Automating the B2B Salesperson Pricing Decisions: Can Machines Replace Humans and When?

Yael Karlinsky-Shichor and Oded Netzer

Abstract

In a world advancing towards automation, we ask whether salespeople making pricing decisions in a high human interaction environment such as business-to-business (B2B) retail, can be automated, and when it would be most beneficial. 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 10% 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, the salesperson generates higher profits when pricing for quotes or clients with unique or complex characteristics. Accordingly, we propose a machine learning Random Forest *hybrid pricing strategy*, that automatically combines the model and 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. Whether it was machinery in production lines substituting for physical work or computer programs substituting for routine data processing, occupations either vanished or were redefined by technology. In the early days of automation, its goal was first and foremost scalability and efficiency. The tasks were well-defined with clear inputs and outputs. More recently, automation has tapped into occupations that require judgment and sense-making, as advances in computational methods in general and artificial intelligence (AI) in particular expanded the limits of automation to include non-routine tasks (Brynjolfsson and McAfee, 2012; Chui et al., 2016). With the limits for automation now lying at aspects of the job that involve perception and manipulation, creative intelligence and social intelligence (Frey and Osborne, 2017), a significant share of occupations are expected to be transformed by automation in the near future (Nedelkoska and Quintini, 2018).

Some recent applications of automation and AI methods have pushed the boundaries of automation to tasks such as screening resumes (Cowgill, 2017), scanning X-ray or CT images to identify irregularities[1], and replacing judges deciding whether defendants will await trial at home or in jail (Kleinberg et al., 2017). Yet, while those 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 information in the resume should contain all (or most) of the information needed to make the judgment.

The question we ask in this research is: Can automation be applied in domains where soft skills and interpersonal interactions have an important role in the decision-making process? Domains in which interpretation of environmental cues can provide valuable information rather than merely noisy cues? Specifically, we investigate the potential and challenges of

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

introducing 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 of expertise and soft skills of salesmanship. On one hand, B2B salespeople's pricing decisions are repetitive and arguably predictable. On the other hand, such pricing decisions involve a high degree of inter-personal communication, long-term relationships and persuasion skills. They involve understanding the state of mind of the client and interpreting behavioral cues in generating price quotes to clients. Accordingly, there is a potential for combining human and machine decisions in B2B pricing.

We use data from 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, the salesperson may identify the client's state of mind and adjust prices according to the salesperson's assessment of the client's willingness to pay. Hence, it is highly unclear whether the pricing process could be automated in this context given the great share of relationship-based communication in the pricing decisions.

Our approach to automation is to create an artificial intelligence version of the B2B salesperson that mimics her past pricing behavior and applies it systematically to new pricing decisions. We create a linear representation of each salesperson in the company (as well as alternative machine learning representations) 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 referred to as *judgmental bootstrapping* in the behavioral judgment literature (Dawes, 1979). Using judgmental bootstrapping to automate the salesperson allows us to not only reveal the salesperson's pricing policy, but also preserve the salesperson's expertise and knowledge.

In order to test the profit-performance of the bootstrap-pricing model relative to that of the salesperson, we worked with the B2B retailer to conduct a real-time pricing field experiment. 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 firm 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 entering the quote details and her own pricing, each salesperson received the price predicted by the model-of-herself and decided whether to keep her own price or adjust it based on the recommendation.

Providing salespeople with price recommendation of their own model in real time led to substantially and statistically significant higher profits than not providing such a recommendation. Specifically, mean profit per line within a quote in the treatment condition is \$9.58 higher relative to the control condition, an increase of 10% in profitability, totaling in added profits to the company of over \$24K during the eight days of the experiment, or over \$1.3 million when extrapolated yearly.

To further explore the potential of automating the B2B salesperson's pricing decisions, we perform several counterfactual analyses, which allow us to overcome some of the limitations of a field experiment (e.g., the salesperson's decision of whether to comply with the model) and simulate different scenarios of automation. 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 whether the client would accept or reject the price

quote at different price points, controlling for possible price endogeneity using a control function approach. 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 5.2% higher than the expected profitability of the salesperson's prices.

Although pure automation performs better than the salespeople in terms of profitability, evidence from the experiment as well as prior research on B2B pricing suggest that in some cases valuable information may be held by the salesperson when making pricing decisions. We propose two methods for creating a pricing hybrid that combines automation and human decision making to increase profitability. First, using our modeling approach we identify cases in which the salesperson is possibly relying on information that the model does not have in making the pricing decision. We estimate an individual hybrid for each salesperson, that combines human and model pricing, depending on the deviation of the salesperson's price from her model. This hybrid pricing scheme leads to an additional increase in profits, overall generating expected profits 6.8% higher than those of the salespeople, and significantly higher than those of pure automation as well (1.5% higher than the model's profits).

For our second pricing hybrid, we train a machine learning Random Forest (RF) model that predicts the difference in expected profitability between the salesperson and her model based on the quote's and client's characteristics (e.g., weight of the order or frequency of purchases by the client). Using a RF model that predicts the difference in expected profits between the salesperson and the model, we allocate each quote to either human or automatic pricing and find that the machine learning RF hybrid generates expected profits 7.4% higher than those of the salespeople. An advantage of the machine learning hybrid approach over the first hybrid approach is that it relies only on the quote and client characteristics and does not require the salesperson to price the quote in order to allocate the quote to a salesperson or the model.

Thus, in this work we demonstrate that salesmanship in B2B is one such occupation that could be transformed by introducing automation to improve its decision making processes.

Through a field experiment and various counterfactual analyses, we show that a *hybrid approach* that uses both automation and human judgment to make pricing decisions generates higher profits to the company than either full automation or pure human pricing. Moreover, our hybrid automation approach not only automates the pricing decision itself, but also the decision of whom should price the quote, the salesperson or the model. The company 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-judgment and machine-learning hybrid pricing schemes. Section 6 provides evidence and discusses 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. 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. B2B pricing decisions remain a relatively understudied topic in the literature. Some recent exceptions include Bruno et al. (2012) who study how reference price in B2B affects pricing and demand behavior, and Zhang et al. (2014) who study how pricing dynamics can affect client relationships in settings similar

to ours. These studies highlight the opportunity in improving B2B salespeople's pricing decisions with the help of econometric models.

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, is an essential part of 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* model, the judgment (e.g., the salesperson's 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 (Hoffman, 1960) that extracts the underlying policy executed by the expert in the decision process. Applications of judgmental bootstrapping 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)

An implicit (yet often strong) assumption underlying the superiority of models over experts in the behavioral judgment literature, is that much of the information available to

the expert is also available to the model, and hence the possible superiority of the model comes from appropriately and consistently weighing the information (Meehl, 1954). 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. The model may be consistent and unbiased, but it lacks possibly important information available to the human decision maker (e.g., information exchanged through interpersonal communication), which may outweigh the value from the increased consistency.

Therefore, the improved prediction of automated judgment over expert judgment is far from obvious when the problem involves potentially important information available only to the expert. Indeed, in our B2B pricing context, on 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, based on a phone conversation with the client, that the order is urgent and the client is willing to pay more for this order. While the model's consistency may lead to better pricing decisions in many cases, in others the model could be missing crucial information. Thus, whether a model of the B2B salesperson would outperform the salesperson in making pricing decisions, is an open empirical question.

One way to assist human decision makers in making better decisions is via decision models (Little, 1970) in the form of decision support systems (DSS). Rich literature on DSS describes the benefits of allowing managers to use an automated system in making decisions (e.g., Sharda et al., 1988; Eliashberg et al., 2000; Lilien et al., 2004). Yet, a common hurdle to the effectiveness of DSS is usage, whether due to complexity (Little, 1970), due to missing (codeable) information in the system (Van Donselaar et al., 2010), or due to behavioral biases of the decision maker (Elmaghraby et al., 2015). Our work goes beyond decision models and support systems not only in automating the salesperson's pricing behavior, but

also in determining when the salesperson should price the quote and when the model should do so with no additional input from the expert. That is, while the goal of DSS is primarily to support the human decision maker, we move from support to automation and allow the model to make decisions autonomously and automatically.

