A Forecasting Face-off for Oil and Gas Spare Parts

Mahmood Serry
Bachelor of Engineering, Cairo University, 2010

and

James Vasa
Bachelor of Science, University of Illinois at Urbana Champaign, 2010

SUBMITTED TO THE PROGRAM OF SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Signature of Author: _____
Department of Supply Chain Management
May 8, 2020

Signature of Author: _____
Department of Supply Chain Management
May 8, 2020

Certified by: _____
Dr. Nima Kazemi
Postdoctoral Associate
Capstone Advisor

Accepted by: _____
Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

A Forecasting Face-off for Oil and Gas Spare Parts

by

Mahmood Serry

and

James Alan Vasa

Submitted to the Program in Supply Chain Management
on May 1, 2020 in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Spare parts demand forecasting is a key activity for asset intensive industries, but it is challenging due to the underlying demand characteristics. Demand is characterized by periods of zero demand arrivals; and the size of the order is variable with large, unexpected spikes. Schlumberger, an oil and gas service manufacturer, is facing the issue of low forecast accuracy for its spare parts, and has challenged the team to improve it. This research uses machine learning techniques to improve demand forecasting accuracy of spare parts for Schlumberger. The methodology of the research starts with classifying the parts into four classes namely: smooth; intermittent; erratic; and lumpy, Then, we apply recommended time series based on the literature for forecasting four classes. The time series forecast was then fed as features along with judgmental forecast and the demand parameters into two different machine learning algorithms, namely Classification and Regression Trees (CART) and Random Forests. Both models showed more than 75% improvement in accuracy over conventional demand forecasting methods when measured by Root Mean Squared Error. This improvement shows the potential benefit of adding human judgement as a parameter into machine learning algorithms when forecasting spare parts.

Capstone Advisor: Dr Nima Kazemi
Title: Postdoctoral Associate

ACKNOWLEDGMENTS

We would like to personally thank Dr. Nima Kazemi, who advised this project as both a friend and a mentor to us on our journey. Professor Yossi Sheffi deserves a special mention as a great assistance both personally and professionally.

-James and Mahmood

First and foremost, I would like to thank my wife for her support and sacrifice on our journey to MIT and beyond. Ellen, your encouragement and love make any challenge surmountable. I appreciate all that you have ever done. And to my children, William and Sarah, may this be proof you can do anything you set your mind to. Sarah, you gave me the best birthday present at MIT, you. Reading and playing with you, William, will always be the highlight of my day.

-James

Omayma, my wife, for her unparalleled love, support and for always believing in me. Mohammad, my son, may you achieve all your dreams and be the man you aspire to be. Mom, dad and Radwa, I could not have done this without you. Dahab, Hameed and Zyad, thank you for taking care of my family and for helping me survive all the stress being away for a year.

-Mahmood

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 RESEARCH MOTIVATION

Companies store spare parts to maintain their equipment. The unavailability of spare parts when needed may harm business continuity, particularly in asset intensive industries that rely heavily on asset availability, such as oil and gas and cement. Spare parts have several unique challenges. First, spare parts demand is often intermittent, with many periods having zero demand arrivals. In addition, demand sizes are often highly variable, or 'erratic'. A large proportion of spare parts are accompanied by both intermittent and erratic demand, which makes the demand 'lumpy' (Boylan & Syntetos, 2010). Second, companies often store a large variety of spare parts families, which makes it hard to identify individual inventory policies for different spares (Guvenir & Erel, 1998).

In the case where actual demand is higher than forecasted demand, the company will experience stockouts which have an immediate short-term revenue loss, interrupted operations, and potentially customer loss (Suryapranata, 2003). Just as there are risks, in over-forecasting demand, there are risks in under-forecasting. If realized demand is less than the forecasted demand, the company will face excess inventory costs and obsolescence risk. In addition, it was estimated that spare parts annual cost is around 2.5% of the purchase cost of assets with a useful life of up to 30 years (Gallagher, Mitchke, & Rogers, 2005). Therefore, a part purchased for $100, would have spare part spend of $75 over its lifetime which means spare part consumption represents

approximately 43% of its lifetime costs. Spare parts inventory management is a challenging task for many companies, but especially so in asset intensive industries.

1.2 PROBLEM STATEMENT

Schlumberger is the sponsoring company of this project. They compete in the oil and gas services industry, an asset intensive industry that requires high availability of parts with the right quantity, in the right place and at the right time. Schlumberger is facing the challenge of low spare parts demand forecasting accuracy. Currently, Schlumberger is using a bottoms-up approach where each region provides a forecast and then the corporate aggregates the individual forecasts from the regions around the globe. Every location uses a different forecasting methodology and demand forecasting is not standardized globally. Often, they use simple techniques such as a moving average forecasting model to forecast the demand of spare parts. Although these models are simple to interpret and implement, they do not provide the accuracy needed for optimal spare parts management, mainly because the large amount of zero demand periods distort the forecast (van Wingerden, Basten, Dekker, & Rustenburg, 2014).

1.3 RESEARCH OBJECTIVE

The objective of this project is to improve the forecasting accuracy of Schlumberger spare parts using machine learning techniques. The project also seeks to compare the forecasting accuracy of the machine learning algorithms with the current forecasting practices at Schlumberger. More accurate forecasting methods enables Schlumberger to be more effective at balancing supply and demand of spare parts. The project first starts by identifying the right features for forecasting the spare part. In contrast to other studies

that only include features pertinent to demand characteristics, this project introduces two new features: 1- human forecasting, which is the forecasted demand made by Schlumberger experts for every Schlumberger's SKU and 2- time series forecasting, which is the demand forecast calculated using time series forecasting methods. We speculate that demand forecasting can be improved by combining these two demand forecasts. Once the features are built, we leverage them in Classification and Regression Tree and Random Forest models.

Our hypothesis was that a more robust forecast methodology, such as ensemble learning, would be better able to predict spare parts demand at Schlumberger and improve the forecasting accuracy. This hypothesis was tested by comparing the accuracy as measured by root mean squared error (RMSE) of the different models. A survey of existing literature shows time series forecasts, machine learning, and ensemble methods have an improved forecasting accuracy for a variety of different companies across various industries.

2 LITERATURE REVIEW

2.1 INTRODUCTION

This section reviews different categories of forecasting methodologies used for spare parts. It begins by describing forecasting time series approaches which are traditionally used to forecast spare parts. Then we briefly review forecasting with human judgment. Both of these will be used as inputs to our machine learning and ensemble models which are surveyed next. Finally, we examine a suite of metrics for measuring spare parts forecasting accuracy.

2.2 CONVENTIONAL FORECASTING METHODS

Both the complexity and importance of forecasting spare parts demand have made many researchers focus on this interesting problem. Demand forecasting is an essential step in determining the right inventory policy for individual spare parts. Conventional spare parts forecasting methods are the methods that use a time series. time series based forecasting tools assume that future demand arrivals could be predicted from previous demand patterns (Hu et al., 2018). One important consideration is that forecasts are seldom accurate and that in this capstone we are more interested in forecasting ranges than the point forecasts. In this section we will cover the time series based forecasting methods of Simple Exponential Smoothing, Croston's Method, and Syntetos Boylan Approximation that were introduced to predict the spare parts demand.

According to Boylan & Syntetos (2010), improving forecasts for spare parts can be broken down into three main steps (see Figure 1):

1. Pre-processing
2. Processing
3. Post-Processing.

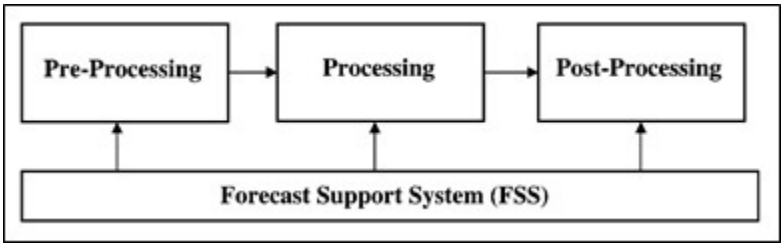

Figure 1: Forecasting Processes (Boylan & Syntetos, 2010)

The pre-processing step is the segmentation of spare parts into collective homogeneous classes with the main purpose of proposing the right forecasting tool for every class. The post processing step, as proposed by researchers, is the process of judgmental adjustment of the demand forecast by humans (Boylan & Syntetos, 2010). This is an often-overlooked step in practice but it is as important as selecting the right forecasting method. To implement such time series based methods for spares at Schlumberger, the spare parts had to go through two main steps: forecasting-based spares classification (preprocessing) and demand forecasting (processing).

2.2.1 FORECASTING BASED SPARES CLASSIFICATION

Different spare parts are accompanied by different demand patterns, which means that different spare parts require different forecasting tools (Heinecke, Syntetos, & Wang, 2013). This shows the need to segment spare parts based on the associated demand patterns. It is necessary to note that as the focus of this project is on spare part classification based on demand pattern, we exclude reviewing the classification for inventory control from the literature review. Readers can refer to Hu et al.'s 2018 review

of operations research as applied to spare parts for further discussions on these two concepts.

Segmentation based on demand patterns is different from classifying spare parts for inventory control. Product classification for demand forecasting uses the underlying demand characteristics for categorization. whereas inventory control partitions based on criticality, annual dollar usage, and other criteria. The two classification schemes are independent of each other. As the sponsoring company is most concerned with improving forecasting, we will only consider forecasted based classification and exclude classification for inventory control from this project.

Multiple spare parts classification schemes have been proposed by researchers. The first study in this line of research was carried out by Williams (1984), who introduced the notion of variance partitioning. Variance partitioning is the partitioning of the demand during lead time (DDLT) into its main parts and then classifying spare parts into “Smooth”, “Slow-moving” and “Erratic”. The study assumed that the demand arrives through a Poisson process and the cutoff values for classification are arbitrary. Johnston and Boylan (1996) bifurcated spares on the basis of how many forecast periods elapsed before demand is realized. Their empirically demonstrated threshold value was 1.25 for intermittent.

