Automation, Decision Making and Business to Business Pricing

Yael Karlinsky Shichor

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Automation, Decision Making and Business to Business Pricing

Yael Karlinsky Shichor

In a world going towards automation, I ask whether salespeople making pricing decisions in a high human interaction environment such as business to business (B2B) retail, could be automated, and under what conditions it would be most beneficial. I propose a hybrid approach to automation that combines the expert salesperson and an artificial intelligence model of the salesperson in making pricing decisions in B2B. The hybrid approach preserves individual and organizational knowledge both by learning the expert's decision making behavior and by keeping the expert in the decision making process for decisions that require human judgment. Using sales transactions data from a B2B aluminum retailer, I create an automated version of each salesperson, that learns the salesperson's pricing policy based on her past pricing decisions. In a field experiment, I provide salespeople in the B2B retailer with their own model's price recommendations through their CRM system in real-time, and allow them to adjust their original pricing accordingly. I find that despite the loss of non-codeable information that is available to the salesperson but not to the model, providing the model's price increases profits for treated quotes by as much as 10% relative to a control condition, which translates to approximately $1.3 million in yearly profits. Using a counterfactual analysis, I also find that a hybrid pricing approach, that follows the model's pricing most of time, but defers to the salesperson's pricing when the model is missing important information is more profitable than pure automation or pure reliance on the salesperson's pricing. I find that in most cases the model's scalability and consistency lead to better pricing decisions that translate to higher profits, but when pricing uncommon products or pricing for unfamiliar clients it is best to use human judgment. I investigate different ways, including machine learning methods, to model the salesperson's behavior and to combine salespeople's expertise as reflected by their automated representations, and discuss implications for automation of tasks that involve soft skills.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

A special thank you to the special people in my life that made this achievement possible:

To my husband and true partner, Ilan, I could not have done this without your continuous support and love. Thank you for always believing in me and for dreaming big for me, for us and for our family.

To my dear parents, Zehava and Itshak Karlinsky, your love and reassurance have been essential to me during this long journey. Thank you for always being there for me even when we were miles apart.

To my beloved children, Yonatan, Daniel and Naomi, who made this impossible and possible at the same time. Thank you for putting everything in life into perspective.

To my advisor, Oded Netzer, from whom I have learned so much. Thank you for being not just my advisor, but my mentor and true friend for years to come.

DEDICATION

To Yonatan, Daniel and Naomi

Be Kind and Do Good

I Believe in You

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 computerization and computational methods expanded the limits of automation to include non-routine tasks (Brynjolfsson and McAfee, 2012; Chui et al., 2016). The limits for automation have now become aspects of the job that involve perception and manipulation, creative intelligence and social intelligence (Frey and Osborne, 2017). While early estimations of the extent to which automation will take over human jobs presented a pessimistic future to the human employee (Frey and Osborne, 2013), the current view is that while some jobs are predicted to be replaced by machines altogether, most occupations will be affected by automation to a limited extent, presenting a combination of human and machine labor. Indeed, a recent OECD report (Nedelkoska and Quintini, 2018) found that automation will significantly change the skill set required for one-third of the jobs reviewed in the study.

Some recent applications of automation and AI have pushed the boundaries of automation to tasks such as screening resumes for white collar jobs (Cowgill, 2017), scanning X-ray or CT images to identify irregularities in the image [1], or replacing judges deciding whether defendants will await trial at home or in jail (Kleinberg et al., 2017). Yet, a common characteristic of the above examples is that while they require a high level of expertise (medical doctors, human resource personnel or court judges), the problem is still relatively well-defined and subjective cues in the environment should play little role in the decision process.

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

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 I ask in this dissertation 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 and not just noise?

Specifically, the objective of my work is to investigate the potential and challenges of introducing automation to one such domain with high importance to marketers: pricing decision making in a business to business (B2B). The B2B market is estimated at trillions of dollars, yet it largely lags behind the business-to-consumer (B2C) market in terms of adoption of 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 often repetitive and arguably predictable. Salespeople in B2B work in a fast-pace environment, making their pricing process almost "automatic" at times. On the other hand, such pricing decisions often involve high degree of inter-personal communication, long-term relationship 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 combination of human and machine decisions in B2B pricing. Furthermore, while in the aforementioned examples the decision was typically binary (e.g., invite for interview or not, await trial at home or in jail), the pricing decision problem is layered: the outcome, profitability to the company, is a non-linear function of the expert's decision variable, pricing.

I explore the trade-off between the benefits of automating the pricing decision and the value of soft skills in the context of salespeople making pricing decisions in a business to business (B2B) environment. I further investigate ways to preserve the human knowledge in building the automation approach and suggest a hybrid pricing scheme that relies on automatic pricing most of time and refers to the salesperson in cases where the value of the information processed and held by the salesperson (e.g., based on interpersonal communication with the client) is likely to be high.

I use data from an aluminum B2B 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 SKUs, customizable products and varying commodity prices, permitting the salespeople to determine prices 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. The salesperson may identify the client's state of mind over a phone conversation and adjust prices according to her assessment of the client's willingness to pay. Accordingly, it is highly unclear whether the pricing process could be automated in this context given the great share of relationship-based communication in the role of the salesperson making pricing decisions.

I propose an approach to automating the B2B salesperson by creating an artificial intelligence version of the salesperson that mimics its past pricing decisions and applies it systematically to new pricing decisions. I 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 to the salesperson when making the pricing decision (e.g., cost of the material, the size of the order, whether a cut is needed or the identity of the client). By modeling past pricing decisions, I estimate the weight given by the salesperson to each observed variable when setting prices. The approach, that uses the decision variable (price margin) rather than the outcome (whether the client accepted the price, or, alternatively, 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 me to not only reveal the salesperson's pricing policy, but also, assuming that the model is correctly specified, preserve the salesperson's expertise and knowledge as well as potentially identify cases where private information existed and guided her pricing.

In order to test the performance of the bootstrap-pricing model relatively to the performance of the salesperson in generating profits to the company, I 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 4,243 SKUs 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. I worked with the firm to integrate my 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. The salesperson could decide whether to adjust the price she offered to the client according to the recommended price or keep her own price.

The field experiment reveals that the effect of providing salespeople with price recommendation of their own model leads to substantially statistically significantly higher profits than not providing such a recommendation. Specifically, I find that relatively to the control condition, mean gross profit per line within a quote in the treatment condition is $9.53 higher, 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. While compliance with the model's recommendations, i.e., cases in which salespeople chose to fully or partially adopt the recommendation, was relatively low, I find that salespeople complied more when pricing for frequently contacted clients or for frequently purchased product categories. This suggests that in those cases the model captured the salesperson's policy better than in other cases, pointing to the potential of a hybrid approach in which the model and the salesperson each address different types of quotes.

To further explore the potential of automating the B2B salesperson's pricing decision, I perform several counterfactual analyses, which allow me 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 full automation of salespeople.

Given alternative pricing schemes (model pricing vs. salesperson pricing), I create a profit counterfactual for each quote. I calculate the expected profit under the model and the salesperson pricing and compare profitability at the quote level. For that purpose, I 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. I find that despite the loss of valuable information available to the salesperson but not to the model, the expected profitability of pure automation (the model prices all quotes) is 5.3% 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 field experiment and the literature on B2B suggest that in some cases the information held by the salesperson could lead to higher-profitability pricing decisions than the model. Using my modeling approach I identify cases in which the salesperson is possibly relying on information that the model does not have in making the pricing decision. I 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's price. This hybrid pricing scheme leads to an additional increase in profit, 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). I find that salespeople that deal with complex quotes (e.g., high cost) are less likely to be replaced by their own model, and discuss the effect of automation on salespeople by their level of expertise.

In addition to the hybrid pricing scheme that combines the model and the salesperson I investigate the performance of pricing schemes based on combination of models of salespeople to preserve expertise. I find that salespeople develop expertise related to specific clients or products, and that aggregating models based on several salespeople leads to better performance than using even a single "best" salesperson's model to make all pricing decisions.

I also explore different, possibly more sophisticated, ways to create an automated version of the salesperson by utilizing machine learning (ML) tools to create the bootstrap-model of the salesperson. I find that, in my application, the simple linear model with client random effect used in the experiment performs well relatively to the more complex ML models both in terms of predictions and increased profitability.

In this work I demonstrate that salesmanship in business to business (B2B) is one such occupation that could potentially be transformed by automation, and should be transformed by introducing automation in order to improve its decision making processes. I show that for an occupation that combines repetitive and predictable tasks with soft skills of sense-making and communication, combining machine and human is superior to letting either the machine or the person perform the task in its entirety. Through a field experiment and various simulated analyses, I 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. As a testimony to the importance of my work, the company is now implementing my model permanently into its CRM system.

The discussion on automation often revolves around its economic value in reducing human labor expenses, but the potential benefit of automation is not only the financial savings associated with replacing a human employee with an automated process. In many cases the algorithm is not only less costly than the expert, but also does the expert's job better than the expert herself. I find that most pricing decisions in B2B are better be made by the model for higher profitability, while the expert salesperson is essential in pricing quotes that are unique and out of the ordinary.

The remaining of the dissertation is organized as follows: In Section 2 I discuss my contribution to existing work on B2B pricing and automation. In the first Essay I describe my approach to automating the salesperson and the details of the field experiment I used to test it. In Section 3 I describe 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, its benefits as well as limitations. In the second Essay I approach the problem of automating the B2B salesperson from a different method that allows me to test the full potential of automation and investigate the conditions to when it works best. In Section 5 I describe the counterfactual analysis used to simulate full automation. In Section 6 I create the hybrid pricing scheme and discuss conditions to when we should use the model and when the salesperson. In Section 7 I suggest alternative ways to combine models of salespeople and in Section 8 I describe alternative methods to modeling the salesperson. I conclude and summarize the two Essays in Section 9 by discussing implications of my findings to sales force automation and other tasks that involve soft skills.

2 B2B Pricing and Automation

My work builds on and contributes to several streams of literature. First, I add to the relatively limited literature on B2B, specifically on B2B pricing. The B2B market was estimated at over $8 trillion in transactions in 2014. More and more sellers face business clients that prefer to interact and place orders via e-commerce (Forrester, 2015). 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 pricing dynamics in settings similar to ours, and found that pricing behavior drives the relationship of clients with the company and subsequently their demand. 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 (Morgan and Hunt, 1994) and variation of prices across clients and across purchases is common (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 in light of the great share that communication holds in revealing information that affects pricing decisions.

Second, the roots for my approach to automation are in the behavioral judgment literature as well as decision models literature. The former stressed the idea that models of experts trumpet experts in judgments and decision making (Meehl 1954; Dawes 1979). For example, Dawes (1971) found that a simple linear model of three components (Graduate Record Examinations, Grade Point Average and a measure of the student's undergraduate institution quality) was more predictive of graduate students ratings than the admissions committee's evaluation based on those three components exactly. In a related application, Wiggins and Kolen (1971) asked students to forecast other first-year-graduates grade point average based on ten cues reported in the students' records. They found that regressing the forecasts of the students on the cues led to prediction of the grades that was more accurate than the students' original forecasts.

This approach is often referred to in the behavioral judgment literature as *judgmental bootstrapping*. It uses the judgment (e.g. students' forecasts of grades), rather than the outcome (e.g. students' actual grades) as the dependent variable in the regression. 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 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).

A strong (yet often implicit) assumption underlying the superiority of models over experts in the behavioral judgment literature, is that most 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 most real-world problems the expert has access to richer information than the model does. The model may be consistent, but it lacks possibly important information available to the human decision maker (e.g., information exchanged through interpersonal communication).

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, salespeople work in a dynamic environment and are exposed to cues which may steer them wrong on a case-by-case judgment. B2B salesforce pricing decisions are based heavily on interactions with the client and 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 overcharge the client. While the model's consistency may lead to better pricing decisions in many cases, in other cases 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 to provide decision models (Little, 1970) in the form of aid tools to the manager, where the goal is to provide a parsimonious and usable tool. Rich literature on decision support systems (DSS) describes the benefits of allowing a manager to use an automated system in making decisions (e.g., Sharda et al., 1988; Eliashberg et al., 2000). 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). My work goes beyond decision models or support systems in two important ways: 1) I automatically learn the weights of the expert by bootstrapping her historical decisions, by that allowing for more efficient and exhaustive extraction of the expert's knowledge, expertise and decision behavior, and 2) realizing that only in some cases the expert's input is beneficial for the decision, while in others the model can in fact make a better decision than the salesperson, I identify cases in which the model should make the decision with no additional input from the expert. That is, while the goal of DSS is primarily to support the human decision maker, I move from support to automation and allow the model to make decisions autonomously and automatically. Furthermore, I argue for automation of the assignment of who, the model or the human expert, should make the decision.