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, 2015). 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 social skills through 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 that policy 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 restrictively, 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 under 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 Appendix A). Aluminum categories vary in terms the shape of the metal (e.g., plates vs. rounds), 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 high degree of autonomy in pricing any product 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

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

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

3.2 Model Specification

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, price margin is a natural choice to build the automated pricing model. Margin is defined per Equation 1 and is calculated

[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.

Table 1: Descriptive Statistics 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
London Metal Exchange (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					

†Calculated at the product category level

at the line level. Because the firm always prices above cost, price margin could range from zero to one, and is somewhat skewed to the left in the observed prices. The average line margin in the data is 41% and the median is 36%. Consequently, we use the logarithmic transformation of price margin as the dependent variable of the margin equation.

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 to create a list of variables hypothesized to affect pricing (see a screenshot of the CRM system in Appendix A, Figure A1). The model includes the following variables:

a. **Product category**. Dummy variables for eight out of nine product categories the retailer sells. We set the category Aluminum - Cold Finish as the baseline category.

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 be different for the same product specification, depending on the relative weight of the

line in the overall order, as salespeople may employ a quantity discount at the quote level.

d. **Cut**. When the client requests for a made-to-order piece, processing is required. 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.

f. **Commodity market prices**. The salesperson has access to the actual market price as published by the London Metal Exchange (LME). We include the most recently published daily LME price per lb. as well as calculation of volatility of LME prices in the week prior to the date of the quote, as a measure of market price variability.

g. **Foot-base products**. While the vast majority of SKUs in the data are stocked and priced by weight (or have a per lb. price conversion in the CRM system), some items, mostly pipes, are stocked in feet and do not have a weight-based price. These items consist of 3.5% of the data. We include a dummy for such items.

h. **Client characteristics**.

   (a) **Priority**. The firm prioritizes each client based on its calculated 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 to become high priority clients (potential is decided based on the salesperson's judgment). We include priority in our model using a set of dummy variables. Note that priority could change over the data window because the client's priority is updated by the firm every six month. We set Priority A as the baseline priority.

   (b) **Recency, frequency and monetary - RFM**. Recency is defined as days since the client's last quote request from the same product category as the focal product priced; 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. The RFM measures are calculated at the category level to capture category-specific purchase habits.[5]

   (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.

[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 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.

i. **Client-salesperson history**. Relationship with the client could affect the salesperson's pricing behavior. On one hand, long term relationship may expose the salesperson to private information about the client. On the other hand, it may bias her pricing decisions (e.g., the salesperson's pricing may become too lenient). As a measure for the relationship of the salesperson with the client 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 that may affect pricing, we include quarter dummies. We set Q1 of 2016 as the baseline.

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 margin for the requested product specification. The margin 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 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.

In the subsequent analyses we use the margins predicted by the individual-salesperson models; however, to get a sense for the effect each variable has on log margin we hereby show and discuss results from a mixed model with client random effect and salesperson fixed effect estimated on the whole sample (see Table 2 for the aggregate regression results and Table A2 in Appendix A for average estimates across the individual-salesperson regressions).

The regression model explains nearly 70% of the variation in the pricing policy. Thus, it is apparent that our automated version of the salesperson is capturing the salespeople's pricing policy well. Indeed, when converting log margin back to margin, the average predicted line

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)
Aluminum - Plates, Aerospace	0.208***	(0.011)
Aluminum - Plates, Commercial	0.388***	(0.010)
Aluminum - Round, Flat, Square Solids	0.346***	(0.010)
Aluminum - Shapes and Hollows	0.386***	(0.010)
Aluminum - Sheets, Aerospace	0.340***	(0.026)
Aluminum - Sheets, Commercial	0.354***	(0.011)
Other Metals	0.128***	(0.018)
Stainless - Other Stainless	0.269***	(0.046)
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 price margins,

Regression includes client random-effect and salesperson fixed effect,

Priority A is the baseline category for priority,

Aluminum - Cold Finish is the baseline for product category,

Q1 of 2016 is the baseline category for the quarter dummies.

margin of 41.96% is very similar to the average observed line margin of 41.14%.

In terms of the model's estimates, we find that when cost increases, the company decreases its margins. However, when the daily metal price increases, the company seems to pass through some of the increase to the consumers (controlling for the cost of the material to the firm). High variability in market prices leads to lower price margins (a negative coefficient for LME volatility). The firm seems to employ quantity discount in margins, such that larger order have lower margins. Similarly, the larger share the line takes of the total order (fewer lines), the higher the margins of that line. As expected, processing (cut) increases margins.

In terms of client behavior, out of the three RFM measures, the company provides lower margins to customers who buy frequently, but salespeople charge higher margins for client whose previous order was large. We find that when clients receives their regular salesperson they receives lower margins, suggesting that relationship building may lead to lower margins. In terms of client priority, priority translates to better margins. When clients gain higher priority, they receive lower margins. Similarly, clients with high potential (Priority P) receive low margins.

Finally, there seems to be a positive time trend for margins. Discussions with the company's CEO confirmed that pricing guidelines changed over time to reflect higher margins across all clients, partly through instruction to request higher margins for low-priority clients (the company is striving to maintain a high quality client base and encourage low-volume clients to quit). This is also reflected by the somewhat higher margins for low priority clients.

4 Randomized Field Experiment

Now that we created an individual model for every salesperson in the company, we can assess the value of automating the salesperson pricing decisions. For that task, we collaborated with the aluminum retailer to conduct a large-scale field experiment. While we could not

completely replace salespeople in making pricing decisions, the company agreed that for a randomly selected set of quote requests, we provide to the salespeople, in real time through their CRM system, price recommendations based on each salesperson's bootstrapped model, and allow them to adjust their original prices accordingly.

4.1 Experimental Design

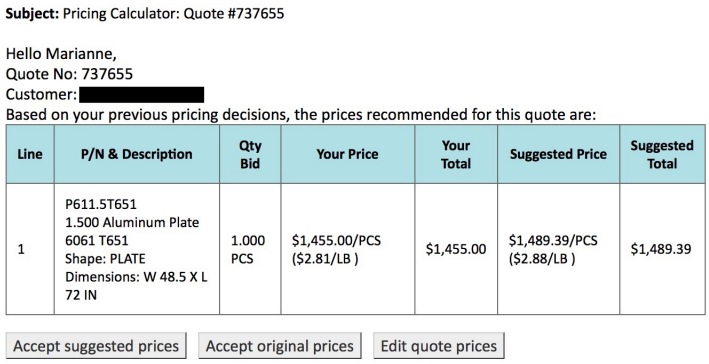
In collaboration with the B2B retailer's information technology team, we created a "price calculator", that upon entering a new quote to the system takes as input the quote, client, and salesperson characteristics. Using Equation 2, in real time, the price calculator outputs the model's margins for each incoming quote 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] We intentionally over-weighted treatment over control with anticipation of low compliance rates. The regular work flow for a quote request by the salespeople is as follows: when a client calls (or emails) for a new quote request, the salesperson enters a 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 she is ready to send the quote for the client's approval, the salesperson generates a price quote document and sends it to the client via email.

The experimental intervention in this process was upon entering the quote information and saving the new quote in the system: for quotes in the treatment group, an email was sent to the salesperson, displaying the text: *Based on your previous pricing decisions, the prices recommended for this quote are:* and below was a table displaying the part number and quantity requested for every line of the quote, as well as the price that the salesperson had just entered to the system, per pound and per unit, and total per line. The last two columns in the email displayed the model's price per pound and per unit, and total per line

[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.