Another widely used method for categorization of spare parts demand is based on cutoff values of Average Demand Interval (P) and Square of Coefficient of Variation (CV^2) of demand sizes (Syntetos & Boylan, 2001). Figure 2 elucidates the proposed cut-off values for P and CV^2 as 1.32 and 0.49 respectively and classification of spare parts into four categories: “Smooth”, “Intermittent”, “Erratic” and “Lumpy”. Smooth

demand has relatively stable demand sizes and relatively stable occurrences of demand. For intermittent demand, the demand size is stable, but the amount of time between each order varies. Erratic demand has demand at regular intervals but the size of demand varies significantly with each instance. Lumpy demand is objectively the most challenging condition, where both the demand sizes and the demand intervals are highly variable.

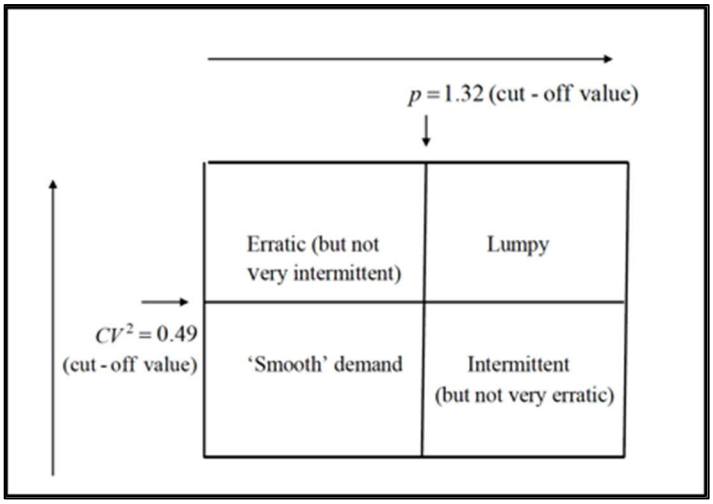

Figure 2: Cutoff Values of SKU Classes (Syntetos & Boylan, 2001)

It is worth noting here that the cutoff values and the classes proposed are mainly used to select the forecasting model for each class. Although we have found other segmentation schemes in practice, such as the one introduced by Kostenko and Hyndman (2006), we have decided to use Syntetos's classification as it was tested empirically by van Wingerden, Basten, Dekker, & Rustenburg (2014) and was shown to include the most important criteria for spare parts classification.

2.2.2 SPARE PARTS DEMAND FORECASTING

Processing is the step of applying the forecasting model to every spare part after it has been assigned a category in the preprocessing step. Classical forecasting tools such as simple exponential smoothing (SES) are shown to overestimate the demand of

intermittent items (Boylan & Syntetos, 2010). Simple exponential smoothing assigns a weight to the most current demand compared to the previous periods. Other forecasting tools were proven to provide better estimates in the case of spare parts. Perhaps the most widely used method is Croston's method (Hu et al., 2018).

Croston's method showed that SES might, sometimes, overestimate the intermittent items' demand by two times what was really demanded. He proposed a new method, that was named after him, where he used different estimates for both the demand intervals and the demand sizes (Croston, 1972). This model was proved to be more robust than moving averages and SES in the case of intermittent demand patterns. Syntetos and Boylan (2001) showed that Croston's method, although robust, is biased. To correct the bias, the authors introduced a correction in a method that is often referred to as Syntetos Boylan Approximation (SBA).

Syntetos, Boylan, & Croston (2005) compared various forecasting methods that include Croston's method and SBA on 3000 intermittent demand items from the automotive industry. After comparing the mean square errors of different forecasting tools on the various classes of spare parts, the authors suggested that Croston's method should be used for "Smooth" demand items; items with low intermittency and low variability in demand sized, and suggested to use SBA for the other three classes.

Zhou and Viswanathan (2011) examined applications of bootstrapping to spare part forecasting and found bootstrapping worked best on a simulated data set, but SBA outperformed it on an actual data set. This was validated by Synetetos, Babi, and Gardner (2015) who noted that SBA and Croston's methods perform just as well as bootstrapping without the burden of computational complexity.

The previous forecasting tools are the most used methods in practice and were proven to be superior to moving average and SES (Gutierrez, Solis, & Mukhopadhyay, 2008; Eaves and Kingsman, 2004; Synetetos, Babi, & Gardner, 2015). As SBA and Croston's method have been designed and validated to overcome the challenges of intermittency and variability inherent within spare parts, they have been successful across a wide variety of industries. Although the industries studied were not oil and gas, the underlying demand characteristics of spare parts are the same. As such, it is reasonable to hypothesize they will most likely help the sponsoring company in its challenge of improving the current forecasting accuracy.

2.3 HUMAN FORECASTS

2.3.1 PURE HUMAN FORECASTING

Whether implicit or explicit, human judgement has always played a role in forecasting. Implicitly, humans have to decide which forecasting methods to use. Explicitly, people can either rely solely on the results of a judgmental forecast, or they revise with results of the statistical forecast. Historically, judgmental forecasting has been the most common even more so the classical models (Cerullo 1975). Pure human forecasts can be classified as one manager's opinion, executive opinion, and a sales force composite opinion. A Delphi Method can also be used to anonymously converge a group of experts though blind votes on the best forecast. Even through the 1990's, the most popular method of forecasting in US corporations was judgmental forecasting because of the accuracy and challenge in getting data for the statistical methods (Sanders and Manrodt 1994). In the current era of big data having access to data and computing power is no longer an issue. Indeed, this led to a decrease in the number of forecasters using

judgement alone, and an increase in the number of forecasters combining human judgment with analytical forecasts (Fildes and Goodwin 2007).

2.3.2 HUMAN ADJUSTED FORECASTS

The combination of these methods has seen mixed results. In cases where analysts have more information than the statistical forecast, human intervention generally improves the accuracy. However, there is a danger in tinkering too much, which decreases the accuracy, because analysts either see patterns that are not there, or there are asymmetric costs for doing so (Goodwin, 1996).

While no studies specifically have addressed spare parts, researchers have looked into parts that have the same underlying demand characteristics of being slow moving and intermittent. Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin (2009) found that negative adjustments aid in accuracy more than positive adjustments, but small positive changes to a zero forecast improve accuracy. Overall, human adjustments increase forecast accuracy in the study. Although limited in scope, these results show the promise of using human judgement in intermittent, erratic, and lumpy demand situations. However, instead of adjusting after the fact, we will use the human forecast as input, and the model will learn the human's biases and hopefully correct for it in machine learning programs outlined in the next section.

2.4 MACHINE LEARNING

2.4.1 INTRODUCTION

Machine learning comprises a set of algorithms that uses data and answers to understand decision rules for a particular problem instead of having decision rules

programmed directly (Roiger and Geatz, 2003). The algorithms learn from the data to improve performance. The past decade has witnessed an advancement in machine learning and artificial intelligence due an increase in both available data and computing power to learn from that data (Bertsimas, 2017). In their review of use cases of machine learning, Chui et. al (2018) refers to demand forecasting as a promising area where machine learning has outperformed traditional time series techniques by ten to twenty percent. Machine learning algorithms listed include CART, Random Forests, and Neural Networks. Finally, we examine ensemble methods which have multiple prediction models and that merge their results.

2.4.2 CART

Classification and Regression Trees (CART) also have an independent variable and several independent variables. If the model is predicting a categorical value such as yes or no, it is called a classification tree. If the model is predicting, as in the case of forecasting, a nominal value, it is a regression tree. Regression trees will be used for forecasting demand as demand is numeric. (Roiger and Geatz, 2003).

CART determines which independent variable to split on and at which value the split occurs in such a way that minimizes range of variation within the independent variable most subject to a limit of the number cases in each bucket. Then the tree is pruned to only include splits that improve fit beyond a specified value. After splitting and pruning are complete, the average of the independent variable is calculated. To determine how this value was derived simply walk down each split of the tree (Bertsimas, 2017).

CART offers many advantages such as making no assumptions about the data and offering clarity of decision rule model structure. However, decision trees can lack

robustness, as small differences in the training set selection may cause different trees and paths through a similar tree and as a result may not perform well on future data sets (Roigers and Geatz, 2003). In spite of these limitations, Shuemli, Bruce, Yahav, and Lichtendah (2018) state CART models have demonstrated good performance in a variety of settings. A South Korean study of military spare parts illustrates the potential viability of using CART for spare parts forecasting. Moon (2013) showed that CART was only beaten by a margin of 2.8% error improvement of the best method and was more easily explained to decision makers. Consistent with this finding, a single classification and regression tree is not often by itself; instead it is used as a building block for more complex methods such as a random forest.

2.4.3 RANDOM FOREST

Random forests can be thought of as one of many extensions of the CART methodology. A random forest is composed of many different trees. To get different trees each time, replacement of the data and randomization is important. Random forests sample the data with replacement to generate unique subsets of training data for each tree. The splits of each tree are determined by a randomly selected set of the independent variables at the end, the random forest takes an average of all the different trees' predictions.

Random forests have seen mixed results when applied to the problem of forecasting. Vairagade, Logofatu, and Muharemi (2019) found random forests to have the strongest overall performance for forecasting demand for supermarket items. However, when random forests were applied to the forecasting of spare parts, the results appear less promising. Babajanivalashed, Babajanivalashedi, Baboli, Shahzad, & Tonadre (2018)

argued that random forests are outperformed by other machine learning algorithms in nearly all cases in a case study of airplane spare parts. Yet random forests have outperformed in other demand forecasting scenarios such as shoes (Kharfan & Chan, 2018). In general, it appears that the research is divided on the efficacy of random forests. This may be due to the underlying structure or quantity of data itself.

2.5 ENSEMBLE METHODS

Ensemble methods comprise different prediction techniques and combine their predictions. This is different from hybrid methods which use different models to predict different parameters of the demand, for example, use neural networks to predict demand arrivals and use simple exponential smoothing to predict demand sizes. However, ensemble learners combine different forecasts of the demand and predict the best forecasted output, for instance, combining human forecast with a time series forecast into a regression tree.