Third, I 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, the demonstration of successful automation of soft skills is still sparse (Deming, 2015). Research in labor economics shows that automation can substitute for 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). Moreover, by definition artificial intelligence representations of human judgment tasks such as judicial or human resource selection decisions (e.g., Kleinberg et al., 2017; Cowgill, 2017) are trained on historical judgments and responses to historical inputs. As a result, they perform fairly well in stationary environments, but fail to appropriately respond to dynamic or non-stationary inputs generated by policy changes or unseen variables, in the absence of dynamic adjustment or re-training of the model (e.g., Ditzler et al., 2015; Lughofer and Sayed-Mouchaweh, 2015).

The salesperson's job is a combination of repetitive, technical calculation of prices based on quote characteristics and the delicate use of social skills through communication to understand the client's state of mind and leverage it to maximize profits. Moreover, while salespeople develop expertise (that the model can learn) with clients or products over time, new clients may approach the company, or existing clients may request for new product specifications, presenting cases unseen before to the salesperson and hence to the model. Indeed, I 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.

Essay I

Automating The B2B Salesperson

3 The Model of the Salesperson

My 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, I estimate a linear regression of previous margins on a set of variables available to the salesperson at the time of decision. Although I 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 and not to find optimal prices that maximize profits. 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 I 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) who use the product as a component in their own product or service. The company sells thousands of stock-keeping units (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 my data, see Table A2). Aluminum categories vary on 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 full freedom in pricing any product on a quote-by-quote basis, providing different prices to different clients or even different prices to the same client over time.

A client may request for a 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 quantity. A quote from a client may include only one SKU or multiple lines (SKUs). After receiving the order's specifications, the salesperson provides a price quote [2]. Orders may be priced per pound, per feet or per unit. Salespeople are guided to maintain high price margins. Although pricing is done by unit or by weight unit, salespeople verify that the price meets margins requirements. The firm calculates price margin for line l in quote q by client i as follows:

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

where c_{lq} is the cost per pound of the material and p_{lqi} is the price per pound provided to 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. I 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

[2]Shipping costs are priced separately, one line per quote, and I don not model those costs.

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

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 quotes per client. Each of 17 salespeople in the sample made on average over 8,000 pricing decisions. A sales order may include one or more product specifications, each line priced by its own. 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 and Table A2 in the Appendix for frequencies of the product categories in the data.

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
Market price per lb.	0.76	0.07	0.68	0.75	0.86
Market price volatility	0.01	0.00	0.00	0.01	0.01
Weight	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
Regular salesperson	0.78	0.31	0.14	0.93	1.00
Total = 139,869					

†Calculated at the product category level

3.2 Model Specification

The performance of salespeople in the firm is measured by profitability margins of their prices. Therefore, price margin is naturally the criterion of the pricing model. Margin is defined as specified above in Equation 1 and is calculated at the line level (a quote may include one or more lines with different part numbers). Price margin could range from zero to one, and is skewed to the left in the data. The average line margin in the data is 41%

and the median is 36%. Consequently, I use the logarithmic transformation of price margin as the dependent variable of the margin equation.

In building the model I attempt to include all the information available to the salesperson at the time of the pricing decision. To identify that information I conducted several interviews with senior management and salespeople in the firm to get an idea of the information flow along the pricing process. I then explored the CRM software salespeople use when determining prices to create a list of variables hypothesized to affect pricing (see a screenshot of the CRM system in the Appendix, Figure A1). The model includes the following variables:

a. **Product category**. Dummy variable for eight out of nine product categories the retailer sells.

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 (specifically, in the data the largest quote has 85 lines). Pricing may be different for the same product specification, depending on whether it is requested in itself or as part of a larger order. In the latter case price may be lower because profit comes from multiple items, therefore the salesperson does not have to imply minimum profitability on a single line. To account for that I include in the margin equation the relative weight of the line out of the overall weight of the quote.

d. **Cut**. When the client requests for a made-to-order piece, processing is required. To account for the additional work required to process large orders, I insert cut to the margin equation as an interaction between the cut dummy variable and $1/weight$.

e. **Cost**. In the sales system, the salesperson observes cost per pound for the requested part number. I observe and include in the model the cost as it appears 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). I 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. **FT-base**. While the 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 FT and do not have a weight-based price. These items consist of 3.5% of the data and a dummy is included to identify them, because their pricing may be different than items priced per pound. For these SKUs margin calculation is based on price and cost per FT.

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 and based on information on competition and local market potential). I include priority in my 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.

   (b) **Recency, frequency and monetary**. Recency is defined as days since the client's last quote request from the same product category, frequency as the client's running average of requests from the product category per week, and monetary as the log of the total amount previously requested from the product category based on the client's quoting and purchase history. Recency, frequency and monetary are calculated at the category level to capture category-specific purchase habits.

In the calculation of recency, frequency and monetary measures I include quotes that were not converted to sales, under the assumption that the client decided to purchase the product somewhere else, nevertheless the quote reflected a pattern of purchase.

(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 endogenous pricing based on the client's identity I include client random effect in the model.

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 in the wrong direction (e.g. the salesperson's behavior may become more lenient). As a measure for the relationship of the salesperson with the client I calculate the proportion of quotes that the salesperson priced out of the total number of quotes received by the retailer from the client. That is, I measure to what extent this is the client's regular salesperson. This is a running ratio calculated up-to-date. The data show that on average, the same salesperson handles the client nearly 80% of the time.

j. **Time dummies**. To control for any time trends that affect pricing, quarter dummies are included in the equation.

3.3 Model Estimation and Results

I 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, I 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} , and salesperson-client random effect, α_{is} for salesperson s and client i[4]:

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

where ϵ_{lqis} is a normally distributed random shock. In the subsequent analyses I use the margins predicted by the individual-salesperson models; however, to get a sense for the effect each variable has on log margin I 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 A3 in the Appendix for average estimates across salespeople based on the individual regressions).

I find that when cost increases, the company decreases its margins. However, when the daily metal price increases, the company seems to increase margins (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. The larger the order, the lower the margin charged. Similarly, the larger share the line takes of the total order, the higher the margins of that line, indicating an order with less items. Processing (cut) increases margins as expected and the positive sign of weight/cut indicates that for smaller quotes the margin increases more due to processing. Lastly, the small number of SKUs that are stocked in feet is priced with lower margins relatively to the majority of items stocked and priced in pounds.

In terms of client behavior, out of the three RFM measures, the company provides lower margins to customers who buy more frequently, but salespeople increase margins based on large previous purchase or quoting for the product category. I find that when clients interact with their regular salesperson they receive lower margins, suggesting that relationship building may lead to lower margins. It also seems like the salespeople's pricing scheme is following the structure of priorities defined by the company. When clients gain higher priority, they

[4]We explored different model specifications (e.g. non-linear specifications of weight, interactions of RFM) which yielded lower or comparable fit and consequently opted for the parsimonious model with the best fit.

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)
FT 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 A	0	(.)
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 - Cold Finish	0	(.)
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)
2016q1	0	(.)
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%	

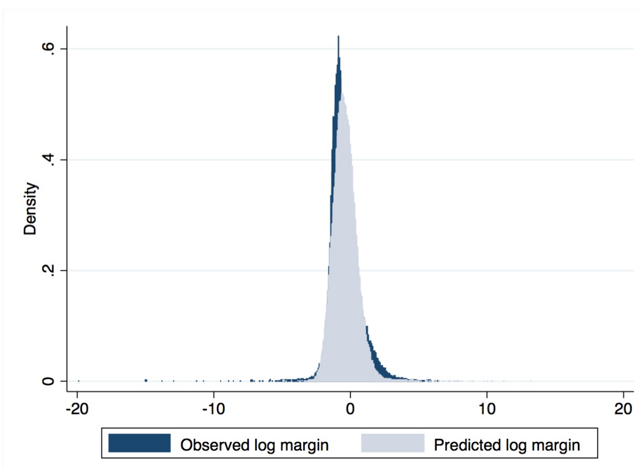
* $p < 0.05$, ** $p < 0.01$,*** $p < 0.001$

Note: regression includes client random-effect and salesperson fixed effect receive lower margins relatively to being at lower priorities.

Finally, it is worth noting the positive time trend captured by the quarter dummies. Across six quarters, there is a consistent increase in average 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 reflected by the somewhat higher margins for low priority clients.

The aggregate model explains way over half of the variation in the pricing policy ($R^2 = 0.67$), suggesting fairly good model fit. Figure 1 shows observed and predicted log margin values based on the individual models estimated separately for each salesperson. It is apparent that the model's specification is capturing salespeople's pricing policy well. Indeed, when converting log margin back to margin, the average predicted line margin of 41.96% is similar to the average observed line margin of 41.14%.

Figure 1: Log Margin Fit



4 Randomized Field Experiment

Now that I created an individual model for every salesperson in the company, I can use those individual bootstrap-pricing models to directly evaluate the causal effect of automation by replacing or aiding the salesforce's pricing task. For that task, I collaborated with the aluminum retailer to conduct a large-scale field experiment. While I did not completely replace salespeople in making pricing decisions, the company agreed that for a randomly selected set of orders, I 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, I created a "price calculator", that upon entering a new quote to the system takes as input the quote, client, and salesperson characteristics and using Equation 2, in real time, 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) [5]. The regular work flow of the salespeople is as follows: when a client calls (or emails) for a new product request, the salesperson enters a new quote into the CRM system. She then saves the quote and is able to edit prices within the quote. When she is ready to send the quote for the client's approval, the salesperson re-opens the system, generates a price quote document and sends it to the client via email.

The experimental intervention in this process was upon 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 (see Figure A3 in the Appendix for

[5] 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.

a screenshot of the email). The salesperson could then either click *Accept suggested prices* to update the sales system to reflect the model prices, *Accept original prices* to keep her original prices, or *Edit*, which would open an edit form (see Figure A5 in the Appendix for a screenshot of the Edit form). In the edit form the salesperson could accept the model's price for only some of the lines, as well as to edit prices 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 A2 in the Appendix.

Because treatment involved an extra step, of evaluating the e-mails 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 A4 in the Appendix 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 A6 in the Appendix for a screenshot of the control condition 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 prior to the experiment).

It is important to note, that when entering her original price quote, the salesperson did not know whether this quote belongs to the treatment or control group (i.e., whether she would receive a price recommendation or not), hence the original price quotes are independent of the experimental design. This unique design gives me knowledge of three data points for each quote, whether it was assigned to the control or the treatment group: the original price set by the salesperson, the model's recommended price (which I 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 only knows the outcome under the different tested policies. This design gives me access to the original pricing decision of the salesperson, before the assignment of treatment has been realized. Knowing that, enables me to better understand patterns of pricing.

I ran the experiment for eight consecutive business days. Prior to the commencement of the experiment, I let the salespeople experience the tool for four business days, during which I adjusted the tool to fit best into their work flow and corrected any technical issues that arose. During those days I visited two out of the three locations the firm has (I visited the NY and PA locations) [6] and made sure salespeople were feeling comfortable with the tool and understood its flow. After excluding missing or erroneous values, as well orders with extreme weight [7] , 2,106 quotes made by 1,053 clients remained in the sample, with a total of 4,243 pricing decisions (some quotes had multiple lines, and each line is a pricing decision). 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 and using the tool, was 10%, and varied across salespeople.

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. I intentionally over-weighted treatment over control with anticipation of low compliance rates and the hope that a higher proportion of treated quotes

[6] Phone conversation were made with sales people in the third location (CA).