Figure 1: Emails Sent to Salespeople during Field Experiment

(a) Treatment Email Format



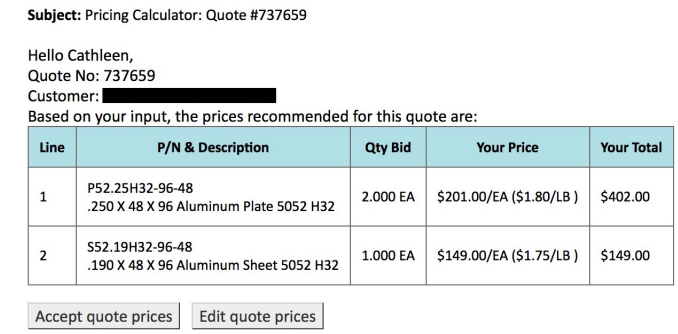
Subject: Pricing Calculator: Quote #737655

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

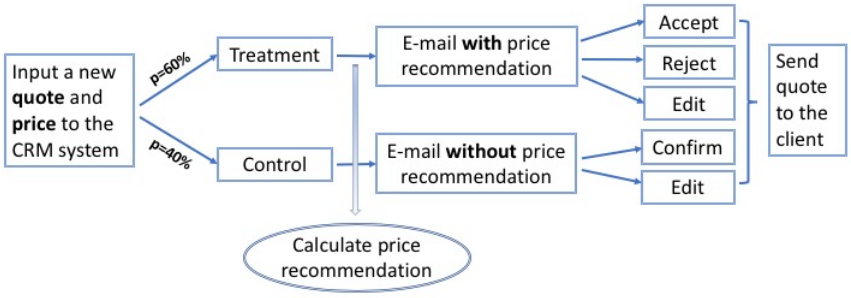
Accept quote prices | Edit quote prices

(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 Appendix B for a screenshot of the treatment Edit form). 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 condition e-mail, the control condition e-mail allowed the salesperson to either *Accept* her

original prices or *Edit*, in which case an edit form, similar to the one of the treatment condition only without recommended prices, was displayed (see Figure A2b in Appendix B for a screenshot of the control Edit form). 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 and continue with generating the price quote document and sending it to the client as she would have done without the experiment. It is important to note, that when

Figure 2: Flow of Field Experiment



entering her original price quote, the salesperson did not know whether this quote belongs to the treatment or control group (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 or 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. Typically in field experiments, the researcher knows the outcome only under the different tested policies. This design gives us access to the original pricing decision of the salesperson, before the assignment of treatment has been realized. Knowing that, enables us to better understand the pricing patterns.

We ran the experiment for eight consecutive business days. 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 locations the firm has (New York and Pennsylvania) and conducted several phone conversations with the third location (California) to make sure salespeople were comfortable with the tool and understood its flow. Our data include 2,106 quotes made during the 8 days of the experiment by 1,053 clients, with a total of 4,244 pricing decisions (some quotes had multiple lines, and each line is a pricing decision).[7] The average compliance level with the tool (i.e. quotes for which salespeople either fully accepted the recommended prices or decided to edit prices based on the recommendation using the tool), was 11%. We note that in our analysis we use intention to treat (price recommendation) as opposed to compliance (the salesperson adopting our price recommendation) because compliance is endogenous. Hence, considering the compliance levels, our results may underestimate the true effect of automation.

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, 59.73% 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. That is, that quotes assigned to treatment group are similar in characteristics to quotes assigned to control.

Table 3 shows the randomization check for different quote variables such as average cost, total weight, number of lines requiring cut and number of lines per quote. We find that randomization was performed correctly, as none of the quote characteristics are statistically significantly different between the two groups. In addition, the groups are not significantly different in the original price set by the salesperson, the model's price and the difference between them. Therefore, we can conclude that no omitted covariate made the salespeople

[7]We excluded from the analysis lines with cost or price per lb. larger than $20 that often relate to irregular orders. When including these data points the results shown in Section 4.2 are similar.

or the model price differently under the two conditions, prior to receiving the treatment.

Table 3: Randomization Check for Quote Statistics

	Control	Treatment	Diff.	Std. Dev	P-Value
Cost per lb.	1.7784	1.7579	0.0205	0.0405	0.6123
Weight	708.6789	694.6924	13.9865	50.4760	0.7817
Cut/weight	0.3072	0.3081	-0.0009	0.0200	0.9626
Total lines	2.0814	1.9706	0.1108	0.0983	0.2597
Original price per lb.	3.4243	3.4433	-0.0189	0.1123	0.8661
Model price per lb.	3.6035	3.6417	-0.0382	0.1163	0.7425
Price difference	0.6998	0.7324	-0.0326	0.0659	0.6210
Number of quotes	848	1,258			

4.1.2 Stable Unit Treatment Value Assumption

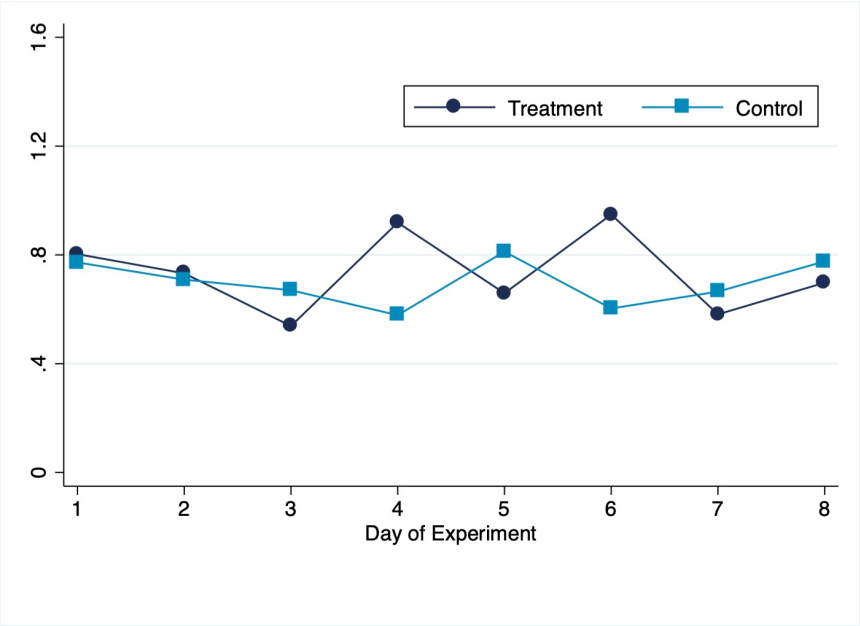
The small number of salespeople in the company was key reason to randomizing at the quote level, rather than at the salesperson level. When choosing a design where some of the salesperson's quotes are treated while others are not, there exists the risk of potential violation of the stable unit treatment value assumption (SUTVA, Rubin 1980) at the quote level. 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. We conduct both aggregate and individual-level time trend analyses to test for possible SUTVA violations.

One possible mechanism through which treatment quotes may contaminate control quotes is through 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 on the next quotes, affecting both future treatment and control quotes.

To evaluate the extent to which learning is affecting pricing, we can compare the difference between the model price and the salesperson's original price over time, for control and treatment quotes. While the model maintains the same pricing rule, if the person learns over the course of the experiment to price more systematically and more similarly to the model recommendation, the difference between the salesperson original prices and

the model's prices should decrease over time. Figure 3 shows that over the duration of the experiment, the difference between model price and the original salesperson price did not change within or between the experimental conditions, suggesting that violations of SUTVA due to learning are likely to be minimal.

Figure 3: Average Difference between Quote Model-Price and Original Price Over Time: Treatment vs. Control



To statistically test possible violations of SUTVA via the effect of one quote on a subsequent, we tested whether the treatment given to a quote affects the pricing by the same salesperson in the following quote. For each line in a quote we regress the price per pound on the set of line characteristics time-varying client characteristics, salesperson fixed effect, salesperson-client random effect, as well a dummy variable indicating *whether the previous quote priced by the salesperson was treated*. If SUTVA violations exist we would expect to find significant effect of the past quote treatment dummy on the pricing of the current quote. The results of the regression show that the treatment given to the previous quote priced by the salesperson did not significantly affect the pricing of the current quote (b=-0.0937, p-value = 0.110). See the Appendix B for full details of this analysis.

4.2 Results

4.2.1 Non-parametric Test

To test the effectiveness of the treatment (recommending to salespeople their model's pricing) we compare the gross profit (GP) between treatment and control orders. 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. Also, note that the mass at zero is not a left truncation of the GP for low GP orders, hence Tobit-type models are not appropriate. 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 order. Considering the interdependence distributional constraints of GP and the non-independence of lines within a quote, 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 profit 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 = \$9.58, $GP_{control}$ = \$93.84, $GP_{treatment}$ = \$103.42, $Z = -1.9692,\ p = 0.049$). Overall, the increase in profits is equal to nearly \$24,000 during the eight days of the experiment, or over \$1.3 million when extrapolated to increase in yearly profits. Thus, automation in the form of recommending salespeople with their own model's prices can result in significant and substantial increase in profitability for the firm.

