Automating the B2B Salesperson Pricing Decisions: A Human-Machine Hybrid Approach

Yael Karlinsky-Shichor and Oded Netzer

Abstract

In a world advancing towards automation, we propose a human-machine hybrid approach to automating decision making in high human interaction environments and apply it in the business-to-business (B2B) retail context. Using sales transactions data from a B2B aluminum retailer, we create an automated version of each salesperson, that learns and automatically reapplies the salesperson's pricing policy. We conduct a field experiment with the B2B retailer, providing salespeople with their own model's price recommendations in real-time through the retailer's CRM system, and allowing them to adjust their original pricing accordingly. We find that despite the loss of non-codeable information available to the salesperson but not to the model, providing the model's price to the salesperson increases profits for treated quotes by 11% relatively to a control condition. Using a counterfactual analysis, we show that while in most of the cases the model's pricing leads to higher profitability by eliminating inter-temporal human biases, the salesperson generates higher profits when pricing special quotes with unique or complex characteristics. Accordingly, we propose a machine learning Random Forest *hybrid pricing strategy*, that automatically allocates quotes to the model or to the human expert and generates profits significantly higher than either the model or the salespeople.

1 Introduction

In the past century, automation has changed the labor market by consistently substituting for predictable and repetitive human tasks. In the early days of automation, its goal was first and foremost scalability and efficiency of well-defined tasks with clear inputs and outputs. Recent advances in computational methods and artificial intelligence (AI) allowed automation to tap into occupations that involve non-routine aspects such as judgment, perception and manipulation, creative intelligence and social intelligence (Brynjolfsson and McAfee, 2012; Chui et al., 2016; Frey and Osborne, 2017). Consequently, automation is bound to transform a significant share of soft-skills based occupations in the near future (Nedelkoska and Quintini, 2018).

Recent applications of automation and AI methods include tasks such as screening resumes (Cowgill, 2017), identifying irregularities in CT scans[1], and replacing judges in deciding whether defendants will await trial at home or in jail (Kleinberg et al., 2018). Yet, while these examples require a high level of expertise (medical doctors, human resource personnel or court judges), the task is still relatively well-defined and subjective cues in the environment should play little role in the decision process. That is, the X-ray image or the resume file should contain all (or most) of the information needed to make the judgment.

In this research we ask whether automation, either in the form of replacing the human agents or supporting them, could be applied to domains where soft skills and interpersonal interactions play an important role in the decision-making process, and where interpretation of environmental cues may provide valuable information. Specifically, we introduce automation to one such domain with high importance to marketers: pricing decision-making in business to business (B2B) retail. The B2B market is estimated at trillions of dollars, yet it largely lags behind the business-to-consumer (B2C) market in adopting technology and automation (Asare et al., 2016). Pricing decisions in B2B are often based on a combination

[1] https://finance.yahoo.com/news/intermountain-healthcare-chooses-zebra-medical-120000157.html

of sales expertise and soft skills. On the one hand, B2B salespeople's pricing decisions are good candidates for automation because they are often repetitive and arguably predictable. On the other hand, such pricing decisions may be difficult to automate because they involve a high degree of inter-personal communication, interpretation of behavioral cues and persuasion skills.

We collaborate with a B2B aluminum retailer, where salespeople interact with business clients on a daily basis and price incoming requests for products to maximize profitability. The company has thousands of stock keeping units (SKUs), customizable products and varying commodity prices, giving salespeople pricing autonomy on a quote-by-quote basis. The pricing process is relationship-based (Zhang et al., 2014), and in determining prices salespeople often respond to case-based information available to them. During the interaction with the client, salespeople may adjust prices according to their assessment of the client's willingness to pay. While salespeople are able to use soft skills that automation lacks, they suffer from a host of human behavioral decision making biases (e.g., Payne et al., 1993) and hence often make different decisions under the same circumstances with no justified reason. Examples of such biases reported in the context of pricing decisions include, for example, higher loss aversion in the afternoon to recover from morning losses (Coval and Shumway, 2005) or inter-temporal incentive scheme misalignment (Misra and Nair, 2011; Larkin and Leider, 2012). Automation can mitigate such inter-temporal biases and perform better than people when consistency is beneficial. However, given its great share of human expertise, it is unclear whether the B2B pricing process could be automated.

We propose a hybrid approach to automation, in which the salesperson and an automated pricing algorithm participate in the pricing process, utilizing the algorithm's reliability in consistently applying pricing rules and eliminating inter-temporal biases, and using human judgment for interpreting non-codeable contextual cues. In a field experiment and in counterfactual simulations we show that combining automated and human pricing can lead to higher profits than using either approach separately.

Our automated algorithm is an AI version of the B2B salesperson that mimics her past pricing behavior and applies it systematically to new pricing decisions. We create a representation of each salesperson in the company by regressing the salesperson's past pricing decisions on different variables observed by the salesperson when making the pricing decision (e.g., cost of the material, order size or the identity of the client). The approach that uses the decision variable (price margin) rather than the outcome (whether the client accepted the price or gross profit conditional on acceptance), is called *judgmental bootstrapping* in the judgment and decision making literature (Dawes, 1979). It allows to easily capture and consistently apply the salesperson's expertise and pricing knowledge.

In order to test the profit-performance of the bootstrap automated-pricing model relative to that of the salesperson, we worked with the B2B retailer to conduct a real-time pricing field experiment. Ideally in a hybrid approach the automated pricing will be the default, with salespeople monitoring prices and handling extreme cases. However, given the large and immediate impact such an experiment could have on the business, we could only provide the automated prices as recommendations. Over the course of 8 business days, involving over 2,000 price quotes and over 4,000 product requests (lines), each incoming quote was randomly assigned to either treatment (receive price recommendation based on the model) or control (do not receive price recommendation) to test the causal effect of providing salespeople with the model-based pricing. We worked with the company to integrate our pricing model for each salesperson into their customer relationship management (CRM) system and provide price recommendations in real time for quotes assigned to the treatment condition. After receiving the price predicted by the model-of-herself in the treatment condition the salesperson could decided whether to accept it, adjust it or keep her original price.

Providing salespeople with price recommendations of their own model led to substantially and statistically significantly higher profits than not providing such recommendations. Specifically, mean profit per line within a quote in the treatment condition is $10.95 higher relative to the control condition, an increase of 11% in profitability, totaling in added profits to the company of over \$26K during the eight days of the experiment, or over \$1.4 million when extrapolated yearly. An IV analysis with treatment as the exogenous instrumental variable demonstrates that treatment (price recommendations) affected client acceptance by anchoring salespeople to offer prices that are closer to the recommended model prices.

To further explore the potential of automating the B2B salesperson's pricing decisions, we perform several counterfactual analyses that allow us to overcome some of the limitations of a field experiment (e.g., price provided as a recommendation, salespeople compliance) and simulate different automation scenarios. Given alternative pricing schemes (model pricing vs. salesperson pricing), we create a profit counterfactual for each quote. For that purpose, we estimate a demand model for quote acceptance controlling for possible price endogeneity using a control function approach, with cost as an instrumental variable. We find that despite the loss of valuable information available to the salesperson but not to the model, the expected profitability of pure automation (use model prices for all quotes) is 4.9% higher than the expected profitability of the salesperson's prices.

Although pure automation performs better than the salespeople in terms of profitability, the nature of B2B pricing suggests that in some cases using human skills will lead to higher profits than using a model. Consequently, we propose a human-machine hybrid that combines automation and human decision making to increase profitability. We train a machine learning (ML) Random Forest (RF) model that predicts the difference in expected profits between the salesperson and her model based on the quote's and client's characteristics (e.g., quote weight or client purchase frequency) and allocate each quote to either human or automatic pricing. The hybrid model generates expected profits that are 7.8% higher than those of the salespeople. Aligned with our theoretical predictions that salespeople "shine" when human judgment is required, we find that the hybrid model allocates to the salesperson quotes that are more complex (e.g., include more items or require processing).

The hybrid model pushes a step forward on the human-machine continuum in automating not only the pricing decision itself, but also the decision of who should price the quote, the salesperson or the model. Rather than allowing the salesperson to decide when to follow the model's prices as in the field experiment, here the model automatically decides, based on the client and order characteristics, whether to price the quote or refer the client to a salesperson to price the quote.

Thus, in this work we demonstrate that a human-machine hybrid approach to automation could transform B2B sales. Through a field experiment and various counterfactual analyses, we show that a using both automation (for routine, codeable cases) and human judgment (for special cases, possibly with soft information involved) to make pricing decisions generates higher profits to the company than either full automation or pure human pricing. The company we collaborated with is currently implementing our model permanently into its CRM system.

The remaining of the paper is organized as follows: Section 2 discusses our contribution to the work on B2B pricing and automation. Section 3 lays out the specification of the bootstrap model of the salesperson and the empirical context for evaluating it. Section 4 describes the field experiment conducted with the company, and Section 5 describes the counterfactual analyses used to create the human-machine hybrid. Section 6 demonstrates how the company's incentive system might affect pricing and its automation. Section 7 concludes by discussing implications of our findings to salesforce automation.

2 B2B Pricing and Automation

2.1 B2B Marketing

Our work builds on and contributes to several streams of literature. We add to the relatively limited literature on B2B marketing (Grewal et al., 2015; Lilien, 2016), and specifically on B2B pricing. The B2B market was estimated at nearly $9 trillion in transactions in 2018. Nevertheless, B2B pricing decisions remain a relatively understudied topic in the literature. Increasingly, sellers face business clients that prefer to interact and place orders via e-commerce (Forrester, 2015, 2018). It is therefore of great interest to examine the possibility of automating pricing decisions in B2B context.

Buyer-seller relationships in B2B are typically long-term and relationship based (Morgan and Hunt, 1994; Lam et al., 2004). Variation of prices across clients and across purchases is common in B2B (Zhang et al., 2014). Consequently, maintaining relationship with clients, responding to clients' needs and understanding their state of mind are essential to the B2B salesperson's job when it comes to making pricing decisions. While automation has gone a long way with respect to emulating human behavior, "the real-time recognition of natural human emotion remains a challenging problem, and the ability to respond intelligently to such inputs is even more difficult" (Frey and Osborne, 2017). Therefore, the potential benefit from automating B2B pricing decisions is unclear.

2.2 Judgmental Bootstrapping, Decision Models and Automation

The roots of our approach to automation lie in the behavioral judgment as well as the decision models literature. The former stressed the idea that models of experts trumpet experts in judgments and decision making (Meehl 1954; Dawes 1979). In a *judgmental bootstrapping* (JB) model, the judgment (e.g., price), rather than the outcome (e.g., profit) is used as the dependent variable in the model of the expert. Consequently, model coefficients reflect the weight that the expert puts on each variable in making the judgment, creating a paramorphic representation of the expert's decision policy (Hoffman, 1960). Applications of JB include predicting students performance (Wiggins and Kolen, 1971), bootstrapping psychiatric doctors (Goldberg, 1970) and financial analysts (Ebert and Kruse, 1978; Batchelor and Kwan, 2007) as well as some limited applications to managerial tasks (Bowman, 1963; Kunreuther, 1969; Ashton et al., 1994)

Why should automation of the salesperson through JB perform better than the expert? Ultimately, JB uses less information (only codable information) and may repeat inefficiencies in the experts past decisions. The reasons and empirical demonstrations for the superior performance of JB over experts proposed in the decision making literature point out to its ability to eliminate inter-temporal biases. People are inconsistent decision makers, often making different judgments under similar circumstances. They tend to "think outside of the box", over-complicate and over-weigh noisy but salient inputs where deviation is not needed. While this sophistication works in odd cases, it hurts reliability in most cases (Dawes et al. 1989; Kahneman 2011). In the context of salesforce such biases may include over-correcting for losses earlier in the day (Coval and Shumway, 2005), over-weighing vivid information communicated during the call, or inter-temporal incentive scheme misalignment (Misra and Nair, 2011). The JB model may perform better in such cases by appropriately and consistently weighing the information according to rules extracted from the human decision policy, and limiting the effect of inter-temporal biases on the human decision maker's judgment. (Meehl, 1954; Armstrong, 2001).

A condition to the superiority of JB over the expert is that the information used in the decision making process is available and codeable for the model to consider. While this may be a reasonable assumption in a stylized clinical experiment, in many real-world problems the expert has access to richer information than the model does. Indeed, in our B2B pricing context, on the one hand, salespeople work in a dynamic environment and are exposed to cues which may steer them wrong on a case-by-case judgment. On the other hand, the interactions with the client may provide valuable and material information for the pricing decision. Salespeople often have the authority to adjust prices based on case-based information. For example, the salesperson may realize during a phone conversation with the client, that the order is urgent and the client is willing to pay more for this order. Adding such information to the pricing decision is important, but if it is over-weighed due to its vividness and recency, such information might steer the salesperson wrong. Nevertheless, while the model's consistency may lead to better pricing decisions in many cases, in others the model could be missing crucial information. Thus, the nature of B2B pricing might call for combining human and automated pricing decisions, with the balance between them being an open empirical question.

Another question is why automate via JB of the salesperson and not based on "optimal" prices based on an estimated model of demand. There are several reasons for this choice. First, in the spirit of directly testing automation of human behavior, we wish to build a machine (automate) of the salespeople themselves as opposed to impose an "optimal" price on them. This approach to automation helps to fix human inter-temporal biases in decision making, but not necessarily systematic inefficiencies of the pricing system. As mentioned above, this approach has deep roots in the behavioral decision making literature. One of its main benefits is that it captures the decision maker's expertise and applies it consistently. Second, deciding on optimal prices requires making strong assumptions about the nature of demand, such as: the degree to which consumers are forward looking or the competitive and regulatory environment. Our approach to automation does not require making such assumptions, as it only builds on salespeople's historical decisions. Third, and related to the second point, our goal is to implement the pricing decisions in real-time. Building a demand model and determining optimal prices given demand at quote occasion, may not be feasible in real-time. As we demonstrate, automation via JB is easily implementable in real-time. Finally, given the sensitivity of replacing salespeople with an automated solution we opted for a model that relies on the individual salesperson expertise rather than what would have been a black-box model to the sales team.

Similar to our experimental settings, decision support systems (DSS) often offer model-based advice to the decision maker (e.g., Sharda et al., 1988; Eliashberg et al., 2000; Lilien et al., 2004). On a subtle, yet possibly critical, reversal to the setting used in our experimental design, the forecast literature has demonstrated that using the model's predictions as default with the expert adjusting it based on domain-expertise led to improvement in accuracy of sales forecasting (e.g., Mathews and Diamantopoulos, 1986; Nikolopoulos et al., 2005). Interestingly, Nikolopoulos et al. (2005) found that large adjustments to the statistical forecast were beneficial, while small adjustments were not, confirming the theoretical justification for using models and human experts for different cases.

Our work goes beyond decision models and support systems not only in automating the salesperson's pricing behavior, but also in determining which cases the salesperson should price and which the model should price with no additional input from the expert. Although our experiment, due to the large and immediate impact the treatment could have on the company's profits, used a more traditional DSS-like setting, our automation hybrid moves from decision support to decision allocation and allows the model to make decisions autonomously and automatically. This design overcomes the challenges of automated advice adoption by experts. To our best knowledge, this is a first application of a hybrid approach to decision making that uses a JB model of the expert in an empirical business application.

We also add to the literature on automation by providing an empirical test for automating the B2B salesperson's job. While automation made a long way in substituting for human tasks, automation of soft skills is still sparse (Deming, 2017). Research in labor economics shows that automation can substitute workers in performing tasks that follow explicit rules, while it complements them in performing non-routine problem solving and communication-based tasks (Autor et al., 2003). The salesperson's job is a combination of repetitive, technical calculation of prices based on quote characteristics, and delicate use of soft skills and communication to understand the client's state of mind and maximize profits. Indeed, we find that using the model to make pricing decisions when a standard pricing formula applies, but building on human skills for making out-of-the-ordinary pricing decisions that require judgment and case-based consideration, generates higher profits than do either the model or the salesperson solely (e.g., Blattberg and Hoch, 1990).

3 The Model of the Salesperson

Our approach to automation is to create a model of each salesperson, that will learn her pricing policy based on her pricing history, and apply it to new incoming quotes. For every salesperson separately, we estimate a model of previous pricing decisions as a function of a set of variables available to the salesperson at the time of decision. Although we observe the outcome of the offered price quote, i.e., whether the client accepted it or not, it is not included in the model, because the goal is to create a judgmental bootstrap model that mimics the salesperson's pricing behavior. Then, the model can be used to replace every salesperson with a consistent and automated version of herself to price a new set of quotes.

3.1 Data

The empirical context and data we use to calibrate the model of the salesperson come from a U.S.-based metals retailer that supplies to local industrial clients. The company has sales teams in three locations in Pennsylvania, New York and California. In each of these locations there is a team of salespeople servicing mostly, but not exclusively, clients from the area. The retailer buys raw aluminum and steel directly from the mills, cuts it according to the specification provided by the client and ships the product to the client. Clients may be small to medium sized industrial firms (e.g., machine shops, fabricators or small manufacturers). The company sells thousands of SKUs in nine product categories, seven of which are sub-categories of aluminum (the other two: stainless steel and other metals, represent less than 2% of the lines in our data, see Table A1 in Web Appendix A). Aluminum categories vary in terms the shape of the metal, their thickness and their designation (e.g., aerospace vs. commercial). Because of the large number of SKUs, the dynamic nature of this industry in terms of varying commodity prices and the high customization of products, there is no price catalog available. The salesperson has a high degree of autonomy in pricing products on a quote-by-quote basis, providing different prices to different clients and even different prices to the same client over time.