Ensemble learning for forecasting uses this same principle. In an extensive review from 2018, Perera, Hurley, Fahimnia, and Reisi (2018) recommends using human judgement in a machine learning ensemble method as an exciting new direction to explore for forecasting. The Makridakis Competition (M-Competition) is hosted by Professor Spyros Makridakis, a renowned expert in the field of forecasting. This competition is considered by many to be the premier forecasting competition in the world. It intends to find what methods of forecasting are the most accurate. In the most recent M Competition, the M4, ensemble and hybrid methods with machine learning outperformed both machines learning and traditional methods (Gilliand 2020). There were

over 100,000 different time series in this competition. However, this data set does not include low volume and intermittent demand items. The researchers are excited to see how well an ensemble learning method with a human element will perform on real data.

2.6 METRICS

There is no common performance metric for evaluating spare forecasting improvement with different studies considering different success criteria (Hu et al., 2018). Nevertheless, certain accuracy metrics do not fit. Mean Absolute Percentage Error (MAPE) is a common way to measure the forecast accuracy of non-spare parts, the first step of calculating MAPE involves subtracting the actual demand from the forecasted demand and then dividing by the actual demand. Often the actual demand for spare parts is zero, and the calculation fails as dividing by zero is undefined. Kim and Kim (2016) propose Mean Arctangent Absolute Percentage Error (MAAPE) which overcomes the limitation of MAPE by measuring slope as angle instead of a ratio. While intellectually interestingly, MAAPE as a metric has not yet gained widespread acceptance among demand forecasting practitioners (Makridakis, Spiliotis, & Assimakopoulos, 2020).

Other accuracy measures have practical concerns as well. Teunter and Duncan (2009) empirically demonstrate a forecast of all zeroes resulting in the best Mean Absolute Error and Mean Absolute Scaled Error. Such a forecast would be meaningless in application. After discussions with the sponsoring company, we decided to use the root mean square error (RMSE). This was mainly because RMSE does not have any of the aforementioned inherent limitations and has a huge impact on inventory policies,

because it is taken as a proxy for the standard deviation of demand, which is a factor in calculating the safety stock.

2.7 CONCLUSION

We have reviewed the different methods proposed for forecasting spare parts. Conventional models such as Croston's methods and SBA are widely accepted to perform better than moving average and SES models, particularly when used to forecast the demand of spare parts with lumpy demand patterns. In order to better define 'perform better', we looked at different metrics to measure improvements in forecasting for spare parts.

Non-conventional methods such as machine learning models and ensemble learning models have all started to gain researchers' attention in the past decade. All papers we reviewed have shown non-conventional methods have potential to show improvements over Croston's and SBA models.

In this project we will study the performance of both conventional and non-conventional models on Schlumberger's real data. Based on the results, we will be able to advise the company on which forecasting models to be used for the different classes of spare parts. According to our review of the literature, we are the first researchers to use an ensemble method of combining a human forecast with traditional time series methods into a machine learning algorithm.

3 DATA AND METHODOLOGY

The following section expounds both the data and the methodology of the project. The Data subsection begins with a synopsis of the data extracts provided by Schlumberger. Each of the extracts needed to be cleaned and given structure to make it meaningful. We start with the demand-related extracts where we review the different demand data sources, challenges, and global aggregation steps. Next, we review the data cleaning for human forecasts. Here, we discuss details on how we mapped job count forecasts to individual spare parts. Finally, we summarize the different variables in the cleaned datasets which concludes the data subsection.

After the data subsection concludes, we introduce how we used the data in the methodology subsection. In this section, we explicate the modeling steps of data preprocessing through demand classification and processing through time series forecasting. Then we combine time series forecasts and human forecasts into machine learning and ensemble learning algorithms. We conclude with a discussion of which metrics should be used to best compare our methods to each other and to the base case of time series forecasting.

3.1 DATA OVERVIEW

The data used in this study is the data related to the parts used by Schlumberger's Drilling & Measurement (D&M) business unit in North America, Latin America and the Middle East over the period of the past 5 years. Schlumberger currently operates in 85 countries and it works with every major international oil company, and directly for most of the petrostates – including Saudi Arabia, Libya, Russia and Turkmenistan. The

sponsoring company provided us with a sample of 30K spare parts from across the D&M network.

The actual spare part demand process flow begins when a field technician orders a spare from the warehouse. Schlumberger records this transaction and calls it consumption. Consumption alone would understate demand. Consider the case where ten parts are stocked and ten are ordered, and the next day ten are ordered again. By itself, consumption would say aggregate demand is ten, but aggregate demand is actually twenty. Schlumberger tracks these instances and aggregates them in a backlog report. Finally, the field technicians can over order relative to what they actually use, and they return unneeded spares back to the warehouse. These are called returns. Forecasted spare part demand would be higher without accounting for these returns. We all agreed to express demand as:

Monthly Demand = Monthly Consumption + Monthly Backlogs - Monthly Returns

Equation 1: Components of Demand

Where **consumption** is what the technicians consumed at the different sites, **backlogs** are the parts that were required but were not available in the warehouse and **returns** are spare parts that the technicians return after they are issued.

This demand data was used to create a conventional time series forecast. In addition to the time series forecast, we used a proxy for Schlumberger's human forecast. The human forecast data provided from the company was in terms of forecasted job counts. Job counts are the number of jobs forecasted over a given time horizon. Jobs correspond to a certain number of field hours that a part is in active use. This forecast was provided at a product family level, which is a group of similar equipment that spare parts support.

In order to match the forecasted job counts with the individual spare parts, the company provided a mapping file which matched the part family with its spares.

During the project, we received 51 files from Schlumberger. Table 1 summarizes the number of files for every category:

Category	Number of files	Total number of records
Consumption and returns	8	3.8M
Backlog	35	700K
Human forecast	7	2.3M
Parts to Family mapping	1	34K

Table 1: Summary of the Data Received

Throughout this research process, we use R as the main software for data analysis and for running machine learning models.

3.1.1 DATA CLEANING FOR CONSUMPTION

The data collected for demand required significant cleaning and manipulation before being ready for analysis. Consumption data came from different systems across North America, Latin America and the Middle East. Every computer system treated consumption in a different way; thus the files that we received were inconsistent and had different formats. This made it data cleaning and aggregation difficult and challenging. Some of the challenges are as the following: During the five year time horizon, the sponsoring company used two different versions of SAP, namely legacy SAP and SAP ITT. Data from legacy SAP was from 2016 to 2019, and SAP ITT contained information

from 2019. There were many differences in the extracted formats that had to be manipulated. For instance, in some files consumption was positive and returns were negative, but in the others, this was reversed. The researchers had extracts of both files and performed cleaning and manipulation in R.

Of special note was the consumption unit of the different materials. We observed different consumption units such as meter, piece, or kilogram used frequently across the dataset for the same item. Some were easy to fix like when there were instances of some consumption in inches, and others in meters. Others were thornier. Items could be measured in discrete units or could be measured in eaches. When these conflicts arose, we checked which one was most commonly used, and converted it to be the best of their ability using common sense. Finally, if they were unable to resolve together, the data was flagged and taken out of the analysis. This was important to get accurate results due to the sparseness of data.

Some spare parts data had to be dropped from the analysis as they did not allow for time series forecasting or demand classification. For instance, we removed all parts that had less than two demand arrivals during the past 5 years. We did so because, for demand classification, we would need to calculate the average demand interval to determine the variability between demand arrivals. If there is only one demand arrival, we cannot calculate the gap as there would be no interval, only a point. Another challenge we encountered and worth mentioning here was the amount of data to be analyzed. Although it seemed trivial to have consumption data for 60 months, the original consumption data file contained data for only nonzero demand in daily buckets. As a result, we had to build a complete time series across all parts for all days over the past 5

years, once completed, the number of rows for the complete time series reached 134M rows from the initial 3.8M rows.

3.1.2 DATA CLEANING FOR BACKLOGS AND RETURNS

As to the backlog data, they were provided in an Excel format. The main challenge of the backlog report was that the number of backlogged parts for each month was not recorded. We noticed that instead the company uses the rolling backlog technique, which accounts for the accumulated backlogs over time. For example, if there was a backlog of two pieces in January, and the backlog was not still fulfilled in February, the backlog will show two in February's report as well, but this would be a double count. If the demand from January of two pieces was filled in February, but there was a new backlog in February, it would also show two pieces in backlog, but would not be double counted. To resolve this issue, the data was transformed so we could capture unique instances of each backlog.

Other than consumption and return, the return data was more straightforward. The return data was recorded daily in the received data and we compiled the daily returns into monthly buckets within each file. in the original files, certain systems used positives to denote returns, and others used negatives. When we merged the files, we ensured all returns had the same polarity as different systems had different conventions.

3.1.3 DATA CLEANING AND MANIPULATION FOR HUMAN FORECAST

Humans are involved in the forecasting process at a high level in the sponsoring company. A planner will look ahead and forecast the material requirements within the planning horizon based on job counts. Job counts represent how many jobs a tool family

will be in active service over a given time. An actual job count is also calculated after the fact.

Due to the unpredictable nature of the oil and gas market, the planner cannot use time series data as the past may not accurately model the future. The planner needs to consider many factors such as: number of open contracts; how many new customers are coming online; and the projected future prices. Given such complexity and such a large number of spares, human forecasting takes place at a less granular level.

The main challenge was the sponsoring company provided the forecast for the tool family level instead of the specific spare part level. Spare parts are replacements parts for a specific tool. A group of similar tools is a tool family. We used a company provided mapping tool which matched the spare to its tool family. In the instances when a part was shared across tool families, we consulted with the company and agreed to drop the parts. An additional consideration was that certain spares are not assigned to a tool family. These parts were assigned the aggregate plan for the category and noted as such in the analysis.

In addition to the above, over time the way the parts are serviced has changed. In certain cases, spare parts can be bundled into a group called a kit, and the kit is assigned a part number. During the time horizon, certain parts moved into a kit and some out of a kit and into individual service. For these and others, the part numbers changed. The researchers used the mapping tool to guarantee the demand for the old part numbers were matched to the new part numbers thus ensuring appropriate continuity.