[7] I do not include in the analysis the following extreme-values and unique lines:

a the top and bottom 1% of orders by weight. The order is priced by a manager if very large or follows some irregular pricing rules if very small.

b Lines with price or cost lower or equal to zero, or missing.

c Lines with price per lb. larger than $20.

d Orders of over 8,000 lbs.

e Contractual clients or clients that interact with the company on a basis close to contractual, based on information from the company's management and the history of the client with the company beginning January 2016 and up to the commencement of the experiment.

will balance out non-compliance behavior. Randomization was done by the company, and as expected, 59.68% 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. I 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, I can conclude that no omitted covariate led to different pricing 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.	1.7993	1.7652	0.0341	0.0426	0.4236
Weight	707.2152	694.5144	12.7008	50.4559	0.8013
cut_ratio	0.3068	0.3076	-0.0008	0.0200	0.9697
Total lines	2.0766	1.9730	0.1036	0.0983	0.2920
Original price per lb.	3.4945	3.4812	0.0133	0.1184	0.9103
Model price per lb.	3.6729	3.6685	0.0043	0.1219	0.9717
Price difference	0.7272	0.7408	-0.0135	0.0682	0.8429
Number of quotes	849	1,257			

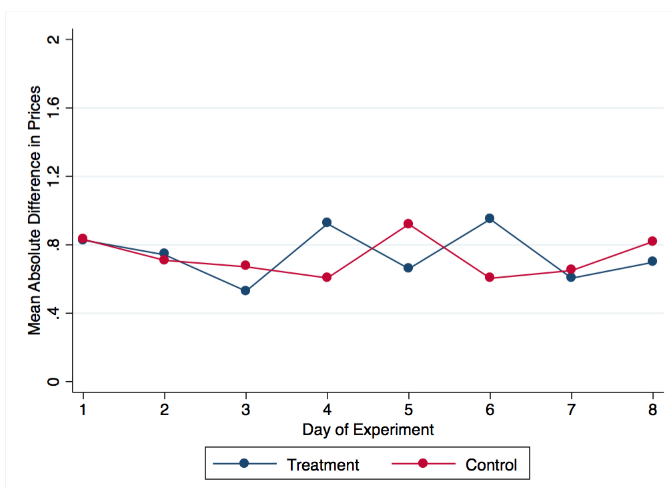
4.1.2 Stable Unit Treatment Value Assumption

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. In what follows I show that there was no spill over of treatment effect on the pricing process during the experiment.

The treatment may affect quotes in the control condition if the salesperson is changing her pricing policy during the experiment even for quotes for which she did not receive a price recommendation. One possible mechanism through which such contamination may happen is learning. If, for example, the salesperson receives a few consecutive treatment emails recommending higher prices than her original prices, she may adjust her pricing upwards on the next quotes, affecting both future treatment and control quotes.

To evaluate the extent to which learning is affecting pricing, I 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 2 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, relieving the concern of learning or violation of SUTVA.

Figure 2: Average Difference between Quote Model-Price and Original Price Over Time: Treatment vs. Control



To further verify that the stable unit treatment value assumption was not violated, I test whether the treatment given to a quote affects the pricing by the same salesperson in the following quote. For each line l of each quote q priced by salesperson s for client i at time t I regress the price per pound p^t_{lqis}, on the set of line characteristics, time-varying client characteristics and salesperson fixed effect, x^p_{lqi} , and salesperson-client random effect, α^p_{is} as well a on $T^{p,t-1}_s$, a dummy indicating the previous quote priced by salesperson s was treated:

$$p^t_{lqis} \sim \alpha^p_{is} + \boldsymbol{\rho_s x^p_{lqi}} + \kappa^p_T \; T^{p,t-1}_s + \epsilon^p_{lqis}, \tag{3}$$

where ϵ^p_{lqis} is a normally distributed random shock. Because this analysis can be done only for the second quote and on by each salesperson, the usable sample size for the regression is 4,206 pricing decisions (and a total of 2,089 quotes). The results of the regression are shown in Table 4 and as desired, the treatment given to the previous quote priced by the salesperson did not affect pricing in the current quote.

Table 4: Price regression

Cost per lb.	1.148***	(0.036)
LME per lb.	-0.546	(6.082)
LME volatility	32.73	(33.023)
Weight (log)	-0.850***	(0.024)
Relative weight	0.737***	(0.084)
Cut/weight	30.89***	(1.059)
Recency	0.0000645	(0.000)
Frequency	-0.111*	(0.046)
Monetary	-0.0153	(0.023)
Regular salesperson	-0.189	(0.116)
FT base	0.168	(0.163)
Previous quote treated	-0.0988	(0.059)
Constant	6.034	(5.187)
Observations	4,206	
R^2	60.8%	

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

4.2 Experiment Results

To test the effectiveness of the experiment I 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 a Tobit type model would not be appropriate. Accordingly, I use a non-parametric test to compare the average GP on both groups: treatment and control. In addition, although randomization was done at the quote level, pricing is done separately, but not independently, for each line within the quote. Consequently, I cluster the standard errors across lines of the same order. Considering the distributional constraints of GP and the non-independence of lines within a quote, I 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 [8]. I find a significant increase of $9.53 in gross profit per line in a quote in the treatment condition vs. line in a quote in the control condition ($GP_{control}$ = \$93.76, $GP_{treatment}$ = \$103.29, $Z = -2.007,\ p = 0.045$) [9]. Overall, the increase in profits is equal to nearly $24,000 during the eight days of the experiment, and more than $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.

4.3 Regression Analysis

In order to further understand the mechanism behind the positive effect of providing price recommendations to quotes in real time, I 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

[8] For the small minority of orders in which some lines were rejected and some were accepted, declined lines will have zero GP.

[9] While the pricing model predicts margins, I measure the treatment effect on the outcome - profits. This is because the company's outcome of interest is profits and indeed profits are a function of price margins. Nevertheless, when performing the non-parametric clustered Wilcoxon rank sum test on line margin, rather than gross profit, the treatment effect is positive and significant $Z = -2.68,\ p = 0.007$

suggested price by the client) and profit level conditional on acceptance of the price [10]. I estimate the following set of equations for the Cragg hurdle model:

$$Pr_{lq} = \delta_{T1} T_q + \delta_{sp,1} I_{salesperson,q} + \delta_{day,1} I_{day,q} + \boldsymbol{\Theta_1 x_{lq}} + \varepsilon_{1lq}, \tag{4a}$$

$$GP_{lq} = \delta_{T2} T_q + \delta_{sp,2} I_{salesperson,q} + \delta_{day,2} I_{day,q} + \delta_{cost} cost_{lq} + \boldsymbol{\Theta_2 x_{lq}} + \varepsilon_{2lq} \tag{4b}$$

Equation 4a describes the client's likelihood of accepting the price quoted for line l within quote q, and equation 4b describes the line's gross profit (the price the client paid for the line minus the cost the of line to the firm) conditional on the client accepting the price. T_q is a dummy variable that equals one if order q was assigned to the treatment condition and zero otherwise, $\boldsymbol{x_{lq}}$ is a vector of line characteristics including line weight and whether the order required a cut (divided by the weight), and the gross profit equation includes the cost of the material as an additional control variable. $I_{salesperson,q}$ are a set of dummy variables to control for salesperson fixed effect and $I_{day,q}$ are a set of dummy variables to control for day of the experiment fixed effects. ϵ_{1lq} and ϵ_{1lq} are normally distributed random shocks.

The results of the analysis are shown in Table 5. 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 \$14.30 per line, controlling for the above salesperson and day fixed effects and for quote characteristics. Thus, I find that the treatment effect 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. Salespeople might make two types of errors in pricing: type I, when they price too high and

[10] 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 profits to the negative domain.

lose the deal, or type II, when they price too low and leave money on the table. I find the the model's effect is in correcting the first type and leading to higher quote conversion.

Table 5: Regression analysis using Cragg Hurdle Model

Variable	**Coefficient**	**Std. err.**
Client acceptance of price		
Treatment	0.171*	(0.073)
Line weight (log)	-0.0739***	(0.020)
Cut / weight	-1.152	(1.036)
Constant	0.133	(0.216)
Line gross profit		
Treatment	0.017	(0.039)
Line weight (log)	0.568***	(0.014)
Cost per lbs.	0.165***	(0.023)
Cut / weight	7.011***	(0.820)
Constant	2.133***	(0.100)
lnsigma		
Constant	-0.625***	(0.038)
Sigma	0.536	(0.020)
Marginal effect	**14.30***	(5.98)
Observations	4,243	
Pseudo R^2	10.88%	

Salesperson and day fixed effects included
* $p < 0.05$, *** $p < 0.001$

To further examine the increased acceptance rate by clients in the treatment condition I compare the model's recommended price to the original price set by the salesperson with respect to the client's response. Indeed, I 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 sale, the model recommended a higher price in only 60.2% of the cases (these proportions are significantly different, $t = 2.2612$, $p = 0.012$). This provides a suggestive evidence that the model's pricing corrects for over-pricing by the salespeople, which may lead to failure to convert the quote, consistent with the results shown in Table 5 of increased acceptance rates in the treatment condition. B2B salespeople often lobby for lower prices (Simester and Zhang, 2014), and indeed I find that the model's prices were higher than those of the salespeople in 62% of the cases. It does seem like salespeople are making both types of errors discussed above; nevertheless, there seems to be a mismatch in the cases for which they are lobbying for lower prices - while the model suggested increasing prices in some cases, the treatment effect comes from correcting over-pricing by the salespeople.

4.4 Compliance Analysis

One of the largest risks when conducting an experiment that requires cooperation of subjects is compliance. Specifically, when offered to rely on algorithmic decision aids, people may demonstrate *algorithm aversion* and limit their use of the aid tool. Among the reasons for this aversion are the belief that humans can reach near-perfection in decision making (Einhorn, 1986) and the belief that human predictions improve through experience (Highhouse, 2008). The latter is especially important when it come to experts decision making. Experts over weigh their experience and expertise in making decisions, and this over-confidence leads to poor predictability (Arkes et al. 1986; Camerer and Johnson 1997). Moreover, when facing the (inevitable) error of the algorithm, people are less likely to trust and use it (Dietvorst et al., 2015).

Over-confidence and dis-trust in the algorithm may pose significant risk when the design relies on salespeople to use their model's price recommendation. While training salespeople with the pricing tool as preparation to the experiment, they expressed great confidence in their own judgment. For example, one salesperson said that he was "not likely to follow the recommended price" because he had already "put a lot of thought into pricing the quote and considered everything there is to consider". Moreover, almost every salesperson I talked with, said that while the tool may be useful for other salespeople, her clients (or the quotes she typically prices) are "different".

Although treatment was assigned at the quote level, salespeople could comply at the line-within-the-quote level. For example, a salesperson could accept the price recommended for one line and reject the price recommended for the other line in a two-items quote. Consequently, compliance rate is calculated as the share of lines for which the salesperson either accepted or partially accepted ("edit") the model's recommendation out of the total number of lines she priced in the treatment group. Overall, compliance at the line level was relatively low, about 10% (248 out of 2,480 lines in the treatment condition).

Because compliance is inherently endogenous to the treatment (e.g., salespeople may be more/less compliant with the treatment for more/less profitable orders) I measured thus far the experiment's results in terms of intention to treat, which is exogenously determined, rather than treatment (based on compliance). Nevertheless, I can examine the cases in which salespeople chose to comply with the price recommendation to possibly understand some of the determinants for their decision. In order to do that, I ran a logit model for the salesperson's probability to comply with the model's price. The utility for salesperson s from complying with the model's price recommendation to quote q by client i is:

$$u_{sqi} = \boldsymbol{\vartheta x_{sqi}} + \nu_{salesperson,q} I_{salesperson,q} + \nu_{day} I_{day,q} + \varphi_{sqi} \tag{5}$$

where $I_{salesperson,q}$ are a set of dummy variables to control for salesperson fixed effect and and $I_{day,q}$ are a set of dummy variables to control for day of the experiment fixed effects. $\boldsymbol{x_q}$ is a set of line and client characteristics that includes: weight, cost per lb., cut, relative weight, recency, frequency and monetary at the product category level, regular salesperson measure, category and client priority.

Assuming that φ_{sqi} is extreme value distributed, the probability that salesperson s will comply with the price recommendation provided for quote q by client i follows the binary logit specification:

$$Pr_{sqi} = \frac{e^{u_{sqi}}}{1 + e^{u_{sqi}}} \tag{6}$$

Table 6 shows the estimation results of the logit regression for the salesperson's decision to comply with the model's recommendation. Salespeople choose to follow the model when they are pricing for their regular clients, suggesting that the model captures the way they price for these clients based on their joint history. In addition, though the effect is not significant, salespeople tend to comply when the relative share of the line in the quote is large, typically indicating of a quote with less lines, i.e. a less complicated quote. Finally, they comply when pricing the most common product category, suggesting that the model was able to capture their pricing policy well for those repeated and fairly predictable cases.