A client may request a price quote via email, fax or by calling the supplier. Although the work flow in the firm allows any available sales agent to pick up the call and provide a price quote, most clients interact with the same salesperson on most purchase occasions.

When requesting for a price quote, the client specifies the requested metal, size of the piece, if cutting is required, and the quantity. A quote from a client may include only one SKU or multiple SKUs, which we define as lines. After receiving the order's specifications, the salesperson provides a price quote[2]. Salespeople are guided and incentivized to maintain high price margins. Although pricing to clients is done by unit or by weight unit, salespeople report to and are evaluated by the management based on price margins. Salesperson s calculates price margin for line l in quote q for client i as follows:

$$m_{lqis} = \frac{p_{lqis} - c_{lq}}{p_{lqis}}, \tag{1}$$

where c_{lq} is the cost per pound that the company paid to buy the material and p_{lqis} is the price per pound provided by salesperson s for client i for line l of quote q[3]. After receiving the price quote, the client decides whether to accept or reject the quote given the price in the quote. In this industry price negotiation beyond the first level negotiation of price quote and acceptance is rare. We verify this empirically by comparing the initial price from the quote to the final invoice price, and find the prices to be identical in over 99% of the cases.

The data include transaction level information of price quotes spanning 16 months from January 2016 to April 2017. The sample includes 3,863 clients with an average of 36 product requests per client[4]. Each of the 17 salespeople in the sample made on average over 8,000 pricing decisions. A sales order may include one or more products (lines), each line is priced separately. The sample includes 67,851 price quotes with an average of about 2 lines per quote, totaling in 139,869 pricing decisions (every line is a "pricing decision"). 56.9% of the

[2]Shipping costs are priced separately as an additional line in the quote. We do not model those costs.

[3]A small number of SKUs are not stocked and priced by weight, but by length. We later account for that in the pricing model

[4]We removed from this analysis clients that had only one quote, and hence do not allow estimating a reliable pricing model, clients defined by the company as either contractual or semi-contractual and rare cases of lines with missing or negative price or cost. Additionally, and following the company's recommendation, we removed orders of over 8,000 lbs. or orders at the bottom 1% of orders by weight. Such orders are treated differently by the company and are often priced by a manager or follow pre-defined rules.

quotes were accepted by the clients (i.e., converted into sales orders). See Table 1 for line level summary statistics of the data.

Table 1: Descriptive Statistics of Quotes and Orders per Line

	Mean	Std. dev.	Lower 10%	Median	Upper 90%
Line margin	0.41	0.20	0.20	0.36	0.72
Price per lb.	4.78	25.06	1.67	2.60	7.19
Cost per lb.	1.98	10.64	1.18	1.40	2.74
LME† price per lb.	0.76	0.07	0.68	0.75	0.86
LME price volatility	0.01	0.00	0.00	0.01	0.01
Weight (in lbs.)	352.30	683.54	16.09	117.00	892.77
Client recency (in days)‡	61.86	207.92	1.00	13.00	120.00
Client frequency (per week)‡	0.62	0.68	0.08	0.41	1.39
Client previous order $ amount (log)‡	6.52	1.39	4.88	6.39	8.37
% of quotes priced by same salesperson	0.78	0.31	0.14	0.93	1.00
Total = 139,869					

†London Metal Exchange

‡Calculated at the product category level

3.2 Model Specification

As mentioned above, to standardize across products and order sizes the firm uses price margins as opposed to price or price per pound to evaluate its pricing strategy. Therefore, we use price margin in building the automated pricing model. Price margins are defined per Equation 1 and are calculated at the line level. Because the firm always prices above cost, price margins could range from zero to one, and are somewhat skewed to the left. The average line price margin in the data is 41% and the median is 36%. Consequently, we use the logarithmic transformation of price margin as the dependent variable in the pricing model.

In building the model we attempt to include all the information available to the salesperson at the time of the pricing decision. We conducted several interviews with senior management and salespeople in the firm to get an idea of the information flow along the pricing process. Additionally, we capture all of the information recorded on the firm's CRM software that salespeople use when determining prices (see a screenshot of the CRM system in Web Appendix A). The model includes the following variables:

a. **Product category**. Dummy variables for eight out of nine product categories the retailer sells (Baseline category is Aluminum - Cold Finish).

b. **Weight**. Log of total line weight in pounds.

c. **Relative weight**. While 57.6% of the quotes include only one line, there may be dozens of product specifications requested within the same quote. Pricing may differ depending on the relative weight of the line in the overall order, due to quantity discount at the quote level.

d. **Cut**. Made-to-order piece often require processing. We include cut in the margin equation as an interaction between the cut dummy variable and $1/weight$.

e. **Cost**. The cost per pound for the requested part number as displayed to the salesperson in the CRM system, which reflects the price the company paid for the material.

f. **Commodity market prices**. Salespeople have access to market prices published by the London Metal Exchange (LME). We include the daily LME price per lb. as well as the volatility of LME prices in the week prior to the date of the quote (measured by the LME standard deviation during the past 5 business days).

g. **Foot-base products**. A dummy variable for whether the product is priced per feet rather than per lb (3.5% of the items).

h. **Client characteristics**.

   (a) **Priority**. The firm prioritizes clients based on orders volume in the preceding twelve months. Priority A is the highest for clients with order volume of at least $100,000, and priority E is the lowest for clients with spending of less than $5000 in the past 12 months. Priority P is given to clients with "E" orders volume that have a potential (judged by the management) to become high priority clients. We include priority in our model using a set of dummy variables. A client's priority may change over the data window because it is updated by the firm every six months (baseline priority is Priority A).

   (b) **Recency, frequency and monetary - RFM**. Recency is defined as days since the client's last quote request from the same product category; frequency is defined as the client's running average of requests from the product category per week; and monetary is defined as the log of the total $ amount of the client's last order in the product category.[5]

[5]In the calculation of RFM measures we include quotes that were not converted to sales, under the assumption that the client decided to purchase the product somewhere else. To initialize the recency and

(c) **Client random effect**. One of the most prominent characteristics of B2B pricing is that prices can vary across clients (Khan et al., 2009). To account for client-specific pricing based on the client's identity we include client random effect in the model.

i. **Client-salesperson history**. Relationship with the client could affect the salesperson's pricing behavior. On the one hand, long term familiarity with the client can increase the salesperson's persuasion power. On the other hand, it may bias her pricing decisions (e.g., pricing may become too lenient). As a measure of the salesperson-client relationship we calculate the proportion of quotes up-to-date that the salesperson priced with the focal client out of the total number of quotes received by the retailer from the client (i.e., we measure to what extent this is the client's regular salesperson). On average, the same salesperson handles the client nearly 80% of the time.

j. **Time dummies**. To control for any time trends, we include quarter dummies (baseline quarter Q1 of 2016).

3.3 Model Estimation and Results

We estimate a linear regression separately for each salesperson to extract the weight each salesperson puts on each variable in setting the price margins for the requested product specification. The price margins equation is specified in Equation 2: for each line l of each quote q priced by salesperson s for client i in the sample, we regress the logistic transformation of the price margin m_{lqis} (as defined in Eq. 1), on the set of line characteristics and time-varying client characteristics, x_{lqi}, as well as salesperson-client random effect, α_{is} for salesperson s and client i

$$log\left(\frac{m_{lqis}}{1-m_{lqis}}\right) \sim \alpha_{is} + \boldsymbol{\rho_s x_{lqi}} + \epsilon_{lqis}, \tag{2}$$

where ϵ_{lqis} is a normally distributed random shock.

Note that the subscript s in Equation 2 means that we estimate Equation 2 for each salesperson s separately. However, to get a sense for the effect each variable has on the log price margins we hereby show and discuss results from a mixed model with client random

monetary variables, if the client purchased before January 2016 we use the last purchase prior to January 2016. If the client is a new client we dropped the first purchase from this analysis and used it to initialize these variables. For frequency we use the running average since the client made their first quote request.

effect and salesperson fixed effect estimated on the whole sample (see Table 2). Table A2 in Web Appendix A reports average estimates across the individual-salesperson regressions).

Table 2: Bootstrap Pricing Model

Variable	**Coefficient**	**Std. err.**
Cost per lb.	-0.003***	(0.000)
LME per lb.	0.860***	(0.076)
LME volatility	-1.454**	(0.462)
Weight (log)	-0.469***	(0.001)
Relative Weight	0.270***	(0.005)
Cut/weight	0.303***	(0.007)
Foot base	-0.232***	(0.009)
Recency	0.00001	(0.000)
Frequency	-0.077***	(0.004)
Monetary (log)	0.003*	(0.001)
Regular salesperson	-0.018*	(0.008)
Priority B	0.010	(0.045)
Priority C	0.042	(0.042)
Priority D	0.189***	(0.047)
Priority E	0.299***	(0.041)
Priority P	0.036	(0.049)
2016q2	0.077***	(0.006)
2016q3	0.095***	(0.007)
2016q4	0.132***	(0.009)
2017q1	0.129***	(0.013)
2017q2	0.157***	(0.016)
Intercept	0.646***	(0.068)
Observations	139,869	
R^2	67.1%	

* $p < 0.05$, ** $p < 0.01$,*** $p < 0.001$
Note: DV is Logit transformed price margins.
Regression includes client random-effect, salesperson fixed effect and category dummies.
Baseline priority - priority A, Baseline quarter - Q1 of 2016.

The automated version of the salesperson captures salespeople's pricing policy well - the regression model explains nearly 70% of the variation in the pricing policy. Indeed, when converting log price margins back to price margins, the average predicted line price margin of 41.96% is very similar to the average observed line price margin of 41.14%.

We find that when cost increases, the salespeople decrease price margins. However, when the daily metal price increases, salespeople seem to pass through some of the increase to the consumers (controlling for the cost of the material to the company). High variability in market prices leads to lower price margins. The salespeople seem to employ quantity discount in pricing, such that larger order have lower price margins. As expected, processing (cut) increases price margins.

With respect to client behavior, the company provides lower price margins to customers who buy frequently, but salespeople charge higher price margins from clients whose previous order was large. We find that clients receive lower price margins from their regular salesperson, suggesting that relationship building may lead to lower pricing. In terms of client priority, when clients gain higher priority, they receive lower price margins.

Finally, there seems to be a positive time trend for margins. This pattern was corroborated by the company's CEO as consistent with the company's strategy of increasing price margins during the data window.

4 Randomized Field Experiment

To assess the value of automating the salesperson pricing decisions through the individual pricing models, we collaborated with the company to conduct a large-scale field experiment. Ideally, the automated prices would replace salespeople's prices altogether. However, due to the immediate impact such a pricing experiment can have on the company's profits, we were only able to provide the model's prices as (real time) recommendations, and allow salespeople to adjust their original prices accordingly.

4.1 Experimental Design

In collaboration with the B2B retailer's information technology team, we created a "price calculator", that upon receiving a new quote calculates the model's predicted price margins (defined in Equation 2) based on the quote, client, and salesperson characteristics. The calculated price per lb. is then displayed in real time as a recommendation to the salesperson. The experimental design randomly allocates incoming quotes into treatment (60% of the quotes) and control (40% of the quotes).[6] The regular pricing work flow is as follows: when a client puts a new quote request, the salesperson enters the new quote information (client ID, SKUs requested, etc.) into the CRM system. The salesperson then provides a price quote, saves it to the system, and is able to edit prices as needed. When done editing, the salesperson generates a price quote document and sends it to the client via email.

In the experimental intervention for quotes in the treatment condition, after the salesperson entered their quoted pricing information, the following message was emailed to the salesperson: *Based on your previous pricing decisions, the prices recommended for this quote are:* and below was a table displaying the product information for every line of the quote, the price that the salesperson had just entered to the system, per pound and per unit, and total per line, as well as the model's price per pound and per unit, and total per line[7] (see Figure 1a for a screenshot of the email). The salesperson could then either click *Accept suggested prices* to update the sales system to reflect the model's prices, *Accept original prices* to keep her original prices, or *Edit*, which would open an edit form (see Figure A2a in Web Appendix B). In the edit form the salesperson could accept the model's price for only some of the lines, as well as edit any price manually. Prices were automatically updated in the sales system, therefore not requiring an extra step on behalf of the salesperson. The full flow of the experiment is depicted in Figure 2.

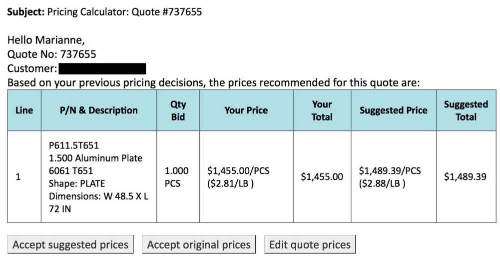
Because treatment involved an extra step of evaluating the original prices, which may, in and of itself, generate higher attention of the salesperson to her pricing decisions, an email was also sent to quotes in the control group. The control e-mail was similar to that of the treatment, except it did not include the columns displaying the model's recommended price (see Figure 1b for a screenshot of the control group e-mail). Similar to the treatment

[6]Due to the relatively small number of salespeople in the company (17 salespeople at the time), randomization was done at the quote level rather than at the salesperson level. We intentionally over-weighted treatment over control with anticipation of low compliance rates.

[7]The company has a minimum price of 30$ per line for high priority clients and 150$ per line for low priority clients. If the model's calculation resulted in a total price lower than that minimum, we adjusted the price per lb. and the total per line to reflect the minimum price.

Figure 1: Emails Sent to Salespeople as Part of the Field Experiment

(a) Treatment Email Format



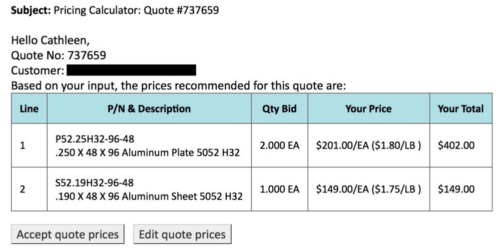
Subject: Pricing Calculator: Quote #737655

Hello Marianne,
Quote No: 737655
Customer: [redacted]
Based on your previous pricing decisions, the prices recommended for this quote are:

Line	P/N & Description	Qty Bid	Your Price	Your Total	Suggested Price	Suggested Total
1	P611.5T651 1.500 Aluminum Plate 6061 T651 Shape: PLATE Dimensions: W 48.5 X L 72 IN	1.000 PCS	$1,455.00/PCS ($2.81/LB)	$1,455.00	$1,489.39/PCS ($2.88/LB)	$1,489.39

Accept suggested prices | Accept original prices | Edit quote prices

(b) Control Email Format



Subject: Pricing Calculator: Quote #737659

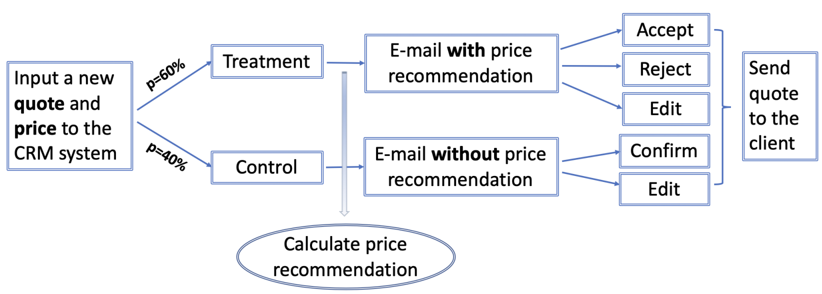
Hello Cathleen,
Quote No: 737659
Customer: [redacted]
Based on your input, the prices recommended for this quote are:

Line	P/N & Description	Qty Bid	Your Price	Your Total
1	P52.25H32-96-48 .250 X 48 X 96 Aluminum Plate 5052 H32	2.000 EA	$201.00/EA ($1.80/LB)	$402.00
2	S52.19H32-96-48 .190 X 48 X 96 Aluminum Sheet 5052 H32	1.000 EA	$149.00/EA ($1.75/LB)	$149.00

Accept quote prices | Edit quote prices

condition e-mail, the control condition e-mail allowed the salesperson to either *Accept* her original prices or *Edit*, in which case a control edit form, similar to the one of the treatment condition only without recommended prices, was displayed (see Figure A2b in Web Appendix B.1). If edited, prices were updated directly in the system. The salesperson's next step in both control and treatment flows was to go back to the system, generate the price quote document and send it to the client as she would have done without the experiment.

Figure 2: Flow of Field Experiment



Note, that when entering her original price quote, the salesperson did not know whether this quote belongs to treatment or control (i.e., whether she will receive a price recommendation or not), hence the original price quotes are independent of the experimental manipulation. This unique design gives us knowledge of three data points for each quote (control and treatment): the original price set by the salesperson, the model's recommended price (which we calculated in both control and treatment, but made available to the salesperson only in the latter) and the final price that the salesperson provided to the client. We use this information in subsequent analyses to shed light on salespeople behavior in the experiment.

Prior to the commencement of the experiment, we let the salespeople experience the tool for four business days, during which we adjusted the tool to fit best into their work flow and corrected any technical issues that arose. During those pre-test days we visited two out of the three company's locations (New York and Pennsylvania) and conducted several phone conversations with the third location (California) to make sure salespeople were comfortable using the tool and understood its flow.

We ran the experiment for eight consecutive business days. Our data include 2,075 quotes by 1,045 clients, with a total of 4,142 pricing decisions (some quotes had multiple lines, and each line is a pricing decision).[8] The average compliance level with the tool (i.e. quotes for which salespeople either fully accepted the recommended prices or edited prices in the direction of the recommendation), was 19.48%. We note that in our analyses we use intention to treat (price recommendation) as opposed to compliance (whether the salesperson adopted our price recommendation) because compliance is endogenous. Hence, considering the compliance levels, our results may underestimate the true effect of automation. We further discuss salespeople compliance behavior in section 4.2.4.

