Project-Based Manufacturing: An Approach for Quote Development

by

Dante Edward Montgomery

B.S. Mechanical Engineering, Georgia Institute of Technology, 2013

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering in Partial Fulfillment of the Requirements for the Degrees of

MASTER OF BUSINESS ADMINISTRATION
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology

May 2020

Signature of Author

MIT Sloan School of Management,
MIT Department of Mechanical Engineering
May 8, 2020

Certified by

Hermano Igo Krebs, Thesis Supervisor
Principal Research Scientist and Lecturer of Mechanical Engineering

Certified by

Roy Welsch, Thesis Supervisor
Eastman Kodak Leaders for Global Operations Professor of Management

Accepted by

Nicolas Hadjiconstantinou, Chairman of the Committee on Graduate Students
MIT Department of Mechanical Engineering

Accepted by

Maura Herson, Assistant Dean, MBA Program
MIT Sloan School of Management

This page has been intentionally left blank

Project-Based Manufacturing: An Approach for Quote Development

by
Dante Edward Montgomery
B.S. Mechanical Engineering, Georgia Institute of Technology, 2013

Submitted to the MIT Sloan School of Management and the MIT Department of Mechanical Engineering on May 8, 2020, in partial fulfillment of the requirements for the degrees of

Master of Business Administration
and
Master of Science in Mechanical Engineering

Abstract

Project-based manufacturing presents unique challenges that highly automated and repetitive production lines do not face. Highly automated production lines do not have much variation in the types of products that are made, and these products are usually manufactured at high volume. This makes it relatively easy for the business to know its marginal cost of production and to assist with pricing strategies.

Project-based manufacturers, on the other hand, tend to specialize in manufacturing products that are highly customized. These environments have much more variability and products are usually produced in low volume. Predicting the cost of production in this scenario is difficult and usually requires a dedicated team that focuses on cost estimating and proposal development for competitive bids. Cost estimating and proposal development is a time-consuming and costly overhead activity and inaccurate proposals can have significant impact to business performance. This project focuses on how data analytics can be used to streamline the estimating process for project-based manufacturers resulting in reduced proposal cycle time and improved accuracy and precision metrics.

This pilot project focuses on one product type with a privately owned project-based manufacturer. Although the scope of this project focuses on one product type, each project is unique as it relates to dimensions, surface contour shapes, technical specifications, and other unique features. The methodology for this project analyzes historical technical data and labor durations and implements a lasso regression model to predict the number of labor hours required for a future project with a given set of technical inputs.

While this project focuses on producing a fit-for-purpose solution for estimating one product type for one company, the process and methodology can be applied more holistically. The findings from this research can be applied to accomplish the same objectives for products across an array of project-based manufacturing industry verticals.

Thesis Supervisor: Roy Welsch
Title: Eastman Kodak Leaders for Global Operations Professor of Management

Thesis Supervisor: Hermano Igo Krebs
Title: Principal Research Scientist and Lecturer of Mechanical Engineering

This page has been intentionally left blank

Acknowledgments

I want to thank both of my academic advisors, Hermano Igo Krebs and Roy Welsch, for their guidance and advice throughout this process. Their knowledge and expertise have provided valuable learning experiences and creative problem-solving strategies that I will be able to leverage for future professional experiences.

I also want to thank the teams at both the private equity firm and the portfolio company for providing the autonomy to define and pursue solutions to the challenge presented while providing the resources necessary to achieve the vision. I especially want to recognize my company supervisor for always being available to talk about progress throughout the project, strategize on next steps, and remove any obstacles that came about. This support made the onsite experience quite fulfilling.

Next, I want to thank the LGO staff and LGO students. I have received great support from this community throughout the two-year experience and they were always available to share ideas about how to tackle challenges I was facing. This has given me a glimpse into how we will continue supporting each other for years to come.

Finally, I want to especially thank my parents and my entire family for always pushing me to work hard and for being my biggest supporters through every goal I have set in my life. I am extremely blessed to have the family and friends that I have to continue helping me achieve my most ambitious goals.

This page has been intentionally left blank

Table of Contents

List of Figures

List of Tables

This page has been intentionally left blank

This page has been intentionally left blank

Introduction

A fundamental part of running a manufacturing business is quantifying and understanding total costs and marginal costs of production. Depending on the manufacturing environment, this can be relatively simple or quite difficult. Nevertheless, every reasonable attempt should be made to accurately quantify and categorize production costs to streamline production planning, predict costs for future production, and stay profitable in a competitive industry. It is important to understand why different manufacturing environments make quantifying and predicting costs more difficult than others and how companies can overcome these challenges.

1.1 Problem Statement and Business Case

The focus of this study was a medium-size project-based manufacturer (hereinafter referred to as the "Company") in a complex manufacturing industry. As a project-based manufacturer, the Company experiences many challenges associated with its manufacturing environment. Every day, the Company receives a handful of requests for quotation for various types of products and solutions. Jobs can be very simple and require only a few production areas or they can be highly complex and require support from every production area with intricate levels of testing and quality assurance. There are three key challenges that the Company faces that may improve commercial operations.

The first challenge for the Company is to reduce the cycle time for proposal development in competitive bids. Typically, the commercial team will only forego a bid if they are technically incapable of providing the solution requested. However, this is rare since their customers typically know the full capabilities of the Company. Therefore, the daily variation in quote volume and complexity affects the team's ability to review larger proposals with management.

This time constraint has the potential to lead to overbidding or underbidding if the review process must be rushed to meet proposal submission deadlines.

The second challenge for the Company is to ensure its estimates are consistently accurate. The Company can easily submit proposals by the deadline set by its customers, but if the cost estimates in the proposals are inaccurate, this will significantly hurt business performance. Underestimating proposals hurt profit margins and overestimating proposals can lead to work lost to competitors. Additionally, the Company plans its operations and resource requirements based on the estimates from successful proposals. Ensuring accuracy and confidence in proposals is vital to maintaining and improving success and profit margin.

Finally, the Company needs to incorporate a data feedback loop into its estimating processes. While the Company has a database of historical proposals and an enterprise resource planning (ERP) system with actual duration and cost data from completed projects, this data is only referenced on an as-needed basis for either exact replicas or very similar projects. As it currently stands, there is no systemized method for feeding data back into the estimating process. Figure 1 displays the current and proposed state for the Company's estimating process.

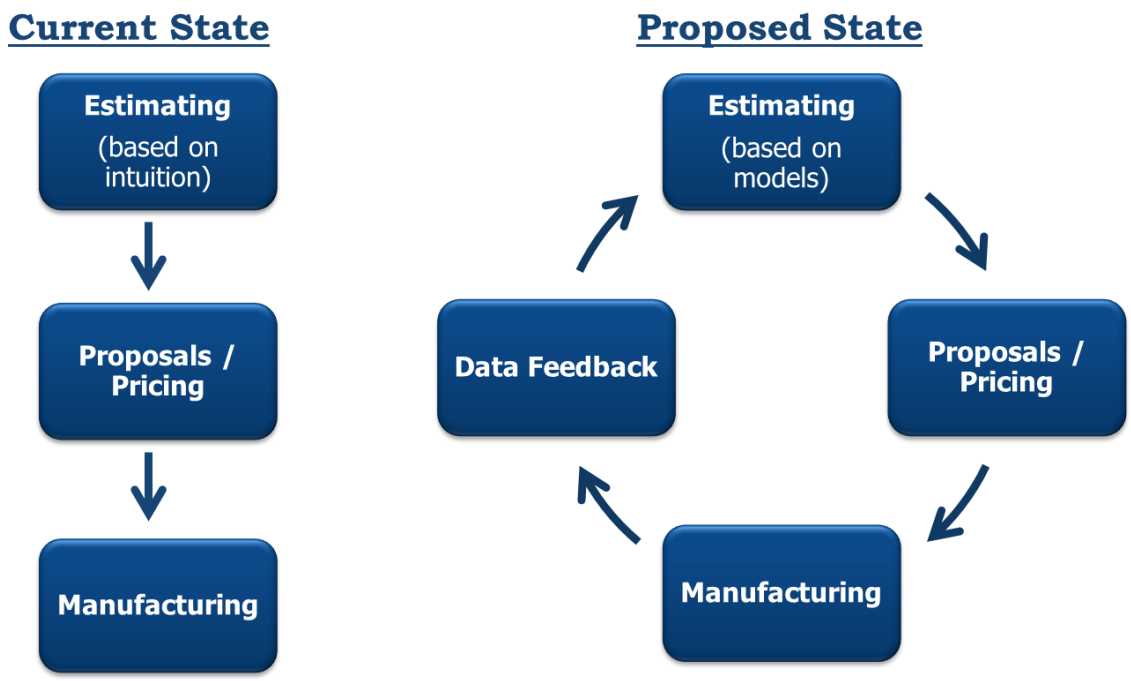

Figure 1. Current and proposed state for data feedback in estimating process

The objective of this project is to address these three challenges by developing a semi-automated estimating model by leveraging data from historical projects. Data was collected from previously completed projects and modeled using regression techniques to predict the time required for future projects. A feedback process was also proposed so that data from ongoing operations would be incorporated into the model and older projects would eventually be removed over time. The final model and estimating process would result in an accelerated proposal development process and improved estimating accuracy.