3.2 FINAL DATA AFTER CLEANING & AGGREGATION

Table 2 and Table 3 explain the final datasets obtained after cleaning, manipulating and aggregating the data to be ready for modeling and analysis.

Demand	
Column	**Explanation**
Month	Demand month
Year	Demand year
Material	Spare part number
Demand	Aggregate global demand, including consumption, backlogs and returns

Table 2: The summary of demand data set after cleaning

Human Forecast	
Column	**Explanation**
Month	Demand month
Material	Spare part number
Tool Family	Group of similar finished product
Forecasted Job Count	Aggregate global forecasted job counts for spare parts

Table 3: The summary of human forecast data set after cleaning

3.3 METHODOLOGY

In this section, we will discuss the methodology. The section is organized as follows: We begin by preprocessing the created time series, then we process using Syntetos Boylan Approximation which forecasts the demand. Next, we compile the human forecast. Finally, we combine each part's category, time series forecast, and human forecast into various ensemble learners. The result of these models was compared to each other and to the basic time series model using Root Mean Square Errors. This flow is shown in Figure 3.

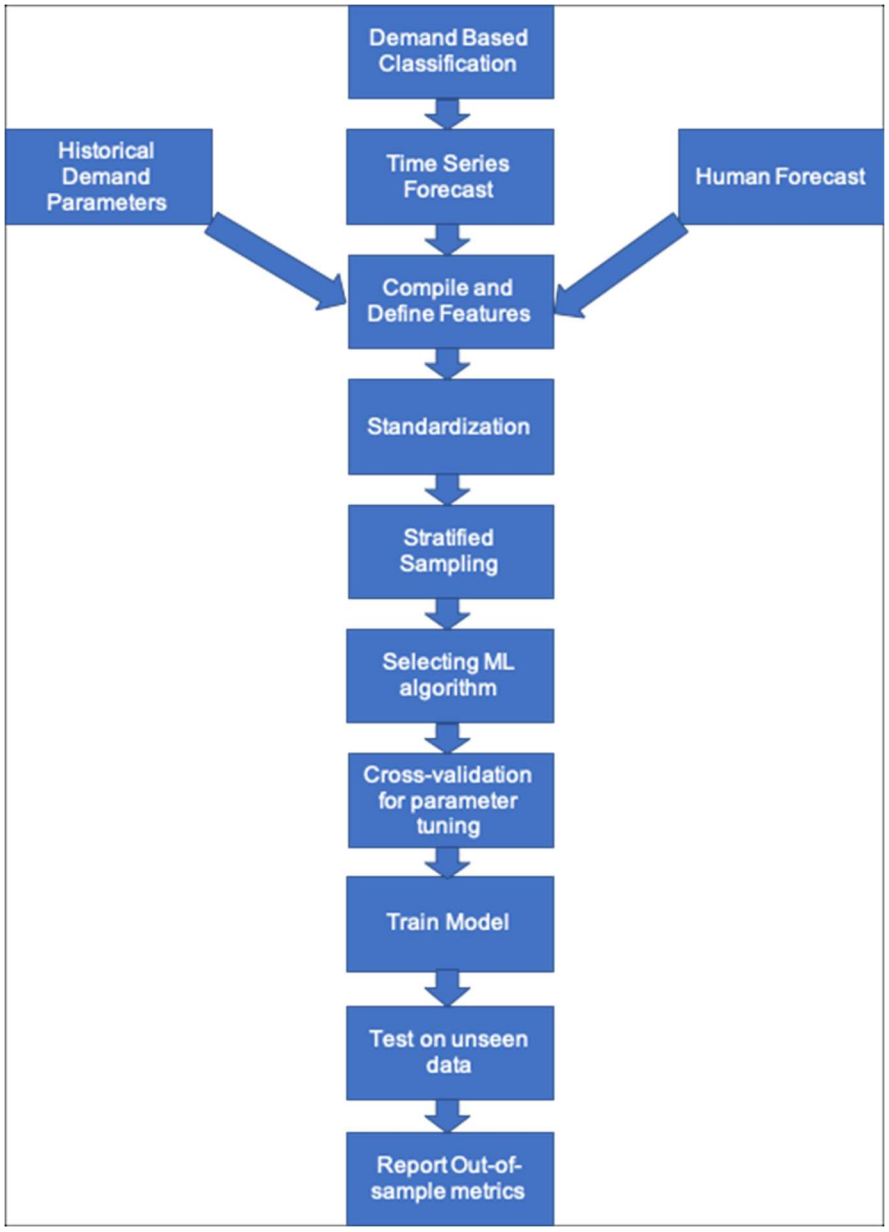

Figure 3: Overview of Methodology

3.3.1 DEMAND BASED CLASSIFICATION & HISTORICAL PARAMETERS

Different spare parts have different demand patterns. Logically, the forecast to approximate these demand patterns should be different as well. In order to define the best-fit time series forecasting method for each item, spare parts were segmented based on the Average Demand Interval (P) and the Coefficient of Variation (CV^2) with the cutoff

values proposed by Syntetos et al. (2005). Though a variety of different methods of cross-cutting the data have been proposed, these cut-off values are the current standard in industry and academia (van Wingerden, Basten, Dekker, & Rustenburg, 2014). By the end of the preprocessing segmentation, every spare part was classified as either Smooth, Intermittent, Lumpy, or Erratic. The accompanied demand parameters, namely Average Demand Interval and Coefficient of Variation of demand size, were used later in the ensemble model as parameters of the historical demand.

3.3.2 TIME SERIES FORECAST

After classifying the different parts into the different classes in preprocessing, we processed the data with time series demand forecasting for the different classes. Croston's method is the most common forecasting method for spare parts with intermittent demand, but Syntetos and Boylan (2001) have shown that the method is biased for nonsmoothed demand. This project used Syntetos and Boylan Approximation method (SBA), which corrected for the bias in Croston's method. For items in the smooth class, the researchers used simple exponential smoothing. For items in the lumpy, erratic, and intermittent categories, we used SBA with the appropriate smoothing constants. The root mean square error of the forecast was used for estimation of the standard deviation of demand over lead time.

3.3.3 HUMAN FORECAST

Forecasted job count data collected from the company which was mapped to the different spare parts in the data cleaning section acted as a proxy for the human forecast. We believed that a better human forecast would have been the aggregate forecast from all locations for all the parts under study; however, the company started aggregating part-

based human forecasts one year before the project and the amount of the data was not enough to include in the machine learning algorithm.

Job counts were used for human forecasts, but the challenge was that there were many forecasts snapshots for every month. This meant that every month, the planners revised their job counts forecast for the next year. After several discussions with the stakeholders, we agreed to utilize a 6-month look ahead forecast; that is for every month, we used the snapshot 6 months earlier. The 6-month look ahead window was justified that although forecasting for a shorter look ahead will be more accurate, the company's policy was to use a 6-month look ahead forecast in order to use in its inventory control system.

Due to limited data availability from the company, we received 24 monthly snapshots for the job counts, and given the 6-month look ahead agreement, we were left with only 18 months of human forecast for every part. We believed that if more data had been provided, the machine learning algorithms would have performed better.

3.3.4 COMPILE DATA AND DEFINE FEATURES

In our ensemble models, the demand is the dependent variable that we are trying to predict and the following features (independent variables) were selected to train and test the model as demonstrated in Table 4:

Feature (Independent Variable)	Description
Time Series Forecast	Forecast generated through traditional methods. For example, Simple Exponential Smoothing and SBA.
Human Forecast	Forecasted job counts provided by the company
Spare Parts Class	Class of spare parts based on the demand characteristics
Average Demand Interval	Average number of months between two demand arrivals
Square of Coefficient of Variability of demand sizes	Square of Coefficient of Variability of demand sizes for periods with realized demand

Table 4: Compiled Features

3.3.5 STANDARDIZATION

Different spares had different unit of measures. For example, some parts are measured in pieces while others are measures in meters or inches. Also, spare parts consumption has wide range of demand points with many outliers in place. This problem might distort predictions in the machine learning algorithms. Before sampling and training the machine learning algorithms, we standardized the numerical features to have a mean of zero and a standard deviation of 1.

3.3.6 STRATIFIED SAMPLING

The final dataset was split and stratified on the demand instances into two data se

1. training dataset: constitutes 80% of the data
2. testing dataset: constitutes 20% of the data

The reason for the split was to measure the model accuracy on unseen data. In our ensemble models, we used the training data set to train our model on the inherent complexities of the data and then measured its performance on the test data set. As such, we reported the RMSE and R-Squared values of the test data. RMSE measures the accuracy of the forecast. R-Squared measures how the well the model fits the data.

Typically, test data usually has a lower R-Squared or goodness of fit than the training data. Reporting the test data performance is more indicative of the model's predictive value.

3.3.7 SELECTION OF MACHINE LEARNING ALGORITHMS

Two machine learning algorithms were used and compared to the base model in order to test our hypothesis. They two algorithms were:

1. Classification and Regression Trees (CART)
2. Random Forests

3.3.8 CROSS VALIDATION FOR TUNING THE PARAMETERS

Cross-validation is a way to tune parameters to arrive at a model with the best predictive power. Cross-validation builds a model for each value of a parameter, and then chooses the parameter value that yields the best out of sample accuracy. This ensured that the model learned correctly and the results were generalizable.

In our models we used cross-validations with 10 folds to tune different parameters for CART and random forest models. For the CART model, we tuned the complexity parameter, and for random forests we tuned the number of trees in the forest as these values control the learning process for their respective algorithm.

3.3.9 Comparisons and Results

In order to achieve this aim, we had to define a single metric to compare different models. We compared the different models based on the RMSE, and the model with the lowest RMSE was recommended to the company. The main challenge in using RMSE was that RMSE was calculated for every part by itself. However, the company required the model to have one number to compare the different models, so we decided to use an

aggregate RMSE across all the parts as a proxy for the aggregate accuracy measure. We believed that as long the metric is homogeneous across all the models, any improvement from a model would have been reflected in the same metric. We acknowledge a forecast bias metric which calculates if there is any consistent difference between the forecast and the actual demand, would be valuable, but Schlumberger requested only an accuracy measure.