Table 6: Logit Model for Salesperson's Decision to Comply with the Price Recommendation

Variable	**Coefficient**	**Std. err.**
Weight (log)	0.061	(0.087)
Cost per lb.	0.211	(0.111)
Cut / weight	-1.683	(4.214)
Relative weight	0.559	(0.327)
Recency	0.0001	(0.001)
Frequency	0.122	(0.205)
Monetary	-0.095	(0.103)
Regular client	0.806*	(0.402)
Aluminum - Cold Finish	0	(.)
Aluminum - Plates, Aerospace	0.940	(0.677)
Aluminum - Plates, Commercial	0.548	(0.645)
Aluminum - Round, Flat, Square Solids	1.244*	(0.612)
Aluminum - Shapes and Hollows	0.521	(0.613)
Aluminum - Sheets, Aerospace	0	(.)
Aluminum - Sheets, Commercial	1.034	(0.645)
Other Metals	-0.550	(1.022)
Stainless - Other Stainless	0	(.)
Priority A	0	(.)
Priority B	-0.112	(0.498)
Priority C	0.630	(0.470)
Priority D	0.740	(0.831)
Priority E	0.032	(0.534)
Priority P	0.674	(0.685)
Constant	-7.634***	(1.537)
Observations	2,145†	
Pseudo R^2	27.80%	

* $p < 0.05$, *** $p < 0.001$
†335 observation dropped from the analysis due to collinearity
Salesperson and day fixed effects included in the model

4.5 Towards a Hybrid Approach to Automation

On one hand, allowing salespeople to make a judgment with regards to when to use the model's price and when not to use it, led to low compliance rates, which possibly limited the treatment effect. On the other hand, it is possible that salespeople decided when to comply with the model intelligibly, using their own prices when they realized that model did not price appropriately, and the model's price when the model "made sense". The salesperson would choose to use her own price when she has valuable information that the model was missing. For example, if the client expressed high urgency for the order over a phone conversation and the salesperson decided to take advantage of the client's need and over-charge him. In this case, the model would have no information of that profit opportunity and would recommend a lower price, which the salesperson would have rejected.

When quoting prices to clients, the B2B salesperson uses pricing rules and heuristics as well as information derived from continuous interactions with clients, relationship development and soft skills of salesmanship. While the repeated and predictable aspect of pricing could arguably be coded and automated (Arkes et al., 1986), human common sense and soft skills are still beyond the reach of automation (Frey and Osborne, 2017). In the behavioral judgment literature (Meehl, 1954), the non-codeable cases were coined as the *broken leg* cases. In those broken leg cases, the salesperson will outperform the model, because the model is missing crucial information that the salesperson has. Depending on the application domain, and possibly on the expert herself, the balance between codeable input and soft non-codeable input will differ.

While testing the individual bootstrap-pricing models in a field experiment provided a direct proof to the causal effect of the automation approach on the company's profitability, low compliance with the model limits my ability to recover the true share of codeable and non-codeable cases in the task of the B2B salesperson. Examining salespeople's compliance patterns provides preliminary and partial evidence that salespeople complied with the model when the client or product were "familiar". Evidently, salespeople in the experiment chose to not use automation in a large number of cases. Nevertheless, it is likely that at least in some of those cases they chose not to comply for the wrong reasons (e.g. over confidence). Therefore, it is still an open question what is the real share of codeable decisions in the B2B salesperson's task which is highly human-interaction based? And in what ways the salesperson and her model could be combined such both the expert and the model's advantages are utilized in a hybrid automation approach? I address these question in Essay II that follows.

Essay II

The Automation Hybrid

5 Counterfactuals Analysis

While the experiment allowed me to investigate, in a direct way, the causal effect of automation on profitability, as with any field experiment there were some limitations and constraints. First, the firm only allowed me to provide the model's prices as a recommendation or a decision support tool to salespeople, rather than replacing them completely in providing price quotes to clients. Particularly given the low compliance levels, this prevented me from fully testing the value of automation. Second, because the salesperson endogenously decided when to comply with the model, I cannot directly assess under which conditions it would be most profitable to use the model and under which condition to defer to the salesperson's pricing. Finally, given the cost involved in running such a price experiment, I 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, I 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 the profitability of different 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 I did not use the client's decision whether to accept or reject the quoted price in creating the automated salesperson (rather, I used the salesperson's decision - price margin), I 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. Therefore, I estimate demand as well as calculate profit counterfactuals at the quote level. Put formally, for each quote q requested by client i, based on observed prices p_{qi} and predicted prices $\hat{p_{qi}}$ (calculated based on the model's predicted margins), I calculate predicted acceptance probabilities, $Pr(p_{qi})$ and $Pr(\hat{p_{qi}})$ respectively. Then I can calculate the expected profit for quote q requested by client i in the following manner:

$$\Pi_{qi} = (p_{qi} - c_q) \times Pr_{qi}(p_{qi}), \tag{7}$$

$$\hat{\Pi}_{qi} = (\hat{p}_{qi} - c_q) \times Pr_{qi}(\hat{p}_{qi}), \tag{8}$$

and compare expected profits based on observed prices (Eq. 7) to expected profits based on predicted prices (Eq. 8). To calculate expected profits I need to estimate the probability of quote acceptance given price (the last term in equations 7 and 8), which is done in section 5.2, but prior to that I explore alternative specifications to the pricing model described equation in Equation 2.

5.1 Data for Counterfactuals

For the purpose of calculating profit counterfactuals, a longer period of data is required, that will facilitate the split of the data into calibration data and validation data. Consequently, I use an earlier 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. For the same reasons described in subsection 4.2 I exclude from the analysis extreme values of price and weight as well as contractual or close-to-contractual clients. Tables A4 - A7 in the Appendix show summary statistics and variables frequencies in the counterfactuals data.

Time Trend While the data used for the experiment for the counterfactual analysis mostly overlap, and therefore both exhibit the increase in margins over time discussed earlier, six months long of transactions were kept for validation. As can be seen in Appendix Table A8, in addition to the continuous increase over time (captured by the quarter dummies), there was also a regime shift in pricing that "kicked-in" during the validation period, a shift that the quarter dummies do not capture. To capture that regime shift of nearly 10% in price margins, I 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's predictions in the validation period for the calculation of profit counterfactuals .

5.2 The Demand Model

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 [11]. I observe multiple transactions per client with both accept and reject outcomes for each client. I estimate a logistic regression model for the client's probability to accept the quote. The control function approach is used to account for endogeneity in the pricing decision (Petrin and Train, 2010).

Following Zhang et al. (2014) who used data from the same retailer to model targeted pricing, I allow reference price to affect the client's decision. Reference price is calculated as the difference between the current price and the average price the client received in the last three quote requests in the category [12]. If the current price is higher than the reference

[11] 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 I handle each of those quotes as two quotes: one accepted, and one rejected.

[12] I 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 results.

price, the difference will be counted as loss; if the current price is lower, the difference will be counted as gain. For every price quote either gain or loss is greater than zero and the other is equal to zero. I calculate reference price by product category, because pricing can vary substantially across categories and to account for different purchase cycles for different product categories. Clients typically order the same product categories in most or all of their quotes. For the control function I use cost, cut and quarter fixed effect as 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.

5.3 Demand Specification

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 to capture individual client price targeting, $cost_q$ 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 utility for client i from quote q is:

$$u_{qi} = \beta_{1i} + \beta_{2i}\ gain_{qi} + \beta_{3i}\ loss_{qi} + \boldsymbol{\beta_z}\ \boldsymbol{z_{qi}} + \gamma\ \Delta P_{qi} + \sigma\ \eta_{qi} + \xi_{2qi}, \tag{10}$$

where

$$gain_{qi} = \begin{cases} ref\_price_{qi} - price_{qi} & \text{if } price_{qi} < ref\_price_{qi} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

$$loss_{qi} = \begin{cases} price_{qi} - ref\_price_{qi} & \text{if } price_{qi} > ref\_price_{qi} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

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 average price for the product category in the last three quote requests. Because a quote may include a request for more than one category, in calculating reference price I first calculate category-based reference price, i.e., the average of the price for the product category in the last three times it was requested, 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. 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 data (where the dummy for every category that was requested in quote j is equal to one). The last two terms prior to the random shock ξ_{2qi} in Equation 10 reflect the specification of the control function approach (Petrin and Train, 2010): $\Delta P_{qi} = p_{qi} - \tilde{p_{qi}}$, 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. Assuming that ξ_{2qi} is extreme value distributed, the probability that client i will accept quote q follows the binary logit specification:

$$Pr_{qi} = \frac{e^{u_{qi}}}{1 + e^{u_{qi}}} \tag{13}$$

5.3.1 Demand Estimation and Results

To estimate the demand model with the pricing control function, I first estimate a random effects model for the control function pricing equation and use the residuals from the control function (ΔP_{qi} in Equation 10) to estimate the demand controlling for possible price endogeneity. I then use Bayesian inference with MCMC sampling to estimate the demand quote acceptance model. I 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 7 and 8, respectively). Higher cost and cut requirements increase the price as expected. 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. Price has a significant effect on acceptance probability (gain seems to have a negative sign as well, but significantly lower in magnitude than that of loss). 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 7: 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	

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

Table 8: Parameter Estimates for Client's Acceptance Decision

Parameter	Mean	Mean SE	Std. dev.	$Q_{2.5}$	$Q_{97.5}$	N_{eff}	Rhat
Intercept	1.299	0.006	0.202	0.901	1.699	1,000.000	1.000
Gain	-0.072	0.001	0.019	-0.106	-0.035	1,000.000	0.999
Loss	-0.136	0.001	0.017	-0.171	-0.103	1,000.000	1.004
Recency	-0.001	0.000	0.000	-0.001	-0.001	1,000.000	1.001
Weight (log)	-0.299	0.000	0.013	-0.327	-0.274	1,000.000	0.999
LME	0.226	0.008	0.248	-0.293	0.704	1,000.000	1.002
LME volatility	0.023	0.001	0.037	-0.049	0.096	1,000.000	0.999
Regular salesperson	0.566	0.002	0.061	0.440	0.682	1,000.000	0.999
Aluminum - Cold Finish	-0.024	0.004	0.090	-0.202	0.152	647.696	0.999
Aluminum - Plates, Aerospace	0.059	0.003	0.086	-0.114	0.236	761.128	0.999
Aluminum - Plates, Commercial	0.327	0.003	0.066	0.203	0.457	509.584	0.999
Aluminum - Round, Flat, Square Solids	0.283	0.003	0.060	0.164	0.401	564.376	1.000
Aluminum - Shapes and Hollows	0.063	0.465	0.718	621.972	0.999		
Aluminum - Sheets, Aerospace	-0.782	0.008	0.249	-1.287	-0.301	1,000.000	0.999
Aluminum - Sheets, Commercial	0.386	0.003	0.073	0.237	0.525	592.980	0.999
Other Metals	0.856	0.005	0.143	0.557	1.126	1,000.000	1.001
Aluminum - Sheets, Commercial	0.473	0.013	0.405	-0.335	1.284	1,000.000	0.999
γ	-0.060	0.000	0.015	-0.089	-0.031	1,000.000	1.000
σ	0.020	0.000	0.011	0.001	0.043	1,000.000	1.000

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

5.4 Profits of Model Pricing Vs. Profits of Salesperson Pricing

So far I have created an individual linear model of each salesperson in my sample calibrated on the first 18 months of quotes priced by the salesperson to predict prices (margins) for quotes in the hold-out sample . I further used the calibration data to estimate a demand model that predicts the client's acceptance behavior as a function of different pricing schemes. Now, using the demand estimates, I can calculate quote acceptance and hence expected profits based on original (observed) prices offered to clients by the salesperson (following Equation 7). Using the bootstrap model in Equation 2 (and converting margins to prices) to calculate the model's prices for each of the hold-out sample quotes I can calculate the expected profits based on the model-of-the-salesperson predicted prices (Equation 8), and compare the two expected profits.