This page has been intentionally left blank

Background

1.2 Types of Manufacturing Environments

Most manufacturing environments can be broadly placed in three different categories. The first environment can be described as repetitive where a production line produces the same item, or very closely related items, over an extended time period. These items typically have production setups that are rarely modified. The marginal cost of producing each item is generally well-known because there is typically a continuous flow of lots of units through the manufacturing environment for reliable data collection. These environments are also able to utilize more automated machinery given the repetitive nature of the process.

At the other end of the spectrum, a second type of environment is project-based manufacturing, also referred to as a "job shop." This environment is highly variable as compared to a repetitive production environment. Rather than having production lines, a project-based manufacturing facility typically has production areas. These production areas focus on specific tasks or skillsets, such as welding, machining, surface finishing, or quality assurance, but each product coming through these production areas will have unique features or manufacturing requirements. This requires the workforce to plan the requirements for each product as a separate project leading to large amounts of variability in time and cost required to manufacture each product. This work environment tends to be highly labor intensive and cannot use as much automated machinery as repetitive manufacturing environments.

The third manufacturing environment can be described as a hybrid between a highly repetitive production line and a job shop. This manufacturing environment has various discrete setups that allow for equipment to be reconfigured to make a suite of different SKUs that have similar characteristics. The more similarities SKUs have, the more it resembles a repetitive manufacturing environment; conversely, the more dissimilar SKUs are, the more it resembles

project-based manufacturing. Table 1 summarizes these three manufacturing environments with an example of a production sequence that may be observed in each. [1]

Table 1. Types of Manufacturing Environments [1]

Environment Type	Production Sequence
Repetitive	A A A A A A A A A A A A
Hybrid	A B A B A C A A B B D A
Project-Based (Job Shop)	A B C D E F G A H A I J K

Project-based manufacturing brings a unique set of challenges for the business. On the shop floor, there must be a team constantly analyzing work in progress, capacity constraints, and personnel requirements to ensure maximum plant utilization and to allow for adjustments before a production area is over capacity. On the commercial side, these jobs are typically bid competitively where the lowest quote, technically competent bidder wins the jobs. Often, if the true market price for a product is not known, manufacturers will submit proposals for these projects by estimating their own cost of production (based on the design and specifications provided by the customer) and add a profit margin to the estimate (also known as cost-plus pricing). These types of projects make estimating production costs extremely important to the business in order to 1) avoid underbidding more expensive jobs and risk losing money, and 2) avoid overbidding and not being awarded the contract at all. This is a challenge the Company faces on a daily basis.

1.3 Overview of the Private Equity Portfolio Company

The private equity firm involved in sponsoring this project is an operationally-oriented middle market private equity firm focused on buying and improving industrial businesses. The Company is held within the firm's portfolio and they provide manufactured solutions for many

customers in a particular industry vertical. It employs around 1,000 employees across four states in the United States and several international locations with a global customer base. The Company, as it now exists, is the result of a series of mergers and acquisitions of various legacy companies over the course of the private equity firm's ownership. Each legacy company prior to acquisition had a particular industry niche, so these acquisitions allowed the new company to become a fully integrated solutions provider to their customers. They are also continuing to make investments into new capabilities to broaden their value proposition.

This project focuses on estimating the time requirements for manufacturing one particular product type within the Company's portfolio. These products are primarily manufactured at one of two locations: these locations will be referred to as Manufacturing Site 1 (MS-1) and Manufacturing Site 2 (MS-2). MS-1 manufactures a higher quantity of these products; however, MS-2 typically manufactures the projects that are physically larger in size given the larger footprint and larger machinery at MS-2. Although manufacturing for these products take place at both sites, the estimating team for this product is located solely at MS-1.

This page has been intentionally left blank

Data Collection

The strategy for this model was to leverage as much historical data as possible from recently completed projects as the basis for the model. Data was collected from projects that were completed from 2017 through 2019. A total of 77 projects across both manufacturing sites were analyzed for use in the estimating model.

1.4 Cost and Labor Data

Cost and labor duration data from completed projects were collected from two sources within the Company's internal network: SAP and Visual. These are both enterprise resource planning (ERP) systems that are designed to integrate various business processes into a centralized system and provide information across departments in real time. Their objectives are to increase productivity, better manage inventory, promote quality in manufacturing, reduce material cost, and manage human resources, among various other uses. [2] [3]

One important feature of these systems, as it relates to the purpose of this research, is that manufacturing execution for each project can be tracked at a high level of detail. As projects progress through the facility each day, every worker records which project they worked on, what type of work they completed, and how long they spent on each task. For example, at the end of a welder's shift, s/he may document that they spent 5 hours welding for Project A and 5 hours welding for Project B. This data is continuously logged in SAP or Visual and stored for as long as the Company deems appropriate per internal retention requirements. This feature provides the user with the capability of running queries to see how many hours were spent in each production area for any given project.

A third system that the Company uses to track projects is QlikView. QlikView is another third-party data integration, data analytics, and data visualization platform to support business operations. [4] The Company uses QlikView to integrate with both SAP and Visual to extract

and analyze data. The following sections will provide more detail on the internal uses of these data systems and the structure of the data that was available.

1.4.1 SAP

SAP has been the ERP used at MS-1 for many years and covers the full breadth of projects collected from this site. MS-2 began using SAP only in January 2019. One of the reports in SAP is set up to track shop floor labor durations and costs for specific tasks for each worker. Table 2 provides an overview of the structure for this report where each category is one of the column headers in SAP. When a worker logs in to the system, they log their hours for each task by project ID (identical to the WBS Element in SAP) with a free form text description of the task completed. All other columns are automatically populated by SAP depending on who is logging the information and the project that is being worked on. To run this report, the SAP user only needs the project ID.

Table 2. Structure of SAP report data

Category	Description
Posting Date	Date entered into SAP
Period	Month of entry
Time of Entry	Time on the date of entry
Employee Name	Employee Name
Personnel Number	Personnel Number
WBS Element*	Project ID
Partner-CCtr	Cost Center
Order	SAP specific number
ParActivity	Activity Group
Object	SAP specific number
CO object name*	Freeform text of the task(s) completed
Total Quantity*	Time spent on this task
Unit of Measure*	in hours
Cost Element	SAP number associated with ParActivity
Cost element descr.*	Description of cost element
Ref Document Number	SAP specific number
Val/COArea Crcy	Cost in US dollars
CO area currency	In US dollar
Value TranCurr	Cost in transaction currency
Transaction Currency	In transaction currency
Purchasing Document	SAP specific number

Of the 21 columns in this SAP report, only five of the columns are useful for this analysis. These five columns are denoted by the asterisk in the category column of Table 2. As mentioned, the 'WBS Element' is identical to the project ID and is the only entry used to run the SAP query. 'CO object name' is the free form text description of the task(s) completed and 'Total Quantity' is the number of hours spent working on this task. This SAP report also includes costs for materials from inventory, allocated overhead costs, and other costs not associated with direct manufacturing labor, so the 'Unit of Measure' is filtered to only include 'HR' indicating that the line item is manufacturing labor hours. Finally, 'Cost element descr.' is a preset description of the department that the worker belongs to.

1.4.2 Visual

Visual was a second ERP that had been used by MS-2 for most of their recent past. Visual was a legacy system at this facility that they continued to use even after the parent company made the acquisition. In January 2019, however, the Company fully transitioned MS-2 to the SAP ERP to have the entire company operating on the same system. Nevertheless, Visual contained the data for projects built at MS-2 prior to 2019, many of which are projects that are larger in scope and necessary for understanding how economies of scale may factor into the analysis and final model. At the time of this data collection Visual was no longer in use and new users were not set up for use in the system. Therefore, to access any data from Visual, QlikView was the source for gathering this information.

1.4.3 QlikView

QlikView is a third-party business intelligence platform that provides data analysis and visualization support to help users better understand business operations and financials. Of the many features of this platform, among the most important for this project is that it consolidates data from SAP and Visual into a single location and data structure on the Internet browser. QlikView was used as the data source for collecting projects that were started at MS-2 prior to January 2019 since these projects were stewarded using Visual and not SAP. Table 3 provides an overview of the data structure in QlikView.

Table 3. Structure of QlikView data

Category
Customer
Job #*
Leg
Op
Status
Op Type
Determinant Path
Progress
Leg Desc
Start
Finish
Days
Could Start
Delay
Description*
Planned Hrs
Actual Hrs*
Rem Hours
Reqs Fulfilled
Resource ID*
Department*
Promise Date
Sched Health

Once again, only those categories denoted by an asterisk were identified as useful for this analysis. The column headers in QlikView are much more self-explanatory as compared to SAP. Similar to 'CO object name' in SAP, 'Description' in QlikView is a free form text of the description of work performed for that line item. Additionally, 'Resource ID' was a more specific job title or job description of a worker within a more general 'Department'

1.5 Technical Data

In addition to the cost and labor data collected from the ERP and business intelligence systems, technical data was also collected for each project. The technical design for projects

dictates the scopes of work; these various metrics and characteristics would be used as explanatory variables for the estimating model.