4 RESULTS AND ANALYSIS

4.1 INTRODUCTION

In this section, we present the parameters and results of the conventional methods as compared to the machine learning and ensemble methods.

4.2 CONVENTIONAL METHODS RESULTS

4.2.1 DEMAND BASED SPARE PARTS CLASSIFICATIONS

The outputs in the figures below were the classification outputs from R. They show that most of the spares under study were either lumpy or intermittent. Figure 4 shows the classification counts in the classic Syntetos and Boylan matrix view, while Figure 5 depicts the same information in an easy to compare bar chart.

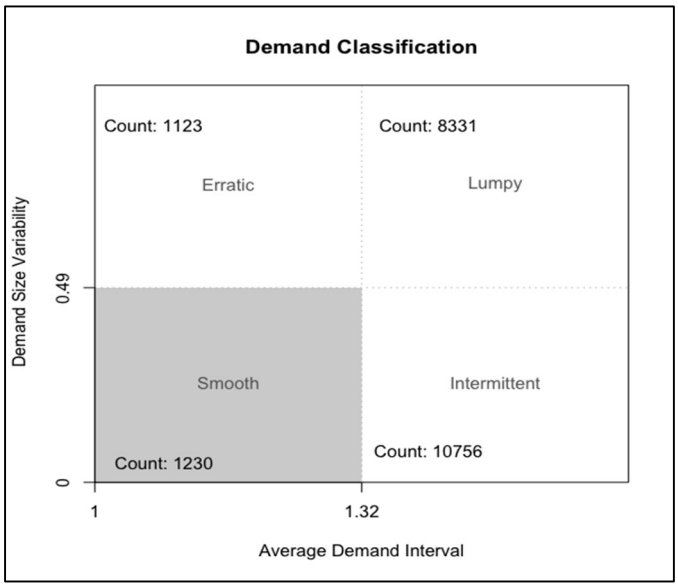

Figure 4: SBA Classification Matrix View

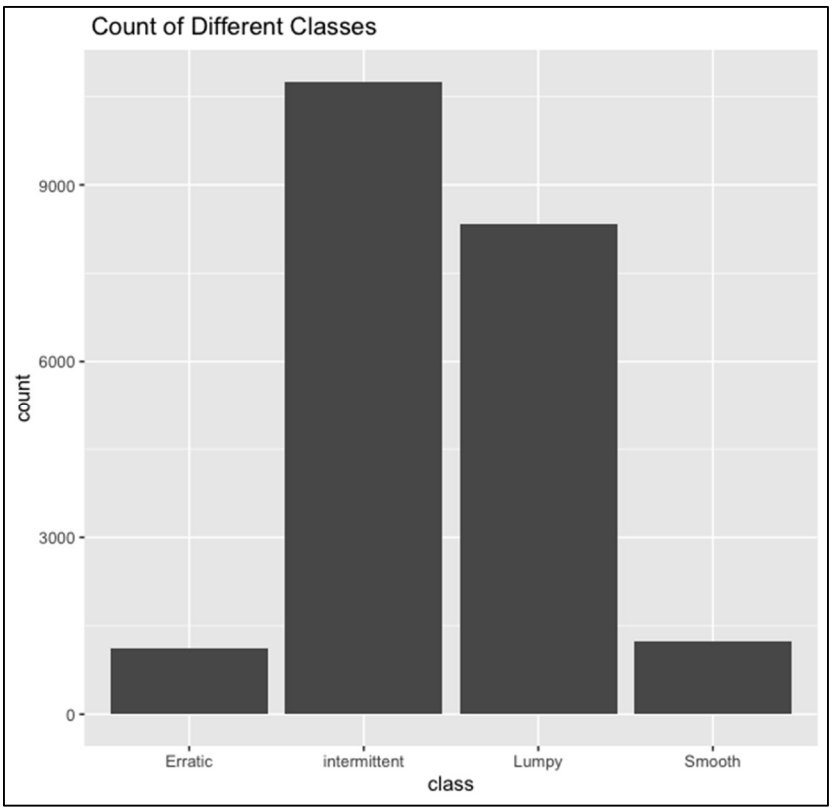

Figure 5: SBA Classification Bar Chart View

4.2.2 TIME SERIES FORECASTING

Simple Exponential Smoothing and SBA methods were used to forecast the rolling demand for the parts under study and the RMSE for every part was reported. As there were 60 globally aggregate demand data points but only 18 data points for human forecasts, we split the data set up into 42 months (for training) and 18 months (for testing). Doing so made the demand and human forecasts points the same in terms of records. We then used 42 months to train different time series models before forecasting. The main purpose of the training was to identify the best in-sample smoothing parameter. Finally, we used the trained model to forecast the demand for every month ahead in an incremental manner. We did that in order to test the models on unseen data. Demand is shown in straight lines and time series forecast is shown in dashed lines for representative

parts for each of the quadrants of SBA in Figure 6. Starting with the lower left and moving clockwise, there is smooth, erratic, lumpy, and intermittent.

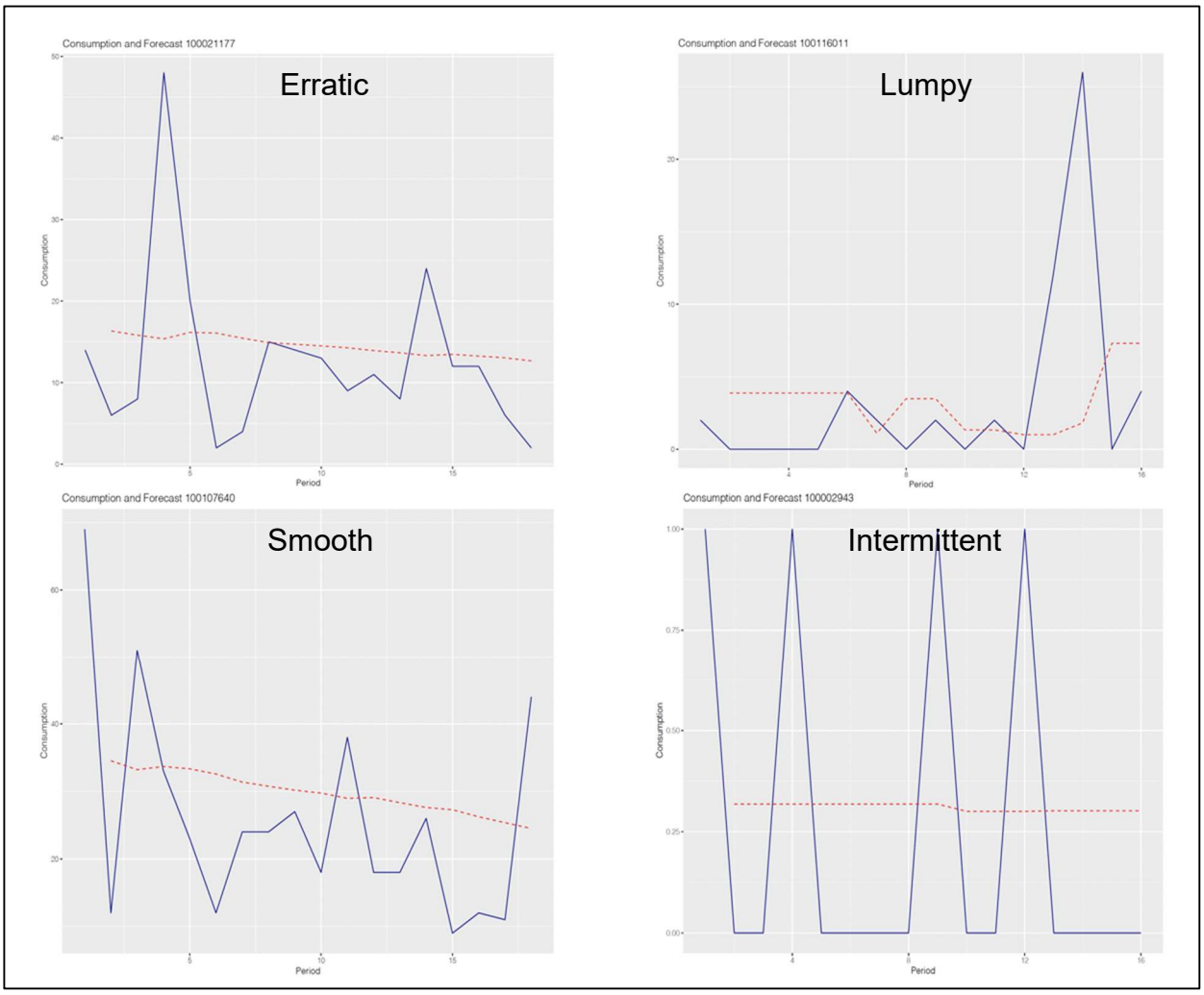

Figure 6: Time Series Forecast and Actual Demand

We can see from the above charts that the forecast is missing the lumpiness or overshoots of the parts demand. However, it is noted that as the forecast is calm then it will have a better impact on inventory policies and positions. In the case of the smooth demand in the lower left corner of Figure 6, Simple Exponential Smoothing method was used. The stability of demand is reflected in the relative stability of its forecast. As one moves clockwise, the erratic demand category has a demand every period, but the size of the demand in each period varies considerably. The forecast mirrors this. For the lumpy

demand, there are a large number of periods with no demand and when the demand arrives its size is not consistent. The forecast lags the high and low spikes as it updates after the fact. In the case of intermittent demand, the forecast is relatively constant. As shown in the chart, demand size variability is low, but the amount of time between each instance is high. SBA assumes demand is equally likely in each period. As a result of these two factors, the forecast is relatively stable. The above models had an aggregate RMSE 316.77.

4.3 ENSEMBLE LEARNING RESULTS

As discussed before, we used machine learning algorithms to learn from both time series and human forecasts. However, before building the machine learning algorithms, we needed to build a data frame that includes all the independent variables and the dependent variable (Demand). Table 5 shows the structure of the final data frame that we used in all the machine learning algorithms.