To calculate the counterfactuals I use the hold-out sample of six months, 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 expected acceptance probability based on the original pricing scheme is 61.1% and that based on the model's pricing scheme is 61.8%.

Using Equations 7 and 8 and aggregating across quotes, I find that the model's pricing scheme generates expected profits that are 5.3% higher than those of the salespeople's pricing scheme: $\Pi[p] = \$2,438,442$ compared to $\Pi[\hat{p}] = \$2,566,329$. 95% posterior confidence intervals (PCI) of the difference across a sample of 100 draws from the MCMC algorithm output do not contain zero. See Table A11 in the Appendix for a full list of posterior expected profits. Real profits for the 6,894 quotes converted to sales were \$2,345,479.

To a large extent, the B2B salesperson's work is based on her soft skills of 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. The model-of-the-salesperson can learn the salesperson's pricing policy and reapply it to new quotes under the limitations of information availability. The model makes the pricing decision purely based on data inputted to it. The model has only knowledge of case-based circumstances that were coded and fed to it, whereas much of the communication aspects in the sales process that may have led the salesperson to alter her prices are often missing from the model. In the behavioral literature, judgmental bootstrapping of the expert was found to perform better than the expert in a wide range of applications, particularly when the prediction environment is unspecified and thus cues lead to inconsistencies in human behavior (Camerer, 1981; Dawes et al., 1989). Moreover, in many applications that compared judges and their bootstrap version, by construction the judge and the model made a prediction based on the same data (e.g., Wiggins and Kolen, 1971). The performance of judgmental bootstrapping has been rarely tested in settings where the judge has access to much richer information than her model, information that can arguably lead to superior decision making on the expert's end. Therefore, it is impressive that in the information-intensive B2B sales environment, a linear judgmental bootstrap model performs better than the expert salesperson, despite the loss of private information exchanged through inter-personal communication with clients.

6 The Hybrid Approach

The B2B environment is highly communication- and relationship-based, and 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 expert 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 and when not to use it, led to low compliance rates, which possibly limited the treatment effect. On the other hand, it is possible that salespeople decided when to comply with the model intelligibly, using their own prices when they perceive that model was missing information and the model's price when its pricing captured the situation well.

How can I identify those cases where the model is doing better from the cases where the salesperson is doing better? my modeling approach can be used to separate the two cases and decide whether the model or the salesperson should price the quote. Because the model created for each salesperson is in fact an automated representation of the salesperson herself, I expect the model to reflect the salesperson's pricing policy, and can assume that if the salesperson deviates from her regular pricing (as predicted by the model), she does so in the presence of meaningful case-based information. I will therefore look at the distance between observed pricing and predicted pricing (as measured by margins) for every pricing decision, and instead of letting the model price always, defer to the salesperson's price when the difference between the salesperson's price and her model's price is relatively large, assuming the large deviation resulted from private information. If the deviation is small, I will account for it as noise and use the model's price.

6.1 Structuring the Hybrid

To structure those hybrid pricing schemes, for each salesperson separately I calculate the standard deviation of the distribution of the differences between observed log margin and predicted log margin (the calculation is done separately for each salesperson based on her own quotes)[13]. I 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 or the salesperson based on how well the model does in predicting

[13]To capture deviations most accurately, I work at the log margin level, the DV of the model-of-the-salesperson.

the input to profitability, namely prices.

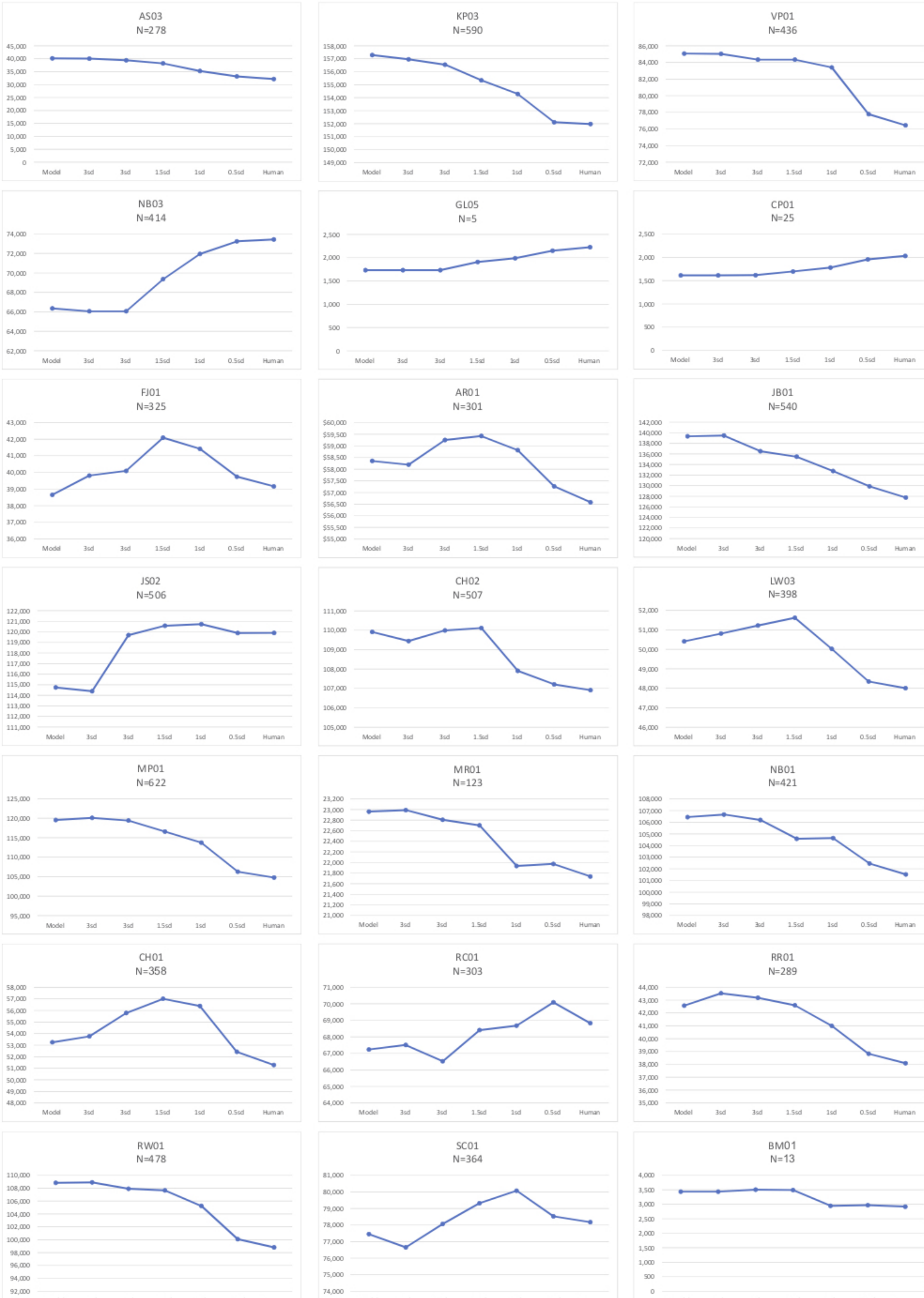
I then calculate expected acceptance probability and expected profits for all the quotes in the hold-out sample, based on the new policy. I create five hybrid pricing schemes, defined by the threshold of deferring to the salesperson: $x = 3$ sd, 2 sd, 1.5 sd, 1 sd or 0.5 sd. Note, that the higher the 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. That is, when the threshold is high, I let the model price more quotes for which the salesperson and the model diverged, than when the threshold is low.

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, that in the experiment too, every salesperson chose her own hybrid based on her own experience and model recommendation.

For the task of deciding which hybrid structure to assign to each salesperson, and for that task only, I re-estimate both the pricing model and the demand model, but this time only on the first 5 quarter of what was previously used for calibration and leave the 6 quarter for prediction. That is, I predict prices and acceptance rates on q2 of 2016 and calculate profit counterfactuals for the seven pricing schemes (full model pricing, full human pricing and five hybrid schemes) for each salesperson. The optimal hybrid structure for each salesperson as estimated in the validation period of q2 of 2016 is then used for choosing each salesperson's hybrid in the original validation period (2016 q3-q4). Note, that because the hybrid structure was estimated based on a different sample, it is not necessarily the most profitable hybrid in the validation.

Figure 3 shows the hybrid structures calculated based on profit counterfactuals in quarter 2 of 2016. For the three salespeople in the top row it is best to completely replace them with their own model; for the three salespeople in the second row it is best to let them

Figure 3: Expected Profits of Pricing Schemes by Salesperson



price all quotes by themselves; and for all other salespeople in rows 3-7 there is an optimal combination of the salesperson and her model that generates the highest profits. For example, for salesperson coded as "CH01" (sixth row, left) the optimal hybrid is the one where she prices about 15% of the quotes and the model prices about 85% of the quotes, based on 1.5 standard deviations of her own distribution of deviations from model predictions.

6.2 Profits of the Hybrid Pricing Scheme

Expected profits in the validation period for the hybrid scheme integrated over all the salespeople, each using her optimal hybrid structure, are 1.5% higher than those of the model, $\Pi[p_{\hat{hybrid}}] = 2,603,719$, $\Pi[\hat{p}] = \$2,566,329$ (95% PCI of the difference across posterior draws do not contain zero, see Table A11 in the Appendix). I find that 82.5% of the quotes are directed to model's pricing in the hybrid pricing scheme, while salespeople price the remaining quotes. Overall, the hybrid scheme generates profits that are 6.8% higher than those of the salespeople themselves, a total increase in profits of more than $165,000 for 11,621 quotes in the hold-out sample.

The fact that the hybrid pricing scheme generates higher profits than either pure automation or the salespeople, supports my conjecture that in some pricing decisions the model's consistency 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. In those cases, in which the trade-off between the information provided by environmental cues and the possible bias caused by the human decision making tilts towards the former, it is then best to let the salesperson make the pricing decision. My 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, I find that due to the mixed nature of their work, that combines rule-based decisions with judgments based on communication and interpersonal interactions, pricing decisions are best automated while leaving the expert in the process and allowing her to price "special" cases.

Three factors may drive the difference between the treatment effect size in the experiment (10%) and the somewhat lower increase in profits in the counterfactual analysis (7%): first, in the counterfactual analysis I use a demand model to estimate profits, and the logit scale parameter tempers the effect of increased profitability; second, in the counterfactual analysis, I use a decision rule to allocate quotes to either the salesperson or her model, possibly missing opportunities for the model to gain higher profits (e.g. in cases that the deviation of the person from the model was large, yet the model's profits were higher); third, while preparing the CRM system to the experiment and integrating the price engine into the work flow of the salespeople, the IT team of the company was able to add ad-hoc improvement to the data collection process, leading to higher-quality input to the model during the experiment.

6.3 Understanding the Hybrid Structure

It is of interest to understand what underlies the assignment of a hybrid structure to a salesperson. I coded the seven hybrid structures as a continuous variable representing the share of the model in the hybrid: the variable can get the values: 100, 99.7, 95.5, 86.6, 68.3, 38.3 or 0, corresponding to full model pricing, 3sd-, 2sd-, 1.5sd-, 1sd-, 0.5sd- hybrid, and full human pricing. I first asked the CEO of the company to classify the level of expertise of each salesperson in the company. Based on his rating for 18 of the 21 salespeople in the data, I divide the salespeople into two groups: lower expertise (N=10) and higher expertise (N=8) salespeople. A one-way ANOVA revealed no significant difference by salesperson rating as determined by the company's CEO on the hybrid structure, i.e., what percentage of quotes should be priced by the model and what percentage by the salesperson ($F(1,16) = 0.66$, $p = 0.427$).

I then ran a linear regression of model-percentage on salesperson characteristics calculated based on the calibration period (in the regression analysis all salespeople are included). For each salesperson s I regress her hybrid structure H_s on the set of salesperson characteristics, x_s^H that includes her tenure in days, number of clients she worked with, average number of pricing decisions she makes per week, and her average quote characteristics (weight, cost and ratio of processing):

$$H_s \sim \boldsymbol{\rho^H x_s^H} + \epsilon_s^H, \tag{14}$$

where ϵ_s^H is a normally distributed random shock. Table 9 shows that the model's share in the hybrid increases with the number of clients the salesperson worked with, suggesting rich history for the model to learn from. Similarly, experience with quotes that require processing increases the share of the model in the hybrid. On the other hand, when salespeople typically handle complex quotes with high cost, the model's share in the hybrid decreases. Finally, it seems that the more productive the salesperson is (number of pricing decisions per week), the less she is being replaced by her model in making pricing decisions, possibly because productivity is correlated with being a high expertise salesperson.