1.5.1 3D Models

All projects have an accompanying 3D model with the exact dimensions and features of the final products. These models were used for two specific objectives. The first objective was for taking physical measurements. Not only were length, width, and height among the measurements, but certain thicknesses, depths, and curvatures were also measured for various calculations. The second objective was to count quantities of key features. Since these were all historical projects, walking to the shop floor to see the work in process was not an option to capture these quantities. Additionally, the PDF drawings that came with the 3D models were not always easy to look at when trying to capture the physical measurements and quantities of key physical features. The 3D models also provided a sense of the complexities and major unique features that needed to be considered for quantifying manufacturing durations for the various production areas.

1.5.2 Other Technical Documents

In addition to the 3D models, the estimating team is provided various technical documents either from the customer or from the Company's internal engineering team with information that is not captured in the 3D models. Some of the technical information includes surface finish specifications, contour tolerances, and quantities for items that were not explicitly presented in the 3D models. The information captured from the 3D models and from the other technical documents is all that was needed to collect the potential explanatory variables.

1.6 QuoteCentral

Internally, the Company uses a customer database system called QuoteCentral that was developed by a third party IT consulting company. This system is used by the estimating team to input the various elements of a proposal, document the basis of the proposal, and generate the necessary paperwork to send to the customer. Every proposal that the Company submits to its customers is stored in QuoteCentral. During the proposal stage, a proposal will have a quote number. If the Company wins the proposal, a project number is generated and tied to the proposal number for stewardship during the design and manufacturing phases.

Table 4 provides the structure of proposal hours in QuoteCentral. The various production areas are generalized to help maintain the anonymity of the Company. The proposals that are stored in QuoteCentral were used as the baseline to compare potential accuracy improvements. In order to maintain consistency and to assess changes in accuracy of the model versus the original proposal, the model had to generate estimates in the same format as QuoteCentral.

Table 4. Structure of Estimated Hours in QuoteCentral

	Production Area (PA)	Estimated Hours
Project Management	PA-01	
	PA-02	
	PA-03	
	PA-04	
	PA-05	
	PA-06	
Fabrication	PA-07	
	PA-08	
	PA-09	
	PA-10	
	PA-11	
	PA-12	
	PA-13	
	PA-14	
	PA-15	
	PA-16	
	PA-17	
Machining	PA-18	
	PA-19	
	PA-20	
	PA-21	
	PA-22	
	PA-23	
	PA-24	
	PA-25	
Finishing	PA-26	
	PA-27	
	PA-28	
	PA-29	
	PA-30	
	PA-31	
	PA-32	
	PA-33	
QA/QC	PA-34	
	PA-35	
	PA-36	
	PA-37	

	Production Area (PA)	Estimated Hours
"Miscellaneous" (Concealed for identity protection)	PA-38	
	PA-39	
	PA-40	
	PA-41	
Planning	PA-42	
	PA-43	
	PA-44	
Total Hours		

1.7 Data Mapping

Prior to this effort, there was no work completed to leverage historical project data for developing estimating tools for this product line. Some work had been completed on a different product line, but it did not leverage detailed SAP data as the basis for analysis (data was manually entered rather than linked to detailed SAP reports). Therefore, there was little prior knowledge or learnings to help determine how much detail was appropriate for developing an accurate estimating model that was still easy to use. Given that there was limited basis to start from, this effort started with the approach of gathering as much detailed information as possible about each project to break down the scope into as many components as possible. For example, one of the production areas that consumes lots of hours is the welding area. Welding can be broken into smaller scopes of work that are all completed within the welding production area. Theoretically, a low level of granularity is possible, but there are some challenges that currently prevent this from being possible.

The biggest challenge is the free-form text aspect of logging hours into the ERP. While free-form text allows for potentially important information to be documented in the day's work, it makes it difficult to run quick analyses in more detail. Every line item in the ERP needed to be mapped to one of the production areas. The 'Cost element descr.' column in SAP helped to

identify some of the major categories but it was not all-encompassing, meaning that there are more production areas in the QuoteCentral breakdown than are categorized in SAP. This challenge meant that a direct mapping from SAP to QuoteCentral was not possible.

To address this, a helper mapping table was created to map every line item from SAP to one of the production areas in QuoteCentral. Each ‘CO object name’ was mapped to the correct production area so that analysis at the production area level could be conducted. This mapping effort needed to happen for every project that was added to the dataset and was largely a manual process. Since there were no previously made mapping tables to convert from SAP exports to a QuoteCentral structure, using a natural language processing model, such as the bag-of-words algorithm, to map the correct categories was not yet possible. However, this is an opportunity that can be pursued in the future and will be discussed in chapter 0 Future Opportunities.

.

Model Development

As outlined in the problem statement, there were three objectives for this estimating model that were always at the forefront of the development process: 1) reduce cycle time for proposal development, 2) maintain or improve estimating accuracy, and 3) ensure an ability for continuous improvement of the model through a feedback process. To accomplish these objectives, the estimating model had to be robust and easy to use while ensuring enough granularity to assess the accuracy of the estimating model versus the baseline (i.e. the original proposals). There also needed to be a seamless methodology for updating the model with new projects as they are completed so that the model accuracy can continue to improve over time.

The estimating model was developed using both R and Microsoft Excel given the ease of use for developing preliminary models as well as the widespread user familiarity with the Microsoft Excel platform. The long-term vision for the Company is to build the model into a cloud-based platform, such as Office 365, or directly into QuoteCentral to enhance the security of the model while maintaining ease of access and use within the Company network.

1.8 Exclusion of Major Unique Scope Items

One key refinement of the model entailed accounting for major unique scope items that require appreciable amounts of manufacturing time. Through exploring the data and speaking with various individuals who support estimating and/or manufacturing operations, there were three scopes of work that needed closer attention. These are not required for most projects, but if they are part of the scope, then they require special attention when developing time estimates. These three scope items have limited consistency from project to project and, therefore, make it difficult to use historical data and statistical relationships to predict durations. To develop an estimating model for these items, additional less intuitive or difficult to measure variables would need to be captured. Therefore, this model development process intentionally excluded durations

associated with these scopes of work. Across the 77 projects used for this model, these unique scope items are only 5.5% of the total manufacturing hours so adding this extra capability was not worth the limited value add for the time allotted to the project.

Given the difficulty of automating the estimation of these unique items, the final scope and intent of the model is to help the team develop estimates for the standard structure of the products with these unique scope items excluded. For most projects, this capability covers the full scope of work from initial raw material preparation through final shipment. For other projects that have these major unique scope items, estimating manufacturing time for these features will be extra requirements that the estimator must complete. They will continue following the traditional approach for estimating time and cost for these features. By using the new model to predict manufacturing durations for the standard structures, the team will save time on the repetitive items and can focus more on the complexities, if they exist.

1.9 Data Exploration

The data collection process resulted in collecting 21 variables across the 77 projects. Of these 21 variables, 7 variables are categorical and the other 14 variables are continuous, numerical variables. All but one of the variables are considered objective variables, meaning that they are observable and measurable. The one subjective variable is a contour complexity categorization that attempts to categorize projects into one of three types of relative perceived manufacturing difficulty: low, medium, and high. These complexity categories are based on how sharp the curves are and how many different directions the contour curves in.

Table 5 lists the 7 categorical variables that were collected and the number of different categories each variable could take on. Table 6 lists the continuous variables and their respective variable inflation factors (VIF). In the data collection process, it was known that many variables

were not independent and that multicollinearity among regressors would be a concern. The variable inflation factor detects multicollinearity and estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. [5] A higher VIF means that a variable is highly correlated with the other predictors. The general rule of thumb varies among sources, but a VIF larger than 10 is consistently considered high among all sources. Therefore, these variables have high levels of multicollinearity and will need to be accounted for to generate a stable model.

Table 5. Categorical variables and the number of categories each can take

Variable Number	Number of Categories
Var-1	2
Var-2	3
Var-3	2
Var-4	2
Var-5	2
Var-6	2
Var-7	3

Table 6. Continuous variables and their variable inflation factors

Variable Number	Variable Inflation Factor
Var-8	142.16
Var-9	139.38
Var-10	15.68
Var-11	144.20
Var-12	7.16
Var-13	3.14
Var-14	2.98
Var-15	58.78
Var-16	65.64
Var-17	43.72
Var-18	489.99
Var-19	309.91
Var-20	110.45
Var-21	80.50

In addition to the need for addressing the issue of multicollinearity, expanding the list of potential regressors was necessary to address the potential for nonlinear relationships between the response variable (Total Hours) and the explanatory variables. All continuous variables were squared, square-rooted, and natural-logged (where it potentially made sense) to expand the list of explanatory variables from 21 to 62. Given the need to decide which variables were most important, reduce the model to as few variables as possible, and resolve the multicollinearity problem, a lasso regression (least absolute shrinkage and selection operator) model was chosen to conduct the analysis.

