORT SYSTEMS) 1. Introduction Everybody's got so much information all day long that they lose their common sense. (Ger-trude Stein) A very important problem in sales forecasting is combining the wisdom of experienced businessmen with statistical analysis. (Lorie 1957) With the availability of more and better sources of data, decision makers must begin to systematically incorporate such information into the decision process. At present, however, we know much more about building and processing databases than about how experts might use such data to improve decision quality. We show how firms can take advantage of the data explosion by combining simple database models with managerial intuition, viewing the two decision inputs in combination rather than in competition. We identify differences between models and intuition, and develop a way to isolate and evaluate intuition. We analyze several on-line business forecasting situations and show that decision quality could have been improved dramatically by relying on both models and intuition. We find that a 50% Model + 50% Manager heuristic improves forecast quality. The essential message is that both statistical and human inputs should guide final decisions, at least given the current state of model building and the difficulty in quantifyinig expert intuition. The paper is applicable to a large constituency in the business community. Forecasters can use "econometric" models effectively only if they have a built-in adjustment mech-anism to capture the changing environment. We argue that managerial forecasts can fill this role and improve predictive accuracy. Thus we will not simply conclude that com-bination rules are best. This is of course the overwhelming conclusion reached in the forecasting literature (Clemen 1989), though most of the empirical work in that literature has focused on a model-model (Granger and Ramanathan 1984) or expert-expert (Ashton and Ashton 1985) combinations. Because the model-expert case is arguably the most common, it deserves special attention. In previous judgment research, experts often have provided little predictive power beyond that contained in models (Camerer 1981), so it was not obvious a priori that model-expert combinations would actually work. Our psy-chological view highlights how the dynamic interplay between model and expert improves \* Accepted by Robert L. Winkler, former Departmental Editor; received September 8, 1988. This paper has been with the authors 4 months for 2 revisions. 887 0025- 1909/90/3608/0887$01.25 Copyright C) 1990, The Institute of Management Sciences This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

888 ROBERT C. BLATTBERG AND STEPHEN J. HOCH forecast accuracy. Models and experts have shared but also unique forecasting aptitudes; models are too consistent and experts are too flexible. By integrating these decision inputs, we can exploit strengths and compensate for weaknesses. 2. A Comparison of Model Predictions and Expert Judgment How do forecasts based on statistical models and expert judgment differ? Both methods operate on target information, but while models do so in a consistent, mechanical fashion, experts rely on intuition. Dictionaries define intuition as the act or process of coming to direct knowledge or certainty without reasoning or inferring, a keen and quick insight. Opinions about intuition have varied greatly over time, glorified as the only certain form of knowledge (the view of Descartes and Spinoza) and castigated for its unreliability (Noddings and Shore 1984). But despite disagreement over its value, there is consensus that purely intuitive judgments represent global decisions about which the decision maker is usually not able to offer a clear and complete justification. The expert has difficulty articulating the bases for intuition beyond a reliance on "gut-feel." The illusiveness of intuition is problematic for managers who have to rationalize their decisions, but (valid) "tacit" knowledge may result from the automatization of decision processes. Evidence on the validity of intuition is equivocal. We briefly review the literatures on (a) comparisons of experts and novices and (b) comparisons of experts to simple quan-titative models in order to highlight the relative strengths and weaknesses of models and expert judgment. Experts versus Novices versus Simple Models Research in cognitive science (Larkin et al. 1980; Lesgold et al. 1988) suggests that experts have highly organized, domain-specific knowledge that allows them to encode complex information; this knowledge results in faster and more accurate performance (Chi, Glaser, and Rees 1981). The judgment and decision making literature presents a less flattering picture of the expert. Experts often have performed no better than novices (Einhorn 1974; Goldberg 1959; Hoch 1988). These results are puzzling since, by defi-nition, experts should be more skillful. Research suggests that experts are better at knowing what questions to ask (diagnosis) than at predicting the future. Research comparing experts to simple actuarial models is much clearer. In studies of medical diagnosis, psychological assessment, financial forecasting, student admissions, and security analysis, actuarial models always produce more accurate forecasts than the experts (Dawes, Faust, and Meehl 1989; Sawyer 1966). The most common paradigm has been to present multiattribute profiles of a target stimulus to the expert and then ask for relevant judgments about some criterion. The models typically have been built using multiple regression, regressing the criterion onto the various predictor variables that make up the multiattribute description. In an early influential study, Meehl (1959) had 29 clinical psychologists rate the mental health of 861 patients described by profiles consisting of 11 scores from a commonly used psychological test (MMPI). The actuarial model (r = 0.46) fit better than the average (r = 0.28) and the best clinician (r = 0.39). Another body of research, known as judgment bootstrapping, has demonstrated an even more interesting finding. A model of the expert's judgment policy is constructed regressing the expert's forecasts onto the multiattribute profiles. Multiple studies (Camerer 1981) have shown that the predictions from the model of the expert (fitted values from the regression) are more highly correlated with the criterion than are the original judgments on which the expert's model is based. Bootstrapping works when the residuals from the model of the expert consist mainly of random variance in judgment; use of the fitted values from the expert's model eliminates a source of variance not predictive of outcomes. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 889 Bootstrapping systematizes judgment, but also discards any intuition the expert may have that is not compatible with the model. The robustness of the bootstrapping result has led some to conclude that model consistency more than compensates for any valid intuition experts might have not captured by the simple model (Dawes et al. 1989). Some valid aspects of judgment, however, are not easily captured by a linear model. Meehl (1954) coined the term "broken leg" cue to describe highly diagnostic cues that occur so infrequently that they are difficult to incorporate into a statistical model. He gave the example of a sociologist predicting whether a professor would go to the movies on a particular night. A sociometric model could be built containing a variety of relevant variables (tenure, children). If, however, the sociologist found out that the professor had just broken his leg, then he could confidently ignore the model and still make an accurate forecast based on a single fact. Models have difficulty incorporating broken leg cues. There may not be enough historical evidence showing the effect of these rare cues; more-over, the modeler could exhaust the available degrees of freedom by attempting to in-corporate all the potential candidates, a problem not faced by experts. Johnson (1988) found that without broken leg cues experts and novices performed at comparable levels and were outperformed by both an actuarial and bootstrapping model. With access to broken leg cues, however, experts improved dramatically, outperforming the bootstrapping model (which ignores broken leg cues) though not the actuarial model (which also ignores broken leg cues). Combining Models and Experts Over the last 20 years an extensive literature on combining forecasts has accumulated (Clemen 1989). In searching for the single "best" combining method (Granger and Ramanathan 1984), the lasting conclusion is that almost any combination of forecasts proves more accurate than the single inputs. Most research has focused on combinations of multiple models or multiple experts, but not model and expert (cf. Conroy and Harris 1987; Lawrence, Edmundson, and O'Connor 1986; Pankoff and Roberts 1967). Models from different sources (e.g., competing econometric forecasting firms) often will have different predictor variables and be calibrated on databases that are only partially redun-dant (so sample size effectively increases). In the face of random error, averaging these different model specifications can improve accuracy. Combining models can also work because of difficulties in consolidating all exogenous variables into a single comprehensive model (Bunn 1988). If such a model is estimated on an unstable covariance matrix, then a combination of submodels which ignore intercorrelations and treat forecast errors as independent can do better. Combinations of experts, on the other hand, increase accuracy because the inconsistencies of one judge tend to cancel out the inconsistencies of another (Hogarth 1978). Studies (Goldberg 1970) have found that an average of experts' forecasts performs comparable to bootstrapping models; both methods eliminate random variance. But what about model-expert combinations? One might argue that combining experts with models is simply another form of combining forecasts. However, in light of the bootstrapping literature, it is not clear that experts would add anything to the model forecasts, since their residual expertise (above that captured by a model) typically has been very small. Alternatively, experts and models have quite different strengths and weaknesses, which might be beneficial for combining purposes; these respective strengths and weaknesses are summarized below. Where Experts Are Weak and Models Are Strong: > Experts display decision biases of perception and evaluation. Models are unbiased. > Experts suffer from overconfidence and are influenced by organizational politics. Models take base-rates into account and are immune to social pressures for consensus. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

890 ROBERT C. BLATTBERG AND STEPHEN J. HOCH > Experts get tired, bored, and emotional. Models do not. > Experts do not consistently integrate evidence. Models optimally weight the evidence. Where Models Are Weak and Experts Are Strong: > Models know only what the expert has told the model-builder. Experts know what questions to ask and can identify new variables. Experts diagnose and predict, models only predict. > Experts are proficient at attribute valuation, providing subjective evaluations of variables that are difficult to measure objectively (Einhorn 1974). > Models are consistent, but as a consequence are also rigid. Experts are inconsistent but are flexible in adapting to changing conditions. > Experts have highly organized, domain-specific knowledge. They may be able to recognize and then interpret abnormal cases containing "broken leg" cues, cues that are very diagnostic but so rare that they are difficult to anticipate and therefore include in a model. Experts and models are both substitutes and complements; substitutes because both take into account much of the same decision relevant information; complements because where one decision input is weak the other is stronger and vice versa. In the next section we offer a method for taking advantage of the consistency offered by a model and the intuition that only can come from an expert. 