4.1.1 Randomization

Every incoming quote was assigned to the treatment group with probability 0.6 or to the control group with probability 0.4. Randomization was done by the company, and as expected, 58.3% of incoming quotes were assigned to the treatment condition. As with any experimental design, the first order of business is to examine that the randomization was preformed correctly. We performed a randomization check for different quote variables such as average cost, total weight, number of lines requiring cut and number of lines per quote, as well as the original price set by the salesperson, the model's price and the difference between them.

[8]We excluded from the analysis approximately 10% of the lines with cost or price per lb. larger than $16 that often relate to irregular orders as well as lines for which the final profit margin was negative (i.e., the price offered to the client was lower than material cost). The results reported in Section 4.2 are robust to including these data points.

We find no statistically significant difference between the treatment and control conditions (all $p-values > 0.23$; See Web Appendix B.2). Therefore, we can conclude that no omitted variables made the salespeople or the model price differently under the two conditions, prior to receiving the treatment.

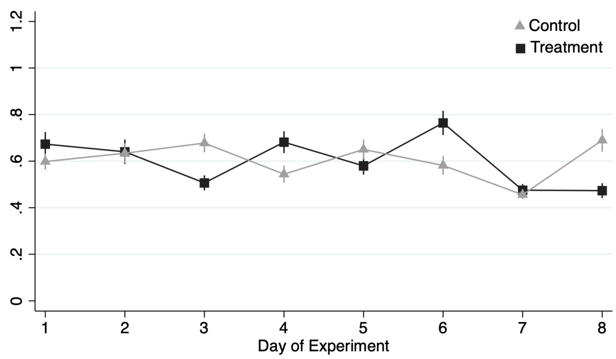
4.1.2 Stable Unit Treatment Value Assumption

The relatively small number of salespeople in the company was key reason to randomizing at the quote level, rather than at the salesperson level. When choosing a design where some of the salesperson's quotes are treated while others are not, there is a risk of violating the stable unit treatment value assumption (SUTVA, Rubin 1980). That is, that treatment of quotes in the treatment group "contaminates" the quotes in the control group because the same salesperson prices both the treatment and the control quotes. One possible mechanism through which such contamination may occur is learning. If, for example, the salesperson receives a few consecutive treatment emails recommending higher prices than her original prices, she may adjust her pricing upwards in the following treatment and control quotes.

To evaluate the extent to which learning is affecting pricing, we compare the difference between the model's price per lb. and the salesperson's original price per lb. over time, for control and treatment quotes. While we expect that the model maintains the same pricing rule, if the salesperson learns over the course of the experiment to price more systematically and more similarly to the model, the difference between the salesperson's original prices and the model's prices will decrease over time. Figure 3 shows the average difference between the model's price and the original salesperson's price over the eight days of the experiment. We see no apparent pattern in the difference between the model and the salesperson pricing in either of the experimental conditions over the course of experiment, suggesting that violations of SUTVA due to learning are likely to be minimal.

To statistically test possible violations of SUTVA via the effect of one quote on a subsequent quote, we tested whether the treatment given to a quote affects the pricing by

Figure 3: Average Difference between Model-Price Per lb. and Original Price Per lb. Over the Eight Days of the Experiment: Treatment vs. Control



the same salesperson in the following quote. For each line in a quote we regress the absolute difference between the model's price per lb. and the salesperson's original price per lb. on a dummy variable indicating *whether the previous quote priced by the salesperson was treated*, controlling for the set of line characteristics, time-varying client characteristics, salesperson fixed effect and salesperson-client random effect. If SUTVA violations exist, the salesperson will price more similarly to the model following a treatment quote, as they can learn from the treatment quote pricing. However, we do not find a statistically significant relationship between whether the previous quote belongs to the treatment condition and the difference between the salesperson's and the model's prices in the current quote ($\beta_{previous\_quote\_treated} = 0.0019, p = 0.959$). See Web Appendix B.3 for full details of this analysis.

Beyond investigating SUTVA, another implication of this analysis is that, at least within the eight days of an experiment, we cannot expect a decision support automation tool such the one we used to train salespeople to become more consistent on their own, without receiving a price recommendation.

4.2 Field Experiment Results

4.2.1 Non-parametric Test

To test the effectiveness of the treatment (providing price recommendation) we compare the gross profit (GP) between treatment and control quotes. GP can go from zero to a large number. Because quotes that were not converted to sales (i.e., the client declined the offered price) have zero GP, the distribution of GP has a mass at zero. Thus, GPs in the treatment and the control are not normally distributed. Accordingly, we use a non-parametric test to compare the GPs between the treatment and control conditions. In addition, although randomization was done at the quote level, pricing is done separately, but not independently, for each line within the quote. To account for such interdependence, we cluster the standard errors across lines of the same quote. Specifically, we use a non-parametric Wilcoxon rank sum test with clustered standard errors for lines within a quote (Datta and Satten 2005, Jiang et al. 2017) to compare mean line gross profits between treatment and control conditions. We find that quotes in the treatment group have a statistically significantly higher gross profits per line relative to quotes in the control group (Diff = \$10.95, $GP_{control}$ = \$94.16, $GP_{treatment}$ = \$105.11, $Z = -2.132,\ p = 0.033$). Overall, the increase in profits corresponds to over \$26,000 for the treated quotes during the eight days of the experiment, and over \$1.4 million when extrapolated to all quotes handled by the company in a year. Thus, automation in the form of recommending salespeople their own model's prices can result in significant and substantial increase in profitability for the company.

4.2.2 Cragg Hurdle Regression Analysis

The positive effect of treatment on profits and margins could come from increasing the number of quotes that were accepted and/or from higher price margins of accepted quotes. In order to further understand the mechanism behind the positive effect of providing price recommendations to quotes in real time, we estimated a Cragg hurdle regression (Cragg, 1971). The Cragg hurdle model enables the estimation of the treatment effect separately on the two observed processes: selection (acceptance of the suggested price by the client) and GP level conditional on acceptance of the price (GP is zero if the client rejects the price offer).[9] Specifically, we use a normalized *log(1+GP)* as DV and define its distribution using the following selection model:

$$f(log(1{+}GP)|\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}) = \begin{cases} \Phi(\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}\boldsymbol{\delta}^{\mathbf{1}})[\Phi(\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}\boldsymbol{\delta}^{\mathbf{2}})/\sigma]^{-1}\phi[log(1{+}GP) - \boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}\boldsymbol{\delta}^{\mathbf{2}})/\sigma]/\sigma, & \text{if } GP > 0, \\ 1 - \Phi(\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}\boldsymbol{\delta}^{\mathbf{1}}), & \text{if } GP = 0, \end{cases} \tag{3}$$

where the top part of the equation reflects the cases in which the client accepted the quote and hence the *GP* is positive, and the bottom part, the selection process in which the quote was rejected by the client. $\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}$ includes a dummy for whether the quote was treated or not, a set of dummy variables to control for day of the experiment fixed effect, line weight, cost per lb. and whether the quote required a cut (divided by the weight).

The results of the Cragg hurdle model analysis are shown in Table 3. Controlling for line characteristics and for day fixed effect, the effect of the treatment (i.e., providing price recommendation to the quote in real time) on the probability that the client will accept the quote is positive and significant. The effect of the treatment on gross profit for the lines that were converted is not significant [10]. Thus, the treatment worked through setting prices that increase the likelihood of the client accepting the quote, but not through setting prices that lead to higher profits given quote acceptance[11].

To investigate the mechanism by which treatment led to increase in quote acceptance

[9]A Tobit II analysis would not be appropriate to separate the effect of treatment on acceptance and profits because the data is not left truncated. Not observing gross profits occurs due to client rejection of the quote an not due to truncation of the firm's profits to the negative domain.

[10]Note that in the Cragg hurdle model the exogeneity assumption is not held in the profit equation.

[11]We find similar results when running the Cragg Hurdle analysis on the treatment variable without the control variables in $\boldsymbol{x}^{\mathbf{1}}_{\boldsymbol{lq}}$.

Table 3: Cragg Hurdle Regression Analysis

Variable	Coefficient	Std. err.
Client acceptance of price		
Treatment	0.154*	(0.076)
Line weight (log)	-0.089***	(0.022)
Cost per lb.	-0.051	(0.040)
Cut / weight	-3.338	(1.558)
Constant	0.473*	(0.191)
Log line gross profit		
Treatment	0.004	(0.009)
Line weight (log)	0.115***	(0.003)
Cost per lbs.	0.039***	(0.006)
Cut / weight	1.541***	(0.222)
Constant	0.973***	(0.022)
$\log(\sigma)$		
Constant	-2.188***	(0.039)
Observations	4,142	
Pseudo R^2	27.99%	

Day fixed effects included
* $p < 0.05$, *** $p < 0.001$

by the client, we run an instrumental variable (IV) analysis for quote acceptance on the absolute value of the difference between the model's price per lb. and the final price per lb. quoted to the client, with treatment as an exogenous IV for the price difference. Because quote acceptance by the client is a binary variable we use a binary IV Probit regression (Amemiya, 1978) with clustered standard errors for lines within a quote. We estimate the following model:

$$P(sale_l = 1) = Probit(\Delta Price_l \beta_1 + \boldsymbol{x_l^2 \beta_2}) \tag{4a}$$

$$\Delta Price_l = I_T \pi_1 + \boldsymbol{x_l^2 \pi_2} + v_l^2, \tag{4b}$$

where $sale_l$ is the client's decision to accept line l (in quote q), $\Delta Price_l$ is the absolute value of the difference between the model's price per lb. and the final price per lb. for line l, and $\boldsymbol{x_l^2}$ is the same set of controls used in Equation 3. The Gaussian function for $\Delta Price_l$ includes the same set of controls $\boldsymbol{x_l^2}$, a treatment dummy I_T and a random shock normally distributed, v_l^2.

The results of the IV analysis are shown in Table 4. As expected, the term that captures the difference between the model and final price has a negative coefficient in the quote acceptance, suggesting that when the salesperson prices closer to the model (final price is more similar to model's price - smaller difference) client acceptance increases, and confirming that the treatment works through making the salesperson's pricing more similar to her model.

In addition, we run an IV regression, in which we include an interaction term between $\Delta Price_l$ and a dummy variable for whether the model recommended a higher price than the salesperson (61.8% of cases). We find that the instrumented price difference variable is more strongly related to quote acceptance when the model recommends lower prices than the salesperson relative to when it recommends higher prices ($\beta_{\Delta Price_l \times I_{model\_higher}} = 1.43, p < 0.001$), suggesting that the model affects quote acceptance by recommending lower prices to the salespeople. The full details of this analysis are shown in Table A5 in the Web Appendix.

Table 4: Instrumental Variables Analysis for Line Conversion

Variable	**Coefficient**	**Std. err.**
$\Delta Price$	-0.938***	(0.048)
Line weight (log)	-0.284***	(0.027)
Cost per lb.	0.174***	(0.039)
Cut / weight	13.86***	(2.275)
Constant	1.655***	(0.205)
Observations	4,142	

Day fixed effects included.
*** $p < 0.001$

While we have now established that following the model's recommendation leads to higher profitability (through increased acceptance), a question may arise: how does the model, by simply mimicking the salesperson's pricing policy, lead to better outcomes? The judgmental bootstrap literature suggests that systematically applying the expert's decision policy will lead to better predictions by mere consistency of re-application of the expert's judgment. Consistent with this account, we find that the coefficient of variation of the model's predicted price margins (0.372) is significantly smaller than that of the salespeople's price margins (0.432; $p < 0.001$). That is, the model leads to lower variance in the pricing decisions. In the following sections we attempt to further investigate how the treatment works by conducting heterogeneous treatment effect as well as salespeople compliance analyses.

4.2.3 Heterogeneity in Treatment Effect

Prediction Intervals. We would expect the model to perform better and help salespeople in situations where the model has more data and hence more accurate predictions. When orders are complex or odd the model predictions are likely to be less accurate and hence less helpful. To investigate this conjecture, we calculated prediction intervals (PIs) for each of the model's price margin recommendations. We then include these mean-centered PIs as main effects and their interaction with treatment (heterogeneity in treatment effects) in $\boldsymbol{x^1_{lq}}$ in Equation 3 in the Cragg analysis. Prediction intervals, by definition, are larger when model covariates are extreme and thus the model's prediction is less certain. Therefore, we would expect the treatment effect to be weaker when intervals are larger.

Indeed, we find that when the PIs are large, the treatment is weaker both for conversion and gross profit, significantly so for gross profit ($\beta_{TreatmetXinterval} = -0.112$, $p = 0.016$) and directionally for conversion ($\beta_{TreatmetXinterval} = -0.052$, $p = 0.904$). That is, as expected, the model's recommendation leads to higher profitability when the model is able to capture the salesperson's past pricing policy and consistently apply it to unseen cases (see Table A6 in Web Appendix B.4 for the full results of this analysis).

Salesperson Characteristics. Theoretically, it would be informative to investigate heterogeneity in treatment effect by salesperson characteristics such as consistency of past pricing decisions, expertise or tenure with the company. However with only 17 salespeople in the experiment, such analyses can be suggestive at best. Directionally, we find that salespeople for whom our model of the salesperson had a higher coefficient of determination, R^2 (i.e., salespeople who were more consistent in the past) had lower treatment effect (lower increase in qoute accpetance and GPs due to treatment (see Figure A3 in Web Appendix B.4). This directional result is consistent with the observation that these salespeople's pricing behavior was more consistent to begin with, hence the model's consistency is contributing less to them.

Behavior Change Following Conversion. So far we have demonstrated that, on average, the treatment leads to price margin reduction and consequently an increase in conversion rate. In this analysis we demonstrate that in some cases the treatment attenuates salespeople's tendency to lower price margins. We run an analysis for the logit transformation of line price margin of line l in quote q by salesperson s, m_{lqs}, and include two measures of previous quotes outcome: I_s^{prev}, a dummy indicating whether the previous quote priced by the salesperson was converted to a sale or not, and $prev\_avg_s$, the average conversion rate of salesperson s in the previous business day. In addition, we include in the analysis I_q^T, a dummy for whether the current line (quote) was treated or not, the interaction between I_T and $prev\_avg_s$ (which is the cleaner of the two conversion measures, given that after a day salespeople are more likely to have full information of a quote's outcome than immediately after pricing the quote). We also include $\boldsymbol{I_{day}}$, a set of day of experiment dummies and fit the following equation:

$$log\left(\frac{m_{lqs}}{1-m_{lqs}}\right) = I_q^T\epsilon_1 + I_s^{prev}\epsilon_2 + prev\_avg_s\epsilon_3 + I_q^T \times prev_avg_s\epsilon_4 + \boldsymbol{I_{day}\eta} + w_{lqs}, \qquad (5)$$

where w_{lqs} is a normally distributed random shock.

The results of of this analysis are shown in Table 5. Supporting previous findings of increased conversion outcome, treatment leads to lower price margins. Success in converting previous quotes has a negative effect on price margins, i.e., salespeople tend to lower prices following a successful sales day, possibly to continue the winning streak. However, the treatment attenuates that inter-temporal behavioral effect, reversing most of the previous day effect.

Table 5: Line Price Margin by Previous Quote or Day Conversion

Variable	**Coefficient**	**Std. err.**
Treatment	-0.0327*	(0.015)
Previous quote accepted	-0.00278	(0.007)
Previous day average conversion rate	-0.0730**	(0.023)
Treatment × prev. day conversion rate	0.0619*	(0.027)
Constant	0.471***	(0.015)
Observations	3,666	

Day fixed effects included.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4.2.4 Compliance Analysis

One of the largest risks when conducting an experiment that requires cooperation of participants is lack of compliance. Specifically, when offered to rely on algorithmic decision aids, people may demonstrate *algorithm aversion* and limit their use of the aid tool. Among the reasons for this aversion are the belief that humans can reach near-perfection in decision making (Einhorn, 1986) and that human predictions improve through experience (Highhouse, 2008). The latter is especially important when it come to experts decision making. Experts tend to over-weigh their experience and expertise, which often leads to poor predictability (Arkes et al. 1986; Camerer and Johnson 1991). Moreover, when facing (inevitable) algorithmic errors, people are less likely to trust and use the algorithm (Dietvorst et al., 2015).

The experimental design, in which salespeople received the model's prices as recommendations and could use it at their discretion, posed a risk of low compliance to our experiment. During the experiment salespeople expressed great confidence in their own judgment. For example, one salesperson said "I am not likely to follow the recommended price because I had already put a lot of thought into pricing the quote and considered everything there is to consider". Moreover, many salesperson said that while the tool may be useful for other salespeople, their clients (or the quotes they typically price) are "different". Overall, salespeople's reluctance to accept the model's price could make it harder to identify the true effect of the treatment.

Our experimental design allows us to assess compliance, because we have information about the salesperson original pricing decision made prior to exposure to treatment (only after the salesperson inputs into the system a price for the new quote, the quote is randomly assigned to treatment or control and the model's price is displayed for treated quotes). Accordingly, in what follows we analyze the compliance patterns to shed more light on the observed treatment effects. However, because compliance is endogenous to the decision maker and to the quote and client characteristics, the analysis in this section is more descriptive than causal.