1.10 LASSO Regression

Lasso regression models are particularly well-suited when explanatory variables have high levels of multicollinearity and when a simple, interpretable model with a small number of variables is preferred over a more complex model. Lasso regression operates with the support of a tuning parameter, λ, that penalizes each coefficient in the regression based on its magnitude.

This algorithm results in driving some coefficients to zero and eliminating them from the model altogether. [6]

Two lasso models were built, tested, and compared to see which strategy best balanced predictive ability with model simplicity. The first lasso model focuses on predicting the total manufacturing hours rather than predicting each production area individually. These hours are then distributed to each of the production areas based on the distribution of hours from the actual dataset. Distributing hours to each production area is important because these are the budgets each production area will use to steward execution. The second lasso model groups the production areas into six production area groupings and a lasso model for each grouping is produced to predict hours at the grouping level. The hours are then distributed to each production area following the same methodology as the first lasso model. Each lasso model was then compared against the baseline model, where the baseline is the Company's current estimating process. The actual hours from SAP were the point of comparison to see whether the lasso models were more effective prediction methods than the baseline. The following explanation describes the process for modeling the first lasso model, but the same process can be followed for the second lasso model as well.

R was the programming language chosen to develop these lasso models. The response variable (Hours) and all 62 explanatory variables were stored in a CSV file and imported into R. A set of ten randomly generated seeds were used to build ten slightly different models to see how model variable selection and coefficients changed across trials. For each seed, the data were partitioned into a training set and a test set where 75% of the projects were in the training set and 25% were in the test set. This resulted in the training set containing 58 projects and the test set

containing 19 projects. The *glmnet* function in R was used to execute the lasso regression on the training set using the following format:

$$lasso\_mod = glmnet(x\_train, y\_train, alpha = 1, lambda = grid)$$

where *grid* is a sequence of 10,000 evenly spaced numbers from 10^{-2} to 10^{10}. This range of values ensured that the optimal *lambda* fell somewhere in between these two extremes and 10,000 samples was the order of magnitude before the algorithm's processing time became slow. *lambda* was optimized using cross validation by executing the two functions below.

$$cv.out = cv.glmnet(x\_train, y\_train, alpha = 1)$$

$$bestlam = cv.out\$lambda.min$$

Figure 2 shows a sample plot of the training mean-squared-error (MSE) as a function of lambda. The optimal *lambda* is where the MSE of the plot in Figure 2 is at its minimum.

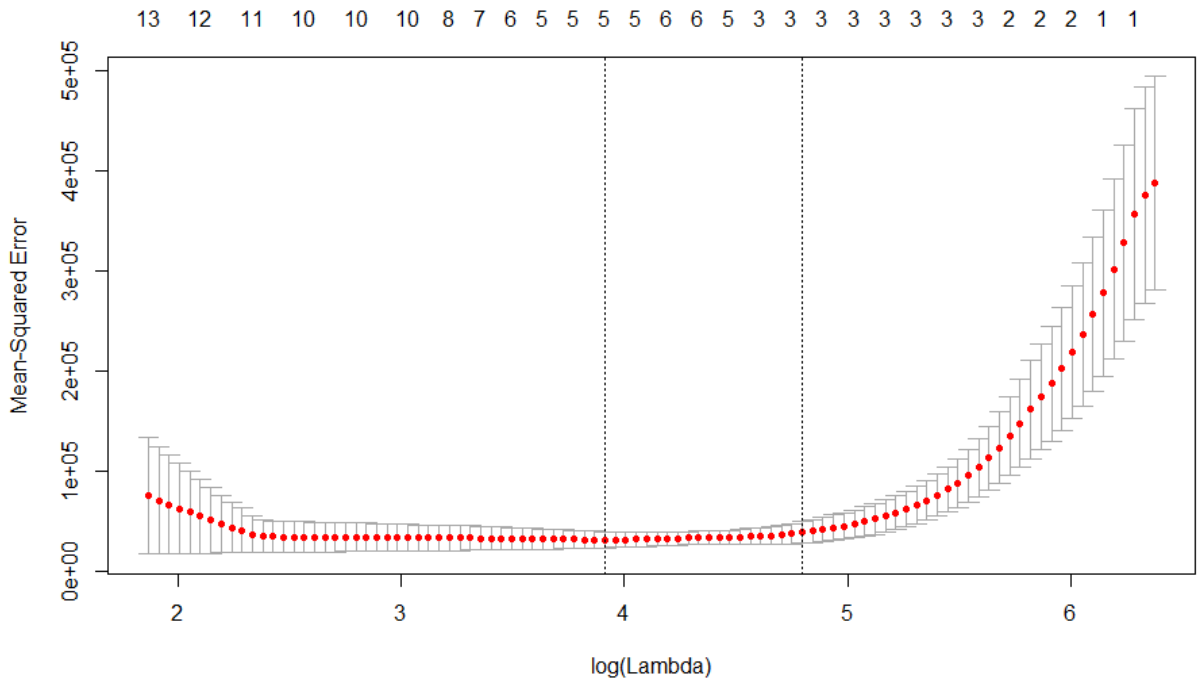

Figure 2. Lasso lambda cross-validation

The optimal *lambda* for each trial is then used to predict the total hours in the test set using the function below:

$$lasso\_pred = predict(lasso\_mod, s = bestlam, newx = x\_test)$$

where *x_test* is the array of explanatory variables in the test set. For each trial, the root mean square error (RMSE) on the test set was always lower than the RMSE of the baseline estimating process. This was an indication that the model could show performance improvements and was worth continuing to pursue.

Next, a lasso model was fit with the same optimal *lambda* in each trial on the full dataset with the following R functions:

$$out = glmnet(x, y, alpha = 1, lambda = grid)$$

$$lasso\_coef = predict(out, type = "coefficients", s = bestlam)$$

where *x* and *y* are the explanatory variables and response variable, respectively, for the full dataset. Table 7 shows the resulting regression model variables and coefficients for the full dataset across all ten trials for the first lasso model. The table also shows the root mean square error for each trial, which averaged 168 across the ten trials. The root mean square error is a metric used to quantify model error and variation; this and other performance metrics will be discussed further in section 1.12 Model Testing and Performance.

Table 7. Lasso regression models for ten random trials in the first lasso model

Trial	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RMSE	186	239	186	139	107	129	150	229	197	122
(Intercept)	79.4	81.1	106	130	76.4	76.0	107	75.2	140	84.5
Var-2	124	122	88.2	56.3	138	139	86.8	139	42.4	117
Var-7	-	-	-	-	-18	-19	-	-21	-	-
Var-19	5.0	5.0	4.7	4.4	5.3	5.4	4.7	5.4	4.3	5.0
Var-20	6.8	6.9	9.7	12.3	5.5	5.5	9.8	5.5	13.4	7.3
Var-21	1.0 E-02	1.0 E-02	9.4 E-03	8.6 E-03	1.0 E-02	1.0 E-02	9.4 E-03	1.0 E-02	8.2 E-03	1.0 E-02
(Var-21)^2	-	-	-	-	3.7 E-09	4.8 E-09	-	6.4 E-09	-	-
sqrt(Var-17)	3.1	3.1	3.1	3.1	3.6	3.6	3.1	3.6	3.1	3.1
sqrt(Var-19)	22.0	22.2	24.7	27.1	19.4	19.5	24.8	19.8	28.2	22.6

The lasso regression performs as expected and eliminates either 54 or 56 of the 62 explanatory variables, leaving only six or eight explanatory variables depending on the trial. However, in these ten trials, *Var-19* appears twice in every trial: once as itself and once as the

square root of itself. Ideally, the final model should not include both *Var-19* and *sqrt(Var-19)* since these variables can be used to explain each other. With the same randomly generated seeds, these ten trials were ran again with *sqrt(Var-19)* removed as an option from the dataset. Table 8 shows the resulting regression models.

Table 8. Lasso regression models for ten random trials with *sqrt(Var-19)* removed

Trial	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
RMSE	188	240	186	114	108	127	146	230	194	119
(Intercept)	133	137	166	151	124	124	155	125	191	142
Var-2	125	120	88.8	104	139	139	100	138	61.5	115
Var-7	-	-	-	-	-27.9	-27.6	-	-26.1	-	-
Var-19	5.8	5.8	5.6	5.7	6.1	6.1	5.6	6.1	5.4	5.7
Var-20	12.2	12.7	15.9	14.3	9.9	9.9	14.8	10.1	18.6	13.2
Var-21	8.6 E-03	8.4 E-03	7.4 E-03	7.9 E-03	9.1 E-03	9.1 E-03	7.7 E-03	9.1 E-03	6.4 E-03	8.3 E-03
sqrt(Var-17)	10.4	10.5	11.4	10.9	9.6	9.6	11.1	9.7	12.0	10.7

While *sqrt(Var-19)* was manually removed from the model, this series of reruns also removed $(Var\text{-}21)^2$ through the algorithm. Despite removal of these two variables, the average root mean square error reduced from 168 to 165 and shows that the performance, based on this metric alone, is not compromised. Additionally, *Var-7* was expected to be an important metric but only appeared in three of the ten trials. Since *Var-7* is a categorical variable and is expected to be significant, a few more trials were run to generate a total of five models that include *Var-7* and then all coefficients were averaged to determine the final model to be used for analysis. Table 9 summarizes the final regression model that was used for the first lasso modeling approach.