Table 9: Hybrid Structure by Salesperson Characteristics

Linear Regression of Model's Share in the Hybrid

Variable	**Coefficient**	**Std. Dev**
Tenure (in days, log)	17.72	(8.524)
Number of unique clients	0.214**	(0.065)
Avg. weight	-0.196	(0.109)
Ratio of lines w/processing	151.1*	(65.744)
Avg. cost per lb.	-102.7*	(37.703)
Avg. pricing decisions per week	-2.01**	(0.629)
Constant	132.4	(87.824)
Observations	21	
R^2	71%	

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

6.4 When Should the Salesperson Price?

After establishing that the hybrid pricing scheme is superior to that of the person or the model, I wish to understand under what conditions the model should price an incoming quote, and under what conditions the person should do it. I regress the difference (mean centered) between the model's expected profit and the salesperson's expected profit on a set of quote and client characteristics, salesperson fixed effects and client random effects [14]. By construction of the DV, when a regression coefficient is positive, the model is generating higher expected profits relatively to the person, and vice versa. Specifically, For each quote q by client i I regress the difference in profits Δ^p_{qi} on the set of quote and client characteristics, x^{Δ}_{qi} that includes weight, cost, market price at the time of quoting, ratio of processing, RFM for the product category, a dummy for each category in the quote and the client's priority, as well as salesperson dummy, I_{person}:

$$\Delta^p_{qi} \sim \boldsymbol{\rho^{\Delta} x^{\Delta}_{qi}} + \rho^I I_{person} + \epsilon^{\Delta}_{qi}, \tag{15}$$

where ϵ^{Δ}_{qi} is a normally distributed random shock.

Table 11 shows the regression results for the difference in profits on the quote and client characteristics. For large quotes (weight), when the quote has multiple lines, or when special processing is required (cut ratio), it is better to refer the quote to the expert salesperson.

These findings support the rational behind the hybrid approach - when a quote is complex or is out of the ordinary in its specifications - follow the human expert's judgment in pricing. These are exactly the cases that the model is not able to address, because in the lack of history to learn from, it does not know how to respond to them **??**. It is important to note, that this analysis allows me to identify extreme, or "broken leg", cases based on observable covariates only. When the model is missing information, I can only identify the gap by the difference between model and salesperson margin, as was done in Section 6 above.

[14] Because the demand model was estimated at the quote level, I conduct this analysis at the quote level as well.

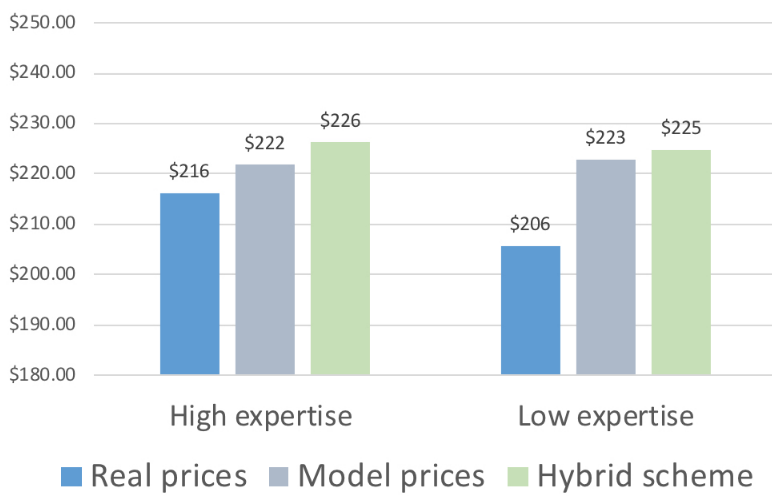
Table 11: Profits Difference Regression Results

Variable	Coefficient	Std. Dev
Weighted cost per lb.	2.102	(1.612)
LME per lb.	-18.77	(51.010)
LME volatility	0.0568	(3.543)
Quote weight	-0.00522**	(0.002)
Lines per quote	-4.260***	(1.102)
Recency†	-0.769	(0.751)
Frequency†	-1.022	(5.264)
Monetary†	1.320	(1.099)
FT base ratio	-4.481	(7.693)
Cut ratio	-9.842*	(4.182)
Aluminum - Cold Finish	5.566	(9.676)
Aluminum - Plates, Aerospace	16.17	(8.992)
Aluminum - Plates, Commercial	1.269	(6.891)
Aluminum - Round, Flat, Square Solids	-17.33**	(6.405)
Aluminum - Shapes and Hollows	-7.558	(6.594)
Aluminum - Sheets, Aerospace	-39.64	(20.936)
Aluminum - Sheets, Commercial	-0.628	(7.179)
Other Metals	5.038	(13.253)
Stainless - Other Stainless	95.29**	(33.148)
Client Priority	1.544	(1.677)
Constant	-4.263	(41.121)
Observations	11,621	
Clients	2,152	

†Quote average
Controlling for salesperson fixed effect
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

I next analyze the model's performance by salesperson's expertise. 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 (all differences are "significant" based on the MCMC draws, that is 95% PCI do not include zero). First, note that consistent with the CEO's classification, the high expertise salespeople generate higher expected profits than the low expertise salespeople. Second, the model-of-the-salesperson improvement over the salesperson is much higher for the lower expertise people than for the high expertise people. This may suggest that the higher expertise salespeople take advantage of private information in the environment more efficiently, and when replaced completely by their model a significant share of private information is lost. The hybrid approach increases the average profit per quote twice for high-expertise salespeople than for low-expertise salespeople, bringing back the lost information and again indicating their better skills in utilizing private information.

Figure 4: Expected profits by Salesperson Expertise



7 Alternative Hybrids

I now turn to explore alternative hybrid structures. First, instead of estimating an individual model for each salesperson, I estimate a single aggregate model for all the salespeople in the company and calculate expected profits for it. The aggregate model's expected profits are similar to those of the hybrid model $\Pi[\hat{p_{agg}}] = 2,606,904$. The hybrid model generated profits by building on the expert's expertise twice: first, by learning from her experience and applying her policy consistently to new quotes, second by choosing to use her when private information possibly existed. The aggregate model is generating higher profits than the salespeople as well as their models by building on the common knowledge of the salespeople as a group (Armstrong, 2001).

7.1 Expertise-based Hybrids

The other extreme of using all salespeople to generate price predictions would be to take a single salesperson and use her model to predict prices for all the quotes in the validation period. What would be a rational way to choose that salesperson? I turn again to the CEO's evaluation and choose two experienced salespeople that he classified as high-expertise salespeople. Both of them have been with the company for over ten years and have a high number of quotes in the calibration (for model training purposes). When applying those salespeople's individual models to all the quotes, both expected profits are lower than even the individual models' profits: $\Pi[\hat{p_{JS02}}] = 2,316,729$ and $\Pi[\hat{p_{RC01}}] = 2,105,641$, compared to $\Pi[\hat{p}] = 2,566,329$. This result indicates, that expertise is much narrower than one would imagine, and a single expertise does not fit all requirements. It is interesting to mention in this context the comment of salespeople (incidentally, one of the two modeled here) during the course of the experiment. She commented that the price recommendation tool could be useful on days when she is out of the office to help her colleagues price quotes for her clients. When a salesperson is absent, her expertise is gone with her, and another salesperson, even with his own set of expertise, cannot immediately replace her. Automating the salesperson by bootstrapping her is therefore a way to capture, code and automatically apply organizational knowledge held by members of the organization.

While using a single high-expertise salesperson to price all quotes was not a successful route, maybe aggregating all the high expertise salespeople will lead to better results. To test this, I ran an aggregate model of only the high expertise salespeople as identified by the CEO of the company. Indeed, the use of aggregation of the better part of the crowd led to the best results so far, with expected profits 12.5% higher than those based on the prices of the salespeople: $\Pi[\hat{p_{high}}] = 2,744,454$. This result shows, that not only the method of aggregation is important, but also the quality of data used - the model is as good as the data inputted to it. Interestingly, when applying each of the high-expertise salespeople models to price all quotes and averaging those predictions, the resulting expected profits are not even close to the aggregate model based on the same group of salespeople ($\Pi[\bar{p_{high}}] = 2,271,161$), leading to the conclusion that in this empirical context aggregation of predictions is inferior to predictions generated on aggregated data.

7.2 Experience-based Hybrid

Finally, I create a hybrid that combines all the elements discussed above: it averages predictions based on the individual expertise (measured by experience) of the salesperson with the product category priced. For every salesperson and product category I calculated the share of quotes that the salesperson priced in the category out of all the quotes in this product category in the data. The individual shares sum up to 1 in each category. I then use every salesperson's model to predict all the quotes in the validation period, but instead of using only one prediction for a quote, I create a weighted average of the salespeople's prices for the quote, that depends on the history of the salesperson in pricing that category. If the salesperson priced a large share of the quotes in the category in the past, her prediction will be weighted accordingly in calculating the price for the quote. Expected profits for this hybrid ($\Pi[p_{cat\hat{e}gory}] = 2,362,292$) do not exceed prior profits by aggregate models or even the individual models.

8 Alternative Pricing Model Specifications

The approach I 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. I 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, 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, I 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 creating 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 I examine the ability of the random effect linear model described in section 3 to capture the salesperson's pricing decisions relative 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). Similarly to the linear regression model, I estimate an individual model separately for each salesperson in the counterfactuals data. Similar to the pricing model described in section 3, for each one of the models I use log-margin as the dependent variable and the same set of variable described in Section 3.2 as predictors. One exception is that because ML methods do not include random effects, I included instead as an additional variable the average log margin per client, as a proxy to client individual effect.

For the implementation of all three ML models I used Python's scikit-learn software (Pedregosa et. al., 2011). To fit each model, I used cross validation on the calibration data to fit hyper-parameters of the model. Specifically, for the Lasso and Ridge I used cross validation to estimate the tuning parameter alpha. For the RF, I 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. I allow the range of the randomized search to vary based on the number of quotes priced by each salesperson (the sample size for each salesperson). Table A10 in the Appendix shows the parameters for which a randomized search was conducted and the the set of parameters that yielded the best score for each salesperson.

I calibrate the three additional 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 model, the Lasso, Ridge and the RF, I 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 (see Equation 16).

$$RMSE(y_i, \hat{y}_i) = \sqrt{\frac{1}{n}\Sigma_{i=0}^{n-1}\Big(y_i - \hat{y}_i\Big)^2} \tag{16}$$

Table 12 shows the RMSE scores for each model for the 21 salespeople in my data as well as simple and weighted (by number of quotes per salesperson) average RMSE scores per model[15]. For every model I report the in- and out- of sample RSME score. First, note that the two linear ML models, the Lasso and Ridge, are comparable in performance to the simple linear model. Second, the random forest perform better than the linear model in predicting prices, however, in profit simulations the linear model leads to higher profitability than that of the random forest ($\Pi[\hat{p_{RF}}] = 2,204,991$, over 10% below the model's prices).

One reason for the superior performance of the simple linear model could be that the RF seems to predict significantly lower prices than the linear model, possibly affecting profits directly (because the increased simulated acceptance does not make up for the lower overall profit). While I find that the simple random effect linear model outperforms the ML models in my application, I encourage future research to explore the ML approach for automation of decision making as the inferiority ML models may be specific to my application.

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

Table 12: Comparison of Models - Fit and Prediction RMSE

	Saleperson	N Train	N Test	Lasso in	Lasso out	Ridge in	Ridge out	RF [16] in	RF out	Linear[17] 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			0.619	0.604	0.612	0.598	0.473	0.557	0.549	0.587
	Weighted average			0.575	0.550	0.571	0.545	0.411	0.486	0.527	0.547

[16] Random Forest

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

9 Discussion

Using a multi-method approach that combines judgmental-bootstrapping from behavioral science, field experiment in which I embed my pricing model into the CRM system of a B2B retailer, and econometric modeling for counterfactual analysis, I demonstrate that pricing decisions in B2B setting can be automated. Specifically, I demonstrate that because such 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 (83% of cases), but allows the salesperson to price complex or irregular quotes does better than either a fully automated pricing model or the salespeople's pricing.