3. Quantifying Managerial Intuition From the perspective of engineering higher quality decisions, it is not necessary to completely understand intuition in order to use it to good advantage. In fact, if experts could tell us the information on which they base their intuition, we would include it in our models if quantifiable. In situations where experts do have valid intuition, all that is necessary is that we can isolate that part of subjective judgment that contains the intuition. Our approach shares similarities with the Lens Model analysis developed by Brunswik ( 1952) used in judgment bootstrapping (Goldberg 1970). We start with a target event (Y) that is a probabilistic function of multiple correlated cues or environmental infor-mation (Xi's). The Xi's are observable predictor variables available to both decision maker and modeler. The expert also makes a prediction (P) of Y, the target event. We begin by building the best fitting model of Y given the Xi cues using multiple regression, Y= X+E. (1) Let M = X,8 be the vector of model predictions. We consider only linear models, though our approach could be adapted to any estimable nonlinear form. On the estimation data, the decision maker cannot integrate (linearly) the same information in a more efficient manner; the model predictions (M) represent an upper bound on the linear information extractable from these environmental cues. The question is whether the expert adds any predictive power above and beyond that of the model. Operationally, we define intuition as the residual portion of an expert's prediction controlling for the predictor variables (the Xi's). Intuition can be isolated by regressing the expert's predictions (P) onto the model's predictions (M), P= yM+ U, (2) with all variables in standardized form (including the residuals), equation (2) can be rewritten as p =aihm+ 1l-ao2u, (3) This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 891 where a equals the correlation between p and m,, and u represents the standardized residuals (Hoch 1987). The residuals, u, contain the "unique" part of the expert's forecast composed of both valid intuition and random error. The valid intuition could result from the expert's ability to pick up omitted variables or nonlinearities and interactions not in the model. Because u and mn are orthogonal, the expert's predictive accuracy, the correlation between y and p [ r( y, p)], can be expressed as, r(y, p) = ar(y, m) + 1-aa2r(y, u), (4) where the correlation r( y, mt) represents the accuracy of the model and the correlation r(y, u) represents the validity of intuition (Hoch 1987).1 With a means of isolating intuition, the question becomes whether we can put u to better use than the experts have in their raw predictions; in other words, can we combine the model (mt) and intuition (u) together in a better manner? On the estimation data, the answer is obviously yes. An optimal combination of model and intuition can be no less accurate than the best of the inputs in isolation. Adding any new variable would increase the overall fit, though this may not hold true when moving on to out-of-sample forecasting. Consider first the optimal combination (y3) of model predictions with expert predictions that is obtained from regressing y onto mh and p, y=b1rm+b2P- (5) The overall fit of the optimal combination of model and expert can be written as R2(y, 9) = b1r(y, mh) + b2r(y, p), (6) where bI and b2 are the relative weights for model and expert. The trade-off between model and intuition is revealed by substituting equation (4) into equation (6), R2(y, 9) = (b, + b2a)r(y, mi) + b2l1 - a2r(y, u). (7) Because the orthogonality of mn and u allows for a unique variance partitioning between model and intuition, equation (7) further simplifies to R2(y, 9) = R2(y, mh) + R2(y, u). (8) Whenever experts have valid intuition [ r( y, u) # 0], model-expert combinations will be more accurate than either of the single inputs. Two useful statistics fall out of this analysis. The first is r( y, u), the validity of expert intuition, which is equivalent to the semipartial correlation between y and p after partialling the model ( m) out of p (Cohen and Cohen 1975). Regular partial correlations, where the model is controlled for in both y and p, can also be calculated [equivalent to the correlation between the residuals in equations ( 1) and (3), r( E, u)]. When squared, the regular partial correlation represents the percent of outcome variance unexplained by the model that can be explained by expert intuition. 4. Database Models and Managerial Intuition Forecasting Situations Two different forecasting situations were studied: (a) buyers' predictions of catalog sales of fashion merchandise, and (b) brand managers' predictions of coupon redemption rates. Statistical models were built and then contrasted against managerial forecasts and ' The residuals here are estimated by controlling for the actuarial model (Ye in Lens Model terms). The Lens Model residuals, z,, are estimated by controlling for the model of the judge (j)). We use this formulation because in combining models and intuition, we wish to use the best available model and the actuarial model always will be at least as good as the model of the judge. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