Table 6 depicts the compliance patterns based on whether the direction in which salespeople changed their price, relative to their original price, is consistent or inconsistent with the model's recommendation[12]. First looking at the control condition, we find that salespeople have an insight into adjusting their price in the right direction. In the control condition salespeople did not see the model's recommendation; yet, they adjusted prices on their own in higher rates in the direction of the model than in the opposite direction (9.52% price decreases vs. 6.25% price increases when the model's price was lower; 16.93% price increases vs. 3.78% price decreases when the model price was higher). Overall, in the control condition, salespeople adjusted their original price in the direction of the model in 14.05% of the cases (64 price decreases, 179 price increases). Turning now to the treatment condition we see an even higher rate of "compliance" with the model's recommendation. In 19.48% of the cases (133 price decreases, 337 price increases; see bold face numbers) the salespeople changed their price in the direction of the model, a lift of 37.7% (5.4% percentage points) in compliance over the control condition.

Table 7 shows quote conversion rates by model recommendation and salesperson behavior. Cases in which the salesperson changed the price in a direction congruent with the model's recommendation are in bold in both treatment (top) and control (bottom). As expected and in line with the results of the Cragg analysis, the largest increase in conversion,

[12]Note, that this measure of compliance based on a price change is conservative because in some of the cases in which the salesperson did not change their price, the model recommended a price similar to the salesperson's price.

Table 6: Compliance Patterns by Model Recommendation

	Model's recommendation	Salesperson's behavior: Decreased price	No change	Increased price	Total
Treatment	**Decrease price**	133	732	44	909
		14.63%	80.53%	4.84%	100%
	Increase price	57	1,110	337	1,504
		3.79%	73.80%	22.41%	100%
	Total	90	1,842	381	2,413
		7.78%	76.34%	15.79%	100%
Control	**Decrease price**	64	566	42	672
		9.52%	84.23%	6.25%	100%
	Increase price	40	838	179	1,057
		3.78%	79.28%	16.93%	100%
	Total	104	1,404	221	1,729
		6.02%	81.20%	12.78%	100%

from 25% in control to 50.38% in treatment, comes from following the model in decreasing the price ($p < 0.001$). When increasing the price following the model's recommendation, we do not expect an increase in conversion because the price was increased (39.17% in treatment vs. 37.99% in control, $p = 0.79$).

Table 7: Conversion Rates by Model Recommendation Compliance

	Model's recommendation	Conversion rates: Decreased price	No change	Increased price	Total
Treatment	**Decrease price**	**50.38%**	48.36%	52.27%	48.84%
	Increase price	59.65%	54.23%	**39.17%**	51.06%
	Total	53.16%	51.90%	40.68%	50.23%
Control	**Decrease price**	**25.00%**	42.40%	26.19%	39.73%
	Increase price	42.50%	47.37%	**37.99%**	45.60%
	Total	31.73%	45.37%	35.75%	43.32%

The off-diagonal, in which salespeople went against the recommendation of the model, also reveals an interesting pattern. In these cases we find significantly higher conversion rates in treatment relative to control. Due to self-selection we can only speculate about the reason for this increase, but these results are consistent with what the judgmental bootstrapping literature calls "broken-leg" cases Meehl (1954). The term "broken-leg" describes a scenario in which a model can successfully predict whether one will go to the movies in any given night, but fails in the rare and unexpected case in which one broke their leg that day, and the model is unaware of the incident. The analogy of the "broken-leg" to our context occurs when salespeople had private information leading to a conviction in a positive outcome. These are likely to be the cases in which the salesperson decides not only to not follow the model, but to change their original price against the model's recommendation. Of course, if this information was codeable it could have been incorporated in the model to improve prediction. To complete the picture of compliance, Table A7 in Web Appendix B.4 shows gross profit by model recommendation and salesperson response. Consistent with the insignificant effect of treatment on gross profits given quote conversion in the Cragg model, we do not find significant differences in gross profits by compliance.

One of the reasons suggested by the judgmental bootstrapping literature for why a model of the expert improves the expert's decision making is that it helps the expert avoid inter-temporal biases due to, for example, reacting to previous successes in independent decisions (Coval and Shumway, 2005). To investigate this issue, we look at whether a salesperson's likelihood to comply with the model depends on the salesperson's success in converting previous quotes. We run a mixed logit regression for whether the salesperson complied with the model or not. Following Table 6 compliance is defined as adjusting the original price in the direction of the model's recommendation, either upwards or downwards [13]. We find that following a conversion of a quote, and even more so following a successful day, salespeople comply less with the model (see Table 8).[14] This could hint to over-confidence, where salespeople are not seeking the model's advice following success as they are confident in their own pricing decisions. Indeed, over-confidence has been demonstrated to be prevalent among salespeople (Bonney et al., 2016). Automation can help mitigate this bias by smoothing out inter-temporal over-confidence due to winning streaks.

Overall, the compliance analysis suggests a moderate level of compliance, which led to

[13]This is a conservative measure of compliance because in some cases the salesperson did not adjust the price and the model's price was similar to the original price.

[14]The insignificant effect of previous quote could be due to the fact that information about whether the quote was converted may not be immediately available.

Table 8: Compliance by Conversion Rates

Variable	Coefficient	Std. err.
Previous day conversion rate	-1.485***	(0.355)
Previous quote conversion	-0.251	(0.140)
Constant	-1.469***	(0.264)
Observations	3,666†	

Day fixed effect and client random effect included.
†Observations from days 2-8 of the experiment.
*** $p < 0.001$

higher quote conversion, particularly when the model recommended to decrease the price. When salespeople decide to go against the model's price, it is often when the quote had a higher chance of conversion, hinting towards the existence of non-codable information. Finally, we find that following a successful day with high proportion of conversions, salespeople tend to become over-confident in their abilities and exhibit low compliance with the model prices.

5 Counterfactual Analyses

While the experiment allowed us to directly investigate the causal effect of automation on profitability, as with any field experiment, there are some limitations and constraints. First, the firm only allowed us to provide the model's prices as a recommendation to salespeople, rather than replace them completely in providing price quotes to clients. Particularly, given the relatively low compliance levels, this prevents us from fully testing the value of automation. Second, because salespeople endogenously decided when to comply with the model, we cannot directly assess under which conditions it would be most profitable to use the model and under which conditions to defer to the salesperson's pricing. Finally, given the cost involved in running such a pricing experiment, we were only able to run the experiment with one bootstrap (linear) pricing model. However, it is possible that more flexible non-linear or machine learning models would better capture the salesperson's pricing policy. To answer these questions, we build a demand model that mimics the client's decision to accept or reject the quote given the quoted price. We then run a set of counterfactuals comparing profitability under different pricing schemes based on versions of full and hybrid automation, and with more flexible machine learning models of the salesperson.

While we did not use the client's decision of whether to accept or reject the quoted price in creating the automated JB model of the salesperson, we do observe it in the data. The client's response can be used to estimate a demand model for aluminum products and predict the client's behavior under different pricing schemes. Note, that while pricing is done at the line level, the client's acceptance decision is typically done at the quote level, either accepting or rejecting all the lines in the quote. Therefore, we estimate demand as well as calculate profit counterfactuals at the quote level.[15] Put formally, for each quote q requested by client i, based on observed prices p_{qi} and predicted prices $\hat{p_{qi}}$ (calculated based on the model's predicted margins), we calculate predicted acceptance probabilities, based on the actual price, $Pr(p_{qi})$, and the model's price, $Pr(\hat{p_{qi}})$. We then calculate for quote q requested by client i:

$$\Pi_{qi} = (p_{qi} - c_q) \times Pr_{qi}(p_{qi}), \tag{6}$$

$$\hat{\Pi}_{qi} = (\hat{p}_{qi} - c_q) \times Pr_{qi}(\hat{p}_{qi}), \tag{7}$$

and compare expected profits based on the difference between Π_{qi} and $\hat{\Pi}_{qi}$.

5.1 Data for Counterfactuals

Because the counterfactual analysis requires leaving hold out data for validation, we use a longer period to estimate demand and price margins models than the period used to estimate the pricing bootstrap model in the experiment. Specifically, we use a data period that spans two years of transactions between 2015 and 2016, using the first eighteen months for calibration and the last 6 months for validation (prediction). Overall, the calibration data include

[15] Only about 5% of the quotes in the sample were partially accepted , i.e., the client accepted the price for some of the lines in the quote and rejected the price for others. In the analysis we handle these quotes as two separate quotes: one accepted, and one rejected.

21 salespeople making 104,336 pricing decisions for 3,787 clients over the course of eighteen months. Table A8 in Web Appendix C shows summary statistics of the counterfactual analyses data.

As discussed previously, the company exhibited a trend of increased margins over time. Specifically, the company enjoyed higher margins since Q1 2016 (See Table A9 in Web Appendix C). We capture such a time trend in the pricing model by including quarterly dummies.[16] Table A10 in Web Appendix C shows the estimates of the pricing model (similar to Table 2 but for the counterfactual calibration data).

5.2 The Demand Model

To calculate expected profits we need to estimate the probability of quote acceptance given price (the last term in Equations 6 and 7). A purchase event is initiated when the client approaches the firm with a request for a price quote for one or more specifications of material, size, weight and cut for aluminum products. The salesperson prices all the lines of the quote and then the client decides whether to accept or reject the price quote. For each client, we observe multiple quote requests and the corresponding accept or reject decisions. We assume that the utility for client i from accepting quote q is:

$$u_{qi} = \beta_{1i} + \beta_{2i}\ p_{qi} + \boldsymbol{\beta_z}\ \boldsymbol{z_{qi}} + \gamma\ \Delta P_{qi} + \sigma\ \eta_{qi} + \xi_{2qi}, \tag{8}$$

where β_{1i} is a random intercept for client i, and p_{qi} is the price offered for quote q made by client i[17]. $\boldsymbol{z_{qi}}$ is a vector of covariates that includes recency (days since the last quote request by client i), regular salesperson (the ratio of quotes priced by the salesperson out of the total number of quotes by this client up to the date of the current quote), log weight of

[16] To extend the time trend to the validation period we multiply the validation period predicted prices by the ratio of the average log price margins in the validation period (q3 and q4 of 2016) to the average log price margins of the last quarter in the calibration period (q2 of 2016).

[17] Estimating the demand model with reference prices instead of price yields similar results.

quote j, LME price on the day of quote j, LME volatility on the week prior to quote j and a set of dummies, one for each product category included in the quote.

To control for possible endogeneity of the price due to either targeted pricing for specific clients or unobserved random shocks that may affect both pricing and demand, we use a control function approach (Petrin and Train, 2010). For the control function we use cost, cut and quarter fixed effect as exclusion instrumental variables that affect acceptance; and client random effect to control for potential endogenous effect in targeting prices to clients based on their estimated likelihood to accept. Specifically, we believe that the cost the company paid for the product is a good instrument for price as its effect on clients' demand should primarily go through the price of the product. One may be worried that cost may be correlated with competitive pricing. However, given that the cost is determined based on the price the company paid when buying the product, and products tend to stay in the company's warehouse for as long as six months, correlation between wholesale price and competitive prices is likely to be low. To further test the validity of this instrument we ran the Hausman specification test adapted for control function estimation (Hausman and McFadden, 1984) for our main IV, cost. The test suggests validity of instrumental variable approach (Chi-Sq=18.26, $p < 0.001$).

The Gaussian control function price equation for client i and quote q is:

$$p_{qi} = \lambda_i + \lambda_{cost}\ cost_q + \lambda_{cut}\ cut_q + \boldsymbol{\lambda_{quarter}}\ \boldsymbol{quarter_q} + \xi_{1qi}, \tag{9}$$

where p_{qi} is the actual price for quote q requested by client i, λ_i is a client i random-effect intercept, $cost_q$ is the cost of the material for quote q, cut_q is the ratio of lines in the quote that require special processing, and $quarter_q$ is a set of dummy variables for five out of the six quarters in the calibration data. ξ_{1qi} is a random shock normally distributed with a zero mean and a variance σ_{1q}.

The last two terms prior to the random shock ξ_{2qi} in Equation 8 reflect the specification of the control function approach. $\Delta P_{qi} = p_{qi} - \tilde{p_{qi}}$, is the residual of the control function price equation, where $\tilde{p_{qi}}$ is the fitted value of Equation 9 for the specific values of quote j and η_{qi} is i.i.d standard normal.

Finally, assuming that ξ_{2qi} is extreme value distributed, the probability that client i will accept quote q follows the binary logit specification:

$$Pr_{qi} = \frac{e^{u_{qi}}}{1 + e^{u_{qi}}}. \tag{10}$$

We estimate the demand model in two stages. First, we estimate control function random effects model to estimate $\Delta P_{qi} = p_{qi} - \tilde{p_{qi}}$; then we use Hamiltonian Monte Carlo (HMC) with No U-turn sampler (NUTS) to estimate the demand model. Web Appendix D includes the full details of the demand model estimation and results. In what follows we use results from the demand model estimation to calculate the profit counterfactuals.

5.3 Profits of Model Pricing Vs. Profits of Salesperson Pricing

Using the price margins model (Equation 2) together with the demand model that predicts the client's acceptance behavior as a function of different pricing schemes, we can compare expected profits based on the model-of-the-salesperson predicted prices and based on salespeople's prices (following Equations 6 and 7).

To calculate the counterfactuals profits, we use the hold-out sample of the last six months of the data, which were not used in estimating the demand or the pricing models, with a total of 11,621 quotes. In the hold-out sample, the observed average price per lb. per quote is \$3.41, and the average predicted price per lb. based on the JB model is \$3.28. We confirm that, similarly to the experiment, the coefficient of variation of the model's prices is smaller than that of observed prices (0.584 vs. 0.693, respectively, $p < 0.001$). The expected acceptance probability based on the original pricing scheme, 61.1%, is comparable to the actual observed acceptance probability, 59.3%, pointing to a reasonable aggregate demand model accuracy.

Using Equations 6 and 7 and aggregating across quotes, we find that the model's pricing scheme generates expected profits that are 4.9% higher than those of the salespeople's pricing scheme ($\Pi[\hat{p}] = \$2,536,058$ compared to $\Pi[p] = \$2,417,149$). This difference is statistically significant, based on the 95% posterior confidence intervals (PCIs) across a sample of 100 draws from the output of the HMC sampler. The actual profits for the same set of quotes were $2,345,479. Thus, consistent with experimental results, but now fully replacing the salespeople with their bootstrap model, the results of the counterfactual analysis demonstrate that the model of the salesperson does better than the salesperson in generating profits for the company. This should not be taken for granted because, as discussed previously, the B2B salesperson's work is based on soft skills, communicating with clients, understanding clients' state of mind, and using those insights to leverage pricing authority to increase profitability. For example, Elmaghraby et al. (2015) discuss the role of environmental information in making pricing decisions in B2B settings. While in the experiment the salesperson could ignore the model-of-the-salesperson in cases where such information dimmed valuable, in the counterfactual analysis the information is completely absent. Accordingly, in the next section we examine a hybrid pricing scheme that allows the salesperson to price some of the quote thus preserving some of the private information that the salesperson has and is not captured by the model.

5.3.1 Sensitivity to Alternative Pricing Models

In addition to the linear JB pricing model we test the robustness of our results to two alternative pricing models: (1) non-linear machine learning-based JB model, and (2) mixed effect model that pulls information across salespeople.

Machine-learning non-linear pricing model - Equation 2 and the analyses described thus far present a linear bootstrap model of the salesperson. However, it is possible that a non-linear machine learning representation of the salesperson would better mimic the salesperson's pricing behavior. Accordingly, in addition to the linear model we estimate several machine learning specifications of the JB price margins function, including linear regularized regressions (Ll and L2) and RF as well as alternative specifications of the weight and RFM variables. We find that despite its relative simplicity, the linear model has better fit and prediction relative to the regularized regression models and slightly worse in-sample fit but similar out-of-sample predictions relative to the RF model. However, we find that the RF model produces worse couterfactual predicted profits relative to the linear model. See Web Appendix E for details.

Mixed effect pricing model - In order to mimic as closely as possible the salesperson in our JB model we estimated a separate model for each salesperson. However, such specification may suffer from over-fitting or low accuracy for salespeople with relatively small number of price quotes (the minimum number of price quotes per salesperson in our counterfactuals training data is 376). Accordingly, in addition to estimating an individual model for each salesperson, we estimated a mixed pricing model with random salesperson effects for the different model variables (in addition to random client intercept). We find that the mixed model's expected profits ($2,575,836) are higher than those of the individual models ($2,536,058), so partially pooling on the knowledge of all salespeople improves performance of the model over the individual models (see details of the mixed pricing model in Web Appendix E.1). We leave for future research the investigation of automation by wisdom of the crowds, pooling information across experts. However, we note that when we create a hybrid between the salesperson pricing and model pricing as we do in the next section, the advantage of the mixed effect pricing model relative to the individual model diminishes.

5.4 The Human-Machine Hybrid Approach

Allowing all quotes to be priced by the salesperson (as in the current practice in the firm and the control condition in our experiment) or all quotes to be priced by the model (as we did in the in the previous section) are two extremes on the continuum of human-machine hybrid automation. The treatment condition in the experiment provides one option for such a hybrid model as the salesperson received the model pricing as a recommendation and could employ judgment on whether to accept or reject the model's pricing. However, in light of the relatively low compliance rates observed in the experiment it is not clear whether the salesperson's judgment on when to comply with the model's pricing was optimal.

On the one hand, allowing salespeople in the experiment to make the judgment of when to use the model's price, might have led to low compliance rates, which possibly limited the possible treatment effect. On the other hand, salespeople may have decided to forgo the model prices when they had valuable information that the model was missing. For example, if the client expressed high urgency for the order over a phone conversation, the salesperson may have decided to take advantage of the client's needs and over-charge them. In these "broken-leg" (Meehl, 1954) cases, the model had no information of the profit opportunity and would have recommended a lower price, which the salesperson would have rejected. Consequently, the salesperson will outperform the model, because the model is missing crucial information that the salesperson has. One distinction between the broken leg example described in Meehl (1954) and our application, is that "broken-leg" cases (when meaningful information is available to the decision maker but not to the model) may not be rare or extreme in our application.