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 margins from 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) for zero-inflated continuous data. 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.[8] Consequently, the distribution of GP can be defined using the following selection model:

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

where the top part of the equation reflects the cases where the client accepted the quote and hence the GP is positive, and the bottom part, the selection process where the quote was rejected by the client. $\boldsymbol{x_{lq}^1}$ includes a dummy for whether the quote was treated or not, a set of dummy variables to control for salesperson fixed effect, a set of dummy variables to control for day of the experiment fixed effects, line weight and whether the order required a cut (divided by the weight). $\boldsymbol{x_{lq}^2}$ includes all the covariates included in $\boldsymbol{x_{lq}^1}$ as well as the cost per lb. of the line.

The results of the analysis are shown in Table 4. Controlling for line's characteristics, and for day and salesperson fixed effects, 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. Overall, the marginal effect of providing a price recommendation to the quote is estimated at \$14.09 per line. Thus, we find that 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.

[8] As mentioned earlier, 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.

Table 4: Cragg Hurdle Regression Analysis

Variable	Coefficient	Std. err.
Client acceptance of price		
Treatment	0.167*	(0.073)
Line weight (log)	-0.0724***	(0.021)
Cut / weight	-0.758	(1.259)
Constant	0.125	(0.219)
Line gross profit		
Treatment	0.019	(0.039)
Line weight (log)	0.568***	(0.014)
Cost per lbs.	0.165***	(0.023)
Cut / weight	7.018***	(0.818)
Constant	2.133***	(0.100)
$\log(\sigma)$		
Constant	-0.625***	(0.038)
Marginal effect	**14.09***	(5.94)
Observations	4,244	
Pseudo R^2	10.88%	

Salesperson and day fixed effects included
* $p < 0.05$, *** $p < 0.01$

To further examine the increased acceptance rate by clients in the treatment condition we compare the difference between the model's recommended price and the original price set by the salesperson for quotes that were accepted and rejected by the client. As expected, we find that when the price quote was converted to a sale, the model's recommendation was higher than the salesperson's price in 63.6% of the cases. However, when the price quote was not converted into a sale, the model recommended a higher price in only 60.2% of the cases (the difference between these proportions is statistically significantly, $z = 2.2765$, $p = 0.011$).

Thus, the model's pricing corrects for over-pricing by the salespeople for quotes that were not converted to sales. B2B salespeople often lobby for lower prices (Simester and Zhang, 2014). Indeed we find that the model's prices were higher than those of the salespeople for most (62%) of the quotes. Nevertheless, there seems to be a mismatch in the cases over- or under- priced by the salespeople. While the model suggested increasing prices in some cases, the treatment effect comes from correcting over-pricing by the salespeople for certain quotes that were eventually lost.

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 or a decision support tool to salespeople, rather than replace them completely in providing price quotes to clients. Particularly, given the 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 price 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 be able to better capture the salesperson's pricing decision. 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, and then run a set of counterfactuals comparing profitability under different pricing schemes based on versions of automation, with different hybrids between the salesperson's pricing and the model pricing and 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 salesperson (rather, we used the salesperson's decision - price margin), 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.[9] Put formally, for each quote q requested by client i, based on observed prices p_{qi}

[9]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

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 can them calculate the expected profit for quote q requested by client i:

$$\Pi_{qi} = (p_{qi} - c_q) \times Pr_{qi}(p_{qi}), \tag{4}$$

$$\hat{\Pi}_{qi} = (\hat{p}_{qi} - c_q) \times Pr_{qi}(\hat{p}_{qi}), \tag{5}$$

and compare expected profits based on observed prices (Equation 4) to expected profits based on predicted prices (Equation 5).

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. 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 21 salespeople making 104,336 pricing decisions for 3,787 clients over the course of eighteen months. Table A4 in Appendix C shows summary statistics of the counterfactuals data.

As discussed previously, the company exhibited a trend of increased margins over time. Specifically, the company enjoyed higher margin since Q1 2016 (See Table A5 in the appendix). We capture such time trend in the pricing model by including quarterly dummies. To extend the time trend to the validation period we calculated the ratio between the average log margin in the validation period (q3 and q4 of 2016) and the average log margin of the last quarter in the calibration period (q2 of 2016), and used it to adjust the model prices for the validation periods. Table A6 in Appendix C shows the estimates of the pricing model (similar to Table 2 but for the counterfactual calibration data).

as two separate quotes: one accepted, and one rejected.

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 4 and 5). A purchase event is initiated by the client who has a need for aluminum supply. The client approaches the firm with a request for a price quote for one or more specifications of material, size, weight and cut. 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.

5.2.1 Demand Specification

We assume that the utility for client i from accepting quote q is:

$$u_{qi} = \beta_{1i} + \beta_{2i}\ gain_{qi} + \beta_{3i}\ loss_{qi} + \boldsymbol{\beta_z}\ \boldsymbol{z_{qi}} + \gamma\ \Delta P_{qi} + \sigma\ \eta_{qi} + \xi_{2qi}, \tag{6}$$

where

$$gain_{qi} = \begin{cases} ref\_price_{qi} - price_{qi} & \text{if } price_{qi} < ref\_price_{qi} \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

$$loss_{qi} = \begin{cases} price_{qi} - ref\_price_{qi} & \text{if } price_{qi} > ref\_price_{qi} \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

β_{1i} is a random intercept for client i, and ref_price_{qi} is the reference price for quote q made by client i, calculated as the difference between the current price and the average price the client received in the last three quote requests in the category.[10]

We model the effect of price on demand as a reference price following Zhang et al. (2014), who used data from the same retailer to model targeted pricing. Because a quote may include

[10]We compared alternative specifications of the reference price, including longer and shorter time windows to define the reference period, as well as time weighted, and order-weight weighted reference prices. All specifications lead to similar or worse model fit.

a request for more than one category, in calculating reference price we first calculate category-based reference price (i.e., the average of the price for the product category in the client's last three quote requests in the category, and then average the category-based reference prices for all categories requested in the current quote based on the relative weight of the category in the quote. If the current price is higher than the reference price, the difference will be counted as loss; if the current price is lower, the difference will be counted as gain. We calculate reference price by product category, because pricing can vary substantially across categories and to account for different purchase cycles for different product categories.

$\boldsymbol{z_{qi}}$ 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 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 category 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.

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 six out of the seven quarters in the 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 6 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 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. 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 calculate the quote acceptance and hence expected profits based on the model-of-the-salesperson predicted prices (following Equation 5) and compare it to the expected profits based on original (observed) prices offered to clients by the salesperson (following Equation 4).

To calculate the counterfactuals 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 pound per quote is \$3.41, and the average predicted price per pound based on the bootstrap model is \$3.28. The corresponding expected acceptance probability based on the original pricing

scheme is 61.1% and that based on the model's pricing scheme is 61.8%. The actual observed acceptance probability was 59.3%.

Using Equations 4 and 5 and aggregating across quotes, we find that the model's pricing scheme generates expected profits which are 5.2% higher than those of the salespeople's pricing scheme ($\Pi[p] = \$2,438,442$ compared to $\Pi[\hat{p}] = \$2,566,329$). This difference is statistically significant, as the 95% posterior confidence intervals (PCI) of the difference across a sample of 100 draws from the output of the HMC sampler do not contain zero. The actual profits for the same set of quotes were $2,345,479.

Thus, consistent with the the results of the experiment, the results of the counterfactual analysis demonstrate that the bootstrap model of the salesperson does better than the salesperson herself in generating profits for the firm. This should not be taken for granted because, as discussed previously, the B2B salesperson's work is based on her soft skills, communicating with clients, understanding their state of mind, and using those insights to leverage her 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 were such information dimmed valuable, in the counterfactual analysis the information is completely absent. Next, we examine a hybrid pricing scheme that preserves some of the private information that the salesperson has and is not captured by the model.