Table 9. Final lasso model with five trials that include *Var-7*

Model	**1**	**2**	**3**	**4**	**5**	**Average**
RMSE	139	139	139	143	141	140
(Intercept)	124	124	125	128	127	126
Var-2	139	139	138	132	134	136
Var-7	-27.9	-27.6	-26.1	-11.7	-17.7	-22.2
Var-19	6.1	6.1	6.1	5.9	6.0	6.0
Var-20	9.9	9.9	10.1	11.1	10.7	10.3
Var-21	9.1 E-03	9.1 E-03	9.1 E-03	8.8 E-03	8.9 E-03	9.0 E-03
sqrt(Var-17)	9.6	9.6	9.7	10.1	9.9	9.8

While this lasso model predicts the total hours across all production areas, each production area needs their own estimates to plan capacity requirements and to properly steward project execution. To account for this, the total hours are allocated across each production area based on the observed distribution in the dataset. Table 10 shows a breakdown of how predicted hours are distributed across the production areas. The distributions are grouped by the three different material types and three contour complexity levels.

Table 10. Allocation of production hours by material and complexity

	Mat-1 Com-L	Mat-1 Com-M	Mat-1 Com-H	Mat-2 Com-L	Mat-2 Com-M	Mat-2 Com-H	Mat-3 Com-L	Mat-3 Com-M	Mat-3 Com-H
PA-01	4%	3%	2%	5%	1%	2%	3%	3%	4%
PA-07	0%	0%	1%	1%	0%	1%	1%	1%	1%
PA-08	3%	8%	8%	4%	6%	8%	5%	7%	3%
PA-09	14%	30%	32%	18%	32%	32%	22%	16%	23%
PA-10	3%	3%	2%	2%	2%	2%	1%	1%	2%
PA-11	3%	4%	3%	3%	2%	3%	3%	2%	3%
PA-12	4%	3%	5%	5%	6%	5%	5%	5%	6%
PA-21	2%	2%	3%	3%	5%	3%	5%	5%	3%
PA-22 PA-23 PA-24 PA-25	26%	14%	14%	19%	18%	14%	18%	15%	20%
PA-26	6%	10%	9%	13%	13%	9%	12%	8%	10%
PA-27	16%	9%	7%	11%	4%	7%	10%	19%	7%
PA-28	2%	2%	2%	2%	2%	2%	0%	0%	0%
PA-29 PA-34	11%	8%	6%	8%	7%	6%	4%	11%	10%
PA-33	1%	2%	1%	1%	1%	1%	4%	2%	3%
PA-34	6%	4%	4%	5%	2%	4%	9%	5%	5%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Allocating predicted hours to each production area using Table 10 was the first strategy for providing estimates for each production area. A second model was created that focused on estimating manufacturing durations for groupings of related production areas rather than solely focusing on the total hours. The objective for this model was to, theoretically, increase accuracy and precision for each of these production area groupings. The same lasso regression approach was conducted for the production area groupings (PAG) in Table 11 that share similar characteristics and human resources.

Table 11. Production area groupings for second lasso model

Production Area Grouping	Production Area
PAG-1	PA-07
	PA-09
	PA-10
	PA-11
	PA-12
PAG-2	PA-08
PAG-3	PA-21 PA-22 PA-23 PA-24 PA-25
PAG-4	PA-26
PAG-5	PA-27
	PA-28
	PA-29 PA-34
	PA-33
	PA-34
PAG-6	PA-01

A sample model of one trial of the lasso regression analysis resulted in the model outlined in Table 12. The first column lists the variables that were selected across all PAGs. Each number represents the coefficients for their respective variables to calculate hours for each PAG. A series of additional trials were run yielding similar variable selections as those in Table 12, but this methodology was abandoned due to an appreciable increase in model complexity and perceived level of effort for updating without an appreciable improvement in predictive performance. This will be discussed further in section 1.12 Model Testing and Performance.

Table 12. Sample of lasso regression across six production area groupings

	PAG-1	PAG-2	PAG-3	PAG-4	PAG-5	PAG-6
Intercept	-95.8	-10.9	73.6	44.7	67.5	12.6
Var-2	117	26	-	-	-	-
Var-11	0.7	-	-	-	-	-
Var-1	-	1.1	-	-	-	-
Var-18	-	1.56E-03	-	-	-	-
Var-17	-	-	0.1	-	-	-
Var-16	-	-	-	0.4	0.4	0.3
Var-19	-	-	1.2	0.2	1.5	-
Var-20	16.8	-	-	-	-	-
sqrt(Var-16)	-	-	-	0.8	-	-
sqrt(Var-17)	-	1.9	-	1.0	-	-
sqrt(Var-18)	-	3.1	-	-	-	-
sqrt(Var-19)	27.3	-	-	-	5.1	-
sqrt(Var-20)	-	0.6	-	-	-	-
sqrt(Var-21)	5.9E-08	-	-	-	-	-
Ln(Var-18)	15.8	-	-	-	-	-

This page has been intentionally left blank

Discussion

1.11 Variable Selection

A fundamental characteristic of the lasso modeling algorithm is that it naturally selects the explanatory variables that provide strong prediction models while eliminating variables that do not add much value to the model's predictive power. For both models, the algorithm eliminates variables without regard to whether the user expects a particular variable or set of variables to be included or not. Two of the continuous variables included in the first lasso model were expected to be included based on an intuitive understanding of the manufacturing process and how highly correlated these variables were with the total hours. However, two of the categorical variables, specifically material and complexity, were deemed to not be as important for predicting total manufacturing hours. The initial expectation was that all three materials and all three complexities would be vital characteristics for predicting manufacturing durations, but the lasso regression only distinguishes one of the three material types and only one of the three complexity levels. If complexity is truly important in reality, then this analysis may benefit from reassessing how complexity is captured and include more projects that are noticeably and measurably more complex than the projects included in this dataset. Additionally, the model would likely benefit from including more projects that were manufactured using the other two materials that are not distinguished from each other in the final model.

The second lasso model, which is the aggregation of six lasso regressions, was much more complex than the first lasso model. Each of the six individual models include anywhere from one to six variables, but the variables chosen are not always intuitive for explanation purposes. Furthermore, each of the six sub-models use different variables in the calculations so the full list of variables in the complete second model is 15 rather than six variables for the first model. Of these 15 variables, eight of them are direct physical measurements or observations as

opposed to four measurements or observations for the first lasso model. This increase in variable count increases the level of effort required to update the model in the future and can only be justified if there is an appreciable improvement in predictive performance.

1.12 Model Testing and Performance

A mix of criteria was used to compare the two lasso models against the baseline, where the baseline is the current estimating process utilized by the Company. The performance criteria included the following five metrics defined below: mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), mean bias error (MBE), and mean percentage error (MPE).

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

$$MAPE = \frac{1}{n}\sum_{j=1}^{n}\frac{\left|y_j - \hat{y}_j\right|}{\left|\hat{y}_j\right|}$$

$$MBE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)$$

$$MPE = \frac{1}{n}\sum_{j=1}^{n}\frac{y_j - \hat{y}_j}{\hat{y}_j}$$

The mean absolute error quantifies the average magnitude of the error, regardless of whether the error is positive or negative. Each residual contributes proportionately to the total amount of error, meaning that larger errors contribute linearly to the overall error. [7] The root

mean square error is a quadratic scoring rule that also measures the average magnitude of the error. It is different from the mean absolute error in that the RMSE is the square root of the average of squared errors. Since errors are squared before being averaged, larger errors contribute more weight in the calculation than smaller errors. RMSE is an important calculation for models where larger errors are particularly undesirable. [8] Whereas MAE and RMSE measure error in units (in this case the units are hours), the mean absolute percentage error measures error in percentage terms, a unitless measurement. MAPE can be used in conjunction with MAE and RMSE to paint a more complete picture of model accuracy, regardless of whether errors are positive or negative. For all three of these calculations, the closer the number is to zero, the more accurate the predictions are.

The mean bias error and mean percentage error are similar calculations to the mean absolute error and mean absolute percentage error, respectively, except the absolute value function is removed. The MBE and MPE are used to determine whether a model typically overestimates or underestimates in its predictions and are not helpful for determining model accuracy. In these calculations, a large negative error and a large positive error with the same magnitude would cancel each other out. Positive calculations for MBE and MPE indicate a bias towards overestimating while negative calculations indicate a bias towards underestimating. The closer the calculations are to zero, the lower the bias. [8]

Table 13 summarizes the results of these calculations for the baseline and for both lasso models. Both lasso models perform better than the baseline across every performance metric except for the mean absolute percentage error where the second lasso model performs about equal to the baseline. The lasso models are generally more accurate, have less variation in the magnitude of errors, and have less bias. This observation supports the hypothesis that advanced

data analytics, and lasso regression in particular, can provide noticeable improvements in estimating performance with the appropriate technical inputs.