892 ROBERT C. BLATTBERG AND STEPHEN J. HOCH a model-expert combination. Models were developed during various consulting projects, and as such represented good faith efforts to identify the "best" model given time and money considerations. Catalog Fashion Sales. In two different firms selling apparel through direct mail catalogs, buyers were responsible for estimating demand for an item at the SKU (stock keeping unit) level. Predicting fashion is extremely difficult-fashions are constantly changing and buying decisions need to be made anywhere from 3 to 6 months in advance of the catalog drop in order to ensure adequate inventory. These catalog companies were interested in improving the ordering process. There were two types of predictor variables (Xi's): (a) characteristics of the item and the way it was merchandised in the catalog; and (b) information about each item from a consumer survey. Variables in the first category included: percent of page devoted to the item; location in the catalog; department (e.g. lingerie); price and percent markup; and other variables like the number of colors. These variables were identified both through discussions with buyers, management, and previous catalog research. Also, a sample of target consumers were shown mocked-up versions of the catalogs. After browsing through the catalog, consumers answered a series of questions about each item such as "Is the item a good value for the money?" and also could purchase any of the items at a 10% discount, allowing us to calculate sample response rates. Buyers made sales forecasts for each of the items in the context of normal decision making about item purchasing and inventory. Buyer forecasts were made after all other decision parameters (price, etc.) had been set. In each firm, multiple buyers were involved in forecasting, though for each item only one buyer made a forecast. The criterion variable (Y) was the number of orders for each item received by each firm, more appropriate than actual sales because it is unaffected by item stock-outs due to inaccurate forecasts. Models were built using OLS regression, regressing the logarithm of orders onto the predictors. Buyer forecasts were also logged. Coupon Redemption Rates. In three other firms, brand managers routinely made predictions about the redemption rates for price-off coupons on frequently purchased consumer packaged goods. Such forecasts are a common aspect of managing promotional activities; accurate forecasts are important because managers need to anticipate the prod-uct's financial liability, and for purposes of choosing among different promotional activities depending on tactical goals (e.g., inducing trial). Predictor variables for these models were identified through interviews with product management and from related published research in the area (Blattberg and Neslin 1990). Predictors included: coupon face value, percent discount, brand, duration of the offer, a product category development index, and media type (FSIs, in-packs). Managers made forecasts at the time of the coupon issue date so all decision parameters had been set. Models were built using weighted least squares regression; the observations were weighted to reflect the number of coupons dropped during the promotion. The criterion and manager forecasts were logged before estimation. Results The final fashion buying models for Company 1 (CO 1) and Company 2 (C02) each contained 11 predictors. The coupon redemption models for Companies 3-5 each con-tained over 30 predictors, most of which were dummy variables for media type and brand. To control for shrinkage, cross-validated models were tested, where the model was fit on half the data and then used to predict the remaining data. To ensure robustness of the results, 10 separate cross-validation analyses, randomly splitting the samples 10 different ways, were conducted for each data set; the reported results represent the average of the 10 analyses. The cross-validation results appear in Table 1. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 893 Model, Manager, and Model + Manager. In all cases, the statistical models fit quite well. Though some shrinkage was observed in the cross-validation analyses, the model fits were still quite good, an average R2 of over 0.55. Managers also displayed substantial expertise, predictive accuracy comparable to that of the models except for CO 1, t( 105) - 2.4, p < 0.01. Hierarchical regression analyses were conducted to assess whether the combination of the model and manager provided a significant incremental increase in accuracy over either decision input in isolation. The overall fit of "model + manager" was assessed by regressing the criterion onto the predictions of the model (M) and the expert (P). The statistical tests are F tests of the differences between the fit (R 2) of the full model (model + manager) and the two reduced models (database model or manager alone). In all cases a combination of the model and the manager led to a significant increase in predictive accuracy over the model or the manager alone, all p's < 0.0001. The fourth column labled i\ shows the increase in R2 that accompanies relying on a combination of model + manager compared to using the best single decision input. i\ averages 0.09 in the cross-validation results, increasing from 0.05 in the complete data sets. This suggests that manager forecasts assume an even more important role when model shrinkage occurs (whether due to overfitting or structural changes in the environment). Model shrinkage averaged almost 13% in the cross-validation analyses. When manager forecasts were combined with model predictions, however, shrinkage was reduced to only 5.5%, resulting in almost 60% less shrinkage in the model + manager combination forecasts. The two decision inputs are complementary. When misspecified models break down during im-plementation, expert judgment can significantly improve predictive accuracy. Several other analyses are of interest. First, there was a significant degree of overlap between manager forecasts and model predictions, average r (M, P) > 0.7 for the five firms. At the same time, however, managers displayed a significant amount of intuition about fashion buying and coupon redemptions. To examine intuition, we calculated semipartial and regular partial correlations between actual sales (Y) and manager forecasts (P) controlling for model predictions. These statistics allow us to assess the incremental contribution of the manager adjusting for the correlation between the manager and the database model. The results appear in the last two columns of Table 1. One thing is very