Ideally, the company would be able to identify and allocate to human pricing these "broken-leg" quotes as they come in, while automatically pricing the other quotes by the model. In order to automatically identify which quotes should be priced by the model and which by the salesperson we trained a machine learning Random Forest (RF: Breiman, 2001) model that predicts whether the salesperson or the model will generate higher profits for each quote based on the characteristics of the quote and the client. Specifically, the dependent variable for the RF model is the difference in expected profits between the salesperson and the model based on the demand model described in Section 5. As independent variables we include the quote and client characteristics used in the pricing model: cost per lb., quote weight (log), LME price per lb., LME volatility, lines per quote, regular salesperson for the client, average quote recency, frequency and monetary (for previous quote), ratio of items priced in non-weight units (FT), ratio of items requiring processing, categories included and client priority[18]. To fit the RF model we used a randomized search with the sixth quarter of the data (quarter 2 of 2016) used for cross-validation to estimate the model's hyper-parameters[19]. We then predict the difference in expected profits between the salesperson and the model for each of 11,621 quotes in the validation period (quarters 3 and 4 of 2016). We allocate a quote to the model if the predicted expected profit based on the model's price is higher than predicted expected profit based on the salesperson's price, and to the salesperson otherwise. Overall, the RF hybrid allocated 68% of quotes to model pricing, with the remaining 32% priced by salespeople. Thus, the RF allocation model recommends a much higher level of suggested compliance relative to the compliance levels observed in the experiment, a possible evidence to algorithm aversion among salespeople in our experiment.

Based on the validation period, we find that the total expected profits of the machine learning RF hybrid are 7.8% higher than those of the salespeople ($\Pi[p_{M\hat{L}\_hyb}] = 2,606,208$ vs. $\Pi[p] = \$2,417,149$) and 3.1% higher than those of the model ($\Pi[\hat{p}] = \$2,536,058$). The differences between the profits of the RF hybrid and the salesperson or the model profits are statistically significant based on the the 95% PCIs. Thus, we find that the human-machine hybrid scheme, in which the majority of the quotes are priced by the model and the remaining one-third of the quotes are priced by the salespeople, leads to higher profits than the two extreme cases (full automation or no automation). This raises the question of which quotes should be allocated to the model and which to salesperson, which we address next.

[18]Machine learning models such as the RF model cannot include client random effects.

[19]We estimated the RF model using Python's scikit-learn package. The estimated values for the hyper-parameters of the RF are: $bootstrap = False$; $max\_depth = 398$; $max\_features = sqrt$; $max\_leaf\_nodes = 578$; $min\_samples\_leaf = 15$; $min\_samples\_split = 15$ and $n\_estimators = 49$.

5.4.1 Understanding the Machine Learning Hybrid

The RF algorithm is a "black box" non-linear prediction tool. To get a first glance into what affects the decision to allocate quote requests to the model, we can look at the feature importance of the RF algorithm. Quote characteristics such as weight, cost, cut and number of lines per quote as well as client characteristics such as RFM and whether the salesperson is the regular salesperson all affect the quote allocation decision (see Web Appendix F.1 for a full list of feature importance). In order to shed more light on the allocation rules used in the RF model, we run a mixed linear regression for the difference between salesperson's expected profits and model's expected profits on the same variables used in the RF model (see details of the model in Web Appendix F.2).

The results of this analysis, shown in Table 9, indicate that salespeople generate higher profits than their bootstrap model when cost per lb. and quote weight have extreme values, either very small or very large. In addition, consistent with the RF feature importance, salespeople are more likely to generate higher profits than the model when there are multiple lines in the quote. These results corroborate the "broken-leg" effect, suggesting that salespeople are more successful in pricing out of the ordinary quotes with special features, and the RF captures that.

In the experiment we found that salespeople were more likely to adjust prices upwards (see Table 6). Similarly, we find that the hybrid model is much more likely to allocate quotes to the model when it prices higher than the salesperson (5,584 out of 7,090 cases, 78.7%) than when it prices lower (2,266 out of 4,531 cases, 50%), see Table 10.

To understand where the improvement in profits is coming from in the hybrid scheme it is useful to compare the observed conversion rates (at the actual salesperson prices) when the model's price is used in the hybrid vs. when the salesperson's price is used. Comparing conversion rates across rows in Table 10 we note that the model provided a lower price than the salesperson when conversion rates were significantly lower. That is, the model (ex-ante without observing conversion) is indeed recommending to lower prices when clients were less

Table 9: Interpreting the Random Forest Quote Allocation:
Mixed Linear Regression of Expected Profits difference
(Salesperson profits minus Model Profits)

Variable	**Coefficient**	**Std. Err.**
Weighted cost per lb.	-16.46***	(4.589)
Weighted cost per lb. squared	1.164**	(0.445)
Quote weight (log)	-60.08***	(7.486)
Quote weight (log) squared	5.054***	(0.693)
LME per lb.	-164.6*	(67.908)
LME per volatility	7.077	(3.888)
Lines per quote	5.882***	(1.060)
Regular salesperson	8.573	(6.748)
Cut ratio	8.477	(4.963)
Quote recency	-0.726	(0.897)
Quote frequency	-7.166	(6.210)
Quote monetary	2.434	(1.267)
FT base ratio	-4.806	(8.783)
Constant	261.9***	(57.546)
Observations	5,829†	

†Based on Quarter 6 that was used for the RF training.
Regression includes client priority, product category and salesperson fixed effects.
Regression includes client random effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

likely to accept the quote. Moreover, comparing conversion rates across columns in Table 10 we see that the hybrid allocation seems to identify those cases in which salespeople do not perform well and allocate them to the model. That is, the acceptance rates (at the human prices) are higher when the model allocates the quote to the human than when it allocates the quotes to the model (57.95% vs. 54.1%, $p = 0.0096$; 64.21% vs. 60.69%, $p = 0.0128$). This analysis points to the hybrid model's ability to appropriately allocate quotes to salesperson or the model based on expected client acceptance.

Table 10: Observed Conversion Rates
by Model Price Hybrid Structure

	Human Used in Hybrid	Model Used in Hybrid
Model Price is Lower	N=2,265 57.95%	N=2,266 54.10%
Model Price is Higher	N=1,506 64.21%	N=5,584 60.69%

The fact that the hybrid allocation model generates higher profits than either pure automation or the salespeople, supports our conjecture that in some pricing decisions the model's consistency in pricing is helpful, while in others there exists private (non-codable) information that the salesperson has but the model does not have. Although the model generated higher expected profits than the salespeople to begin with, the hybrid led to an additional significant increase in profits, by diverging some of the quotes to human pricing.

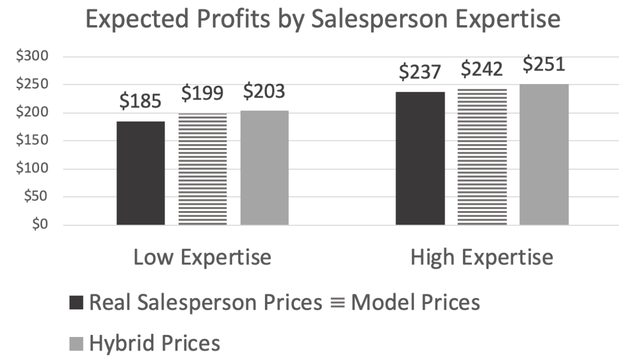
To further investigate the idea that the model performs better in most cases by applying the salesperson's pricing policy consistently, but that in some "broken-leg" cases private information allows the salesperson to generate higher profits, we create another human-machine hybrid that uses the distance between the person's and the model's price margin predictions to determine who should price the quote. If the deviation between the salesperson's pricing and the model's pricing is relatively small, we assume that the deviation is due to noisy signals that led to inter-temporal biased decision, and allocate the quote to the model. However, when the salesperson's pricing is very different from the model's pricing, we assume that the salesperson had private information that made them deviate from their standard course of pricing, and allocate the quote to the salesperson. We call this model a human judgment hybrid (see full details of how we constructed the human judgment hybrid and its results in Web Appendix F.3). We find that the human judgement hybrid's profits are significantly higher than those of the salespeople or the model (6.7% and 1.7% higher, respectively). While we do not have information about the "broken-leg" circumstances that led salespeople to largely deviate from the model in some case, the size of the deviation serves as a good proxy to the existence of private salesperson information that entails usage of human judgement, whereas in the majority of cases consistent pricing by the model was more beneficial.

Our findings provide an empirical evidence, in the context of B2B pricing, to the idea discussed in labor economics, that while automation can substitute for predictable and rule-based human labor, it can only complement human labor that is largely based on social and emotional skills (Autor et al. 2003, Autor 2015). Specifically, for salespeople making pricing decisions in a B2B context, we find that due to the mixed nature of their work, that combines rule-based decisions with judgments based on communication and interpersonal interactions, a combination of human pricing for "special" cases with automation of pricing for the majority of the cases outperforms full automation. Additionally, our proposed hybrid approach not only automates the pricing decision but also the decision of when to automate, i.e. when to use the model's prices.

5.4.2 Hybrid Structure by Salesperson Expertise

Another factor that could affect the hybrid allocation decision is the expertise of the salesperson. Upon our request, the CEO of the company rated 18 of the 21 salespeople in the data by level of expertise, dividing them into two groups: lower expertise (N=10) and higher expertise (N=8) salespeople. Figure 4 shows average expected profits per quote by expertise group based on original prices, model's prices and hybrid prices. First, consistent with the CEO's classification, the high expertise salespeople generate higher expected profits relative to the low expertise salespeople. Second, the model-of-the-salesperson improvement over the salesperson is significantly higher for the low expertise salespeople than for the high expertise salespeople ($14.85 vs. $5.21, $p < 0.001$). With the model's benefit coming from avoiding inter-temporal biases in pricing, it makes sense that its effect will be larger for less experienced salespeople whose pricing behavior may be more susceptible to contextual influences. In that sense, automation can serve as an equalizer lowering the expertise gap. Finally, the hybrid led to a significantly higher improvement in pricing for the high expertise vs. low expertise salespeople ($9.17 vs. $3.42, $p < 0.001$). This is consistent with expert salespeople being able to better leverage non-codable information (or better price unusual quotes).

Figure 4: Expected Profits by Salesperson Expertise



6 Salesperson Incentives and Automation

Designing a salesforce compensation program that fully aligns the company's incentives with agents' incentives is a complicated task (e.g., Chung et al. 2013, Kim et al. 2019). Salespeople in our settings are compensated with a base salary and a fixed percentage of their total monthly gross profit. The percentage paid to them is contingent on reaching one of three personal gross profit targets ($50K, $60K and $80K) as well as the whole branch reaching a group target. Maintaining a reasonable level of profit margin is embedded in the company's work process and is monitored on both a regular and a case-by-case basis by the management. While the company's goal is to maximize profitability levels (rather than sales), salespeople may adopt a short-sighted strategy of increasing sales by lowering margins in order to close more deals. Indeed, previous research suggests that salespeople in B2B settings often lobby internally for lower prices (Simester and Zhang, 2014).

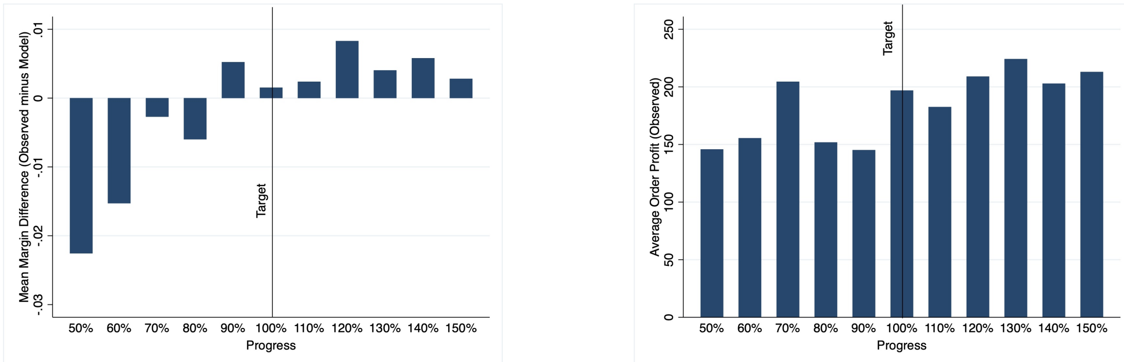
The structure of the incentives system may introduce systematic inter-temporal biases to the salesperson's pricing behavior. Hence, we did not include compensation variables in our model of the salesperson. In this section we present evidence that: (1) one reason that automation of the salesperson improves profitability is by correcting inter-temporal biases generated by the incentive program, and (2) indeed it would not be beneficial to include the salesperson's behavior with respect to the incentive program when creating a model that mimics the salesperson's pricing policy.

In order to understand how the incentive system may affect the salesperson's pricing behavior, we look at the difference between the salesperson's price margin and the model's price margin at the line level (where the model is specified in Subsection 3.2 and does not include incentive variables) with respect to the salesperson's progress towards her bonus target. Of the three targets defined in the incentives program, we set the monthly target to be the one closest to the actual total gross profit that the salesperson made that month. We calculate the progress of the salesperson towards the target as the total of gross profits accumulated since the beginning of the month up until the day of the quote divided by the target value. Because progress may exceed the target, the progress may be larger than 1. To focus on quotes that may be most affected by the incentive program, we analyze quotes priced in the last ten days of the month and when salespeople were between 50% and 150% progress with respect to their monthly target.

Figure 5a shows the average difference in line price margins between the salesperson and the model around the incentive goal mark. On average, salespeople under-price relatively to the model when being far from their target, and increase price margins upon getting closer or passing their target. The average difference in price margins between the salesperson and the model is significantly higher after passing the target (-0.004 before vs. 0.005 after, $p < 0.001$). The difference stems from an increase in salespeople's price margins (42.5% after vs. 41.7% before passing the target, $p = 0.084$) while the model's price margins are not different before and after passing the target (42.18% before vs. 42.03% after $p = 0.64$). We find that not only on average salespeople price margins are lower before passing their bonus target, but also salespeople are more likely to price lower than the model before reaching their target (56.11% before target vs. 53.56% after, $p = 0.021$) confirming that there is a change in the pricing behavior of salespeople when reaching their monthly target. In a mixed linear regression of price margins difference on distance from the goal (controlling for quote and client characteristics), we confirm the positive effect (negative to distance from goal) of progressing towards the target on the price margins difference ($\beta_{progress\_before} = -0.027, p =$

0.041). The full details of this analysis are provided in Web Appendix G.

Figure 5: Salespeople Pricing and Profits by Progress towards the Incentive Bonus Target



(a) Margin Difference by Progress

(b) Profit by Progress

In addition, we compare the difference in **observed** quote profits before and after reaching the target and find that on average profits are significantly higher after passing the target vs. before (\$206.35 vs. \$170.46, $p = 0.004$). Figure 5b shows average **observed** quote profits by progress. Overall, the difference in observed profits shows that although salespeople provide lower price margins relatively to the model before reaching their target, they do so wrongfully, as the lower prices do not lead to higher profitability. Both margins and profits increase after passing the target, suggesting that pricing behavior in earlier stages is biased downwards by the incentive system and that the model could help "push" salespeople towards more profitable prices.[20]

Finally, to confirm that the incentive variables should be excluded from the model, we estimated the model of the salesperson specified in Subsection 3.2 with the incentive variables (see Web Appendix G for details). We calculate profit counterfactuals for the prices predicted by this "non-normative" model, and indeed find that its expected profits are significantly lower than those of the original model, $\Pi[\hat{p_{incentives}}] = \$2,035,067$ vs. $\Pi[\hat{p}] = \$2,417,149$.

[20]Because of the short duration of the experiment, we could not directly study the effect of recommending the model's prices on "de-biasing" the incentive program effects in the experiment.

7 Summary and Discussion

Algorithmic pricing transformed the way sellers set prices, and in some domains, mainly in business to consumers (B2C) context, almost fully replaced human pricing. Yet, in some cases algorithmic pricing can lead to extreme failures (e.g., when the price of a book in Amazon peaked to $24 million[21], or when Delta Airlines was accused of price gouging during Hurricane Irma[22]).

The B2B market lags behind the B2C market in adopting automation (Asare et al., 2016). To a large extent pricing processes in B2B still rely on human labor, and soft skills, such as communication or salesmanship, are believed to be essential to B2B sales. In this paper we examine whether in high human-relationship environments such as B2B pricing, in which salespeople provide individual price quotes to customers, models can assist to, or even replace, human pricing. Using a multi-method approach that combines a field experiment, in which we embed AI-based algorithmic pricing into the CRM system of a B2B retailer, and econometric modeling for counterfactual analysis, we demonstrate that pricing decisions in B2B settings can be automated by modeling the salesperson and re-applying her pricing policy automatically to new pricing decisions. Providing salespeople with automated price recommendation in a real-time led to an 11% increase in profits to the company. Moreover, in a counterfactual analysis we show that because B2B pricing decisions involve a high degree of soft skills and salesmanship expertise, a hybrid model that prices incoming quotes most of the time, but allows the salesperson to price complex or irregular quotes, performs better than either full automated pricing or pure human pricing. The hybrid approach uses the model's scalability and consistency for most pricing decisions, and human judgment for unique cases that possibly involve non-codeable information. Such an approach allows to mitigate extreme algorithmic pricing failures as the one described above.