5.3.1 Alternative Pricing Models

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 margins function, including linear regularized regressions (Ll and L2) and RF as well as alternative specifications of the weight and RFM variables. The linear model has better fit and predic-

tion relative to the regularized regression models and slightly worse fit relative to the RF model. However, the RF model has worse predicted profits relative to the linear model. See Appendix E for details.

5.4 The Hybrid Approach

In light of the low compliance rates observed in the experiment there may be a reason to believe that within the full range of quotes, some quotes should in fact be priced by the salesperson in order to generate higher profits. On one hand, allowing salespeople in the experiment to make a judgment with regards to when to use the model's price, led to low compliance rates, which possibly limited the treatment effect. On the other hand, it is possible that salespeople may have decided to forgo the model prices when the salesperson 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 decide to take advantage of the client's need and over-charge him. In this case, the model had no information of the profit opportunity and would have recommended a lower price, which the salesperson would have rejected. The non-codeable cases are called *broken leg* cases in the behavioral judgment literature. The term broken leg, coined by Meehl (1954), describes a scenario in which a model can successfully predict whether one will go to the movies in any given night, but will fail in the rare and unexpected case in which one broke their leg that day, and the model is unaware of the incident. In those broken leg cases, the salesperson will outperform the model, because the model is missing crucial information that the salesperson has.

The question is: how can we identify those cases where the model is doing better from the cases where the salesperson is doing better? We propose two hybrid pricing schemes. The first hybrid scheme uses the deviations in the salesperson's price relatively to her model's price to identify broken leg cases in which the salesperson may have had important information that was not available to the model. The second hybrid scheme uses the features of the quote and characteristics of the client to train a machine learning RF model that predicts who will

generate higher profits: the model or the salesperson, and allocates new quotes accordingly to either human or model pricing.

5.5 Human-Judgment Hybrid

Our modeling approach can be used to identify the broken leg cases. 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 pricing and predicted pricing (as measured by margins) for every pricing decision, and rather than pricing all price quotes by the model, we 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 margin and predicted log margin[11]. 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 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

[11]To capture deviations most accurately, we work at the log margin level, as in the model-of-the-salesperson.

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). Note similarly that in the experiment, different salespeople exhibit different compliance levels and hence different hybrids.

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 selected 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.

Figure A3 in Appendix F shows the hybrid structures for each salesperson. We find that for three salespeople it is best to completely replace them with their own model; for three salespeople it is best to let them price all quotes by themselves; and for all other salespeople (15 salespeople) there is an optimal combination between every salesperson and her model that generates the highest profits. For example, for salesperson coded as SP16 the optimal hybrid is one where she prices about 15% of the quotes and the model prices about 85% of

the quotes. Across salespeople, we find that our approach recommends allocating 89.7% of the quotes to the model and 10.3% to the salesperson. Due to the relatively small number of salespeople it is not meaningful to conduct a statistical analysis of the proportion of quotes replace by the model in the hybrid for each salesperson by salesperson characteristics. However, anecdotally we find that salespeople who were rated (prior to the analysis) by the CEO as having an above average expertise were somewhat less likely to be replaced by the model (high expertise salespeople had on average 85.5% of their quotes priced by the model and low expertise salespeople had 93.6% of their quotes priced by the model[12]).

5.5.1 Profits of the Human-Judgment Hybrid

Expected profits in the validation period for the hybrid scheme integrated over all the salespeople, are 1.5% higher than those of the model and 6.8% higher than those of the salesperson, $\Pi[\hat{p_{human\_hyb}}] = 2,603,719$, $\Pi[\hat{p}] = \$2,566,329$, $\Pi[p] = \$2,438,442$ (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.

5.5.2 Understanding the Human-Judgment Hybrid

It is informative to understand which type of quotes were directed to human pricing (i.e., in which type of quotes the deviations of the salesperson from the model were large). We ran a mixed binary logit model for the probability that the hybrid uses the salesperson's judgment to price the quote[13] as a function of a set of client and quote characteristics. The probability that the salesperson's price is used in the human-judgment hybrid is:

$$Pr_{1sqi} = \frac{e^{\alpha_i^{Hyb1} + \boldsymbol{\rho^{Hyb1}} \boldsymbol{z_{qi}^{Hyb1}}}}{1 + e^{\alpha_i^{Hyb1} + \boldsymbol{\rho^{Hyb1}} \boldsymbol{z_{qi}^{Hyb1}}}}, \tag{11}$$

[12]Based on 18 salespeople evaluated by the CEO.

[13]Because the demand model was estimated at the quote level, we conducted this analysis at the quote level as well.

where α_i^{Hyb1} is client random effect and $\boldsymbol{z_{qi}^{Hyb1}}$ is a set of quote and client characteristics that includes: 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.

Table 5 shows the result of the mixed logit regression for whether the hybrid uses human pricing for the quote. We see that the salesperson's is more likely to be used when the quote has special characteristics, such as multiple lines or high processing ratio. In addition, quotes by high priority clients (the highest priority is 1 and the lowest is 6) are more likely to be priced by salespeople. With regards to weight, salespeople are more likely to price the lower weights, which is in line with the company's policy to charge minimum prices for small orders (i.e., not follow the regular pricing rules). Remember, that by construction the human-judgment hybrid uses the salesperson's price when it is relatively different than the model's price. Therefore, the analysis shown in Table 5 reflects cases in which the salesperson largely deviated from the model's price, possibly due to important information that the salesperson had but the model did not.

Hybrid Structure by Salesperson Expertise Finally, we analyzed the model's performance by salesperson's expertise. We asked the CEO of the company to classify the level of expertise of each salesperson in the company. The CEO of the company rated 18 of the 21 salespeople in the data, 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 pricing, based on the model's pricing and based on the hybrid approach. First, note that consistent with the CEO's classification, the high expertise salespeople generated higher expected profits relative to the low expertise salespeople. Second, the model-of-the-salesperson improvement over the salesperson was much higher for the lower expertise people than for the high expertise people. This may suggest that the

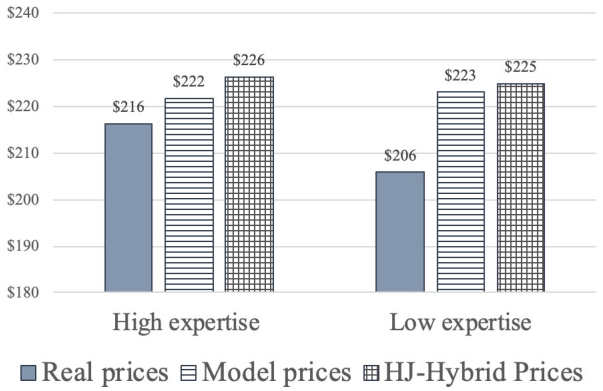
Table 5: Mixed Logit Model for Using Salesperson Price in Human-Judgment Hybrid

Variable	**Coefficient**	**Std. Err.**
Weighted cost per lb.	0.0223	(0.041)
Quote weight (log)	-0.123***	(0.033)
LME per lb.	2.110	(1.475)
LME volatility	0.0144	(0.104)
Lines per quote	0.358***	(0.031)
Regular salesperson	0.322	(0.167)
Recency†	0.00183	(0.023)
Frequency†	-0.281	(0.182)
Monetary†	-0.0113	(0.032)
FT base ratio	0.435	(0.226)
Cut ratio	0.461***	(0.128)
Client Priority	-0.373***	(0.071)
Constant	-3.871**	(1.213)
log(variance) of random intercept	1.998***	(0.079)
Observations	11,621	

†Quote average
‡Regression includes product category dummies
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

higher expertise salespeople take advantage of private information in the environment more efficiently, and when replaced completely by the model a significant share of private information is lost. Finally, the hybrid approach increased the average profit per quote twice as much for high-expertise salespeople than for low-expertise salespeople, again indicating their better skills in utilizing private information. These difference are statistically significant based on the 95% posterior confidence intervals (PCI) across a sample of 100 draws from the HMC algorithm output do not contain zero.