clear: in all five data sets, managers demonstrate a substantial degree of intuition about the nonlinear aspects of their forecasting tasks. On average the validity of intuition is over 0.31 in the hold-out samples. These results differ markedly from almost all previous studies of expert judgment, where typically intuition is quite low. A reanalysis of 15 bootstrapping studies by Camerer (1981) found much lower nonlinear intuition, an average validity of less than 0.06 (comparable to the fifth column in Table 1) .2 Managers TABLE 1 Regression Results Comparing the Predictive Accuracy of the Model, the M1fanager, and a Combination of the Model + the Manager Unexplained R2 of Validity of Variance Picked Up Cross-validation Hold- R2 of R2 of Model Intuition by the Manager out Sample Model Manager + Manager i\ r(y, u) [r(y, p.n,)]2 Company 1 (n= 108) 0.47 0.30 0.53 0.06 0.25 12% Company 2 (n = 100) 0.63 0.67 0.74 0.07 0.33 29% Company 3 (n = 203) 0.56 0.52 0.66 0.10 0.31 22% Company4(n= 173) 0.71 0.74 0.83 0.11 0.34 40% Company 5 (n= 1008) 0.39 0.39 0.50 0.11 0.32 17% 2 Bootstrapping residuals (z5 = y- j35)a lways are more highly correlated with the criterion than the residuals (ul = Ys -Ye) used here because rz = r(ye, z5) =r(y, u)/r(zS, u). This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

894 ROBERT C. BLATTBERG AND STEPHEN J. HOCH also were able to pick up a substantial amount of the variance not explained by the model, 24% in the hold-out samples, corroborating the significant improvements in ac-curacy due to combination of model and manager. These results provide a much more flattering picture of expert judgment than most previous studies. Stabilizing Models with Manager Forecasts. The cross-validation results for both the catalog and coupon redemption data suggest that managerial forecasts can play an im-portant role in the decision making process by providing a stabilizing influence on model forecasts which serves to reduce model shrinkage. Managers may be able to take into account structural changes in the decision environment that models cannot detect. Al-though models can be reparameterized periodically, sudden changes in the data generating process (possibly signalled by broken leg cues) will be difficult to detect without the extended history required for recalibration. To examine whether managers can indeed anticipate structural changes not detected by the statistical models, temporally-based analyses of C04 and C05 were conducted using information about coupon issue date. The data sets covered 20 months. Models were initially fit to the first 10 months of data, and then these models were used to forecast coupon redemptions in the succeeding months. The results appear in Table 2. During the "fitted" period the models fit very well; however, in the forecast period, model fits decreased significantly, C04, z = 4.04, p < 0.001, and, for C05, z = 2.32, p < 0.01. This suggests nonstationarity (or trend) in the data and that the model is unable to account for whatever structural changes occurred. In contrast, manager forecasts were robust across time periods. Managers successfully picked up environmental changes, and so in the forecast period manager's forecasts explained almost 38% of the variance in re-demptions not captured by the model [r(y, p.mn)]2 By spending 10 months collecting data (and then building a model), the two firms could have increased forecast accuracy by 12% (R2 increase of 0.06) above manager forecasts. Alternatively, by combining manager forecasts with existing database models calibrated on past data, accuracy could have been increased by 38% (R2 increase of 0.17). Discussions with managers indicated that they believed that over time there had been a general decline in coupon redemption rates, whether due to increased use as a promotional tool by competitors or changes in buying patterns. Even if redemptions were declining (something not obvious in the data), it is not clear how this fact could have been incorporated into our models. Optimal and Heuristic Weighting of Model and Manager. The optimal trade-off be-tween model and manager can be expressed in percentage terms (the percentages represent a ratio of the standardized beta weights), where the percents would reflect the relative weights applied to a linear combination of the standardized forecasts of model and man-ager. On average the optimal trade-off in the hold-out samples is 50:50, ranging from 64:36 for CO 1 to 44:56 for C02 and C04. These model/manager tradeoffs can be trans- TABLE 2 Time-Based Redemption Rate Analyses Comparing Model, Manager, and Model + Manager Unexplained R2 of Validity of Variance Picked Up R2 of R2 of Model Intuition by the Manager Time Period Model Manager + Manager i\ r(y, u) [r(y, p.h)]2 Model Estimation: Company 4 (n = 174) 0.87 0.74 0.90 0.03 0.24 28% Company 5 (n = 1048) 0.42 0.41 0.51 0.09 0.31 17% Future Forecast: Company 4 (n = 172) 0.61 0.74 0.80 0.06 0.44 49% Company 5 (n = 968) 0.36 0.45 0.52 0.07 0.41 26% This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 895 lated into unique model/ intuition tradeoffs using equation (7). Part of the valid variance in manager forecasts is already contained in model predictions. The 50:50 split between model and manager translates into a 70:30 split between model and intuition. The robustness of the 50:50 heuristic was tested through a simulation where the relative weights given to model and manager were systematically varied in increments of 10% from 100% model to 100% manager. The cross-validation results are shown for each firm in Figure 1. The simulation demonstrates that the utility of combining the two decision inputs is relatively insensitive to the exact weights applied to each. Substantial drop-offs in fit occurred only at the extremes, where most or all of the emphasis was given to one of the decision inputs. Robustness with respect to the exact weighting of model and manager is expected in this case given the relatively high correlation between the two inputs. Using