We propose a machine learning approach to automating the allocation of incoming quotes

[21] https://www.wired.com/2011/04/amazon-flies-24-million/

[22] https://www.nytimes.com/2017/09/17/travel/price-gouging-hurricane-irma-airlines.html

to the salesperson or to the model. The machine learning algorithm automatically predicts who, the salesperson or the model, will generate higher profits, and allocates each quote accordingly. The human-machine hybrid performs significantly better than pure model pricing in generating profits to the company, with an increase of over 7.8% in profits over pure human pricing. By using machine learning to automatically identify who should price the quote we lay the grounds to a hybrid automation solution that utilizes the benefits of automation in overcoming inter-temporal human biases, but preserves human expertise and experience gained by salespeople in the company over time. Our empirical analysis shows that for the B2B salesperson making pricing decisions, the balance between substitution and complementarity is key to automation. We argue that automation should be used not only to make the pricing judgment in some cases, but also to determine who should be making the decision, the machine or the salesperson.

Our research bridges between the behavioral judgment literature and marketing science literature by building a pricing judgmental bootstrapping model (Dawes 1971), and demonstrating, using both a field experiment and econometric modeling, how such a model could be applied in real-world settings to address a major business problem. The performance of judgmental bootstrapping has been rarely tested in repeated business decision making, and in settings where the expert has access to richer information than the model-of-the-expert, information that can arguably lead to superior decision making on the expert's end. Moreover, our research bridges theory and practice, by demonstrating via a pricing field experiment how automation can improve the profitability of a B2B retailer. Indeed, following our experiment, the B2B retailer we collaborated with is adding our pricing model to their CRM system to provide price recommendations to salespeople for all incoming quotes. In the longer term, and based on our work, the firm is considering to use our hybrid model to move to an online sales process, which automates both the prices presented to clients online and the decision of whether to present an online price or a "call an agent" message. We call for future research to further explore these two degrees of automation.

In our empirical application we find that using a linear judgmental bootstrapping to "teach" the model how to price works better than more advanced machine learning models of the salespeople. An advantage of the linear model is its simplicity, which is particularly important given that the company will need to occasionally re-run the model to update model parameters. Nevertheless, we encourage future research to explore the performance of machine learning relatively to linear models in automating human decision making in other contexts. Additionally, we encourage future research to explore automation using profit maximizing prices as opposed to a judgmental bootstrapping approach that mimics the expert. Such automation would need to make assumptions about demand and is likely to be more complicated for the firm to routinely estimate and optimize.

Using a hybrid automation approach that complements the salesperson with a model of herself can have far-reaching implication for preserving organizational knowledge in a work environment characterized by high salesforce turnover rates[23]. Salespeople develop expertise and familiarity not only with the product they sell, but also with their regular clients. By learning the salesperson's pricing policy and applying it automatically, the tool serves not only as a pricing aid, but also as a knowledge management mechanism, a means to preserve organizational knowledge and specific expertise within the organization, and to mitigate loses in case of salesforce turnover (Shi et al., 2017). Conversations with salespeople in the company echo the benefits of the approach. For example, one salesperson commented during the course of the experiment: "when I am not in the office, other salespeople can use my tool's recommendations to price my quotes. Currently they are not willing to take my quotes because it takes them too long to price them, so I am losing business when I am not here". Future research could further explore the use of automation to preserve organizational knowledge and mitigate the negative consequences of personnel turnover and absences.

Our analysis explored the potential of automation in B2B salesforce pricing decisions

[23]https://radford.aon.com/insights/articles/2016/Turnover-Rates-for-Sales-Employees-Reach-a-Five-Year-High

using a field experiment and secondary data from a metal B2B retailer. Future research could explore the generalizability of these findings to other B2B retail domains, and to other managerial decision making. Potential applications include other retail environments such as building supplies (Bruno et al., 2012), or special expertise in B2B services such as consulting, legal services or architectural services. The degree to which the hybrid model would fit such environments and the share of transactions that should be allocated to automation would depend on how structured the transactions are and how common "broken leg" cases are in each context. Our automation approach can flexibly accommodate different levels of automation that are appropriate for each domain.

One limitation of our field experiment was the relatively low compliance of the salespeople with the tool, which possibly underestimates the potential effect of automation. People, and especially experts, are often averse to using algorithms to aid them in decision making (Arkes et al. 1986; Camerer and Johnson 1991). Compliance may limit the effectiveness of any tool that relies on experts' willingness to use it. Specifically, if a hybrid approach is adopted and usage is in the discretion of the expert, the approach's effectiveness will depend on compliance patterns. We postulate that a bootstrap-type model is likely to facilitate higher compliance rates relative to a normative model because it mimics the salesperson's behavior as opposed to some "optimal" algorithmic behavior. Future research could further explore the role of compliance in automation in general and in hybrid automation in particular.

In summary, our research provides first empirical evidence to the potential of automating the human intensive work of B2B salesforce. It suggests that although the B2B salesperson is traditionally perceived as indispensable, some salespeople tasks could be automated. By automating parts of the pricing task the company could not only reduce costs associated with maintaining its sales team, but also increase profitability due to better-quality pricing decisions. Moreover, we show that the decision of when to use human expert pricing to override the model could, in and of itself, be automated. We hope this research will spark further investigation of this promising direction.

References

Amemiya, T. (1978). The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society*, pages 1193–1205.

Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1):93–110.

Armstrong, J. S. (2001). Judgmental bootstrapping: Inferring experts' rules for forecasting. In *Principles of Forecasting*, pages 171–192. Springer.

Asare, A. K., Brashear-Alejandro, T. G., and Kang, J. (2016). B2B technology adoption in customer driven supply chains. *Journal of Business & Industrial Marketing*, 31(1):1–12.

Ashton, A. H., Ashton, R. H., and Davis, M. N. (1994). White-collar robotics: Levering managerial decision making. *California Management Review*, 37(1):83–109.

Autor, D. H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives*, 29(3):3–30.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Batchelor, R. and Kwan, T. Y. (2007). Judgemental bootstrapping of technical traders in the bond market. *International Journal of Forecasting*, 23(3):427–445.

Blattberg, R. C. and Hoch, S. J. (1990). Database models and managerial intuition: 50% model+ 50% manager. *Management Science*, 36(8):887–899.

Bonney, L., Plouffe, C. R., and Brady, M. (2016). Investigations of sales representatives' valuation of options. *Journal of the Academy of Marketing Science*, 44(2):135–150.

Bowman, E. H. (1963). Consistency and optimality in managerial decision making. *Management Science*, 9(2):310–321.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bruno, H. A., Che, H., and Dutta, S. (2012). Role of reference price on price and quantity: insights from business-to-business markets. *Journal of Marketing Research*, 49(5):640–654.

Brynjolfsson, E. and McAfee, A. (2012). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy.* Brynjolfsson and McAfee.

Camerer, C. F. and Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly. *In K. A. Ericsson & J. Smith (Eds.), Toward a general theory of expertise: Prospects and limits*, 195–217.

Chui, M., Manyika, J., and Miremadi, M. (2016). Where machines could replace humans—and where they can't (yet). *McKinsey Quarterly*, 7.

Chung, D. J., Steenburgh, T., and Sudhir, K. (2013). Do bonuses enhance sales productivity? a dynamic structural analysis of bonus-based compensation plans. *Marketing Science*, 33(2):165–187.

Coval, J. D. and Shumway, T. (2005). Do behavioral biases affect prices? *The Journal of Finance*, 60(1):1–34.

Cowgill, B. (2017). Automating judgment and decision making: Theory and evidence from resume screening. *Columbia Business School working paper*.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

Datta, S. and Satten, G. A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association*, 100(471):908–915.

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

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 132(4):1593–1640.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Ebert, R. J. and Kruse, T. E. (1978). Bootstrapping the security analyst. *Journal of Applied Psychology*, 63(1):110.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment*, 50(3):387–395.

Eliashberg, J., Jonker, J.-J., Sawhney, M. S., and Wierenga, B. (2000). Moviemod: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science*, 19(3):226–243.

Elmaghraby, W., Jank, W., Zhang, S., and Karaesmen, I. Z. (2015). Sales force behavior, pricing information, and pricing decisions. *Manufacturing & Service Operations Management*, 17(4):495–510.

Forrester (2015). Threats To Their Traditional Sales Force Will Change The Focus For B2B Marketers. Death Of A (B2B) Salesman. Technical report, Forrester.

Forrester (2018). Mapping The $9 Trillion US B2B Online Commerce Market. Technical report, Forrester.

Frey, C. B. and Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin*, 73(6):422.

Grewal, R., Lilien, G. L., Bharadwaj, S., Jindal, P., Kayande, U., Lusch, R. F., Mantrala, M., Palmatier, R. W., Rindfleisch, A., Scheer, L. K., et al. (2015). Business-to-business buying: Challenges and opportunities. *Customer needs and Solutions*, 2(3):193–208.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica: Journal of the Econometric Society*, pages 1219–1240.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1(3):333–342.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, 57(2):116.

Jiang, Y., He, X., Lee, M.-L. T., Rosner, B., and Yan, J. (2017). Wilcoxon rank-based tests for clustered data with r package clusrank. *arXiv preprint arXiv:1706.03409*.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan.

Khan, R., Lewis, M., and Singh, V. (2009). Dynamic customer management and the value of one-to-one marketing. *Marketing Science*, 28(6):1063–1079.

Kim, M., Sudhir, K., Uetake, K., and Canales, R. (2019). When salespeople manage customer relationships: Multidimensional incentives and private information. *Journal of Marketing Research*, 56(5):749–766.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kunreuther, H. (1969). Extensions of bowman's theory on managerial decision-making. *Management Science*, 15(8):B–415.

Lam, S. Y., Shankar, V., Erramilli, M. K., and Murthy, B. (2004). Customer value, satisfaction, loyalty, and switching costs: an illustration from a business-to-business service context. *Journal of the Academy of Marketing Science*, 32(3):293–311.

Larkin, I. and Leider, S. (2012). Incentive schemes, sorting, and behavioral biases of employees: Experimental evidence. *American Economic Journal: Microeconomics*, 4(2):184–214.

Lilien, G. L. (2016). The b2b knowledge gap. *International Journal of Research in Marketing*, 33(3):543–556.

Lilien, G. L., Rangaswamy, A., Van Bruggen, G. H., and Starke, K. (2004). Dss effectiveness in marketing resource allocation decisions: Reality vs. perception. *Information Systems Research*, 15(3):216–235.

Mathews, B. P. and Diamantopoulos, A. (1986). Managerial intervention in forecasting. an empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, 3(1):3–10.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press.

Misra, S. and Nair, H. S. (2011). A structural model of sales-force compensation dynamics: Estimation and field implementation. *Quantitative Marketing and Economics*, 9(3):211–257.

Morgan, R. M. and Hunt, S. D. (1994). The commitment-trust theory of relationship marketing. *The Journal of Marketing*, 58(3):20–38.

Nedelkoska, L. and Quintini, G. (2018). Automation, skills use and training. Technical Report 202, OECD Report.

Nikolopoulos, K., Lawrence, M., Goodwin, P., and Fildes, R. (2005). On the accuracy of judgmental interventions on forecasting support systems. Technical report, Lancaster University Management School.

Payne, J. W., Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge university press.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1):3–13.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Sharda, R., Barr, S. H., and MCDonnell, J. C. (1988). Decision support system effectiveness: a review and an empirical test. *Management Science*, 34(2):139–159.

Shi, H., Sridhar, S., Grewal, R., and Lilien, G. (2017). Sales representative departures and customer reassignment strategies in business-to-business markets. *Journal of Marketing*, 81(2):25–44.

Simester, D. and Zhang, J. (2014). Why do salespeople spend so much time lobbying for low prices? *Marketing Science*, 33(6):796–808.

Wiggins, N. and Kolen, E. S. (1971). Man versus model of man revisited: The forecasting of graduate school success. *American Psychological Association*, 19(1):100–106.

Zhang, J. Z., Netzer, O., and Ansari, A. (2014). Dynamic targeted pricing in B2B relationships. *Marketing Science*, 33(3):317–337.

Appendices

A Pricing Model

Figure A1: Screenshot of the CRM System

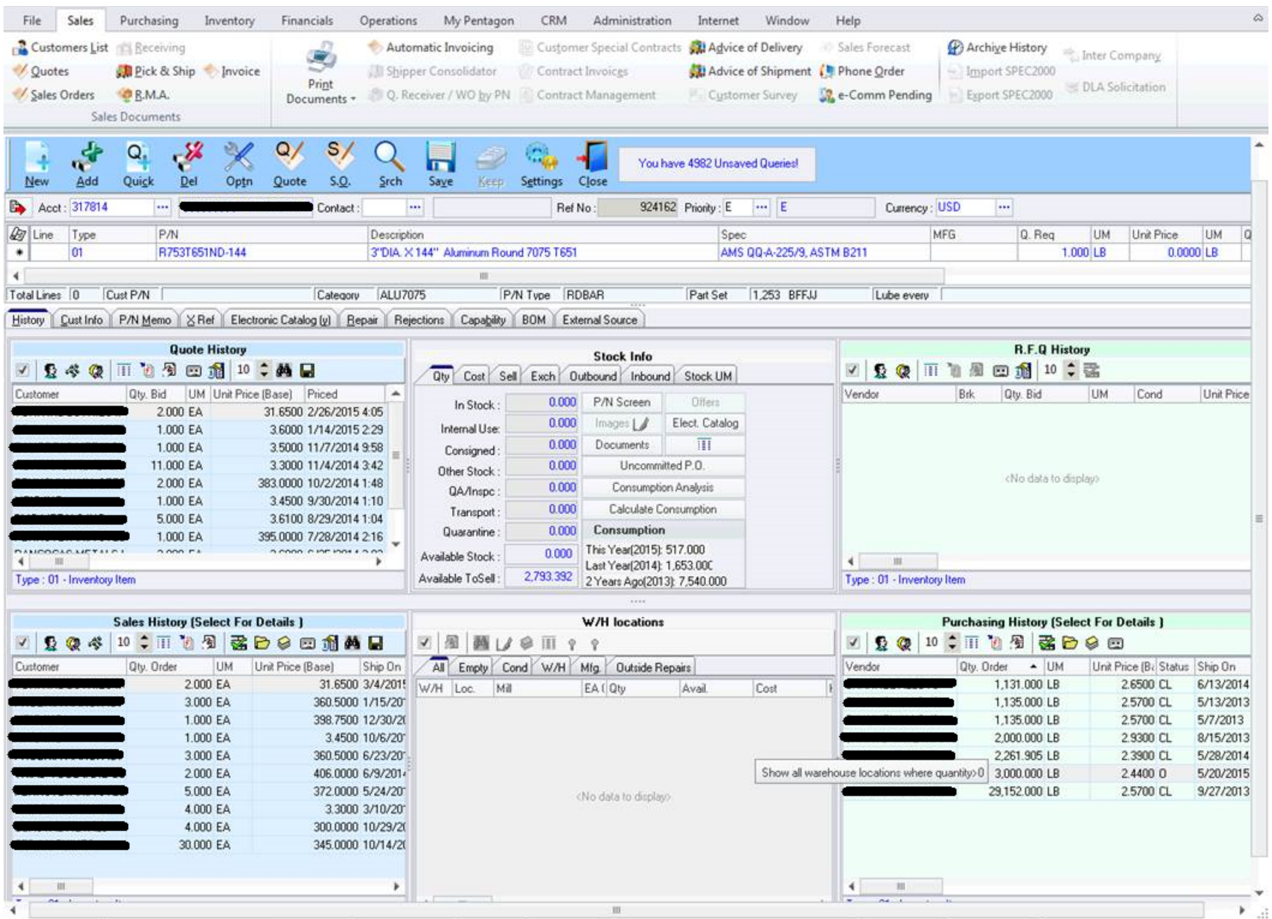


Table A1: Summary of Product Categories in the Data

	N	Frequency	Cum. freq.
Aluminum - Cold Finish	5,293	3.78	3.78
Aluminum - Plates, Aerospace	8,448	6.04	9.82
Aluminum - Plates, Commercial	32,355	23.13	32.96
Aluminum - Round, Flat, Square Solids	35,634	25.48	58.43
Aluminum - Shapes and Hollows	37,340	26.70	85.13
Aluminum - Sheets, Aerospace	614	0.44	85.57
Aluminum - Sheets, Commercial	17,526	12.53	98.10
Other Metals	2,480	1.77	99.87
Stainless - Other Stainless	179	0.13	100.00
Total	139,869	100.00	