Figure 4: Expected Profits by Salesperson Expertise



5.6 Machine Learning Hybrid

The results of the human-judgment hybrid reported in Section 5.5 demonstrate that a pricing scheme that combines the model and the salesperson is superior to either full automation or full human pricing. However, from a practical point of view the approach is limited because it requires knowledge of the salesperson's pricing decision to realize whether such private information existed or not. In the analysis reported in Table 5, we showed that those quotes that were referred to human pricing had some unique characteristics. Specifically, these quotes tend to have multiple lines, required more processing and were made by higher priority clients. Ideally, the company would be able to identify those quotes as they come in and refer only these quotes to human pricing, while automatically pricing the other quotes by the model.

In order to automatically identify the quotes that should be priced by the model or 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 was the difference in expected profits between the salesperson and the model based on the demand model described in Section 5. As independent variables we included the same variables used in the analyses of the human-judgment hybrid: 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[14]. For the implementation of the RF model we used the Python's scikit-learn software (Pedregosa et al., 2011). To fit the RF model we used a randomized search cross-validation with same calibration period used in the human hybrid model and with the sixth quarter used for cross-validation to estimate the hyper-parameters related to number of trees,

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

max tree depth, number of leafs, maximum feature allowed in a tree[15]. Feature importance of the RF algorithm is available in Appendix G. 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 profits with the model prices is higher than profits with the salesperson prices and to the salesperson otherwise. Overall, the RF hybrid allocated 66% of quotes to model pricing, with the remaining 34% priced by salespeople.

Based on the validation period, we find that the total expected profits of the machine learning RF hybrid are 7.4% higher than those of the salespeople, $\Pi[p_{\hat{ML\_hyb}}] = 2,618,240$ vs. $\Pi[p] = \$2,438,442$ and 2% higher than those of the model, $\Pi[\hat{p}] = \$2,566,329$. The differences between the profits of the RF hybrid and the salesperson or the model profits are statistically significant based on the the 95% posterior confidence intervals (PCI) across a sample of 100 draws from the HMC algorithm output. The performance of the machine learning hybrid is very similar to that of the human hybrid.

The fact that the two hybrid pricing schemes generate profits higher 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 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. 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

[15]The estimated values for the hyper-parameters of the RF are: $bootstrap = False$; $max\_depth = 303$; $max\_features = sqrt$; $max\_leaf\_nodes = 317$; $min\_samples\_leaf = 20$; $min\_samples\_split = 33$ *and* $n\_estimators = 18$.

with judgments based on communication and interpersonal interactions, a combination of human pricing for "special" cases and automation of pricing for the majority of the cases outperforms full automation.

6 Salesperson Incentives and Automation

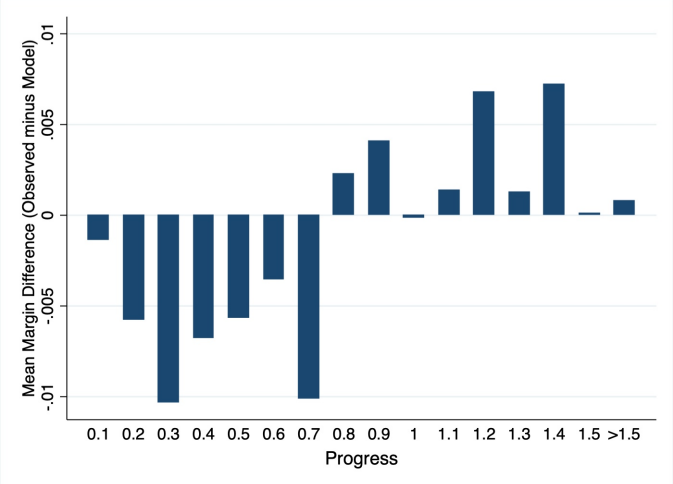
Designing a salesforce compensation program that fully aligns the company's incentives with agents' incentives is a complicated task. Different components of the compensation program may help the firm align its own incentives with those of the salespeople (see for example Chung et al. 2013, Kim et al. 2018). 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 policy 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 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 indeed it would not be beneficial to automate the salesperson's behavior with respect to the incentive program when creating a model the imitates the salesperson's pricing policy. We start by reminding the reader that while in about 62% of pricing decisions in the experiment the model's price was higher than that of the salesperson, the added profitability due to using the model's recommendations came from those cases where the salesperson over-priced and the model corrected the

over-pricing (see the analysis in Subsection 4.2.2 for reference). We find similar evidence in the counterfactual analysis, where roughly 60% of the pricing decisions the model prices are higher than the salesperson prices. That is, we observe both under-pricing and over-pricing behaviours of salespeople.

In order to understand how the incentive system may affect the salesperson's pricing behavior, we looked at the difference between the salesperson's pricing decision (line price margin) and the model's pricing decision (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 calculated 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 the target. Because progress may exceed the target the progress may be larger than 1. Figure 5 shows the average difference in line margins between the salesperson and the model for each progress percentile. It is apparent that on average salespeople under-price relatively to the model when being far from their target, and upon getting closer or passing their target they increase price margins.

Figure 5: Difference in Margin by Progress towards Bonus Target



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:

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

where md_{lqis} is the margin difference between salesperson s and her model for line l of quote q by client i, α_{is}^{md} is client i random effect, $\boldsymbol{x_{lqi}^{md}}$ is a set of line characteristics and time-varying client characteristics, I_s^{md} 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 are included in the regression are: $progress\_before$, defined as $1 - progress$ if the target had not been reached and zero otherwise; $progress\_after$, defined as $progress - 1$ if the target had been reached and zero otherwise; and $branch_passed$, a dummy for whether the group goal was met by the branch or not.

The results of the regression shown in Table 6 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. In addition, the difference between the salesperson and the model is smaller after the group target is met, i.e., salespeople increase prices relatively to the model after the team has reached the goal. Overall, the evidence suggest that salespeople under-price relatively to their model when they are personally or collectively far away from the goal and to some extent correct that pricing bias after reaching the target(s).

The model-free evidence suggests that salespeople are affected, possibly negatively, by the incentive system set by the company. Nevertheless, and in order to confirm that excluding the incentive variables from the model was the right decision, we estimated the model of the salesperson specified in Subsection 3.2 with the three incentive variables described in this section. We calculated profit counterfactuals for the prices predicted by this "non-normative"

Table 6: Line Margin Difference (Observed minus Model)

Variable	Coefficient	Std. Err.
Progress before ind. target	-0.0100***	(0.002)
Progress after ind. target	-0.00172	(0.002)
Branch target passed	-0.00974***	(0.002)
Line weight (log)	-0.0227***	(0.000)
Cost per lb.	-0.00575***	(0.001)
LME per lb.	0.0461*	(0.022)
LME volatility	-0.00298	(0.002)
Cut required	-0.00292	(0.002)
FT base	0.00497	(0.003)
Recency	-0.000252	(0.000)
Frequency	-0.00551***	(0.002)
Monetary	0.00317***	(0.000)
Constant	0.0684***	(0.018)
Observations	35,575	
R^2	11.66%	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Regression includes salesperson fixed effects, client priority and product category.

model, and indeed found that its expected profits are significantly lower than those of the original model, $\Pi[p_{\hat{incentives}}] = \$2,050,578$ vs. $\Pi[\hat{p}] = \$2,566,329$.

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. However, in some cases algorithmic pricing can lead to extreme failures (e.g., when the price of a book in Amazon peaked to $24 million[16], or when Delta Airlines was accused of price gouging during Hurricane Irma[17]).

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,

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

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

in which salespeople provide individual price quotes to customers, models can assist 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 setting 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 a 10% 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, inter-personal communication, and salesmanship expertise, a hybrid model that prices the incoming quotes most of the time, but allows the salesperson to price complex or irregular quotes performs better than either a fully automated pricing model or the salespeople's pricing. The hybrid approach permits scalability and consistency, for most pricing decisions, but uses human decision making for unique cases that require an expert's judgment. Such an approach allows to mitigate extreme algorithmic pricing failures as the one described above.