Between my field experiment and counterfactual analysis, I find an improvement of 7-10% in profitability due to automation. My research bridges between the behavioral judgment literature and marketing science literature by building a pricing judgmental bootstrapping model (Dawes, 1979), and demonstrating using both filed experiment and econometric modeling how such a model could be applied in real-world settings to address a major business problem. Moreover, my research bridges theory and practice, by demonstrating via a pricing field experiment how automation can improve the profitability of a B2B retailer. Indeed, following my experiment, the B2B retailer I collaborated with is adding my pricing model to the CRM system to provide price recommendations to salespeople as input to their pricing decisions for all incoming quotes.

My empirical analysis shows that for the B2B salesperson making pricing decisions, the balance between substitution and complementarity is key to automation. I adopt two levels of automation. First, the automated model is a model of the salesperson. Second, based on case-based information I identify when to let the model price a quote and when to let the person price it. I argue that automation should be used not only to make the pricing judgment in some cases, but also to determine what are those cases (and in contrast, what are the cases that should be delegated to the human expert). In the longer term, and based on my work, the firm I worked with is considering moving 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 and agent" button based on the specific quote. I call for future research to further explore these two degrees of automation.

In my empirical application I find that using judgmental bootstrapping to "teach" the model how to price works better than more advanced ML methods such as RF and regularized regressions. I believe that some of the reasons for the superiority of the simpler and parsimonious model in my case are related to the characteristics of my specific application and that ML should be further explored as a tool to automating decision making.

Using a hybrid automation approach that complements the salesperson with a model of herself could have far-reaching implications for preserving organizational knowledge in a market environment characterized by high sales force turnover rates[18]. Salespeople develop expertise and familiarity not only with the product they sell, but also with the clients they regularly work with (as shown by the analysis in Section 7). 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 company. Conversations with salespeople in the company echo the benefits of the approach. For example, one salesperson said: "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.

My analysis explored the potential of automation in B2B salesforce pricing decisions using secondary data and field experiment with a metal B2B retailer. Future research should explore the generalizability of these findings to other B2B retail domains, and to other

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

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 what is the likelihood of "broken leg" cases. My automation approach can flexibly accommodate different levels of automation that are appropriate for each domain.

One limitation of experimental settings is that the relatively low compliance of the salesforce with the tool possibly underestimates the potential effect of automation. In a larger context, 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. 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" behavior. The issue of compliance may further highlight the benefits of automating not only the pricing decision but also the decision of when to direct execution to the expert and when to force automation.

In summary, my research provides first empirical evidence to the potential of automating the human intensive work of B2B salesforce. It suggests that while the B2B salesperson was traditionally perceived as indispensable, some salespeople, and to some extent of the daily work all salespeople, 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. I 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.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of forecasting*, pages 417–439. Springer.

Asare, A. K., Brashear-Alejandro, T. G., and Kang, J. (2016). B2b technology adoption in customer driven supply chains. *Journal of Business & Industrial Marketing*, 31(1):1–12.

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

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

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

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

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. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, 27(3):411–422.

Camerer, C. F. and Johnson, E. J. (1997). The process-performance paradox in expert judgment: How can experts know so much and predict so badly. *Research on judgment and decision making: Currents, connections, and controversies*, 342.

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

Cowgill, B. (2017). Automating jedgment and decisionmaking: Theory and evidence from resume screening.

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. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American psychologist*, 26(2):180.

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.

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

Ditzler, G., Roveri, M., Alippi, C., and Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25.

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

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

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

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

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

Frey, C. B. and Osborne, M. (2013). The future of employment. *How susceptible are jobs to computerisation.*

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.

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.

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

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

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

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.

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.

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

Lughofer, E. and Sayed-Mouchaweh, M. (2015). Adaptive and on-line learning in non-stationary environments.

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*, pages 20–38.

Nedelkoska, L. and Quintini, G. (2018). Automation, skills use and training. (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.

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.

Zhang, J. Z., Netzer, O., and Ansari, A. (2014). Dynamic targeted pricing in b2b relationships. *Marketing Science*, 33(3):317–337.

Appendices

Table A1: Log Margin by Quarter

	Mean
2016q1	-0.51
2016q2	-0.44
2016q3	-0.40
2016q4	-0.33
2017q1	-0.29
2017q2	-0.25
Average	-0.38

Table A2: 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 A3: Average of Individual Models Estimates

	Mean	Std. dev.	Lower 10%	Median	Upper 90%
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					

Figure A1: Screenshot of the CRM System

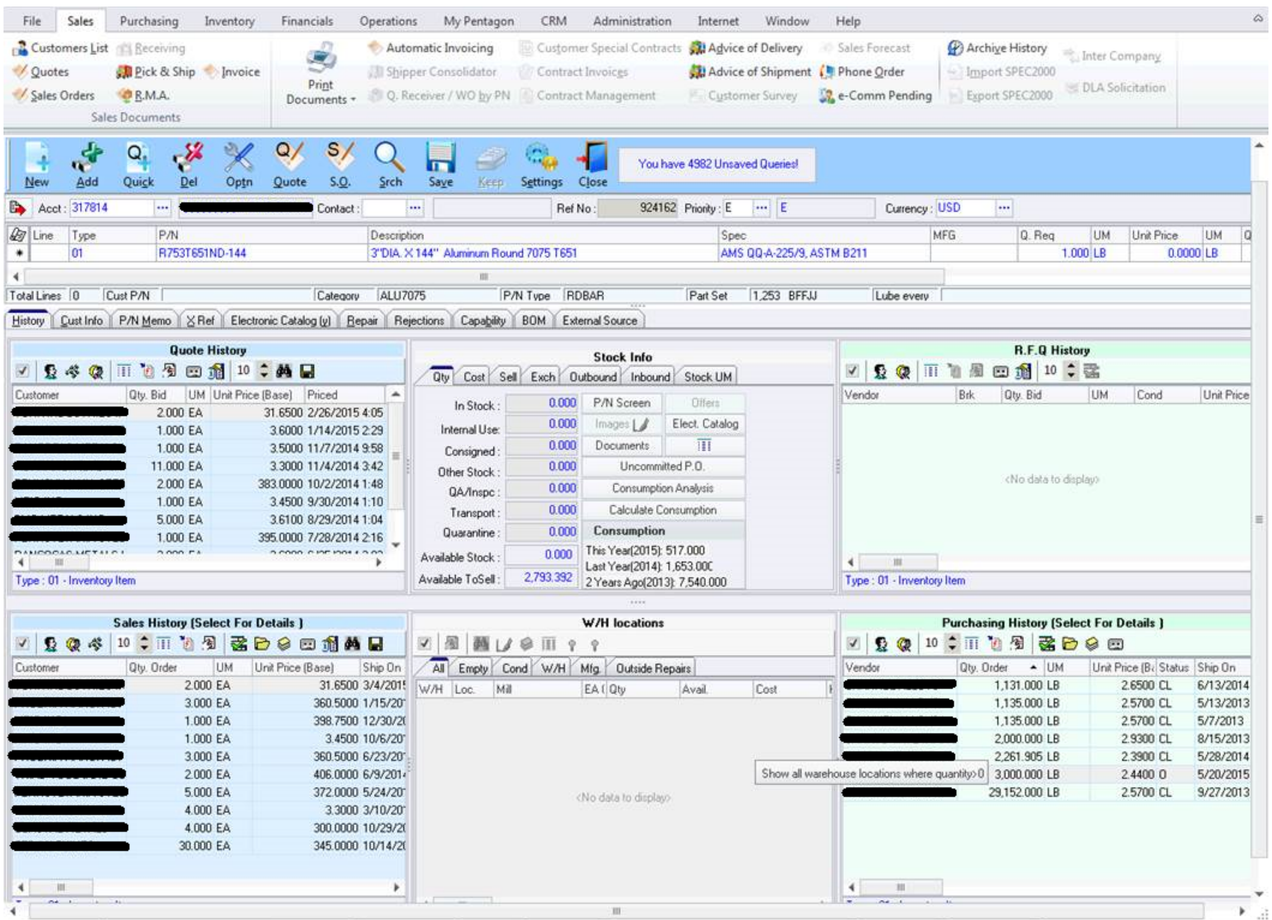


Figure A2: Flow of Field Experiment

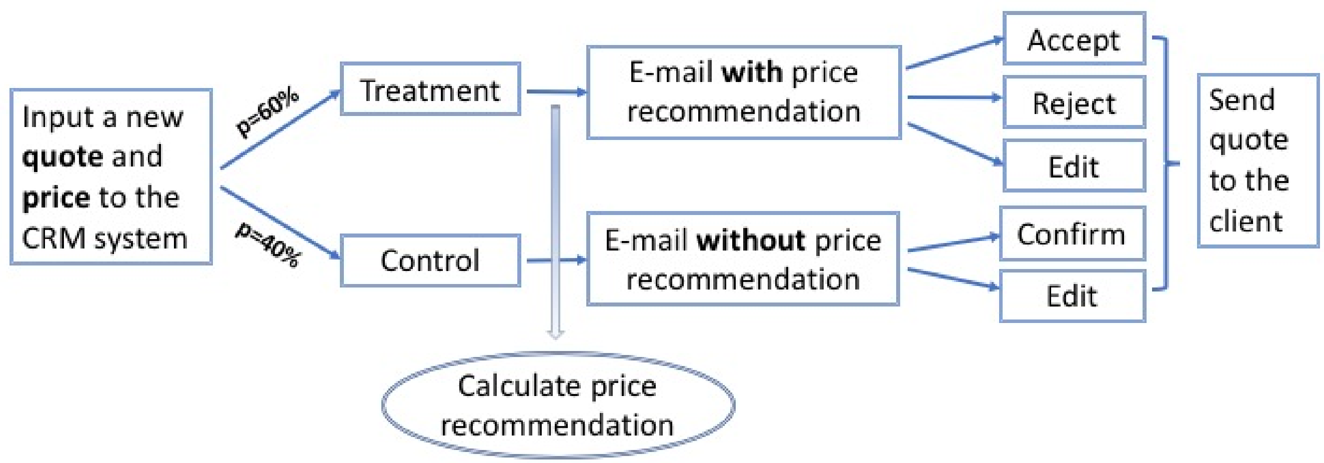


Figure A3: Treatment Email Format

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

Figure A4: 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

Figure A5: Treatment Edit Form

Figure A6: Control Edit Form

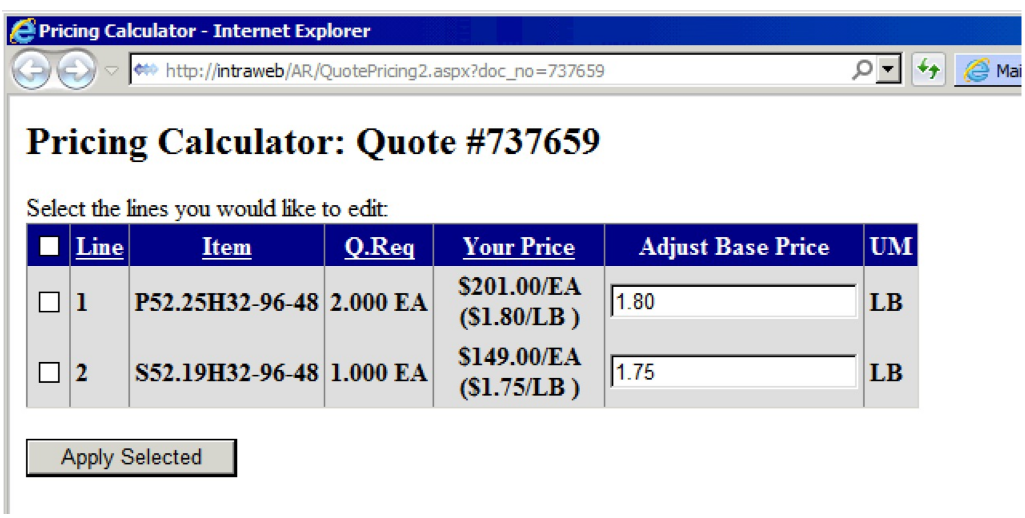


Table A4: Summary Statistics per Line in Counterfactuals Data

	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 = 104,487					

§Line margin calculated as specified in Equation 1

†Calculated at the product category level

Table A5: Summary of Category Variable in Counterfactuals Data

	N	Frequency	Cum. freq.
AluminumColdFinish	3,788	3.63	3.63
AluminumPlatesAerospace	4,921	4.71	8.34
AluminumPlatesCommercial	22,673	21.70	30.03
AluminumRoundFlatSquareSolids	31,741	30.38	60.41
AluminumShapesandHollows	28,001	26.80	87.21
AluminumSheetsAerospace	246	0.24	87.45
AluminumSheetsCommercial	11,603	11.10	98.55
OtherMetals	1,421	1.36	99.91
StainlessOtherStainless	93	0.09	100.00
Total	104,487	100.00	