When comparing the two lasso models, the first lasso model performs noticeably better than the second lasso model in every performance category. This result was expected given that the first lasso model was built using the total, bottom line durations whereas the second model is six sub-models for six production area groupings that were aggregated to develop the total predicted durations.

Table 13. Lasso model performance versus baseline

	Baseline	Lasso Model 1	Lasso Model 2
Root Mean Square Error	413	140	187
Mean Absolute Error	247	111	147
Mean Absolute Percentage Error	24%	17%	25%
Mean Bias Error	152	5	5
Mean Percentage Error	19%	8%	16%

The performance metrics between the two lasso models were also assessed for each of the production area groupings. The objective for comparing these two models was to see if performance improved by modeling production area groupings individually rather than allocating these durations from a single model that predicted total hours for the entire project. Table 14 through Table 18 summarize the performance metrics for five of the six production area groupings as compared to the allocation approach used for the first lasso model. Program management and production planning was the sixth production area grouping but this grouping was not compared between the two models because there are not any strong relationships between the program management durations and any technical characteristics, measurements, or calculations.

The performance differences between the two models at the production area grouping level are inconsistent. In some areas, the second lasso model performs better than the first model and in other areas the first model still outperforms the second model. Even when one model performs better than the other for a given production area grouping, the performance improvement is usually not very significant. Given this inconsistency, it cannot be reliably determined which model is a better performer at the production area grouping level. However, the simplicity of the first model as compared to the second model, both in terms of the number of variables required and the level of effort required to provide model updates, makes the first model much more preferred.

Table 14. Production Area Grouping 1 performance metrics

	Lasso Model 1	Lasso Model 2
Root Mean Square Error	104	85
Mean Absolute Error	71	63
Mean Absolute Percentage Error	28%	33%
Mean Bias Error	-11	0
Mean Percentage Error	14%	16%

Table 15. Production Area Grouping 2 performance metrics

	Lasso Model 1	Lasso Model 2
Root Mean Square Error	30	28
Mean Absolute Error	18	18
Mean Absolute Percentage Error	35%	40%
Mean Bias Error	-1	0
Mean Percentage Error	12%	20%

Table 16. Production Area Grouping 3 performance metrics

	Lasso Model 1	Lasso Model 2
Root Mean Square Error	141	145
Mean Absolute Error	74	73
Mean Absolute Percentage Error	34%	38%
Mean Bias Error	-23	-37
Mean Percentage Error	9%	11%

Table 17. Production Area Grouping 4 performance metrics

	Lasso Model 1	Lasso Model 2
Root Mean Square Error	44	54
Mean Absolute Error	24	33
Mean Absolute Percentage Error	51%	114%
Mean Bias Error	-7	0
Mean Percentage Error	34%	100%

Table 18. Production Area Grouping 5 performance metrics

	Lasso Model 1	Lasso Model 2
Root Mean Square Error	104	92
Mean Absolute Error	68	60
Mean Absolute Percentage Error	25%	36%
Mean Bias Error	-10	-4
Mean Percentage Error	11%	23%

1.13 Observations

There are several important observations to notice in this analysis. The first important observation is looking at the mean absolute error (MAE) and root mean square error (RMSE) for both the baseline model and the lasso model. The MAE for the baseline is about twice the size of the MAE for the lasso model and the RMSE for the baseline is over 2.5 times the size of the RMSE for the lasso model. These metrics help provide evidence that the lasso model is both

more accurate given the lower error metrics and that the lasso model does a better job reducing the magnitude of large errors since the RMSE does not increase substantially above the MAE as compared to the baseline.

A second important observation is that the original proposals show a strong bias towards overestimating the number of labor hours required to manufacture these products. On average, the original proposals, which were largely based on the estimators' prior experiences and personal judgement, allocated 19% more hours on average than were required to complete the project. This equated to an average of 152 hours surplus for each project. The bias for the lasso model was 8% high and only five hours high on average across the dataset. Logically, it is more likely that estimators would be conservative and allocate too many hours for the proposal rather than estimate aggressively and risk not having enough hours. If the operations team is executing a project and they are going to run overbudget without having had any major execution challenges, they will often question and put pressure on the estimator that was responsible for the proposal. However, if the operations team has a budget surplus, the estimator will rarely get questioned.

Systematically overestimating on proposals also has a negative impact on the business. The Company is fortunate that they still won the bids for the overestimated projects in this dataset. However, it is likely that the Company failed to win competitive bids for some other projects due to overestimation. If there was not a strong bias for overestimating proposals, the Company's success would likely increase on bids that are priced strictly on a cost-plus basis.

Finally, it is important to observe how the performance metrics change based on the project size. Table 19 shows the performance metrics for both the baseline and for the first lasso model in three groupings: all projects, projects that required more than 500 hours to manufacture,

and projects that required more than 1000 hours to manufacture. The baseline either maintains its performance or performs worse across every performance metric. The mean absolute error and root mean square error both increase as projects grow. The metric that is the most relevant point of comparison is the mean absolute percentage error. This error grows from 24% in the full dataset to 30% for projects larger than 1000 hours. This shows that the absolute error is growing at a faster rate than project size and that the baseline performs worse as projects become larger. Conversely, the lasso model shows some improvements as projects grow. Again, as expected, the mean absolute error and root mean square error both grow as projects become larger, but they do so at a slower rate than the baseline. Simultaneously, the mean absolute percentage error for the lasso model decreases from 17% to 11% to 9% across the three project size segments. As previously mentioned, it is important for the business and manufacturing operations to improve estimating performance for larger projects given the financial risks associated with large estimating errors on larger projects.

Table 19. Performance metrics of lasso and baseline models by project size

	Baseline	Model	Baseline > 500 hrs	Model > 500 hrs	Baseline > 1000 hrs	Model > 1000 hrs
RMSE	413	140	502	156	726	194
MAE	247	111	336	124	580	159
MAPE	24%	17%	24%	11%	30%	9%
MBE	152	5	200	-25	283	-95
MPE	19%	8%	19%	1%	18%	-5%
	n = 77		n = 51		n = 23	

Although the lasso model generally performs better than the baseline as projects grow larger, the lasso model still shows an inherent vulnerability when estimating larger projects. There appears to be an inherent bias for slightly underestimating larger projects where the mean

percentage error for projects larger than 1000 hours was -5% and the mean bias error was -95 hours. With only 23 projects larger than 1000 hours in this dataset, there is an opportunity to improve on this metric by adding a variety of projects that meet this size threshold to improve upon this underestimation bias.

Figure 3, Figure 4, and Figure 5 graphically show the mean percentage errors of the lasso regression against the baseline as projects grow in size. The errors of the lasso model predictions are shown in green while the errors from the baseline are shown in red. The solid lines show the mean percentage error and the dashed lines are calculations for two standard deviations to visually represent the variation in the magnitude of the error across the dataset. While the size of the band between the two standard deviations tend to get smaller for the lasso model as projects become larger, this same band increases for the baseline as projects become larger.