We propose two methods to identify who should price an incoming quote, the model or the salesperson. In the first method we rely on the difference between the salesperson pricing and the model's pricing to identify cases in which the salesperson had important private information. In the second method we train a machine learning algorithm on the quote's and client's characteristics. The machine learning algorithm predicts automatically who, the salesperson or the model, will generate higher profits and allocates the quotes accordingly. Both hybrid schemes perform significantly better than pure model pricing in generating profits to the company, with an increase of over 7% in profits over pure human pricing. By using machine learning to automatically identify who should price the quote we lay the ground to an automation solution that utilizes the benefits of automation 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 automatically determine who should be making the decision, the machine or the salesperson.

Our research bridges between the behavioral judgment and marketing science literatures 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 client online and the decision of whether to present an online price or a "call an agent" button based on the specific quote. We call for future research to further explore these two degrees of automation.

In our empirical application we find that using 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. In the experiment, and more generally in the application of our approach by the firm, we use the bootstrap model to recommend prices to the firm's salespeople by embedding our model into the company's CRM system. The company will also need to occasionally re-run the model to update the parameters. All of which favor a parsimonious, interpretable, and easy to implement linear specification for the model. Additionally, such linear bootstrap models have been successfully used in the past to automate human decision making (Dawes, 1979; Dawes et al., 1989).

Nevertheless, we encourage future research to explore the performance of machine learning relatively to linear models in automating human decision making in other contexts.

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[18]. 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 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 likely "broken leg" cases are in each context. Our automation approach can flexibly accommodate different levels of automation

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

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 itself, be automated. We hope this research will spark further investigation of this promising direction.

References

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.

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.

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.

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. (2015). The growing importance of social skills in the labor market. Working Paper 21473, National Bureau of Economic Research.

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

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.

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*.

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. (2018). When salespeople manage customer relationships: Multidimensional incentives and private information. *COWLES FOUNDATION DISCUSSION PAPER NO. 3022*.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human decisions and machine predictions. Technical report, National Bureau of Economic Research.

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.

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.

Little, J. D. (1970). Models and managers: The concept of a decision calculus. *Management Science*, 16(8):B–466.

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

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. *OECD Report*, (202).

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.

Van Donselaar, K. H., Gaur, V., Van Woensel, T., Broekmeulen, R. A., and Fransoo, J. C. (2010). Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56(5):766–784.

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

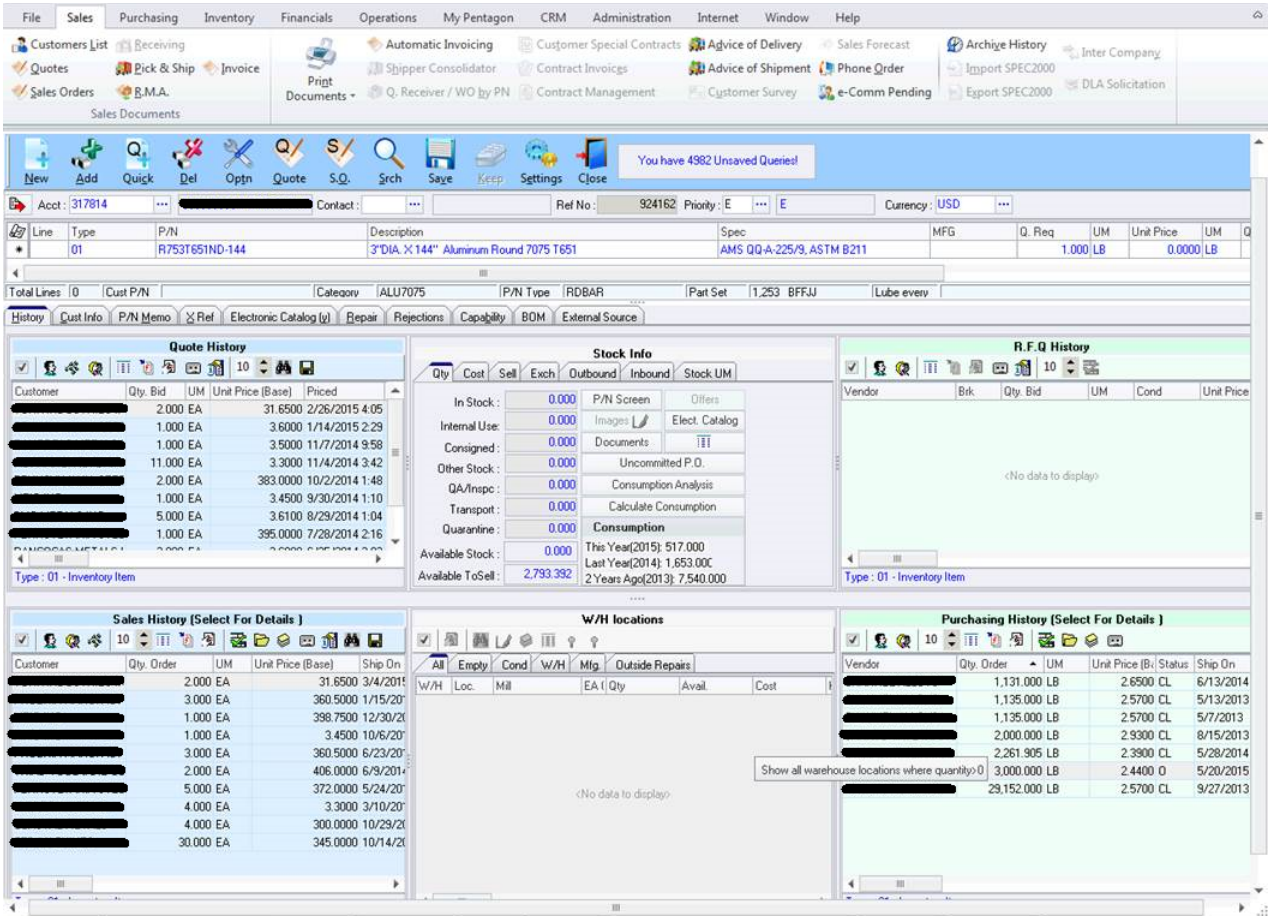


Table A1: Summary of Product Categories in the Data

	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					

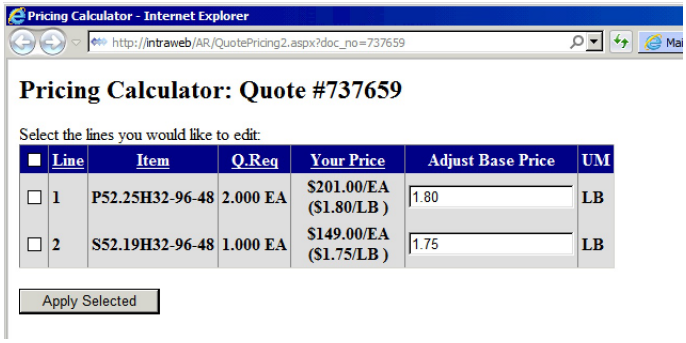
B Field Experiment

B.1 Field Experiment Forms

Figure A2: Field Experiment Edit Forms

(a) Treatment Edit Form

(b) Control Edit Form



B.2 Field Experiment SUTVA Analysis

In this section 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 price per pound p^t_{lqis}, on the set of line and time-varying client characteristics, x^p_{lqi}, salesperson fixed effect, salesperson-client random effect, α^p_{is} as well a on $T^{p,t-1}_s$, dummy indicating whether the previous quote priced by salesperson s was treated:

$$p^t_{lqis} \sim \alpha^p_{is} + \boldsymbol{\rho_s x^p_{lqi}} + \kappa^p_T \; T^{p,t-1}_s + \epsilon^p_{lqis}, \tag{12}$$

where ϵ^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 4,207 pricing decisions. The results of the regression are shown in Table A3. The coefficient of interest is the coefficient kappa of the previous quote treated. We find that there was no significant effect of previous treatment on current period pricing, suggesting that no significant learning due to past treatment occurred on the part of the salespeople.