a 50:50 rule for each of the five firms, in lieu of the optimal weights (indicated by the dots), resulted in only about a 1% decrease in R2 on average. The Value of Less Sophisticated Models. Our analyses show that model and manager forecasts are complementary sources of information that increase in predictive accuracy when considered in tandem. Managerial judgments are more adaptable to new circum-stances and therefore can help to stabilize the performance of models in changing decision environments. Models provide a consistent information source that compensates for the inconsistency inherent in human judgment. The final set of analyses examines potential benefits from combining less sophisticated models with managerial judgment. The basic analytic strategy involved building of systematically "degraded" models. The data were reanalyzed using degraded models containing 50% of the variables contained in the full models. For each firm the "2" models represent the average fit of 10 models each constructed from random subsets of variables in the full model. The 2- models explain only 63% of the variance captured by the full models. However, a com-bination of manager + 2 -model still results in significant improvements in accuracy over the best single decision input-an average increase in R2 of 8.8% (A = 0.07). Degraded models containing less than 20% of the variables in the full model were also constructed. Because 20%-models were more likely not to contain one of the key predictors, their fit .80- . - \_ C04 .70' C0 I.. .60- 0~ ~ ~~~~~~~~~~0 >s 50- o 50:50 .30- | Co1 100% 80:20 60:40 40:60 20:80 100% Model Manager Relative Weighting of Model and Manager FIGURE 1. Predictive Accuracy for Different Weighting Combinations of Model and Manager. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

896 ROBERT C. BLATTBERG AND STEPHEN J. HOCH was poor, an average R2 of 0.18 (28% of the variance explained by the full models). These weak models still improved the accuracy of manager forecasts in isolation, increasing R2 by over 3%, and suggest that model-expert combinations may be worthwhile even in the early stages of model building. Despite an obviously naive model, the consistency inherent in such a model may improve the accuracy of forecasts based solely on intuition. 5. Discussion Our findings suggest an encouraging view of the complementarity of database models and managerial intuition. Managers displayed quite high levels of intuition; they picked up almost 25% of the variance left unexplained by our models. We still do not know where the intuition comes from, but it is clear that it would be foolish to disregard it and rely solely on a statistical model for future forecasting. Although the optimal combination of model and manager will always be more accurate on the estimation data, this need not be the case on hold-out samples or if a heuristic weighting rule is employed. The combination of model + manager increased predictive accuracy by a substantial degree (average A > 0.09 in the hold-out samples). The inclusion of the manager provided the added benefit of dramatically decreasing model shrinkage in hold-out samples. Across the five firms, the goodness of fit of the models and managers varied substantially, a range in R2 for the model of 0.24 and for managers of 0.44. But in each case the 50:50 heuristic improved predictive accuracy. Why did our experts display substantial intuition? We will discuss four possible reasons: (a) building of "naive" models; (b) use of realistic tasks where performance matters; (c) presence of certain artificial factors inflated expert performance; and (d) existence of valid intuition in business forecasting. Although our remarks are speculative, we hope that a discussion of each will put into perspective when and where one might find similar improvements in decision making quality by relying on models and intuition. Naive Models? It is possible that we built overly simplistic models that make the experts look better on a relative basis and inflate the validity of intuition. Although what constitutes a "good" model depends on the phenomenon being predicted, our models fit quite well (R2 = 0.55 on cross-validated samples), much better than most actuarial models reported in the bootstrapping research previously demonstrating the superiority of models over expert judgment. Moreover, extensive discussions with the buyers and managers did not reveal any omitted factors that could be easily incorporated into the models, only qualitative considerations ("fashion orientation"). Nonlinear transformations possibly could have improved model fits, but without prior theory, the possibilities are infinite. Better models could have been built, but given time, money, and potential payout, we feel reasonably confident that the results are not due to obviously naive modeling efforts. Realistic Task Where the Peiformance Matters? Our results were obtained in the context of everyday decision making, offering a "ve-ridical" task (Johnson 1983) where experts could tap into domain-specific knowledge. In the typical experimental study, experts are asked to make judgments about artificial stimuli, usually depicted in multiattribute rather than holistic form (cf. Phelps and Shan-teau 1978). Experts have available to them exactly the same information as does the model-they cannot take advantage of any skill they may have at identifying other in-formation not incorporated in the model. Therefore, the only way that experts can perform better than the model is if they can execute a better information integration policy. By definition, we know that experts cannot execute a better "linear" policy on the estimation sample; therefore, they must apprehend the nonlinear aspects of the given attributes. Previous research clearly has demonstrated that experts are not proficient at this. Whereas This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 897 previous studies have tended to place the burden on the expert to figure out the exact weighting of