Table A6: Summary of Quarter Variable in Counterfactuals Data

	N	Frequency	Cum. freq.
2015q1	15,156	14.51	14.51
2015q2	16,894	16.17	30.67
2015q3	15,920	15.24	45.91
2015q4	15,580	14.91	60.82
2016q1	19,998	19.14	79.96
2016q2	20,939	20.04	100.00
Total	104,487	100.00	

Table A7: Summary of Client Priority Variable in Counterfactuals Data

	N	Frequency	Cum. freq.
A	43	1.14	1.14
B	150	3.96	5.10
C	643	16.97	22.07
D	240	6.34	28.41
E	2,598	68.59	96.99
P	114	3.01	100.00
Total clients	3,788	100.00	

Table A8: Log Margin by Quarter

	Mean
2015q1	-0.77
2015q2	-0.75
2015q3	-0.75
2015q4	-0.80
2016q1	-0.49
2016q2	-0.41
2016q3	-0.39
2016q4	-0.35
Total	-0.57

Table A9: Log Margin Regression Results - Counterfactuals

Variable	**Coefficient**	**Std. err.**
Cost per lb.	-0.136***	(0.003)
Market price per lb. (LME)	0.561***	(0.081)
Volatility	-0.012**	(0.006)
Weight (log)	-0.386***	(0.002)
Relative Weight	0.435***	(0.006)
Cut/weight	2.416***	(0.046)
Recency	0.001	(0.001)
Frequency	-0.052***	(0.007)
Monetary	-0.0004	(0.002)
Regular salesperson	-0.072***	(0.011)
FT-base	0.017	(0.012)
2015 q1	0	(.)
2015 q2	0.013*	(0.007)
2015 q3	0.063***	(0.010)
2015 q4	0.065***	(0.013)
2016 q1	0.422***	(0.013)
2016 q2	0.491***	(0.011)
Aluminum Cold Finish	0	(.)
Aluminum Plates Aerospace	0.024	(0.015)
Aluminum Plates Commercial	0.079***	(0.013)
Aluminum Round Flat Square Solids	-0.078***	(0.012)
Aluminum Shapes and Hollows	0.075***	(0.013)
Aluminum Sheets Aerospace	0.289***	(0.041)
Aluminum Sheets Commercial	0.003	(0.014)
Other Metals	0.285***	(0.022)
Stainless - Other Stainless	0.119*	(0.066)
Priority A	0	(.)
Priority B	0.038	(0.063)
Priority C	0.037	(0.058)
Priority D	0.140**	(0.061)
Priority E	0.213***	(0.056)
Priority P	-0.001	(0.067)
Constant	0.843***	(0.111)
Observations	104,487	
R^2	50.90%	

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

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

Table A10: Random Forest Parameters for Individuals Models

	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	209	auto	314	10	13	33
2	AS03	3,089	TRUE	67	auto	50	15	11	25
3	BM01	376	TRUE	6	auto	9	11	14	14
4	CH01	5,817	TRUE	474	auto	333	11	26	50
5	CH02	7,422	TRUE	43	auto	420	15	33	62
6	CP01	393	TRUE	38	auto	16	11	37	13
7	FJ01	1,309	TRUE	112	auto	95	12	46	21
8	GL05	432	TRUE	18	auto	6	14	34	18
9	JB01	8,727	TRUE	475	auto	510	10	14	35
10	JS02	6,842	TRUE	356	auto	193	10	40	43
11	KP03	8,565	TRUE	432	auto	616	13	22	71
12	LW03	2,927	TRUE	210	auto	127	16	20	50
13	MP01	11,349	TRUE	892	auto	203	12	13	84
14	MR01	1,633	TRUE	127	auto	33	12	10	20
15	NB01	6,567	TRUE	99	auto	161	14	37	78
16	NB03	8,127	TRUE	611	auto	760	11	12	84
17	RC01	5,587	TRUE	221	auto	459	11	20	57
18	RR01	5,007	TRUE	310	auto	344	12	13	57
19	RW01	5,558	TRUE	486	auto	186	11	23	60
20	SC01	3,223	TRUE	54	auto	314	14	27	36
21	VP01	6,091	TRUE	69	auto	459	10	26	40

Table A11: Expected Profits based on the MCMC Draws

MCMC Draw		Expected Profits			
	Real Prices	Model Prices	Hybrid	Model/Real diff	Hybrid/Model diff
1	2,509,275	2,643,120	2,677,890	133,845	34,769
2	2,421,120	2,543,845	2,582,036	122,725	38,191
3	2,281,141	2,399,473	2,433,212	118,332	33,739
4	2,388,444	2,513,429	2,552,308	124,986	38,879
5	2,296,706	2,432,166	2,461,634	135,460	29,468
6	2,366,734	2,495,772	2,526,384	129,038	30,612
7	2,393,633	2,517,745	2,552,488	124,113	34,743
8	2,580,732	2,715,524	2,754,530	134,792	39,006
9	2,337,840	2,459,506	2,494,527	121,667	35,021
10	2,471,102	2,595,786	2,638,711	124,685	42,925
11	2,518,022	2,653,983	2,687,765	135,961	33,782
12	2,423,245	2,543,281	2,583,239	120,036	39,958
13	2,487,309	2,617,121	2,660,673	129,813	43,552
14	2,798,162	2,941,113	2,987,768	142,951	46,655
15	2,711,549	2,853,891	2,894,214	142,341	40,324
16	2,235,357	2,353,683	2,385,909	118,326	32,226
17	2,066,306	2,171,396	2,205,098	105,090	33,702
18	2,462,738	2,590,344	2,625,025	127,606	34,681
19	2,330,748	2,451,255	2,489,540	120,506	38,286
20	2,721,056	2,869,725	2,906,867	148,670	37,142
21	2,443,793	2,566,776	2,608,395	122,983	41,620
22	2,606,458	2,734,557	2,778,992	128,099	44,435
23	2,498,867	2,634,181	2,673,279	135,313	39,099
24	2,319,509	2,440,034	2,478,607	120,525	38,573
25	2,442,089	2,566,818	2,605,428	124,728	38,611
26	2,552,214	2,683,583	2,723,673	131,369	40,089
27	2,191,488	2,304,524	2,340,053	113,036	35,529
28	2,435,344	2,565,166	2,601,215	129,821	36,050
29	2,411,644	2,532,377	2,576,201	120,733	43,824
30	2,342,712	2,464,584	2,498,898	121,871	34,315
31	2,473,824	2,604,951	2,640,670	131,127	35,719
32	2,457,175	2,593,787	2,628,067	136,612	34,280
33	2,292,466	2,411,633	2,451,550	119,168	39,916
34	2,546,359	2,685,348	2,720,465	138,989	35,117
35	2,397,970	2,515,494	2,556,320	117,524	40,826
36	2,805,056	2,962,264	2,999,717	157,208	37,454
37	2,441,081	2,562,624	2,603,810	121,543	41,187
38	2,427,381	2,551,517	2,587,305	124,137	35,788
39	2,367,236	2,498,161	2,528,576	130,925	30,415
40	2,433,506	2,552,469	2,592,461	118,964	39,992
41	2,413,536	2,539,935	2,578,986	126,400	39,050
42	2,304,999	2,424,089	2,461,416	119,090	37,327
43	2,358,414	2,489,704	2,522,848	131,291	33,144
44	2,398,905	2,528,493	2,565,159	129,588	36,666
45	2,250,792	2,366,795	2,403,662	116,003	36,868
46	2,379,403	2,503,996	2,540,783	124,594	36,787
47	2,595,640	2,729,623	2,770,899	133,983	41,277
48	2,465,508	2,588,177	2,627,408	122,669	39,231
49	2,230,852	2,351,338	2,386,810	120,486	35,472
50	2,498,871	2,629,130	2,670,019	130,259	40,889
51	2,273,031	2,386,293	2,425,725	113,263	39,432

Continued...

Draw		Expected Profits			
	Real Prices	Model Prices	Hybrid	Hybrid/Real diff	Hybrid/Model diff
52	2,425,358	2,561,738	2,596,503	136,380	34,766
53	1,996,191	2,099,275	2,133,033	103,084	33,758
54	2,398,169	2,540,052	2,567,838	141,883	27,786
55	2,397,342	2,520,914	2,556,772	123,573	35,858
56	2,382,114	2,510,795	2,542,458	128,681	31,663
57	2,359,820	2,485,143	2,522,315	125,323	37,172
58	2,273,655	2,397,903	2,429,274	124,248	31,371
59	2,616,843	2,756,900	2,791,515	140,058	34,615
60	2,344,650	2,467,472	2,500,813	122,822	33,340
61	2,495,997	2,635,632	2,671,304	139,635	35,672
62	2,472,838	2,601,459	2,639,222	128,621	37,763
63	2,295,593	2,414,137	2,448,109	118,544	33,973
64	2,591,249	2,726,806	2,765,361	135,556	38,556
65	2,572,983	2,704,872	2,745,457	131,889	40,585
66	2,563,634	2,700,010	2,738,020	136,377	38,010
67	2,615,014	2,760,125	2,796,866	145,111	36,741
68	2,065,515	2,173,164	2,205,403	107,649	32,240
69	2,388,137	2,522,137	2,553,050	134,000	30,914
70	2,659,924	2,798,908	2,842,553	138,984	43,644
71	2,751,232	2,897,006	2,941,841	145,775	44,835
72	2,512,991	2,635,010	2,681,093	122,019	46,084
73	2,457,267	2,578,455	2,620,897	121,188	42,442
74	2,612,001	2,759,110	2,796,650	147,110	37,540
75	2,685,755	2,821,403	2,867,774	135,648	46,371
76	2,230,704	2,346,771	2,380,139	116,067	33,368
77	2,398,620	2,524,328	2,562,497	125,709	38,169
78	2,573,896	2,716,858	2,751,616	142,963	34,758
79	2,548,226	2,678,326	2,717,353	130,100	39,027
80	2,619,486	2,761,768	2,801,259	142,282	39,492
81	2,669,537	2,811,184	2,852,854	141,647	41,670
82	2,698,402	2,834,867	2,878,004	136,465	43,137
83	2,241,264	2,355,142	2,390,894	113,878	35,752
84	2,399,072	2,519,091	2,561,338	120,019	42,248
85	2,150,359	2,268,255	2,294,431	117,896	26,176
86	2,422,194	2,544,834	2,580,857	122,640	36,023
87	2,622,885	2,764,747	2,801,249	141,862	36,502
88	2,422,612	2,553,159	2,587,542	130,548	34,383
89	2,279,326	2,406,497	2,436,192	127,171	29,695
90	2,457,870	2,584,613	2,623,781	126,743	39,168
91	2,289,730	2,411,992	2,448,327	122,262	36,336
92	2,193,704	2,307,439	2,342,607	113,734	35,168
93	2,230,659	2,361,282	2,387,622	130,623	26,340
94	2,499,708	2,617,139	2,662,751	117,431	45,612
95	2,425,748	2,548,964	2,590,883	123,216	41,919
96	2,551,292	2,678,931	2,716,663	127,638	37,733
97	2,451,348	2,571,603	2,613,063	120,256	41,460
98	2,667,523	2,803,678	2,847,644	136,156	43,965
99	2,634,217	2,765,999	2,811,476	131,782	45,477
100	2,374,266	2,496,436	2,529,539	122,170	33,103
Average	2,438,443	2,566,329	2,603,719	127,887	37,390