Table A3: Price regression

Variable	**Coefficient**	**Std. err.**
Cost per lb.	1.151***	(0.036)
LME per lb.	1.241	(6.009)
LME volatility	26.96	(32.657)
Weight (log)	-0.790***	(0.024)
Relative weight	0.690***	(0.083)
Cut/weight	38.57***	(1.262)
Recency	0.0000587	(0.000)
Frequency	-0.122**	(0.046)
Monetary	-0.0147	(0.023)
Regular salesperson	-0.183	(0.117)
Foot base	0.165	(0.160)
Previous quote treated	-0.0937	(0.059)
Constant	4.222	(5.125)
Observations	4,207	
R^2	61.06%	

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

C Counterfactuals Data

Table A4: 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 A5: 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

Table A6: 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 A	0	(.)
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 Cold Finish	0	(.)
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 q1	0	(.)
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

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 6) to estimate the demand controlling for possible price endogeneity. We then used 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 A7 and A8, respectively). As expected, higher cost and cut requirements increase the price. With respect to clients' quote 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. For reference price and consistent with loss aversion we find a much stronger effect for loss relative to gains. Overall, the demand model predicts acceptance probability in the hold-out sample to be 61.1% compared to observed conversion rate of 59.3% .

Table A7: 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	

Model with client random effect
*** $p < 0.001$

Table A8: Parameter Estimates for Client's Acceptance Decision

Parameter	**Mean**	**Mean SE**	**Std. dev.**	$Q_{2.5}$	$Q_{97.5}$
Intercept	1.299	0.006	0.202	0.901	1.699
Gain	-0.072	0.001	0.019	-0.106	-0.035
Loss	-0.136	0.001	0.017	-0.171	-0.103
Recency	-0.001	0.000	0.000	-0.001	-0.001
Weight (log)	-0.299	0.000	0.013	-0.327	-0.274
LME	0.226	0.008	0.248	-0.293	0.704
LME volatility	0.023	0.001	0.037	-0.049	0.096
Regular salesperson	0.566	0.002	0.061	0.440	0.682
Aluminum - Cold Finish	-0.024	0.004	0.090	-0.202	0.152
Aluminum - Plates, Aerospace	0.059	0.003	0.086	-0.114	0.236
Aluminum - Plates, Commercial	0.327	0.003	0.066	0.203	0.457
Aluminum - Round, Flat, Square Solids	0.283	0.003	0.060	0.164	0.401
Aluminum - Shapes and Hollows	0.063	0.465	0.718		
Aluminum - Sheets, Aerospace	-0.782	0.008	0.249	-1.287	-0.301
Aluminum - Sheets, Commercial	0.386	0.003	0.073	0.237	0.525
Other Metals	0.856	0.005	0.143	0.557	1.126
Aluminum - Sheets, Commercial	0.473	0.013	0.405	-0.335	1.284
γ	-0.060	0.000	0.015	-0.089	-0.031
σ	0.020	0.000	0.011	0.001	0.043

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 to code the model into the 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., 2017).

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 log-margins as the dependent variables 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 variable the average log 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 A9 shows the parameters for which a randomized search was conducted and the 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 margins of each line as a risk metric corresponding to the expected value of the squared error.

Table A10 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[19]. For every model we report the in sample and out of sample RSME score. 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 effect, which has a significant share in explaining variance in pricing decisions. The random forest model, on the other hand, performs better both in- and out-of-sample.

We also compared, 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 random forest model's prices generated expected profits about 14% lower than those of the linear model ($\Pi[RF] = \$2,204,991$

[19] All scores are based on models' predictions before adjusting for the regime shift observed in the validation period.

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 random forest 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 is 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 A9: 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 A10: Comparison of Models - Fit and Prediction RMSE

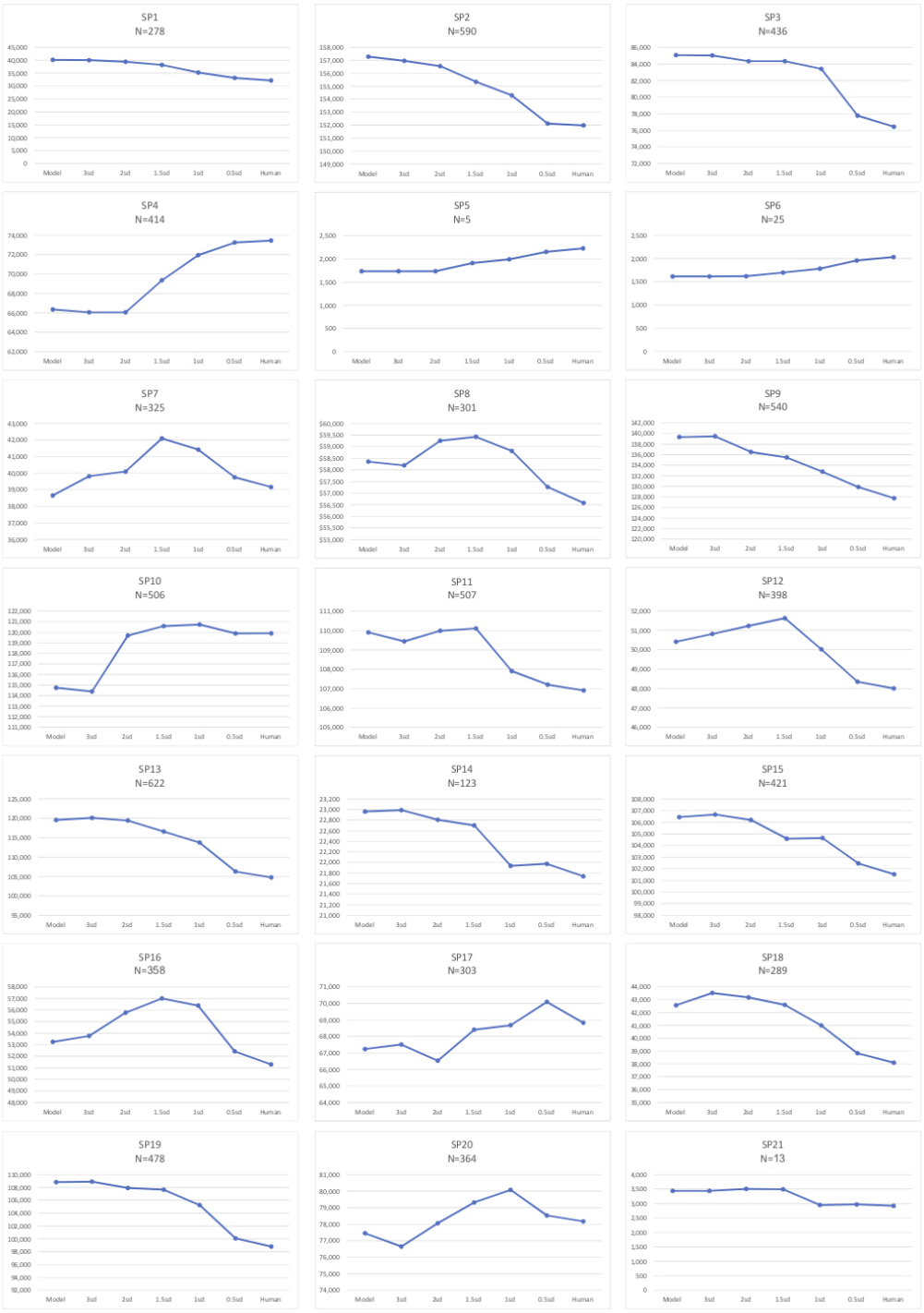
	Saleperson	N Train	N Test	Lasso in	Lasso out	Ridge in	Ridge out	RF[20] in	RF out	Linear[21] in	Linear out
1	AR01	5,295	2,030	0.643	0.559	0.641	0.555	0.447	0.540	0.588	0.537
2	AS03	3,089	1,079	0.647	0.511	0.639	0.499	0.495	0.494	0.575	0.523
3	BM01	376	82	0.584	0.878	0.569	0.825	0.538	0.914	0.496	0.872
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

[20]Random Forest

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

F Human-Judgment Hybrid - Breakdown by Salesperson

Figure A3: Expected Profits by Pricing Schemes



G Machine Learning Hybrid Analysis

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 features with the highest importance in determining the prediction are the weight of the quote, cost per lb., and dollar amount of previous quote. These are followed by number of lines per quote, frequency of purchases by the client and the ratio of quotes priced by the salesperson for this client. The full ranking of feature importance is displayed in Figure A4.

Figure A4: Feature Importance in Random Forest