all the pre-specified cues, our study, in contrast, probably placed more of the burden on the modeler to specify all the appropriate cues available in naturally occurring decision environments. But since the burden will usually be on the modeler in most on-line forecasting situations, our results may be fairly representative. One ad-ditional point-bootstrapping of the experts did not work except for CO1; the benefits of consistency (linearization) could not compensate for the loss of a substantial amount of valid intuition. Our experts also understood the incentives for performing well, so motivation was high. In experimental situations, this may not always be the case. More importantly, fatigue, boredom, and temporary distractions were less likely to influence performance. In laboratory studies multiple judgments are required in short time spans ( 1-2 hours), whereas our experts could elect to deliberate extensively over each prediction. Also, managers in the coupon redemption studies may have learned from experience, as they did receive some outcome feedback over time. Artificial Factors Inflate Decision Maker Accuracy? In discussing why the earnings forecasts of management are consistently superior to those of analysts and extrapolation methods (Armstrong 1983), Brown ( 1988) identifies three factors: self-selection of events to be predicted; inside information; and control over the phenomenon being forecast. Although our study required experts and models to make forecasts for all events, some self-selection may be at work-fashion items and coupon offers characterized by high uncertainty or low expectations of success could have been censored early in the process. Experts also had inside information, not the Machiavellian variety available to corporate officers, but they clearly had more information available to them than did the models. We have elected to label this inside information as intuition and see it as a valuable decision input. Finally, although expert and model forecasts were made after all decision parameters (price, media type) were established, experts probably exerted some control because expectations may have guided their setting of some of decision parameters. Expectations may produce outcomes that are at least partly self-fulfilling. These factors would be inoperative in a randomized experiment, but in most on-line forecasting situations decision makers do exert some control and as a consequence will display higher predictive accuracy. Whether one chooses to label the reasons for this improved performance as artificial or real is irrelevant to the decision to rely on a model-expert combination. We see this as another reason why model-builders cannot ignore expert forecasts. Truly Valid Intuition? If we assume that our models are adequate and that our experts had only minimal control over future outcomes, then one is left with the fact that our buyers and product managers demonstrated valid intuition. Our experts explained almost 10% incremental variance in the target events which represents almost 25% of the outcome variance not captured by the statistical models. At the same time, experts and models were highly correlated, indicating that a substantial part of judgment was linear in form and also fairly consistent. So the question becomes why do decision makers display so much valid intuition about these two forecasting situations? It could be that our experts were capable of the categorical thinking not easily simulated by a linear statistical model, possibly taking account of interactions and nonlinear predictors. They may have had access to cues that either could not be quantified or were perceived at a less conscious, pattern-recognition level. Also, experts may have been able to interpret abnormal cases ("broken leg"-like omitted variables) when encountered even though they could not anticipate them a priori. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

898 ROBERT C. BLATTBERG AND STEPHEN J. HOCH Weather forecasters also have consistently demonstrated forecasting prowess, and in a situation where the forecaster has no control over the target event. Although not fore-casting a physical system, our task had characteristics similar to those that Murphy and Brown ( 1984) believe make weather forecasting amenable to the development of expertise. The tasks were circumscribed, each target event being characterized by similar background information. Experts had substantial experience with the forecasting domain, were com-fortable with the response scales, and had access to a variety of general industry analyses and other experts' forecasts. Conclusions In two forecasting situations where managers made real-time forecasts, we found that statistical models and managerial judgment achieved about the same level of predictive accuracy. We also found that a combination of model + manager outperformed either decision input in isolation. Models and managers have complementary skills. Models combine complex data in a consistent and unbiased manner. Managers have additional insight that the model cannot incorporate, such as the state of the economy, fashion trends, idiosyncratic features of an item, and shifting coupon redemption patterns. Man-agers may pick up a "broken leg" cue, so rare, that it would never be anticipated by a model. The trick is to incorporate, model consistency and managerial insight into one forecast. Models are inflexible, making them less accurate as environments change. Man-agers, on the other hand, may tend to be too adaptive and overreact to current devel-opments. Model-manager combinations can increase adaptivity while placing a regressive, but needed, upper bound on that adaptivity. Thus, model and manager may stabilize each other. Given the well-recognized limits to human information processing capacity and the explosion of new data sources, managers need to move away from intuition as the sole basis for decision making. Intuition needs to be made "less intuitive." Until more is known about how to