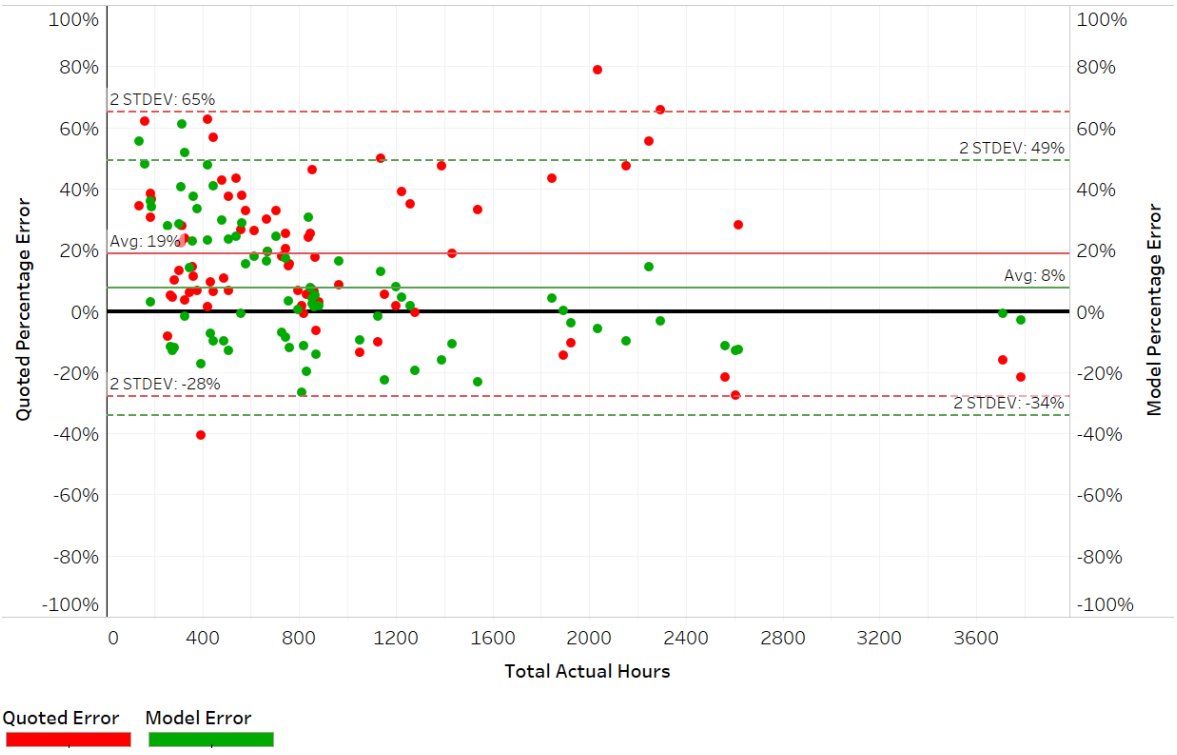

Figure 3. Mean percentage error of lasso regression vs baseline for all projects

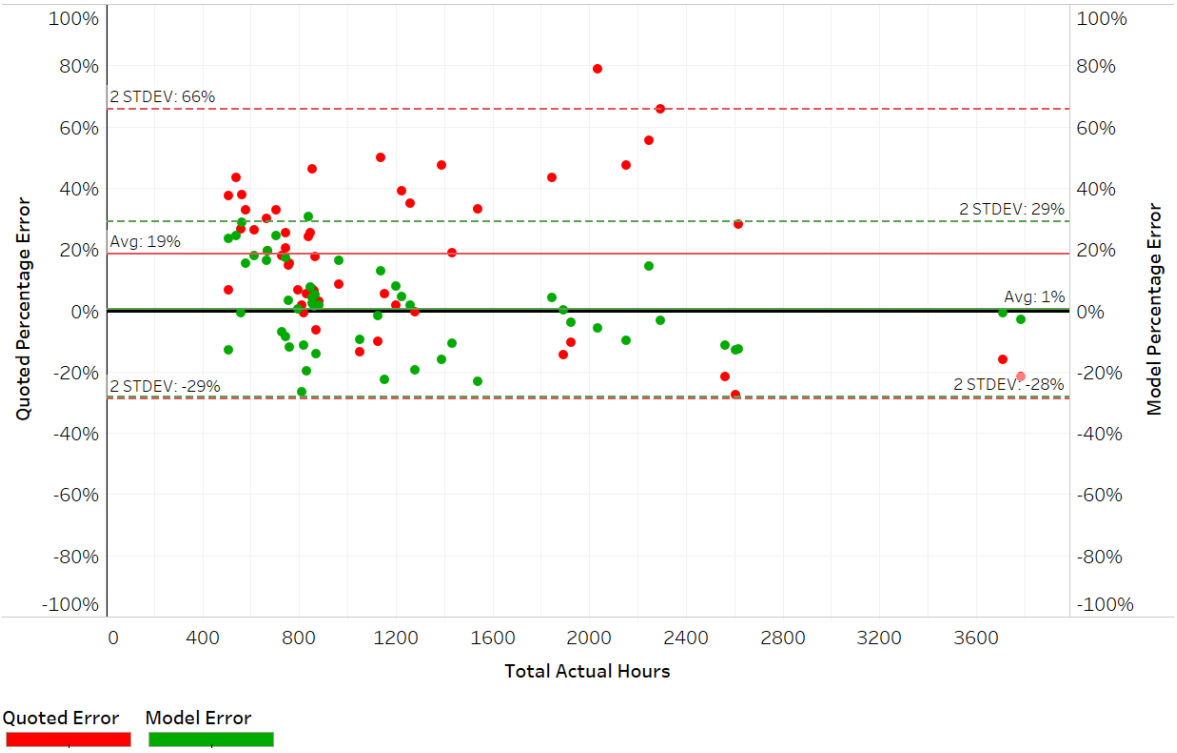

Figure 4. Mean percentage error of lasso regression vs baseline projects > 500 hours

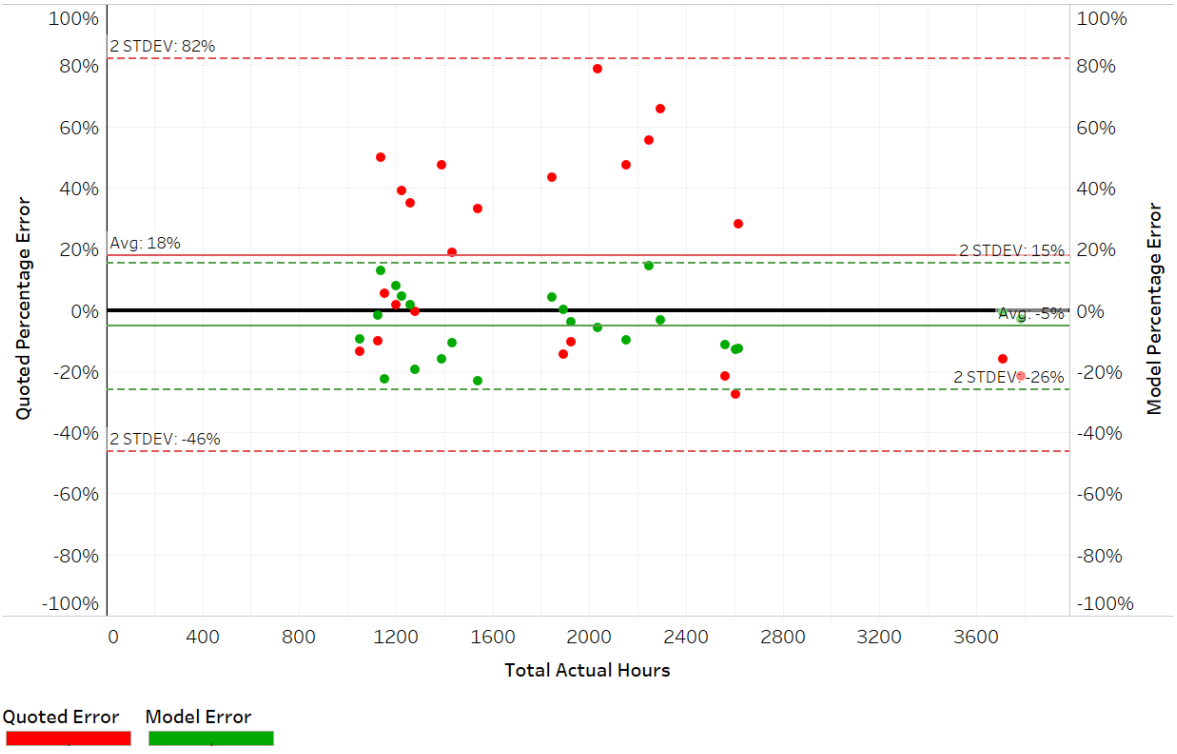

Figure 5. Mean percentage error of lasso regression vs baseline projects > 1000 hours

These observations also have significant implications on the business and can be demonstrated through this simple example. Let's say that an estimator predicts that Project A will require 500 hours of manufacturing time, but it actually took 625 hours to complete. The estimate was 125 hours short and translates to a -25% error. Now let's say that an estimator predicts that Project B will require 2000 hours to complete, but it actually took 2500 hours. The estimate in this case was 500 hours short but also translates to a -25% error. The impact of underestimating Project B is much higher than the impact of underestimating Project A. Underestimating Project B is more likely to cause production areas to become overcapacity and lead to schedule delays not only for this project, but also for other projects in the queue. The Company is then more likely to outsource some portions of the work scope to relieve capacity, usually resulting in cost increases to meet schedule requirements. Finally, assuming the proposal took a cost-plus pricing approach, unless there was a substantial profit markup (which is highly unlikely in this competitive market), the Company is now taking a financial loss to provide this product to the customer.

Overestimating on larger projects also has negative implications for the business. One point that was mentioned earlier is that the Company may be failing to win some bids if a competitor is delivering proposals that are both lower cost and more accurate. A second point is that the Company plans capacity needs weeks in advance of when the capacity is needed. These planning efforts leverage the hours from successful bids to plan requirements for each production area. Overestimating hours on larger projects can lead to underutilization in the future. This is particularly problematic if the Company made advanced plans to outsource some work and signed contracts to complete work elsewhere when they have the capacity to complete more work in-house. The negative impacts of overestimation or underestimation increase nonlinearly

as projects grow so it is in the Company's best interest to ensure mean absolute percentage error decreases as projects grow in size.

1.14 Final Excel Model Structure

While the lasso regression analysis was conducted using R, the final model was implemented using Microsoft Excel given the ease of use for the team. The Excel model contains 17 inputs as listed in Table 20; five of the input names are concealed to help protect the industry and identity of the manufacturer. These 17 inputs are used to generate the remaining explanatory variables in the analysis.

Table 20. Final Microsoft Excel model inputs

Build Location
Material
Length (in)
Average Width (in)
Height (in)
Depth (in)
Base Structure Type
Surface Finish
Surface Complexity (1-3)
Surface Style
Surface Thickness (in)
Support Thickness
Hidden Measurement 1
Hidden Measurement 2
Hidden Measurement 3
Hidden Measurement 4
Hidden Measurement 5

The build location is the only input that was not included as an explanatory variable for lasso regression. The build location only affects one of the production areas. Manufacturing Site

2 (MS-2) has an in-house capability that Manufacturing Site 1 (MS-1) does not have. All projects require this scope of work but the duration requirements for this work do not significantly depend on any physical measurements or properties of the product. Thus, if the project is proposed for manufacturing at MS-2, then a constant number of hours is added to the estimate. If the project is executed at MS-1, then the estimator must add the cost for outsourcing to the proposal.