Variable	Explanation
Demand	Aggregated monthly demand to be forecasted
Time Series Forecast	Forecast generated through traditional methods. For example, Croston and SBA.
Human Forecast	Forecasted job counts provided by the company
Spare parts class	Class of spare parts based on the demand characteristics
Average Demand Interval	Average number of months between two demand arrivals
Square of Coefficient of Variability of demand sizes	Square of Coefficient of Variability of demand sizes for periods with realized demand

Table 5: Final Data Frame

4.3.1 CLASSIFICATION AND REGRESSION TREES (CART)

The first machine learning algorithm we test is the CART model, which splits the tree to minimize impurity at the leaf nodes. This process is controlled by the minbucket parameter and the complexity parameter. The tree grows and splits as long the minimum number of leaves at the end of a split (minbucket) is satisfied. The resulting tree is then pruned using the complexity parameter, which cuts splits that do not improve R-squared by at least the complexity parameter value. Logically, then as the complexity parameter gets larger, the fewer splits are in the resulting tree. We used 10-fold cross validation to tune the complexity parameter of the tree as it controls the final shape more so than the minbucket. Figure 7 shows the output for the cross validation and the selected complexity parameter was 0.001, as this was the parameter with the highest R-Squared.

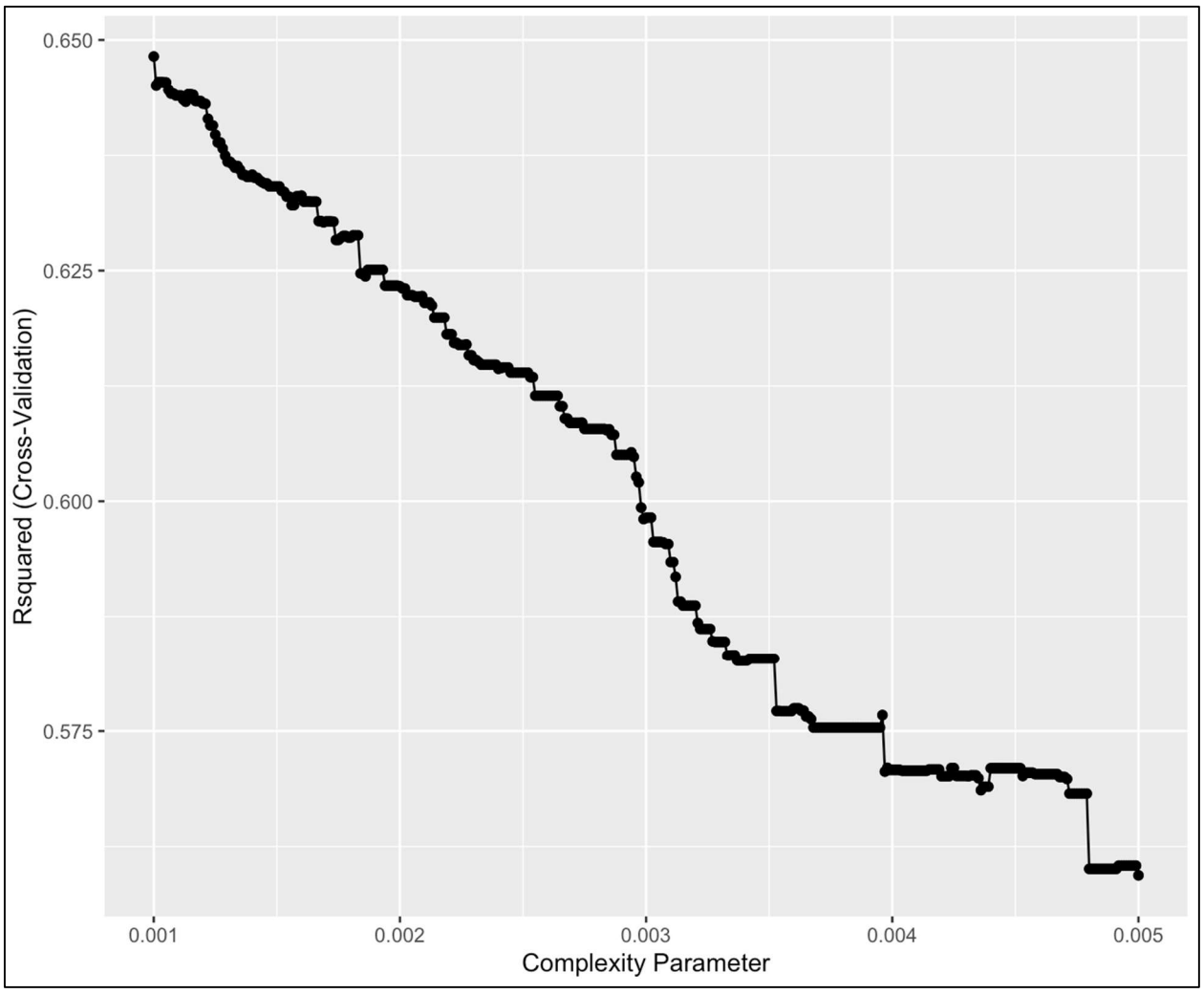

Figure 7: Cross Validation of Complexity Parameter

The dependent variable for the CART model was Demand. The independent variables used were: Time Series Forecast; Human Forecast; Class; Average Demand Interval, and Coefficient of Variability of Demand Sizes. Figure 8 shows both the independent variable split on and the associated value of the split in the CART model. This resulting model had an R-squared of 0.609 and RMSE of 74.9. There is a significant improvement over the time series forecast in terms of RMSE. The CART model improves on the time series model with 76% reduction from 316 to 74.9.

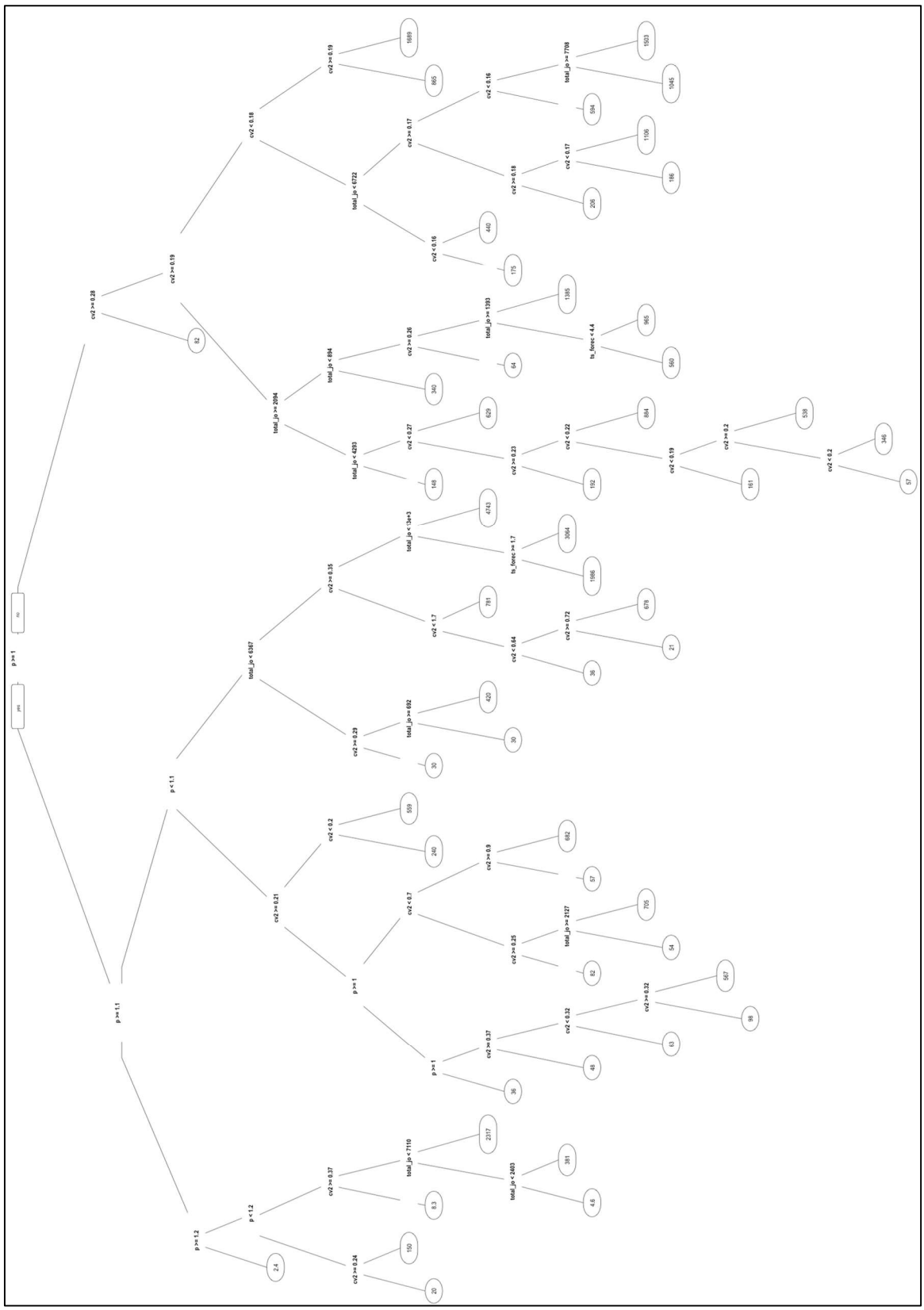

Figure 8: CART Dendrogram

4.3.2 RANDOM FORESTS

Below we test a random forest model and compare it to the previous model. There are three main parameters in random forests in this package in R. Ntree is the number of trees in the random forest. We used the default value of five hundred trees beyond which have increasingly marginal returns. For each tree in the random forest, we specify a minbucket called nodesize. To be consistent with best practice for regression, we used a default value of one. Also, for each tree, we consider only a certain number of variables from the original set to split on; this is represented by mtry. However, before applying the algorithm we executed cross validation to tune the number of variables examined at each split of the tree within the random forest as shown in Figure 9. The random forest package in R states that the model is most sensitive to this parameter. There appears to be a sweet spot of nine variables per split.