Table A2: Average Estimates of 17 Individual Pricing Models

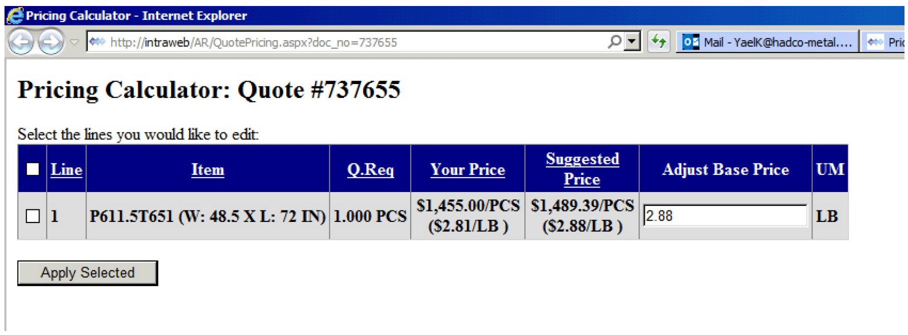
	Mean	Std. dev.	Lower 10 salesperson%	Median salesperson	Upper 90 salesperson%
Client intercept	0.87	0.82	0.01	0.87	2.18
Cost per lb.	-0.05	0.03	-0.10	-0.05	-0.01
Market price per lb.	0.64	0.91	-0.36	0.88	1.40
Market price volatility	-2.08	5.91	-7.37	-2.27	5.96
Weight (log)	-0.47	0.07	-0.57	-0.45	-0.41
Relative weight	0.28	0.11	0.12	0.27	0.41
Cut / weight	0.85	0.67	0.16	0.79	1.72
FT base	-0.13	0.16	-0.40	-0.10	0.05
Recency	0.00	0.00	-0.00	0.00	0.00
Frequency	-0.07	0.04	-0.12	-0.06	-0.02
Monetary	0.00	0.01	-0.01	0.00	0.02
Regular salesperson	0.02	0.13	-0.14	0.02	0.21
2016q2	0.09	0.09	0.03	0.08	0.20
2016q3	0.13	0.19	0.03	0.08	0.19
2016q4	0.18	0.22	0.01	0.12	0.30
2017q1	0.19	0.28	-0.04	0.15	0.34
2017q2	0.24	0.35	0.02	0.11	0.43
Priority B	-0.01	0.14	-0.17	0.02	0.15
Priority C	0.04	0.11	-0.08	0.05	0.18
Priority D	0.19	0.18	0.06	0.18	0.36
Priority E	0.25	0.15	0.06	0.24	0.45
Priority P	0.04	0.24	-0.22	-0.03	0.40
Aluminum Plates Aerospace	0.11	0.14	-0.09	0.13	0.25
Aluminum Plates Commercial	0.30	0.13	0.15	0.27	0.50
Aluminum Round Flats Squares Solids	0.24	0.13	0.07	0.25	0.42
Aluminum Shapes and Hollows	0.29	0.12	0.14	0.29	0.49
Aluminum Sheets Aerospace	0.17	0.30	-0.23	0.17	0.50
Aluminum Sheets Commercial	0.26	0.14	0.11	0.26	0.44
Other Metals	0.36	0.27	0.09	0.34	0.80
Stainless Other Stainless	0.54	0.69	0.00	0.28	1.09
Total Salespeople = 17					

B Field Experiment

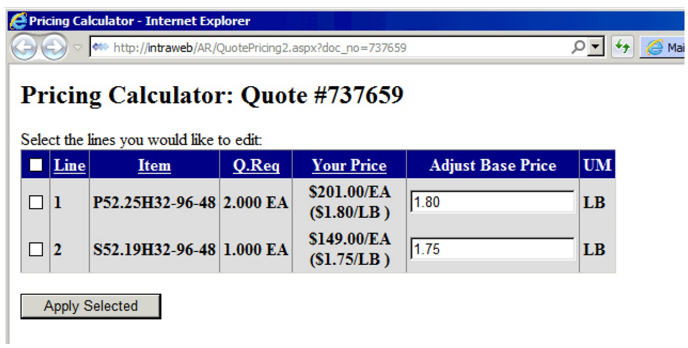
B.1 Field Experiment Forms

Figure A2: Field Experiment Edit Forms

(a) Treatment Edit Form



(b) Control Edit Form



B.2 Field Experiment Randomization Check

Table A3: Randomization Check for Quote Statistics

	Control	Treatment	Diff.	Std. Dev	P-Value
Cost per lb.	1.7323	1.7191	0.0132	0.0323	0.6834
Weight	718.1085	703.9762	14.1323	51.1599	0.7824
Cut/weight	0.3064	0.3064	0.0001	0.0202	0.9964
Total lines	2.0647	1.9531	0.1116	0.0976	0.2530
Original price per lb.	3.2691	3.2473	0.0217	0.0903	0.8099
Model price per lb.	3.4339	3.4294	0.0045	0.0871	0.9591
Price difference	0.5917	0.5927	-0.0010	0.0416	0.9806
Number of quotes	837	1,238			

B.3 Field Experiment SUTVA Analysis

In this appendix we provide details of the stable unit treatment value assumption (SUTVA) analysis of the field experiment. For each line l of each quote q priced by salesperson s for client i at time t we regress the absolute difference between the model's price per lb. and salesperson's original price per lb., Δp^t_{lqis}, on the set of line and time-varying client characteristics, $x^{\Delta p}_{lqi}$, salesperson fixed effect, salesperson-client random effect, $\alpha^{\Delta p}_{is}$ as well a on $T^{\Delta p,t-1}_s$, dummy indicating whether the previous quote priced by salesperson s was treated:

$$\Delta p^t_{lqis} \sim \alpha^{\Delta p}_{is} + \boldsymbol{\rho_s x^{\Delta p}_{lqi}} + \kappa^{\Delta p}_T \, T^{\Delta p,t-1}_s + \epsilon^{\Delta p}_{lqis}, \tag{11}$$

where $\epsilon^{\Delta p}_{lqis}$ is a normally distributed random shock. After removing the first quote for each salesperson, which was used to initialize the previous treatment dummy, the usable sample size for the regression is 4,105 pricing decisions. The results of the regression are shown in Table A4. The coefficient of interest is the coefficient kappa of the previous quote treated. We do not find a significant relationship between whether the previous quote was a treatment quote and the difference between the salesperson price per lb. and the model's price per lb. in the current pricing decision, suggesting that no significant learning due to past treatment occurred on the part of the salespeople.

Table A4: Absolute Difference between
Original- and Model- Price per lb.

Variable	**Coefficient**	**Std. err.**
Cost per lb.	0.232***	(0.030)
LME per lb.	9.771**	(3.656)
LME volatility	-33.86	(19.676)
Weight (log)	-0.319***	(0.016)
Relative weight	-0.0410	(0.051)
Cut/weight	17.23***	(1.005)
Recency	0.0002*	(0.000)
Frequency	-0.0377	(0.023)
Monetary	0.0288*	(0.014)
Regular salesperson	0.0763	(0.060)
Foot base	0.104	(0.103)
Previous quote treated	0.00185	(0.036)
Constant	-6.622*	(3.122)
Observations	4,105	
R^2	33.75%	

* $p < 0.05$, ** $p < 0.01$,*** $p < 0.001$
Controlling for salesperson, product category,
and client priority fixed effect and client random effect.

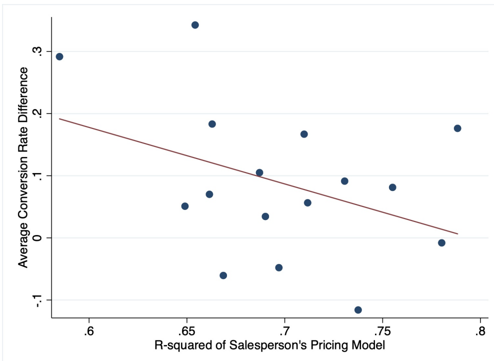
B.4 Additional Analyses

Table A5: Instrumental Variables Analysis
for Quote Conversion (with model
recommends higher price interaction)

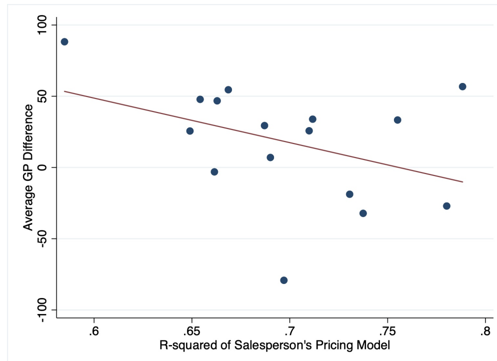
Variable	**Coefficient**	**Std. err.**
$\Delta Price$	-1.575***	(0.102)
Model higher	-0.883***	(0.054)
Model higher $\times$ $\Delta Price$	1.429***	(0.109)
Line weight (log)	-0.211***	(0.026)
Cost per lb.	-0.00381	(0.029)
Cut / weight	24.41***	(3.414)
Constant	1.926***	(0.179)
Observations	4,142	

Day fixed effects included.
*** $p < 0.001$

Figure A3: Heterogeneous Treatment Effect
(Average Difference between Treatment and Control)
by Salesperson Model R^2



(a) Conversion rates



(b) Gross Profit

Table A6: Cragg Analysis with Interaction
between Treatment and Prediction Intervals

Variable	Coefficient	Std. err.
Client acceptance of price		
Treatment	0.151*	(0.075)
Prediction interval	-0.930**	(0.353)
Treatment X interval	-0.052	(0.430)
Line weight (log)	-0.088***	(0.022)
Cost per lbs.	-0.026	(0.041)
Cut / weight	-3.337*	(1.552)
Constant	0.422*	(0.190)
Log line gross profit		
Treatment	0.003	(0.009)
Prediction interval	0.160***	(0.038)
Treatment X interval	-0.112*	(0.047)
Line weight (log)	0.115***	(0.003)
Cost per lbs.	0.037***	(0.006)
Cut / weight	1.523***	(0.220)
Constant	0.976***	(0.021)
$\log(\sigma)$		
Constant	-2.198***	(0.039)
Observations	4,142	
Pseudo R^2	29.45%	

Day fixed effects included
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A7: Line Gross Profit by Model Recommendation and Salesperson Behavior

		Line Gross Profit			
	Model's recommendation	**Decreased price**	**No change**	**Increased price**	**Total**
Treatment	**Decrease price**	\$200.63	\$155.92	\$182.33	\$163.74
	Increase price	\$90.07	\$72.59	\$55.78	\$69.67
	Total	\$168.97	\$105.70	\$70.39	\$105.11
Control	**Decrease price**	\$116.27	\$150.76	\$144.56	\$147.11
	Increase price	\$50.68	\$62.04	\$55.49	\$60.50
	Total	\$91.04	\$97.81	\$72.16	\$94.16

C Counterfactuals Data

Table A8: Summary Statistics per Quote Line in the Data used for the Counterfactuals Analysis

	Mean	Std. dev.	Lower 10%	Median	Upper 90%
Line margin§	0.36	0.19	0.17	0.32	0.65
Price per lb.	3.32	2.51	1.70	2.49	5.67
Cost per lb.	1.82	1.01	1.26	1.57	2.68
LME per lb.	0.73	0.06	0.67	0.72	0.82
LME volatility	0.74	0.34	0.34	0.67	1.20
Weight	265.00	473.36	15.14	98.57	675.95
Recency†	0.88	2.57	0.01	0.20	1.80
Frequency†	0.42	0.43	0.06	0.28	1.00
Monetary†	6.34	1.38	4.69	6.23	8.16
Regular salesperson	0.83	0.28	0.33	0.97	1.00
Cut required	0.32	0.47	0.00	0.00	1.00
Feet base	0.03	0.18	0.00	0.00	0.00
Sale (quote converted)	0.64	0.48	0.00	1.00	1.00
Total = 104,336					

§Line margin calculated as specified in Equation 1

†Calculated at the product category level

Table A9: Line Margin by Quarter in the Data used for the Counterfactuals Analysis

	Mean
2015q1	0.333
2015q2	0.338
2015q3	0.341
2015q4	0.334
2016q1	0.393
2016q2	0.409
2016q3	0.411
2016q4	0.419
Total	0.375

D Demand Model Estimation

Demand Estimation and Results

To estimate the demand model with the pricing control function, we first estimate a random effects model for the control function pricing equation and use the residuals from the control function (ΔP_{qi} in Equation 8) to estimate the demand controlling for possible price endogeneity. We then use Bayesian inference with HMC sampling to estimate the demand quote acceptance model. Convergence of the sampler was assessed using a Rubin Gelman convergence diagnostic (Gelman et al., 1992). We estimate the demand model on the first 18 month of the data, on the same sample used to estimate the model of the salesperson, and leave the remaining 6 months of quotes for validation. Parameter estimates for the control function and acceptance decision are mostly significant and in the expected direction (see Tables A11 and A12, respectively). As expected, higher cost and cut requirements increase the price. With respect to clients' quote acceptance, higher price reduces the likelihood of acceptance. Larger quotes are less likely to be converted. If the client hasn't been ordering for a while (large recency), the client is less likely to accept the quote. When working with the regular salesperson, the client is more likely to accept the quote. Overall, the demand model predicts acceptance probability in the hold-out sample to be 60.8% compared to observed conversion rate of 59.3% .

Table A10: Bootstrap Pricing Model for Counterfactuals Analysis

Variable	**Coefficient**	**Std. err.**
Cost per lb.	-0.136***	(0.003)
Market price per lb. (LME)	0.562***	(0.081)
Volatility	-0.012**	(0.006)
Weight (log)	-0.385***	(0.002)
Relative Weight	0.434***	(0.006)
Cut/weight	2.423***	(0.046)
Foot base	0.018	(0.012)
Recency	0.001*	(0.001)
Frequency	-0.052***	(0.007)
Monetary (log)	-0.0004	(0.002)
Regular salesperson	-0.070***	(0.011)
Priority B	0.037	(0.064)
Priority C	0.038	(0.059)
Priority D	0.142**	(0.062)
Priority E	0.216***	(0.058)
Priority P	0.0001	(0.068)
Aluminum Plates Aerospace	0.022	(0.015)
Aluminum Plates Commercial	0.078***	(0.013)
Aluminum Round Flat Square Solids	-0.079***	(0.012)
Aluminum Shapes and Hollows	0.074***	(0.013)
Aluminum Sheets Aerospace	0.288***	(0.041)
Aluminum Sheets Commercial	0.002	(0.014)
Other Metals	0.283***	(0.022)
Stainless - Other Stainless	0.117*	(0.066)
2015 q2	0.013*	(0.007)
2015 q3	0.063***	(0.010)
2015 q4	0.064***	(0.013)
2016 q1	0.422***	(0.013)
2016 q2	0.491***	(0.011)
Intercept	0.843***	(0.111)
Observations	104,336	
R^2	62.11%	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: regression includes client random-effect and salesperson fixed effect

Baseline priority - Priority A.

Baseline category - Aluminum Cold Finish.

Baseline quarter - 2015 q1.

Table A11: Control Function Regression Results

Variable	Coefficient	Std. err.
Client intercept	0.997***	0.03
Cost per lb.	1.379***	0.009
Cut ratio	0.452***	0.024
2015 Q1	-0.455***	0.032
2015 Q2	-0.463***	0.028
2015 Q3	-0.423***	0.028
2015 Q4	-0.497***	0.028
2016 Q1	-0.042***	0.026
2016 Q2	0	(.)
REML criterion	131,823	

Client random effect included
*** $p < 0.001$

Table A12: Parameter Estimates for Client's Acceptance Decision

Parameter	Mean	Mean SE	Std. dev.	$Q_{2.5}$	$Q_{97.5}$
Intercept	**1.359**	0.007	0.219	0.944	1.777
Price	**-0.084**	0.001	0.015	-0.115	-0.055
Recency	**-0.001**	0.000	0.000	-0.001	-0.001
Weight (log)	**-0.317**	0.000	0.013	-0.342	-0.291
LME	0.461	0.008	0.264	-0.062	0.999
LME volatility	0.022	0.001	0.039	-0.055	0.097
Regular salesperson	**0.576**	0.002	0.058	0.463	0.685
Aluminum - Cold Finish	0.135	0.004	0.094	-0.051	0.316
Aluminum - Plates, Aerospace	0.184	0.004	0.092	-0.011	0.360
Aluminum - Plates, Commercial	**0.371**	0.003	0.063	0.256	0.490
Aluminum - Round, Flat, Square Solids	**0.278**	0.003	0.057	0.162	0.385
Aluminum - Shapes and Hollows	**0.593**	0.003	0.060	0.476	0.715
Aluminum - Sheets, Aerospace	**-0.658**	0.007	0.286	-1.204	-0.102
Aluminum - Sheets, Commercial	**0.389**	0.003	0.069	0.257	0.521
Other Metals	**1.132**	0.006	0.143	0.855	1.405
Stainless - Other Stainless	0.770	0.011	0.457	-0.103	1.696
γ	**-0.053**	0.001	0.019	-0.086	-0.014
σ	**0.009**	0.000	0.007	0.000	0.026

Posterior means and standard deviations are calculated across the HMC draws.
Estimates in bold indicate a significant effect.

E Alternative Pricing Model Specifications

The approach we took to automate the salesperson in the model used in the experiment was to bootstrap the salesperson's past pricing decisions and reapply the learned pricing policy systematically to new pricing decisions. We chose a simple linear model, as opposed to more flexible non-linear models, to automate the salesperson for two reasons. First, keeping in mind that the model would be used by the company to recommend prices to its salespeople in real time, and the company's intention to implement the price recommendation permanently in their system, which will require their IT team to occasionally re-run the model and code it into their CRM system, we chose a parsimonious, interpretable, and easy to implement linear specification for the model. Second, previous research has shown the robustness of simple linear model of human decision making (Dawes, 1979; Dawes et al., 1989).

However, it is possible that other, non-linear or machine learning (ML) specifications, will capture the salesperson's pricing process better, hence create a better model of the salesperson. Indeed, ML has been recently used to automate decision making in several domains, such as human resource screening (Cowgill, 2017) or judicial decisions (Kleinberg et al., 2018).

Accordingly, in this section we compare the random effect linear model described in section 3 to three alternative ML models: two linear regularization models - the Lasso and Ridge regression models, and one non-linear model - Random Forest (RF: Breiman, 2001) model. Similar to the linear regression model, we estimate an individual pricing model separately for each salesperson using the counterfactuals data. For each one of the models we use the logit transformed price margins as the dependent variable and the same set of variable described in Section 3.2 as predictors. One exception is that because ML methods cannot accommodate random effects, we included instead as an additional predictor the average log price margin per client, as a proxy for client individual effect.