1.15 Management Considerations

While advanced data analytics has potential for improving the estimating and proposal development process, there are some challenges that company management must be aware of. These challenges include ensuring that the Company has the skills necessary to update and maintain the model, developing trust in the new data-driven estimating process over the historical "human-intuition" process, and providing the appropriate incentives to ensure data is accurate and reliable. Expansion of the model applications, change management, and growing the culture of continuous improvement and data-driven decision making will be necessary to capitalize on the full potential of this initiative.

While the final model is implemented in Microsoft Excel, the analysis and model development were executed in R. In order to update this model with new projects, and to expand the applicability of this methodology to other product types, the Company will need a human resource who knows how to manipulate the R code and understand the outputs generated. Updating the model for the product type focused on for this project will not be a major concern since the code has already been written. The administrator simply needs to update the CSV file that is loaded into R with the new dataset and run the code. The output is a simple copy and paste from R to the Excel file. However, as the Company explores expanding this methodology to

other product types, this code will need to be manipulated to fit the new products and will likely require the administrator to have some understanding of R.

In addition to having a basic understanding of R, the person who develops new models for other product types will need to have an understanding of manufacturing processes and have skills in R programming, Microsoft Excel, and 3D modeling software. These skills are typically readily available with LGO students, especially if they come from a mechanical or aerospace engineering program. There is also an opportunity for young engineers who have strong data analysis skills to challenge themselves and take on some of these responsibilities. Most undergraduate mechanical and aerospace engineering programs include statistics and 3D modeling as part of the core curriculum. The challenge would be for them to learn the fundamentals of R and implement lasso regression (or other modeling techniques) where appropriate. The concept of lasso regression is based on linear regression and multiple regression which is covered in most statistics courses. With this model as a go-by, a bright, young engineer can learn from this methodology and apply it to other product types while having the opportunity to learn more about the company, engage with various stakeholders, and develop their leadership and communication skills.

For this model (and the estimating methodology more generally) to have long term success, the various stakeholders need to develop trust and confidence in the estimating process. It is important for the Company to understand that no process and no estimating methodology is perfect. However, it does appear that this methodology provides some performance improvements over the existing process for the applicable product types and sizes and that the proposal development cycle time is reduced by at least 50%. It can be easy to blame the model and deem it as useless or unreliable if it ever has a big miss. In fact, the model will have a higher

frequency of underestimation since it fundamentally removes the systematic conservative bias that human estimators have.

There are two things that will help the Company and the key stakeholders develop confidence in this estimating methodology and approach. The first is to educate the key stakeholders on the fundamentals of how the model works and how it performs relative to the baseline process. This education can take the form of a lunch and learn series to demonstrate (at the appropriate level of detail) how the model was built, how it performs relative to the baseline, and what potential operational and financial opportunities exist if the company was able to fully transition their estimating methodology. The second thing that will help build trust and confidence is time. Initially, the Company should continue to use the baseline estimating process in parallel with the new estimating model and continue comparing the performance of both. Over time, if the new model continues to perform better than the baseline, the team will continue building confidence and can fully transition to the semi-automated model for applicable projects.

Finally, this model relies on consistent and reliable data as the foundation for model development. If unreliable data from historical projects is incorporated into the analysis, then the outputs from the model will also be unreliable. For long-term sustainability, the company needs to ensure the correct incentives are in place to ensure reliable data is entered into SAP during manufacturing. There is a potential for the Company to have a challenge regarding data reliability. Shop floor workers might log hours to projects in SAP in a fashion that is most beneficial for them politically rather than accurately log hours to projects as it actually happened. More specifically, if a production area is about to overrun their budgeted hours for a project, it is not unreasonable for a worker to log more hours to another project that has a budget surplus. Thus, a project that should have overrun, or should have overrun by a more significant amount,

may have been worse if some hours were inaccurately allocated to another project. Simultaneously, a project that finished at or under budget may have inappropriately received hours from a project that was running over budget. This practice would compromise the data quality and make the model less reliable.

This risk of data reliability is particularly important since the model fundamentally removes the conservative estimating bias they have seen historically. More projects will overrun their budget than they've seen previously. To prevent the practice of misallocating hours from occurring, the Company must ensure that there is not a culture of blame and punishment, but rather a culture that promotes data accuracy so that management and other office workers can learn from the operational challenges, adjust processes, and improve estimating models. Senior management should continuously reinforce the importance of accurate data management and that nobody will be punished or negatively treated for any individual project going over budget. A more holistic evaluation across a larger set of projects should be the norm.

Another approach to help ensure data reliability is to prevent the shop floor workers, and possibly even the program managers, from knowing what their budget is for each individual project. Removing this anchor could help promote consistency across the portfolio of projects. Program managers and shop floor workers would focus solely on meeting the required delivery dates to ensure customer satisfaction and the operations managers would focus on ensuring each production area has the necessary resources to complete the work on time. The hours would then be a true representation for the level of effort required to manufacture each project.

Future Opportunities

The Company can continue improving on this study to enhance accuracy, speed, and breadth of estimating. These improvements include modifying the 3D modeling capabilities to accelerate data collection, implementing natural language processing algorithms to support SAP data mapping, and expanding the scope of this modeling methodology to other product types.

When developing this model, data collection was unquestionably the most time-intensive requirement. For each project, data had to be collected from 3D models and other PDF files and manually entered into an Excel spreadsheet. For each project, this could take around 30 minutes, assuming the data was readily available. There was also data that ideally would have been useful to include in the dataset but were either too time-consuming to reliably capture or too difficult to measure consistently across projects. A potentially big opportunity exists if the Company can modify its 3D modeling software to generate an export into a spreadsheet in the same format as the technical inputs in the model. These automated exports would include both physical measurements of the product itself as well the quantities of key features. Automating this process would save time for both the estimators and the administrative user who is responsible for capturing the technical data during model updates.

Continuing with the opportunity to accelerate the model updating process, a natural language processing algorithm can be implemented into the machine learning model to prevent manually mapping unique ‘CO object name’ descriptions to the various production areas. This research found that the workers are not consistent in their descriptions of the work that is completed for each project. Most projects had some unique descriptions (some were just the result of typos) requiring a manual data mapping step to be implemented into the model updating process. A natural language processing algorithm can help alleviate this challenge. While there were over 1500 unique ‘CO object name’ descriptions, many of those that fell into the same

category contained specific keywords or were part of the same ‘ParActivity’ to indicate the correct production area to map to. The mapping table developed for this project can now be used to train a language processing algorithm to automate the SAP data mapping process. Furthermore, the Company can help this effort by changing the ‘CO object name’ from free form text to a dropdown option when workers log their time in SAP. This standardization would also prevent the need for this additional algorithm.

Finally, this project focused on only one of the Company’s products, but they provide many products and solutions for their customers. This modeling methodology can be expanded to other product types following a similar process of collecting technical data from 3D models and time data from SAP and using lasso regression analysis to develop estimating models. Similar to this project, the developer may need some creativity in defining the explanatory variables and using logic and intuition to create variables from a series of other simple measurements. If the Company can apply a streamlined and semi-automated estimating methodology across its portfolio with modern data analytics methodologies, they will continue speeding up the estimating process, improving estimate accuracy, and removing estimator bias.

Bibliography

[1] B. Goldense, "The 5 Types of Manufacturing Processes," Endeavor Business Media, LLC., 24 August 2015. [Online]. Available: https://www.machinedesign.com/community/contributing-technical-experts/article/21831946/the-5-types-of-manufacturing-processes#:~:text=. [Accessed 18 March 2020].

[2] SAP SE, "Enterprise Resource Planning (ERP) and Finance," 2020. [Online]. Available: https://www.sap.com/products/erp-financial-management.html. [Accessed 18 March 2020].

[3] Infor , "Order-Driven Manufacturing ERP Solutions," [Online]. Available: https://www.infor.com/products/visual. [Accessed 18 March 2020].

[4] Qlik, "QlikView - Powerful Integrative Analytics & Dashboards," 2020. [Online]. Available: https://www.qlik.com/us/products/qlikview. [Accessed 18 March 2020].

[5] S. Glen, "Variance Inflation Factor," 21 September 2015. [Online]. Available: https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/. [Accessed 22 March 2020].

[6] M. Oleszak, "Regularization: Ridge, Lasso and Elastic Net," DataCamp Inc., 12 November 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net. [Accessed 18 March 2020].

[7] C. Pascual, "Tutorial: Understanding Regression Error Metrics in Python," 26 September 2018. [Online]. Available: https://www.dataquest.io/blog/understanding-regression-error-metrics/. [Accessed 3 March 2020].

[8] J. Wesner, "MAE and RMSE — Which Metric is Better?," Medium, 23 March 2016. [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d. [Accessed 23 March 2020].