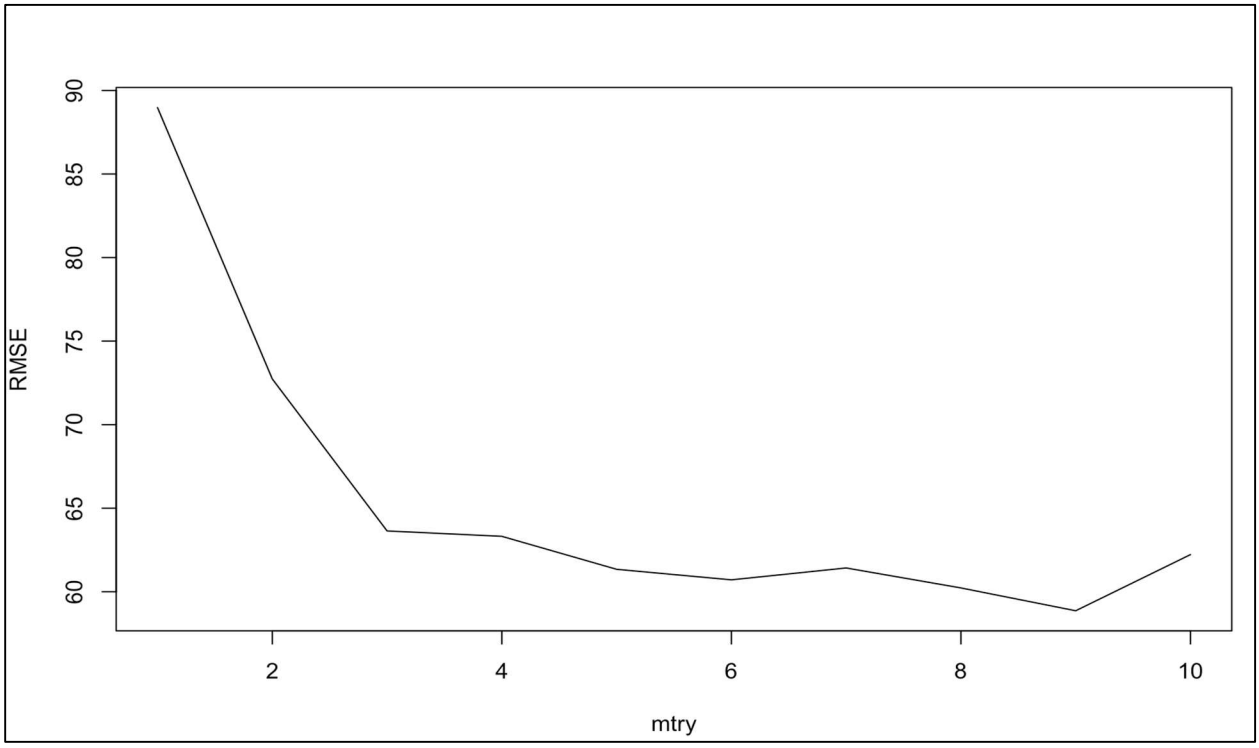

Figure 9: Tuning of mTry

For the random forest, the dependent variable is the Demand. The independent variables were identical to the CART model: Time Series Forecast; Human Forecast; class; Average Demand Interval; and Coefficient of Variability of Demand Sizes. Parameters for the model were: 80 trees for the number of trees in the forest, 9 variables per split; and a minimum of 20 observations in each terminal node. The tuned model outputted a R-Squared of 0.736 and an RMSE of 61.5

This is the best model tested during the project, yet it is more complicated one than the CART method. The R-squared is 0.736, which shows an improvement over the CART model in terms of fitting the data. Similarly, the RMSE improved from the CART model with a change from 74.9 to 61.5.

An advantage of the random forest model is we can see which variables improve the prediction accuracy the most when averaged across all trees in the forest. Figure 10 shows the importance of variables based on the random forest. One can see that the coefficient of variability of demand sizes is the most important variable in determining the forecasted demand. Note that the order is more important and interpretable than the value of the metric itself.

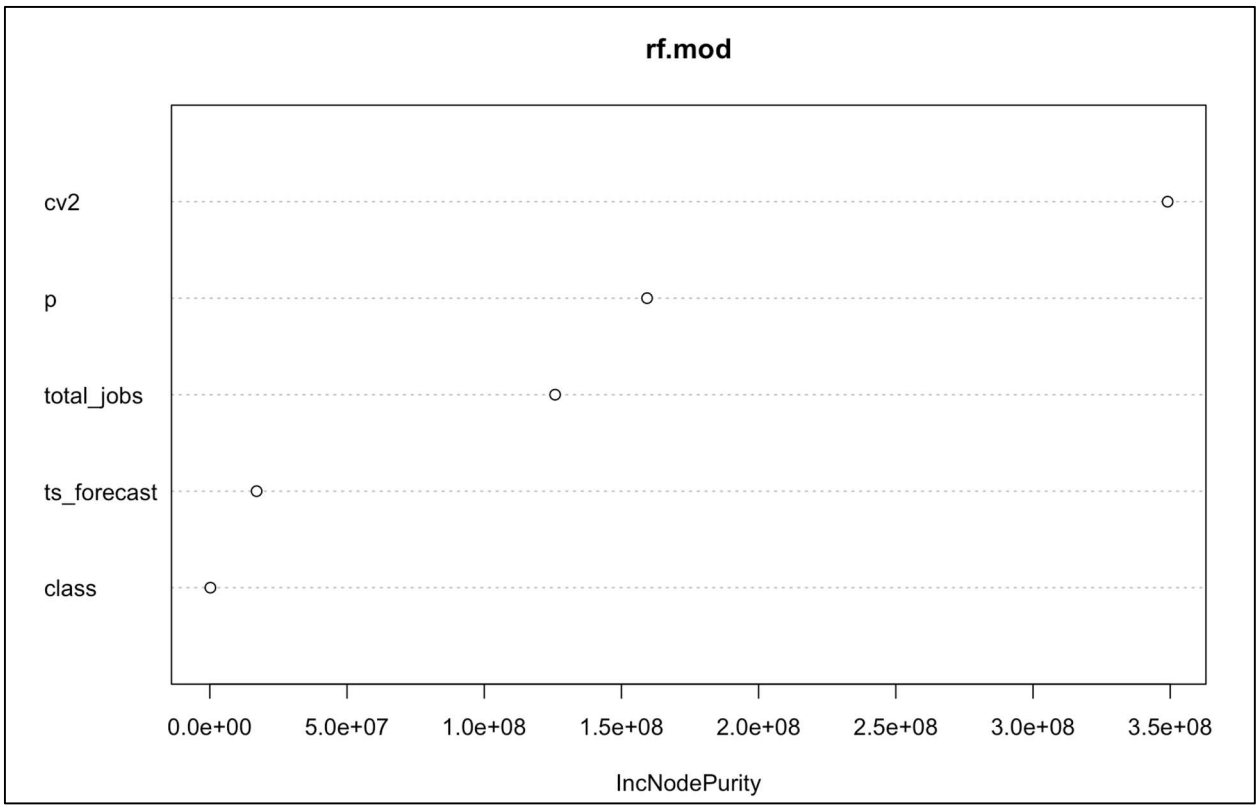

Figure 10: Random Forest Ordering of Variable Importance

4.4 COMPARISON OF RESULTS

Overall, both the CART model and the random forest model which incorporated the human forecast show a marked improvement over the time series model alone when measured by RMSE. RMSE is a measure of forecast accuracy where lower is better. However, the goodness of fit as measured by R-squared is not particularly strong, but the problem is very difficult as we are predicting the value of the demand if it occurs at all. R-squared measures how well the model fits the data as given by out of sample performance. When it is closer to 1, this is better. This can be seen in Table 6. It also appears the added complexity of the random forest resulted in only a marginal

improvement to the forecasting accuracy as given by RMSE and out of sample goodness of fit as given by R-squared when compared to the CART model.

Model #	Model Type	Time Series	Human Forecast	Class	P	CV^2	R^2	RMSE
1	SBA & SES	x						316.77
2	CART	x	x	x	x	x	0.609	74.9
3	Random Forest	x	x	x	x	x	0.736	61.5

Table 6: Comparison of Models Head to Head

5 DISCUSSION

In this project, we investigated multiple models to improve the forecasting accuracy for the sponsoring company. We analyzed the demand patterns of more than 30K SKU across thousands of locations and compared conventional forecasting models with new ensemble models that incorporate human forecasts and conventional methods into different machine learning algorithms. In this section we summarize the challenges faced, discussions of the results, recommendations, and project proposals for the sponsoring company.

5.1 CHALLENGES AND IMPLICATIONS

1. Non-standardized database systems and different treatment for demand across the company's locations => Challenging to implement standard process
2. Massive amount of data which constituted more than 80% of the time for cleaning and manipulation => Hard for company to implement a change in the way they forecast
3. Substantial amount of parts is not classified into respective tool families => Lack of data means parts had to give an average value; more data would mean better forecasts
4. Human forecasts are provided in terms of jobs for tool families and not for single parts => More granular data could improve what parts are actually demanded
5. Low number of snapshots for job count forecasts => Reduced data means less learning for the algorithms

5.2 FUTURE DIRECTIONS

We have shown that incorporating time series forecasts with human forecasts and other demand parameters into machine learning models is superior to time series forecasting methods by themselves. Our view of the results is that both time series and human forecasts are inherently biased in different directions, and we believe that machine learning and artificial intelligence could learn from these biases and provide more accurate forecasts in the area of spare parts.

The challenge in implementing such ensemble models depends on their interpretability by humans and implementability by different companies. For instance, it is simple to explain how Croston or SBA algorithms work, but it is more challenging to explain why the CART model decided on selecting these specific decision trees to define the forecast. It is even more complex to explain the Random Forest logic in determining the forecasted demand and what are the important variables to include.

Another area of thought is how current ERP and planning systems lack such machine learning algorithms in demand forecasting areas. We have seen Croston's method in some ERP systems, but we believe that including other modern forecasting models could help companies improve their inventory management policies and save millions of dollars tied up in inventory. Regardless of improvement method, a practical first step in quantifying the benefit is understanding current state through baselining the current forecast accuracy.