build better models, the 50% Model + 50% Manager decision heuristic is a nonoptimal but pragmatic solution offering three key advantages: (a) sim-plicity-managers do not need to understand or develop models, so the natural orga-nizational separation of modelers and managers can continue; (b) palatability-managers retain a considerable amount of control over the decision making process; and (c) ac-curacy-a combination of model + manager will be more accurate than the individual decision inputs. Model-manager combinations are easily incorporated into existing decision support systems. If managers (possibly working with modelers) can identify the informational basis for exceptions to the model, the model refinement process could be improved. Future research might investigate the efficacy of a procedure where, like weather forecasters who modify model predictions on the fly when not jibing with subjective assessments (Murphy and Brown 1984), managers make adjustments to model predictions rather than making independent forecasts. Whether or not an interactive procedure such as this would lead to an improvement (or decrement) over a mechanical combination rule is an open question, but the opportunity for training managers to use the system is intriguing given the success of weather forecasters.3 Order of authorship is alphabetical. The authors thank Ken Hammond, Scott Hawkins, Robin Hogarth, Josh Klayman, George Loewenstein, and Jay Russo for comments. References ARMSTRONG, J. S., "Relative Accuracy of Judgmental and Extrapolative Methods in Forecasting Annual Earn-ings," J. Forecasting, 2(1983), 437-447. ASHTON, A. H. AND R. H. ASHTON, "Aggregating Subjective Forecasts: Some Empirical Results," Management Sci., 31 (1985), 1499-1508. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions

DATABASE MODELS AND MANAGERIAL INTUITION 899 BLATTBERG,R. C. AND S. NESLIN, Sales Promotions, Prentice-Hall, Englewood Cliffs, NJ, 1990. BROWN, L. D., "Comparing Judgmental to ExtrapolativeF orecasts:I t's Time to Ask Why and When," Internat. J. Forecasting, 4 (1988), 171-173. BRUNSWIK, E., "The Conceptual Framework of Psychology," in International Encyclopedia of Unified Science, Vol. 1, No. 10, University of Chicago Press, Chicago, 1952. BUNN, D. W., "Combining Forecasts," European J. Oper. Res., 33 (1988), 223-229. CAMERER, C., "General Conditions for the Success of Bootstrapping Models," Organizational Behavior and Human Performance, 27 (1981), 411-422. CHI, M. T. H., R. GLASER AND E. REES, "Expertise in Problem Solving," in R. J. Sternberg (Ed.), Advances in the Psychology of Intelligence, Erlbaum, Hilldale, NJ, 1981. CLEMEN, R. T., "Combining Forecasts: A Review and Annotated Bibliography," Internat. J. Forecasting, 4 ([989), 559-584. COHEN, J. AND P. COHEN, Applied Multiple Regression/Correlation Analysis of the Behavioral Sciences, Erlbaum, Hillsdale, NJ, 1975. CONROY, R. AND R. HARRIS, "Consensus Forecasts of Corporate Earnings: Analysts' Forecasts and Time Series Methods," Management Sci., 33 (1987), 725-738. DAWES, R. M., D. FAUST AND P. E. MEEHL, "Clinical Versus Actuarial Judgment," Science, 243 (1989), 1668-1674. EINHORN, H. J., "Cue Definition and Residual Judgment," Organizational Behavior and Human Performance, 12 (1974), 30-49. GOLDBERG, L., "The Effectiveness of Clinician's Judgments: The Diagnosis of Organic Brain Damage from the Bender-Gestalt Test," J. Consulting Psychology, 23 (1959), 25-33. , "Man vs. Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences," Psychological Bulletin, 73 (1970), 422-432. GRANGER, C. W. J. AND R. RAMANATHAN, "Improved Methods of Forecasting," J. Forecasting, 3(1984), 197-204. HOCH, S. J., "Perceived Consensus and Predictive Accuracy: The Pros and Cons of Projection," J. Personality and Social Psychology, 53 (1987), 221-234. , "Who Do We Know: Predicting the Interests and Opinions of the American Consumer," J. Consumer Res., 15 (1988), 315-324. HOGARTH, R. M., "A Note on Aggregating Opinions," Organizational Behavior and Human Petformance, 21 (1978), 40-46. JOHNSON, E. J., "Expertise and Decision Under Uncertainty: Performance and Process," in M. T. H. Chi, R. Glaser, and M. J. Farr (Eds.), The Nature of Expertise, Erlbaum, Hillsdale, NJ, 1988. JOHNSON, P. E., "What Kind of Expert Should a System Be?," J. Medicine and Philosophy, 8 (1983), 77-97. LARKIN, J., J. McDERMOTT, D. P. SIMON AND H. A. SIMON, "Expert and Novice Performance in Solving Physics Problems," Science, 208 (June 1980), 1335-1342. LAWRENCE, M. J., R. H. EDMUNDSON, AND M. J. O'CONNOR, "The Accuracy of Combining Judgmental and Statistical Forecasts," Management Sci., 32 (1986), 1521-1532. LESGOLD, A., H. RUBINSON, P. FELTOVICH, R. GLASER AND D. KLOPFER, "Expertise in Complex Skills: Diagnosing X-Ray Pictures," in M. T. H. Chi, R. Glaser, and M. J. Farr (Eds.), The Nature of Expertise, Erlbaum, Hillsdale, NJ, 1988. LORIE, J. H., "Two Important Problems in Sales Forecasting," J. Business, 30 (1957), 172-179. MEEHL, P. E., Clinical Versus Statistical Prediction, University of Minnesota Press, Minneapolis, 1954. ,("A Comparison of Clinicians with Five Statistical Methods of Identifying Psychotic MMPI Profiles," J. Consulting Psychology, 6 (1959), 102-109. MURPHY, A. H. AND B. G. BROWN, "A Comparative Evaluation of Objective and Subjective Weather Forecasts in the United States," J. Forecasting, 3 (1984), 369-393. NODDINGS, N. AND P. J. SHORE, Awakeningt he InnerE ye: Intuitioni n Education, Teacher's College Press, New York, 1984. PANKOFF, L. D. AND H. V. ROBERTS, "Bayesian Synthesis of Clinical and Statistical Prediction," Psychological Bulletin, 80 (1968), 762-773. PHELPS, R. H. AND J. SHANTEAU, "Livestock Judges: How Much Information Can An Expert Use?," Orga-nizational Behavior and Human Petformance, 21 (1978), 209-219. SAWYER, J., "Measurements and Prediction, Clinical and Statistical," Psychological Bulletin, 66 (1966), 178- 200. This content downloaded from 164.67.137.168 on Wed, 26 Feb 2014 18:04:49 PM All use subject to JSTOR Terms and Conditions