For the implementation of all three ML models we used Python's scikit-learn software (Pedregosa et. al., 2011). To fit each model, we used cross validation on the calibration data to fit hyper-parameters of the model. Specifically, for the Lasso and Ridge we used cross validation to estimate the tuning parameter alpha. For the RF, we used a randomized search cross-validation to estimate the hyper-parameters related to number of trees, max tree depth, number of leafs, maximum feature allowed in a tree. We allow the range of the randomized search to vary based on the number of pricing decisions made by each salesperson (the sample size for each salesperson's model). Table A13 shows the parameters for which a randomized search was conducted and the set of parameters that yielded the best score for each salesperson.

We calibrate the three ML models on the same data described in 5.1, covering 18 months and use the last six months of 2016 for prediction. To compare the four models - linear, Lasso, Ridge and the RF models - we calculated for each model the root mean-squared-error (RMSE) between the predicted and observed logit transformation of price margins of each line as a risk metric corresponding to the expected value of the squared error.

Table A14 shows the RMSE scores for each model for the 21 salespeople in our data, as well as simple and weighted (by number of quotes per salesperson) average RMSE scores per model. For every model we report the in-sample and out-of-sample RSME scores. First, we see that the two ML linear models (Lasso and Ridge), perform worse than the simple linear model, possibly due to the loss of the client random effects, which has a significant share in explaining variance in pricing decisions. The RF model, on the other hand, outperforms the other models both in- and out-of-sample.

We also calculated, using the counterfactual analysis, the predicted profitability of the ML models relative to the simple linear model and find that the linear model leads to the highest profitability among all four models. Specifically, the RF model's prices generated expected profits about 14% lower than those of the linear model ($\Pi[RF] = \$2,204,991$ compared to $\Pi[\hat{p}] = \$2,566,329$). One possible reason for the difference in profits is the lower predicted price per lb., on average, of the RF relative to the linear model ($Pr[RF] = \$3.08$ compared to $Pr[\hat{p}] = \$3.28$).

Thus, overall, we find that in our application the simple random effect linear model is performing better than the alternative ML models in generating profits to the company. Nevertheless, we encourage future research to explore the ML approach for automation as some of the limitations of the ML models may be specific to our application.

Table A13: Random Forest Hyper-parameters for each Individual Salesperson Pricing Model

	Salesperson code	N Train	bootstrap	max_depth	max_features	max_leaf_nodes	min_samples_leaf	min_samples_split	n_estimators
1	AR01	5,295	TRUE	398	auto	118	10	29	52
2	AS03	3,089	TRUE	176	auto	243	11	27	8
3	BM01	376	TRUE	29	auto	27	12	29	13
4	CH01	5,817	TRUE	88	auto	221	13	12	48
5	CH02	7,422	TRUE	381	auto	418	11	25	72
6	CP01	393	TRUE	15	auto	20	13	15	8
7	FJ01	1,309	TRUE	36	auto	99	10	40	28
8	GL05	432	TRUE	34	auto	11	18	13	14
9	JB01	8,727	TRUE	352	auto	363	10	20	55
10	JS02	6,842	TRUE	616	auto	296	10	19	81
11	KP03	8,565	TRUE	23	auto	679	13	16	84
12	LW03	2,927	TRUE	250	auto	279	13	13	53
13	MP01	11,349	TRUE	924	auto	788	10	16	47
14	MR01	1,633	TRUE	97	auto	77	13	24	28
15	NB01	6,567	TRUE	260	auto	386	11	10	53
16	NB03	8,127	TRUE	315	auto	506	10	37	81
17	RC01	5,587	TRUE	107	auto	105	14	20	35
18	RR01	5,007	TRUE	332	auto	133	14	16	57
19	RW01	5,558	TRUE	429	auto	428	14	25	20
20	SC01	3,223	TRUE	47	auto	104	11	30	41
21	VP01	6,091	TRUE	19	auto	607	11	15	71

Table A14: Comparison of Models - Fit and Prediction RMSE

	Saleperson	N Train	N Test	Lasso in	Lasso out	Ridge in	Ridge out	RF[24] in	RF out	Linear[25] in	Linear out
1	AR01	5,295	2,030	0.643	0.559	0.641	0.555	0.447	0.540	0.588	0.537
2	AS03	3,089	1,079	0.647	0.511	0.639	0.499	0.495	0.494	0.575	0.523
3	BM01	376	82	0.584	0.878	0.569	0.825	0.538	0.914	0.496	0.872
4	CH01	5,817	1,879	0.611	0.644	0.609	0.636	0.376	0.466	0.576	0.651
5	CH02	7,422	2,401	0.563	0.546	0.562	0.545	0.429	0.488	0.515	0.544
6	CP01	393	214	1.287	1.105	1.226	1.162	1.144	1.034	0.878	0.768
7	FJ01	1,309	573	0.591	0.523	0.584	0.522	0.469	0.461	0.548	0.521
8	GL05	432	6	0.612	0.685	0.598	0.684	0.618	0.916	0.529	0.686
9	JB01	8,727	2,810	0.461	0.454	0.455	0.443	0.303	0.385	0.424	0.453
10	JS02	6,842	2,336	0.511	0.449	0.509	0.445	0.396	0.438	0.454	0.453
11	KP03	8,565	3,075	0.474	0.454	0.472	0.448	0.346	0.423	0.424	0.450
12	LW03	2,927	2,398	0.590	0.514	0.585	0.515	0.451	0.475	0.537	0.531
13	MP01	11,349	3,445	0.565	0.560	0.564	0.559	0.383	0.447	0.529	0.564
14	MR01	1,633	698	0.631	0.592	0.625	0.575	0.531	0.578	0.527	0.573
15	NB01	6,567	2,143	0.560	0.610	0.559	0.611	0.428	0.532	0.508	0.631
16	NB03	8,127	2,736	0.701	0.608	0.695	0.590	0.497	0.599	0.662	0.563
17	RC01	5,587	1,953	0.553	0.512	0.547	0.499	0.372	0.447	0.516	0.490
18	RR01	5,007	1,137	0.587	0.632	0.586	0.631	0.402	0.420	0.555	0.651
19	RW01	5,558	2,158	0.608	0.571	0.607	0.566	0.412	0.457	0.551	0.581
20	SC01	3,223	1,267	0.670	0.697	0.663	0.692	0.497	0.694	0.618	0.704
21	VP01	6,091	1,292	0.553	0.572	0.548	0.565	0.389	0.490	0.513	0.579
	Average RSME			0.619	0.604	0.612	0.598	0.473	0.557	0.549	0.587
	Weighted average RSME			0.575	0.550	0.571	0.545	0.411	0.486	0.527	0.547

[24] Random Forest

[25] Linear Random Effects model as specified in Equation 2

E.1 Mixed Pricing Model

In addition to estimating the bootstrap pricing model individually for each salesperson, we estimated a mixed bootstrap model that partially pools information across salespeople. Specifically, for each line l of each quote q priced by salesperson s for client i, we estimated:

$$log\left(\frac{m_{lqis}}{1-m_{lqis}}\right) \sim \alpha_i + \boldsymbol{\rho_s x^1_{lqi}} + \boldsymbol{\kappa x^2_{lqi}} + \xi_{lqis}, \tag{12}$$

where α_i is client random intercept, $\boldsymbol{x^1_{lqi}}$ is a vector of random salesperson coefficients that includes line weight, cost per lb., LME price per lb., LME volatility, relative weight of line in quote, cut divided by weight, recency, frequency and monetary and a measure of client-salesperson relationship. $\boldsymbol{x^2_{lqi}}$ is a vector of fixed effect dummies for client priority, product category, quarter, whether the line is priced by feet and salesperson. Finally, ξ_{lqis} is a normally distributed random shock.

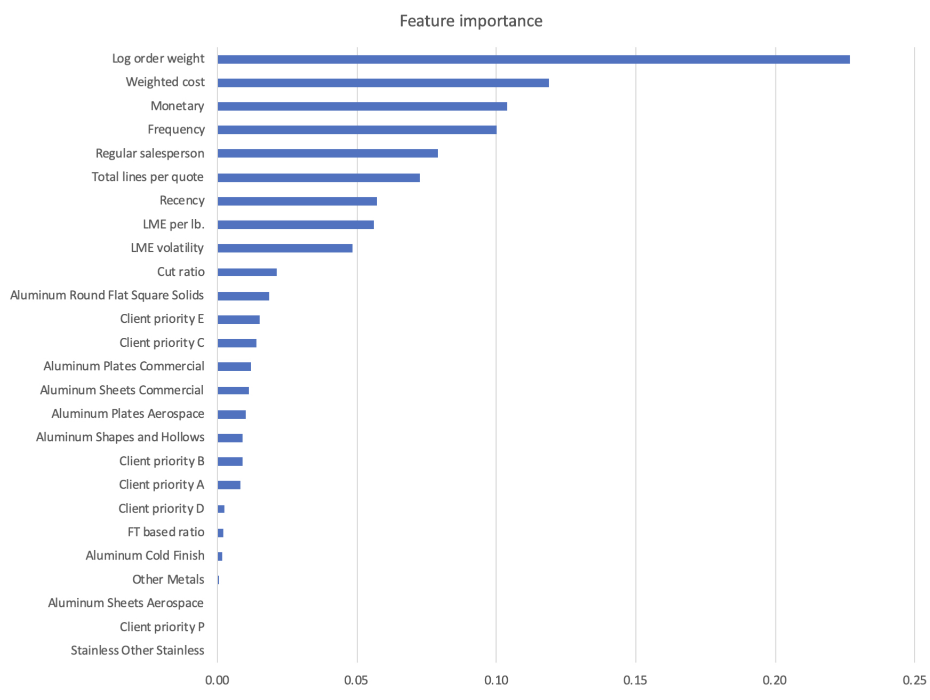
Using the estimates of the mixed bootstrap model, we calculate expected profits for each quote in the the validation set as described in Section 5. We find that for the 11,261 quotes of the validation period the mixed model's expected profits (\$2,575,836) are higher than those of the individual models (\$2,536,058), so partially pooling on the knowledge of all salespeople improves performance of the model over the individual models. The hybrid profits based on the mixed model are slightly higher than those based on the individual bootstrap models (\$2,622,831 vs. \$2,606,208).

F Additional Hybrid Analyses

F.1 RF Feature Importance

To gain some understanding with respect to which quote and client characteristics influence the RF algorithm allocations of quotes to model or salesperson pricing we look at the feature importance of the RF and find that the most important feature in determining the prediction is the weight of the products ordered in the quote. It is followed by cost per lb., dollar amount of previous quote and frequency, as well as the ratio of quotes quoted by this salesperson for the client and number of lines per quote. The full ranking of feature importance is displayed in Figure A4.

Figure A4: Feature Importance in Random Forest



F.2 Understanding the Hybrid Quote Allocation

In order to shed more light on the allocation rules used in the RF model, we run a mixed linear regression on the same variables used in the RF model and with the same DV, difference in expected profits between the model and the salesperson:

$$\Delta\Pi_{qis} \sim \alpha_i^{\Delta\Pi} + \boldsymbol{\rho_s x_{qi}^{\Delta\Pi}} + \beta_{sp}\ I_s^{\Delta\Pi} + \epsilon_{qis}^{\Delta\Pi}, \tag{13}$$

where $\Delta\Pi_{qis}$ is the difference between model and salespeople expected profits for quote q by client i price by salesperson s, $\boldsymbol{x_{qi}^{pd}}$ is a vector of quote (in addition to cost per lb. and log quote weight, we added to the linear regression quadratic cost and log weight terms, to capture possible non-linear effects in those variables, that the RF is able to capture by its non-linear nature) and client time-varying characteristics, I_s^{pd} are salesperson dummies and ϵ_{qis}^{pd} is a normally distributed random shock.

F.3 The Human Judgment Hybrid

Because the model created for each salesperson is in fact an automated representation of the salesperson herself, we expect the model to reflect the salesperson's pricing policy, and can assume that if the salesperson's pricing substantially deviates from her regular pricing (as predicted by the model), she does so in the presence of meaningful case-based information. We will therefore look at the distance between observed and predicted price margins for every pricing decision, and defer to the salesperson's price when the difference between the salesperson's price and her model's price is relatively large.

To structure the judgment-based hybrid pricing scheme, for each salesperson separately, based on her own quotes, we calculate the standard deviation of the distribution of the differences between observed log price margin and predicted log margin[26]. We structure a new pricing policy, that follows the model's margin if the salesperson's margin is within x standard deviations away from the model's margin, but follows the salesperson's margin if the distance is larger than x standard deviations. It is important to note, that the hybrid

[26]To capture deviations most accurately, we work at the level of the logit transformation of price margin, as in the model-of-the-salesperson.

policy uses the input (difference in price margin) rather than the output (profits) to create the pricing hybrid. Thus, the process does not simply create a hybrid in which the model is chosen when the model leads to higher profitability and the salesperson is chosen when the salesperson leads to higher profitability. The hybrid approach chooses the model based on deviation in the pricing policy.

We then calculate expected acceptance probability and expected profits for all the quotes in the hold-out sample, based on the new policy. We create five hybrid pricing schemes for each salesperosn, defined by the threshold of deferring to the salesperson: $x = 3$ sd, 2 sd, 1.5 sd, 1 sd or 0.5 sd. Note, that the higher the standard deviation threshold, the higher the proportion of quotes priced by the model and lower the proportion of quotes that are priced by the salesperson in the hybrid.

Each salesperson may have a different hybrid structure: for one salesperson expected profits may be highest if she prices about 60% of the quotes and model prices the remaining 40% (i.e., her optimal hybrid is the one based on sd = 0.5), while for another salesperson expected profits may be highest if the salesperson prices only 5% of the quotes and the model prices the rest (i.e., the hybrid based on 2 sd's).

For the task of deciding the hybrid threshold for each salesperson, we estimate the pricing and demand models only on the first 5 quarter of the calibration period, leaving the sixth quarter in the calibration in order to estimate hybrid threshold in a cross-validation fashion. That is, we predict prices and acceptance rates for q2 of 2016 and calculate for each salesperson the profit counterfactuals for seven different levels of hybrid thresholds (all quotes priced by the model; the salesperson prices quotes for which the difference between the model and the salesperson prices is +/- 3 sd, 2 sd, 1.5 sd, 1 sd or 0.5 sd away from the mean; and all quotes are priced by the salesperson). We then select the hybrid threshold that maximize profits in the sixth month of the calibration data, and use that threshold in the predicting profits in the validation period.

Expected profits in the validation period for the hybrid scheme integrated over all the salespeople, are 1.7% higher than those of the model and 6.7% higher than those of the salesperson, $\Pi[p_{hum\hat{a}n\_hyb}] = 2,578,852$, $\Pi[\hat{p}] = \$2,536,058$, $\Pi[p] = \$2,417,149$ (95% PCI of the difference between the hybrid profits and both the model and salesperson profits across posterior draws does not contain zero). Overall, the judgment-based hybrid generates profits that are significantly higher than those of the model alone or and salespeople themselves.

G Salespeople Incentives

To further understand how progress with respect to the bonus target affects the pricing behavior of the salesperson we estimated for every line l in quote q by client i priced by salesperson s in the validation period the following mixed linear regression model:

$$\Delta m_{lqis} \sim \alpha_i^{\Delta m} + \boldsymbol{\rho_s x_{lqi}^{\Delta m}} + \beta_{before}\ progress\_before + \beta_{after}\ progress\_after \\ + \beta_{br\_passed}\ branch\_passed + \beta_{sp}\ I_s^{\Delta m} + \epsilon_{lqis}^{\Delta m},$$

where Δm_{lqis} is the price margin difference between salesperson s and her model for line l of quote q by client i, $\alpha_{is}^{\Delta m}$ is client i random effect, $\boldsymbol{x_{lqi}^{\Delta m}}$ is a set of line characteristics (cost, weight, LME and volatility, cut, total lines per quote, RFM, FT base and category dummies) and time-varying client characteristics (client priority), $I_s^{\Delta m}$ are a set of dummy variables to control for salesperson fixed effect and ϵ_{lqis} is a normally distributed random shock. The three incentive variables included in the regression are:

$$progress\_before = \begin{cases} 0 & if\ target\ reached \\ 1 - progress & if\ target\ is\ not\ reached \end{cases}$$

$$progress\_after = \begin{cases} progress - 1 & if\ target\ reached \\ 0 & if\ target\ is\ not\ reached \end{cases}$$

$$branch\_passed\ = \begin{cases} 1 & if\ branch\ target\ reached \\ 0 & if\ branch\ target\ not\ reached \end{cases}$$

The results of the regression shown in Table A15 confirm that the further away the salesperson is from her target, she prices lower relatively to her model (note, that $progress_{before}$ is coded such that it is large when progress is low. However, after passing the target, there is no significant effect to progress.

Table A15: Line Margin Difference
(Observed minus Model)

Variable	**Coefficient**	**Std. Err.**
Progress before ind. target	-0.0265*	(0.013)
Progress after ind. target	-0.00828	(0.010)
Branch target passed	-0.0379	(0.057)
Line weight (log)	-0.0220***	(0.001)
Cost per Pound	-0.00824***	(0.001)
LME per lb.	0.0223	(0.049)
LME volatility	0.00404	(0.003)
Cut required	0.00194	(0.004)
Recency	0.000231	(0.001)
Frequency	-0.000414	(0.005)
Monetary	0.00232*	(0.001)
FT base	0.0143*	(0.007)
Constant	0.138	(0.071)
Observations	8,311	
R^2	12.18%	

* $p < 0.05$, *** $p < 0.001$

Regression includes salesperson, client priority and product category fixed effects.

Regression includes client random effects.