Since Schlumberger does not have in depth records of previous forecasts, it is impossible to compare the time series model, CART, and random forest model to it. It

would be highly advisable for them to track the forecast, so we can better measure and improve upon its performance.

5.3 RECOMMENDATIONS

5.3.1 Company Recommendation

Combining time series forecasts with human forecasts can improve forecasting accuracy massively. There is no one size fits all model for spare parts forecasting and there are financial benefits from investing in better forecasting tools, especially in an asset intensive company like Schlumberger.

In Table 1, we can see that the random forest model is the best model in terms of forecasting accuracy, yet it is a very complicated one in terms of interpretability and implementability with only marginal improvements to the CART model. We recommend implementing the CART model, as it includes interpretable decision rules for forecasting. We would also like to emphasize that the model should not be static and should be revised with new data during specific time intervals, monthly or quarterly, for example. This is mainly because it is a machine learning algorithm that needs to be retrained on real data, in order to capture the inherent bias and variability in the different forecasting methods. The model should also be tested on a sample of the company's spares before rolling it out to the whole company's family of parts.

Although the main purpose of the project was to improve the forecasting accuracy, we collected some useful data that could be shared with the company. These data could be thought of as a bi-product of our analysis. For example, we identified the spare parts that, in aggregate, did not have any movement in the past 5 years. These parts could be

at a very high risk of obsolescence and might require different management techniques. We also defined parts that are very slow moving, moving only once in the past 5 years. These parts were very hard to forecast statistically because there was not enough demand data on them and they also might require different management techniques.

The company's management should be aware that a better forecasting model by itself is not the solution for high inventory levels, although an important piece of the inventory management puzzle. Based on our discussions we defined some areas for improvements that could be tackled in future research projects:

1. **Lead Time modelling improvements project:** Currently, the company uses the quoted lead times from vendors in its inventory policy calculations. Variable lead times have an impact on safety stocks that have to be taken into consideration along with the forecasting error. We believe that future projects could help the company define better models for lead times across the whole company.
2. **Multi-echelon inventory optimization project:** The company has thousands of locations across the globe, and the quoted lead times from the different echelons in the supply chain can have an impact on the safety stock held at the different echelons. We believe that optimizing the inventory at Schlumberger based on a multi-echelon optimization framework will have a profound impact on the inventory positions performance at the company.

5.3.2 Research Recommendations

We believe that more research is required in the area of ensemble learning models for spare parts. Conventional time series forecasting methods, when combined with human judgement and machine learning algorithms could lead to improved forecasting

metrics in the area of spare parts. Some machine learning models that could be investigated in future research are Neural Networks, Deep Neural Networks, and Support Vector Machines.

6 CONCLUSION

In this project, we proposed an innovative way to incorporate human judgement into current best practices in spare parts demand forecasting. To begin, we calculated the historical demand parameters of both the Average Demand Interval and the Square of Coefficient of Variation of each SKU. Then we categorized each SKU into four classes using Syntetos and Boylan's classification. Next, we created a time series forecast for each part using the recommended conventional method in each class.

In talking with experts from the oil and gas industry inside of Schlumberger, we were convinced that there is substantial value in adding their input to the forecast as well. This is mainly driven by the challenge of predicting demand in an industry whose high volatility implies the past cannot prognosticate the future. To reconcile, we combined the judgmental forecast with the conventional time series forecast and the associated historical demand parameters in two different machine learning algorithms.

By utilizing the above procedure, we were able to show a significant improvement over the baseline Synetos and Boylan method in both the CART and random forest. The traditional metrics had a Root Mean Squared Error of 316.77. The CART model reduced these errors to 74.9 while the random forest reduced them to 61.5. This reduction is an indicator that adding human judgement into a model has merit. This corroborates Franses and Legerstee (2012), who state the best models formally incorporate past expert performance explicitly along with traditional statistical forecasting at the SKU level so they can improve a poor performing initial model. However, we have used machine learning

to model this interaction and they have specified to model it within a time series forecast only.

We recommend Schlumberger run a pilot study using the CART model and measure its performance in terms of both accuracy and precision. We recommend the CART model for ease of interpretation and implementability. CART's decision rules are transparent which makes explaining to management easier. Random forests typically offer better performance but whose inner workings are not easily explainable. In our case, the random forest only had marginal improvement over CART, so we recommend CART. Furthermore, the model will need to be rerun every month so it can continue to learn, and the CART model is easier to implement with the existing tools at Schlumberger.

The search to improve spare parts forecasting accuracy will continue to be an active and challenging area for both academia and industry. We believe the approach in this capstone shows the potential of using human judgement to improve spare parts demand forecasting in machine learning.

REFERENCES

Boylan, J. E., & Syntetos, A. A. (2010). Spare parts management: A review of forecasting research and extensions. IMA Journal of Management Mathematics, 21(3), 227–237.

Babajanivalashedi, Rez, Baboli, Armand, Shahzad, Muhammad Kashif, & Tonadre, Romy. (2018). A Predictive Approach to Define the Best Forecasting Method for Spare Parts: A Case Study in Business Aircrafts' Industry. 773–778.

Bertsimas, Dimitris. 15.071 The Analytics Edge. Spring 2017. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Kharfan, M., & Chan, V. (2018). Forecasting Seasonal Footwear Demand Using Machine Learning. Massachusetts Institute of Technology

Chui, M, Manyika, J., Miremadi, M., Henke, C., Chung, L., & Valley, S. (2018). Notes from the AI frontier: Insights from hundreds of use cases (p. 32) McKinsey Global Institute.

Cerullo, Michael. (1975). Sales forecasting practices: A survey. Managerial Planning, 24(5), 33–39.

Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. Journal of the Operational Research Society, 23(3), 289–303.

Eaves, A., & Kingsman, B. (2004). Forecasting for the ordering and stockholding of spare parts. Journal of the Operational Research Society, 55(4), 431–437

Fildes, Robert, & Goodwin, Paul. (2007). Against your better judgment? How Organizations can improve their use of management judgement in forecasting. Interfaces, 37(6), 570–576.

Franses, Philip Hans, & Legerstee, Rianne. (2013). Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? International Journal of Forecasting, 29(1), 80–87.

Gallagher, T., Mitchke, M. D., & Rogers, M. C. (2005). Profiting from spare parts. The McKinsey Quarterly, February 2005

Gilliland, Michael. (2020). The Value Added by Machine Learning Approaches in Forecasting. International Journal of Forecasting, 36(1), 161–166.

Goodwin, Paul. (1996). Statistical correction of judgmental point and forecasts. International Journal of Forecasting, 24(5), 85–99.

Gutierrez, R. S., Solis, A. O., & Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. International Journal of Production Economics, 111(2), 409–420.

Guvenir, H. A., & Erel, E. (1998). Multicriteria inventory classification using a genetic algorithm. European Journal of Operational Research, 105(1), 29–37.

Heinecke, G., Syntetos, A. a., & Wang, W. (2013). Forecasting-based SKU classification. International Journal of Production Economics, 143(2), 455–462.

Hu, Q., Boylan, J. E., Chen, H., & Labib, A. (2018). OR in spare parts management: A review. European Journal of Operational Research, 266(2), 395–414.

Johnston, F., & Boylan, J. (1996). Forecasting for Items with Intermittent Demand. The Journal of the Operational Research Society, 47(1), 113-121.

Kim, Sungil, & Kim, Heeyoung. (2016). A new metric of absolute percentage error for intermittent demand forecasts. International Journal of Forecasting, 32(3), 669–679.

Kostenko, A. V., & Hyndman, R. J. (2006). A note on the categorization of demand patterns. Journal of the Operational Research Society, 57(10), 1256–1257.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74.

Moon, S. (2013). Predicting the Performance of Forecasting Strategies for Naval Spare Parts Demand: A Machine Learning Approach. Management Science and Financial Engineering, 19(1), 1–10.

Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. European Journal of Operational Research, 274(2), 574–600.

Roiger, Richard J., & Geatz, Michael W. (2003). Data Mining A tutorial-based primer. Pearson Education.

Sanders, N. R., & Manrodt, K. B. (1994). Forecasting Practices in US Corporations: Survey Results. Interfaces, 24(2), 92–100.

Shmueli, G., Bruce, P., Yahav, I., Patel, N., & Lichtendahl, K. (2018). Data Mining for Business and Analytics Concepts, Techniques, and Applications in R. John Wiley & Sons, Inc.

Suryapranata, A. (2003). Forecasting Framework for Inventory and Sales of Short Life Span Products. Technology and Policy Analysis. Deflt University of Technology.

Syntetos, A. A., Babai, M. Z., & Gardner, E. S. (2015). Forecasting intermittent inventory demands: Simple parametric methods vs. Bootstrapping. Journal of Business Research, 68(8), 1746–1752.

Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. International Journal of Production Economics, 71(1), 457–466.

Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. Journal of the Operational Research Society, 56(5), 495–503.

Syntetos, A. A., Nikolopoulos, K., Boylan, J, Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. International Journal of Production Economics, 118(1), 72–81.

Teunter, R, & Duncan, L. (2009). Forecasting intermittent demand: A comparative study. Journal of the Operational Research Society, 60(3), 321–329.

Vairagade, Navneet, Logofatu, Doina, Leon, Florin, & Muharemi, Fitore. (2019). Demand Forecasting Using Random Forest and Artificial Neural Network for Supply Chain Management. In Nguyen, N, Chbeir, R, Exposito, E, Aniorté, P, & Trawiński, B (Eds.), Computational Collective Intelligence. (Vol. 11683). Springer, Cham.

van Wingerden, E., Basten, R. J. I., Dekker, R., & Rustenburg, W. D. (2014). More grip on inventory control through improved forecasting: A comparative study at three companies. International Journal of Production Economics, 157, 220–237.

Williams, T. M. (1984). Stock Control with Sporadic and Slow-Moving Demand. Journal of the Operational Research Society, 35(10), 939–948

Zhou, C., & Viswanathan, S. (2011). Comparison of a new bootstrapping method with parametric approaches for safety stock determination in service parts inventory systems. International Journal of Production Economics, 133(1), 481–485.